Diss. ETH No. 20918

Information Theoretic Modeling of Dynamical Systems: Estimation and Experimental Design

A dissertation submitted to
ETH Zurich

for the degree of
Doctor of Sciences

presented by
Alberto Giovanni Busetto
Dott. Mag. in Ingegneria Informatica,
Università degli Studi di Padova
born November 26, 1983 *in* Venice
citizen of the Italian Republic

accepted on the recommendation of
Prof. Dr. Joachim M. Buhmann, examiner
Prof. Dr. Manfred Morari, co-examiner
Prof. Dr. Jörg Stelling, co-examiner

2012

Abstract

Dynamical systems are mathematical models expressing cause-effect relations of time-varying phenomena. This thesis focuses on learning dynamical systems from empirical observations. Three settings are considered: unsupervised, supervised, and active learning. The unifying goal is to extract predictive information from data.

A method is introduced to cluster time-series and perform model validation. The method addresses order and model selection with the principle of Approximation Set Coding [Buhmann (2010)]. Experimental verification is performed in the context of relational clustering of temporal gene expression profiles. The results demonstrate wide applicability and consistency with the Bayesian Information Criterion. Then, discrete dynamic transitions are reconstructed from high-dimensional time-series with an unsupervised approach. The approach, based on Hidden Markov Models over Gaussian Mixtures, is applied to predict cell morphology classes from time-resolved microscopy data. Experimental validation with fluorescent markers and screening data demonstrates accurate identification of human cell phenotypes. Reported results highlight competitiveness and increased objectivity in comparison to supervised approaches based on user labeling.

In the supervised setting, clustering is employed to improve conventional particle filtering for generalized state estimation. Preventive clustering mitigates the inevitable divergence of resampling for sequential Monte Carlo methods. Supervised learning with dynamic

Bayesian Networks is employed to model human learning for effective treatment of learning disabilities. In dyslexia, the model predicts forgetting, focus and receptive states of the subject on the basis of input behavior. In dyscalculia, numerical cognition is enhanced through model-based adaptive training.

In the context of active learning, the thesis focuses on the near-optimal design of experiments for dynamical system modeling. An efficient method is introduced to select informative time points and measurable quantities. The design method is guaranteed to yield near-optimal informativeness with a polynomial number of evaluations of the objective function. The method builds on previous work on submodular active learning [Krause and Guestrin (2005)] and achieves the best possible constant approximation factor, unless P=NP [Feige (1998)]. Experimental design is applied to the reconstruction of cell signaling networks in systems biology.

The introduced contributions highlight fundamental analogies between learning and communication. In conclusion, the results demonstrate that predictive models can be built from efficient strategies of information transmission over a noisy channel. On the basis of statistical arguments, the presented results formalize and automate aspects of the hypothetico-deductive method of scientific inquiry.

Acknowledgments

"We are nothing without the work of others our predecessors, others our teachers, others our contemporaries."

— J. R. Oppenheimer

I dedicate the thesis to my wife Simonetta, my parents Patrizia & Giovanni and my brother Gianluca. They support me with love, patience, and knowledge. My sincerest gratitude goes to my grandparents Cesira, Imelde, Dobrillo and Egisto, who encouraged my studies and made them possible. Without them and their work this thesis would not have been written. My thankful thoughts are addressed to the loving memory of Cesira and Egisto. My deepest appreciation goes to all my teachers (and in particular to Antonio Rosino), for conveying a spirit of adventure in regard to research and scholarship. Among them, I am grateful to Joachim M. Buhmann for offering me the opportunity to pursue my doctoral studies. I feel honored to have been part of the Machine Learning Laboratory with Brian V. McWilliams, Cheng Soon Ong, David Balduzzi, Elias August, Gabriel Krummenacher, Kay H. Brodersen, Ludwig Busse, Manfred Claassen, Morteza Haghir Chehreghani, Peter Orbanz, and Thomas Fuchs. These are amazing persons from whom I learned a lot. I would like to thank my colleagues and friends Elias Zamora-Sillero, Francesco Dinuzzo, Irene Otero-Muras, Mikael Sunnåker, Nelido Gonzales-Segredo, Riccardo Porreca, Rajesh Ra-

maswamy, Sarvesh Dwivedi and Sotiris Dimopoulos. I had a great time interacting with them; they are a group of creative, hard-working, cheerful and helpful individuals who enriched my experience at ETH Zurich. I am particularly grateful to my collaborators A. Hauser, D. Stekhoven, G.-M. Baschera, T. C. Käser and Q. Zhong. Guidance, assistance, and insightful discussions provided by Marcus Hutter, Volker Roth, Peter Grünwald, Marcus Gross, Daniel W. Gerlich, Andreas Krause, and especially by the members of the thesis committee Jörg Stelling and Manfred Morari, were greatly appreciated. I wish to acknowledge my friends Nicola Carlon, Ruggero Dalla Santa, Francesco Gibaldi, Andrea Gesmundo, and Andrea Ierace for supporting my enthusiasm in science and for the amazing time that we have had together. A further word of acknowledgment goes to the memory of Giuseppe Dalla Santa and Riccardo Carlon.

The collaboration work presented in this thesis has been financed in part by the following agencies and institutions. The Machine Learning Laboratory received funding from the SystemsX.ch initiative (LiverX and YeastX projects), evaluated by the Swiss National Science Foundation, and support by the DFG-SNF research cluster FOR916. The laboratory headed by D. W. Gerlich received funding from the European Community's Seventh Framework Programme FP7/2007-2013 under grant agreements no. 241548 (MitoSys) and no. 258068 (Systems Microscopy), from a European Young Investigator award of the European Science Foundation, from an EMBO Young Investigator Programme fellowship to D. W. Gerlich and from the Swiss National Science Foundation. J. P. Fededa was funded by an EMBO long-term fellowship. The study on dyscalculia was funded by the CTI-grant 11006.1 and the BMBF-grant 01GJ1011. The work on dyslexia was funded by the CTI-grant 8970.1. Together with the co-authors of the respective works, we thank C. S. Ong, J. M. Buhmann, S. Dimopoulos, D. Scheder, C. Sommer, E. Zamora-Sillero, J. Stelling, U. Sauer, M. Kast, K. H. Brodersen, and B. Solenthaler for helpful suggestions and/or feedback on the published manuscripts, M. Held for processing image data, and R. Stanyte for user annotations of image data.

I acknowledge these organizations, institutions, and initiatives:

- Eidgenössische Technische Hochschule Zürich, CH
- Università degli Studi di Padova, IT
- Scuola Matematica Interuniversitaria, IT
- University of Illinois at Urbana-Champaign, US
- Stanford University, US
- Massachusetts Institute of Technology, US
- California Institute of Technology, US
- University of Cambridge, UK
- Centrum Wiskunde & Informatica, NL
- International Centre for Mathematical Sciences, UK
- Competence Center for Systems Physiology and Metabolic Diseases, CH
- Max-Planck-Gesellschaft zur Förderung der Wissenschaften, DE
- The Royal Society of London for Improving Natural Knowledge, UK
- Swiss National Science Foundation, CH
- Deutsche Forschungsgemeinschaft, DE
- Life Science Zurich Graduate School, CH
- SystemsX.ch, CH
- Society for Industrial and Applied Mathematics
- Institute of Electrical and Electronics Engineers
- Functional Genomics Center Zurich, CH
- Free Software Foundation Europe
- Free Software Foundation
- European Science Foundation
- European Molecular Biology Organization
- Public Library of Science
- Wikimedia Foundation

Contents

Chapter 1

Introduction

> *"It is those who know little, and not those who know much, who so positively assert that this or that problem will never be solved by science."*
>
> — C. R. Darwin

This chapter starts by outlining the content of the thesis. Focus and motivation are described in the first paragraph. The second paragraph gives an introductory survey of the relevant literature, and is followed by a first positioning of the work. The organization of the manuscript and the main contributions of the thesis are described in the following. For clarity, the last section of the introduction contains an informal explication of the terminology used through the text, as well as of the most commonly used abbreviations. Before the bibliography, the nomenclature consists of a topic-wise description of the formal notation.

Focus of the thesis. The thesis focuses on modeling dynamical systems from empirical observations. Dynamical systems are powerful tools to predict and control time-varying phenomena. At present, dynamical models are employed in a variety of domains, such as physics,

economics, biology, medicine, and engineering. They are fundamental: the foundation of entire scientific theories relies upon them. Dynamical systems are particularly useful models which express cause-effect relations between the interacting components of a system.

This work entertains the idea that the value of a dynamical system depends, among other things, on its predictive capacity with respect to an objective. In the hands of a modeler, dynamical systems can be employed as mathematical tools useful for prediction. Ultimately, the expected quality of such predictions depends on the amount of available information regarding the relevant aspects of the problem. The modeler extracts such information from empirical observations as well as, when possible, by incorporating previous knowledge. The modeling process often involves both the estimation of the internal states over time, as well as of the structural interactions between the components of the system state. The models are, in practice, identified through a combination of assumptions and accumulated evidence. The central question is: on the basis of finite and noisy observations, which dynamical system should be selected for a certain application? In this thesis, the value of a dynamical system is established by its ability to perform accurate predictions within the application scope. To make the goal precise, the evaluation of the predictive power requires a formal definition of success in prediction. On the basis of statistical arguments, the presented results aim at formalizing and automating aspects of the hypothetico-deductive method of scientific inquiry [Whewell (1837)]. Information theory provides a formal framework to quantify uncertainty in terms of transmission rates over communication channels. The theory constitutes the central theme and the mathematical cornerstone on which the following results are based upon.

Main motivation. By construction, dynamical models include time as an independent variable. The models are able to incorporate regularities beyond those exhibited by primarily static (or stationary) phenomena. Modeling of dynamical systems can be defined as the

discipline which aims at selecting dynamic models from empirical observations. This research field has a long and successful history, and currently offers important questions which remain open to further research and improvement. A relevant epistemological question is: how to perform induction? In technical terms, the question becomes: how to build a learning agent? At present, learning agents are based on estimation techniques which are empirically evaluated within the scope of a given task. Some approaches are more general than others, and are applicable to multiple concrete problems [Nelles (2001)]. Reasonable claims of universality are based on an assumption which is often implicit: phenomena of interest do not exhibit absolute arbitrariness. They do exhibit special regularities which, however, might be partially unknown to the modeler. Dynamical models aim at capturing the regularities which can be expressed as interactions between the constituting components of the system. In this sense, modeling consists of capturing such regularities from a finite set of empirical observations. When such observations are obtained in a controlled setting, the measurements qualify as experimental data. Ideally, the models obtained from data reflect the available evidence and the assumptions taken by the modeler. In many applications, however, data acquisition is a significantly resource-demanding process. Direct inspection of the inner workings of the system is often not possible to the modeler, and thus datasets may consist of scarce and noisy indirect observations. The limitation is severe because the expected quality of the model predictions is constrained by that of the available data. How to reliably extract genuine regularities from scarce data? On the one hand, the modeler aims at capturing as many regularities as possible from the available observations. On the other hand, the modeler should filter out the spurious noise effects to avoid prediction errors. Noise filtering is necessary to avoid confusing noisy fluctuations as genuine regularities. In a prediction scenario, there exists a justified tradeoff between the two antagonistic goals: the optimal balance yields the lowest error rate. How to define and calculate such a tradeoff? The available answers are well-justified but incomplete

[Nelles (2001); Bishop (2006); MacKay (2003)]. The results presented in this thesis aim at contributing to this active field of research.

Introductory Survey of the Field. A variety of mathematical methods are employed to model dynamical systems. In its essence, the field can be seen as a branch of control engineering which significantly overlaps with statistics and machine learning. In many cases, it shares not only the goals of these disciplines, but also the mathematical tools [Bishop (2006); MacKay (2003)]. There exists a rich and comprehensive literature of the field, which is condensed in this brief introduction. Precise surveys are postponed to the respective sections. For now, it suffices to refer the reader to a field of large applicability: Model Predictive Control (MPC) [García and D. M. Prett (1989)]. MPC is theoretically sound and practically useful: it makes a direct use of explicit and separately identifiable models to control physical processes. Of direct relevance to the topics of this thesis are two settings: that of modeling with non-linear systems and of learning under significant uncertainty. The former setting exhibits challenges whose nature is primarily computational. In many practical problems, the computational limitations of estimation are enormously aggravated by the fact there are no known regularities which can provide an advantage for optimization. However, there also exist cases in which such knowledge is available. This is often the case for concrete applications with well-studied models [Nelles (2001)]. High uncertainty remains, however, a separate issue: it is exhibited when the modeler has limited access to data which are very noisy. In such cases, the modeler may not be able to satisfy an important requirement: that of quantifying and assessing the uncertainty associated with the results.

Systems biology is a domain in which the described conditions constitute the norm, rather than the exception. On the one hand, bio-medical experimentation is particularly resource-demanding. On the other hand, the investigated phenomena seem to exhibit exceptional complexity. As a research field, modeling of dynamical systems overlaps with active learning when actions are possible. The field

of system identification covers the design of experiments aimed at predictive modeling [Pronzato (2008); Atkinson and Donev (1992)]. Experimental design can be seen as a research area at the interface with statistics. It encompasses the design of passive strategies, as well as of active ones. In the passive setting, the modeler selects a subset of measurable quantities from a pool of available candidates. This case is in contrast to that in which interventions are performed by an agent, which might, for instance, exercise its agency through the actuation of input perturbations [He and Geng (2010)]. A comprehensive literature covers both passive as well as active design strategies [Pronzato (2008); Chaloner and Verdinelli (1995)]. Established procedures have been recently applied to related domains, such as those involving Partial Differential Equations (PDEs) models [Banks and Rehm (2013)].

In summary, a central theme is that of defining and selecting the optimal trade-off between model informativeness and estimation stability [Buhmann (2010)]. The issue is of theoretical as well as practical importance: which model best generalizes the available data?

Positioning of the work. The work presented in the next chapter interfaces with multiple research areas: unsupervised (cluster validation), supervised (parameter estimation and model selection), and active learning (experimental design) [Nelles (2001); Bishop (2006)]. The central topics in the thesis are unified by a common theme: information theory provides a useful framework to evaluate the predictive power of dynamic models [Cover and Thomas (1991); MacKay (2003)]. From a mathematical perspective, the thesis is based upon three main frameworks: probability, information, and systems theory. The contributions presented here are primarily methodological and exhibit wide applicability. They are motivated, however, by open questions in systems biology and human learning.

Organization The thesis is organized as follows. This chapter is introductory: it describes structure and main contributions of the

thesis. The second chapter contains the necessary background and lists the main assumptions. The background consists of basic notions from systems theory, probability theory, and information theory. Results are organized in three chapters: supervised, unsupervised, and active learning of dynamical systems. The conclusion discusses the reported results, and provides an analysis of limitations and potential improvements.

1.1 Contributions

> *"If I have seen further, it is by standing on the shoulders of giants."*
>
> — I. Newton

The contributions encompass multiple aspects of information theoretic modeling of dynamical systems. The organization follows the structure of the three chapters described below. Detailed results are reported in the respective sections and summarized in the conclusion.

- **Unsupervised learning:** (Chapter 3)
  - **Clustering of time series and validation** (Sec. 3.1)
    Objective 1: which measured trajectories are statistically distinguishable? The task consists of selecting cost models for clustering multivariate time series.
    Motivation: when one can select between cost models, it is often a central issue to decide which model best generalizes the available observations. Approximation Set Coding is a recently introduced principle which exhibits the potential to address the issue of model selection in this setting [Buhmann (2010)].
    Contribution: an ASC-based method is introduced to perform order and model selection of costs for relational clustering of time series.
  - **Modeling of high-dimensional sequences** (Sec. 3.2)
    Objective 2: how to build a dynamic model from high-dimensional data sequences without supervision? The task of estimating the dynamic transition function aims at capturing the behavior in a space of statistically distinguishable system states.
    Motivation: human labeling of sequential data such as image patches is not only time-consuming, but also tends to lack self-consistency. Automated procedures are highly

desirable because they enable objective inference with reproducible and coherent results.
Contribution: a method is introduced with the aim of reconstructing the dynamics of a system which evolves in a high-dimensional space. The method is applied to microscopy image analysis; validation is performed with cell cycle data and compared with supervised results.

- **Supervised learning:** (Chapter 4)
  - **Preventive resampling for generalized state estimation:** (Sec. 4.1)
    Objective 3: how to reliably estimate the parameters of a dynamic system from time series? Bayesian approaches such as particle filtering approximate the posterior distribution of the parameters through sampling. However, they tend to suffer progressive information loss due to sample impoverishment and approximation degeneracy.
    Motivation: particle filters are established techniques whose applicability is limited by computational constraints. Improving the efficiency of the techniques is important to extend the scope of practical applicability of the methods.
    Contribution: a preventive approach is introduced to mitigate the information loss due to divergent approximations based on conventional resampling for generalized state estimation.
  - **Quality assessment of heuristic solutions for global optimization problems:** (Sec. 4.2)
    Objective 4: what is the relative position of a solution obtained from a given heuristic approach to optimization? Optimization problems are often computationally hard to solve. The available solutions obtained heuristically might not be globally optimal. The task is to estimate how many solutions are better than the best available heuristic approximation.

Motivation: many tasks are formulated as optimization problems exhibiting unknown regularity. In such cases, it is useful to know at least the uncertainty associated with the relative position of the available solutions obtained with heuristic procedures. Such results are directly applicable to statistical estimation and to the design of experiments.
Contribution: a probabilistic bound is obtained to empirically quantify the uncertainty regarding the quality of optimized solutions. The bound is based on the argument of the proof of the symmetrization lemma from statistical learning theory [Bousquet et al. (2004)].

- **Modeling human learning dynamics for the treatment of dyslexia and dyscalculia:** (Sec. 4.3)
  Objective 5: how to model aspects of human learning to develop targeted treatments for disabilities? Ideally, the model should capture the learning rate of a subject and other time-varying quantities from on-line observations. On the basis of such information, a software tutoring system can be employed to optimize the process of learning by selecting appropriate actions.
  Motivation: feature selection is one of the main challenges in modeling human learning. Predictive features are important to estimate knowledge states and improvement over time. Targeted treatment of learning disabilities benefits from model-based selection of effective interactions.
  Contribution: treatment of dyslexia and dyscalculia is improved by employing dynamic models. Experimental validation demonstrates the predictive power of the models and their ability to enhance human learning.

- **Active learning** (Chapter 5)

  - **Near-optimal design of experiments for modeling with dynamical systems:** (Sec. 5.1)
    Objective 6: which observations improve estimation with

dynamical system? The modeler aims at actively tuning experimental variables to perform better predictions.
Motivation: experiments are often resource-demanding. The rational allocation of such resources is important to achieve high efficiency in learning.
Contribution: a near-optimal design method is introduced to select informative time points and measurable quantities. The method exhibits formal performance guarantees of near-optimality, which are proven by building on previous work [Krause and Guestrin (2005)].

The introduced content results from research collaborations with the co-authors of the following articles. Such content appears in the peer-reviewed articles:

- A. G. Busetto, M. Sunnåker, J. M. Buhmann
  Computational Design of Informative Experiments in Systems Biology
  in "System Theoretic and Computational Perspectives in Systems and Synthetic Biology", (ed.) G.-B. Stan, V. Kulkarni, K. Raman, Springer [in press] (2013),

- M. Sunnåker, A. G. Busetto*, E. Numminen*, J. Corander, M. Foll, C. Dessimoz
  Approximate Bayesian Computation
  PLOS Computational Biology, [in press] (2012),

- Q. Zhong, A. G. Busetto, J. P. Fededa, J. M. Buhmann, D. W. Gerlich
  Unsupervised Modeling of Cell Morphology Dynamics for High-throughput Time-lapse Microscopy
  Nature Methods, 9, pp. 711–713 (2012),

- T. C. Käser-Jacober, A. G. Busetto, G.-M. Baschera, J. Kohn, K. Kucian, M. von Aster, M. Gross
  Modelling and Optimizing the Process of Learning Mathematics
  Lecture Notes in Computer Science, 7315, Springer, Proc. Intelligent Tutoring Systems, pp. 389–398 (2012),

- M. Haghir Chehreghani, A. G. Busetto, J. M. Buhmann
  Information Theoretic Model Validation for Spectral Clustering
  Journal of Machine Learning Research, Proc. Int. Conf.

*contributed equally.

on Artificial Intelligence and Statistics, pp. 495–503 (2012),

- G.-M. Baschera, A. G. Busetto, S. Klingler, J. M. Buhmann and M. Gross
  Modeling Engagement Dynamics in Spelling Learning
  Lecture Notes in Computer Science, 6738, Springer, Proc. Artificial Intelligence in Education, pp. 31–38 (2011)
  (Best Student Paper Award),

- A. G. Busetto, J. M. Buhmann
  Stable Bayesian Parameter Estimation for Biological Dynamical Systems
  IEEE CS Press, Proc. Int. Conf. on Computational Science and Engineering, pp. 148–157 (2009)
  (Best Paper Award),

- A. G. Busetto, C. S. Ong, J. M. Buhmann
  Optimized Expected Information Gain for Nonlinear Dynamical Systems
  ACM Series, 382, Proc. Int. Conf. on Machine Learning, pp. 97–104 (2009),

- A. G. Busetto, J. M. Buhmann
  Structure Identification by Optimized Interventions
  Journal of Machine Learning Research, Proc. Int. Conf. on Artificial Intelligence and Statistics, pp. 57–64 (2009),

and, for collaborations with advised students, on the master theses:

- A. Hauser,
  advised by A. G. Busetto and supervised by J. M. Buhmann

Entropy-based Experimental Design for Model Selection in Systems Biology
ETH Zurich, Department of Computer Science (2009),

- G. Krummenacher,
  advised by A. G. Busetto and supervised by J. M. Buhmann
  Large-scale Experimental Design Toolbox for Systems Biology
  ETH Zurich, Department of Computer Science (2010).

1.2 Terminology and Abbreviations

> *"My difficulty is only an – enormous – difficulty of expression."*
>
> — L. Wittgenstein (transl.)

For clarity, this section explains some terms used through the text. These informal definitions serve the purpose of introducing the reader to the setting of the study. Formal definitions are introduced in the respective sections. The interdisciplinary nature of the work requires an additional effort to explicate the terms which are shared among several disciplines. The author apologizes in advance for slight abuses of terminology. Particular emphasis has been placed on terms with overloaded (and context-specific) connotations, such as "model" and "hypothesis".

- *Computation:* the process of deterministic execution of a finite sequence of symbolic operations. The thesis deals with computation in the abstract, that is regardless of physical implementation. Results are based on the notion of digital computation and, more precisely, on the relation between abstract and concrete computation expressed by the Church-Turing Thesis [Church (1932); Turing (1937)].

- *Measurement:* the process of obtaining and recording numerical data from the studied phenomenon. The observational quantities obtained through the operation of an experimental apparatus are referred to as measurement data.

- *Epistemic Agent:* a learning entity which is capable of computation and action. The agent exhibits internal consistency, means-end coherence, and consistency with belief acquired through passive or active observations [Bratman (1987)]. In this work, the epistemic agent defines a formal modeling system. Such system is capable to process information, retrieve and record data, perform measurements and interventions.

- *Hypothesis:* a candidate explanation for a phenomenon. Hypotheses must be amenable to statistical testing and are consistent with all observations available to the epistemic agent. A priori, candidate hypotheses formalize aspects of a phenomenon which are not conclusively explained on the basis of previous data. Hypotheses are "simple" in the sense that they represent individual candidate explanations [Grünwald (2007)].

- *Model:* mathematical description shared by a set of hypotheses regarding a phenomenon. In contrast to hypotheses, models are sets of candidate explanations with common properties (such as, for instance, exhibiting identical functional forms but with different parameters). In the statistical literature, hypotheses are also known as "simple hypotheses". This definition is in contrast to models, which are called "composite hypotheses".

- *Data:* structured aggregates of measurement observations available to the epistemic agent. In this work, data are represented and processed numerically.

- *Belief State:* internal state of an epistemic agent which measures the subjective plausibility of events. In learning, beliefs are defined over the set of hypotheses, that is the hypothesis class. Belief states are self-consistent and time-varying: their update follows the rules of inference and depends on the available data. Whereas belief states a priori are subjective, justified belief states a posteriori follow deterministically from the priors.

- *Dynamical System:* time-dependent hypothesis which aims at capturing relations between internal states, inputs and outputs of a physical system (that is the data generator). Dynamical systems are able to express cause-effect relations between the variables.

- *A Priori:* the belief state of an epistemic agent before the observation of data.

- *A Posteriori:* the belief state of an epistemic agent after the update of the prior on the basis of the newly available data, which consist of single or multiple measurement instances.

The following abbreviations appear in the text:

Alg.	: Algorithm
Def.	: Definition
Fig.	: Figure
Obj.	: Objective
Sec.	: Section
Tab.	: Table
ASC	: Approximation set coding
BIC	: Bayesian information criterion
CC	: Correlation clustering
CI	: Confidence interval
CSE	: Constant shift embedding
DBN	: Dynamic Bayesian network
EM	: Expectation maximization
ESS	: Effective Sample Size
FN/TP	: False negative/positive
HMM	: Gaussian mixture model
GS	: Gold standard
HMM	: Hidden Markov model
IID	: Independent identically distributed
IVP	: Initial value problem
MAP	: Maximum a posteriori
MCMC	: Markov chain Monte Carlo
MDL	: Minimum description length
MPC	: Model predictive control
ODE	: Ordinary differential equation
PC	: Pairwise clustering
PCA	: Principal component Analysis
PDE	: Partial differential equation
RNAi	: RNA interference
SD	: Standard deviation
SDE	: Stochastic differential equation
SMC	: Sequential Monte Carlo
SVM	: Support vector machine
T3C	: Temporal constrained combinatorial clustering
TN/TP	: True negative/positive

Chapter 2

Background

> *"Those who are in love with practice without knowledge are like the sailor who gets into a ship without rudder or compass and who never can be certain whether he is going."*
>
> — L. Da Vinci (transl.)

This chapter recapitulates basic notions from systems theory, probability theory, and information theory. The expert reader can skip the first three sections, which cover the minimal introductory background. The end of the chapter lists the main assumptions on this work. Despite the simplistic nature of the assumptions, they constitute a useful starting point to clarify the scope of the reported results. For reasons of space, these sections gloss over technical subtleties. Further insights are left to the specialized literature in the respective fields. The presented notions are covered in considerable depth by the literature [Hopcroft et al. (2007); Li and Vitányi (1997); MacKay (2003); Jaynes (2004); Cover and Thomas (1991); Nelles (2001)].

2.1 Systems Theory

> *"Science is what we understand well enough to explain to a computer. Art is everything else we do."*
>
> — D. E. KNUTH

The term dynamical system typically refers to the model of a natural or artificial phenomenon which is referred to as the physical system. In engineering applications, dynamical systems are used for control, design, diagnosis, simulation and optimization. In a learning setting, a dynamical system Σ can be seen as a tool aimed at prediction. The learned system captures aspects of interests of the studied physical system Σ^*. Dynamical systems are often modeled through the combination of first-principle assumptions and experimental data [Nelles (2001)].

2.1.1 State-space Modeling

Among the existing alternative definitions, this thesis defines dynamical systems as computable state-space models. State-space models express input-output relations in terms of causal effects between the internal states of a system. In this study, however, the modeler is unable to observe the inner workings through direct inspection.

Let $\mathcal{X}$ denote the state space, that is the set of all possible distinguishable states x of a system. The state space might be continuous or discrete, depending on the case. Let $n_x \in \mathbb{N}^{>0}$ denote the dimension of the state space.

Let $\mathcal{T} \subset \mathbb{R}$ be the discrete set of time points t_i from the initial time point t_0, $i \in \mathbb{N}$. Every pair of time points $(t_i, t_j) \in \mathcal{T}^2$ obeys the total order $t_i > t_j$ induced by the ordering $i > j$ of the indexes.

The transition function $\mathrm{F}\colon \mathcal{X} \times \mathcal{T} \to \mathcal{X}$ maps current states onto consequent ones (in some cases, the definition may be restricted to

time-invariant transition functions). The function is the map

$$\mathrm{F}\colon (t_i, x(t_i)) \mapsto x(t_{i+1}). \tag{2.1}$$

Definition 1. *A dynamical system Σ is defined as*

$$\Sigma \coloneqq (\mathcal{X}, \mathcal{T}, F), \tag{2.2}$$

where $\mathcal{X}$ is the state space, and $\mathcal{T}$ is a set of time points. The system obeys the transition function $F(x,t)$.

It is worth noting that, with the provided formulation, the memory of the system solely consists of the configuration expressed by the current state. Definition 1 captures a simplistic notion of causal deterministic behavior. It is, however, in theory sufficient for our purposes. The definition, in fact, captures a rich set of possible behaviors [Li and Vitányi (1997)]. The scope of the definition includes a dynamical system capable of universal computation: a universal Turing machine U [Turing (1937)]. Universal Turing machines are theoretical systems which are able to emulate the behavior of any other Turing machine without any loss of information [Hopcroft et al. (2007)]. The Church-Turing thesis states that the set of all numerical functions amenable to effective computation coincides with the class of partial recursive functions (that are those calculated by Turing machines) [Turing (1937); Church (1932); Li and Vitányi (1997)]. When Σ defines a universal Turing machine, the function F implements a universal partial recursive function [Li and Vitányi (1997)].

In its full generality, systems theory also considers systems beyond those of Def. 1. The set of considered systems includes, for instance, stochastic and continuous dynamical systems (with continuous time and state space). In the thesis, such systems are considered within the limits of their effective numerical approximation. In the case of stochastic and continuous systems, the results that follow apply to their numerical approximation [Stoer (2002)]. Consistently with that, data representation and information processing are intended to be algorithmic in nature.

The transition function F defines explicitly the causal relations among the components of the state. It also defines implicitly the trajectory of the system over time. The initial condition of the system Σ is denoted as

$$x_0 \coloneqq x(t_0) \in \mathcal{X}. \tag{2.3}$$

The trajectory of length $s+1$ for the system Σ from x_0 is

$$\varphi \coloneqq (x(t_0), \ldots, x(t_s)). \tag{2.4}$$

So far, the definitions described only autonomous dynamical systems, that are systems without input interventions. When the activity of the epistemic agent is limited to passive modeling, the transition function already incorporates by definition all external inputs and influences (through time-dependency, for instance). There are other cases, however, which are interesting in the context of active learning. When the agent can interact with the system, input interventions are considered explicitly in the definition of Σ.

In the non-autonomous case, the dynamical system Σ_t is subject to series of instantaneous input interventions

$$u \coloneqq (u(t_0), \ldots, u(t_s)). \tag{2.5}$$

Each intervention is denoted as $u(t) \in \mathcal{U}$, for the intervention space $\mathcal{U}$. The interventions influence the behavior of the system through the transition function. In the active setting, the definition of F is extended to

$$\mathrm{F}\colon (t_i, x(t_i), u(t_i)) \mapsto x(t_{i+1}). \tag{2.6}$$

2.1.2 Differential Models

In the rest of the manuscript, continuous dynamical systems are defined in terms of differential equations. For systems of ordinary differential equations (ODEs), the system Σ_{ODEs} is defined as follows.

Definition 2. *A system of ODEs Σ_{ODEs} is given by*

$$\frac{dx(t)}{dt} = F_{ODEs}(x(t), \theta), \tag{2.7}$$

where the function F_{ODEs} denotes the system of equations defining the infinitesimal increments in the trajectory over the state space.

The reader should note that system of ODEs may exhibit time-dependency as well. The numerical approximation of such system is performed up to a tolerance which is fixed a priori. The results that follow assume negligible numerical errors for the discretization of the system. Analogous definitions are introduced for systems of Stochastic Differential Equations, which are introduced in Sec. 3.2.

The parameter vector of Eq. (2.7) is denoted as $\theta \in \Theta$, and defined over the parameter space Θ. The functional form $\mathrm{F}_{\mathrm{ODEs}}$ is an example of a model M, that is a family of hypotheses sharing the functional form of Eq. (2.7) in the range of parameters defined by Θ. When modeling, the model class $\mathcal{M}$ is assumed to be fixed a priori. It may, however, still be data-dependent, as in clustering. From a model M, hypotheses are identified with their individual parameters $\theta \in \Theta$. In a learning scenario, the modeler is often assumed to know $\mathcal{X}$ and $\mathcal{T}$. In this setting, the modeler aims at estimating from the data the function F^* of the physical process $\Sigma^* \coloneqq (\mathcal{X}, \mathcal{T}, \mathrm{F}^*)$.

For a given initial condition x_0, one can determine the integral solution of the system over time. The task constitutes a conventional initial value problem (IVP). Whereas the solution of IVPs is straightforward for discrete cases, the definition of continuous IVPs may require careful restrictions on the set of allowable functions describing the infinitesimal transitions in the state space. All results in the thesis assume that the necessary conditions for the well-posedness of IVPs are satisfied. Well-posedness is defined in the sense of Hadamard[1]. The requirement extends to the numerical approximations of Eq. (2.7) and of other differential systems such as those described by delay or

[1]Informally, a problem is well-posed when its solution exists, is unique, and depends continuously on the data [Hadamard (1902)].

partial differential equations (DDEs and PDEs, respectively). Alternative formulations are also considered through the text, including hidden Markov models and dynamic Bayesian networks. The formal definition of these model classes is postponed to the relevant chapters.

2.1.3 Noisy Time Series Data

In the thesis, the epistemic agent is able to perform imperfect measurements of the physical system Σ^*. The experimental observations available to the modeler consist of finite datasets which are noisy. Noise terms are here denoted as ν and distributed according to the respective distributions $N(\mathcal{N})$. The noise variables are defined over the noise space $\mathcal{N}$. When measuring the state of Σ^* at time point t_i, for every i, the observer obtains the readout samples

$$y(t_i) \coloneqq h(x(t_i), t_i, \nu_i) \tag{2.8}$$

where

$$h\colon \mathcal{X} \times \mathcal{T} \times N \to \mathcal{Y}. \tag{2.9}$$

The measurement space is denoted as $\mathcal{Y}$. Measurements are taken at the time points

$$\mathcal{T}^{\downarrow} \subseteq \mathcal{T}, \tag{2.10}$$

where the finite sample size is

$$n \coloneqq |\mathcal{T}^{\downarrow}|. \tag{2.11}$$

Individual trajectories measured at time points $\mathcal{T}^{\downarrow}$ are denoted as

$$\psi \coloneqq (y(t_j), \ldots, y(t_n)), \tag{2.12}$$

with $1 \leq j \leq n$. Time series are obtained by combining trajectories with the respective sampling points, giving

$$\Psi \coloneqq \{(t_j, y(t_j))\}_{t_j \in \mathcal{T}^{\downarrow}}. \tag{2.13}$$

Data availability and noise level are defined by the experimental setting. Individual experiments are defined as $\varepsilon \in \mathbf{E}$, over the space of experimental settings $\mathbf{E}$.

Definition 3. *The experimental setting* $\varepsilon \in \mathbf{E}$ *can be defined as*

$$\varepsilon \coloneqq (\mathcal{T}^{\downarrow}, N, h). \tag{2.14}$$

Experimental design aims at selecting the best ε according to a given objective. In practice, the process of design amounts to setting the tunable parameters of ε to the most informative values. Section 5.1 adopts a simpler definition for ε: the experiment consists of a set of indexes which refer to individual observations, thus indirectly operating on h.

2.2 Probability Theory

> *"One sees, from this Essay, that the theory of probabilities is basically just common sense reduced to calculus."*
>
> — P.-S. Laplace (transl.)

This section recalls the basics of probability theory, which plays a central role in many (but not all) approaches to learning [Bishop (2006)]. Currently, probability theory remains the prevalent framework to quantify and update the justified belief states of epistemic agents. It is, however, not the only option available to the modeler. An alternative to probability theory would be given, for instance, by possibility theory [Dubois and Prade (1988)] (which is based on fuzzy set theory [Zadeh (1978)]). Another alternative could be Dempster-Shafer theory of belief functions [Dempster (1968); Shafer (1976)]. In contrast to probability theory, transferable belief model theory [Smets and Kennes (1994)] separates belief from decisions[2]. Discussing about the relative merits of alternative approaches to quantify belief would be epistemologically interesting and intellectually valuable, but the topic goes beyond the scope of this work. The thesis is based on probabilistic grounds, which also constitute the foundation of (classical) information theory [Cover and Thomas (1991)].

Probability theory has been formalized according to alternative sets of axioms. At this point, it is important to note that the framework of probabilistic inference is not a choice made ad-hoc: there exist different sets of axioms for belief quantification which invariably lead to the conventional rules of probability [Bishop (2006)]. Alternative formalizations of probability theory are different in philosophy and purpose. Among others, there exist theories of Ramsey [Ramsey (1931)], Kolmogorov [Kolmogorov (1933, 1965)], Cox

[2]The formalization rejects the validity of betting arguments which are widely employed to justify probabilistic belief.

[Cox (1946, 1961)], Good [Good (1950)], Savage [Savage (1961)], de Finetti [de Finetti (1970)] and Lindley [Lindley (1982)]. They suggest different interpretations[3], but the distinguishing technical details seem to be negligible for most practical purposes. As a first drastic simplification, it is customary to define the two main interpretations of probabilities as either classical (sometimes called frequentist) or Bayesian. It is important to highlight, however, that no single classical or Bayesian interpretation exist [Bishop (2006); Jaynes (2004)].

2.2.1 Modeling Justified Belief

The thesis subscribes to the Pólya-Cox axioms, which assign probabilities on logical grounds. Pólya-Cox axioms satisfy three minimal desiderata of rationality and consistency [Cox (1946); Jaynes (2004)]. To formalize these ideas, let us denote $B(\rho)$ as the degree of belief in proposition ρ, and $B(\rho|\varrho)$ as the degree when ϱ is true. In brief, the axiomatic desiderata are [Jaynes (2004); MacKay (2003)]:

> **PC-A1:** Probability values are divisible, comparable, bounded, and depend on the available information[4]:
>
> $$B\colon \Xi \to [0,1], \tag{2.15}$$
>
> where Ξ is a set of propositions.
>
> **PC-A2:** The degree of belief in a proposition $\rho \in \Xi$ is a function of its negation $\neg\rho \in \Xi$:
>
> $$B(\rho) = \chi[B(\neg\rho)],$$
>
> for a certain function χ.

[3] Even the term "interpretation" is not strictly applicable in its typical sense to this topic. A more definition would be "explication" [Carnap (1945)], since there does not exist a single formal system of probability. This topic is philosophically important and deserves attention, but goes beyond the scope of this work.

[4] Probabilities are here arbitrarily (yet conventionally) normalized between $B(\textsc{False}) = 0$ and $B(\textsc{True}) = 1$. The choice is without any loss of generality.

PC-A3: The degree of belief in the conjunction of $\rho \in \Xi$ and $\varrho \in \Xi$ is a function of the degrees of plausibility of ρ given ϱ and that of ϱ alone:

$$B(\rho \wedge \varrho) = \zeta[B(\rho|\varrho), B(\varrho)],$$

for a function ζ.

In the limit of absolute certainty, these requirements are consistent with Aristotelian deductive logic obeying a Boolean algebra. Notably, the formalization generalizes Aristotelian logic by including weak (but still valid) syllogisms [Jaynes (2004)]. The axioms, in conjunction with the additional differentiability condition for ζ, determine uniquely the set of valid reasoning rules (up to normalization) [Jaynes (2004)]. Invariably, one has that $B(\rho) \equiv p(\rho)$. The obtained rules consist of the (familiar) sum and product rules of conventional probability calculus. In the end, all existing definitions of probability invariably lead to the same set of rules. In probability theory, the justification of any set of axioms is currently not undisputed [Jaynes (2004)]. However, it is noteworthy that the Pólya-Cox system is in essential agreement with that derived from Kolmogorov's axioms. The difference between the two is primarily epistemological: Pólya-Cox axioms provide a foundation based on logic [Jaynes (2004)].

Several of the results in this thesis are better understood within a Bayesian framework, while others, such as Approximation Set Coding (ASC), not necessarily. In the Bayesian setting, distributions quantify the belief state of the epistemic agent [Bishop (2006); Jaynes (2004)]. Bayesian belief states are thus subjective, yet not arbitrary. In fact, they are justified on the basis of the available evidence [Jaynes (2004)] and coincide under the same priors. The Bayesian perspective is in contrast to the interpretation of probabilities in terms of frequencies of random and repeatable events. In all cases, probabilities are intended as carriers of information[5]. The isomorphism of the alternative def-

[5]To be more precise, it would be better to talk about probabilities in terms of carriers of incomplete information, that is uncertainty. This subtle but impor-

initions of probability makes the axiomatic differences of negligible impact for the overwhelming majority of practical purposes [Bishop (2006); Jaynes (2004)].

2.2.2 Bayesian Inference

Combining the sum and product rules, one obtains the cornerstone of Bayesian inference: Bayes' theorem. For two random variables M and D, Bayes' rule is given by

$$p(M|D) \equiv \frac{p(D|M)p(M)}{p(D)}. \tag{2.16}$$

Letting model and data be denoted, respectively, as M and D, the terms of Bayes' theorem are conventionally referred to as

- $p(M)$: prior probability
  (probability of model M before observing data D);
- $p(M|D)$: posterior probability
  (probability of model M after observing data D);
- $p(D|M)$: likelihood
  (probability of generating data D from model M);
- $p(D)$: the evidence
  (general probability of observing data D).

The class of models, that is the sample space for M, is denoted as $\mathcal{M}$ and called the model class. The calculation of the model posterior involves the marginalization for all individual hypotheses. In fact, taking hypotheses H from the hypothesis class $\mathcal{H}$ into account, one has

$$p(M|D) \equiv \sum_{H \in \mathcal{H}} p(M, H|D). \tag{2.17}$$

tant distinction is clarified in the next section, which introduces basic ideas from information theory.

The data space $\mathcal{D}^*$ indicates the space of datasets, which consists of all possible sequences of samples $y(t_i)$ of size n obtainable from Σ^* through the measurement process of Eq. (2.8). In the rest of the thesis, H will indicate hypothetical transition functions F. Similarly, M will indicate families of hypothetical transition functions sharing the same functional form (but, for instance, subject to different assignments of the parameters).

2.2.3 Modeling Uncertainty

A significant advantage of Bayesian inference is the direct inclusion of previous knowledge in terms of prior probabilities. In the case of independent experiments, the posterior of the last iteration constitutes the prior of the current one without suffering any loss of information. The Bayesian reliance on priors is, however, also the subject of heated controversies [Bishop (2006)]. One might ask: where are the priors coming from? Are they chosen according to mathematical convenience or on the basis of previous evidence? These questions are important because the effect of the prior on the posteriors might be significant. Multiple solutions have been proposed to solve the issue of assigning priors. All these solutions, in a way, are attempts at modeling ignorance.

The agnostic learner wants to avoid assigning zero priors to elements of $\mathcal{H}$. Imposing such prior would make any posterior zero by default, irrespective of the data. No evidence would be strong enough to modify the belief state for models which are excluded a priori. This problematic situation is avoided by Cromwell's rule [Lindley (1991)]. The rule states that the assignment of 1 or 0 to prior probabilities should be exclusively restricted to statements which are logically true or false, such as logical propositions.

Other approaches to assign priors require the transfer of aggregate statistics from previous experiments. The technical way to incorporate into the prior some non-probabilistic information is delicate and deserves special attention. To set the prior, Bayes rule must be com-

plemented by subscribing to one or more external assumptions. In the agnostic case, one could apply the principle of indifference. Informally, it states that, given insufficient reasons to distinguish individual hypotheses, all available candidates should be considered equally plausible [Keynes (1921)]. Formally, the principle sets the prior to the uniform distribution[6]. More generally, non-informative priors are attempts at exercising minimal influence on posteriors. Among other issues, they might even be improper due to lack of normalization. Assigning a prior gets even more problematic when densities are transformed according to non-linear changes of variables [Bishop (2006); Berger (1985)]. Jeffreys' rule addresses this issue by constructing invariant non-informative prior distributions on the parameter space. Such distributions exhibit invariance under a class of reparameterizations. The rule assigns priors which are proportional to the square root of the determinant of the Fisher information [Jaynes (2004)].

When only partial information is available, the principle of maximum entropy provides a way to incorporate the available testable information[7]. As shown later, testable information has to be amenable to statistical verification. In the continuous case, which is introduced below, the application of the principle of maximum entropy requires the specification of an invariant measure function. The requirement is necessary to avoid dependency on the choice of the parameters [Jaynes (2004)]. Informally, the principle of maximum entropy states that no additional information should be presumed (the formal definition is postponed to the next section). Notably, several well-known distributions are obtainable from maximum entropy arguments. This set of distributions includes uniform, exponential, Gaussian, Laplace and Gibbs distributions [Bishop (2006)]. With awareness regarding the limitations of maximum entropy arguments, the principle may be employed as an additional assumption. In principle, there exists an

[6] Additional considerations are required in the case of non-bounded hypotheses classes [Jaynes (2004)]. In such cases, transformation invariance may become a particularly dangerous issue.

[7] One has to be careful, however, on how testable information is defined and obtained [Jaynes (1957a,b); Shannon and Weaver (1963)]

elegant way to define a general prior which is, in a technical sense, objective and universal [Rathmanner and Hutter (2011)]. Such distribution, named Solomonoff-Levin distribution, specifies a prior over the set of computable functions [Solomonoff (1964a,b)]. The universal prior is

$$\mathrm{M}(\bar{z}) \propto \sum_{\mathrm{pg:U[pg]}=\bar{z}*} 2^{-\mathrm{len}(p)}, \tag{2.18}$$

where $\bar{z} = z_1 z_2 \dots z_n$ is a string of length n. In the equation, the terms $z_i \in \mathcal{Z}$ for all $i = 1, \dots, n$, indicate symbols from the alphabet $\mathcal{Z}$. Now, $\bar{z}*$ denotes the subset of strings of arbitrary (but always finite) length having $\bar{z}$ as a prefix. The minimal program pg of length len(pg) outputs string $\bar{z}$ when emulated by the universal Turing machine U [Rathmanner and Hutter (2011); Solomonoff (1964a,b)]. The prior $\mathrm{M}(\bar{z})$ relies on quantities which cannot be computed even in principle, and thus requires practical approximations. Under rather minimal assumptions, inference based on the universal prior can, however, be regarded as a gold standard [Rathmanner and Hutter (2011)]. At present, the application of such concepts remains an area of active research [Li and Vitányi (1997); Rathmanner and Hutter (2011)].

The likelihood function $p(D|M)$ constitutes a probability with respect to D, but not with respect to M (due to lack of normalization). It plays a central role both in Bayesian inference through Bayes' rule, and in the classical framework. In classical statistics, conclusions are often drawn according to the principle of maximum likelihood [Fisher (1922)]. The normalization of $p(D|M)$ is given by Bayes' rule. Bayesian agents perform inference by calculating posteriors from prior and likelihood (given the data). In the Bayesian setting, the evidence term is not as fundamental as the prior and the likelihood; in fact, it constitutes just a normalization constant (for a given dataset). In practice, the evidence can be calculated as

$$p(D) \equiv \sum_{M \in \mathcal{M}} p(D|M)p(M), \tag{2.19}$$

where $\mathcal{M}$ is the hypothesis class, that is the set of candidate models available to the modeler.

So far, probabilities have been defined over propositions. Equivalently, they may be expressed as functions on sets of events (consistently with established formulations). It is possible to extend the given definitions to continuous random variables, obtaining similar properties for density functions. In the easy univariate case, a probability density may be defined as

$$p(z \in (a,b)) \coloneqq \int_a^b p(z)\,\mathrm{d}z, \tag{2.20}$$

where $z \in \Omega$ denotes the value taken by a continuous random variable Z. Here, (a,b) denotes the continuous interval on the real line $\mathbb{R}$. Sum and product rules of probability apply to densities as well, and also to combinations of discrete and continuous variables [Bishop (2006)]. Similar properties extend the univariate definitions to multivariate settings [Jaynes (2004)]. For mean and covariance parameters μ and Σ, the multivariate normal distribution of dimension r is defined as

$$\mathcal{N}_{\mathrm{or}}(z|\boldsymbol{\mu},\boldsymbol{\Sigma}) \coloneqq \frac{1}{(2\pi)^{r/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(z-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right).$$

The sample space is $\Omega = \mathbb{R}^r$ and the covariance matrix $\boldsymbol{\Sigma}$ is symmetric and positive definite. With a slight (but conventional) abuse of notation, the rest of the thesis will homogeneously refer to both probability distributions and densities as $p(Z)$. The distinction is typically clear from the context. If Z is a random variable and $z \in \Omega$ is an element of the sample space Ω, one may for simplicity use $p(z)$ or $p(Z)$ rather than $p(Z = z)$.

2.3 Information Theory

> *"Information can tell us everything. It has all the answers. But they are answers to questions we have not asked, and which doubtless don't even arise."*
>
> — J. Baudrillard (transl.)

This section starts by defining two central concepts: self-information and entropy. These concepts are consistent with commonly accepted notions of uncertainty [MacKay (2003); Cover and Thomas (1991)]. The Shannon self-information content for the outcome z in the sample space Ω of the random variable Z is given by

$$\mathrm{h}(z) \coloneqq -\log_2 p(z), \tag{2.21}$$

where $p(z)$ is the probability distribution for Z [Shannon (1948); Shannon and Weaver (1963)]. Except otherwise specified, the base of the logarithm will remain fixed to 2. The choice of measuring information in Bits is arbitrary and without any loss of generality[8].

2.3.1 Uncertainty Quantification

Shannon entropy is a measure of the uncertainty associated with the distribution of the random variable Z. Informally, entropy measures the missing information over a weighted ensemble of possible outcomes.

The Shannon entropy of the random variable Z is given by

$$\mathrm{H}[p] \coloneqq \mathbb{E}[\mathrm{h}(z)] \equiv -\sum_{z\in\Omega} p(z)\log_2 p(z), \tag{2.22}$$

that is by the expected self-information for the ensemble.

[8] There will be other cases in which the base of the logarithm is e, leading to alternative measures in Nats.

In a communication scenario, Shannon entropy arises as an answer to the following technical question: what is the average length of the shortest description of the random instance emitted by a stationary source [Shannon (1948); Cover and Thomas (1991)]? Shannon-Khinchin axioms provide the foundation for the derivation of Shannon entropy from a set of minimal requirements [Khinchin (1957); Shannon and Weaver (1963)]. Defining the probabilities $p_i \coloneqq p(z_i)$ with $z_i \in \Omega$ for all $i = 1, \ldots, n$, the axioms are defined for an uncertainty measure S [Khinchin (1957); Suyari (2004)]:

SK-A1: for every $n \in \mathbb{N}^{>0}$, $S(\bar{p})$ is continuous with respect to the argument $\bar{p} \coloneqq (p_1, \ldots, p_n)$.

SK-A2: for a given $n \in \mathbb{N}^{>0}$, the point of global maximum for $S(\bar{p})$ is $\bar{u}_n \coloneqq (1/n, \ldots, 1/n)$ (that gives the uniform distribution $\mathcal{U}_{nif}(\Omega)$).

SK-A3: the function S is additive with respect to every $p_{ij} \geq 0$, that is

$$S(p_1 1, \ldots, p_{nm_n}) = S(\bar{p}) + \sum_{i=1}^{n} p_i S\left(\frac{p_{i1}}{p_i}, \ldots, \frac{p_{im_i}}{p_i}\right), \quad (2.23)$$

for all $i = 1, \ldots, n$ and all $j = 1, \ldots, m_i$, where $p_i \coloneqq \sum_{j=1}^{m_i} p_{ij}$.

SK-A4: the function S is expandable, that is

$$S(p_1, \ldots, p_n, 0) = S(p_1, \ldots, p_n). \quad (2.24)$$

These requirements are met by a single function, which is Shannon entropy (up to scaling)[9]. One has, in fact, that $S(\bar{p}) \propto \mathrm{H}[p]$. Let $\mathcal{Z}^*$ denote the set of strings of arbitrary length composed of symbols from the alphabet $\mathcal{Z}$.

[9]Please note that, since $\lim_{p_i \to 0} p_i \log p_i = 0$, the work subscribes to the convention of taking $p_i \log p_i = 0$ for $p_i = 0$.

The prefix-free Kolmogorov complexity $\mathrm{K}(\bar{z})$ [Solomonoff (1964a,b); Kolmogorov (1965); Chaitin (1966); Li and Vitányi (1997)] is defined for the string $\bar{z} \in \mathcal{Z}^*$ as

$$\mathrm{K}(\bar{z}) \coloneqq \min_{\{\mathrm{pr}:\ \mathrm{U}[\mathrm{pr}]=\bar{z}\}} \mathrm{len}[\mathrm{pr}], \tag{2.25}$$

where U is a prefix-free universal Turing machine.

Notably, the choice of the universal machine affects the program length by at most a constant number of bits [Li and Vitányi (1997)]. A fundamental notion relates the Kolmogorov complexity to Shannon entropy. In fact [Li and Vitányi (1997)],

$$\sum_{\bar{z}\in\mathcal{Z}^*} p(\bar{z})\mathrm{K}(\bar{z}) \leq \mathrm{H}[p] + \mathrm{K}(p) + O(1), \tag{2.26}$$

for every recursive probability p over $\mathcal{Z}^*$.

Differential entropy exhibits similar (but not exactly the same) properties holding for Shannon entropy [Bishop (2006); Li and Vitányi (1997)]. The differential version is obtained by quantizing the continuous random variable and eliminating the logarithmic term (which diverges in the infinitesimal quantization limit).

Definition 4. *The differential entropy of the density p is*

$$H[p] \coloneqq -\int_\Omega p(z) \log p(z)\, dz. \tag{2.27}$$

For precision and consistency with the principle of maximum entropy, the thesis considers differential entropy with respect to the invariant measure function $\mathrm{m}(z)$. The precise (and preferable) definition of differential entropy then becomes [Jaynes (2004)]

$$\mathrm{H}_\mathrm{m}[p] \coloneqq -\int_\Omega p(z) \log \frac{p(z)}{\mathrm{m}(z)}\, \mathrm{d}z. \tag{2.28}$$

As for probabilities, the thesis homogenizes the notation for discrete and continuous cases. This overload is acceptable since the case is

typically clear from the context[10].

On the basis of these definitions, it is now possible to formalize the principle of maximum entropy (to incorporate it as an additional axiom). Let T denote the available testable information (that is, amenable to statistical verification), then

> **ME-A1:** the candidate distribution over the models in the hypothesis class is given by
>
> $$p_{\mathrm{ME}} = \arg \max_{\{p:\Upsilon[p]=T\}} \mathrm{H}[p]. \tag{2.29}$$
>
> where $\Upsilon[p]$ is the function producing statistics as T from p.

The principle it not only used to set priors, but also as a stand-alone principle for model specification [Jaynes (1957a)].

2.3.2 Learning and Communication

A fundamental quantity which relates two probability distributions p and q is the relative entropy [Kullback and Leibler (1951)], which is sometimes called Kullback-Leibler divergence.

Relative entropy is defined as follows:

$$\mathrm{KL}[p \parallel q] \coloneqq \sum_{z\in\Omega} p(z) \log \frac{p(z)}{q(z)}. \tag{2.30}$$

The interpretation of Eq. (2.30) is the following: given the approximating distribution q of the unknown stationary source p, $\mathrm{KL}[p \parallel q]$ is the expected number of additional bits required for communication. The equation is obtained under the assumption that transmission is performed with respect to an optimal coding for q (rather than for p, which is the source generator). It is important to note that relative

[10] Depending on the context, entropy is denoted either as $H[p]$ or as $H(z)$. The former notation highlights the dependency on the distribution, whereas the latter emphasizes the random variable associated with the distribution.

entropy is not symmetric, that is $\mathrm{KL}[p \parallel q] \not\equiv \mathrm{KL}[q \parallel p]$, and thus cannot be a distance [Kullback and Leibler (1951); Cover and Thomas (1991)]. Another fundamental quantity from information theory is the mutual information, which measures the statistical dependence of two random variables Z_1 and Z_2 (with respective sample spaces Ω_1 and Ω_2). The mutual information between two random variables is

$$\mathrm{I}(Z_1, Z_2) \coloneqq \sum_{z_1 \in \Omega_1} \sum_{z_2 \in \Omega_2} p(z_1, z_2) \log \frac{p(z_1, z_2)}{p(z_1)p(z_2)}. \tag{2.31}$$

Mutual information is a measure of statistical dependence, in the sense that

$$\mathrm{I}(z_1, z_2) = 0 \iff z_1 \perp z_2, \tag{2.32}$$

where $\perp$ indicates statistical independence, that is

$$p(z_1, z_2) = p(z_1)p(z_2). \tag{2.33}$$

Conditional entropy is defined as

$$\mathrm{H}(Z_1|Z_2) \coloneqq -\sum_{z_1 \in \Omega_1} \sum_{z_2 \in \Omega_2} p(z_1, z_2) \log p(z_1|z_2). \tag{2.34}$$

Hence, mutual information can be seen as the relative entropy between the joint distribution and the product of the marginal distributions:

$$\begin{aligned} \mathrm{I}(Z_1, Z_2) &\equiv \mathrm{KL}[p(z_1, z_2) \parallel p(z_1)p(z_2)] \\ &\equiv \mathrm{H}(Z_1) + \mathrm{H}(Z_2) - \mathrm{H}(Z_1, Z_2) \\ &\equiv \mathrm{H}(Z_1) - \mathrm{H}(Z_1|Z_2), \end{aligned} \tag{2.35}$$

where the term $\mathrm{H}(Z_1|Z_2)$ denotes the conditional entropy. The last equivalence of Eq. (2.35) offers a particularly valuable interpretation. Mutual information can be seen as the reduction of uncertainty about Z_1 as a consequence of observing Z_2 [Bishop (2006); Cover and Thomas (1991)]. The differential versions of the introduced quantities share fundamental similarities with their discrete counterparts[11].

[11] They also exhibit important differences [Bishop (2006); Cover and Thomas (1991); Jaynes (2004)].

The differential relative entropy, mutual information, and conditional entropy are, respectively, the following:

$$\begin{aligned} \mathrm{KL}[p \parallel q] &:= \int_{\Omega} p(z) \log \frac{p(z)}{q(z)} \, \mathrm{d}z \\ \mathrm{I}(Z_1, Z_2) &:= \int_{\Omega_1} \int_{\Omega_2} p(z_1, z_2) \log \frac{p(z_1, z_2)}{p(z_1) p(z_2)} \, \mathrm{d}z_1 \mathrm{d}z_2 \\ \mathrm{H}(Z_1 | Z_2) \quad - &:= \int_{\Omega_1} \int_{\Omega_2} p(z_1, z_2) \log p(z_1 | z_2) \, \mathrm{d}z_1 \mathrm{d}z_2; \end{aligned} \tag{2.36}$$

in all these, the random variables are continuous. As for the differential entropy, there exist analogous (and preferable) versions which include an invariant measure function.

2.4 Main Assumptions

> *"Making assumptions simply means believing things are a certain way with little or no evidence that shows you are correct, and you can see how this can lead to terrible trouble."*
>
> — L. SNICKET in "The Austere Academy"

The main assumptions of this work are explicitly stated below. This section contains informal and simplistic formulations. Rigorous specifications and exceptions are introduced in the respective sections.

- **There exists a data generating model.** The studied phenomenon admits a formal (that is, mathematical) representation Σ^*, that is called the physical system. The system is associated with a model M^* which indicates the transition function of the system, as in Eq. (2.2). The model may or may not (depending on the case) be an element of the hypothesis class of an epistemic agent. The evaluation of experimental design, for instance, considers both settings. Nonetheless, data are always assumed to be consistently generated according to Σ^*. Whereas guarantees and designs are formulated with respect to M^*, results and model predictions are evaluated empirically with external testing. The assumption is in contrast, for instance, to the view of extreme empiricism for which phenomena are not even in principle amenable to formal representation [van Fraassen (1980)]. *Justification:* the assumption provides a framework to evaluate learning rates and to provide statistical guarantees.

- **Belief states may be subjective, but inference is objective.** The conclusions of epistemic agents coincide when prior information, evidence, and inference method are shared by the agents. The assumption means that the belief state is $B(M|D) \coloneqq \Psi(B(M), D)$ for a certain function Ψ. Moreover,

conclusions are derived taking into account all available evidence: data points are not selected and no presumed evidence can be incorporated if it is not already in the prior. The assumption is in contrast, for instance, to radical subjectivism [Dancy (1985)].
Justification: the assumption is useful to justify the derivation of objective conclusions from subjective belief states and applies to Bayesian inference.

- **The dynamic behavior of the physical system is separable from the belief state of the modeler.** With the exclusion of active interventions, the belief state of an agent exercises no effect on the behavior of the physical system (with the additional requirement that the definition of the physical system does not include the epistemic agent). It means that, for every $t_i \in \mathcal{T}$, the state $x(t_j)$ (for $j > i$) of Σ^* may depend exclusively on the belief state of the agent at times $\{\ldots, t_{i-1}, t_i\}$ through active interventions. The assumption is in contrast to, for instance, EEG feedback systems in which the definition of the physical system includes the agent itself [Engstrom et al. (1970)].
Justification: the assumption simplifies the modeling process by imposing independence of the behavior of the physical system from the belief of the (passive) agent.

- **The measurement apparatus exercises negligible interference on the physical system.** As in classical mechanics, for every $t_i \in \mathcal{T}$, the state $x(t_i)$ does not depend on the operations of obtaining, storing, and processing data D produced at time t_i (assuming all of them to be instantaneous processes with respect to $\mathcal{T}$). Future states of Σ^* may, however, be a function of interventions selected on the basis of previously available data. The assumption is in contrast, for instance, to the Copenhagen interpretation of quantum mechanics [Brock (2003)].
Justification: the assumption simplifies the inference process by

making the behavior of the physical system independent from the effects of the measurement process.

- **The physical process is causal.** The future behavior of the physical system can be described solely as the function of current and past states. The assumption, implicitly incorporated in Eq. (2.2), imposes that transitions to future states $x(t_{i+1})$ may depend only on states and interventions acting at time points t_j with $j < i$. In other words, anticipatory effects are excluded. The assumption is in contrast, for instance, to anti-causal effects in batch image processing.
  Justification: the assumption significantly restricts the set of possible behaviors for the physical system.

- **Consecutive experiments are independent.** The observed outcomes of separate experiments are conditionally independent given the behavior of the physical system and of the measurement apparatus. This assumption requires that

$$\mathrm{I}(D_i, D_j | M^*, \varepsilon) = 0 \tag{2.37}$$

  for every measured dataset with $i \neq j$. As before, M^* denotes the data generator and $\varepsilon \in \mathbf{E}$ is the experimental setting. The assumption is in contrast, for instance, to the case in which memory effects are not negligible between experiments (and are not already captured by M^*).
  Justification: the assumption enables the recursive combination of evidence resulting from sequences of experiments.

- **Axiomatic foundations.** Unless otherwise specified, the work subscribes to

  - Zermelo-Fraenkel set theory (with the axiom of choice) [Zermelo (1908); Hardin and Taylor (2008)],
  - the Church-Turing Thesis [Church (1932); Turing (1937)],

- Pólya-Cox axioms
  (**PC-A1,2,3**) [Jaynes (2004)],
- Shannon-Khinchin axioms
  (**SK-A1,2,3,4**) [Khinchin (1957)],
- the Principle of Maximum Entropy
  (**ME-A1**) [Jaynes (1957a)].

An additional disclaimer should be made here. Overall, the thesis maintains a pragmatic standpoint. In spite of the apparent simplicity of the definitions, their application to real problems often requires intricate justifications to maintain precision regarding the range of validity of the obtained conclusions. The thesis is guided by the underlying rationale of the theory, while empirical results are evaluated externally in the context of the respective application domains.

Chapter 3

Unsupervised Learning

"Science is the belief in the ignorance of experts."

— R. P. Feynman

3.1 Time Series Clustering and Validation

Clustering is one of the cornerstones of unsupervised data analysis. It proves particularly useful in the first stages of modeling dynamical systems, for instance to extract compressed information regarding the distribution of trajectories in the state space. In brief, the goal of clustering is to select informative label assignments on the basis of the available observations. The fundamental modeling question is: which model should be selected? When models are seen as tools aimed at prediction, the question ultimately relies on a measure of information. Ideally, the modeler should select cluster assignments which are informative, as well as reasonably stable under the fluctuations induced by the noise. The central idea is the following: the modeler selects the model yielding the highest reliable information capacity. Models are predictive if they are able to consistently distinguish candidate solutions on the basis of the data. The capacity quantifies the degree of

statistical detail which is extracted by the model. The solutions consist, in this setting, of cluster assignments from the hypothesis class $\mathcal{C}$. The hypothesis class contains all possible assignments of samples to cluster labels.

This section introduces a method to cluster time series and validate the obtained results[1]. The introduced method is based on Approximation Set Coding, a recently introduced principle for model validation [Buhmann (2010)]. In ASC validation, consistently with most of statistical learning theory, the task of model selection is formulated with respect to a class of cost models. The class is given to the modeler a priori and might consist of costs such as those from correlation and pairwise clustering. Each cost model expresses a data-dependent preference towards certain assignment solutions. For each model, rather than selecting the individual best solution, ASC aims at selecting the best set of assignments which are statistically indistinguishable. By doing so, ASC enables

- the quantification of the informativeness of the optimal set of cluster assignments and
- the selection of the best cost model, that is the one maximizing the informativeness in the prediction task.

The results of cluster validation with ASC may also be employed as initial preprocessing steps for the final aim of supervised learning [Nelles (2001)]. Figure 3.1 illustrates the position of exploratory data analysis in a diagram which represents a drastic simplification of the modeling process.

3.1.1 The Principle of Approximation Set Coding

In one of its conventional formulations, clustering aims at partitioning objects into clusters according to a cost model R. In a general setting, data available to the modeler are denoted as $D = \{d_j\}_{j=1}^{\bar{n}}$ and consist of $\bar{n}$ individual samples.

[1] Parts of this section appear in [Chehreghani et al. (2012)].

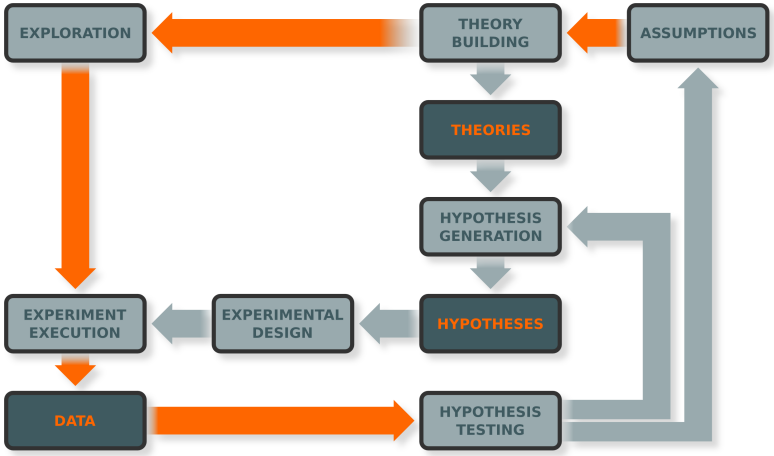


Figure 3.1: Schematic diagram illustrating exploratory data analysis and hypothesis testing in the context of modeling.

For time series analysis, the samples $d_i := \phi_i$ correspond to measured trajectories obtained from sampling a dynamical systems at arbitrarily given points in time. Individual samples are aggregated into sequences as

$$\phi_i = (y_i(t_1), \ldots, y_i(t_n)), \tag{3.1}$$

where n is the dimension of the feature space and $y_i(t)$ is the i-th component of the measurement vector taken at time point t_j, for $1 \leq j \leq s$. The data space, that is the set of datasets, is denoted as $\mathcal{D}^*$. Datasets are assumed to be finite in size, but arbitrarily large.

Given the data, the modeler minimizes a cost model. Costs are often explicitly defined by the modeler, but they may also be expressed implicitly in terms of terminating algorithmic procedures.

Definition 5. *A clustering assignment is a mapping of samples from D to labels $\lambda \in \Lambda$ following*

$$c\colon D \to \Lambda \tag{3.2}$$

with $c\colon d \mapsto \lambda$.

The assignments are elements of the hypothesis class $\mathcal{C}(D)$. The class is data-dependent, as it maps individual data points to the assigned labels.

For every trajectory $j = 1, \ldots, l$, let V_k denote the set of indexes which correspond to samples assigned to the k-th cluster. Formally, the sets are

$$V_k \coloneqq \{i \in [1, \ldots, n] \colon c(i) = \lambda_k\} \tag{3.3}$$

for $k = 1, \ldots, K$ and $K = |\Lambda|$.

Such index sets partition the sample indexes $\{1, \ldots, \bar{n}\} \subset \mathbb{N}^{>0}$ into K subsets. In this setting, individual models are identified by their associated cost functions.

Definition 6. *A cost model is a function evaluating hypothesis c given data D as*

$$R\colon \mathcal{C}(D) \times \mathcal{Y}^* \to \mathbb{R}. \tag{3.4}$$

Implicitly, cost models specify regularities which can be extracted from the data. The goal of model validation is to assess the degree of matching between the regularities expressed by the model and those exhibited by the data. Given a dataset and a cost model, the process of learning terminates as soon as an optimal solution is found. Depending on the cost model, on the algorithm, as well as on the available computational resources, the solutions which are found might be optimal in a local or global sense.

At present, there already exist several established principles and procedures for model selection. Among these, it is worth mentioning Minimum Description Length [Rissanen (1978)], Kolmogorov Structure Function [Kolmogorov (1974)], BIC [Schwarz (1978)], AIC [Akaike (1974)], Minimum Message Length [Wallace and Boulton (1968)], Solomonoff's Induction [Solomonoff (1964a,b)], PAC [Valiant (1984)], as well as PAC-Bayesian generalization bounds [Seldin and Tishby (2010a)]. These approaches rely on convincing justifications from in-

formation theory, probability and statistical learning theory[2]. Approximation Set Coding (ASC) shares the aim of the mentioned approaches, but with a different goal: that of selecting models by measuring the informativeness associated with the best set of indistinguishable solutions. Rather than evaluating individual solutions or global distributions over the hypothesis class, ASC extracts sets of predictive hypotheses. This study focuses on ASC for clustering [Buhmann (2010)]. In this setting, clustering assignments are partitioned into equivalence classes induced by noise fluctuations. The classes depend on the model: under the same noise conditions, some models are better than others at distinguishing sets of solutions. In this sense, ASC operates on the effective resolution induced by the noise.

An informal justification for ASC is the following. For a given model, learning can be seen as finding the best tradeoff between informativeness and stability of the solutions. Previous results emphasized the importance of estimation stability alone [Dudoit and Fridlyand (2002); Lange et al. (2004)], although not undisputedly [S. Ben-David (2006)]. Other appeals to stability in estimation for data analysis are based upon the concepts of well-posedness [Hadamard (1902); Kocha and Tataru (2001)] and experimental repeatability. One requirement is undisputed: in controlled settings, experiments are expected to produce similar readouts under similar conditions. Yet estimation stability alone is not enough. By itself, it does not take into account the information contained in the data. A model should not only replicate the data, but also generalize to other datasets. The modeler must be careful: there exists an overfitting risk when the amount of available information is overestimated. What is the optimal tradeoff? Reductions of estimation stability are tolerable when compensated by appropriate improvements in predictive power (measurable, for instance, in terms of generalization capacity) [Vapnik (1982); Tishby et al. (1999)]. If the modeler selects only few solutions, the selection is informative but possibly unstable [S. Ben-David (2006)]. In

[2]Discussing the individual merits of established approaches goes beyond the scope of this thesis.

fact, the set of optimal solutions for a given dataset might not coincide exactly with those selected by another dataset sampled from the same (stationary) source. The opposite tends to happen when too many solutions are selected. Sets consisting of many solutions are certainly more stable with respect to resampling, but are not particularly informative. ASC formalizes this tradeoff idea on the basis of a fundamental analogy between learning and communication. The optimal tradeoff is the one which maximizes the capacity to transmit set indexes through a noisy channel.

Empirical Minimizers and Approximation Sets

Formally, let us consider a setting with two datasets: D^1 and D^2. Each dataset is of size $\bar{n}$ and is independently sampled from the same stationary source. The asymptotic equipartitioning property guarantees that, with overwhelming probability, the datasets tend towards typicality as the sample size grows [Cover and Thomas (1991)] The requirement of two separate datasets is not limiting the generality of the approach. When a single dataset instance is available and the individual samples drawn independently from the same distribution, a two sample set scenario can be effectively obtained through random splitting of data D into sets D^1 and D^2. The scenario can be extended straightforwardly to cases in which more than two instances exist. The conclusions are analogous and there is no loss of generality.

In the learning scenario, individual solutions are evaluated on the basis of the data through the cost model. According to the prevalent convention from statistical learning theory, smaller costs indicate preferable solutions. The total ordering of the solutions is, therefore, directly induced by the cost model. The empirical minimizers are the optimal solutions for the minimization of the cost model $R(c|D^m)$ (for $m = 1, 2$) and are defined as follows:

$$\mathcal{C}^{\perp}(D^m) \coloneqq \arg \min_{c \in \mathcal{C}(D^m)} R(c|D^m). \tag{3.5}$$

For simplicity of notation, the cost of the empirical minimizers is

denoted as

$$R^{\perp}(D^m) \coloneqq R(c^{\perp}|X^m) \tag{3.6}$$

for $c^{\perp} \in \mathcal{C}^{\perp}(D^m)$, $m = 1, 2$.

To assess the predictive capacity of a model, one has to estimate the variability of the optimal solutions with respect to that of the data D^m. Due to limited knowledge, average-case analysis is not possible at this point: the data distribution is unknown to the modeler and very hard to estimate directly (predominantly due to the high dimension of the data space). Instead, performing operations directly on the solution space might be a tractable task. Even without average-case analysis, it is still possible to say something regarding the range of plausible fluctuations in the data on the basis of the empirical evidence provided by at least two instances. How to compare solutions? The two hypothesis classes are dataset-dependent. Then, to transfer solutions between instances one has to choose a mapping function from $\mathcal{C}(D^1)$ to $\mathcal{C}(D^2)$. The definition of the mapping function is a modeling choice which is necessary to assess the generalization capacity from training to test data.

The mapping function performs the mapping from the first dataset instance to the second one. It is defined as

$$\psi : \mathcal{C}(D^1) \to \mathcal{C}(D^2). \tag{3.7}$$

Relating solutions between hypothesis classes, ψ defines how the modeler intends to generalize across instances (by mapping to the nearest neighbor, for instance). For convenience of notation, subsets of solutions $A \subseteq \mathcal{C}(D_1)$ are mapped to other subsets through the sample-wise application of ψ as

$$\psi \circ A \coloneqq \{\psi(a),\ a \in A\} \subseteq \mathcal{C}(D_2). \tag{3.8}$$

Because of noise, the set of mapped empirical minimizers $\psi \circ \mathcal{C}^{\perp}(D^1)$ do not necessarily coincide with the set $\mathcal{C}^{\perp}(D^2)$. The intersection between the two is given by

$$\left(\psi \circ \mathcal{C}^{\perp}(D^1)\right) \cap \left(\mathcal{C}^{\perp}(D^2)\right). \tag{3.9}$$

The cardinality of the intersection might be small or even empty. The locations of the empirical minimizers, in fact, tend to diverge as the noise level increases.

The central idea of ASC is that of selecting larger sets of cluster solutions from the hypothesis classes to avoid instability in the estimation. Approximation sets are defined as sets of solutions which are γ-minimal.

Definition 7. *The approximation sets (AS) are*

$$\mathcal{C}_\gamma(D^m) \coloneqq \{c \in \mathcal{C}(D^m) \colon R(c|D^m) \leq R_\perp(D^m) + \gamma\}, \tag{3.10}$$

for the respective dataset instances indexed by $m = 1, 2$.

The variable γ defines how close is the cost of the accepted solutions with respect to the cost $R_\perp(D^m)$ of the empirical minimizers. Hence, the elements of the approximation sets are at last γ-optimal.

The selection of an appropriate γ permits the optimization of the tradeoff between selection stability and the informativeness of the solutions. When $\gamma = 0$, the approximation sets consist solely of the respective empirical minimizers, thus yielding unstable results. By contrast, a too large value of γ forces the approximation sets to include a large number of solutions from the hypothesis class. Large selections tend to yield results which are stable. However, the conclusions are uninformative if the selections are too large.

Learning and Communication

The optimal γ is found through information-theoretic optimization. ASC entertains the idea that learning and communication share fundamental properties: performing predictions can be seen as coding with error control and correction. The communication analogy proceeds as follows: in the sender-receiver scenario, distinguishing individual solutions based on data corresponds to transmitting messages over a noisy channel. The ability to discriminate solutions through applied transformations of the data reflects the communication capacity between the sender and the receiver. In learning as well as

in communication, success ultimately depends on the coding strategy (for fixed noise levels). For a given γ, the process of communication is defined as follows (Algs. 1 and 2). In brief, the receiver aims at decoding permutations $\sigma \in \Sigma$ which are applied by the sender to the second instance of the data. The set of indexes for the code problems is denoted as $\mathbb{I}_\sigma$.

Coding and transmission are described as follows. The applied trans-

Algorithm 1: Coding scheme.

Data: cost model R, γ, set of potential transformations Σ, hypothesis class $\mathcal{C}(D^1)$, dataset instance D^1.
Result: coding scheme.

1 SENDER and RECEIVER share R;
2 SENDER and RECEIVER share D^1;
3 SENDER calculates the γ-approximation set $\mathcal{C}_\gamma(D^1)$;
4 RECEIVER calculates the γ-approximation set $\mathcal{C}_\gamma(D^1)$;
5 **forall the** $\sigma \in \Sigma$ **do**
6 SENDER generates the transformed training optimization problem with minimization objective $R(c|\sigma \circ D^1)$;
7 SENDER shares with RECEIVER the transformed training optimization problem with minimization objective $R(c|\sigma \circ D^1)$;
8 SENDER and RECEIVER calculate the γ-approximation set $\mathcal{C}_\gamma(\sigma \circ D^1)$;
9 **end**

formation σ_{sel} is only available to the sender. However, the receiver can employ a decoding rule based on the maximization of the intersection between the available approximation sets to estimate $\widehat{\sigma}$. Decoding is possible because $\sigma_{\text{sel}} \circ X_2$ is known by the receiver as well. In contrast, σ_{sel} and X_2 remain known solely by the sender. The transmission scheme of Alg. 2 can be used to identify σ_{sel} with a certain error rate. Codebook vectors are defined from the approximation

Algorithm 2: Transmission scheme.

Data: cost model R, γ, set of potential transformations Σ, approximation sets $\mathcal{C}_\gamma(\sigma \circ D^1)$ for all $\sigma \in \Sigma$, hypothesis classes $\mathcal{C}(D^1)$ and $\mathcal{C}(D^2)$, mapping function ψ, dataset instances D^1 and D^2.

Result: set of estimated $\widehat{\sigma}$.

forall the σ_{sel} *selected by the* SENDER **do**

  SENDER applies σ_{sel} to D^2;

  SENDER sends $\sigma_{\text{sel}} \circ D^2$ to RECEIVER;

  RECEIVER calculates the γ-approximation set given $\sigma_{\text{sel}} \circ D^2$;

  RECEIVER estimates σ_{sel} with $\widehat{\sigma}$ through the decoding rule:

$$\widehat{\sigma} \in \arg\max_{\sigma \in \Sigma} \left| \psi \circ \mathcal{C}_\gamma(\sigma \circ D^1) \cap \mathcal{C}_\gamma(\sigma_{\text{sel}} \circ D^2) \right| ; \qquad (3.11)$$

end

sets. A large value for γ corresponds to few distinct vectors for coding, whereas small values produce higher error rates for decoding. The (asymptotic) transmission rate defines the tradeoff between estimation stability and informativeness on the basis of the same currency: bits of information. The goal of communication is that of achieving optimal communication (that is, maximally reliable and informative transmission).

When $\widehat{\sigma} \neq \sigma_{\text{sel}}$, the receiver performs the wrong decoding. The noise fluctuations produced a communication error. The intersection between approximation sets is given by

$$\Delta\mathcal{C}_{\gamma}^{q} \coloneqq \left(\psi \circ \mathcal{C}_{\gamma}(\sigma_q \circ D^1)\right) \cap \mathcal{C}_{\gamma}(\sigma_{\text{sel}} \circ D^2) \tag{3.12}$$

for all $\sigma_q \in \Sigma$. This term indicates the intersection between the q-th training approximation set and the test set, for $q \in \mathbb{I}_{\sigma}$.

Bounding of Error Probability

At this point, it is possible to bound the decoding error probability $p(\widehat{\sigma} \neq \sigma_{\text{sel}}|\sigma_{\text{sel}})$ [Buhmann (2010)]. Due to the union bound, one has that

$$p(\widehat{\sigma} \neq \sigma_{\text{sel}}|\sigma_{\text{sel}}) \leq \sum_{q \in \mathbb{I}_{\sigma}} P\left(|\Delta\mathcal{C}_{\gamma}^{q}| \geq |\Delta\mathcal{C}_{\gamma}^{s}| \middle| \sigma_{\text{sel}}\right). \tag{3.13}$$

The bound is necessary since direct evaluation of the error probability might be challenging and no closed-form solution is known. By introducing the indicator function

$$\mathbb{I}_{\{a\}} = \begin{cases} 1 & a \text{ is true} \\ 0 & a \text{ is false} \end{cases}, \tag{3.14}$$

the bound in Eq. (3.13) becomes

$$p(\widehat{\sigma} \neq \sigma_{\text{sel}}|\sigma_{\text{sel}}) \leq \sum_{q \in \mathbb{I}_{\sigma}} \mathbb{E}\mathbb{E}_{\sigma_{\text{sel}}}\left[\mathbb{I}_{\{|\Delta\mathcal{C}_{\gamma}^{q}| \geq |\Delta\mathcal{C}_{\gamma}^{s}|\}} \middle| \sigma_{\text{sel}}\right]. \tag{3.15}$$

In the equation, the first expectation is taken with respect to the pair of instances $\mathcal{D}^1$ and $\mathcal{D}^2$. It constitutes an annealed approximation. Since

$$\mathbb{I}_{\{|\Delta\mathcal{C}_\gamma^q| \geq |\Delta\mathcal{C}_\gamma^s|\}} = \mathbb{I}_{\{\log|\Delta\mathcal{C}_\gamma^q| \geq \log|\Delta\mathcal{C}_\gamma^s|\}}, \tag{3.16}$$

one has that, because $\mathbb{I}_{\{x\geq 0\}} \leq \exp(x)$ the bound becomes

$$\mathbb{E}_{\sigma_{\text{sel}}}\left[\mathbb{I}_{\{|\Delta\mathcal{C}_\gamma^q| \geq |\Delta\mathcal{C}_\gamma^s|\}} \middle| \sigma_{\text{sel}}\right] \leq \frac{|\mathcal{C}_\gamma(D^1)|\,|\mathcal{C}_\gamma(D^2)|}{|\Sigma|\,|\Delta\mathcal{C}\gamma^s|}. \tag{3.17}$$

Since the σ are identically distributed and sampled independently, the upper bound for the error probability of decoding is given by

$$p(\widehat{\sigma} \neq \sigma_{\text{sel}}) \leq (|\Sigma| - 1)\exp\left(-\bar{n}\mathcal{I}_\gamma(\sigma_{\text{sel}}, \widehat{\sigma})\right), \tag{3.18}$$

where the mutual information $\mathcal{I}_\gamma(\sigma_{\text{sel}}, \widehat{\sigma})$ follows the definition below. The mutual information for two approximation sets is given by

$$\mathcal{I}_\gamma(\sigma_{\text{sel}}, \widehat{\sigma}) \coloneqq \frac{1}{\bar{n}} \log\left(\frac{|\Sigma|\,|\Delta\mathcal{C}_\gamma^s|}{|\mathcal{C}_\gamma(D^1)|\,|\mathcal{C}_\gamma(D^2)|}\right). \tag{3.19}$$

At this point, it is possible to determine the optimal γ as follows. The optimal approximation threshold is

$$\gamma^* \in \arg\max_{\gamma\in[0,\infty)} \mathcal{I}_\gamma(\sigma_{\text{sel}}, \widehat{\sigma}). \tag{3.20}$$

The described procedure provides to the epistemic agent:

- the set of γ-optimal cluster solutions which are statistically indistinguishable, as well as
- an absolute measure of the informativeness of the cost model R, which is called approximation capacity.

Definition 8. *The approximation capacity (AC) for cost R and datasets D^1 and D^2 is*

$$AC[R|D^1, D^2] \coloneqq \mathcal{I}_\gamma^*(\sigma_{sel}, \widehat{\sigma}). \tag{3.21}$$

The approximation capacity can be used to compare and select costs from a set of candidates. The class of candidate cost models is the model class

$$\mathcal{R} \subseteq \{R : \mathcal{C}(D) \times \mathcal{Y}^* \to \mathbb{R}\}. \tag{3.22}$$

In practical applications, the set of candidate models may consists of different functional forms as well as of alternative parametrization of the same function.

3.1.2 Cluster Validation of Multivariate Time Series

In clustering, ASC can then be employed for

- model selection (selecting between alternative costs models, such as K-Means, pairwise clustering, or others),
- model order selection (selecting the number of clusters for an individual cost model).

When analyzing multivariate time series, the modeler might employ clustering to abstract, denoise, or compress trajectories in the measurement space. Informally, clustering addresses the question: what is the effective resolution of the data? The modeler can select from the plethora of alternative methods which is available in the literature. This study concerns the selection of relational cost models for clustering multivariate time series. Relational clustering (sometimes referred to as spectral clustering) is a general approach whose theoretical and practical importance is well recognized [Burnham and Anderson (2002)]. Characterizing time points by relations rather than vectors may prove particularly useful when the modeler does not know a priori a predictive similarity measure between trajectories. Until recently, model selection in relational clustering has been performed primarily according to heuristics and expert knowledge [Lange et al. (2004); Slonim et al. (2005)]. In this context, the Bayesian Information Criterion (BIC) is a simple and well-justified model selection method whose effectiveness has been extensively proved [Schwarz (1978); Bishop (2006)].

Definition 9. *The BIC cost is defined as follows:*

$$BIC[M|D] := -2 \log p(D|M) + \varpi(M) \log \bar{n} \tag{3.23}$$

where $\varpi(M)$ is the number of parameters for model M.

The direct application of BIC to relational clustering seems to be problematic. From the definition, BIC requires the specification of the number of free parameters. In its standard form, it can be applied directly only to finite dimensional parameter spaces. In relational clustering, the number of (effective) free parameters is not available to the modeler. Furthermore, the effective dimension may grow as a function of the sample size. As a substitute, one could employ the principle of Minimum Description Length (MDL), provided that an appropriate strategy for coding is available [Rissanen (1978); Grünwald (2007)]. Alternatively, PAC-Bayesian analysis could provide generalization bounds for model order selection [Seldin and Tishby (2010b)].

The following part concerns the automatic determination of the number of clusters as well as the selection of the cost model for clustering time series with ASC [Chehreghani et al. (2012)]. The rest of the section is organized in three parts: model order selection, model selection, and application to scientific data analysis in systems biology. The cost model class consists of pairwise clustering (PC) [Hofmann and Buhmann (1997a)] and correlation clustering (CC) [Bansal et al. (2004)]. The obtained results are compared with the selections provided by the stability criterion [S. Ben-David (2006); Dudoit and Fridlyand (2002); Lange et al. (2004)] and by BIC [Schwarz (1978)]. The model validation task is formalized as follows.

Objective 1. *Given data D, select the cost model in $R^* \in \mathcal{R}$ with the highest approximation capacity $AC[R|D^1, D^2]$.*

Approximation with Boltzmann Factors

Up to logarithmic corrections, the cardinality of the approximation sets can be estimated by borrowing the concept of canonical ensembles from statistical mechanics [Buhmann (2010)].

The partition function approximates the cardinality of a set as

$$|\mathcal{C}_\gamma(D^m)| \doteq \mathrm{Z}^m \coloneqq \sum_{c \in \mathcal{C}(D^m)} \exp\left(-\beta R(c|D^m)\right), \tag{3.24}$$

for $m = 1, 2$, where β is the inverse computational temperature and $a_n \doteq b_n$ denotes the asymptotic relation

$$\lim_{n \to \infty} \frac{1}{n} \log\left(\frac{a_n}{b_n}\right) = 0. \tag{3.25}$$

In Eq. (3.24) and (3.26), the exponential weights are the Boltzmann factors. The cardinality of the intersections can be similarly approximated by taking

$$|\Delta\mathcal{C}_\gamma(D)| \doteq \mathrm{Z}^{12} \coloneqq \sum_{c \in \mathcal{C}(D)} \exp\left(-\beta(R(c|D^1) + R(c|D^2))\right). \tag{3.26}$$

The inverse temperature β can be normalized by imposing that the average cost in $\mathcal{C}_\gamma(D)$ must yield $R_\perp(D) + \gamma$.

Relational Clustering

For a cost model R, let K denote the number of clusters. Furthermore, let X_{ij} denote to the pairwise similarity between sample i and j, for $1 \leq i, j \leq \bar{n}$. For each sample and for each cluster, the given potential h_{ik} quantifies the cost of assigning the i-th sample to cluster k, for $1 \leq i \leq \bar{n}$ and $1 \leq k \leq K$.

In the case of K-Means [Steinhaus (1957); MacQueen (1967)], the cost function minimizes the within-cluster sum of squares giving

$$R_{\mathrm{km}}(c|D) \coloneqq \sum_{k=1}^{K} \sum_{i \in V_k} \|\phi_i - \boldsymbol{\mu}_{c(i)}\|^2, \tag{3.27}$$

where $i \in V_k$ when $c(i) = \lambda_k$, and $\boldsymbol{\mu}_{c(i)}$ denotes the mean of the cluster assignment for ϕ_i. K-Means can be reduced to a model in which the potentials are given by

$$h_{i,c(i)} = \|\phi_i - \boldsymbol{\mu}_{c(i)}\|^2, \tag{3.28}$$

for all $i = 1, \ldots, \bar{n}$. The potentials consists of squared Euclidean distances between data vectors and cluster centroids. K-means is one of the most popular choices for clustering due to its simplicity. The generality of the model is restricted to Euclidean vector data. Relational clustering, in contrast, goes beyond this limitation: it clusters data in terms of pairwise similarities (or dissimilarities). In relational cost models, the potentials $\{h_{i,c(i)}\}$ may not be directly available to the modeler from the formulation. Thus, equivalent potentials are calculated through a mean-field approximation, as described below.

Relational clustering is conventionally defined in terms of an attributed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. The vertex set and the edge sets are respectively denoted as $\mathcal{V}$ and $\mathcal{E}$. The aim is that of clustering

$$\mathcal{G}_u \coloneqq \{i \colon c(i) = u\} \tag{3.29}$$

for $1 \leq u \leq K$. The set of edges between elements of cluster $\mathcal{G}_u$ and $\mathcal{G}_v$ is

$$\mathcal{E}_{uv} \coloneqq \{(i,j) \colon c(i) = u \wedge c(j) = v\}. \tag{3.30}$$

Rewritten in terms of potentials, the weight sums of Eqs. (3.24)

and (3.26) become

$$
\begin{aligned}
Z^m &= \sum_{c\in\mathcal{C}(X^m)} \exp\left(-\beta \sum_{i=1}^{\bar{n}} h^m_{i,c(i)}\right) \\
&= \prod_{i=1}^{\bar{n}} \sum_{k=1}^{K} \exp\left(-\beta h^m_{ik}\right) \\
Z^{12} &= \sum_{c\in\mathcal{C}(X^2)} \exp\left(-\beta \sum_{i=1}^{\bar{n}} (h^1_{i,c(i)} + h^2_{i,c(i)})\right) \\
&= \prod_{i=1}^{\bar{n}} \sum_{k=1}^{K} \exp\left(-\beta \left(h^1_{ik} + h^2_{ik}\right)\right).
\end{aligned}
$$

In the equations, the dataset D^m is substituted by the pairwise similarity matrix X^m, for indexes $m = 1, 2$. Similarly, the potential h^m_{ik} denotes the potential in the m-th instance. The cardinality $|\Sigma|$ is estimated as the entropy of the empirical minimizer assignment as

$$
\frac{1}{n} \log |\Sigma| \simeq -\sum_{k=1}^{K} P_k \log P_k, \tag{3.31}
$$

where P_k is the probability of the k-th cluster in $c^\perp \in \mathcal{C}^\perp(X^1)$. In general, the solution for β satisfying the necessary condition for maximization $\frac{d\mathcal{I}_\beta}{d\beta} = 0$ might be analytically unavailable. The computational cost of ASC is dominated by the estimation of the weight sums in the approximation of the partition functions. Markov Chain Monte Carlo (MCMC), as well as variational approaches, are directly applicable to the cases presented here. Subsampling might prove particularly useful with large datasets.

Model selection and model order selection are both based on the maximization of the approximation capacity described in Alg. 3. For model order selection, the initial number of cluster is denoted as $K_{\max}$. The detailed algorithmic procedures for model order selection and

cost model selection are given by Algs. 4 and 5, respectively. For model selection, let $\mathcal{R} = \{R_l\}_{l=1}^{L}$ indicate the class of candidate cost models, each with its own respective hypothesis class dependent on the similarity matrices X^m, for $m = 1, 2$.

Algorithm 3: Calculation of AC.

Data: computational temperature range $\mathcal{W}$, permutations Σ, k, cost model R, set of hypothesis classes $\mathcal{C}_k(X^m)$ and similarity matrices X^m for $m = 1, 2$.

Result: $\mathrm{AC}_k[R|X^1, X^2]$ for model order k.

forall the $\beta^{-1} \in \mathcal{W}$ **do**
    calculate the potentials h_{ik} at β^{-1} (either by annealed Gibbs sampling or annealed mean-field approximation);
    calculate the mutual information $\mathcal{I}_\beta^k$;
end
Compute the approximation capacity $\mathrm{AC}^k[R|X^1, X^2] \coloneqq \max_\beta \mathcal{I}_\beta^k$.

Algorithm 4: Model order selection.

Data: computational temperature range $\mathcal{W}$, set of transformations Σ, order range $1, \ldots, K_{\max}$, cost model R, set of hypothesis classes $\{\mathcal{C}_k(X^m)\}_{k=1}^{K_{\max}}$ and similarity matrices X^m for $m = 1, 2$.

Result: optimal model order k^*.

for $k = 2$ *to* K_{max} **do**
    Compute the approximation capacity $\mathrm{AC}^k[R|X^1, X^2]$ for k;
end
Select $\{k^*\} \coloneqq \arg\max_{1 \le k \le K_{\max}} \mathrm{AC}^k$.

The subsection proceeds by comparing two established cost models for relational clustering: pairwise and correlation clustering.

Algorithm 5: Model selection.

Data: computational temperature range $\mathcal{W}$, set of transformations Σ, order range $1, \ldots, K_{\max}$, cost model R, set of hypothesis classes $\{\mathcal{C}_k(X^m)\}_{k=1}^{K_{\max}}$ for all models in $\mathcal{R}$ and similarity matrices X^m for $m = 1, 2$.

Result: optimal model order k^*.

forall the $R \in \mathcal{R}$ **do**
  Select $\{k^*\} \coloneqq \arg\max_{1 \le k \le K_{\max}} \mathrm{AC}^k$ for the cost R;
  Compute $\mathrm{AC}^{k^*}[R|X^1, X^2]$;
end
Select $\{R^*\} \coloneqq \arg\max_{R \in \mathcal{R}} \mathrm{AC}^{k^*}[R|X^1, X^2]$.

Definition 10. *Given the similarity matrix X, the pairwise clustering cost [Hofmann and Buhmann (1997b)] for K clusters is given by*

$$R_{pc}(c|X) \coloneqq -\frac{1}{2} \sum_{k=1}^{K} |\mathcal{G}_k| \sum_{(i,j) \in \mathcal{E}_{kk}} \frac{X_{ij}}{|\mathcal{E}_{kk}|}. \tag{3.32}$$

For pairwise clustering, there exists a way to calculate exactly the potentials without distortion of the solutions through constant shift embedding (CSE) [Roth et al. (2003)]. CSE embeds relational data into a high-dimensional Euclidean vector space. In summary, CSE takes advantage of the fact that $R_{\mathrm{pc}}(c|X)$ sums the average similarities per cluster weighted by size. Since the cost is shift-invariant, no distortion of the solutions is introduced by adding a constant to all pairwise similarities. This operation just adds a constant multiplied by n to the costs of each solution. Consistent cost shifting neither changes the total ordering induced by the cost, nor the obtainable approximation capacities. The samples are embedded into a kernel space of dimension $n - 1$. The components X_{ij} of the similarity matrix are then interpreted as scalar products between the two vectors representing the objects i and j, for $1 \le i, j \le \bar{n}$.

Dissimilarities are obtained from similarities through the transformation

$$D_{\text{diss}} \coloneqq -X + \varrho, \tag{3.33}$$

where ϱ is a constant. In the kernel space, pairwise clustering solutions share the same K-means costs [Roth et al. (2003)]. Hence, the calculation of the weight sums is performed exactly by employing the analytic solution available for K-means [Roth et al. (2003)].

As shown before, for factorial models the calculation of the potentials is straightforward. For non-factorial models, the potentials are calculated through approximations. This situation applies to correlation clustering, for which no product form is known in terms of weight sums as in Eqs. (3.24) and (3.26). The approximation capacity is then calculated by mean-field approximation as follows [Chehreghani et al. (2012)]. Essentially, the cost function of correlation clustering sums the object disagreements, which are given by the negative intra-cluster edges and the positive inter-cluster ones. The cost of correlation clustering for K clusters is

$$\begin{aligned} R_{\text{cc}}(c, X) \coloneqq & \frac{1}{2} \sum_{k=1}^{K} \sum_{(i,j) \in \mathcal{E}_{kk}} (|X_{ij}| - X_{ij}) \\ & + \frac{1}{2} \sum_{k=1}^{K} \sum_{v=1}^{k-1} \sum_{(i,j) \in \mathcal{E}_{kv}} (|X_{ij}| + X_{ij}). \end{aligned} \tag{3.34}$$

The mean-field approximation holds for non-factorial models such as correlation clustering [Chehreghani et al. (2012)]. It is noteworthy that, in this case, the mean-field approximation admits an analytic form. In other cases, one could approximate the weight sums by sampling. Since the Boltzmann weights do not assume a product form, the potentials h_{ik} are determined by employing a factorial form to approximate the Gibbs distribution

$$p_{\text{Gibbs}}(c) \coloneqq w(c, X)/Z, \tag{3.35}$$

where the Boltzmann weights are

$$w(c, X) \coloneqq \exp(-\beta R(c|X)) \tag{3.36}$$

and Z is the normalizing partition function. The mean-fields of the approximating distributions correspond to the adjustable parameters. Given the potentials, the cluster assignments are conditionally independent. The distribution for the factorial form is defined with respect to the set

$$\mathcal{Q} \coloneqq \left\{ \mathbf{Q} : \mathbf{Q}(c) = \prod_{i=1}^{\bar{n}} q_{i,c(i)}, \quad q_{i,c(i)} \in [0,1] \right\}. \tag{3.37}$$

The best approximation is the one minimizing the relative entropy with respect to $p_{\text{Gibbs}}(c)$. Formally, it minimizes

$$\begin{aligned} \text{KL}\left[\mathbf{Q} \parallel P_{\text{cc}}\right] &= \sum_{c \in \mathcal{C}(X)} \mathbf{Q} \log \frac{\mathbf{Q}}{\exp(-\beta(R_{\text{cc}} - F_{\text{cc}}))} \\ &= \sum_{i=1}^{\bar{n}} \sum_{k=1}^{K} q_{ik} \log q_{ik} + \beta \mathbb{E}_{\mathbf{Q}}[R_{\text{cc}}] - \beta F_{\text{cc}}. \end{aligned} \tag{3.38}$$

The free energy is given by

$$F_{\text{cc}} \coloneqq -\beta \log Z \tag{3.39}$$

Notably, the free energy does not depend on q_{ik}. Satisfying the normalization constraint

$$\sum_{k=1}^{K} q_{ik} = 1 \tag{3.40}$$

for all $i = 1, \ldots, \bar{n}$, the minimization of $\text{KL}\left(\mathbf{Q} \parallel P_{\text{cc}}\right)$ is performed with respect to q_{ik}. One proceeds by imposing the necessary condition

$$\begin{aligned} 0 &= \frac{\partial}{\partial q_{ik}} \left[\text{KL}\left[\mathbf{Q} \parallel P_{\text{cc}}\right] + \sum_{j=1}^{\bar{n}} \Lambda_j \Big(\sum_{k=1}^{K} q_{jk} - 1 \Big) \right] \\ &= \sum_{c \in \mathcal{C}(D)} \prod_{j \leq \bar{n} : j \neq i} q_{j,c(j)}^{\mathbb{I}_{\{c(i)=k\}}} R_{\text{cc}} + \frac{1}{\beta}(\log q_{ik} + 1) + \Lambda_i. \end{aligned} \tag{3.41}$$

which determines the mean-field assignments

$$q_{ik} = \frac{\exp(-\beta h_{ik})}{\sum_{k'} \exp(\beta h_{ik'})}, \tag{3.42}$$

where

$$h_{ik} = \mathbb{E}_{\mathbf{Q}_{i\to k}}[R_{\mathrm{cc}}]. \tag{3.43}$$

The term $\mathbb{E}_{\mathbf{Q}_{i\to k}}[R_{\mathrm{cc}}]$ denotes the expectation taken over all configurations which assign object index i to the k-th cluster. Finally, the mean-field approximation of correlation clustering yields

$$\begin{aligned} h_{ik} =& \frac{1}{2} \sum_{j \leq \bar{n}: j \neq i} (|X_{ij}| + X_{ij})(1 - \mathbb{E}_{\mathbf{Q}_{i\to k}}[\mathbb{I}_{\{c(j)=k\}}]) \\ &+ \frac{1}{2} \sum_{j \leq \bar{n}: j \neq i} (|X_{ij}| - X_{ij}) \mathbb{E}_{\mathbf{Q}_{i\to k}}[\mathbb{I}_{\{c(j)=k\}}] + \varrho \\ =& \frac{1}{2} \sum_{j \leq \bar{n}: j \neq i} (|X_{ij}| + X_{ij})(1 - q_{jk}) \\ &+ \frac{1}{2} \sum_{j \leq \bar{n}: j \neq i} (|X_{ij}| - X_{ij}) q_{jk} + \varrho. \end{aligned} \tag{3.44}$$

where ϱ indicates an additive constant. The cost is decomposed into contributions depending on the i-th sample and on the cost of all other samples. Each q_{ik} is hence influenced by terms depending on the i-th sample.

Through mutual conditioning, mean-fields and probabilities can thus be effectively approximated for each β. This step is performed iteratively with an algorithm following the Expectation-Maximization procedure [Bishop (2006)], as described by Alg. 6.

Proof-of-concept for PC and CC

Below are a couple of examples for illustration. In the following, $L = 2$ and $\mathcal{R} = \{R_{\mathrm{pc}}, R_{\mathrm{cc}}\}$. First, let us consider pairwise clustering with two datasets X^1 and X^2 of dimension $\bar{n} = 800$ each. Samples

Algorithm 6: AC mean-field approximation for CC.

Data: computational temperature range $\mathcal{W}$, set of transformations Σ, order k, hypothesis classes $\mathcal{C}_k(X^m)$ and similarity matrices X^m for $m = 1, 2$.

Result: Computation of AC for CC.

```
repeat
    Calculate q_ik^(t) as a function of h_ik^(t-1);
    Calculate h_ik^(t) for given q_ik^(t);
until convergence (by iterating over t);
Calculate the weight sums of Eqs. (3.24) and (3.26);
Approximate AC^k[R_cc|X^1, X^2] for k.
```

are drawn independently from four isotropic Gaussian sources in two dimensions. The means of the sources are $\boldsymbol{\mu} = \{(\pm 4, \pm 4)\}$, the shared covariance is the scalar matrix $\boldsymbol{\Sigma} = 5\mathbf{I}_2$ (where $\mathbf{I}_2$ is the identity matrix of dimension 2), and the component parameters are $\pi_k = 1/4$ for $k = 1, \ldots, 4$. The similarity matrices X are obtained by calculating the pairwise squared Euclidean distances between objects:

$$D_{\mathrm{diss}(ij)} = \| \phi_i - \phi_j \|_2^2 := \sum_{r=1}^{n} \left(y_i(t_r) - y_j(t_r) \right)^2 , \tag{3.45}$$

for $i, j = 1, \ldots, \bar{n}$. The potentials $h_{i,c(i)}$ are calculated by annealed Gibbs sampling for a number of initial clusters up to $K_{\max} = 10$, for $1 \leq i \leq \bar{n}$. The mutual information is calculated as a function of β using the potentials, yielding the approximation capacity by maximization.

Figure 3.2 reports data and results for the example. The data are shown in Fig. 3.2(a), whereas Fig. 3.2(b) plots the trajectories of the mean of each cluster as a function of the computational temperature. As the system cools down (that is with larger β), the positions of the cluster centroids tend to diverge. The value of β is indicated by the colors of the trajectories. For $\beta \approx 0$, the color approaches dark blue

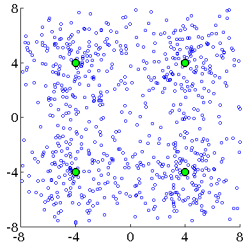
and, viceversa, values of $\beta \gg 1$ are associated to the red color. At high temperatures, all centroids tend to coincide since the process cannot distinguish but a single cluster. The contrary holds at very low temperatures, that is when too many clusters are estimated. The cluster locations are highly susceptible to the influence of the fluctuations and thus exhibit significant instability in estimation. At the optimal inverse temperature, which is denoted by β^*, there is a split from four to eight clusters (transition from green to yellow). The optimizer estimates the correct number of clusters with the optimal temperature, that is the one yielding the approximation capacity. The AC for a different number of clusters is shown in Fig. 3.2(d). Whereas the mutual information is an objective function, BIC values (represented in Fig. 3.2(c)) constitute costs to be minimized. It is noteworthy that ASC and BIC yield consistent results in this case. The plot in Fig. 3.2(e) shows the mutual information as a function of β. Its maximum defines the AC of the cost model R_{pc}. At the optimal value of the inverse computational temperature $1/\beta^*$, the approximation capacity reaches the bitrate of ≈ 1.5 bits per sample. In the noiseless case, the bitrate of the AC approaches 2 bits per sample (that is the logarithm of the number of distinct sources).

As a second example, let us consider correlation clustering with the introduced mean-field approximation. The datasets are pairs of correlation graphs X^m, $m = 1, 2$, with $\bar{n} = 1500$ and 5 source clusters. For varying level of noise, the datasets are generated as follows. At first, the underlying structure is determined by setting a positive similarity label $(+1)$ to intra-cluster edges and a negative one (-1) to inter-cluster edges. Then, each edge in $\mathcal{E}_{uv}$, for $v \neq u$, is randomly flipped to $+1$ with an independent and identically distributed (IID) Bernoulli probability $\mathcal{B}_{ern}(\epsilon|\xi)$. The parameter ξ is fixed to the value 0.35 and defines the complexity of the underlying structure. As a last step, the value of each edge $\mathcal{E}_{uv}$, for $v \neq u$, is replaced by a random value with probability η. The noise parameter η determines the observational noise. The samples are mapped by the ψ function which

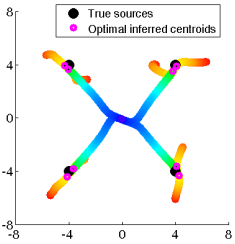
identifies samples according to their associated index. The order of the samples coincides in all data instances. The pairs are generated at noise levels ranging from 0.75 to 0.95.

The mean-field algorithm is executed with ten random initializations per model order. Initial values are varying from 1 to 10. The mutual information is calculated at each round by taking the best results in terms of costs. The approximation capacity is calculated selecting the mutual information for the best numerical approximation of β^*. The quality of the mean-field approximation is numerically verified by checking consistency with estimates obtained by Gibbs sampling.

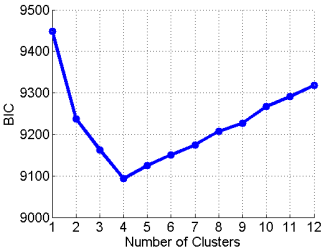
As shown by Fig. 3.3, a noise level of $\eta = 0.75$ does not present major challenges. Gibbs sampling and mean-field approximation invariably select the correct number of clusters. In case the initialization contains too many clusters, the superfluous ones are left empty. A noise level of $\eta = 0.85$ makes the task more complicated but still learnable. In this setting, significant variations due to noise are visible in the plot, which orders the samples according to the cluster assignment of the final selection. The approach determines the correct number of clusters. Higher levels of noise induce a reduced approximation capacity. Figure 3.3(b) shows that the stability criterion, based on the instability measure [Lange et al. (2004)], yields results which are consistent with ASC both for $\eta = 0.75$ and $\eta = 0.85$. The setting with $\eta = 0.95$ exhibits high level of noise. In this case, the edge labels are effectively almost entirely random. Figure 3.3(a) shows that the structure which can be extracted from the data is not learnable any more and, therefore, only a single valid cluster is obtained. Estimation instability, which is an undefined measure for a single cluster, cannot be evaluated in this case.



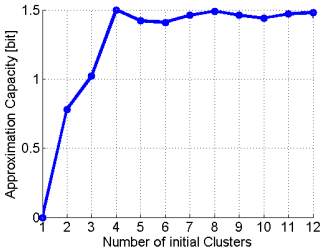
(a) Available data



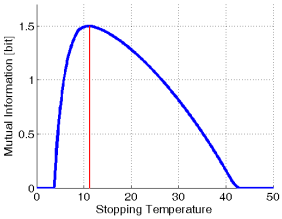
(b) Cluster trajectories



(c) BIC scores

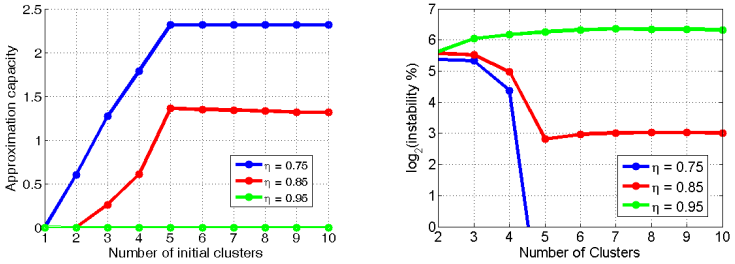


(d) Approximation Capacity



(e) Mutual Information

Figure 3.2: AC score (Fig. 3.2(d)) and BIC cost (Fig. 3.2(c)) for pairwise clustering as a function of the number of clusters K for the synthetic data (Fig. 3.2(a)). The model selection principles yield consistent results on the data. The calculation of the mutual information (Fig. 3.2(e)) for pairwise clustering has been performed with annealed Gibbs sampling as a function of β. Figure 3.2(b) shows the influence of the stopping temperature for annealed optimization on the localization of the cluster centroids (Figures from [Chehreghani et al. (2012)]).



(a) Approximation Capacities (b) Instability Measures

Figure 3.3: On the left, the approximation capacities are estimated for correlation clustering with $\xi = 0.35$ and noise parameters η taking the values0.75, 0.85 and 0.95. On the right, the instability measure yields consistent results with ASC in two out of three settings. In the remaining case with $\eta = 0.95$, instability cannot be compared: the measure is not defined for a single cluster (Figures from [Chehreghani et al. (2012)]).

Clustering of Time Series

The introduced method for model selection and model order selection is now applied to cluster the time-varying states of a biological process. The experimental data consist of readouts of gene expression profiles obtained at uniformly spaced time-points. Published data are extracted from the female digestive gland of *Mytilus galloprovincialis* [Banni et al. (2011)], an organism studied to assess the impact of environmental pollutants. The challenging nature of the application domain motivates the introduced approach: the choice of an appropriate metric to analyze trajectories is far from obvious. Relational clustering overcomes this limitation by comparing the informativeness of the solutions through ASC.

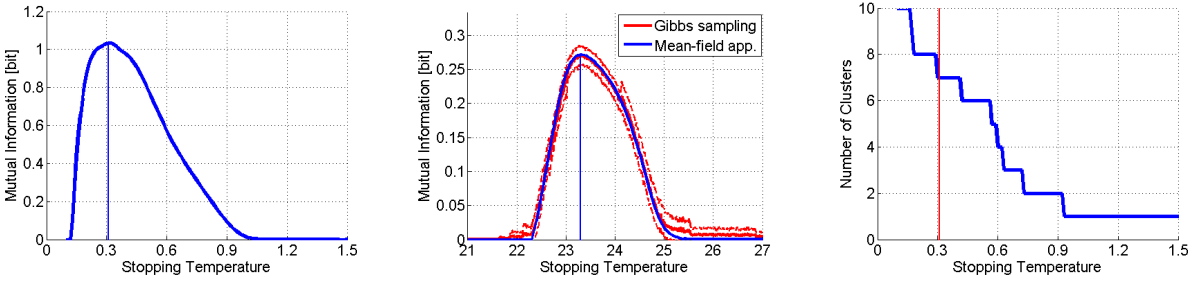
The available dataset consists of relative gene expression measurements for 12 consecutive months, with the first sample (January) taken as a reference. The dataset contains information which is valuable to the biologists to study how seasonal environmental changes affect physiology across the annual cycle of the organism [Banni et al. (2011)]. The samples consist of $\bar{n} = 295$ differentially expressed genes (with intensities evaluated on a log-scale). There exists a single dataset, from which two similarity matrices are obtained by splitting. The separation is performed by interleaving the features, thus capturing the statistical dependence and yet avoiding the risk of undersampling small clusters. Feature splitting takes advantage of the nature of the data, associating the months *(Mar, May, Jul, Sep, Nov)* to the first sample set, and *(Apr, Jun, Aug, Oct, Dec)* to the second. The similarity matrices are calculated from the respective datasets by taking the Pearson correlation coefficient for each pair of genes.

The approximation capacity is calculated by initializing the number of clusters to 10 and then by computing the mutual information at different (inverse) computational temperatures. The approximation capacities yielded by pairwise clustering and correlation clustering are plotted in Fig. 3.4 as a function of the respective inverse temperatures. The capacities demonstrate the ability of pairwise clustering over correlation clustering to capture predictable regularities in the data. The

results are revealing: under the same noise, pairwise clustering is able to discover more statistically valid structure from data than correlation clustering. On this dataset, ASC validates pairwise clustering as ≈ 3.5 times more informative than correlation clustering. The respective approximation capacities for the optimal model order are, in fact, $\mathrm{AC}_{\mathrm{pc}} = 1.03$ and $\mathrm{AC}_{\mathrm{cc}} = 0.272$. Figure 3.4(b) shows the accuracy of the mean-field approximation compared with Gibbs sampling. Whereas pairwise clustering is able to discover 7 valid clusters at optimal temperature, correlation clustering is able to detect only 2. The number of valid clusters identified by pairwise clustering at varying temperature is plotted in Fig. 3.4(c). Correlation clustering is unable to discover more since it exhibits a strong bias which favors clusters of equal size. By contrast, pairwise clustering is unbiased to size. The consistency between the optimal solution of pairwise clustering and correlation clustering is substantial: $\approx 3/4$ of the co-clustered pairs in pairwise clustering are members of the same cluster in correlation clustering. For completeness, the model selection results obtained with ASC are compared to those yielded by other established model selection criteria: BIC and the (in)stability criterion. The former is useful, well-justified and simple to apply. However, it lacks a well-understood applicability in cases where the effective number of parameters is unknown. This analysis exhibits such a limitation which is due to lower effective dimension in the feature space. Thus, the application of BIC is not straightforward in the case of pairwise clustering and even less so for correlation clustering. For pairwise clustering, the BIC score has been computed according to the effective number of dimensions, that is the ratio between the trace and the largest data eigenvalue [Kirkpatrick (2009)]. The stability criterion is a heuristic approach which shares the spirit of cross-validation [Lange et al. (2004)]. In contrast to BIC, the stability criterion is directly applicable to both cost models, but exhibits the severe limitation of being applicable only to alternatives with comparable informativeness. Figures 3.6(b) and 3.6(c) illustrate the results obtained by BIC and the stability criterion. Once again, it is worth noting the consistency between ASC

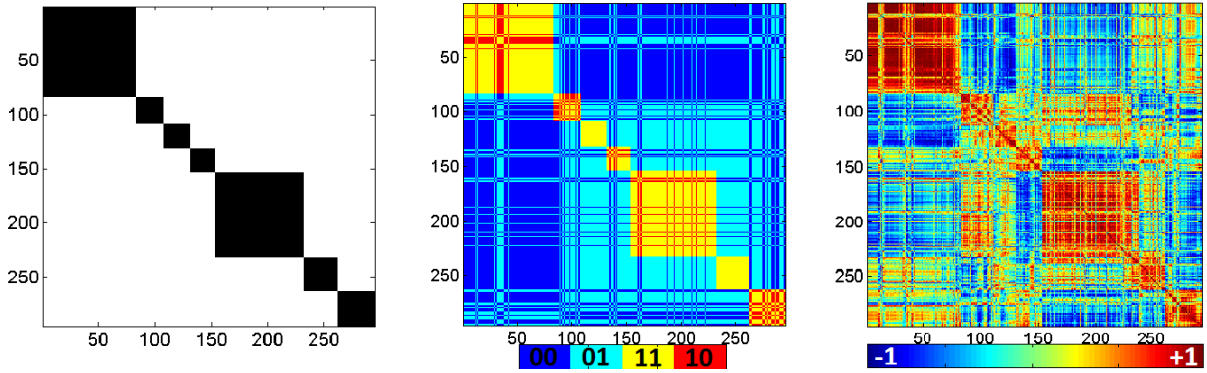
and BIC. The individual assignment of the samples to the clusters are reported in Fig. 3.5. The similarity matrix based on correlations is plotted in Fig. 3.5(c). In both figures, the samples are sorted according to increasing numerical value of the cluster labels. The induced order highlights the co-clustering structure of the optimal pairwise clustering solution, which is visible in Fig. 3.5(a). Figure 3.5(b) puts in contrast the co-clustering blocks of pairwise clustering and correlation clustering (according to the permutation of the indexes obtained for PC). Let us denote the co-clustering matrices for pairwise clustering and correlation clustering as $\mathbf{H}_{\mathrm{pc}}$ and $\mathbf{H}_{\mathrm{cc}}$, respectively. The values of H_{ij} for samples i and j, with $1 \leq i, j \leq \bar{n}$, is 1 when the label is identical, and 0 otherwise. Subscribing to this convention, Fig. 3.5(b) employs the following encoding for the pair $(H^{pc}_{ij}, H^{cc}_{ij})$: *'yellow'* for $(1,1)$, *'red'* for $(1,0)$, *'light blue'* for $(0,1)$, *'dark blue'* for the pair $(0,0)$. In brief, $(0,0)$ and $(1,1)$ denote assignment agreement, while $(0,1)$ and $(1,0)$ denote disagreement. Compared to correlation clustering, pairwise clustering is able to discover valid clusters at a higher level of resolution. The enhanced statistical detail is also consistent with the coarser results obtained through correlation clustering. Each cluster identifies an equivalence class of trajectories which are statistically indistinguishable. Elements of different clusters are, by contrast, distinguishable. The aggregate trajectories for the optimal pairwise clustering solution are reported in Fig. 3.7. The trajectories are normalized with respect to the intra-cluster means and with unitary variance. Such pre-processing of the trajectories is useful to perform parameter estimation and model selection, as discussed in the next chapters.

In summary, the task of model selection has been addressed by ASC to estimate the optimal tradeoff between solution stability and informativeness. ASC exhibited consistency with BIC in the analyzed dataset of gene expression dynamics. In this application, pairwise clustering is able to capture three times more information than correlation clustering thanks to its unbiasedness with respect to cluster size.



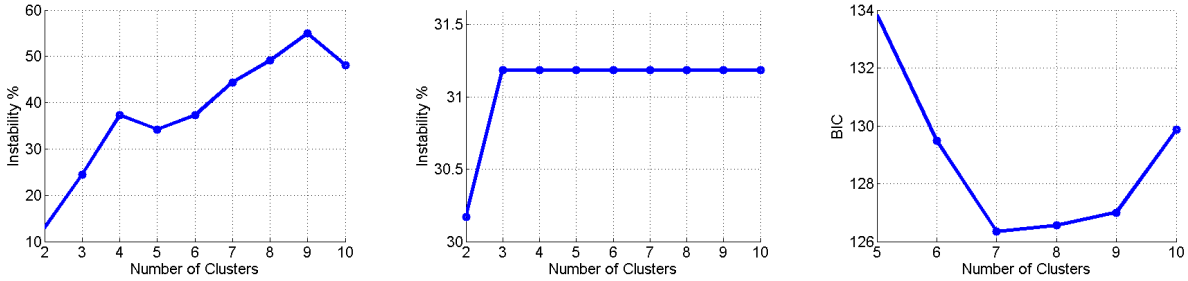
(a) Pairwise clustering (b) Correlation clustering (c) Number of clusters for PC

Figure 3.4: AC for PC (Fig. 3.4(a)) and CC (Fig. 3.4(b)) applied to time series of gene expressions from the annual cycle of *Mytilus galloprovincialis* [Banni et al. (2011)]. Figure 3.4(b) compares the estimated mutual information with mean-field approximation and with Gibbs sampling. The number of clusters for PC at different temperatures is shown in Fig. 3.4(c) (Figures from [Chehreghani et al. (2012)]).



(a) Pairwise Clustering (b) PC Vs CC (c) Correlation matrix

Figure 3.5: The optimal solution for pairwise clustering is plotted in Fig. 3.5(a) (black indicates co-clustering of the elements). Results from pairwise clustering and correlation clustering are reported and compared in Fig. 3.5(b) (with indexes permuted consistently with PC assignments). Colors are assigned according to the following rule: the first bit refers to pairwise co-clustering and the second one to correlation co-clustering. The similarity matrix of pair-wise correlations obtained from the measurement data is plotted in Fig. 3.5(c) (Figures from [Chehreghani et al. (2012)]).



(a) Instability for PC (b) Instability for CC (c) BIC for PC

Figure 3.6: The value of the instability measure is computed for PC (Fig. 3.6(a)) and CC (Fig. 3.6(a)). Figure 3.6(c) reports the results obtained for the BIC cost, which is calculated for the effective dimension. Notably, the results obtained with ASC are consistent with those of BIC (Figures from [Chehreghani et al. (2012)]).

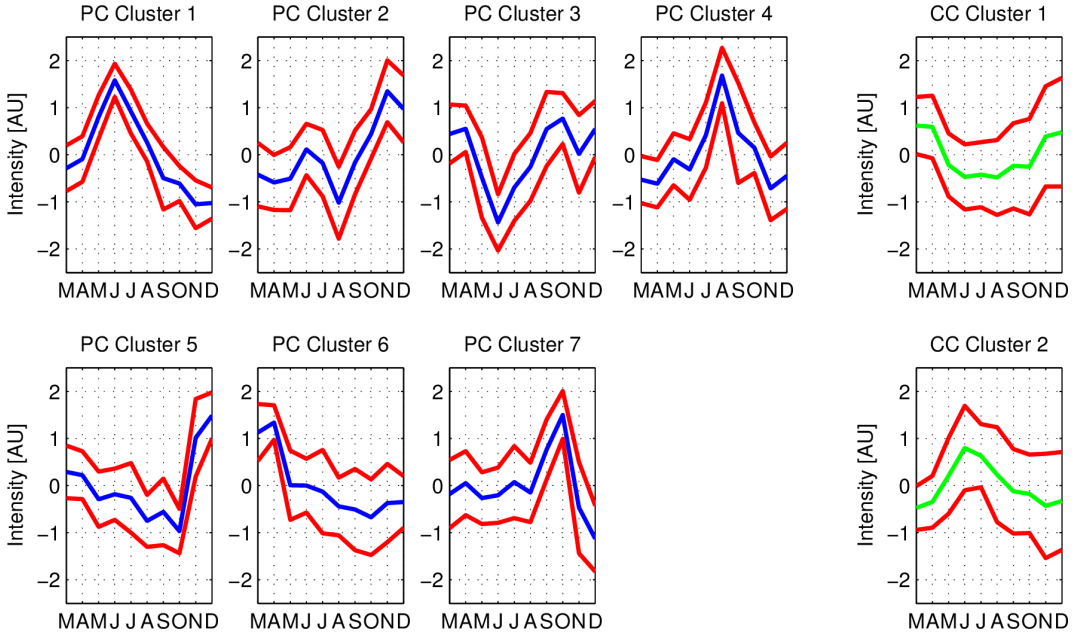


Figure 3.7: Trajectories for each cluster from the optimal cluster assignment given by pairwise (left) and correlation clustering (right). The time frame ranges from March (M) to December (D). Cluster means are plotted in blue (PC) and green (CC), while normalized SDs appear in red. The comparison of the trajectories demonstrate the higher resolution of pairwise clustering compared to correlation clustering.

3.2 Modeling of High-dimensional Sequences

"The great tragedy of Science – the slaying of a beautiful hypothesis by an ugly fact."

— T. H. Huxley

This section introduces a method for the unsupervised modeling of discrete dynamical systems[3]. Whereas the previous section aimed at quantizing the set of trajectories in the state space, the goal of this section is to model the dynamic transitions between states. The study is motivated by the current limitations in the analysis of cellular phenotypes from image data [Zhong et al. (2012)].

In combination with large-scale perturbation by RNA interference (RNAi), live-cell microscopy is a particularly informative approach to discover gene function [Conrad and W. (2010); Goshima et al. (2007); Collinet et al. (2010); Schmitz et al. (2010); Neumann et al. (2010)]. State-of-the-art methods analyze cell morphologies and phenotypes with supervised statistical models, that is with approaches which completely rely on user annotation. In such settings, the data which are available to the modeler consist of large sequences of time-resolved microscopy images. So far, the task of modeling cell morphologies has been formulated mainly as a conventional supervised learning problem [Neumann et al. (2010); Boland and Murphy (2001); Held et al. (2010); Harder et al. (2009); Wang et al. (2008); Loo et al. (2007); Jonesa et al. (2009); Conrad et al. (2011)]. By subscribing to the supervised framework, previous approaches implicitly limited the applicability of cell-based screening to cases in which extensive knowledge regarding morphology is already available. There exists another significant practical limitation associated with supervised approaches: human labeling exhibits low consistency. This problem is enormously aggravated when researchers compare results originating from different laboratories (due to bias and subtle change of controlled conditions). The following annotation-free method overcomes these two

[3]Parts of this section appear in [Zhong et al. (2012)].

issues. The method enables the automatic prediction of cell morphology dynamics from time-resolved images [Zhong et al. (2012)]. The application target of the method is the full automation of data labeling in image-based systems biology. The methodological goal is to increase objectivity, reproducibility, and validity of the results. Being unsupervised, the method has the additional advantage of being significantly faster, since it does not rely on human labor. The method is applied to unsupervised modeling of human cell phenotypes with diverse fluorescent markers and screening data. Experimental results in various experimental conditions verify the accuracy of the method, which is highly competitive even compared to state-of-the-art supervised labeling [Zhong et al. (2012)].

3.2.1 Modeling the Cell Cycle

The cell cycle exhibits an important regularity: a quasi-periodic behavior. A central modeling question is: which related aspects are amenable to predictive modeling? The obtained model demonstrates that both morphological and dynamical aspects of the process can be predicted effectively by dynamic modeling. The method is based on Temporal Constrained Combinatorial Clustering (TC3), which takes advantage of the quasi-periodicity of the cycle to initialize Gaussian mixture models (GMMs). The individual cell dynamics are captured by hidden Markov models (HMMs) through the incorporation of spatial information. Figure 3.8 illustrates the pipeline designed for unsupervised modeling of cell cycle dynamics.

Let C_j indicate the j-th cell in a population of size l. Furthermore, let x indicate the time-varying state of the cell. Assume that each cell behaves according to the following Itō stochastic differential equation (SDE):

$$\mathrm{d}x^{(j)}(t) = f(x^{(j)}(t), t, \theta)\mathrm{d}t + \sigma(x^{(j)}(t), t)\mathrm{d}\mathbf{W}_t, \tag{3.46}$$

for an unknown function f and with a standard Wiener process $\mathbf{W}_t$ whose time-dependent diffusion coefficient is $\sigma(x(t), t)$ [Wilkinson (2006)].

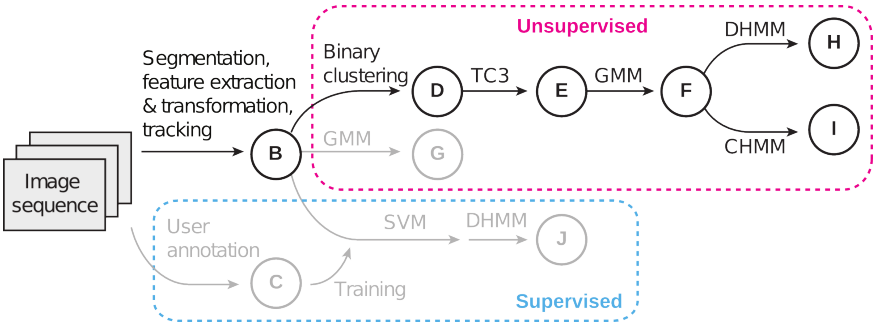


Figure 3.8: Overview of the processing pipeline for unsupervised modeling of cell cycle dynamics from microscopy images. The obtained unsupervised results are compared with the ones obtained through the supervised workflow (Figure from [Zhong et al. (2012)]).

According to this formulation, future states follow the current state with indeterminacy due to the stochastic transitions.

Measurements are performed simultaneously by sampling the trajectories $\phi^{(j)}$ of each cell, obtaining data $D = \{(t_i, y(t_i))\}_{i=1}^{n}$. At time point t_i, the measurement for the j-th cell is given by

$$y^{(j)}(t_i) = h(x^{(j)}(t_i), \nu_i^{(j)}), \tag{3.47}$$

where $h(x, \nu)$ is the measurement function for a fixed setting $\varepsilon \in \mathbf{E}$. In such setting, h is assumed to be a time-invariant non-linear vector function of the state and of the experimental error ν_i. The error distribution is estimated from the data.

The cell population is measured in sequences of images recorded by a digital microscope. The first step in the analysis is that of cell separation and segmentation. Images of the population are processed by extracting features to detect and capture the current state for each individual cell. Let $\Phi = (\phi^{(1)}, \ldots, \phi^{(l)})$ indicate the set of l synchronized multivariate time series, where each measured trajectory $\phi^{(j)}$ is a sequence of division phases of an individual cell recorded

between time frames t_1 and t_n:

$$\phi^{(j)} \coloneqq ((t_1, y^{(j)}(t_1)), \ldots, (t_n, y^{(j)}(t_n))). \tag{3.48}$$

Multivariate time series are thus represented as sequences of numerical arrays [Held et al. (2010)]. Each element of Φ is a measurement of the feature vector $\phi^{(j)} \in \mathbb{R}^{r \times n}$ in the space of r dimensions describing physiological states over time points.

At each frame, high-dimensional features can be extracted by preprocessing techniques. The obtained features are statistically dependent and their relation is, in general, nonlinear. Nonetheless, correlation can be taken as a first-order approximation to study the structure of the system. Principal Component Analysis (PCA) [Pearson (1901)], for instance, may be used to reduce dimension while preserving fundamental characteristics of the data [Pham and Tran (2006); Wang et al. (2007)]. PCA offers the possibility to transform some originally correlated feature variables into combinations which are uncorrelated. In combination with that, feature normalization can be performed by z-score standardization [Wasserman (2003)]. The noisy measurements of the trajectories PCA-reduced to r dimensions are denoted as $\widehat{\phi}$. The tensors are obtained from keeping all features accounting for up to 99% accumulative explained variance [Zhong et al. (2012)].

Definition 11. *Each dataset is a third order tensor (l-by-n-by-r) constructed as follows:*

$$\mathbf{Y} \coloneqq \begin{bmatrix} \widehat{y}^{(1)}(t_1) & \widehat{y}^{(1)}(t_2) & \cdots & \widehat{y}^{(1)}(t_n) \\ \widehat{y}^{(2)}(t_1) & \widehat{y}^{(2)}(t_2) & \cdots & \widehat{y}^{(2)}(t_n) \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{y}^{(l)}(t_1) & \widehat{y}^{(l)}(t_2) & \cdots & \widehat{y}^{(l)}(t_n) \end{bmatrix}. \tag{3.49}$$

where $\widehat{y}^{(j)}(t_i)$ is the r-dimensional reduction of the measured samples through PCA, with $1 \leq i \leq n$ and $1 \leq j \leq l$.

Objective 2. *Estimate the phase labels with the hidden Markov model Σ^* on the basis of the available dataset (that is the measurement tensor $\mathbf{Y}$).*

The full algorithmic procedure employed to address Obj. 2 is outlined in Alg. 7. Segmentation, feature extraction, and tracking are the employed pre-processing steps.

Algorithm 7: GMM-HMM estimation.

Data: $r, K \in \mathbb{N}^{>0}$, time series $\Phi^{(j)}$ for all $j = 1, \ldots, l$ cells.
Result: HMM model Σ and reconstructed trajectories.

```
apply PCA to Φ;
assign r PCA components to φ̂;
for j = 1 to l do
    cluster φ̂^(j);
    initialize GMM with TC3;
end
estimate GMM with EM;
estimate stochastic transition matrix A with the Baum-Welch
algorithm;
reconstruct Viterbi paths;
```

Constrained Clustering

After PCA dimension-reduction, the modeler extracts the mitotic subgraph from the data. This initial step already captures the two main cell division phases through binary clustering. It circumvents the harder task of directly clustering all cell division phases. Clustering is performed with K-means, that is by minimizing the cost of Eq. (3.27) with $K = 2$. The obtained clusters correspond to the two initial components of a Gaussian mixture model [Richard O. Duda (2000)]. To take advantage of the temporal regularities in quasi-periodic sequences of observations, TC3 is employed to initialize the GMMs. In

contrast to the previous section, clustering is now applied to partition samples over time.

For $K, n \in \mathbb{N}^{>0}$, the set of indexes $\{1, \ldots, n\}$ can be partitioned into K distinct clusters. If index i and j, with $j > i$, are in the same cluster $V_k \neq \emptyset$, for $k = 1, \ldots, K$, then r must be in V_k as well.

Definition 12. *Let $\zeta(n, K)$ be the number of possible clustering assignments of the measured time series $\{(t, \widehat{y}(t_i))\}_{i=1}^{n}$ into K clusters. The function obeys the recurrence relation*

$$\zeta(n, K) \coloneqq \begin{cases} \zeta(n-1, K-1) + \zeta(n-1, K) & 1 < K < n \\ 1 & K = 1 \vee K = n \\ 0 & K > n. \end{cases} \tag{3.50}$$

The complete enumeration of all possible assignments of n indexes into K clusters is given by the Stirling numbers of the second kind [Fortier and Solomon (1966); Jensen (1969)], giving

$$\mathrm{S}(n, K) \coloneqq \frac{1}{K!} \sum_{j=0}^{K} (-1)^{K-j} \binom{K}{j} j^n. \tag{3.51}$$

Conveniently, the temporal structure of the task significantly constrains the space of possible cluster solutions to be considered. As a consequence, the growth of $\zeta(n, K)$ is modest compared to that of general combinatorial clustering [Jain and Dubes (1988); Hastie et al. (2001)]. Solving the recurrence relation of Eq. (3.50), one has that

$$\zeta(n, K) \equiv \binom{n-1}{K-1} < \binom{n}{K} \leq \mathrm{S}(n, K). \tag{3.52}$$

In many cases, the number of constrained combinations satisfies the inequality

$$\binom{n}{K} \ll \mathrm{S}(n, K), \tag{3.53}$$

with the exception of cases close to the boundary conditions (that are close to $K = 1 \vee K = n$). For instance, when $n = 19$ and $K = 4$,

$$\zeta(19, 4) = 816 \ll 10^{10} \approx \mathrm{S}(19, 4), \tag{3.54}$$

which already constitutes a difference of seven orders of magnitude.

In the analysis, TC3 is applied twice. Firstly, it is applied to all cell time series in mitotic and non-mitotic subgraphs after binary clustering. The step already reduces the number of possible assignments. The search space is further reduced by the second application, which imposes a minimal size for clusters. By imposing the minimum cardinality $\min_k \ |V_k| = q$ one has that the size becomes

$$\zeta(n - (q - 1)K, K). \tag{3.55}$$

In the case of the cell cycle, the search space is thus relatively limited. Optimization can be performed exhaustively to set the best initial conditions for the GMM.

For all cells, that is for all $j = 1, \ldots, l$, the intra-cluster scatter of the trajectories is minimized with respect to the cost

$$\mathrm{W}(c|D) \coloneqq \sum_{k=1}^{K} \sum_{j \in V_k} \| \widehat{\phi}^{(j)} - \boldsymbol{\mu}_k \|_2^2, \tag{3.56}$$

where the clustering assignment $c \in \mathcal{C}(D)$ associates label λ_k to cluster set V_k for all $k = 1, \ldots, K$. Recalling Eq. (3.27), $\mathrm{W}(c|D)$ is the K-means cost for the measured trajectories. Equation (3.56) measures dissimilarities in terms of squared Euclidean distances from the cluster centroids $\boldsymbol{\mu}_k$ (for the PCA-reduced trajectories).

When $K = 6$ the cluster labels can be associated with the conventional names of the cell cycle phases [Zhong et al. (2012)]. Sequentially

ordered, the names of the labels in Λ are

$$\begin{aligned}
\lambda_1 &= \text{inter(-phase)},\\
\lambda_2 &= \text{pro(-phase)},\\
\lambda_3 &= \text{prometa(-phase)},\\
\lambda_4 &= \text{meta(-phase)},\\
\lambda_5 &= \text{ana(-phase)},\\
\lambda_6 &= \text{telo(-phase)}.
\end{aligned} \tag{3.57}$$

Definition 13. *The labeled dataset obtained from the population is*

$$D_\lambda \coloneqq \{(\widehat{\phi}^{(j)}, c(j))\}_{j=1}^{l}, \tag{3.58}$$

with $c(j) \in \Lambda$ *for all* $j = 1, \ldots, l$.

Discrete States: Mixtures and Transitions

Modeling proceeds by assigning a Gaussian component from a GMM to each phase of the cycle [Pham and Tran (2006); Wang et al. (2007)]. Discrete states are identified with the latent variables of the GMM, hence yielding $\mathcal{X} = \Lambda$.

Definition 14. *The mixture model for the measured trajectories is given by*

$$p_{GMM}(\phi) \coloneqq \sum_{k=1}^{K} \pi_k \mathcal{N}_{orm}(\phi | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{3.59}$$

where π_k *are the mixing coefficients satisfying normalization*

$$\sum_{k=1}^{K} \pi_k = 1. \tag{3.60}$$

To fit the GMM, the initial value of the mixing coefficients is given by the probabilistic membership of the samples. Means are initialized to the class discriminative sample means from D_λ. The maximum

likelihood estimation of the GMM is performed with the Expectation Maximization (EM) algorithm for fixed K.

On the basis of the obtained GMM structure, two models are obtained: a discrete and a continuous hidden Markov model [Rabiner (1989)].

Definition 15. *A stochastic transition matrix is the matrix*

$$\mathbf{A} := [a_{i,j}] \in [0,1]^{K \times K}, \tag{3.61}$$

with transition probabilities $a_{i,j}$ such that

$$\sum_{j=1}^{K} a_{i,j} = 1, \ i \in \mathcal{Y} \tag{3.62}$$

for all $i = 1, \ldots, K$.

In general, HMMs are represented in terms of a single (often discrete, but possibly continuous) random state variable $x(t) \in \mathcal{X}$.

Definition 16. *The HMM Σ_{HMM} is defined by*

- *the transition model $p(x(t_i)|x(t_{i-1}))$,*
- *the observation model $p(y(t_i)|x(t_i))$,*
- *and the initial state distribution $p(x(t_0))$.*

The stochastic transitions between the K hidden states of a discrete HMM are governed by the matrix $\mathbf{A}$. Measured emissions are noisy, satisfying the assumption of identically distributed and conditionally independent noise terms. In the continuous case, the observation density for each state is given by the corresponding Gaussian component $\mathcal{N}_{\text{orm}}(\phi|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. Transition probabilities are constrained to traverse hidden states in a specific forward order by imposing

$$a_{i,j} = 0, \tag{3.63}$$

for $j < i$ or $j > i + 1$ and with

$$a_{i,j} \neq 0, \tag{3.64}$$

with $j = 1$ and $i = K$. By applying the Baum-Welch algorithm [Baum et al. (1970)] in combination with the Viterbi algorithm [Viterbi (1967)], parameters and hidden states are estimated from the dataset D_λ. The estimated states of the learned HMM system Σ are denoted as $\tilde{x}(t_i)$ for $i = 1, \ldots, n$. When a priori domain knowledge is not available, validation of model order can be performed with ASC [Chehreghani et al. (2012)].

3.2.2 Verification and Application to Real Data

The method is evaluated in its application to diverse fluorescent markers and screening data for the classification of cell phenotypes [Zhong et al. (2012)]. Precisely, reference data consist of time-lapse microscopy images of human tissue culture cells (HeLa "Kyoto" cells). The fluorescent chromatin marker employed in the study is the histone H2B-monomeric mCherry. Object detection, tracking of cells over time, and feature extraction are carried out using the image analysis framework CellCognition [Held et al. (2010)]. Detailed image segments with user-labeled phases are shown in Fig. 3.9.

The applied aim of modeling is that of predicting cell morphology labels from image data. In this experimental setting, the number of distinct phases is known to the modeler. The labels correspond to interphase and to the phases of mitosis: prophase, prometaphase, metaphase, anaphase and telophase.

Figure 3.10 shows that user annotation exhibits significant inconsistency. The image visualizes the degree of dissimilarity of three trained biologists with respect to a gold standard (GS). The standard is given by the consensus approximation to ground truth obtained through a majority vote of multiple user label annotations. The dissimilarity matrix reveals relatively modest inconsistency for the same user on different days. By contrast, the inconsistency in the anno-

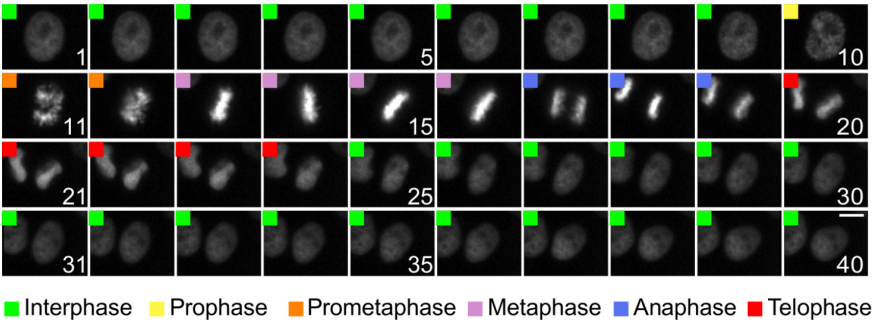


Figure 3.9: Fluorescence images from time-lapse microscopy of a live HeLa cell with user-labeled morphology phases. The cell expresses Histone 2B fused to mCherry. The time lapse is of 4.6 min per frame with a scale bar of 10 μm (Figure from [Zhong et al. (2012)]).

tations performed by different users is statistically and scientifically significant.

The analysis is based on seven image sequences with 326 cell division events. The images are sampled uniformly in time with time-lapse interval of 4.6 min. Segmentation is performed with CellCognition [Held et al. (2010)]. The total number of cells is $l = 13,040$. The result of the extraction of the principal components is reported in Fig. 3.11.

After the application of PCA, the labels obtained from the data with various methods are compared. Figure 3.12 visualizes the labels with the color-coding of Fig. 3.9. Each of the $l = 51$ rows in the matrices is a labeled cell trajectory in $n = 40$ consecutive time points from $\mathcal{T}^{\downarrow}$. The GMM alone exhibits a poor match with respect to user annotation. The direct application of the GMM results in low predictive capacity, predominantly due to sensitivity to local maxima and to initial conditions. In Fig. 3.12, comparison with other approaches demonstrates the benefits of incorporating additional knowledge for effective learning of cell morphologies. To improve classification accuracy, TC3 is used to initialize the GMM as described in the pre-

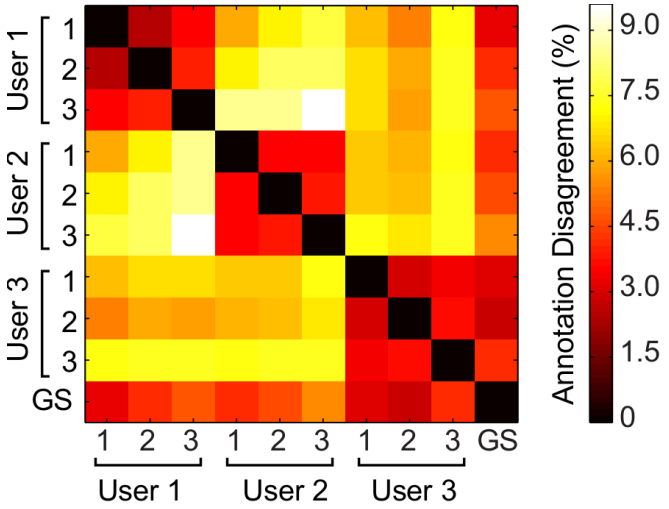


Figure 3.10: Label dissimilarity for three different annotators (Users 1 to 3) with respect to the gold standard obtained by majority vote. The plot shows significant inconsistency between different users (Figure from [Zhong et al. (2012)]).

vious subsection. Further improvements are obtained by performing state reconstruction with the estimated hidden Markov model. The stochastic transition matrix $\mathbf{A}$ identifies the dynamics of the system Σ, while individual cell morphologies are predicted on the basis of the observation densities. The analysis show that the obtained model predicts well the dynamics of the cell cycle, and yields labels which are consistent with user annotation.

For completeness, the performance of the unsupervised method which combines TC3, GMM and HMM is compared with the state-of-the-art supervised method: support vector machine classification with HMM correction. Overall, the SVM approach yields relatively low error rates in post-classification inspection. The largest contribute to the uncertainty of state estimation is due to the regions of transition

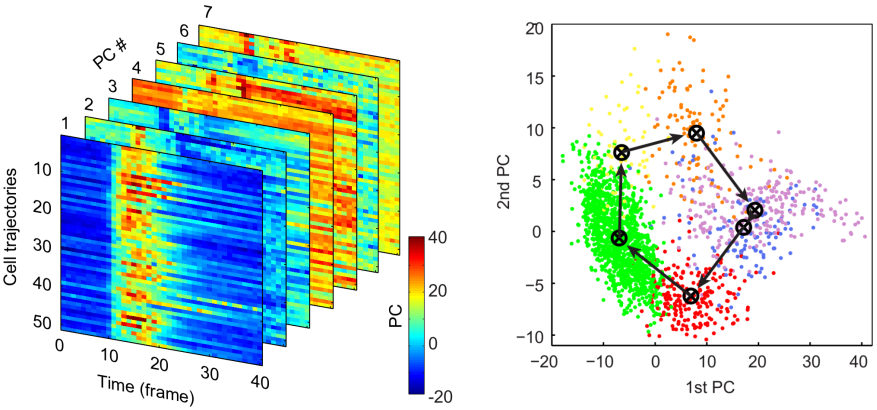


Figure 3.11: On the left, the plot visualizes the first seven principal components of 51 cell trajectories. On the right, the distribution of the samples is projected with respect to the two main principal components. Color-coding is consistent with that of Fig. 3.9. Dots and black crosses indicate, respectively, individual cell objects and sample means per component according to human annotation. Temporal progression is indicated by the cyclic arrows (Figure from [Zhong et al. (2012)]).

between phases. Ambiguity in transitional cell morphology makes the estimation of absolute error rates particularly challenging. Validation is performed as follows: an SVM is independently trained on cell trajectories from 7 image sequences, using the other 6 image sequences as test datasets. Detailed results of precision, recall, and F-score are reported in Tab. 3.1. Figure 3.13 visualizes the total accuracy of each approach. In the analysis, true positives, true negatives, false positives, and false negatives are respectively denoted as TP, TN, FP,

FN. The measures are

$$\begin{aligned} \text{Precision} &:= \frac{\text{TP}}{\text{TP}+\text{FP}}, \\ \text{Recall} &:= \frac{\text{TP}}{\text{TP}+\text{FN}}, \\ F\text{-score} &:= 2\frac{\text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}}. \end{aligned} \tag{3.65}$$

The results show that supervised SVM learning with HMM error correction yields slightly lower mean total accuracy than the unsupervised method, as shown in Fig. 3.13. Overall, unsupervised learning achieves classification accuracy comparable to that of supervised methods and is more objective [Zhong et al. (2012)].

Application of the modeling approach to additional markers in live imaging of HeLa cells expressing EGFP-tagged proliferating cell nuclear antigen (PCNA) yields labels that closely match user annotation with 86.7% total accuracy [Zhong et al. (2012)]. Total accuracy of 75.5% is achieved when modeling mitotic spindle dynamics of HeLa cells expressing fluorescently tagged α-tubulin. This analysis demonstrates that there exists a non-negligible degree of confusion between midbody and interphase morphologies [Zhong et al. (2012)].

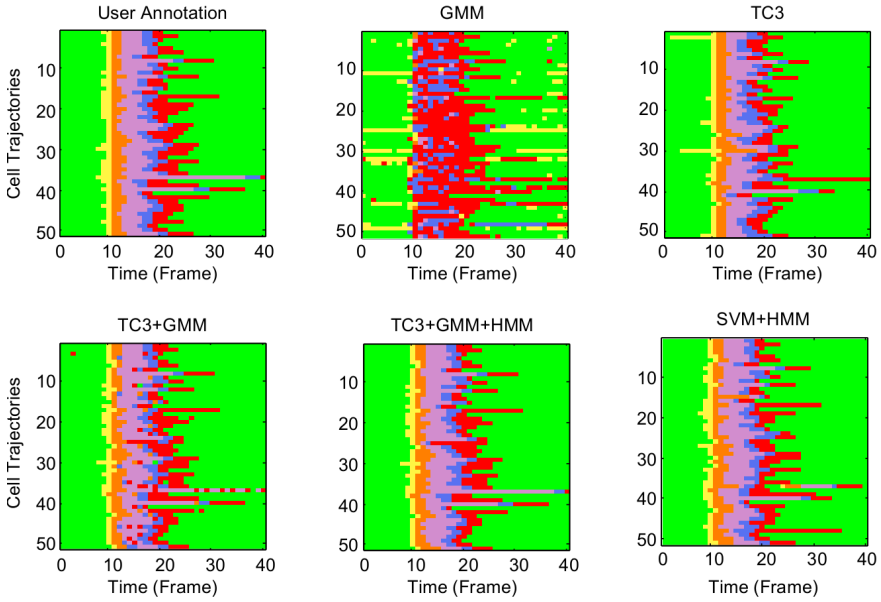


Figure 3.12: Organized plot of the labeled trajectories of all cells from the image containing the example in Fig. 3.9. Progressing clock-wise from the upper-left corner, the plots visualize the labels obtained from GMM (alone), TC3 (alone), GMM initialized by TC3, HMM extension of GMM with TC3, and SVM followed by HMM error correction (Figure from [Zhong et al. (2012)]).

Table 3.1: Method comparison: precision, recall, and F-score

Approach			**Precision**			
	Interphase	Prophase	Prometa	Metaphase	Anaphase	Telophase
TC3	93.32 ± 0.73	73.90 ± 5.08	70.02 ± 2.99	87.66 ± 2.17	49.30 ± 4.42	69.90 ± 5.39
TC3+GMM	94.86 ± 0.84	85.43 ± 2.81	77.69 ± 3.00	98.12 ± 0.56	74.57 ± 8.54	79.14 ± 5.33
TC3+GMM+DHMM	95.97 ± 0.83	83.53 ± 2.07	91.47 ± 2.45	96.82 ± 0.92	80.57 ± 7.67	84.57 ± 5.28
TC3+GMM+CHMM	96.68 ± 0.61	83.59 ± 2.64	85.65 ± 3.07	94.82 ± 0.98	77.60 ± 7.63	77.60 ± 4.36
SVM+DHMM	95.46 ± 0.50	76.54 ± 1.69	80.42 ± 2.24	93.89 ± 0.86	79.49 ± 2.76	78.17 ± 4.29
			Recall			
	Interphase	Prophase	Prometa	Metaphase	Anaphase	Telophase
TC3	98.21 ± 0.38	68.96 ± 3.17	83.98 ± 4.23	62.16 ± 3.10	62.23 ± 3.24	57.20 ± 2.84
TC3+GMM	99.18 ± 0.30	80.97 ± 3.88	83.24 ± 4.79	79.69 ± 4.46	70.36 ± 5.53	77.27 ± 4.84
TC3+GMM+DHMM	99.51 ± 0.32	82.75 ± 4.13	84.43 ± 2.96	88.24 ± 3.63	80.22 ± 6.24	79.50 ± 5.09
TC3+GMM+CHMM	97.89 ± 0.75	85.11 ± 3.27	85.90 ± 4.62	85.90 ± 3.25	72.67 ± 5.51	83.06 ± 4.77
SVM+DHMM	97.20 ± 0.41	79.35 ± 2.99	85.08 ± 1.39	84.90 ± 2.51	81.19 ± 2.14	76.23 ± 3.47
			F**-score**			
	Interphase	Prophase	Prometa	Metaphase	Anaphase	Telophase
TC3	95.69 ± 0.41	70.74 ± 3.39	75.68 ± 2.19	72.37 ± 2.20	54.65 ± 3.87	62.37 ± 3.53
TC3+GMM	96.96 ± 0.37	82.64 ± 2.41	79.70 ± 2.95	87.51 ± 2.73	72.12 ± 6.97	77.77 ± 4.71
TC3+GMM+DHMM	97.69 ± 0.36	82.84 ± 2.62	87.64 ± 2.35	92.05 ± 2.00	80.03 ± 6.79	81.51 ± 4.70
TC3+GMM+CHMM	97.25 ± 0.23	83.98 ± 2.04	85.10 ± 2.89	89.86 ± 1.54	74.89 ± 6.46	79.90 ± 4.10
SVM+DHMM	96.32 ± 0.33	77.66 ± 1.64	82.48 ± 0.89	89.05 ± 1.50	79.99 ± 1.37	76.21 ± 2.31

Means and standard errors of the means for each phase and method. The final model is selected on the basis of grid search with 5-fold cross-validation. In the evaluation, ground truth is defined by single user annotation (Table from [Zhong et al. (2012)]).

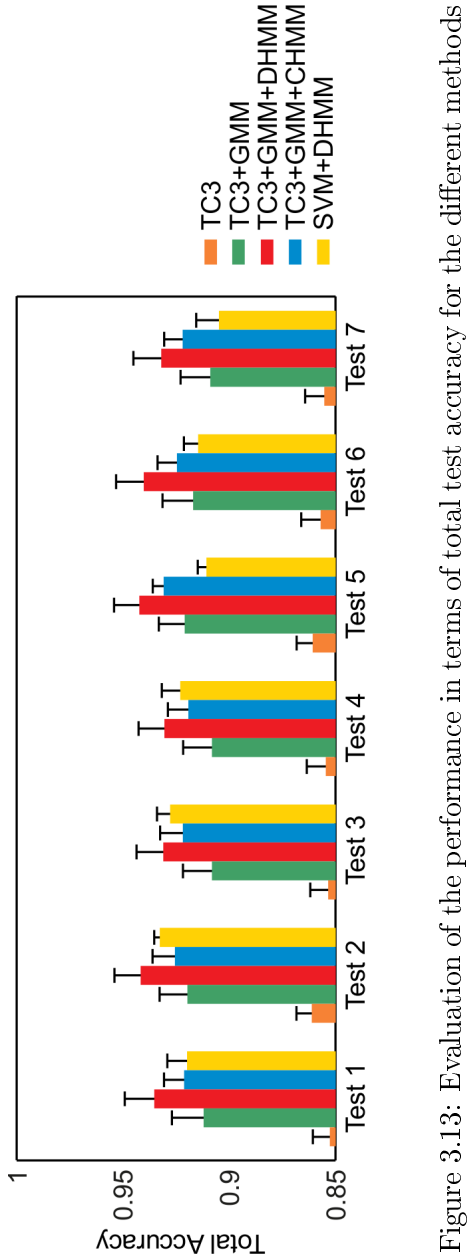


Figure 3.13: Evaluation of the performance in terms of total test accuracy for the different methods (means and standard error of the means) (Figure from [Zhong et al. (2012)]).

Chapter 4

Supervised Learning

"Prediction is very difficult,
especially about the future."

— N. Bohr (attr.)

4.1 Clustered Filtering

Given a model M and data D, parameter estimation can be formulated as the task of evaluating parameters on the basis of the data[1]. In the Bayesian setting, the aim is to calculate the posterior distribution $p(\theta|D)$ over the parameter space Θ. When the model M corresponds to a parametric form for the transition function f of a dynamical system, θ incorporates all free parameters of f. Recalling Def. 2 and Eq. (3.46), the following section refers to the parameters of $\mathrm{F}_{\mathrm{ODEs}}$ for ODEs, as well as for the deterministic component of SDEs.

Objective 3. *Calculate the parameter posterior $p(\theta|D)$ for the free parameters $\theta \in \Theta$ given the data D.*

[1] Parts of this section appear in [Busetto and Buhmann (2009)].

Since model selection is often performed taking into account ranges of plausible parameter values, parameter estimation plays an important role in system identification. Uncertainty in parameter estimates might propagate significantly to higher levels of abstraction. In many cases, it is convenient to reduce parameter estimation to generalized state estimation. One can, in fact, consider parameters as additional components of the generalized state space $\mathcal{X}_\theta \coloneqq \mathcal{X} \times \Theta$. Generalized states are defined as follows.

Definition 17. *The generalized state $x_\theta \in \mathcal{X}_\theta$ is obtained by incorporating the parameters as extensions of the state components in*

$$x_\theta(t) \coloneqq \begin{bmatrix} x(t) \\ \theta \end{bmatrix}, \tag{4.1}$$

where the parameter vector $\theta \in \Theta$ contains all free parameters for a certain functional form f identified by model M.

The parameters are then considered as additional states which are not time-varying and whose initial condition is unknown by the modeler [Nelles (2001); Doucet and Tadič (2003)]. The transition function of the extended model is denoted as f_θ. The role of supervised learning in the modeling process is highlighted in Fig. 4.1.

4.1.1 Avoiding Approximation Divergence

In almost all non-linear cases of practical interest, no analytic formulation is available for the parameter posterior. As a consequence, the estimation is typically performed numerically, and the quality of the inference process is limited by the available computational resources. In general, the dominant bottlenecks are the evaluation of the likelihood and the high-dimensional integration of the evidence [Bishop (2006)].

There exist multiple application domains in which linear-Gaussian assumptions rarely yield acceptable results. Together with economics and chemistry, systems biology makes no exception [Kitano (2002)].

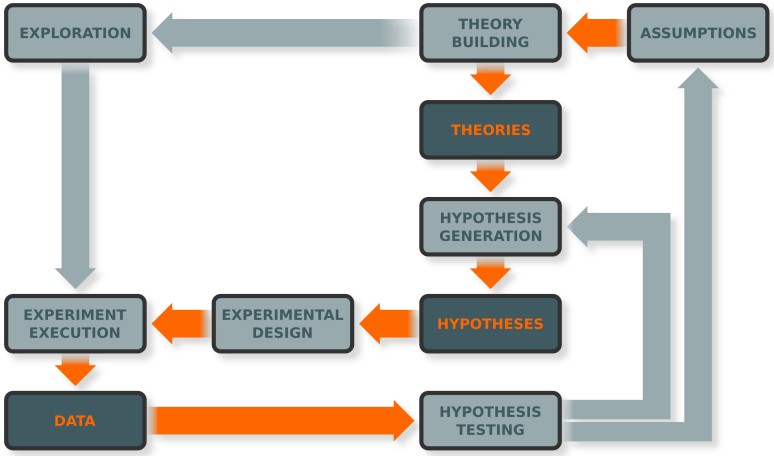


Figure 4.1: This reduced representation of the modeling process specifies the position of the supervised approaches discussed in this section.

In such domains, particle filters are generally applicable approaches which often enable effective state-space estimation [Gordon et al. (1993); Doucet and Tadič (2003)]. Alternatively, the modeler could aim at approximate Bayesian computation methods, which approximate the likelihood function, for instance by performing rejection sampling with a given tolerance [Sunnåker et al. (2012)]. Exact-likelihood sampling approaches are approximate inference techniques which are more computationally intensive, but exhibit the potential to reproduce rather complex multimodal belief states over $\mathcal{X}_\theta$ [Banga (2008)]. Particle filtering is based on sequential resampling of the posterior and are asymptotically correct as the number of samples grows. However, straightforward filtering suffers from well-known degeneracy problems since the resampling phase exhibits a systematic loss of information [King and Forsyth (2000)]. Precisely, the method

produces the approximate posterior

$$\tilde{p}(\theta|D) \simeq p(\theta|D) \tag{4.2}$$

which exhibits a tendency to diverge from the correct posterior $p(\theta|D)$. The quality of the approximation can be quantified in terms of relative entropy, that is a fundamental measure of information loss. Recalling Eq. 2.30, the approximation should aim at minimizing the number of additional bits which is given by $\mathrm{KL}[p \parallel \tilde{p}]$. Increased computational power is sufficient but not necessary to address the issue of estimation divergence. This section introduces a method to mitigate the problem of divergence by keeping track of multimodal posteriors [Busetto and Buhmann (2009)].

The learning scenario is the following. Time series data Φ of sample size n are given to the modeler, who knows the functional form of the physical system Σ^*. The parameters of the model are uncertain and subject to a given prior density $p(\theta)$. The parameter posterior follows from Bayes' rule as

$$p(\theta|\Phi) = \frac{p(\Phi|\theta)p(\theta)}{p(\Phi)}, \tag{4.3}$$

and the same holds for the posterior of the extended space $p(x_\theta|\Phi)$. The goal is the approximation of the posterior given the data at time t_n. At any time point, the posterior incorporates all the knowledge available to the modeler. The posterior is the solution of the optimal filtering problem [Arulampalam et al. (2002)]. Assuming IID noise, the likelihood can be factorized and, consequently, Bayes' rule can be calculated recursively for each sample point. This property permits incremental storing of the data and enables online inference. Recursive Bayesian filters are based on the iteration of two stages: update and propagation. During the former, new data are incorporated into the belief state. During the latter, the belief state is propagated with its uncertainty following drift and diffusion between measurements.

The belief state at time t is denoted as $p_t(x_\theta|\Phi_t)$, where

$$\Phi_t \coloneqq \{(t_i, y(t_i))\}_{i=1}^{j} \tag{4.4}$$

is the time series until the largest $j \leq n$ such that $t_j \leq t$ with all $t_i \in \mathbb{T}^{\downarrow}$. In the time interval $[t_j, t_{j+1})$, that is between measurements, the belief state follows the dynamics of the Kolmogorov forward equation of the system [Risken (1996)]. For a system of SDEs, the probability in the extended state space is given by the Fokker-Planck equation

$$\frac{\partial p_t(x_\theta|\Phi_t)}{\partial t} = -\nabla \cdot [f_\theta(x_\theta, t)p_t(x_\theta|\Phi_t)] + \Delta[\text{Đ}(x_\theta, t)p_t(x_\theta|\Phi_t)] \tag{4.5}$$

from a given initial condition $p_t(x_\theta|\Phi_t)$ and where $\text{Đ}(x_\theta, t)$ is the extended diffusion tensor [Risken (1996)]. Equation (4.5) is a drift-diffusion equation: the drift follows the deterministic component, while the diffusion is due to the stochastic terms of the SDEs. In the case of ODE models, only the drift term exists, yielding the simpler form

$$\frac{\partial p_t(x_\theta|\Phi_t)}{\partial t} = -\nabla \cdot [f_\theta(x_\theta, t)p_t(x_\theta|\Phi_t)]. \tag{4.6}$$

As soon as the new sample $(t_i, y(t_i)) = \Phi_i \setminus \Phi_{i-1}$ is available, the time-varying posterior is recursively updated following

$$p_{t_i}(x_\theta|\Phi_{t_i}) = \frac{p(y(t_i)|x_\theta, t_i)p_{t_i}(x_\theta|\Phi_{t_{i-1}})}{p(y(t_i|t_i))}, \tag{4.7}$$

where $p_{t_i}(x_\theta|\Phi_{t_{i-1}})$ is the solution of the Kolmogorov forward equation from the initial condition $p_{t_{i-1}}(x_\theta|\Phi_{t_{i-1}})$. Algorithm 8 describes recursive Bayesian filtering given the initial belief state $p_{t_1}(x_\theta)$.

Sequential Monte Carlo (SMC) methods numerically approximate through sampling the stages of propagation and update as in Alg. 8. The posterior density is approximated by a discrete set of n_p weighted (pseudo-)random samples. For compactness, let $y_{1:k}$ denote the sequence $y(t_1), \ldots, y(t_k)$ and let $(x; t_{1:k})$ denote $x_\theta(t_1), \ldots, x_\theta(t_k)$ for $k \leq n$. Then, the update follows

$$p(x; t_k|y_{1:k}) = \frac{p(y_k|x; t_{1:k})p(x; t_k|y_{1:k-1})}{p(y_k|y_{1:k-1})}. \tag{4.8}$$

Algorithm 8: Recursive Bayesian filtering.

Data: Φ and $p_{t_1}(x_\theta)$, f and Đ
Result: $p_{t_n}(x_\theta|\Phi)$.

1 **for** $i = 1$ *to* n **do**
2 | in $[t_{i-1}, t_i)$: calculate the Kolmogorov forward equation from $p_{t_{i-1}}(x_\theta|\Phi_{t_{i-1}})$;
3 | at t_i: calculate the posterior $p_{t_i}(x_\theta|\Phi_{t_i})$;
4 **end**

Definition 18. *Let us denote as* $\{x^i, w^i(t)\}$ *a random measure for the posterior* $p(x_{\theta,0:k}|\{y_j : t_j < t_k\})$*, for* $i = 1, \ldots, n_p$*. The measure is approximated by*

$$\tilde{p}(x_{\theta,0:k}|\{y_j : t_j < t_k\}) \coloneqq \sum_{i=1}^{n_p} w_k^i \delta(x_{\theta,0:k} - x_{\theta,0:k}^i). \tag{4.9}$$

The weights are normalized to one

$$\sum_{j=1}^{n_p} w^j = 1 \tag{4.10}$$

and obtained according to importance sampling [Doucet (1998)]

$$w_k^i \propto \frac{p(x_{\theta,0:k}|\{y_j : t_j < t_k\})}{q(x_{\theta,0:k}|\{y_j : t_j < t_k\})}. \tag{4.11}$$

The proposal importance density can be factorized as

$$\begin{aligned} q(x_{\theta,0:k}|\{y_j : t_j < t_k\}) =& q(x; t_k|x_{\theta,0:k-1}, \{y_j : t_j < t_k\}) \\ & \cdot q(x_{\theta,0:k-1}|\{y_j : t_j < t_{k-1}\}). \end{aligned} \tag{4.12}$$

The factorization leads to the recursive update of the time-varying weights

$$w_k^i \propto w_{k-1}^i \frac{p(y_k|x_{\theta,k}^i) p(x_{\theta,k}^i|x_{\theta,k-1}^i)}{q(x_{\theta,k}^i|x_{\theta,k-1}^i, y_k)}. \tag{4.13}$$

At this point, the system of SDEs can be numerically integrated (for instance with the Euler-Maruyama method) to sample from

$$q(x^i_{\theta,k}|x^i_{\theta,k-1}, y_k) = p(x; t_k|x^i_{\theta,k-1}). \tag{4.14}$$

Iteratively, this means that

$$w^i_k \propto w^i_{k-1} p(y_k|x^i_{\theta,k}), \tag{4.15}$$

where the likelihood in $p(y_k|x^i_{\theta,k})$ is defined by the measurement process (assumed to be known).

After several iterations, however, most of the weight mass might already concentrate on only few samples. In this divergent behavior, the overwhelming majority of the particles are updated despite their overall negligible contribution to the approximation [King and Forsyth (2000)]. The problem leads to significant waste of the available computational resources. Two techniques are typically employed to counteract the divergence: importance resampling and careful design of proposal densities. In standard resampling, the new set of samples $\{x^{i*}_{\theta,k}\}^{n_p}_{i=1}$ is obtained from the approximated density. Resampling is useful and, in many cases, even necessary, but suffers from two limitations: sample impoverishment and unimodal divergence. The former issue refers to the tendency towards a systematic reduction of the diversity of the sample set. Divergence is due to the inability to maintain multimodal densities: there is a systematic bias which favors unimodal densities. The presented approach addresses both issues with resampling through clustering and appropriate proposal densities [Busetto and Buhmann (2009)].

Resampling from the approximating discrete density involves a loss of information. Over time, iterated resampling behaves as a generally inhomogeneous Markov chain [King and Forsyth (2000)]. The chain exhibits a tendency to make the sample set collapse into a single mode, diverging from the possibly multimodal posterior. For illustration, consider the following simple example in the state space $\mathcal{X}_{\{A,B\}} = \{A, B\}$. Let $2m$ be the available number of samples, and

$m_A(t)$ the number of samples which are in state A at time t. The probability distribution $p(x) = \mathcal{B}_{ern}(x|1/2)$ is time-independent. At time t, the time-varying approximation is

$$\tilde{p}(x) \simeq \frac{m_A(t)}{2m}\mathbb{I}_{x=A} + \frac{2m - m_A(t)}{2m}\mathbb{I}_{x=B}. \tag{4.16}$$

The approximation is accurate when $m_A(t) \approx m$. Since one has that

$$\tilde{p}(m_A(t_i)|m_A(t_{1:i-1})) = \tilde{p}(m_A(t_i)|m_A(t_{i-1})), \tag{4.17}$$

the process of resampling constitutes a Markov chain which is homogeneous and has the $2m + 1$ states $0, \ldots, 2m$. By resampling with replacement,

$$p(m_A(t_i) = j) = \sum_{k=0}^{2m} p(m_A(t_{i-1}) = k)\binom{2m}{j}\frac{k}{2m}^{j}\frac{2m-k}{2m}. \tag{4.18}$$

The two degenerate states 0 and $2m$ are attractors and, consequently, $m_A(t)$ inevitably tends to diverge from the correct value [King and Forsyth (2000); Busetto and Buhmann (2009)]. Figure 4.2 visualizes the divergence probability of iterative resampling.

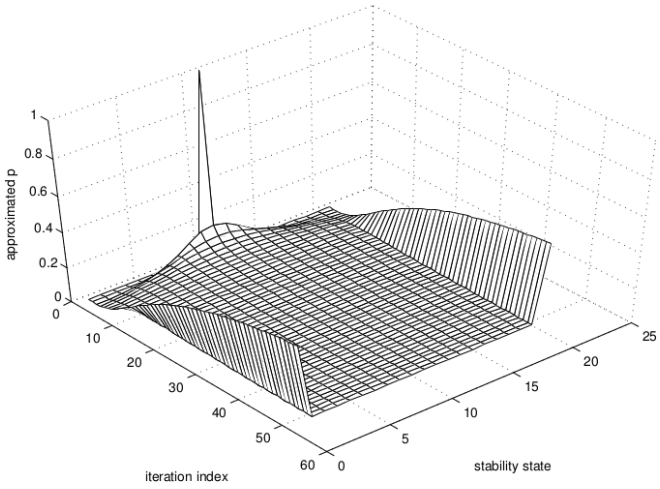


Figure 4.2: Starting from a balanced approximation with $m_A(1) = m = 10$, iterative resampling rapidly increases the probability of reaching one of the attractors at the boundaries. In the case study of Eq. (4.16), the two degenerate representations ($m_A = 0$ and $m_A = 20$) are probabilistically inevitable with consecutive resampling (Figure from [Busetto and Buhmann (2009)]).

Let t_i be the time point at which the new measurement $y(t_i)$ is available to the modeler and let $t_i^- \in \mathbb{T}$ indicate the previous time step. In the propagation phase, the posterior predicts the value at t_i on the basis of the data accumulated until t_i^- through Eq. (4.5) (or, more generally, through the forward equation of the system). At this point, the samples are grouped into K clusters. Their indexes are partitioned into the sets V_k for $k = 1, \ldots, K$. Clustering is performed with weighted K-means [Tseng (2007)]. At time t_i^-, every cluster has $|V_k| = p_k$ samples and

$$\sum_{k=1}^{K} p_k = n_p. \tag{4.19}$$

To maintain multiple modes, resampling is performed by keeping the cardinality of each cluster constant over time. The process is based on the estimation, for every cluster, of two multivariate normal components. The component $\mathcal{N}_k^-$ is obtained at time t_i^-, while $\mathcal{N}_k^+$ is obtained at time t_i. Independently, each component is employed as a proposal distribution for the importance resampling of the density from time t_i^-. The centroids $\boldsymbol{\mu}_k$ of the components define a Voronoi tessellation of partitions $\{(\mathcal{X} \times \Theta)_k\}_{k=1}^K$ of the parameter space Θ [Aurenhammer (1991); Busetto and Buhmann (2009)]. New samples are taken within each cluster, keeping p_k constant in each partition $(\mathcal{X} \times \Theta)_k$. Each new sample x_θ^j is sampled from $\mathcal{N}_k^-$ and accepted if $\mathbb{I}_{x_\theta^j \in (\mathcal{X}\times\Theta)_k} = 1$. When accepted, the weight is computed according to the ratio

$$w^{j+} \propto \frac{\mathcal{N}_k^-(x_\theta^{j-})}{\mathcal{N}_k^-(x_\theta^{j-})}, \tag{4.20}$$

where x^{j-} denotes a sample of the extended state at time t_i^-. The procedure guarantees a constant number of samples in every partition. The weight mass tends to concentrate in regions with significant contributions to the approximation. The procedure is outlined in Alg. 9.

How to measure the quality of the approximation? Effective Sample Size (ESS) can be employed to measure the quality of a sampled approximation [Doucet and Johansen (2011)].

Definition 19. *The ESS is given by*

$$N_{eff} \coloneqq \frac{n_p}{1 + \mathbb{V}[w_i^{*j}]}, \tag{4.21}$$

where

$$w_i^{*j} = \frac{p(x_{\theta i}^j | y_{1:i})}{q(x_{\theta i-1}^j, y_i)}. \tag{4.22}$$

Since the quantity cannot be evaluated exactly, one can evaluate the following approximation

$$\widehat{\mathrm{N}}_{\mathrm{eff}} \coloneqq \frac{1}{\sum_{j=1}^{n_p} (w_i^j)^2}. \tag{4.23}$$

Algorithm 9: Filtering with preventive clustering.

Data: Σ, $x_\theta(t_0)$, Φ, Θ, K.
Result: $\tilde{p}(x_\theta)$.

for $i = 1, \ldots, n$ **do**
  draw n_p samples from $\tilde{p}_t(x_\theta)$ with importance sampling;
  propagate the samples according to the forward equation of Σ;
  weighted K-means of the samples x_θ^j, $j = 1, \ldots, n$;
  for $k = 1, \ldots, K$ **do**
    estimate components $\mathcal{N}_k^-$ and $\mathcal{N}_k^+$;
    resample x_θ^j within each cluster keeping p_k fixed;
    recalculate weights w^{j+} from importance sampling as in Eq. (4.20);
  end
end

Cases in which $\widehat{\mathrm{N}}_{\mathrm{eff}} \approx \mathrm{N}_{\mathrm{eff}}$ is small are pathological and indicate degeneracy in the approximation. It has also been proved that, as $\mathbb{V}[w_i^{*j}]$ does not increase with time, $\mathrm{N}_{\mathrm{eff}}$ cannot increase without resampling [Doucet (1998)]. Small values of $\widehat{\mathrm{N}}_{\mathrm{eff}}$ indicate that, among all updated samples, very few exhibit non-negligible contributions to the approximation. In other words, $\widehat{\mathrm{N}}_{\mathrm{eff}}$ is a practical estimate of the effective number of contributing samples in the approximation. Bottlenecks in ESS result in serious instability of the modes, since they aggravate the tendency to drift towards the unimodal attractors [Busetto and Buhmann (2009)]. Standard techniques for resampling are able to regenerate the weights of the samples obtaining intervals exhibiting high ESS. However, they also create ESS bottlenecks. Resampling with preventive clustering mitigates the problem by drastically reducing frequency and intensity of the bottlenecks. Figure 4.3 illustrates preventive resampling with an proof-of-concept. In the figure, the Bayesian update of the weights of the samples from prior (upper

plot) to posterior (lower plot) is shown on the left. After the update, only a fraction of the samples maintain a non-negligible weight. In contrast, on the right, resampling is performed with preventive clustering. From the same prior (upper plot), resampling is performed before the update (plot in the middle). The resampled posterior (lower plot) exhibits a sample concentration in regions of higher density, thus retaining a large number of effective samples.

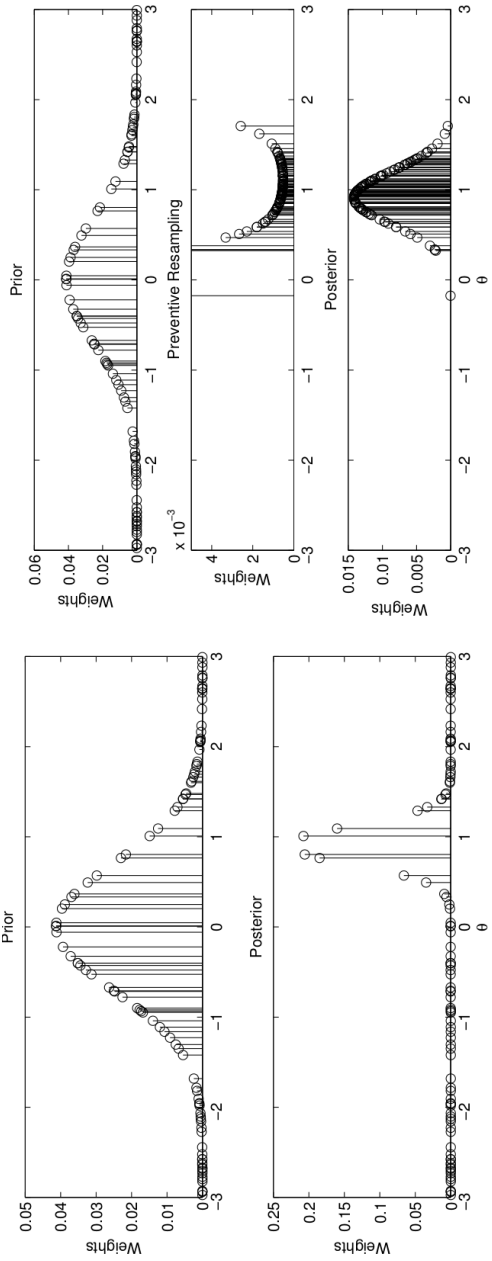


Figure 4.3: On the left, standard updating from prior (top) to posterior (bottom) concentrates the weights on few samples. In contrast, on the right, preventive cluster resampling regenerates the samples from the prior (top) by concentrating them in regions with high posterior density (bottom). The intermediate resampling step (on the middle of the plots on the right) maintains a large number of effective samples (Figure from [Busetto and Buhmann (2009)]).

4.1.2 Evaluation and Biological Application

The application of techniques from Bayesian inference is expanding rapidly in several domains. This section focuses on the motivating application, which comes from computational systems biology [Wilkinson (2007)]. The following considerations apply to general systems of chemical equations. In the considered domain, MCMC approaches have been remarkably successful in a variety of applications [Golightly and Wilkinson (2008)], but remain rather computationally intensive. The established approaches known as SMC suffer the limitations described before.

Let us consider a dynamical system modeling the behavior of a biochemical reaction network at thermodynamic equilibrium. Let n_c denote the number of distinguishable chemical species $\mathbf{C}_i$ in the system, for $i = 1, \ldots, n_c$. There are n_r chemical reactions $\mathbf{R}_i$, for $i = 1, \ldots, n_r$. Assuming thermal equilibrium and homogeneous spatial distribution, one can indicate reaction $\mathbf{R}_j$ as

$$R_j \colon \mathbf{N}_{1j}^{\text{in}}\mathbf{C}_1 + \ldots + \mathbf{N}_{n_cj}^{\text{in}}\mathbf{C}_{n_c} \longrightarrow \mathbf{N}_{1j}^{\text{out}}\mathbf{C}_1 + \ldots + \mathbf{N}_{n_cj}^{\text{out}}\mathbf{C}_{n_c} \quad (4.24)$$

where each $\mathbf{N}_{ij}^{\text{in}}$ is the element at the i-th column and j-th row of the input stoichiometric matrix $\mathbf{N}^{\text{in}} \in \mathbb{N}^{n\times r}$. Similarly, each $\mathbf{N}_{ij}^{\text{out}}$ is an element of the output stoichiometric matrix $\mathbf{N}^{\text{out}} \in \mathbb{N}^{n\times r}$.

Definition 20. *The stoichiometric matrix $\mathbf{N} \in \mathbb{N}^{n\times r}$ associated with a biochemical reaction network is given by*

$$\mathbf{N} \coloneqq \mathbf{N}^{out} - \mathbf{N}^{in}. \quad (4.25)$$

The stoichiometric matrix represents the net effects of all involved reactions in the considered physical system. Let $v_j(\mathbf{C}, \theta_j)$ denote the rate law associated with reaction R_j, for all j, which depends on the current state $\mathbf{C}$ and on parameter θ_j.

The chemical master equation quantifies the probability that, for all $t \in \mathbb{T}$, the biochemical system is in state $\mathbf{C}$. The state is discrete

and describes the time-varying number of molecules for each of the n_c species. The chemical master equation gives the transition probabilities for the dynamics of the biochemical system. Its solutions can be computed explicitly only in rare cases of practical interest [Wilkinson (2011)]. Dividing by the finite volume $\Delta\sigma$, the chemical species concentrations can be defined as

$$x_{ci} = \frac{\mathbf{C}_i}{\Delta\sigma}. \tag{4.26}$$

At time $t + \Delta t$, the master equation gives, up to $o(\Delta t)$

$$\begin{aligned} p_c(t + \Delta t) \simeq & \sum_{j=1}^{n_r} p(\mathbf{C} - \mathbf{N}_j; t) v_j(\mathbf{C} - \mathbf{N}_j, \theta_j) \Delta t \\ & + p(\mathbf{C}; t) \left(1 - \sum_{j=1}^{n_r} v_j(\mathbf{C}, \theta_{\mathbf{j}}) \Delta t \right), \end{aligned} \tag{4.27}$$

where $\mathbf{N}_j$ denotes the j-th column of the stoichiometric matrix. Taking the the infinitesimal limit for $\Delta t \to 0$, the chemical master equation is given by

$$\frac{\partial p(\mathbf{C}; t)}{\partial t} = \sum_{j=1}^{n_r} (p(\mathbf{C} - \mathbf{N}_j; t) v_j(\mathbf{C} - \mathbf{N}_j, \theta_j) - p(\mathbf{C}; t) v_j(\mathbf{C}, \theta_{\mathbf{j}})) \tag{4.28}$$

which is a continuous-time discrete-state Markov process. A drift-diffusion equation is obtained by truncating the Taylor expansion of Eq. (4.28) to the second order, giving the Langevin equation

$$\begin{aligned} \frac{\partial p(x_{ci}; t)}{\partial t} = & - \sum_{i=1}^{n_c} \frac{\partial}{\partial x_{ci}} \mu_i(x_c) p(x_{ci}; t) \\ & + \frac{1}{2} \sum_{i=1}^{n_c} \sum_{j=1}^{n_c} \frac{\partial^2}{\partial x_{ci} \partial x_{cj}} D_{cij}(x_c) p(x_{ci}; t), \end{aligned} \tag{4.29}$$

where the drift term is given by

$$\mu_i(x_c) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \mathbb{E}[x_{ci}(t + \Delta t) - x_{ci}(t) | x_c(t) = x_c], \tag{4.30}$$

and the diffusion tensor has elements

$$D_{cij}(x_c) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \mathbf{C}[\{\Delta x_{ci}(t)\}, \{\Delta x_{cj}(t) | x_c(t) = x_c\}], \tag{4.31}$$

with differences defined as $\Delta x_{ci}(t) = x_{ci}(t + \Delta t) - x_{ci}(t)$. Under weak conditions [Wilkinson (2011)], the dynamics follows the systems of SDEs

$$\mathrm{d}x_c(t) = \mu(x_c(t), \theta)\mathrm{d}t + D^{1/2}(x_c(t), \theta)\mathrm{d}\mathbf{W}_t, \tag{4.32}$$

with

$$\mu(x_c(t), \theta) = \mathbf{N}v(x_c, \theta) \tag{4.33}$$

and

$$D(x_c(t), \theta) = \mathbf{N}\mathrm{diag}[v(x_c, \theta)]\mathbf{N}^T. \tag{4.34}$$

Estimating the parameters of the chemical master equation involves challenging computations and, therefore, the Langevin equation is employed as an effective approximation [Golightly and Wilkinson (2008)]. The quality of the approximation increases as the Markov process exhibits smaller jumps and slower variations in the solution of $p(x_c; t)$ with respect to the state x_c.

Generalized state space estimation can be performed as described at the beginning of this section, that is by extending the state space $\mathcal{X}$ to incorporate the parameters in $\mathcal{X} \times \Theta$. To improve resampling, the modeler can subject the extended state space to an invertible transformation [Busetto and Buhmann (2009)]. The transformation makes all components of the state space time-varying, still maintaining x_θ the result of a bijective function of x and θ for any time point. The transformation could, for instance, consider dynamic fluxes as well as Hill variables as additional states of the biochemical reaction network [Bullinger et al. (2008)]. The most common functional forms for the reaction rate laws are mass action for signal transduction, Michaelis Menten for metabolic pathways and Hill for gene regulation [Szallasi et al. (2006); Wilkinson (2011); Farina et al. (2006)]. All these forms

and their products can be expressed as

$$v_j(x_c, \theta) = v_{j,\text{nom}} \prod_{i=1}^{n_c} \frac{c_i^{\nu_{j,i}}}{K_{j,i}^{\eta_{j,i}} + x_{ci}^{\eta_{j,i}}}, \tag{4.35}$$

where the parameters $v_{j,\text{nom}}, K_{j,i}, \nu_{j,i}$ and $\eta_{j,i}$ are elements of the parameter vector. This formulation is remarkably expressive and able to model realistic combinations of activation, inhibition, and cooperative behavior. By expressing the new $x_\theta = [x_c, v, m]$ as a function of fluxes and Hill variables, one can write the transformed dynamics as

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t m_{j,i}} &= f_m(x_\theta(t), \theta) \\ \frac{\mathrm{d}}{\mathrm{d}t v_j} &= f_v(x_\theta(t), \theta) \end{aligned} \tag{4.36}$$

with

$$\begin{aligned} f_m(x_\theta(t)) &= \eta_{j,i} x_{ci}^{(\eta_{j,i}-1)} \frac{\mathrm{d}x_{c_i}}{\mathrm{d}t} \\ f_v(x_\theta(t)) &= v_j \sum_{i=1}^{n_c} \left(v \nu_{j,i} \frac{1}{x_{ci}} \frac{\mathrm{d}x_{ci}}{\mathrm{d}t} - \frac{1}{m_{j,i}} \frac{\mathrm{d}m_{j,i}}{\mathrm{d}t} \right), \end{aligned} \tag{4.37}$$

where $m_{j,i} = K_{i,j}^{\eta_{i,j}} + c_j^{\eta_{i,j}}$ and v_j denote, respectively, the components of m and v [Bullinger et al. (2008)]. Given the initial condition, the system of SDEs can be written as

$$\mathrm{d} \begin{bmatrix} x_c \\ m \\ v \end{bmatrix} = \begin{bmatrix} \mathbf{N} v \\ f_m(x_c, m, v) \\ f_v(x_c, m, v) \end{bmatrix} \mathrm{d}t + \begin{bmatrix} \mathbf{I} \\ \mathbf{0} \end{bmatrix} D([x_c, m, v], \theta)^{1/2} \mathrm{d}\mathbf{W}_t. \tag{4.38}$$

The parameters, which are unknown to the modeler, appear as extended initial conditions.

The described approach is applied to a dynamical system which models the circadian clock [Busetto and Buhmann (2009)]. The system is a modified version of the Goodwin model for the molecular mechanisms of the circadian clock of Neurospora and Drosophila

[Goodwin (1965)]. The model has interesting features: Hill and mass action kinetics, as well as an inhibitory feedback loop. The system is defined as the system of SDEs with deterministic function

$$\begin{aligned}
\frac{\mathrm{d}x_{c1}(t)}{\mathrm{d}t} &= 3 - v_1 - v_2 - v_4 \\
\frac{\mathrm{d}x_{c2}(t)}{\mathrm{d}t} &= v_1 - v_3 \\
\frac{\mathrm{d}x_{c3}(t)}{\mathrm{d}t} &= v_2 - v_3 \\
\frac{\mathrm{d}x_{c4}(t)}{\mathrm{d}t} &= 6v_3.
\end{aligned} \tag{4.39}$$

In the equation, the rate laws are

$$\begin{aligned}
v_1 &= \theta_1 x_{c1} \\
v_2 &= \theta_2 x_{c1} \\
v_3 &= \theta_3 x_{c2} x_{c3} \\
v_4 &= \theta_4 \frac{x_{c4}^2}{x_{c4}^2 + \theta_5^2}.
\end{aligned} \tag{4.40}$$

The stochastic terms are defined by the elements of the covariance matrix of $\mathbf{W}$, which are $D_{ij} = \delta_{ij} 10^{-3}$ for $i, j \neq 1$ with $D_{11} = 10^{-2}$. Figure 4.4 illustrates with a diagram the reaction network of the double Goodwin model. The initial conditions are $x_c(t_0) = [3, 0.6, 0.4, 0.8]^T$ and the nominal values for the parameters are $\theta = [7, 1, 10, 3, 1]^T$.

The measurement noise is additive, white, and normal with variance $\sigma_N^2 = 0.6$ per component of the measurement space. Only x_{c1} and x_{c4} are directly measurable, whereas x_{c2} and x_{c3} are measured as a linear combination. For all $t_i \in \mathbb{T}^{\downarrow}$, the available measurements are obtained from

$$y(t_i) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x_{c1}(t_i) \\ x_{c2}(t_i) \\ x_{c3}(t_i) \\ x_{c4}(t_i) \end{bmatrix} + \begin{bmatrix} \nu_1(t_i) \\ \nu_2(t_i) \\ \nu_3(t_i) \end{bmatrix}, \tag{4.41}$$

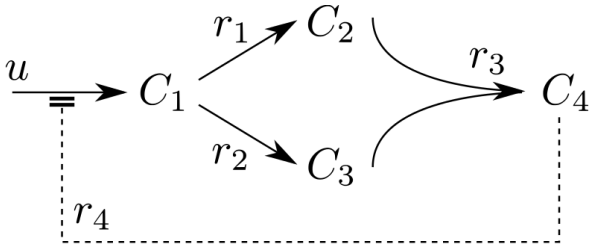


Figure 4.4: Diagram illustrating the reaction network of the double Goodwin model (Figure from [Busetto and Buhmann (2009)]).

with $\nu_j(t_i)$ IID according to $N = \mathcal{N}(\nu_j|0, \sigma_N^2)$ for $j = 1, 2, 3$ and $i = 1, \ldots, n$. An individual trajectory of the system, as well as a noisy realization of readout data are visualized in Fig. 4.6. Figure 4.5 compares the $\widehat{\mathrm{N}}_{\mathrm{eff}}$ obtained with different resampling strategies as a function of time. The introduced approach regenerates the ESS, thus mitigating the bottlenecks introduced by standard resampling.

On the left, Fig. 4.7 contains the plots of the marginalized posteriors for each parameter. The surface on the right represents the marginalized joint posterior for θ_1 and θ_2. In the calculation, $n_p = 2 \cdot 10^3$ samples. Figure 4.8 illustrates the effect of preventive cluster resampling in the parameter space. The approach avoids the situation in which particles with different coordinates in the extended space collapse into the same position in the parameter space. Preventive clustering maintains the equilibrium in multimodal cases, as shown in Fig. 4.9. As visible in Fig. 4.7, the posterior density remains bimodal in the parameter space. The plots of Fig. 4.9 show the marginalized distribution for $\theta_1 - \theta_2$ of the number of samples (with $n_p = 200$) staying the largest mode. With time, standard resampling diverges in the overwhelming majority of the cases by accumulating more and more samples with each iteration. Clustering helps by keeping the number of particles constant in both regions of the parameter

space. The final distributions are obtained by generating 500 IID datasets. Figure 4.10 shows the marginal posterior for $\theta_1 - \theta_2$ over time, demonstrating the persistence of the two modes with cluster resampling.

In summary, the introduced approach yields two beneficial properties: it mitigates sample degeneracy by preemptive resampling, and contrasts unimodal attractors through clustering. It is worth highlighting the importance of selecting the appropriate number of clusters K, a task which can be addressed by ASC validation (as shown in the previous chapter).

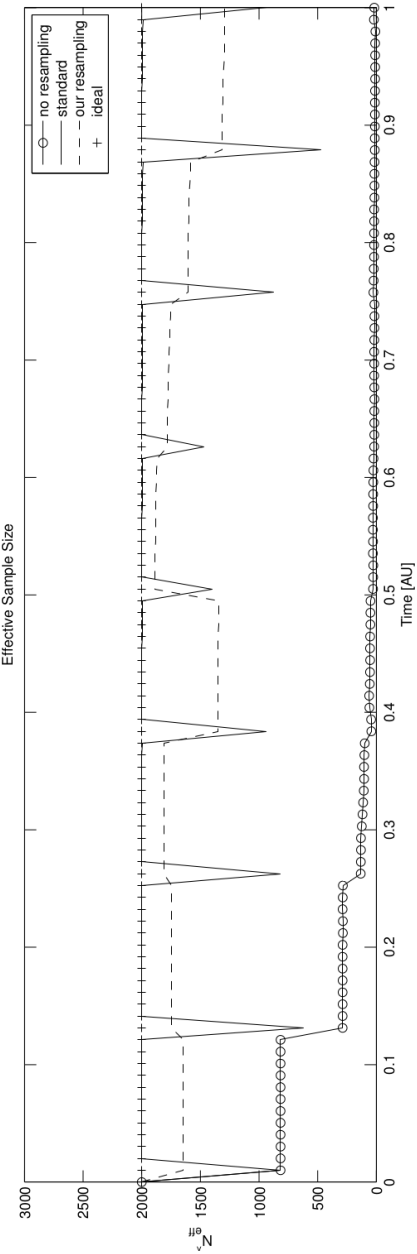


Figure 4.5: Estimated effective sample size subject no resampling, standard resampling and predictive cluster resampling as a function of time (Figure from [Busetto and Buhmann (2009)]).

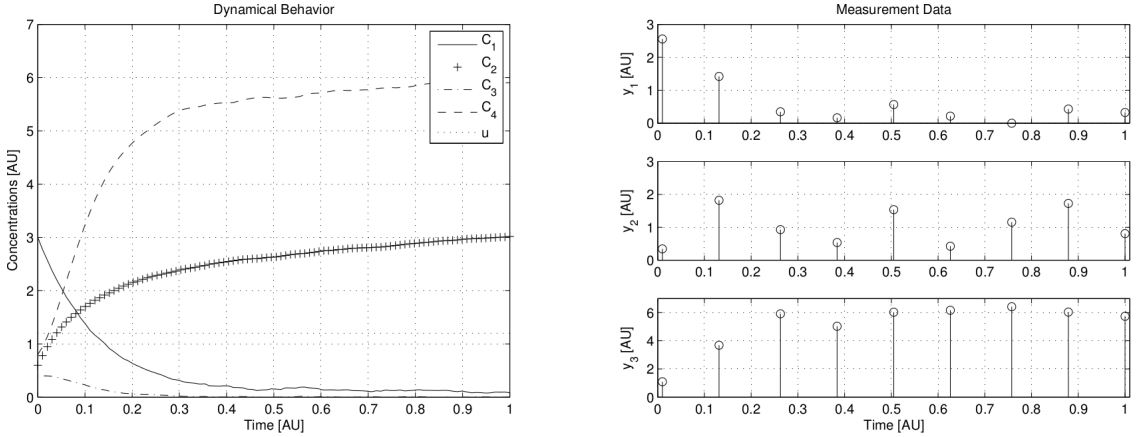


Figure 4.6: On the left, the plot shows a trajectory of the system from the initial condition. On the right, a realization of the data is visualized with uniform spacing between samples chosen for simplicity (Figure from [Busetto and Buhmann (2009)]).

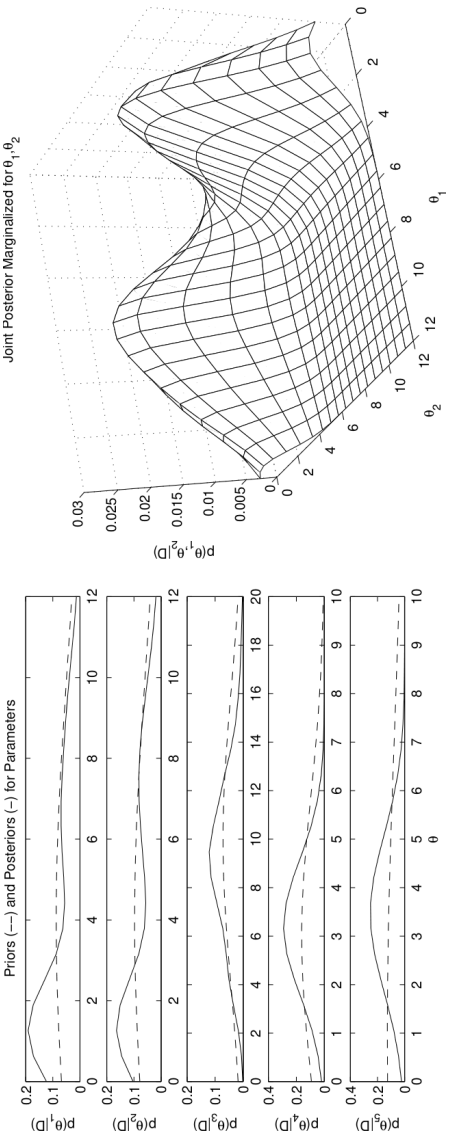


Figure 4.7: Marginalized priors and posteriors for each parameter (left). The marginalized joint posterior for θ_1 and θ_2 is shown on the right (Figure from [Busetto and Buhmann (2009)]).

Figure 4.8: Visualization of the weighted samples for θ_1 and θ_2 without (left) and with (right) preventive cluster resampling. The calculation on the right illustrates the mitigation of the problem of sample impoverishment in the parameter space (Figure from [Busetto and Buhmann (2009)]).

Figure 4.9: In the studied case, the posterior is bimodal in the parameter space. The plot shows how many of the $n_p = 200$ samples persist in the largest mode over 500 experiments with independently sampled datasets. On the left, standard resampling exhibits the tendency to accumulate all samples in one mode. In contrast, the right plot shows that the equilibrium is stable with cluster resampling (Figure from [Busetto and Buhmann (2009)]).

Figure 4.10: Visualization of the marginalized joint posterior for $\theta_1 - \theta_2$ as a function of time for a given dataset. In contrast to standard resampling (left), cluster resampling (right) maintains the the bimodal posterior (Figure from [Busetto and Buhmann (2009)]).

4.2 Evaluation of Optimized Solutions

> *"I believe that we do not know anything for certain, but everything probably."*
>
> — C. Huygens

Many estimation tasks are defined in terms of a cost function $R(c|D)$, which is minimized with respect to $c \in \mathcal{C}$ given the data. Minimization is performed over a search space of candidate solutions. The problem can also be equivalently formulated as an optimization problem in which the objective function $F_{\text{obj}}(c) \coloneqq -R(c|D)$ is maximized. Practical examples of this setting are those of regularized maximum likelihood and maximum a posteriori estimation. In practice, optimization is performed by the means of digital computation, effectively on a finite search space $\mathcal{C}$ (for instance, that of bounded float-point numbers). Despite being finite, the space is typically so large that it would be impractical to find the globally optimal solution c^* by inspection. In principle, one could take advantage of the regularities of the objective function to make the problem tractable (for instance, by excluding regions which are a priori known to yield suboptimal scores). There as cases, such as in submodular optimization [Nemhauser et al. (1978)], in which certain algorithms produce optimized solutions whose scores are guaranteed to be near-optimal. In practice, however, in most cases no regularities are known about the objective. Because of that, the modeler has access to the best solution $\bar{c}$ obtained from heuristic algorithms. In this general setting, one thing can be said for sure: the optimized solution $\bar{c} \in \mathcal{C}$ is the best solution among the evaluated candidates. To the modeler, such heuristic solution is often the best available approximation of the globally optimal solution c^*. Computational limitations, in fact, may limit significantly the number of possible evaluations of the objective function. In many cases of practical interest, no guarantees are available to the modeler for a general optimization task with objective $F_{\text{obj}}(\bar{c})$. However, even when nothing can be said regarding the value

of the optimized score, the relative position of the optimized solution can still be estimated with some additional empirical observations.

Definition 21. *For a finite search space $\mathcal{C}$, the relative position $pos(\bar{c})$ of the solution $\bar{c} \in \mathcal{C}$ is given by the ratio*

$$pos(\bar{c}) \coloneqq \frac{|\mathcal{C}_{\bar{c}}^{+}|}{|\mathcal{C}|}, \tag{4.42}$$

where

$$\mathcal{C}_{\bar{c}}^{+} \coloneqq \{c \in \mathcal{C} \colon F_{obj}(c) \geq F_{obj}(\bar{c})\} \tag{4.43}$$

is the set of solutions which are better or equal than $\bar{c}$ according to the total order induced by F_{obj}.

If a solution has $\text{pos}(\bar{c}) = \alpha$, then it can be said that it is in the top $\alpha \cdot 100\%$ for $F_{\text{obj}}(c)$.

Objective 4. *For the objective function $F_{obj}(c)$ over the search space $\mathcal{C}$, what is the relative position $pos(\bar{c})$ of a given optimized solution $\bar{c}$?*

4.2.1 Bound Derivation

The central idea is to employ a bound to estimate the relative position of the best available solution $\bar{c}$, which is given by the heuristic algorithm. The estimation can be performed through sampling, and the uncertainty can be bounded by a Chebyshev-type inequality. The results obtained here are an application of ideas employed as an argument in the proof of the symmetrization lemma from statistical learning theory [Bousquet et al. (2004)].

In the optimization setting, the heuristic algorithm terminates and yields the solution $\bar{c}$. At that point, let the available computational resources allow the additional evaluation of n_p samples. Let then c^j, for $j = 1, \ldots, n_p$, denote the IID samples drawn from the uniform distribution $\mathcal{U}_{nif}(\mathcal{C})$. The question is: what can be said about $\text{pos}(\bar{c})$ on the basis of these additional n_p evaluations? Let us introduce

the indicators $\mathbb{I}^j_{c^j \in \mathcal{C}^+_{\bar{c}}}$ for $j = 1, \ldots, n_p$. The indicators are random variables as well, IID drawn from the the Bernoulli distribution

$$\mathcal{B}_{ern}(\mathbb{I}^j_{c^j \in \mathcal{C}^+_{\bar{c}}} = 1 | \text{pos}(\bar{c})) = \text{pos}(\bar{c}). \tag{4.44}$$

One can, at this point, introduce the additional random variable

$$k_p = \sum_{j=1}^{n_p} \mathbb{I}^j_{c^j \in \mathcal{C}^+_{\bar{c}}} \sim \mathcal{B}_{in}(k_p | n_p, \text{pos}(\bar{c})), \tag{4.45}$$

where $\mathcal{B}_{in}(k_p | n_p, \text{pos}(\bar{c}))$ denotes the Binomial distribution

$$\mathcal{B}_{in}(k_p = k | n_p, \text{pos}(\bar{c})) = \binom{n}{k} \text{pos}(\bar{c})^k (1 - \text{pos}(\bar{c}))^{n_p - k}. \tag{4.46}$$

From the equations above, one has that the relative position $\text{pos}(\bar{c})$) can be estimated as the ratio

$$\widehat{\text{pos}}(\bar{c}) \coloneqq \frac{k_p}{n_p}. \tag{4.47}$$

Essentially, such ratio counts the proportion of uniformly sampled solutions which yield better scores than $F_{\text{obj}}(\bar{c})$. If $\text{pos}(\bar{c})$ is small (that is when its position is close to that of the point of global optimum), the modeler can expect $k_p = 0$ for small n_p with high probability.

Theorem 1. *For any $F_{obj}(c)$ over the search space $\mathcal{C}$, the absolute error between $pos(\bar{c})$ and the estimated $\widehat{pos}(\bar{c})$ satisfies the bound*

$$p(|pos(\bar{c}) - \widehat{pos}(\bar{c})| \geq \epsilon) \leq \frac{1}{4 n_p \epsilon^2}, \tag{4.48}$$

for any $\epsilon > 0$ and $n_p \in \mathbb{N}^{>0}$.

Proof. Chebyshev's inequality gives

$$p(|\mathbb{E}[Z] - Z| \geq \epsilon) \leq \frac{\mathbb{V}[Z]}{\epsilon^2}, \tag{4.49}$$

for the random variable Z. For $Z = k_p/n_p$, one has that, since $k_p \sim \mathcal{B}_{in}(k_p|n_p, \text{pos}(\bar{c}))$,

$$\begin{aligned} \mathbb{E}[k_p] &= n_p \text{pos}(\bar{c}) \\ \mathbb{V}[k_p] &= n_p \text{pos}(\bar{c})(1 - \text{pos}(\bar{c})) \end{aligned} \tag{4.50}$$

and, thus, that

$$\begin{aligned} \mathbb{E}\left[\frac{k_p}{n_p}\right] &= \text{pos}(\bar{c}) \\ \mathbb{V}\left[\frac{k_p}{n_p}\right] &= \frac{\text{pos}(\bar{c})(1 - \text{pos}(\bar{c}))}{n_p} \leq \frac{1}{4n_p}. \end{aligned} \tag{4.51}$$

Finally, the bound is obtained with

$$p\left(\left|\frac{k_p}{n_p} - \text{pos}(\bar{c})\right| \geq \epsilon\right) \leq \frac{\text{pos}(\bar{c})(1 - \text{pos}(\bar{c}))}{N\epsilon^2} \leq \frac{1}{4n_p\epsilon^2}. \tag{4.52}$$

□

When $k_p = 0$, as in many cases of practical interest, the bound simplifies to

$$p(|\text{pos}(\bar{c})| \geq \epsilon) \leq \frac{1}{4n_p\epsilon^2}, \tag{4.53}$$

and can be interpreted as follows: the probability that the optimized solution $\bar{c}$ is not in the $\epsilon \cdot 100\%$ is smaller than $1/(4n_p\epsilon^2)$.

The bound is also useful considering the dependence of $R(c|D)$ on the fluctuations of the data, so that the relative position reflects the variability due to the noise. In the context of statistical learning theory, the upper bound is calculated with the intention of estimating the deviation of empirical measures [Bousquet et al. (2004)]. In this setting, the bound is used to quantify the uncertainty regarding the relative position of a solution obtained through heuristic optimization. To employ such result for estimation, one has to note that, by itself, the bound does not address the problem of overfitting.

4.3 Modeling Human Learning Dynamics

"All models are wrong, but some are useful."

— G. E. P. Box

This section introduces two models of human learning aimed at treating dyslexia and dyscalculia, respectively[2].

Objective 5. *On the basis of interaction data D, select models M of human learning to design effective treatments for dyslexia and dyscalculia.*

For dyslexia, the model of engagement dynamics in spelling learning is obtained from experimental data consisting of recorded user-computer interactions. The model relates patterns of typed inputs to three aspects of human learning. It relates learning rates to engagement dynamics, predicting focused and receptive states, as well as forgetting. The final model has been selected from a set of candidate dynamic Bayesian networks (DBNs) which are estimated from data [Murphy (2002)]. The available dataset contains more than $150,000$ complete inputs recorded over several months through a training software for spelling.

The second model is introduced to enhance the numerical cognition of children with developmental dyscalculia through a computer-based training program. The model aims at predicting aspects of cognitive processes to control stimulation levels. The purpose is to optimize the learning process with targeted selection of informative exercises. The optimization process is based on a dynamic Bayesian network model which represents domain knowledge and predicts the level of accumulated knowledge of the subject over time. Estimation is employed to perform action selection on the basis of the state of knowledge of the subject.

[2]Parts of this section appear in [Baschera et al. (2011); Käser-Jacober et al. (2012)].

4.3.1 Modeling for Dyslexia Treatment

Sensor data [Cooper et al. (2010); Heray and Frasson (2009)] and input data [Baker et al. (2005); Johns and Woolf (2006); Arroyo and Woolf (2006)] are useful sources for modeling learning and affective dynamics. These sources differ in quality and quantity: sensor data are more direct and comprehensive, whereas input data are more general. The extraction of input logs, in fact, is not limited to laboratory experimentation. Predictive modeling and control are possible through software tutoring systems, since large and well-organized sample sets are available to the modeler. Inputs can be automatically recorded onto log files to capture informative features such as time lags and hints per problem, as well as interval lengths between consecutive solution attempts [Arroyo and Woolf (2006)].

Experimental Setting

The following results are obtained with data coming from Dybuster, a multimodal spelling software designed for dyslexic children [Gross and Vögeli (2007)]. During training, the tutoring system selects and generates sequences of test words which are prompted orally. Answers are typed in by the user, with acoustic signals notifying errors when incorrect letters are typed. Interactions are time-stamped and stored in comprehensive log files. Data are available from large-scale studies spanning an interval of several months [Kast et al. (2007); Baschera et al. (2011)]. The participants are German-speaking subjects in the age group between 9 and 11 years, trained four times a week for three months. The length of each session ranges from 15 to 20 minutes, summing up to ≈ 950 minutes of interactive training per user. Complete training is available for 28 dyslexic and 26 control subjects, for a total of $159,699$ records of entered words, inputs, errors, and respective timestamps.

Feature Selection and Modeling

Table 4.1 shows the main features extracted from the log data. The selection includes input and error behavior, timing, and controller-induced setting variations. The features measure quantities which are consistent with those of previous studies [Baker et al. (2005); Johns and Woolf (2006); Arroyo and Woolf (2006)]. The engagement state of the subject is estimated on the basis of the available data without external direct assessment. Error repetitions are employed to evaluate learning rates as well as forgetting. The analysis is restricted to phoneme-grapheme matching errors. Such errors belong to a category representing missing knowledge in spelling [Baschera and Gross (2010)]. In total, $14,892$ error observations are available in D. The features are processed according to the assumption that sustainable progress is separable in terms of time scales from local effects due to transitions between emotional and motivational states. Long-term variations are extracted from the time series ϕ through user-dependent regression under the assumption of additive white normal noise. Initial feature processing include scaling, outlier detection, low-pass and variance filtering [Baschera et al. (2011)]. The relation between the processed features $\widehat{D}$ and the error repetition γ_r is estimated via LASSO logistic regression with 10-fold cross-validation [Bishop (2006)]. Table 4.3.1 reports the values of the estimated parameters θ and the significance for the selected features. The model based on the processed features exhibits a better BIC score of -6369 compared to that obtained with unprocessed regression -6742. The selected features highlight three main time-varying components which influence the state of knowledge of the user:

- *focus:* indicating whether the subject is focused or distracted;
- *receptivity:* indicating the level of receptiveness of the subject;
- *forgetting:* indicating whether the subject is forgetting.

Non-focused states exhibit larger rates of non-serious errors which are due to lapses of concentration. However, these errors are less likely

to be committed again at later repetitions (that is, they exhibit lower error repetition probability). In contrast, non-receptive states inhibit learning, thus causing higher error repetition probabilities. Forgetting of learned spelling, which is due to time decay and interference between error and repetition, increases the error repetition probability as well.

The components of the parameter vector θ of the logistic regression indicate the relation between the individual features and the error repetition probability. The parameter $\theta_{\mathbf{EF}} = 0.06$ demonstrates that there exists a dependency between higher than expected error frequency and lower error repetition probability. In fact, non-focused subjects tend to commit more errors, but the errors are typically non-serious. By contrast, when the answer is complete but contains errors (that is, when $\mathbf{FC} = 0$), the error repetition probability increases ($\theta_{\mathbf{FC}} = -0.49$). The fact indicates that the subjects which do not correct the errors before answering are less likely to learn the correct spelling.

In the estimation setting, the model class consists of dynamic Bayesian networks. Such models are dynamical systems defined in terms of a Bayesian network graph structure which does not change over time[3]. The main difference between HMMs and DBNs is that, whereas the former class represents the state $x(t) \in \mathcal{X}$ by a single random variable, the latter allows the state space to be represented in factored form [Murphy (2002)]. DBNs do, in this sense, generalize HMMs. In DBNs, the hidden state is represented in terms of a set of (discrete or continuous) n_h random variables. In contrast to HMMs, which are defined by their transition model, the n_h state components and the n_o measured emissions are modeled as sets of random variables subject to conditional distributions described by a two-slice Bayes net.

Definition 22. *In a DBN Σ_{DBN}, transitions and observations are*

[3]time-varying structures are also possible but are here seen as extensions of the conventional definition [Murphy (2002)]

described by the conditional distributions

$$p(Z(t_i)|Z(t_{i-1})) = \prod_{j=1}^{n_h+n_o} p(Z^j(t_i)|Pa(Z^j(t_i))), \tag{4.54}$$

where $Z(t_i)$ denotes the node Z at time t_i (hidden or observed) and $Pa(Z)$ are the parents of Z [Murphy (2002)].

Let x_{F} and x_{R} denote, respectively, the state components (nodes in a DBN) indicating focus and receptiveness. Three candidate models are tested as on the basis of data $\widehat{D}$:

- $(x_{\mathrm{F}} \perp x_{\mathrm{R}})$ x_{F} and x_{R} are mutually independent;
- $(x_{\mathrm{F}} \leftharpoondown x_{\mathrm{R}})$ x_{F} depends on x_{R};
- $(x_{\mathrm{F}} \rightharpoondown x_{\mathrm{R}})$ x_{R} depends on x_{F}.

The parameters of the dynamic Bayesian network are estimated with EM (Bayes Net Toolbox for MATLAB [Murphy (2001)]. On the basis of $\widehat{D}$, the respective BIC costs of the models are

$$\begin{aligned} \mathrm{BIC}[x_{\mathrm{F}} \perp x_{\mathrm{R}}|\widehat{D}] &= -724,111 \\ \mathrm{BIC}[x_{\mathrm{F}} \leftharpoondown x_{\mathrm{R}}|\widehat{D}] &= -718,654 \\ \mathrm{BIC}[x_{\mathrm{F}} \rightharpoondown x_{\mathrm{R}}|\widehat{D}] &= -718,577. \end{aligned} \tag{4.55}$$

BIC selects the model Σ_{DBN} in which the receptive component depends on the focus of the subject. Figure 4.11 plots the diagram of the selected model. The joint probability distribution for x_{F} and x_{R} is reported on a log-scale in Fig. 4.12 (on the left). It is worth noting that, when the subject is fully focused, complete non-receptiveness is negligible. In contrast, it is not uncommon for the subjects to be non-focused despite being in an overall receptive state. This effect might be due, for instance, to temporary distraction. The error repetition probability is plotted on the right of Fig. 4.12 conditioned on the two components x_{R} and x_{F}. The top surface is obtained during

forgetting, and shows a greater offset in states of high focus. The available evidence is consistent with the assumption that in states of low focus the subjects tend to commit predominantly non-serious errors, since the correct spelling is already known. In contrast to forgetting, which has low impact on the error repetition probability, non-receptive states are associated with higher rates of error repetition. The effect, though, is reduced for non-serious errors in states with reduced focus. Table 4.3.1 reports parameter values for the selected DBN [Baschera et al. (2011)]. A conclusive remark can be made regarding the age-dependence of the engagement states. Non-receptive and non-focused states are more frequently found in subjects below the median of approximately 10 years (with p-value < 0.001). For this group, the prevalence of non-receptive states is 24.2% and that of non-focused states is 32.5%. In contrast, for subjects above the median the respective prevalence is 20.0% and 27.0% [Baschera et al. (2011)].

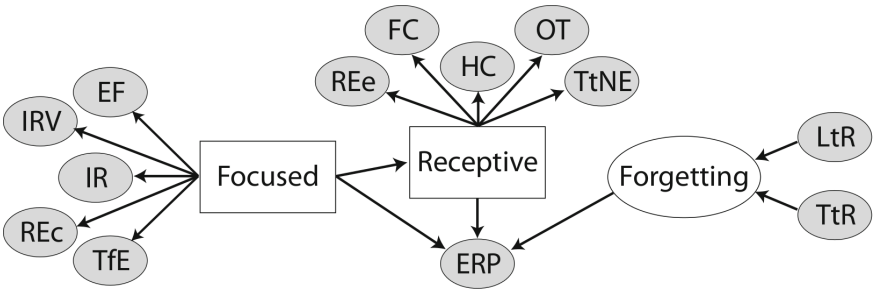


Figure 4.11: Diagram representing the dynamic Bayesian network selected by BIC. Dynamic components are indicated by rectangles. Shaded nodes are directly observed (Figure from [Baschera et al. (2011)]).

Feature	Description
Timing	
Input **R**ate	Keystrokes per second.
Input **R**ate **V**ariance	Seconds per keystroke variance.
Think **T**ime	Time between dictation of a word to first typed letter.
Time **f**or **E**rror	Time between last correct input to error.
Time **t**o **N**otice **E**rror	Time from error to first corrective action.
Off **T**ime	Longest time interval between consecutive letter inputs.
Inputs & Errors	
Help **C**alls	Number of help calls for dictation repetition.
Finished **C**orrectly	Indicator that all errors are corrected when enter key is pressed.
Same **P**osition **E**rror	Indicator for multiple errors for a single letter.
Repetition **E**rror	Input state of the previous input for the same test word (*Correct / Erroneous / Not Observed).*
Error **F**requency	Relative entropy between observed and expected error distributions estimated over five inputs with the student model Baschera and Gross (2010).
Controller Induced	
Time **t**o **R**epetition	Time from erroneous input to respective word repetition.
Letters **t**o **R**epetition	Number of entered letters from erroneous input to respective word repetition.

Table 4.1: Extracted features with abbreviations in bold (Table from [Baschera et al. (2011)]).

Feature	**Processing**	θ	**sig.**	$\mathbf{p_1}[\%]/\boldsymbol{\mu}$	
Focused State				focused	non-f.
EF	Exp	0.06	2e-4	0.16	-0.34
IR	Log - DevC - LearnC - Var	-0.12	4e-6	-0.41	0.87
IRV	Log - DevC - LearnC	-0.22	2e-11	-0.36	0.78
REc		-0.28	8e-8	45%	32%
TfE	Log - DevC - LearnC - LowP	-0.50	1e-9	-0.13	0.28
Receptive State				receptive	non-r.
FC		-0.49	1e-7	95%	88%
HC	Split(zero/non-zero)	0.29	2e-4	4%	28%
OT	Log - DevC - LearnC - LowP	0.27	1e-9	-0.35	1.20
REe	LowP	0.20	1e-9	0.07	-0.24
TtNE	Exp - DevC - LearnC	-0.18	1e-5	0.11	-0.36
Forgetting					
TtR	Exp	-0.29	2e-8		
LtR	Log	0.34	1e-9		

Table 4.2: Estimated parameters and significance for features selected by LASSO logistic regression. The influence of the engagement states on the features modeled in the dynamic Bayesian network is shown in the last two columns: probability p_1 of being TRUE for binary nodes and estimated mean $\boldsymbol{\mu}$ for Gaussian nodes (Table from [Baschera et al. (2011)]).

Figure 4.12: On the left, the plot shows, on a logarithmic scale, the joint probability distribution of focused (F) and receptive (R) states. On the right, the surfaces correspond to the error repetition probabilities conditioned on the engagement states for forgetting (top) and non-forgetting (bottom) (Figure from [Baschera et al. (2011)]).

4.3.2 Optimization for Dyscalculia Treatment

Computer-assisted learning is helpful to overcome learning disabilities [Käser-Jacober et al. (2012)]. Conventional learning therapy can be complemented by computer-based approaches which rely on formal modeling. Domain knowledge can, in fact, be directly incorporated into the DBN to enhance numerical cognition. Developmental dyscalculia is a learning disability with an estimated prevalence ranging from 3% to 6% [Shalev and von Aster (2008)]. The disability systematically affects the acquisition of arithmetic skills [Shalev and von Aster (2008)]. Due to its diffusion, large amounts of genetic, neurobiological, and epidemiological data are available to the modeler [Käser-Jacober et al. (2012)]. With developmental dyscalculia, the subjects are not only impaired by the disability itself, but tend to develop additional anxiety and aversion to mathematics [Rubinsten and Tannock (2010)]. Furthermore, the subjects face individual challenges which are mostly subject-dependent. Hence, training individualization has the potential of being particularly beneficial for treatment. The introduced model-based optimization system aims at improving success as well as at making the subject more motivated. The introduced results demonstrate substantial improvements verified by internal and external experimental evidence.

Learning Environment

Current models of neuropsychological development postulate the existence of cognitive representational modules. These modules are task-specific and localized in circumscribed regions of the brain. Recent studies identified specific functions related to adult cognitive number processing and calculation [Dehaene (2011)]. A reference framework is given by Dehaene's triple-code model, which is based on three distinct representational modules. The modules are associated to complementary aspects of number processing: understanding based on verbal, symbolic, analogue magnitudes operations [Dehaene (1992)]. Furthermore, compelling justifications exist for two additional assumptions.

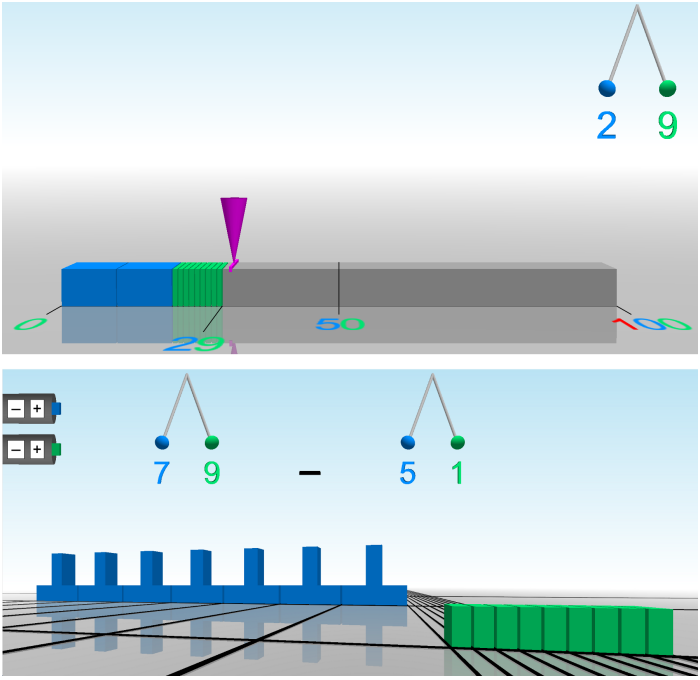


Figure 4.13: Screenshots of two games: "landing" on the left and "plus-minus" on the right (Figure from [Käser-Jacober et al. (2012)]).

The first is that modules develop hierarchically over time [von Aster and Shalev (2007)]. The second is that, as the level of mathematical understanding grows, the overlap of different number representations increases as well [Kucian and Kaufmann (2009)]. Learning rates are subject-dependent and influenced by several factors, such as the development of other cognitive skills and biographical aspects [von Aster and Shalev (2007)]. Because of intrinsic variability, model-based training optimization is necessary to better adapt to the individual subject through structured curricula with targeted stimuli.

The training system is based upon a hierarchical structure consisting of multiple games. Each game focuses on a specific skill taken from one of the two main training areas. The first area specializes on number representations and understanding. The second one primarily trains cognitive operations and numerical procedures. The difficulty of each game is reflected by the respective position in the hierarchical structure, which contains tasks of increasing complexity. To reinforce links between different number representations, numerical properties are associated to auditory and visual cues (color, form, and topology), while numerical values are associated to visual representations (blocks and positions in the place-value system) [Käser-Jacober et al. (2012)]. For illustration, the screenshots of two games are shown in Fig. 4.13. The game on the left is designed to train number representation and understanding. The one on the right aims at training and automating arithmetical operations.

Student Modeling and Action Selection

Training decisions are made online by the pedagogical module, which is a software system that adaptively selects games and configurations. Selections are made on the basis of the current knowledge state, which is estimated on the basis of the accumulated user inputs. The estimation of the time-varying learning state $x(t)$ is performed with respect to the student model, a fixed DBN which incorporates domain knowledge [Käser-Jacober et al. (2012); Dehaene (1992); von Aster and Shalev (2007); Geary et al. (1992); Ostad (1997, 1999)]. The network consists of a directed acyclic graphical model which satisfies Def. 22. The structure of the student DBN model, containing 100 skills, is visualized in Fig. 4.14. Each skill $x_{\mathrm{s}} \in \mathcal{X}_{\mathrm{s}}$ is a node in the network. A directed connection from skill $x_{\mathrm{s}A}$ to $x_{\mathrm{s}B}$ indicates that mastering the former is a prerequisite for learning the latter. Following the conditional network structure, one has that the probability of having $x_{\mathrm{s}A}$

in a certain learning state is

$$p(x_{\mathrm{s}}) = \prod_{x_{\mathrm{s}j} \in \mathcal{X}_{\mathrm{s}}} p(x_{\mathrm{s}j} | \mathrm{Pa}(x_{\mathrm{s}j})). \tag{4.56}$$

The state of each skill cannot be directly observed, and thus the system performs online inference on the basis of the answers to the posed task questions. Error patterns are identified from the measured actions of the subject, and matched with those available on the bug library. The generation of remediation skills is triggered by the detection of the respective errors. Denoting observations as y_{E}, the net posteriors $p(x_{\mathrm{s}A}|y_{\mathrm{E}k})$ are updated taking into account the observations coming the k-th task using the sum-product algorithm (through the libDAI library for discrete approximate inference in graphical models [Mooij (2010)]). The initial condition of all such probabilities is initialized to the uniform distribution, subscribing to the principle of indifference. The DBN has a memory of five time points.

Training interventions are rule-based and non-sequential. Learning paths are adapted to the estimated state of knowledge expressed by $p(x_{\mathrm{s}})$ at each time point. Three actions are available to the controller:

- Go Back: train a precursor skill;
- Stay: train the current skill (by generating a new task for the same skill);
- Go Forward: train a successor skill.

Decisions are based on the state of the skill net posteriors $p(x_{\mathrm{s}}|y_{\mathrm{E}})$ as described below. The posterior is compared to the values of an upper threshold p_u and a lower one p_l. The action Stay is selected when the probability for skill x_{s} is in the interval $[p_l, p_u]$. Above and below the interval, the training system selects Go Back and Go Forward,

respectively. The thresholds are time-varying and follow

$$\begin{aligned} p_l &= p_{l(\text{init})} \cdot l_c^{n_c} \\ p_u &= p_{u(\text{init})} \cdot u_c^{n_c}, \end{aligned} \tag{4.57}$$

where n_c denotes the number of accumulated samples for the tested skill, and parameters (l_c, u_c) and initial conditions $(p_{l(\text{init})}, p_{u(\text{init})})$ are determined heuristically [Käser-Jacober et al. (2012)]. The aim of these decision rules is to balance two learning aspects. On the one hand, the system has to guarantee a sufficient number of tests per skill. On the other hand, it avoids frustration in case of repeated failure. The decision trees for the action selection are shown in Fig. 4.15.

Experimental Results

The quality of the control system based on the student model is assessed on the basis of external testing. Input data are collected from two large-scale studies from Germany and Switzerland [Käser-Jacober et al. (2012)]. Both studies are conducted with cross-over design. The set of participating subjects $\mathbb{U}$ is partitioned into two groups: one starting the training immediately, the other waiting. Groups are mapped according to intelligence score, gender and age (second to fifth grade of elementary school). The participants are German-speakers and attend normal public schools in the respective countries. Overall, the subjects exhibit difficulties in learning mathematical concepts, as indicated by the following below-average performances in arithmetic tests [Haffner et al. (2005)]:

	T-score	*SD*
Addition	35.4	7.1
Subtraction	35.4	7.9

Available data consists of 33 complete logfiles for participants trained with five 20-minutes sessions a week for 6 weeks. On average, 29.84 sessions are completed per user (SD 2.87, minimum 24, maximum

36.96), solving a total of 1562 tasks (SD 281.53, minimum 1011, maximum 2179). The number of solved tasks per session is 52.37 (SD 7.9, minimum 37.8, maximum 68.1).

Definition 23. *A skill x_{sA} is a key skill for the subject $u \in \mathbb{U}$ if u goes back to a precursor skill x_{sB} of x_{sA} at least once before passing x_{sA}. The set of key skills for u is denoted as $\mathcal{K}_u$.*

Key skills are the hardest skills to pass. The set of key skills $\mathcal{K}_{\mathrm{u}}$ reflects the particular difficulties encountered by an individual subject. The tutoring system adjusts rapidly to the estimated state of knowledge of the subject. In fact, starting the training from the easiest skill, subjects encounter the first key skill after solving 144.3 tasks on average (SD 113.2, minimum 10, maximum 459), that is after an average of 1.95 sessions (SD 1.63, minimum 0.08, maximum 6.48).

For each subject $\mathrm{u} \in \mathbb{U}$, improvement is quantified as the learning rate over $\mathcal{K}_{\mathrm{u}}$ on the basis of all the available samples. The improvement is estimated with a non-linear mixed effect model employing one group per user and key skill [Pinheiro and Bates (1995)]. Denoting sample correctness as $\bar{y}_i \sim \mathcal{B}_{ern}(\theta_i)$, zero-mean normal noise terms as ϵ_i, and normalized sample indexes x_i between $[0, 1]$, learning is estimated as

$$\theta_i = \frac{1}{1 + e^{-(\mathbf{b_0} + \mathbf{b_1} x_i + \epsilon_i)}}. \tag{4.58}$$

Table 4.3 reports the estimation results, while Fig. 4.16 illustrates the improvement over time.

The analysis of the 533 Go Back cases shows that the option is beneficial in two ways: it reduces the error rate, and it increases the learning rate. Evaluation is performed for all cases in which a subject trains with a certain skill (samples x_b), goes back to one or several easier skills, and finally passes them to come back to the current skill (samples x_a). For each k-th case, the correct rates over time $c_{a,k}$ and $c_{b,k}$ are estimated separately for x_a and x_b. Logistic regression is performed with bootstrap aggregation (with resampling parameter B

	$\mathbf{b_0}$	$\mathbf{b_1}$
value (SD)	0.09 (0.06)	1.0 (0.06)
sig.	0.16	<1e-4
95% CI	[-0.073, 0.21]	[0.89, 1.11]

Table 4.3: Estimated coefficients of the non-linear mixed effects model with standard deviations, significance (sig.), and confidence intervals (Table from [Käser-Jacober et al. (2012)]).

	μ	**sig.**	**99% CI** of μ	**SD** σ	**99% CI of** of σ
$\bar{d}$	0.1494	<1e-6	[0.1204, 0.1784]	0.2593	[0.2403, 0.2814]
$\bar{r}$	0.3758	<1e-6	[0.3236, 0.4280]	0.4662	[0.4319, 0.5059]

Table 4.4: Improvement statistics related to Go Back actions. The mean is denoted as μ (Table from [Käser-Jacober et al. (2012)]).

= 200) [Breiman (1996)]. For the k-th case, the direct improvement is

$$d_k \coloneqq c_{b,k} - c_{a,k}, \tag{4.59}$$

that is the difference between the correct rates for $x_a = 0$ and $x_b = 1$. The average direct improvement is denoted as $\bar{d}$. Similarly, r_k indicates the learning rate improvement as the difference in learning rate over $c_{a,k}$ and $c_{b,k}$. The average learning rate improvement is denoted as $\bar{r}$. Figure 4.17 visualizes the histograms for $\bar{d}$ and $\bar{r}$ and their normal approximation. Both averages are positive and a two-sided t-test indicates that their difference from zero is statistically significant, as shown in Tab. 4.4.

Two external tests are employed to evaluate training effects:

- HRT is a paper-pencil test in which the subjects solve as many addition (subtraction) tasks as possible within a time frame of 2 minutes. Tasks are ordered according to increasing difficulty [Haffner et al. (2005)];
- AC is an arithmetic test in which the subjects solve addition (subtraction) tasks of increasing difficulty in a time frame of 10 minutes. The test exists both in paper-pencil as well as in computer-based versions.

The analysis compares the effects during training and waiting periods, respectively denoted as T_c and W_c. Results are available for 33

training subjects (26 females, 7 males) and 32 waiting subjects (23 females, 9 males). Whereas pre-tests indicate no significant difference between the groups, Tab. 4.5 shows that training induces a significant improvement in subtraction (for both HRT and AC) and no improvement is found after waiting.

There exists additional evidence of learning with respect to a specific indicator: the relative amount of training time spent by the subjects for subtraction tasks. Subtraction is considered the main indicator for numerical understanding [Dehaene (2011)]. During training, 62% (73% for key skills) of arithmetical tasks consist of subtractions. Consistent results are obtained for the improvement in the context of number line representation.

Accuracy analysis is performed using the non-linear mixed model based on a Poisson distribution of the deviance $y_i \sim \mathcal{P}_{ois}(\lambda_i)$ in which the parameter θ_i is given by

$$\lambda_i = e^{\mathbf{b_0}+\mathbf{b_1}x_i+\epsilon_i} \tag{4.60}$$

with zero-mean additive normal noise terms ϵ_i. In the estimation, fitting is performed using a single group per user. Numerical results, reported in Tab. 4.6 and plotted in Fig. 4.18, indicate that the subjects achieved greater accuracy in positioning numbers on a line [Käser-Jacober et al. (2012)].

Figure 4.14: Diagram representing the student model as a dynamic Bayesian network. On the right, the zoom highlights the subgraph containing addition skills (Figure from [Käser-Jacober et al. (2012)]).

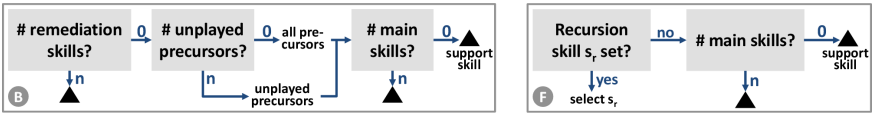


Figure 4.15: Simplified decision diagram for Go Back and Go Forward options (Figure from [Käser-Jacober et al. (2012)]).

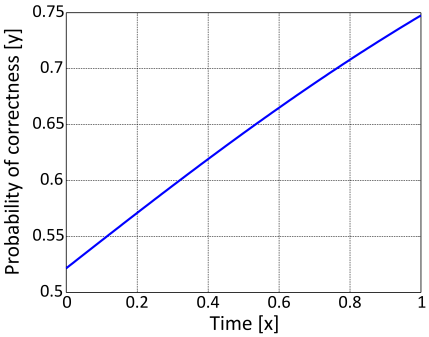


Figure 4.16: For the key skills, the percentage of correctly solved tasks exhibits an increase of 22.6% over the training interval (Figure from [Käser-Jacober et al. (2012)]).

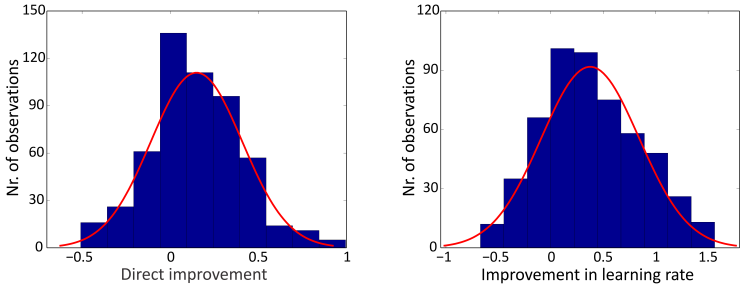


Figure 4.17: Histograms and normal approximation of the direct and learning rate improvements $\bar{d}$ and $\bar{r}$ (Figure from [Käser-Jacober et al. (2012)]).

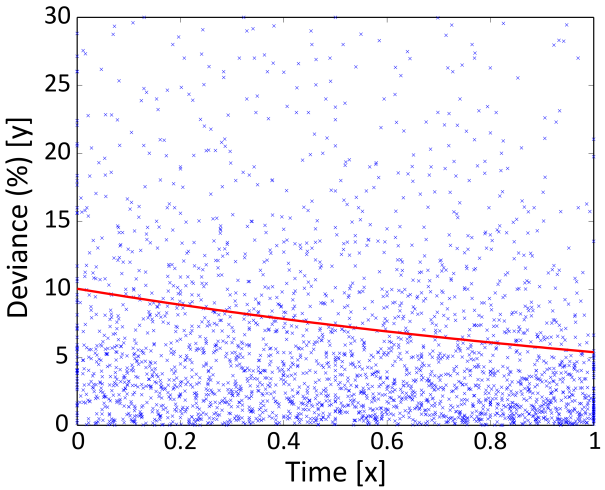


Figure 4.18: Landing accuracy in the range between 0 and 100 increases over time (Figure from [Käser-Jacober et al. (2012)]).

	Cond.	Pre-Score(SD)	Post-Score(SD)	sig.	Comparison
HRT	T_c	12.9 (5.38)	16.7 (5.3)	1.5e-8	2.6e-5 (2.9e-5)
	W_c	14.84 (6.47)	15.06 (5.87)	0.72	
AC	T_c	50.53 (27.25)	60.63 (26.3)	4.5e-4	1.9e-3 (2.0e-3)
	W_c	55.18 (25.24)	52.9 (27.74)	0.42	

Table 4.5: Comparison of test improvement during training and waiting. T-test results on the improvements assuming same and different variances are reported in the last column.

	$\mathbf{b_0}$	$\mathbf{b_1}$
Estimate(SD)	2.3 (0.07)	-0.63 (0.02)
sig.	<1e-4	<1e-4
95% CI	[2.17, 2.44]	[-0.67, -0.58]

Table 4.6: Estimated coefficients of the non-linear mixed effects model.

Chapter 5

Active Learning

"Doubt is the origin of wisdom."

— R. Descartes (transl.)

5.1 Design of Near-optimal Experiments for the Selection of Dynamical Systems

The previous sections focused on modeling given the data. This chapter reports results in the setting of active learning[1].

These contributions find their position in the context of optimal experimental design, that is the field in which the variables of an experiment are tuned to yield maximally informative results. Design optimization is motivated by the fact that, for many applications, the quality of the obtainable predictions is severely limited by data scarcity. In the hypothetico-deductive method of scientific inquiry, optimal experimental design finds its place in the iterative loop linking experimentation to modeling and viceversa. Figure 5.1 illustrates

[1] Parts of this section appear in [Hauser (2009); Krummenacher (2010); Busetto et al. (2009, 2013)].

experimental design at the interface between the process of inference and the execution of new experiments. Informally, the task of experimental design can be formulated as the selection of preparations, interventions, and observations which, on the basis of the available knowledge, yields the maximal amount of information. The general objective is the following.

Objective 6. *Given a set of candidate experiments* $\mathbf{E}$*, a class of dynamical systems functions* $\mathcal{F} = \{f_i\}_{i=1}^m$ *and prior information* $p(f)$*, select the most informative experiment* $\varepsilon^* \in \mathbf{E}$ *to maximize predictive capacity.*

This section concerns the design of experiments aimed at selecting dynamical systems from empirical data in a Bayesian setting. Precisely, the aim is to jointly select a set of informative time points and measurable quantities.

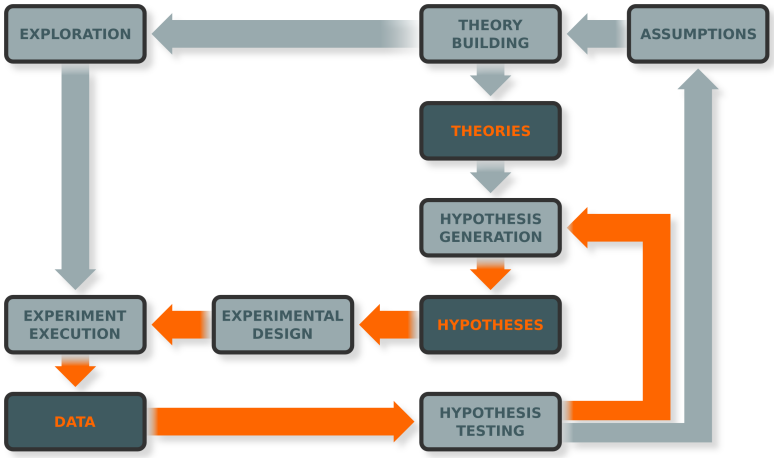


Figure 5.1: Schematic diagram illustrating the place of experimental design in the iterative loop between modeling and experimentation.

At present, there remain a number of open questions regarding experimental design aimed at selecting dynamical systems [Kreutz and

Timmer (2009); Faller et al. (2003); Myung and Pitt (2009)]. By contrast, extensive and conclusive results already exist for design aimed at parameter estimation [Bandara et al. (2009); Faller et al. (2003)]. The design of experiments for dynamical system selection remains a challenging field especially when the hypothesis class $\mathcal{F}$ contains strongly nonlinear systems. In systems biology and other application domains, in fact, $\mathcal{F}$ often contains nonlinear systems which are not well approximated through local linearization [Nelles (2001); Kitano (2002); Balsa-Canto et al. (2008)]. For such settings, the set of Bayesian techniques which are available to the designer has grown significantly in recent years, particularly in the domains of neuroimaging and biochemical modeling [Busetto et al. (2009); Kramer and Radde (2010); Daunizeau et al. (2011); Steinke et al. (2007)]. Methods rooted in classical statistics are employed as well, and among them ensemble noncentrality constitutes a reference approach [Skanda and Lebiedz (2010); Atkinson and Fedorov (1975); Ponce de Leon and Atkinson (1991)].

This section introduces a method which is based on previous results from the fields of active learning and optimization [Krause and Guestrin (2005, 2007); Krause et al. (2008); Nemhauser et al. (1978); Feige (1998)]. Building on such results, one can prove that, under conditions which are satisfied for systems of ODEs, the expected information gain is a submodular function [Krause and Guestrin (2005)]. As a consequence of that, it is possible to prove that a greedy method yields near-optimal solutions efficiently (that is, with a polynomial number of evaluations) [Krause and Guestrin (2005); Nemhauser et al. (1978)]. Furthermore, the introduced method dominates all other efficient techniques, unless P=NP. The result is proved by reducing the dynamical system scenario to a known setting which provides the best constant approximation factor to the value of the optimal solution [Krause and Guestrin (2005); Feige (1998)]. The design method exhibits general applicability and is motivated by biological questions. It addresses a challenging task in systems biology: the identification of the Target-of-Rapamycin (TOR) pathway, an open prob-

lem of high relevance for understanding metabolic operation [Kuepfer et al. (2007)]. Numerical evaluation demonstrates that the introduced method yields solutions with almost optimal informativeness when applied to glucose tolerance modeling. The design method has been recently applied to ongoing work in phosphoproteomics with the goal of modeling the key control mechanisms of nuclear phosphorylation.

5.1.1 Theoretical Results

Consider the model class of dynamical systems expressed in terms of systems of ODEs as in Def. 2, with known initial conditions and parameters. In the learning scenario, the task is that of estimating the transition function f^* of Σ^*. The prior $p(f)$ incorporates all available information for all candidate dynamical systems $f_i \in \mathcal{F}$. In the concrete application to biochemical modeling, for instance, f is the vector-valued function of a set of ODEs describing the dynamics of a chemical reaction network with fixed rates parameters and known kinetic constants. Let the pair of indexes (i, j) indicate to the noisy measurement

$$y_j(t_i) = x_j(t_i) + \nu_{ij}, \tag{5.1}$$

that is the observation at time point $t_i \in \mathcal{T}^{\downarrow}$ of the j-th component of the state vector $x(t_i) \in \mathcal{X}$. The noise terms ν_{ij} are independent and distributed according to the respective arbitrary distributions N_{ij}.

For simplicity, let us define the space of (i, j) index pairs as

$$Y_{ij} \coloneqq \{1, \ldots, n\} \times \{1, \ldots, n_x\}. \tag{5.2}$$

In this setting, each candidate experiment $\varepsilon \in \mathbf{E}$ consists of a set of index pairs $(i, j) \in Y_{ij}$. Each experiment ε is, hence, an element of the powerset $\mathbf{E} = \mathbb{P}(Y_{ij})$. The random variable D_ε denotes the dataset obtained from measuring $y_j(t_i)$ for all index pairs in ε. Formally, the dataset obtained through ε is

$$D_\varepsilon = \{(t_i, x_j(t_i)) \in \mathcal{T}^{\downarrow} \times \mathcal{X} \colon (i, j) \in \varepsilon\}. \tag{5.3}$$

At this point, the question is: how to measure the information gain? The example of Fig. 5.2 appeals to intuition: informative belief states concentrate the mass of the posterior on few candidate models. The update is a function of the data; as new independent samples are available, new evidence is incorporated by using previous posteriors as priors. Uninformative datasets induce negligible updates of the prior.

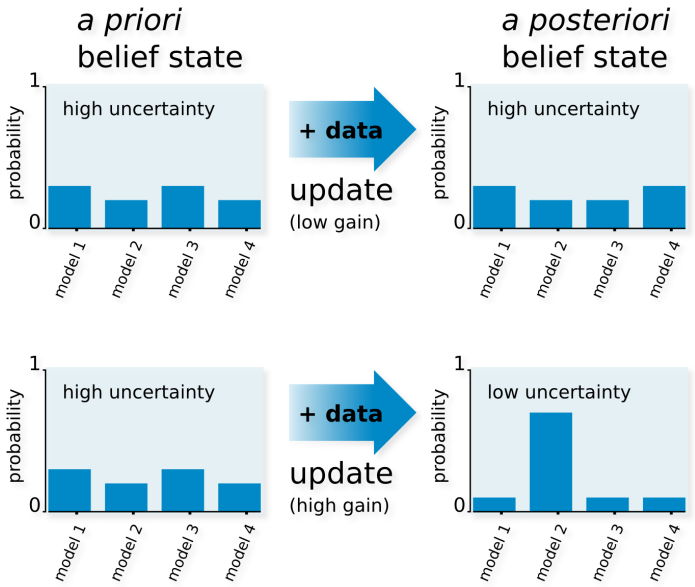


Figure 5.2: Updates from the same prior to the posteriors induced by two datasets. On the top, both initial and final belief states are uninformative: the update yields low information gain. In the bottom, instead, the illustration shows that the posterior concentrates the probability mass on a single model, thus yielding a high information gain).

The relative entropy between prior and posterior has interesting properties: it measures the information loss when the prior is em-

ployed as an approximation to the posterior [Baldi and Itti (2010)]. Such quantity is called information gain and is defined as follows

$$\mathrm{IG}(D_\varepsilon|p(f)) \coloneqq \mathrm{KL}[p(f|D_\varepsilon) \parallel p(f)], \tag{5.4}$$

where $p(f|D_\varepsilon)$ is, in the case of dynamical systems, found by solving a system of differential equations. In the context of communication, the information gain corresponds to the loss of information incurred when the dataset is ignored. Extraordinary evidence yields large information gain, strongly revising the belief state of the observer (as illustrated by Fig. 5.2). For any experiment ε, $\mathrm{IG}(D_\varepsilon|p(f))$ depends on the experimental data. The data instances, however, are a priori unknown to the modeler. Nonetheless, predictions can be made on the basis of available information. By taking advantage of all partial information, the task of designing experiments becomes a decision problem under uncertainty. The objective is to maximize the expected information gain, where the expectation is taken over all possible experimental outcomes induced by ε. Weighting the informativeness of individual outcomes by their respective probabilities, one has

$$\mathrm{I}(\varepsilon) \coloneqq \mathrm{I}(f, D_\varepsilon) = \mathbb{E}_{D_\varepsilon}\left[\mathrm{IG}(D_\varepsilon|p(f))\right], \tag{5.5}$$

which is the mutual information between D_ε and f. Such quantity measures the degree of statistical dependence between data and hypotheses [Busetto et al. (2009); Daunizeau et al. (2011); Chaloner and Verdinelli (1995)]. The optimization task becomes the following decision-theoretic problem: maximize the score of Eq. (5.5) with respect to ε, and subject to the constraint

$$|\varepsilon| \leq \kappa \leq n \cdot n_x, \tag{5.6}$$

for a given κ. The cardinality constraint reflects the fact that the epistemic agent can select only a reduced set of measurements. It is important to note that the following complexity results refer to the number of evaluations of the objective $\mathrm{I}(\varepsilon)$.

The solution of the optimization task can be approximated efficiently with formal worst-case performance guarantees [Nemhauser

et al. (1978); Feige (1998); Krause and Guestrin (2005)]. Starting from the individual sample yielding the largest mutual information, the designer proceeds by greedily selecting the pair of indexes (i, j) maximizing the contribution to the objective in combination with the current selection. Algorithm 10 formally describes the design process.

Algorithm 10: Greedy experimental design for the selection of dynamical systems (joint selection of time points and measurable quantities).

Data: set of hypotheses $\mathcal{H}$ with respective initial conditions and assigned parameters, prior $p(f)$, maximum number of measurements $\kappa \leq n \cdot n_x$, noise distributions N_{ij}.

Result: selection $\bar{\varepsilon}$.

initialization: $\varepsilon_0 \coloneqq \emptyset$;
forall the $f \in \mathcal{H}$ **do**
  forall the $(i, j) \in Y_{ij}$ **do**
    calculate $(t_i^{(f)}, x_j^{(f)}(t_i))$;
  end
end
for $k = 1$ *to* κ **do**
  $\varepsilon_k \coloneqq \varepsilon_{k-1} \cup \arg\max_{(i,j) \in Y_{ij} \setminus \varepsilon_{k-1}} \mathrm{I}(\varepsilon_{k-1} \cup \{(i, j)\})$;
end
$\bar{\varepsilon} \coloneqq \varepsilon_\kappa$;

By reduction to the active learning setting of graphical models [Krause and Guestrin (2005)], the following properties of the method can be proved.

Theorem 2. *The method described by Algorithm 10 finds a near-optimal joint subset $\bar{\varepsilon}$ of time points and measurable quantities for the selection of dynamical systems with a polynomial number of evaluations of $I(\cdot)$. Such approximate design satisfies the formal guarantee*

$$I(\bar{\varepsilon}) \geq \left(1 - \frac{1}{e}\right) I(\varepsilon^*), \tag{5.7}$$

where $\varepsilon^* \coloneqq \arg\max_{\varepsilon \in \mathbb{E}:\ |\varepsilon| \leq \kappa} I(\varepsilon)$. Moreover, the provided constant approximation factor is the best one for polynomial-time algorithms, unless P=NP.

Proof. Let $\mathcal{F} = \{f_k\}_{k=1}^m$ denote the hypothesis class of transition functions for dynamical systems defined on the state space $\mathcal{X}$. The trajectory of each dynamical system is determined, given $x_0 \in \mathcal{X}$ and $\theta \in \Theta$, by the respective integral solution of the system of ODEs defined by $f \in \mathcal{F}$. Let Φ_k indicate the noisy evaluation of the integral solution under the measurement model of Eq. (5.1), that is

$$\Phi_k = \{(t_i, y(t_i))\}_{i=1}^n. \tag{5.8}$$

Let a graphical model be constructed by having the set of nodes

$$\mathcal{V} = \{f\} \cup \{y_j(t_i)\}_{(i,j) \in Y_{ij}}, \tag{5.9}$$

where the $y_j(t_i)$ are those of Φ_k. The sets $\{f\}$ and $\{y_j(t_i)\}_{(i,j) \in Y_{ij}}$ are disjoint and, furthermore, one has that

$$\mathrm{I}(y_j(t_i), y_{j'}(t_{i'}) | f, x_0, \theta) = 0 \tag{5.10}$$

for all $(i,j) \neq (i',j') \in Y_{ij}$, because the measurements follow Eq. (5.1) given the deterministic solution of the IVP in Eq. (2.7) for fixed θ. The optimization of the mutual information with respect to ε then

becomes a special case of the known submodular setting for graphical models [Krause and Guestrin (2005)]. This is because the variables $\{y_j(t_i)\}_{(i,j)\in Y_{ij}}$ are independent given f. One has that $\mathrm{I}(\varepsilon)$ is a submodular set function, that is

$$\mathrm{I}(\varepsilon_1) + \mathrm{I}(\varepsilon_2) \geq \mathrm{I}(\varepsilon_1 \cup \varepsilon_2) + \mathrm{I}(\varepsilon_1 \cap \varepsilon_2), \tag{5.11}$$

and non-decreasing on $(i,j) \in Y_{ij}$, and $\mathrm{I}(\emptyset) = 0$. In this setting, the greedy method of inclusion of Alg. 10 with unit cost per observation selects κ elements yielding the constant approximation factor of $(1 - 1/e)$ [Krause and Guestrin (2005); Nemhauser et al. (1978)]. Such factor is also the best for polynomial algorithms, unless P=NP [Krause and Guestrin (2005); Feige (1998)]. □

5.1.2 Empirical and Applied Results

This subsection reports empirical evaluations and application to systems biology. First, the design method is numerically evaluated in a controlled set of frequency and time point selection experiments. Numerical evaluation with the Bergman glucose tolerance models shows that the obtained solutions yield tight approximations of the optimal informativeness. Secondly, the introduced method is compared with alternative approaches in a controlled setting with an external measure of success. Finally, design is applied to select dynamical systems of cell signaling pathways. Such results are biologically relevant to understand metabolic control operation. Submodular optimization is performed with SFO, the toolbox for submodular function optimization [Krause (2010)]. It is noteworthy that reliable parameter estimation constitutes an important requirement for successful design. This issue is important but goes beyond the scope of this study.

Frequency and Time Point Selection

Bergman glucose tolerance models are phenomenological dynamical systems which predict the effects of insulin on the degradation of blood glucose [Bergman et al. (1979)]. The study considers a set of

four models (I,IV,V,VI). These models are nonlinear, but relatively simple and well-understood. Reliable parameter estimates are based on comprehensive empirical observations [Hauser (2009); Bergman et al. (1979)].

The plot in Fig. 5.3 shows the expected information gain as a function of sampling frequency in the range of $[0, 1]$ samples/min. Noise terms are assumed to be additive and normally distributed. The plot shows that mutual information is a function subject to diminishing returns with respect to sample frequency. Precisely, uniform sampling at approximately 1/300 Hz already yields more than 90% of the experimentally available information. In the experimental setting, mutual information is calculated by marginalizing the uncertain parameters through unscented filtering [Julier and Uhlmann (2004)]. For solving the updated Bayesian drift equation with standard errors of 10^{-2} nats, the unscented approximation is significantly faster (on average between 40 to 400 times faster) than standard sequential MC sampling. For comparable numerical errors, in fact, SMC required at least 10^4 samples [Hauser (2009)].

With the same class of dynamical systems (models I,IV,V,VI), the greedy solution is compared with the optimal one (obtained by inspection) for the selection from a pool of $n = 20$ time points with $\kappa = 3, 4, 5$. Table 5.4 reports the solutions and shows that, not only the greedy solutions yield comparable informativeness, but they are also composed of very similar time points. Greedy solutions are, thus, effectively indistinguishable from the optimality for all practical purposes (below error tolerance) [Krummenacher (2010)].

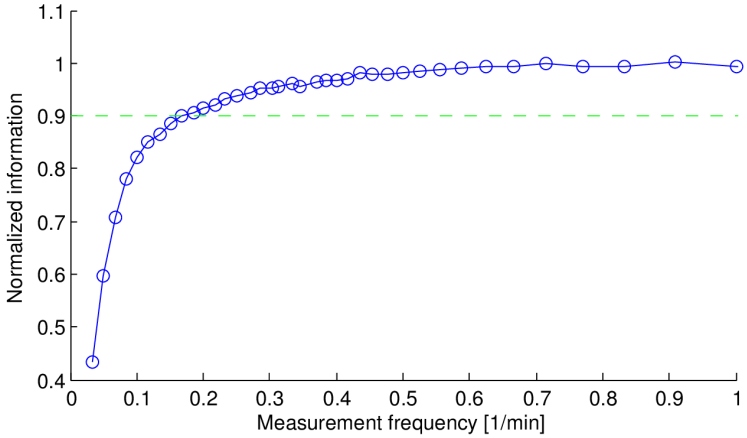


Figure 5.3: Already $\approx 90\%$ of the experimentally available information for modeling glucose tolerance dynamics is obtainable with a sampling frequency of 0.2 min^{-1}. The numerical approximation exhibits standard errors of 10^{-2} nats (not visible). In the plot, the mutual information is normalized to one (Figure from [Hauser (2009)]).

κ	**Greedy Solutions**	**Optimal Solutions**	$\mathbf{I}_{\mathbf{greedy}}$ [nats]	$\mathbf{I}_{\mathbf{optimal}}$ [nats]
3	{31, 34, 37}	{34, 37, 40 }	1.0004 ± 0.004	1.0009
4	{13, 31, 34, 37}	{10, 34, 37, 40}	1.0910 ± 0.004	1.0940
5	{10, 31, 34, 37, 40}	{10, 34, 37, 40, 43}	1.1564 ± 0.016	1.1585

Figure 5.4: Calculated mutual information for subsets of measurement time points with different cardinality constrains $\kappa = 2, 3, 4$. Measurement time points are selected from the set of indexes $\{1, 4, \ldots, 60\}$. Greedy solutions are almost optimal (Figure from [Krummenacher (2010)]).

5.1.3 Application and Comparison

For the benchmark task, the set of hypotheses $\mathcal{F}$ consists of candidate models for the Target-of-Rapamycin (TOR) pathway in yeast. TOR is a cell signaling structure which is highly conserved and whose mammalian homologue is implicated in cancer, cardiovascular diseases, autoimmunity, as well as metabolic disorders. Models are built by combining a core model with 18 hypothetical elementary extensions [Kuepfer et al. (2007)]. The core model represents the consensus based on all known molecular interactions from the inhibition of TOR kinases to the activation of PP2A. The extensions consists of additional reactions corresponding to putative mechanistic explanations of the biochemical system. In principle, not all the elementary extensions are mutually exclusive [Raman and Wagner (2011)]. By combining the core with compatible extensions, the study considers a representative class of 200 hypothetical dynamical systems. The systems exhibit heterogeneous complexity and the noise terms of Eq. (5.1) are additive normal [Krummenacher (2010)]. In the design, the $n_x = 24$ chemical species which are shared by all models are considered measurable candidates for the selection. Figure 5.5 shows the expected information gain by increasing the number of selected chemical species (from left to right). For completeness, offline and online submodular optimization bounds are reported as well [Krause and Guestrin (2007); Minoux (1978)]. Notably, the complex Tap42pP-PPA2, which is known for its central role [Kuepfer et al. (2007)], exhibits the highest individual information content. The selection of the readouts is performed by sampling a maximum of 50 regularly spaced time points in the dynamic range of 0 to 1.4 [AU] (arbitrary time units consistent with [Kuepfer et al. (2007)]). The number of candidate experiments amounts to $|\mathbf{E}| = 1200$.

The empirical success rate is the external measure employed to compare the introduced approach, which is Bayesian, with classical alternatives [Atkinson and Donev (1992)]. Comparison with other approaches is performed in two settings: realizable and non-realizable. In the former setting, the hypothesis class $\mathcal{F}$ contains the data gen-

erator f^*. By contrast, the non-realizable scenario considers the case in which $f^* \notin \mathcal{F}$. Overall, the success rate is estimated over 10^3 numerical simulations with synthetic data.

In each iteration, the method selects a set of measurable components of the TOR models. Each candidate model is employed to generate data with normal noise (variance corresponding to half of the concentration). In the realizable case, the success rate is calculated as the fraction of matches between the best a posteriori $f \in \mathcal{F}$ and f^*. In the non-realizable case, the success rate is defined as the fraction of matches between the best a posteriori model and the hypothesis in $\mathcal{F}$ which minimizes the information loss for all shared species (in terms of relative entropy). Figure 5.6 compares the success rate obtained with the introduced methods with those of alternative approaches (on the left, with the ensemble method). The analysis highlights the fact that significant computational load is the main practical disadvantage of ensemble methods for model selection. In fact, ensemble methods exhibits a bottleneck due to parameter fitting: all fitted parameter configurations are tested against what is assumed to be the correct model. The procedure is so resource-intensive that the hypothesis class is limited to four models with two unknown parameters and two unknown initial conditions.

On the right of Fig. 5.6, the analysis proceeds with the non-realizable setting, which captures the fact that hypothesis classes are approximations of reality. When $f^* \notin \mathcal{F}$, ensemble non-centrality is not directly applicable because it assumes that the true model is among the candidates (and performs selections with respect to it). The alternative cost which is here considered is the average residual sum of squares. In Fig. 5.6, the results on the right are obtained with all 200 models by measuring 50 equally spaced time points.

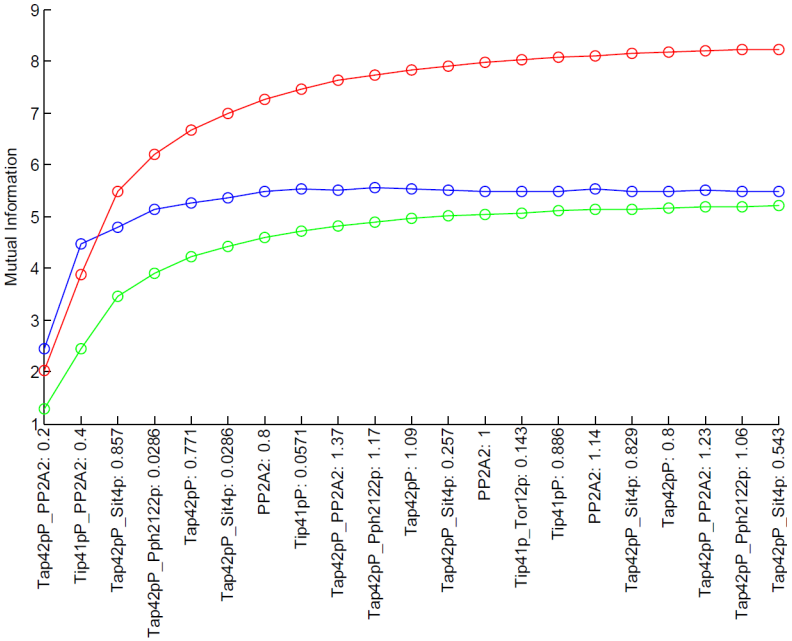


Figure 5.5: Mutual information for greedy selection of components of the measurement space is plotted in green, while online and offline optimization bounds appear in blue and red, respectively (Figure from [Krummenacher (2010)]).

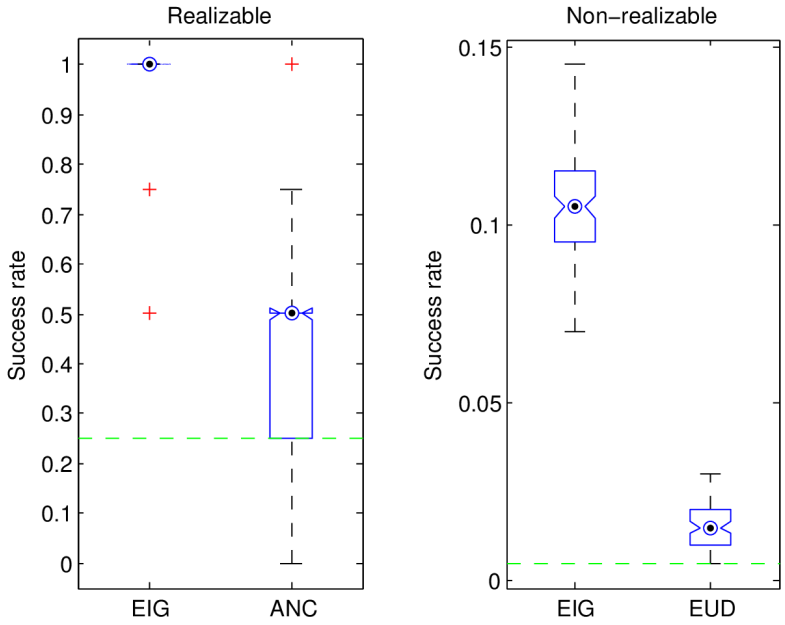


Figure 5.6: Success rate comparison for the selection of TOR pathway models in a controlled setting. The realizable case appears on the left, non-realizable one on right. The abbreviations EIG, ANC, and EUD denote, respectively, the designs obtained with expected information gain, ensemble non-centrality and sum of Euclidean distances. Rates are defined in the range $[0, 1]$, that is from complete lack of success to complete success. The plot on the right offers a relative interpretation of success, since the maximal rate achievable for the sample size is unknown (Figure from [Krummenacher (2010)]).

Chapter 6

Conclusion

> *"We can only see a short distance ahead, but we can see plenty there that needs to be done."*
>
> — A. Turing

The thesis introduced methods to model dynamical systems from data in the settings of unsupervised, supervised and active learning. The results are better understood within the hypothetico-deductive method of scientific inquiry: modeling opens the way to further experimentation, which in turn provides additional evidence for predictive modeling. Below are summarized the main contributions, together with an assessment of the respective limitations and an outline of future research directions.

Clustering of Time Series and Validation.
Results: Section 3.1 presented an effective method to perform model and order selection for relational clustering of time series. The method is based on the principle of Approximation Set Coding [Buhmann (2010)]. In the application to temporal gene expression profiles, the method showed consistency with the Bayesian Information Criterion and wide applicability. In the considered experimental setting, pairwise clustering provides approximately three-times more information

available for prediction than correlation clustering. In the reported results, pairwise clustering with correlation similarity extracts more information from the data; it generalizes better than correlation clustering the available trajectories.
Limitations and Outlook: The thesis does not provide a study of the theoretical relation between ASC and other principles for model selection. It is not known whether alternative approaches can be reduced to the information theoretic framework of ASC. This topic deserves attention and is the subject of active research.

Modeling of High-dimensional Sequences.
Results: Section 3.2 introduced an unsupervised approach to reconstruct the transition function of a dynamical system. Hidden Markov models derived from Gaussian mixture models yield accurate predictions of the trajectories extracted from pre-processed data. The approach is applied to model the dynamic behavior of the cell cycle from images sequences. Validation with biological data demonstrates that the unsupervised method is highly competitive to supervised approaches based on trained human labeling. Comprehensive testing provides evidence that the introduced approach significantly improves the objectivity of the results.
Limitations and Outlook: This study relies on pre-processing with PCA. In the general case, it would be useful to perform model and order selection directly with ASC.

Preventive Resampling for Generalized State Estimation.
Results: Section 4.1 proposed a resampling approach which mitigates the approximation divergence of conventional particle filtering. Numerical results show systematic increment of the effective number of samples.
Limitations and Outlook: In high-dimensional filtering, the task of selecting the appropriate number of clusters might become intractable. Future work will focus on bounding techniques to estimate the quality of the empirical approximations.

Quality Assessment of Heuristic Solutions for Global Optimization Problems.
Results: In Sec. 4.2, a Chebyshev-type bound estimates the relative position of an optimized solution on the basis of a limited amount of additional empirical evaluations. The bound provides formal probabilistic guarantees for the estimation error of the relative position of heuristic solutions with respect to the total order induced by the objective function.
Limitations and Outlook: The probabilistic bound is exact, but it remains to be seen if tighter bounds are generally available. Importantly, the bound is only useful to estimate the relative position of a solution, and says nothing regarding the closeness of the obtained objective with respect to its maximum value. For this purpose, current work focuses on approximate bounds taking into account further regularities of the objective function.

Modeling Human Learning Dynamics for the Treatment of Dyslexia and Dyscalculia.
Results: Section 4.3 introduced a predictive model for the dynamics of engagement in spelling learning. The model is a dynamic Bayesian network which relates typing errors to levels of receptiveness, focus, and forgetting of the subject. For dyscalculia treatment, a dynamic Bayesian network model is employed to improve learning of mathematical concepts through optimized cognitive stimulation. Extensive results with external validation provide evidence of effective learning improvement.
Limitations and Outlook: The presented work is limited to basic aspects of human learning. Whereas data scarcity constitutes a primary source of uncertainty, modeling could be improved by incorporating additional knowledge through transfer learning. Future work will attempt at generalizing the models to other aspects of human learning.

Near-optimal Design of Experiments for Modeling with Dynamical Systems.
Results: Section 5.1 introduced an efficient method to design experiments for selecting dynamical systems with formal guarantees of near-optimal informativeness. The method builds on recent results from submodular optimization in machine learning [Krause and Guestrin (2005)]. With a polynomial number of evaluations, the method yields a design solution which achieves the best constant approximation factor, unless P=NP. Numerical evaluation shows that quasi-optimal solutions are efficiently found to address modeling questions from systems biology.
Limitations and Outlook: The presented method has not been applied to the design of input interventions, which are useful tools for learning. It would be valuable to know which classes of interventions can be efficiently selected for informative design. Furthermore, the methods requires reliable estimates of the parameters of the dynamical systems in the hypothesis class. Finding such values is a separate task which is computationally challenging.

In conclusion, the results presented in this thesis highlight the fundamental analogies between learning and communication. Such perspective yields a rich theory and abundant applications. The introduced contributions facilitate the transfer of information theoretic concepts to guide and improve learning of dynamical systems. Quoting Francis Bacon:

> *"The eye of the understanding is like the eye of the sense: for as you may see great objects through small crannies or levels, so you may see great axioms of nature through small and contemptible instances."*

Nomenclature

"The world is a beautiful book,
but of little use to him who cannot read it."

— C. O. Goldoni

The thesis conforms to the following notation.

	Fundamentals:
$\emptyset$	empty set
$\mathbb{N}$	natural numbers with zero
$\mathbb{N}^{>0}$	natural numbers without zero
$\mathbb{Z}$	rational numbers
$\mathbb{R}$	real numbers
$\mathbb{R}^{\geq 0}$	non-negative real numbers
$\mathbb{R}^r$	r-dimensional Euclidean space
$\rho \wedge \varrho$	logical And
$\rho \vee \varrho$	logical Or
$\neg\rho$	logical negation
$a \in A$	element of a set
(a_1, a_2)	open interval
$[a_1, a_2]$	closed interval
$\{a\colon \ldots\}$	set builder
$A_1 \cup A_2$	set-theoretic union
$\delta(a)$	Kronecker/Dirac delta
$\lvert A\rvert$	cardinality

$A_1 \cap A_2$	set-theoretic intersection	
$A_1 \setminus A_2$	set-theoretic minus	
$A_1 \times A_2$	set-theoretic product	
$\mathbb{P}(A)$	power set	
$\rho_1 \Longleftrightarrow \rho_2$	double implication	
$\mathrm{d}g(a)$	infinitesimal increment	
$\mathrm{d}g(a)/\mathrm{d}a$	first derivative	
$\partial g(a_1, a_2)/\partial a_1$	partial derivative	
$\int_A g(a)\mathrm{d}a$	integral	
$\mathbf{W}_t$	Wiener process	
$\sigma(g(a))$	diffusion coefficient	
$\nabla \cdot [g(a)]$	divergence operator	
$\Delta[g(a)]$	diffusion operator	
$\mathbb{I}_A$	indicator function	Eq. (3.14)
ρ	logical proposition	Sec. 2.2
Ξ	set of propositions	Sec. 2.2
$\parallel v \parallel_r$	vector norm	Eq. (3.45)

	Computation:	
$\mathcal{Z}$	alphabet	Sec. 2.3
$z \in \mathcal{Z}$	symbol	Sec. 2.3
$\mathcal{Z}^*$	set of strings	Sec. 2.3
$\bar{z} \in \mathcal{Z}^*$	string	Sec. 2.2
pg	minimal program	Sec. 2.3
U[pg]	universal Turing Machine	Sec. 2.1
len(pg)	length of a program	Sec. 2.3
$\mathrm{K}(\bar{z})$	Kolmogorov (prefix-free) complexity	Eq. (2.25)

	Matrices and Graphs:	
$\mathbf{M}^T$	matrix transposition	
$\mathbf{M}^{-1}$	matrix inversion	
$\lvert\mathbf{M}\rvert$	matrix determinant	
$\mathrm{diag}[v]$	diagonal matrix	
$\mathbf{0}$	zero matrix	
$\mathbf{I}$	identity matrix	
$\mathrm{Pa}(v)$	node parents	
$\mathcal{G}$	graph	Sec. 2.3
$\mathcal{V}$	set of nodes	Sec. 2.3
$\mathcal{E}$	set of edges	Sec. 2.3
$\mathcal{E}_{ij}$	edge	Eq. (3.30)

	Information Theory:	
$S(p)$	uncertainty measure	Sec. 2.3
$\mathrm{h}(a)$	self-information	Sec. 2.2
$\mathrm{H}[p]$	entropy	Sec. 2.2
$\mathrm{KL}[p_1 \parallel p_2]$	relative entropy	Eq. (2.30)
$\mathrm{I}(Z_1, Z_2)$	mutual information	Eq. (2.31)
$\mathrm{H}(Z_1 \lvert Z_2)$	conditional entropy	Eq. (2.34)
$p_{\mathrm{ME}}(Z)$	maximum entropy distribution	Eq. (2.29)
$p_{\mathrm{Gibbs}}(Z)$	Gibbs distribution	Eq. (3.35)
$\mathrm{M}(\bar{z})$	universal prior	Sec. 2.2
$\mathrm{IG}(D_\varepsilon \lvert p(f))$	information gain	Eq. (5.4)
$\mathrm{I}(\varepsilon)$	design objective	Eq. (5.5)

	Relations and Functions:	
$v := g(a)$	definition	
$a_1 > (\geq) a_2$	grater (or equal) than	
$a_1 < (\leq) a_2$	smaller (or equal) than	
$a_1 \gg a_2$	much greater than	
$a_1 \ll a_2$	much smaller than	
$a_1 \approx a_2$	approximately	
$a_1 \simeq a_2$	similarity	
$g\colon A_1 \to A_2$	function	
$g_1 \equiv g_2$	equivalence	
$g_1 \propto g_2$	proportionality	
$O(g)$	asymptotic behavior	
a	factorial	
$\binom{a_1}{a_2}$	binomial coefficient	
$\log a$	logarithm	
$\exp a$	exponential function	
$\max_A g(a)$	maximum	
$\min_A g(a)$	minimum	
$\arg\min_A g(a)$	(set or element) argument of the maximum	
$\arg\max_A g(a)$	(set or element) argument of the minimum	
$\sum_A a$	sum over a set	
$\prod_A a$	product over a set	
$g_2 \circ g_1$	function composition	
$g_1 \doteq g_2$	asymptotic equivalence	Eq. (3.25)
$\mathrm{S}(n, K)$	2^{nd} kind Stirling numbers	Eq. (3.51)

	Statistical Mechanics:	
$Z \in \mathbb{R}^{\geq 0}$	partition function	Sec. 3.1
F	free energy	Eq. (3.39)
$w(c, X)$	Boltzmann weights	Eq. (3.36)

	Dynamical Systems:	
Σ	dynamical system	Sec. 2.1
Σ^*	physical system	Sec. 2.1
Σ_{HMM}	Hidden Markov model	Def. 16
Σ_{DBN}	dynamic Bayesian network	Def. 22
$\mathcal{X}$	state space	Sec. 2.1
$\mathcal{F}$	class of transition functions	Sec. 5.1
$n_x \in \mathbb{N}^{>0}$	dimension of the state space	Sec. 2.1
$\mathcal{T}$	time interval	Sec. 2.1
$t \in \mathcal{T}$	time point	Sec. 2.1
$t_0, t_s \in \mathbb{R}$	initial and final time points	Sec. 2.1
$\mathrm{F}(x,t) \in \mathcal{F}$	transition function	Sec. 2.1
$\mathrm{F}^* \in \mathcal{F}$	transition function of Σ^*	Sec. 2.1
$x \in \mathcal{X}$	system state	Sec. 2.1
$s \in \mathbb{N}^{>0}$	number of time points	Sec. 2.1
$\varphi \in \mathcal{X}^s$	system trajectory	Sec. 2.1
$x_0 \in \mathcal{X}$	initial condition	Sec. 2.1
Σ_{ODEs}	system of ODEs	Def. 2
Θ	parameter space	Sec. 2.1
$\theta \in \Theta$	parameter vector	Sec. 2.1
$f(x,t,\theta) \in \mathcal{F}$	deterministic update	Def. 2
$\mathcal{U})$	intervention space	Sec. 2.1
$u(t) \in \mathcal{U}$	instantaneous intervention	Sec. 2.1
$u \in \mathcal{U}^2$	series of interventions	Sec. 2.1
$\mathbf{A}$	stochastic transition matrix	Eq. (15)
$\mathcal{X}_\theta$	generalized state space	Sec. 4.1
$x_\theta(t) \in \mathcal{X}_\theta$	generalized state	Def. 17
$f_\theta \in \mathcal{F}$	extended transition function	Sec. 4.1
$Đ(x_\theta, t)$	diffusion tensor	Sec. 4.1
$n_h \in \mathbb{N}^{>0}$	number of DBN state components	Sec. 4.3
$n_o \in \mathbb{N}^{>0}$	number of DBN readout components	Sec. 4.3

	Probability and Statistics:	
$\mathbb{E}[Z]$	expectation	
$\mathbb{V}[Z]$	variance	
$\mathbb{C}[Z_1, Z_2]$	covariance	
$n \in \mathbb{N}^{>0}$	sample size	
$Z_1 \multimap Z_2$	statistical dependence	
$p_1 \perp p_2$	statistical independence	Sec. 2.3
Ω	sample space	Sec. 2.2
$B(\rho)$	plausibility measure	Sec. 2.2
$p(Z)$	probability (distribution/density)	Sec. 2.2
$\bar{p}$	probability vector	Sec. 2.2
$\mathcal{M}$	model class	Sec. 2.1
$M \in \mathcal{M}$	model	Sec. 2.1
$p(M\|D)$	posterior	Sec. 2.2
$p(D\|M)$	likelihood	Sec. 2.2
$p(M)$	prior	Sec. 2.2
$p(D)$	evidence	Sec. 2.2
$\mathcal{H}$	hypothesis class	Sec. 2.2
$H \in \mathcal{H}$	hypothesis	Sec. 2.2
$\Upsilon[p]$	statistics functional	Sec. 2.3
T	testable statistics	Sec. 2.3
$\varpi(M)$	number of model parameters	Def. 9
BIC$[M\|D]$	Bayesian Information Criterion	Def. 9
$\mathcal{Q}$	set of factorial distributions	Eq. (3.37)
$\mathbf{Q}(Z)$	factorial distribution	Eq. (3.37)
$\boldsymbol{\mu} \in \mathbb{R}^r$	mean parameter	Sec. 2.2
$\boldsymbol{\Sigma} \in \mathbb{R}^{r \times r}$	covariance parameter	Sec. 2.2
$\mathcal{N}_{\text{orm}}(z\|\boldsymbol{\mu}, \boldsymbol{\Sigma})$	multivariate normal distribution	Sec. 2.2

$\tilde{p}(\theta\|D)$	approximate posterior	Def. 18
$n_p \in \mathbb{N}^{>0}$	MC sample size	Sec. 4.1
$w^j \in [0,1]$	sample weight	Def. 18
$\mathrm{N}_{\mathrm{eff}} \in \mathbb{R}^{\geq 0}$	ESS	Def. 19
$\widehat{\mathrm{N}}_{\mathrm{eff}} \in \mathbb{R}^{\geq 0}$	approximate ESS	Eq. (4.23)
$\mathcal{U}_{nif}(\Omega)$	uniform distribution	Sec. 2.2
$\mathcal{B}_{ern}(z\|\theta)$	Bernoulli distribution	Sec. 4.2
$\mathcal{B}_{in}(z\|\theta_1,\theta_2)$	Binomial distribution	Sec. 4.2
$\mathcal{P}_{ois}(z\|\theta)$	Poisson distribution	Sec. 4.2
	Measurement Processes:	
$\mathbf{E}$	set of experiments	Def. 3
$\varepsilon \in \mathbf{E}$	experimental setting	Def. 3
$\bar{\varepsilon} \in \mathbf{E}$	near-optimal design	Sec. 5.1
$\varepsilon^* \in \mathbf{E}$	optimal design	Sec. 5.1
$\nu \in \mathcal{N}$	noise instance	Sec. 2.1
ν_i	noise variable	Sec. 2.1
$\mathcal{N}$	noise space	Sec. 2.1
N	noise distribution	Sec. 2.1
$\mathcal{Y}$	measurement space	Sec. 2.1
$y(t_i) \in \mathcal{Y}$	measured sample	Eq. (2.8)
$\widehat{y} \in \mathbb{R}^r$	PCA-reduced sample	Sec. 3.2
ν_i	noise variable	Sec. 2.1
h	measurement function	Eq. (2.8)
$\mathcal{T}^{\downarrow} \subseteq \mathcal{T}$	sampled time points	Sec. 2.1
$\phi \in \mathcal{Y}^n$	measured trajectory	Sec. 2.1
$\phi_i \in \mathcal{Y}^n$	measured component	Eq. (3.1)
$\widehat{\phi} \in \mathbb{R}^{r\times n}$	PCA-reduced trajectory	Sec. 3.2
$\Phi \in (\mathcal{T}^{\downarrow} \times \mathcal{Y})^{s^{\downarrow}}$	measured time series	Sec. 2.1
$\mathbf{Y} \in \mathcal{Y}^{l\times s}$	measurement tensor	Def. 11
$\tilde{x}(t_i) \in \mathcal{X}$	estimated states	Sec. 3.2
$\widehat{D}$	processed features	Sec. 4.3
$\kappa \in \mathbb{N}^{>0}$	maximum number of samples	Sec. 5.1

	Clustering:	
$\bar{n} \in \mathbb{N}^{>0}$	clustering sample size	
$K \in \mathbb{N}^{>0}$	number of clusters	Sec. 3.1
d_i	clustering sample	Sec. 3.1
$\mathcal{D}^*$	dataset space	Sec. 3.1
$D \in \mathcal{D}^*$	dataset	Sec. 3.1
$\mathcal{C}(D)$	clustering hypothesis class	Sec. 3.1
$c \in \mathcal{C}(D)$	clustering assignment	Def. 5
$R(c\|D) \in \mathcal{R}$	cost model	Def. 6
Λ	Label space	Def. 5
$\lambda \in \Lambda$	cluster label	Def. 5
V_k	set of elements in the k-th cluster	Sec. 3.1
$R_{\mathrm{km}}(c\|D)$	K-means clustering cost	Eq. (3.27)
$h_{i,c(i)}$	clustering potentials	Def. 3.1
P_k	cluster probability	Eq. (3.31)
X_{ij}	similarity between samples i and j	Sec. 2.3
X	similarity matrix	Sec. 2.3
$R_{\mathrm{pc}}(c\|D)$	pairwise clustering cost	Def. 10
$R_{\mathrm{cc}}(c\|D)$	correlation clustering cost	Eq. (3.34)
$\mathbf{H}$	co-clustering matrix	Sec. 2.3
H_{ij}	co-clustering indicator	Sec. 2.3
D_{diss}	dissimilarity matrix	Eq. (3.33)
$\mathbb{E}_{\mathbf{Q}_{i\to k}}[R_{\mathrm{cc}}]$	expectation over assignments	Sec. 3.1
$\eta \in \mathbb{R}^{\geq 0}$	noise level	Sec. 3.1
$\zeta(s,k) \in \mathbb{N}^{>0}$	counted assignments	Def. 12
$\mathrm{W}(c)$	intra-cluster cost	Eq. (3.56)
Y_{ij}	set of index pairs	Sec. (5.1)

	Optimization:	
Λ	Lagrange multiplier	
$c^{\perp} \in \mathcal{C}^{\perp}(D)$	empirical minimizer	Eq. (3.5)
$R^{\perp}(D) \in \mathbb{R}$	cost of the empirical minimizer	Eq. (3.6)
$\gamma \in \mathbb{R}^{\geq 0}$	approximation threshold	Def. 7
$F_{\text{obj}}(c)$	objective function	Sec. 4.1
$c^* \in \mathcal{C}$	globally optimal solution	Sec. 4.1
$\bar{c} \in \mathcal{C}$	optimized solution	Sec. 4.1
$\text{pos}(\bar{c}) \in [0,1]$	relative position	Def. 21
$\widehat{\text{pos}}(\bar{c}) \in [0,1]$	estimated $\text{pos}(\bar{c})$	Eq. (4.47)

	Approximation Set Coding:	
$K_{\max}$	initial number of clusters	Sec. 2.3
$\psi(D)$	mapping function	Sec. 3.1
$\mathcal{C}_{\gamma}(D) \subseteq \mathcal{C}(D)$	approximation set	Def. 7
$\Delta\mathcal{C}_{\gamma}$	intersection set	Eq. (3.12)
Σ	set of permutations	Alg. 1
$\sigma \in \Sigma$	permutation	Alg. 1
$\sigma_{\text{sel}} \in \Sigma$	sent permutation	Alg. 2
$\widehat{\sigma} \in \Sigma$	estimated permutation	Alg. 2
$\mathbb{I}_{\sigma}$	set of index permutations	Sec. 3.1
γ^*	optimal threshold	Eq. (3.20)
$\mathcal{I}_{\gamma}(\sigma_{\text{sel}}, \widehat{\sigma})$	ASC information	Eq. (3.19)
$\text{AC}_k[R\|X^1, X^2]$	approximation capacity	Def. 8
$\mathcal{R}$	cost model class	Sec. 3.1
$\beta^{-1} \in \mathcal{W}$	computational temperature	Sec. 3.1
$\mathcal{W}$	computational temperature range	Alg. 3
$1/\beta^* \in \mathcal{W}$	optimal temperature	Sec. 3.1
k^*	optimal model order	Alg. 4

	Biology and Human Learning:	
$l \in \mathbb{N}^{>0}$	population size	Sec. 3.2
D_λ	labeled population data	Eq. (3.57)
$p_{\mathrm{GMM}}(\phi)$	GMM of trajectories	Def. 14
π_k	mixing coefficients	Def. 14
$\mathbf{C}_i$	chemical species	Sec. 4.1
$\mathbf{R}_i$	chemical reaction	Sec. 4.1
$n_c \in \mathbb{N}^{>0}$	number of chemical species	Sec. 4.1
$n_r \in \mathbb{N}^{>0}$	number of chemical reactions	Sec. 4.1
$\mathbf{N}^{\mathrm{in}} \in \mathbb{N}^{n\times r}$	input stoichiometric matrix	Sec. 4.1
$\mathbf{N}^{\mathrm{out}} \in \mathbb{N}^{n\times r}$	output stoichiometric matrix	Sec. 4.1
$\mathbf{N} \in \mathbb{N}^{n\times r}$	stoichiometric matrix	Def. 20
$v_j(\mathbf{C}, \theta_j)$	reaction rate law	Sec. 4.1
γ_r	error repetition	Sec. 4.3
x_{F}	focused component	Sec. 4.3
x_{R}	receptive component	Sec. 4.3
$\mathcal{X}_{\mathrm{s}}$	skill set	Sec. 4.3
$x_{\mathrm{s}} \in \mathcal{X}_{\mathrm{s}}$	learnable skill	Sec. 4.3
$p_u \in [0,1]$	upper threshold	Sec. 4.3
$p_l \in [0,1]$	lower threshold	Sec. 4.3
$\mathbb{U}$	set of subjects	Sec. 4.3
$\mathrm{u} \in \mathbb{U}$	subject	Sec. 4.3
$\mathcal{K}_{\mathrm{u}}$	set of key skills	Def. 23
$\bar{y}_i$	sample correctness	Sec. 4.3
$\bar{d}$	average direct improvement	Eq. (4.59)
$\bar{r}$	average learning rate improvement	Sec. 4.3
T_{c}	training period	Sec. 4.3
W_{c}	training period	Sec. 4.3

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Arroyo, I. and Woolf, B. (2006). Inferring learning and attitudes from a Bayesian network of log file data. In *Proc. AIED*, pages 33–40.

Arulampalam, M. S., Maskell, S., Gordon, N., and Clapp, T. (2002). A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing*, 50:174–188.

Atkinson, A. C. and Donev, A. N. (1992). *Optimum Experimental Design*. Oxford Science Publications.

Atkinson, A. C. and Fedorov, V. V. (1975). Optimal design: Experiments for discriminating between several models. *Biometrika*, 62(2):289–303.

Aurenhammer, F. (1991). Voronoi diagrams - a survey of a fundamental geometric data structure. *ACM Computing Surveys*, 23(3):345–405.

Baker, R. S., Corbett, A. T., and Koedinger, K. R. (2005). Detecting student misuse of intelligent tutoring systems. In *Springer Lecture Notes in Computer Science, 3220 Proc. ITS*, pages 531–540.

Baldi, P. F. and Itti, L. (2010). Of bits and wows: A Bayesian theory of surprise with applications to attention. *Neural Networks*, 23(5):649–666.

Balsa-Canto, E., Alonso, A. A., and Banga, J. R. (2008). Computational procedures for optimal experimental design in biological systems. *IET Systems Biology*, 2(4):163–172.

Bandara, S., Schlöder, J. P., Eils, R., Bock, H. G., and Meyer, T. (2009). Optimal experimental design for parameter estimation of a cell signaling model. *PLoS Computational Biology*, 1:e1000558.

Banga, J. R. (2008). Optimization in computational systems biology. *BMC Systems Biology*, 2(47).

Banks, H. T. and Rehm, K. L. (2013). Applied mathematics letters. *Experimental Design for Distributed Parameter Vector Systems*, 26(1):10–14.

Banni, M., Negri, A., Mignone, F., Boussetta, H., Viarengo, A., and Dondero, F. (2011). Gene expression rhythms in the mussel Mytilus galloprovincialis (lam.) across an annual cycle. *PLoS ONE*, 6(5):e18904.

Bansal, N., Blum, A., and Chawla, S. (2004). Correlation clustering. *Machine Learning*, 56:89–113.

Baschera, G.-M., Busetto, A. G., Klingler, S., Buhmann, J. M., and Gross, M. (2011). Modeling engagement dynamics in spelling learning. In *Springer Lecture Notes in Computer Science, 6738 Proc. AIED*, pages 31–38.

Baschera, G.-M. and Gross, M. (2010). Poisson-based inference for perturbation models in adaptive spelling training. *International Journal of Artificial Intelligence in Education*, 20(4):333–360.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171.

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

Bergman, R. N., Ider, Y. Z., Browden, C. R., and Cobelli, C. (1979). Quantitative estimation of insulin sensitivity. *American Journal of Physiology*, 236(6):G667–G677.

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

Boland, M. V. and Murphy, R. F. (2001). A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells. *Bioinformatics*, 17(12):1213–1223.

Bousquet, O., Boucheron, S., and Lugosi, G. (2004). Introduction to statistical learning theory. In Bousquet, O., von Luxburg, U., and Rätsch, G., editors, *Advanced Lectures on Machine Learning*, volume 3176 of *Lecture Notes in Computer Science*, pages 169–207. Springer.

Bratman, M. (1987). *Intention, Plans, and Practical Reason*. Harvard University Press.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.

Brock, S. (2003). *Niels Bohr's Philosophy of Quantum Physics in the Light of the Helmholtzian Tradition of Theoretical Physics*. Logos Verlag.

Buhmann, J. M. (2010). Information theoretic model validation for clustering. In *Proceedings of the 2010 IEEE International Symposium on Information Theory*, pages 1398–1402.

Bullinger, E., Fey, D., Farina, M., and Findeisen, R. (2008). Identifikation Biochemischer Reaktionsnetzwerke: ein Beobachterbasierter Ansatz. *Automatisierungstechnik*, 56(5):269–279.

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Inference: A Practical Information-Theoretic Approach, 2nd ed.* Springer, New York.

Busetto, A., Sunnåker, M., and Buhmann, J. M. (2013). *System Theoretic and Computational Perspectives in Systems and Synthetic Biology*, chapter Computational Design of Informative Experiments in Systems Biology. (in press).

Busetto, A. G. and Buhmann, J. M. (2009). Stable Bayesian parameter estimation for biological dynamical systems. In *Proceedings of the 12th IEEE International Conference on Computational Science and Engineering*, pages 148–157. IEEE Computer Society Press.

Busetto, A. G., Ong, C. S., and Buhmann, J. M. (2009). Optimized expected information gain for nonlinear dynamical systems. In *Proceedings of the 26th International Conference on Machine Learning*, pages 97–104.

Carnap, R. (1945). The two concepts of probability: The problem of probability. *Philosophy and Phenomenological Research*, 5(4):513–532.

Chaitin, G. (1966). On the length of programs for computing finite binary sequences. *Journal of the ACM*, 13(4):547–569.

Chaloner, K. and Verdinelli, I. (1995). Bayesian experimental design: a review. *Statistical Science*, 10:273–304.

Chehreghani, M. H., Busetto, A. G., and Buhmann, J. M. (2012). Information theoretic model validation for spectral clustering. In *Journal of Machine Learning Research, Proceedings of AISTATS 2012*, pages 495–503.

Church, A. (1932). A set of postulates for the foundation of logic. *Annals of Mathematics*, 33(2):346–366.

Collinet, C., Stöter, M., Bradshaw, C. R., Samusik, N., Rink, J. C., Kenski, D., Habermann, B., Buchholz, F., Henschel, R., Mueller, M. S., Nagel, W. E., Fava, E., Kalaidzidis, Y., and Zerial, M. (2010). Systems survey of endocytosis by multiparametric image analysis. *Nature*, 464:243–249.

Conrad, C. and W., G. D. (2010). Automated microscopy for high-content RNAi screening. *Journal of Cell Biology*, 188(4):453–461.

Conrad, C., Wünsche, A., Tan, T. H., Bulkescher, J., Sieckmann, F., Verissimo, F., Edelstein, A., Walter, T., Liebel, U., Pepperkok, R., and Ellenberg, J. (2011). Micropilot: Automation of fluorescence microscopy-based imaging for systems biology. *Nature Methods*, 8:246–249.

Cooper, D. G., Muldner, K., Arroyo, I., Woolf, B. P., and Burleson, W. (2010). Ranking feature sets for emotion models used in classroom based intelligent tutoring systems. In *Springer Lecture Notes in Computer Science, 6075 Proc. UMAP*, pages 135–146.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. Wiley-Interscience.

Cox, R. T. (1946). Probability, physics, and reasonable expectation. *American Journal of Physics*, 14:1–13.

Cox, R. T. (1961). *The Algebra of Probable Inference*. Johns Hopkins University Press.

Dancy, J. (1985). *Introduction to Contemporary Epistemology*. Blackwell.

Daunizeau, J., Preuschoff, K., Friston, K., and Stephan, K. (2011). Optimizing experimental design for comparing models of brain function. *PLoS Computational Biology*, 11(7):e1002280.

de Finetti, B. (1970). *Theory of Probability.* Wiley and Sons.

Dehaene, S. (1992). Varieties of numerical abilities. *Cognition*, 44:1–42.

Dehaene, S. (2011). *The Number Sense: How the Mind Creates Mathematics.* Oxford University Press.

Dempster, A. P. (1968). A generalization of Bayesian inference. *Journal of the Royal Statistical Society, B*, 30:205–247.

Doucet, A. (1998). On sequential Monte Carlo methods for Bayesian filtering. Technical report, Department of Engineering, University of Cambridge, Cambridge.

Doucet, A. and Johansen, A. M. (2011). A tutorial on particle filtering and smoothing: Fifteen years later. In *Oxford Handbook of Nonlinear Filtering.*

Doucet, A. and Tadič, V. (2003). Parameter estimation in general state-space models using particle methods. *Annals of the Institute of Statistical Mathematics*, 55(2):409–422.

Dubois, D. and Prade, H. (1988). *Possibility Theory, An Approach to Computerized Processing of Uncertainty.* Plenum.

Dudoit, S. and Fridlyand, J. (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology*, 3(7).

Engstrom, D. R., London, P., and Hart, J. T. (1970). Hypnotic susceptibility increased by eeg alpha training. *Nature*, 227:1261–1262.

Faller, D., Klingmüller, U., and Timmer, J. (2003). Simulation methods for optimal experimental design in systems biology. *Simulation*, 79(12):717–725.

Farina, M., Findeisen, R., Bullinger, E., Bittanti, S., Allgower, F., and Wellstead, P. (2006). Results towards identifiability properties of biochemical reaction networks. In *Proceedings of the Conference on Decision and Control*, pages 2104–2109.

Feige, U. (1998). A threshold of Ln n for approximating set cover. *Journal of the ACM*, 45(4):634–652.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society, A*, 222:309–368.

Fortier, J. J. and Solomon, H. (1966). Clustering procedures. In Krishnaiah, P. R., editor, *Multivariate Analysis*, pages 493–506. Academic Press.

García, C. E. and D. M. Prett, M. M. (1989). Model predictive control: Theory and practice – a survey. *Automatica*, 25(3):335–348.

Geary, D. C., Bow-Thomas, C. C., and Yao, Y. (1992). Counting knowledge and skill in cognitive addition: A comparison of normal and mathematically disabled children. *Journal of Experimental Child Psychology*, 54:372–391.

Golightly, A. and Wilkinson, D. J. (2008). Bayesian inference for nonlinear multivariate diffusion models observed with error. *Computational Statistics and Data Analysis*, 52(3):1674–1693.

Good, I. J. (1950). *Probability and the Weighing of Evidence*. Hafners.

Goodwin, B. C. (1965). Oscillatory behavior in enzymatic control processes. *Advances in Enzyme Regulation*, 3:425–438.

Gordon, N. J., Salmond, D. J., and Smith, A. F. M. (1993). Novel approach to nonlinear/non-gaussian Bayesian state estimation. *IEEE Proceedings on Radar and Signal Processing*, 140(2):107–113.

Goshima, G., Wollman, R., Goodwin, S. S., Zhang, N., Scholey, J. M., Vale, R. D., and Stuurman, N. (2007). Genes required for mitotic spindle assembly in Drosophila S2 cells. *Science*, 316:417–421.

Gross, M. and Vögeli, C. (2007). A multimedia framework for effective language training. *Computer & Graphics*, 31:761–777.

Grünwald, P. D. (2007). *The Minimum Description Length Principle*. MIT Press.

Hadamard, J. (1902). Sur les problèmes aux dérivées partielles et leur signification physique. Technical report, Bullettin of the University of Princeton.

Haffner, J., Baro, K., Parzer, P., and F., R. (2005). *Heidelberger Rechentest (HRT): Erfassung Mathematischer Basiskompetenzen im Grundschulalter*. Hogrefe Verlag.

Harder, N., Mora-Bermúdez, F., Godinez, W. J., Wünsche, A., Eils, R., Ellenberg, J., and Rohr, K. (2009). Automatic analysis of dividing cells in live cell movies to detect mitotic delays and correlate phenotypes in time. *Genome Research*, 19:2113–2124.

Hardin, C. S. and Taylor, A. D. (2008). A peculiar connection between the axiom of choice and predicting the future. *American Mathematical Monthly*, 115(2):91–96.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer.

Hauser, A. (2009). Entropy-based experimental design for model selection in systems biology. Master's thesis, advised by A. G. Busetto and supervised by J. M. Buhmann, Deparment of Computer Science, ETH Zurich.

He, Y. B. and Geng, Z. (2010). Journal of machine learning research. *Active Learning of Causal Networks with Intervention Experiments and Optimal Designs*, 9:2523–2547.

Held, M., Schmitz, M. H. A., Fischer, B., Walter, T., Neumann, B., Olma, M. H., Peter, M., Ellenberg, J., and Gerlich, D. W. (2010). CellCognition: Time-resolved phenotype annotation in high-throughput live cell imaging. *Nature Methods*, 7(9):747–754.

Heray, A. and Frasson, C. (2009). Predicting learner answers correctness through brain-waves assessment and emotional dimensions. In *Proc. AIED*, pages 49–56.

Hofmann, T. and Buhmann, J. (1997a). Pairwise data clustering by deterministic annealing. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 19(1):1–14.

Hofmann, T. and Buhmann, J. M. (1997b). Pairwise data clustering by deterministic annealing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(1):1–14.

Hopcroft, J., Motwani, R., and Ullman, J. (2007). *Introduction to Automata Theory, Languages, and Computation.* Pearson/Addison Wesley.

Jain, A. K. and Dubes, R. C. (1988). *Algorithms for Clustering Data.* Prentice-Hall.

Jaynes, E. T. (1957a). Information theory and statistical mechanics. *Physical Rewiew Series II*, 106(4):620–630.

Jaynes, E. T. (1957b). Information theory and statistical mechanics ii. *Physical Rewiew Series II*, 109(2):171–190.

Jaynes, E. T. (2004). *Probability Theory.* Cambridge University Press.

Jensen, R. E. (1969). A dynamic programming algorithm for cluster analysis. *Operations Research*, 17:1034–1057.

Johns, J. and Woolf, B. (2006). A dynamic mixture model to detect student motivation and proficiency. In *Proc. AAAI*, pages 163–168.

Jonesa, T. R., Carpentera, A. E., Lamprechtb, M. R., Moffat, J., Serena, J. S., Greniera, J. K., Castorenod, A. B., Eggertd, U. S., Roota, D. E., Gollandc, P., and M., S. D. (2009). Scoring diverse cellular morphologies in image-based screens with iterative feedback and machine learning. *Proceedings of the National Academy of Sciences*, 106(6):1826–1831.

Julier, S. J. and Uhlmann, J. K. (2004). Unscented filtering and nonlinear estimation. *Proceedings of the IEEE*, 92(3):401–422.

Käser-Jacober, T. C., Busetto, A. G., Baschera, G.-M., Kohn, J., Kucian, K., von Aster, M., and Gross, M. (2012). Modelling and optimizing the process of learning mathematics. In *Springer Lecture Notes in Computer Science, 7315 Proc. ITS*, pages 389–398.

Kast, M., Meyer, M., Vögeli, C., Gross, M., and Jäncke, L. (2007). Computer-based multi-sensory learning in children with developmental dyslexia. *Restorative Neurology and Neuroscience*, 25(3-4):355–369.

Keynes, J. M. (1921). *A Treatise on Probability*. MacMillan and Co.

Khinchin, A. I. (1957). *Mathematical Foundations of Information Theory*. Dover.

King, O. D. and Forsyth, D. A. (2000). How does CONDENSATION behave with a finite number of samples? In *Proceedings of the European Conference on Computer Vision*, pages 695–709.

Kirkpatrick, M. (2009). Patterns of quantitative genetic variation in multiple dimensions. *Genetica*, 136:271–284.

Kitano, H. (2002). Computational systems biology. *Nature*, 420:206–210.

Kocha, H. and Tataru, D. (2001). Well-posedness for the Navier-Stokes equations. *Advances in Mathematics*, 157(1):22 –35.

Kolmogorov, A. N. (1933). *Grundbegriffe der Wahrscheinlichkeitrechnung*. Ergebnisse Der Mathematik.

Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1(1):4–7.

Kolmogorov, A. N. (1974). Complexity of algorithms and objective definition of randomness. In *Talk at Moscow Mathematical Society Meeting (transl. from Russian by L. A. Levin)*, Moscow.

Kramer, A. and Radde, N. (2010). Towards experimental design using a Bayesian framework for parameter identification in dynamic intracellular network models. In *Proceedings of the International Conference on Computational Science*, pages 1645–1653.

Krause, A. (2010). SFO: A toolbox for submodular function optimization. *Journal of Machine Learning Research*, 11:1141–1144.

Krause, A. and Guestrin, C. (2005). Near-optimal nonmyopic value of information in graphical models. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*.

Krause, A. and Guestrin, C. (2007). Near-optimal observation selection using submodular functions. In *Proceedings of the 22nd Conference on Artificial Intelligence*, pages 1650–1654.

Krause, A., McMahan, B., Guestrin, C., and Gupta, A. (2008). Robust submodular observation selection. *Journal of Machine Learning Research*, 9:2761–2801.

Kreutz, C. and Timmer, J. (2009). Systems biology: Experimental design. *FEBS Journal*, 276:923–942.

Krummenacher, G. (2010). Large-scale experimental design toolbox for systems biology. Master's thesis, advised by A. G. Busetto and supervised by J. M. Buhmann, Deparment of Computer Science, ETH Zurich.

Kucian, K. and Kaufmann, L. (2009). A developmental model of number representation. *Behavioral and Brain Sciences*, 32:313–373.

Kuepfer, L., Peter, M., Sauer, U., and Stelling, J. (2007). Ensemble modeling for analysis of cell signaling dynamics. *Nature Biotechnology*, 25(9):1001–1006.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86.

Lange, T., Braun, M., Roth, V., and Buhmann, J. M. (2004). Stability-based validation of clustering solutions. *Neural Computation*, 16(6):1299–1323.

Li, M. and Vitányi, P. (1997). *An Introduction to Kolmogorov Complexity and Its Applications*. Springer.

Lindley, D. (1991). *Making Decision*. Wiley.

Lindley, L. J. (1982). Scoring rules and the inevitability of probability. *International Statistical Review*, 50:1–26.

Loo, L. H., Wu, L. F., and Altschuler, S. J. (2007). Image-based multivariate profiling of drug responses from single cells. *Nature Methods*, 4(5):445–453.

MacKay, D. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297.

Minoux, M. (1978). Accelerated greedy algorithms for maximizing submodular set functions. *Optimization Techniques*, 7:234–243.

Mooij, J. M. (2010). Libdai: A free & open source C++ library for discrete approximate inference in graphical models. *Journal of Machine Learning Research*, 11:2169–2173.

Murphy, K. P. (2001). The bayes net toolbox for MATLAB. *Computing Science and Statistics*, 33:2001.

Murphy, K. P. (2002). *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, University of California, Berkeley.

Myung, J. I. and Pitt, M. A. (2009). Optimal experimental design for model discrimination. *Psychological Review*, 116(3):499–518.

Nelles, O. (2001). *Nonlinear System Identification*. Springer.

Nemhauser, G. L., Wolsey, L. A., and Fisher, M. L. (1978). An analysis of approximations for maximizing submodular set functions. *Mathematical Programming*, 14(1):265–294.

Neumann, B., Walter, T., Hériché, J. K., Bulkescher, J., Erfle, H., Conrad, C., Rogers, P., Poser, I., Held, M., Liebel, U., Cetin, C., Sieckmann, F., Pau, G., Kabbe, R., Wünsche, A., Satagopam, V., Schmitz, M. H., Chapuis, C., Gerlich, D. W., Schneider, R., Eils, R., Huber, W., Peters, J. M., Hyman, A. A., Durbin, R., Pepperkok, R., and Ellenberg, J. (2010). Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. *Nature*, 464:721–727.

Ostad, S. A. (1997). Developmental differences in addition strategies: A comparison of mathematically disabled and mathematically normal children. *British Journal of Education Psychology*, 67:345–357.

Ostad, S. A. (1999). Developmental progression of subtraction strategies: A comparison of mathematically normal and mathematically disabled children. *European Journal of Special Needs Education*, 14:21–36.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2:559–572.

Pham, T. and Tran, D. (2006). Gaussian mixture and Markov models for cell-phase classification in microscopic imaging. In *Proceedings of the 2006 IEEE/SMC International Conference on System of Systems Engineering*, pages 328–333.

Pinheiro, J. C. and Bates, D. M. (1995). Approximations to the log-likelihood function in non-linear mixed-effects models. *Journal of Computational and Graphical Statistics*, 4(1):12–35.

Ponce de Leon, A. C. and Atkinson, A. (1991). Optimum experimental design for discriminating between two rival models in the presence of prior information. *Biometrika*, 78(3):601–608.

Pronzato, L. (2008). Optimal experimental design and some related control problems. *Automatica*, 44(2):303–325.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Raman, K. and Wagner, A. (2011). Evolvability and robustness in a complex signalling circuit. *Molecular Biosystems*, 7(4):1081–1092.

Ramsey, F. (1931). *Truth and Probability*. Humanities Press.

Rathmanner, S. and Hutter, M. (2011). A philosophical treatise of universal induction. *Entropy*, 13(6):1076–1136.

Richard O. Duda, Peter E. Hart, D. G. S. (2000). *Pattern Classification*. Wiley-Interscience Publication.

Risken, H. (1996). *The Fokker-Planck Equation: Methods of Solutions and Applications*. Springer.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14:465–471.

Roth, V., Laub, J., Kawanabe, M., and Buhmann, J. M. (2003). Optimal cluster preserving embedding of non-metric proximity data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1540–1551.

Rubinsten, O. and Tannock, R. (2010). Mathematics anxiety in children with developmental dyscalculia. *Behavioral and Brain Functions*, 6(46).

S. Ben-David, U. von Luxburg, D. P. (2006). A sober look at clustering stability. In *Springer Verlag LNAI 4005 Proceedings of COLT*, pages 5–19.

Savage, L. J. (1961). The subjective basis of statistical practice. Technical report, Department of Statistics, University of Michigan, Ann Arbor.

Schmitz, M. H., Held, M., Janssens, V., Hutchins, J. R., Hudecz, O., Ivanova, E., Goris, J., Trinkle-Mulcahy, L., Lamond, A. I., Poser, I., Hyman, A. A., Mechtler, K., Peters, J. M., and Gerlich, D. W. (2010). Live-cell imaging RNAi screen identifies PP2A-B55alpha and importin-beta1 as key mitotic exit regulators in human cells. *Nature Cell Biology*, 12(9):886–893.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464.

Seldin, Y. and Tishby, N. (2010a). Pac-Bayesian analysis of co-clustering and beyond. *Journal of Machine Learning Research*, 11:3595–3646.

Seldin, Y. and Tishby, N. (2010b). Pac-Bayesian analysis of co-clustering and beyond. *Journal of Machine Learning Research*, 11:3595–3646.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press.

Shalev, R. and von Aster, M. G. (2008). Identification, classification, and prevalence of developmental dyscalculia. *Encyclopedia of Language and Literacy Development*, pages 1–9.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423.

Shannon, C. E. and Weaver, W. (1963). *The Mathematical Theory of Communication*. University of Illinois Press.

Skanda, D. and Lebiedz, D. (2010). An optimal experimental design approach to model discrimination in dynamic biochemical systems. *Bioinformatics*, 26(7):939–945.

Slonim, N., Atwal, G. S., Tracik, G., and Bialek, W. (2005). Information-based clustering. *Proceedings of the National Academy of Sciences*, 102:1827–1830.

Smets, P. and Kennes, R. (1994). The transferable belief model. *Artificial Intelligence*, 66:191–234.

Solomonoff, R. (1964a). A formal theory of inductive inference, part i. *Information and Control*, 7(1):1–22.

Solomonoff, R. (1964b). A formal theory of inductive inference, part ii. *Information and Control*, 7(2):224–254.

Steinhaus, H. (1957). Sur la division des corps matériels en parties. *Bulletin de l'Academie Polonaise des Sciences*, 4(12):801–804.

Steinke, F., Seeger, M., and Tsuda, K. (2007). Experimental design for efficient identification of gene regulatory networks using sparse Bayesian models. *BMC Systems Biology*, 1(51).

Stoer, J., B. R. (2002). *Introduction to Numerical Analysis*. Springer.

Sunnåker, M., Busetto (eq.), A. G., Numminen (eq.), E., Corander, J., Foll, M., and Dessimoz, C. (2012). Approximate bayesian computation. *PLoS Computational Biology*, page (in press).

Suyari, H. (2004). Generalization of shannon-khinchin axioms to nonextensive systems and the uniqueness theorem for the nonextensive entropy. *IEEE Transactions on Information Theory*, 50(8):1783–1787.

Szallasi, Z., Stelling, J., and Periwal, V. (2006). *System Modeling in Cell Biology: from Concepts to Nuts and Bolts*. MIT Press.

Tishby, N., Pereira, F. C., and Bialek, W. (1999). The information bottleneck method. In *Proceedings of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377.

Tseng, G. C. (2007). Penalized and weighted K-means for clustering with scattered objects and prior information in high-throughput biological data. *Bioinformatics*, 23(17):2247–2255.

Turing, A. M. (1937). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, 2(42):230–265.

Valiant, L. (1984). A theory of the learnable. *Communications of the ACM*, 27:1134–1142.

van Fraassen, B. (1980). *The Scientific Image*. Oxford University Press.

Vapnik, V. N. (1982). *Estimation of the Dependencies Based on Empirical Data*. Springer.

Viterbi, A. J. (1967). Error bounds for convolution codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13:260–269.

von Aster, M. G. and Shalev, R. (2007). Number development and developmental dyscalculia. *Developmental Medicine and Child Neurology*, 49:868–873.

Wallace, C. and Boulton, D. (1968). An information measure for classification. *Computer Journal*, 11(2):185–194.

Wang, M., Zhou, X., King, R., and Wong, S. (2007). Context based mixture model for cell phase identification in automated fluorescence microscopy. *BMC Bioinformatics*, 8(32).

Wang, M., Zhou, X. B., Li, F. H., Huckins, J., King, R. W., and Wong, S. T. C. (2008). Novel cell segmentation and online SVM for cell cycle phase identification in automated microscopy. *Bioinformatics*, 24(1):94–101.

Wasserman, L. (2003). *All of Statistics: A Concise Course in Statistical Inference*. Springer.

Whewell, W. (1837). *History of the Inductive Sciences*. John W. Parker.

Wilkinson, D. J., editor (2006). *Stochastic Modelling for Systems Biology*. Chapman & Hall/CRC.

Wilkinson, D. J. (2007). Bayesian methods in bioinformatics and computational systems biology. *Bioinformatics*, 8:109–116.

Wilkinson, D. J. (2011). *Stochastic Modelling for Systems Biology*. Chapman and Hall.

Zadeh, L. A. (1978). Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1:3–28.

Zermelo, E. (1908). Untersuchungen über die Grundlagen der Mengenlehre i. *Mathematische Annalen*, 65:261–281.

Zhong, Q., Busetto, A. G., Fededa, J. P., Buhmann, J. M., and Gerlich, D. W. (2012). Unsupervised modeling of cell morphology dynamics for time-lapse microscopy. *Nature Methods*, 9:711–713.

Alberto Giovanni Busetto

2007–2012:
Doctoral Studies
ETH Zurich, Department of Computer Science & CC-SPMD;

2002–2007:
Laurea Specialistica (e Triennale) in Ingegneria Informatica
Università degli Studi di Padova;

1997–2002:
Diploma di Maturità Scientifica, PNI
Liceo Scientifico G. B. Benedetti, Venezia.

Sommario

I sistemi dinamici sono modelli matematici che esprimono relazioni di causa-effetto riguardanti fenomeni soggetti a variazione temporale. Questa tesi si concentra sulla stima di sistemi dinamici sulla base di osservazioni empiriche. Si considerano tre scenari: stima senza supervisione, con supervisione e attiva. L'obiettivo unificante è l'estrazione di informazione predittiva dai dati.

Viene introdotto un metodo per il raggruppamento di serie temporali e la validazione statistica di modelli. Il metodo si propone di risolvere le questioni di selezione d'ordine e di modello con il principio della Codifica tramite Insiemi di Approssimazione [Buhmann (2010)]. La verifica sperimentale è perseguita nell'ambito di raggruppamento relazionale per profili temporali di espressione genetica. I risultati dimostrano vasta applicabilità e congruenza col Criterio d'Informazione Bayesiano. Inoltre, le transizioni dinamiche discrete sono ricostruite sulla base di serie temporali d'alta dimensionalità tramite un approccio privo di supervisione. L'approccio, che si basa su Modelli Markoviani Nascosti su Miscele Gaussiane, trova applicazione nella predizione di classi morfologiche da dati di microscopia che tengono conto del fattore tempo. La conferma sperimentale con marcatori fluorescenti e cernita dei campioni dimostra la capacità di accurata identificazione dei fenotipi cellulari umani. I risultati riportati evidenziano competitività e miglioramento dell'oggettività rispetto agli approcci supervisionati basati sulla catalogazione operata da parte dell'utente.

Nello scenario con supervisione, il processo di raggruppamento trova impiego nel perfezionamento del filtraggio convenzionale di campioni al fine di ottenere stime di stato generalizzate. Il raggruppamento preventivo mitiga l'inevitabile divergenza del ricampionamento impiegato nei metodi sequenziali di tipo Monte Carlo. La stima supervisionata con reti Bayesiane dinamiche è utilizzata nella modellistica dell'apprendimento umano con il fine di curare efficacemente alcune disabilità dell'apprendimento. Per la dislessia, il modello è in grado di predire rate di dimenticanza, livelli di concentrazione e inclinazione all'apprendimento del soggetto sulla base del comportamento misurato. Per la discalculia, la cognizione numerica è migliorata per mezzo di una terapia adattabile ottenuta sulla base del modello.

Nel contesto della stima attiva, la tesi si concentra sulla progettazione quasi ottima di esperimenti finalizzati alla modellistica di sistemi dinamici. Viene introdotto un metodo efficiente per selezionare quantità misurabili e istanti temporali particolarmente informativi. Si garantisce la quasi ottima informatività del metodo di progettazione, il quale richiede un numero polinomiale di valutazioni della funzione obiettivo. Il metodo si basa su lavoro precedente nell'ambito della stima attiva sotto-modulare [Krause and Guestrin (2005)] e ottiene il migliore fattore costante di approssimazione possibile, a meno che non valga P=NP [Feige (1998)]. La progettazione di esperimenti trova applicazione nella ricostruzione di reti cellulari segnalatrici nell'ambito della biologia dei sistemi.

I contributi presentati sottolineano le analogie fondamentali tra stima e comunicazione. In conclusione, i risultati dimostrano che modelli predittivi possono essere costruiti sulla base di efficienti strategie di trasmissione d'informazione tramite un canale rumoroso. Sulla base di argomentazioni di natura statistica, i risultati esibiti formalizzano e automatizzano aspetti del metodo ipotetico-deduttivo di investigazione scientifica.