HELMHOLTZ
ASSOCIATION
Alliance on Systems Biology
HelmholtzZentrum münchen
German Research Center for Environmental Health

TECHNISCHE
UNIVERSITÄT
MÜNCHEN

From low-dimensional model selection to high-dimensional inference: tailoring Bayesian methods to biological dynamical systems

Sabine Carolin Hug

Juni 2014

TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl M12 (Mathematische Modellierung biologischer Systeme)

"From low-dimensional model selection to high-dimensional inference: tailoring Bayesian methods to biological dynamical systems "

Sabine Carolin Hug

Vollständiger Abdruck der von der Fakultät für Mathematik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzende: Univ.-Prof. Claudia Czado, Ph.D.

Prüfer der Dissertation:

1. Univ.-Prof. Dr. Dr. Fabian J. Theis
2. Univ.-Prof. Dr. Oliver Junge
3. Prof. Dr. Mark Girolami, University of Warwick / UK

   (nur schriftliche Beurteilung)

Die Dissertation wurde am 26.06.2014 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 05.12.2014 angenommen.

Für Alexander und Josef

Acknowledgments

I have always liked the little glimpse of the personality of a scientist that is revealed by the acknowledgments in their thesis, to see who shared this special time with them. When writing this thesis, I have thus looked forward to thanking all the people who have accompanied me for the last four years and contributed to this thesis. So here they come!

Sincere gratitude to my supervisor Fabian Theis for giving me the chance to learn so many interesting new things in his group while doing my PhD work! Thank you, Fabian, for putting your trust in me. And thank you for the many opportunities I had to meet interesting people and work on exciting projects.

A thank you is also due to my second reviewer Oliver Junge, who also joined my thesis committees and thus helped steer this thesis into the direction it has finally taken, for his guidance and advices.

I am also grateful to Mark Girolami for taking the time to read this thesis and write an evaluation on it and for always providing interesting lines of thought when we met at conferences.

Thank you to my collaboration partners Weibo Li, Matthias Greiter, Philipp Hoppe, Michael Halter and Julie Bachmann for trusting me with your data and thus successful collaborations.

Daniel, thanks a million! Without you, this thesis would be very different, of that I am sure. Thank you for patiently answering my many questions, for making working with you so much fun and educating at the same time. Your mentoring of me was much appreciated!

Andreas, thank you for many interesting discussions of "profiles versus MCMC". Also writing a paper with you taught me a lot.

Christiane, thank you for being a wonderful team leader. Thank you for always taking time for me, for your mathematical precision and for proof-reading this thesis so thoroughly. Any typos left are certainly my fault!

Jan, I really enjoyed working with you and thus I look much forward to being a Postdoc in your team and hopefully learn a few of the amazingly many things that you know. I am sure working with you will be both educating and entertaining!

Michi, thank you for our nice project on the single cells. I feel that we did a fair trade of mathematics and bioinformatics knowledge. Carsten, also your suggestions and guidance on the project were much appreciated.

Sabrina and Ferdi, thanks for being the best roomies in the whole world. I immensely enjoyed sharing an office with you, drinking tea, eating fruit and discussing everything from science to soccer. Thanks for patient assistance with Illustrator and tolerating my wish for 25°C in the office.

I thank the whole institute for the really enjoyable working atmosphere! It was always fun to organize something for you. I enjoyed evenings of Barbecue and tablesoccer very much and would like like to thank all my tablesoccer teachers for their effort.

Everybody who has already written a thesis might agree with me that the non-scientific support by family and friends is as important as the scientific one by supervisors and colleagues. So many thanks to all of you, for always supporting me and not complaining too much when I didn't have enough time for you in the last few months.

I especially want to thank my family for being a constant source of encouragement and strength.

Alexander, thank you! (hug)

Abstract

Ordinary differential equations (ODEs) have become a valuable tool for mathematically representing biological processes as dynamical systems. However, inference of the associated parameters for fitting the differential equations to measurement data is not straightforward. Furthermore, if the structure of the ODE itself is uncertain, i.e. if several competing models exist, model selection methods have to be applied. In the last few years a steady increase in the use of Bayesian methods for both parameter and model inference could be noted. These are especially suited to inference in biological systems, since they can efficiently handle the occurring issues such as small numbers of observables, considerable parameter correlations, non-identifiabilities and nonlinearities. Due to the complex nature of the problems at hand, it is however almost never possible to use Bayesian methods "out of the box", but they always have to be adapted to the specific task at hand, often due to scaling or dimensionality issues. This was thought to be hard to impossible in higher dimensional systems. In this work, we are able to show that nevertheless efficient inference is possible with the newly improved methods that we introduce. More precisely, we refine several existing Bayesian methods, ranging from an adaptive scheme for the computation of high-dimensional integrals required for model selection to multi-chain Metropolis-Hastings algorithms for high-dimensional parameter inference. We then present a range of examples spanning from simple models with lots of data where model inference is difficult to highly complex, high-dimensional models where parameter inference is challenging.

Zusammenfassung

Gewöhnliche Differentialgleichungen sind ein wertvolles Instrument um biologische Prozesse mathematisch als dynamische Systeme zu repräsentieren. Allerdings ist die Inferenz der zugehörigen Parameter alles andere als einfach, wenn an Messdaten gefittet werden soll. Ist zudem die Struktur der Differentialgleichung an sich unsicher, d.h. wenn mehrere konkurrierende Modelle existieren, müssen Modellselektionsmethoden angewandt werden. In den letzten Jahren wurden Bayesianische Methoden sowohl für Parameter- als auch Modellinferenz immer beliebter. Diese sind besonders geeignet für die Inferenz in biologischen Systemen, da sie effizient mit auftretenden Problemen wie etwa einer kleinen Anzahl beobachtbarer Komponenten, bedeutenden Parameterkorrelationen, Nichtidentifizierbarkeiten und Nichtlinearitäten umgehen können. Wegen der komplizierten Natur biologischer dynamischer Systeme ist es jedoch nur in den seltensten Fällen möglich, vorhandene Methoden als fertige Standardmethoden anzuwenden, sondern sie müssen an die jeweils vorliegende spezielle Problemstellung angepasst werden, oftmals wegen Skalierungs- oder Dimensionalitätsproblemen. Das wurde für schwer bis unmöglich in höherdimensionalen Systemen gehalten. In dieser Arbeit sind wir in der Lage zu zeigen, dass nichtsdestotrotz effiziente Inferenz möglich ist dank der neu weiterentwickelten Methoden, die hier präsentiert werden. Genauer gesagt verfeinern wir eine Reihe existierender Bayesianischer Methoden. Das Spektrum reicht hierbei von adaptiven Schemata zur Berechnung hochdimensionaler Integrale, die für die Modellselektion benötigt werden, bis hin zu Multiketten-Metropolis-Hastings-Algorithmen für hochdimensionale Parameterinferenz. Wir präsentieren eine Sammlung verschiedener Anwendungsbeispiele, die von einfachen Modellen mit vielen Datenpunkten, wo Modellinferenz schwer

ist, bis zu hochkomplexen, hochdimensionalen Modellen, wo Parameterinferenz allein eine Herausforderung darstellt, reicht.

Contents

1

Introduction

If you can't explain it simply,
you don't understand it well enough.

Albert Einstein

1.1 Scientific question of this thesis

At first glance, mathematics and biology might seem to be very two very differently oriented sciences. In mathematics there is usually a unique and provable solution for a problem or none at all, while in biology often many contradictory theories exist even for simple processes. While many principles of mathematics have stood unchanged for centuries, in biology, new theories for interactions are proposed on a daily basis and almost no two biological entities are ever exactly the same due to biological variability. However, scientific discoveries starting from the 20th century have shown that nevertheless the application of mathematics to biological problems is extremely worthwhile

and can yield many exciting new insights.

Gaining knowledge in biology is usually only possible by doing experiments. This of course can have several disadvantages. First of course, experiments are tainted with measurement noise, which has to be considered in order to get reliable results. Second, often not all components in the biological system are measurable due to technical or monetary limitations. On top of that, there is the biological variability between individuals. Nevertheless many new discoveries have been made thanks to experiments.

Despite and also because of all these difficulties, biological systems are a very interesting and challenging task for applied mathematics. The complexity of biological systems encourages mathematicians to develop better methods, to the profit of both sides.

This successful collaboration began when scientists started applying statistical methods for analyzing biological data. More and more complex approaches have been applied successfully since roughly the 1940s, see Hodgkin & Huxley [1952]; Turing [1952]; von Foerster [1959]; Welch [1947]. Wiener [1948] was one of the first to introduce a systems view. Kitano [2002a,b] then provides an overview over the field of **systems biology**.

For understanding a biological process at the systems level, often models are build for examining the process behavior as a whole. However, care has to be taken when building a model. As Box & Draper [1987] so nicely put it: "Essentially, all models are wrong, but some are useful." Thus it should be clear before building a model what the aim of such a model should be, e.g. whether its purpose is in-depth understanding of system behavior or the prediction of responses to different stimuli (Kitano [2002b]). Models being useful here means that even though models always are simplified views of the complex biological system, they can still offer mechanistic insights. The art here is to choose the right level of abstraction to make the model still representative, but computationally feasible. Lawrence *et al.* [2010] also call this a modeling compromise. The example they provide illustrates this point very well: Even if it is known that a system exhibits dynamical behavior, the assumption that the chemical reaction rates are large relative to the interval at which the system is observed allows us to ignore the dynamic behavior. This can result in enormous computational savings, even if it is clear that measurements of the system at smaller time intervals would refute the

model hypotheses immediately. Still the simplification can be well justified if it allows to answer a specific question.

An ideal way to include several such competing modeling hypotheses is through a Bayesian framework, where individual hypotheses are sustained through probability distributions over parameters. The Bayesian approach allows assessing the probability of a hypothesis together with its underlying parameter uncertainty, which is an important advantage. For practical reasons, it is usually not feasible to consider all possible complexity levels of hypotheses (Lawrence *et al.* [2010]). In real world applications, one thus often is constrained to only a few model hypotheses.

For understanding a system, it is an important first step to identify all components and their interactions, the biological network. As Kitano [2002b] points out, it is then especially important to understand the dynamics of a biological system, not only the network wiring at the basis of the system. The author likens this to a static street map, which does not give any answers to questions such as where traffic patterns emerge and why and how they can be controlled.

For modeling the dynamics of a system, differential equations are a very useful tool (Aldridge *et al.* [2006]) as they can give new mechanistic insights, e.g. in Aldridge *et al.* [2011]; Kollmann *et al.* [2005]; Swameye *et al.* [2003]. They can be used to analyze the evolution and maintenance of cellular functionality over time. The size of the model can vary dramatically depending on the questions asked, from simple linear models of one component to large and complex nonlinear systems of differential equations.

All differential equation models have in common that they are usually parameter dependent, e.g. on rate constants, initial conditions etc. These parameters have to be inferred to fit the model to the data in question (Ljung [1999]), we thus have to perform **parameter inference**.

If the structure of the biological system itself is uncertain, we are faced with the problem of structural inference, also called **model selection**. The target here is to choose among a finite set of candidate models the one which best explains the data. Often these candidate models correspond to different hypotheses which interactions are present in the biological system (Lawrence *et al.* [2010]). For structural inference a variety of indicators are applied, originating from various fields (Kirk *et al.* [2013]).

Parameter inference and also structural inference can be called an inverse problem or reverse engineering. Parameter inference assesses the values and uncertainty of parameters within the model, while structural inference compares several such models. Not only the size of the model, but consequently also the number of parameters can vary dramatically between models, from one or two to several hundred. Thus parameter estimation is found everywhere in systems biology (Engl *et al.* [2009]) and extensive research has gone into the field (Horbelt *et al.* [2002]), producing a wide array of parameter estimation procedures. These are often optimization based, using different cost functions, such as likelihoods. A recent overview over different methods has been given in Villaverde & Banga [2014].

Tackling both parameter inference and model selection at the same time is possible with Bayesian statistics (Bayes [1763]). Bayesian methods found increasing popularity due to their ability to extract information from uncertain and noisy data as well as the possibility to include prior knowledge to yield the posterior distribution of the model parameters given the provided measurement data. Analytical inference of this distribution becomes quickly infeasible, yet Markov chain Monte Carlo (MCMC) methods can be ideally applied, as these can deal with considerable parameter correlations, non-identifiabilities and nonlinearities. Pioneering work using Monte Carlo methods comes from Battogtokh *et al.* [2002]; Brown & Sethna [2003]; Sanguinetti *et al.* [2006]. It was made possible by advances in computer technology and the development of Markov chain Monte Carlo methods such as the Metropolis-Hastings algorithm (Hastings [1970]; Metropolis *et al.* [1953]), which is very generally applicable. Brown & Sethna [2003]; Lawrence *et al.* [2010]; Wilkinson [2006] are also examples of successful Bayesian parameter inference in differential equation models. Bayesian model selection methods often also use MCMC methods for approximating the marginal likelihoods of the models given the data.

When using MCMC methods especially for ordinary differential equation (ODE) models from systems biology, one soon realizes that it is not possible to apply standard algorithms "out of the box". Instead, MCMC methods have to be tuned to the model at hand. Sometimes even new methods have to be developed for reliable inference. This is also due to issues of MCMC concerning scalability, as it is often not clear how to implement an efficient MCMC method in high-dimensional spaces.

Parameters are not always "well-behaved", sometimes due to limitations of the data like low time resolution or measurement noise (Gutenkunst *et al.* [2007]; Komorowski *et al.* [2011]) and/or to deeper underlying problems of the model that are the focus of identifiability analysis (Cobelli & DiStefano [1980]; Little *et al.* [2010]; Raue *et al.* [2009]). A strict analysis of parameter and prediction uncertainty has to take place for obtaining reliable and meaningful results. This is often also called uncertainty analysis.

In this thesis, we focus on efficient inference tailored to the biological system at hand. For this we improve existing Bayesian inference methods for both parameter and model inference. The advancement of the methods is driven by the issues occurring in the examined biological systems, which are among others small numbers of observables, considerable parameter correlations, non-identifiabilities and nonlinearities. We have found that due to the complex nature of the problems at hand, it is almost never possible to use Bayesian methods "out of the box". Instead, they always have to be adapted to the specific task at hand, which was thought to be hard to impossible in higher dimensional systems. It is for example well known that MCMC methods do not scale well with increasing sampling dimensions. In this work, we are able to show that nevertheless efficient inference is possible with the improved methods that we introduce. All in all, we refine several existing Bayesian methods, ranging from an adaptive scheme for the computation of high-dimensional integrals required for model selection to multi-chain Metropolis-Hastings algorithms for high-dimensional parameter inference.

We present in detail three application examples from systems biology. We go from doing model selection on two small ODE models with lots of data, where we improved existing model selection methods, over a medium sized application to a large ODE system with many parameters, where we applied a new method for MCMC. Each of the three applications poses its own difficulties and all three require different methodology for handling them.

In summary, the aim of this thesis is to show how Bayesian methods can and have to be improved to yield reliable inference results in varying biological systems. For this, we also combine Bayesian methods with other mathematical methods such as numerical quadrature and optimization-based uncertainty analysis. In the following two sections,

we will first give an overview of the presented topics and then give a review of the main scientific contributions of this thesis.

1.2 Overview of this thesis

This thesis is separated into two parts. The first part, comprising chapters 2 to 5, focuses on the methodological aspects. This includes an overview over already known methods as well as methods newly extended in this thesis. The second part, comprising of chapters 6 to 8, presents the biological applications to which the presented methodology was applied.

Thus in the first chapter after this one, Chapter 2, we will give a brief overview over the necessary prerequisites for the rest of this thesis. This includes some basics on probability distributions, Markov chains and numerical integration, as well as a brief introduction to molecular biology. The main tool for modeling processes in molecular biology considered in this thesis are ordinary differential equations, which are hence also introduced. We also give a short introduction of non-Bayesian parameter estimation in such ordinary differential equation systems.

Chapter 3 introduces Bayesian inference, from prior and posterior distributions to the analysis of identifiability. While identifiability analysis as such does not have to be Bayesian, it can also be done in a Bayesian way with the help of profile posteriors, which are also introduced.

Markov chain Monte Carlo algorithms are presented in Chapter 4. This includes well-known methods as the Metropolis-Hastings algorithm as well as new methods. One of these is the Adaptive Metropolis Parallel Hierarchical Sampling, which was newly constructed for one of the applications presented in this thesis.

The main methodological novelty is contained in Chapter 5. Here we first present Bayesian model selection through the computation of Bayes factors. We then introduce a new adaptive variant for calculating them through thermodynamic integration and evaluate the new scheme and several other well-known model selection techniques on a numerically tractable model selection example.

In Chapter 6 we present the first application of the methods doing model selection on single-cell time-lapse microscopy data. While the ordinary differential equation models in this application are small, the inference is nevertheless interesting, since single cells are fitted individually but in parallel, leading to large numbers of independent parameters that have to be inferred.

Chapter 7 presents model selection on a medium sized problem of ten versus eleven coupled linear ordinary differential equations, with twelve and fifteen parameters respectively. It compares models for Zirconium processing in the human body on the basis of compartmental models. We analyze thermodynamic integration in this context and present additionally the results of an identifiability analysis.

In Chapter 8 we are faced with a high-dimensional system of 113 parameters in a model for 25 molecular components. Here already the parameter inference is very challenging and standard MCMC algorithms fail. We show how multi-chain algorithms can be applied successfully and how a combination of Bayesian inference with identifiability analysis is worthwhile for obtaining reliable results.

Finally, Chapter 9 presents a discussion of the presented methods and applications. We draw general conclusions and show potential targets for future research.

1.3 Main scientific contributions

The main scientific contributions of this thesis are

- the introduction of a variant of multi-chain MCMC methods, the Adaptive Metropolis Parallel Hierarchical Sampling,
- the adaptive Simpson's rule for calculating marginal likelihoods through thermodynamic integration for model selection,
- inference, identifiability analysis and model selection in single-cell data based on individual treatments of single cells,
- analysis of thermodynamic integration results on two medium-sized linear multi-compartmental models combined with identifiability analysis and

- the proof-of-principle that Bayesian parameter inference of over 100 parameters in a biological system is possible.

These contributions were in part already published in peer-reviewed journals. Some parts of this thesis will thus correspond to or be in parts identical with the following publications:

- D. Schmidl*, **S. Hug***, W.B. Li, M.B. Greiter and F.J. Theis (2012). Bayesian model selection validates a biokinetic model for Zirconium processing in humans. BMC Systems Biology, 6(1), 95.
- **S. Hug***, A. Raue*, J. Hasenauer, J. Bachmann, U. Klingmüller, J. Timmer and F.J. Theis (2013). High-dimensional Bayesian parameter estimation: case study for a model of JAK2/STAT5 signaling. Mathematical Biosciences, 246(2), 293-304.
- **S. Hug**, D. Schmidl, W.B. Li, M.B. Greiter and F.J. Theis. Uncertainty in Biology: a computational modeling approach, chapter Bayesian model selection methods and their application to biological ODE systems, *in revision*
- **S. Hug**, M. Schwarzfischer, J. Hasenauer, C. Marr and F.J. Theis. An adaptive method for calculating Bayes factors using Simpson's rule, *in revision*
- M. Schwarzfischer, O. Hilsenbeck, B. Schauberger, **S. Hug**, A. Filipczyk, P.S. Hoppe, M. Strasser, F. Buggenthin, J.S. Feigelman, J. Krumsiek, D. Loeffler, K.D. Kokkaliaris, A.J.J. van den Berg, M. Endele, S. Hastreiter, C. Marr, F.J. Theis and T. Schroeder. Single-cell quantification of cellular and molecular behavior in long-term time-lapse microscopy, *in preparation*

We mark specifically in the beginning of a chapter where this is the case. Asterisks in the list indicate a shared first authorship.

The content of the first of these papers is also in part contained in another thesis (Schmidl [2012]), as this was a joint first author work. While the focus of Dr. Schmidl was more on the copula-based sampling of the models and the insights that can be gained from the posterior distribution, the author of this thesis performed the analysis of the thermodynamic integration and the identifiability analysis of the models. Furthermore, the author contributed the analysis of an additional model variant.

Also the second paper is a joint first author work, thus the content of this paper is also to a small part contained in another thesis (Raue [2013]). The contribution by the author of this thesis is the MCMC and its interpretation and evaluation, while Dr. Raue provided the profile likelihoods.

Further scientific contributions

In addition to the publications specified above, the author of this thesis participated in several other research projects, which are in some cases not directly related to the content of this thesis or where the contribution was more minor. These projects have led to the following publications:

- **S. Hug** and F.J. Theis (2012). Bayesian inference of latent causes in gene regulatory dynamics. In Latent Variable Analysis and Signal Separation, 520-527. Springer Berlin Heidelberg.
- D. Schmidl, C. Czado, **S. Hug**, F.J. Theis (2013). A vine-copula based adaptive MCMC approach for efficient inference of dynamical systems. Bayesian Analysis 8(1),1-22.
- D. Schmidl, C. Czado, **S. Hug**, F.J. Theis (2013). Rejoinder on: A vine-copula based adaptive MCMC approach for efficient inference of dynamical systems. Bayesian Analysis 8(1),33-42.
- A. Raue, M. Schilling, J. Bachmann, A. Matteson, M. Schelke, D. Kaschek, **S. Hug**, C. Kreutz, B.D. Harms, F.J. Theis, U. Klingmüller and J. Timmer (2013). Lessons learned from quantitative dynamical modeling in systems biology. PloS one 8(9), e74335.
- C. Vehlow, J. Hasenauer, A. Kramer, A. Raue, **S. Hug**, J. Timmer, N. Radde, F.J. Theis and D. Weiskopf (2013). iVUN: interactive Visualization of Uncertain biochemical reaction Networks. BMC Bioinformatics 14(Suppl 19), S2.

Part I

Methods

2

Mathematical and biological basics

This chapter introduces some notions from molecular biology that are relevant for this thesis. Furthermore, some notation, basic probability distributions, Markov chain theory and numerical integration are introduced, followed by basics about dynamical systems in biology and non-Bayesian parameter estimation in such systems are presented.

This chapter presents necessary preliminaries for the understanding of the subsequent chapters. Especially the Markov chain theory will be needed for understanding the required properties of MCMC algorithms presented in Chapter 4.

2.1 Molecular biology

In this section we shortly introduce some basic notions from biology relevant to our applications. These are from molecular biology in general, and more specifically about stem cells and signaling pathways, since these are important applications in this thesis, see Chapters 6 and 8.

2.1.1 Basics of molecular biology

The discovery of deoxyribonucleic acid (DNA) as the coding blocks for all living organisms has opened a tremendous amount of new avenues for biological research (Choudhuri [2003]; Olby [1974]). The Human Genome Project (Collins *et al.* [2003]) has decoded the entire human DNA in 2003, yet still many of the mechanistic interplays and dynamics of genetics are unclear. We do know that one single molecule contains the entire genetic information for the organism (Alberts *et al.* [2002]). While not everything is known about the mechanics, we still know that a **gene** is a small stretch of DNA that contains the genetic code for a protein (Pearson [2006]). Humans have about 20000 to 25000 of these genes. The information in a gene is used for building a protein through the two steps of **transcription** and **translation** (Crick [1970]). Transcription means that the so-called transcription factors read the gene of the DNA strand and initiate the assembly of RNA (ribonucleic acid). There are several distinct types of RNA, for example messenger RNA (mRNA) or micro RNA (miRNA). Not for all of them their complete function is clear yet. What is known however is that in the translation step, which takes place in the cell's cytoplasm, proteins are assembled according to the information in the RNA. Proteins are in this case assembled from smaller building blocks called amino acids. Two recent studies have shown that the human proteome contains about 20000 different proteins (Kim *et al.* [2014]; Wilhelm *et al.* [2014]).

Many cell processes are in some way or the other regulated by such proteins, protein complexes or peptides (short sequences of amino acids). Examples are numerous, ranging from structure proteins for cell skeletons, transcriptions factors, enzymes regulating metabolism or growth factors like cytokines and hormones that stimulate proliferation and cell growth, see e.g. Frixione [2000]; Hartwell & Weinert [1989]; Latchman [1997]; Lodish *et al.* [2012].

2.1.2 Signaling pathways

We know that proteins regulate the important functions in living organisms, but we now want to go into more detail of how they do that. One primary way is via **cellular signaling**. In this, signals are transmitted from cell to cell by proteins, small peptides, lipids or even single amino acids. A few mechanisms of signaling are known:

first is direct cell-to-cell transfer of molecules, for example notch signaling (Artavanis-Tsakonas *et al.* [1999]). The second mechanism is by the secretion of molecules from the signaling cell. The receiving cell reacts to the signal through its receptors on the cell surface, see e.g. Lodish *et al.* [2012]. More precisely: the signaling molecule binds to the extracellular receptor of the receiving cell. This triggers a biochemical reaction cascade where information is transported through the cell membrane into the cell. Inside the cell, a receptor associated kinase or kinase domain gets activated, which in turn activates other intracellular proteins or other signaling molecules. These can then be transported to a target, e.g. the nucleus of the cell, where they might for example control transcription of their target genes. In general, this works by phosphorylation or dephosphorylation steps of proteins (Kowarsch [2011]), which can be imagined like switching the proteins on or off. This whole second mechanism is also called **cellular signaling pathway** or simply **signaling pathway**.

A very important example for such a signaling pathway is the JAK/STAT signaling pathway, which will also be examined in more detail in a later chapter of this thesis. JAK/STAT is an important pathway for gene regulation, which is why it is much investigated and of major scientific interest (Aaronson & Horvath [2002]; Swameye *et al.* [2003]). Disruption of the pathway has been associated with diseases such as leukemia, bronchial asthma or cancer (Igaz *et al.* [2001]). The two key players in the pathway are the kinase JAK (Janus kinase), which has four variants in mammals, and the transcription factor STAT (Signal Transducer and Activator of Transcription), which has seven variants in mammals. The pathway can be triggered by about fifty signaling molecules of different types such as cytokines, growth factors or hormones. Important examples include the epidermal growth factor (EGF), interferons like INFα, INFβ and INFγ, Interleukin-6 (IL-6), or the hormone erythropoietin, which is known by many under its abbreviation Epo for its use in blood doping.

An example of the JAK/STAT signaling pathway can be seen in Figure 2.1. In our example, Epo regulates erythropoiesis, the production of red blood cells. Epo binds to its cognate receptor, which leads to rapid activation of JAK2 phosphorylation followed by phosphorylation of the latent transcription factor STAT5. The quantitative link between the integral STAT5 response in the nucleus and survival of erythroid progenitor cells has recently been elucidated (Bachmann *et al.* [2011]). The broad dynamical range

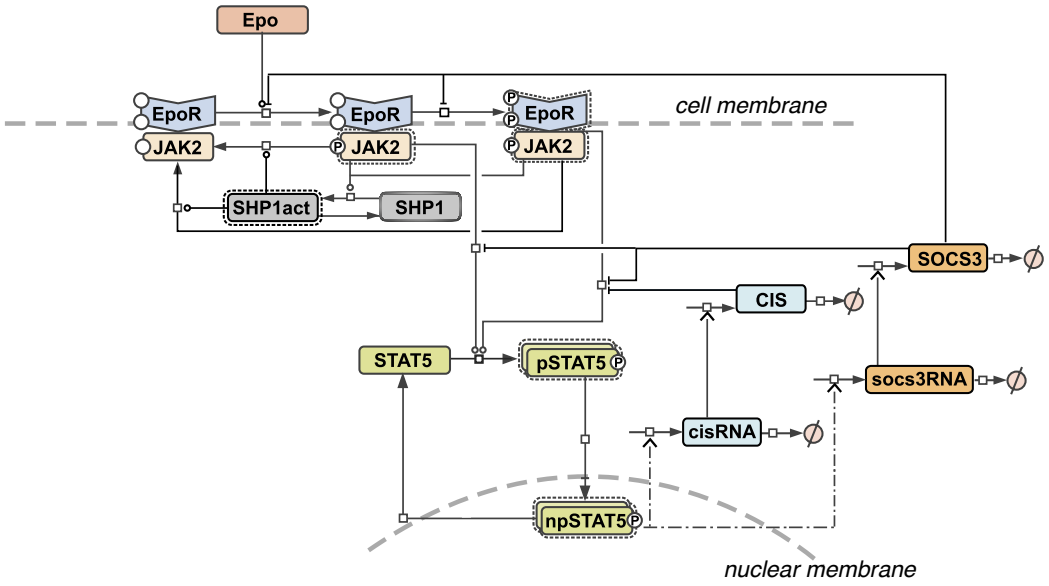


Figure 2.1: Dynamical model of the Epo induced JAK2/STAT5 signal transduction pathway, adopted from Bachmann *et al.* [2011]. The hormone erythropoietin (Epo) binds to its membrane receptor (EpoR) and subsequently leads to receptor phosphorylation (pEpoR) and to phosphorylation of its associated Janus kinase (JAK2, pJAK2). Receptor phosphorylation is balanced by activation of a phosphatase (SHP1, SHP1act). Active EpoR/JAK2 complexes lead to phosphorylation of the Signal Transducer and Activator of Transcription (STAT5, pSTAT5) that transmits the signal to the nucleus (npSTAT5). In the nucleus, STAT5 leads to target gene expression that induces pro-survival signals and self-regulating negative feedbacks. In this case, two regulator proteins and their respective mRNAs are involved, Suppressor Of Cytokine Signaling (SOCS3) and the Cytokine-Inducible SH2-containing protein (CIS).

of Epo concentrations up to 1000-fold in vivo (Becker *et al.* [2010]) requires a stringent regulatory system. This model will be examined in more detail in Chapter 8.

2.1.3 Hematopoietic stem cells

Signaling between cells can lead for example to cell division. Every living organism needs a means to produce new cells, e.g. to replace dead cells, to keep the correct balance between different cell types and to react to injuries. The production of new cells is mediated by a special type of cells, the **stem cells** (Becker *et al.* [1963]). They can be characterized through two main properties: (i) they are pluripotent, i.e. they have the ability to become different types of specialized cells and (ii) they can self-renew indefinitely and thus stay in their pluripotent state even after cell division. How and why stem cells do or do not differentiate, i.e. become more specialized cells, is however still poorly understood.

One very interesting subtype of stem cells are hematopoietic stem cells (HSCs) responsible for replenishing all necessary types of blood cells through the process of hematopoiesis (Orkin & Zon [2008]). The process of differentiating HSCs is thought to be a lineage tree, where the HSCs on top of the tree differentiate through several intermediate cell types to finally give rise to specialized blood cells such as erythrocytes or platelets, as can be seen in Figure 2.2.

At the end of this lineage tree, one can find erythropoiesis, the generation of red blood cells, the erythrocytes. As introduced in the previous section, this process is heavily regulated in cells via the JAK/STAT signaling pathway. This will be discussed further in Chapter 8.

When cells differentiate, they become more and more specialized and gradually lose their ability to self-renew. The exact mechanisms of lineage decisions are not fully understood. It is known that cellular decisions can be triggered by external stimuli or intrinsic factors (Rieger & Schroeder [2007]). Transcription factors specific to each lineage play an important role in the differentiation process through auto-activation and mutual inhibition. For the decision between granulocyte/macrophage progenitors (GMPs) and megakaryocyte/erythrocyte progenitors (MEPs), the transcription factors PU.1 and Gata1 have been suggested as key players (Graf & Enver [2009]). For both

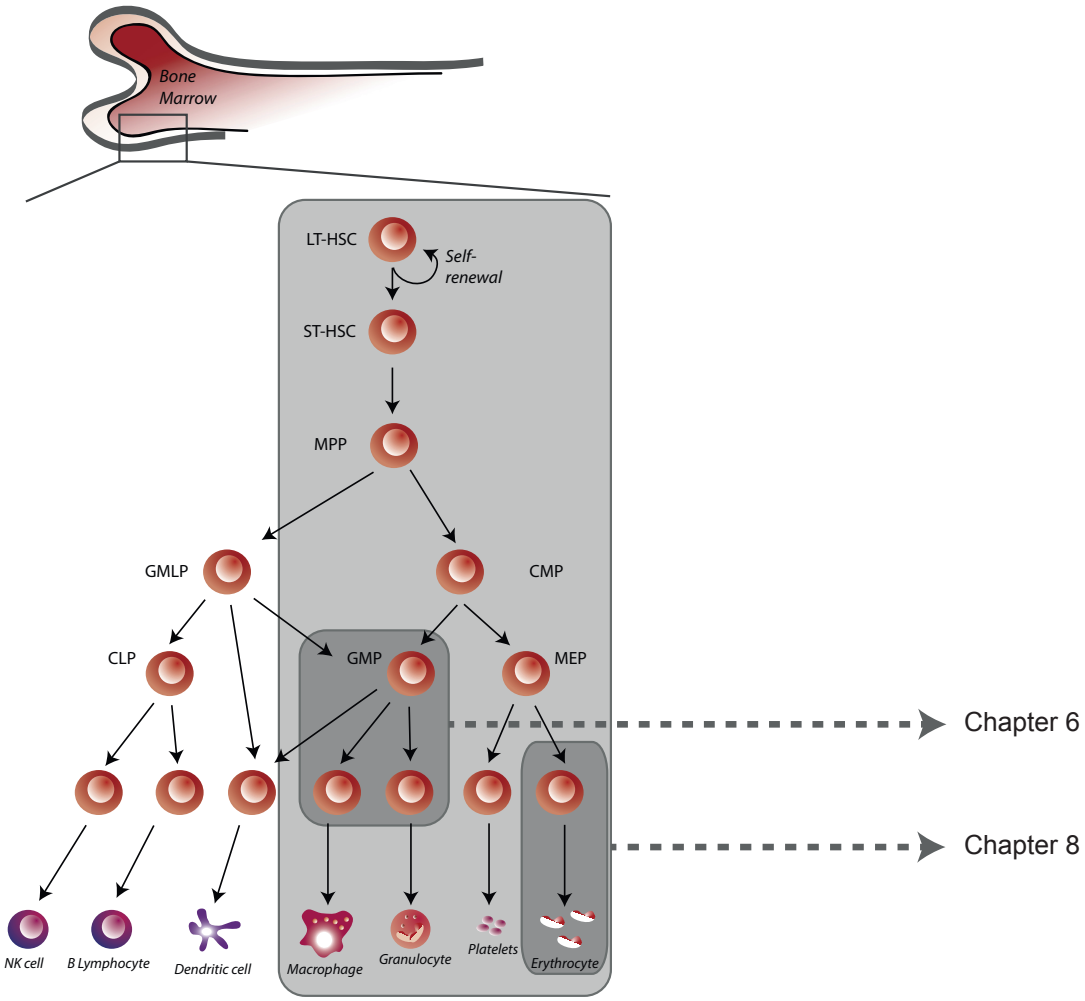


Figure 2.2: Differentiation tree of hematopoietic stem cells, adopted from Rossi *et al.* [2012] and Schwarzfischer [2013]. Starting with HSCs, the cells start to differentiate into more specialized cell types and lose their ability to self-renew in the process. In Chapter 6 we focus on the granulocyte/macrophage progenitor (GMP) cells, while in Chapter 8 the production of erythrocytes is studied.

transcription factors the ability to inhibit the opposing player, transcriptional self-activation and the ability to shut down opposing downstream lineage-specific genes are known (Nerlov *et al.* [2000]; Zhang *et al.* [1999, 2000]). As also pointed out in Schwarzfischer *et al.* [2014], the data on this differentiation is contradictory although the system is heavily studied (Foster *et al.* [2009]). For this reason, continuous single-cell investigations are crucial for gaining insight into the complex underlying biological processes and for dispersing the existing discrepancies.

2.1.4 Time-lapse microscopy for molecular biology

As advocated in the previous section, single-cell data is a highly desirable data type for the study of many systems, not only hematopoietic stem cells. Recent advances in microscopy technology as well as the development of sophisticated software for curating the raw data have lead to the availability of time-lapse single-cell microscopy data. While both microscopy and photography are century-old techniques, their combination for biological processes is still a quite recent development. A milestone for this was the discovery of the green fluorescent protein (GFP), whose emission of green light upon stimulation with ultra violet light can be detected with fluorescence microscopy (Shimomura *et al.* [1962]). This has paved the way for determining protein locations and protein abundances. Full lineages of cells can nowadays be tracked through continuous filming of cells. However, for this new hard- and software is necessary, since huge amounts of data are generated which cannot be manually curated. This has lead to the development of two tools, TTT and QTFy, for efficient quantification of single-cell time-lapse microscopy data (Schwarzfischer *et al.* [2014]). The typical work flow of these tools can be seen in Figure 2.3. For details on the tools and the methods underlying the image quantification, please refer to Schwarzfischer *et al.* [2014]. Most importantly, the individual cells have to be tracked and the fluorescence intensity has to be normalized, segmented and quantified.

Quantitative time-resolved data obtained through these tools is then the basis for single-cell inference of parameters in a dynamical system representing the behavior of the single cell. More specifically, the dynamical properties of the transcription factor PU.1 in GMP cells will be elucidated in Chapter 6. Under special experimental setups, also fully automated quantification might be possible, such as for example in Halter *et al.*

[2007]. There GFP was measured in fibroblast cells, which will also be analyzed in Chapter 6.

2.2 Basic notions and notation

We will now only very briefly introduce the required mathematical notation, since we assume that the reader is already more or less familiar with these.

Vectors and matrices will be indicated by bold letters throughout this thesis, while regular font letters with subscript indices denote their elements. We will display Markov chains as $\{\boldsymbol{X}^{(j)}\}_{j\in I}$ for some index set I. Here the superscript (j) denotes the j^{th} element of the Markov chain. A superscript i denotes a model index of a parameter vector. For example, we will later use the notation $\theta_s^{i,(j)}$, this is the s-th element of the vector $\boldsymbol{\theta}$, belonging to the model with index i, and the j^{th} element of the according Markov chain. Usually, a dimensionality will be written e.g. as $\mathbb{R}^{\mathfrak{d}_u}$ for the dimension of u.

2.2.1 Probability distributions

The theory of probabilities is very rich and delves deep for example also into measure theory. We restrict ourselves to only introducing the necessary notations that will be used later on. The necessary background knowledge, such as the definition of a probability density function, can be found e.g. in Grimmett & Stirzaker [2001].

We will now briefly introduce the most important distributions that we will use in the course of this thesis, which are the uniform distribution, the normal distribution, the lognormal distribution, the gamma distribution and the triangular distribution. While the uniform distribution finds use mostly as prior distribution for the parameters that we want to infer, normal, lognormal and gamma distributions play an important role both as priors and for the error model of our inference. The triangular distribution appears as prior distribution in one application example. The normal distribution will also feature prominently in the Metropolis-Hastings algorithm to be introduced in Chapter 4.1. An illustration of the five distributions can be seen in Figure 2.4. As the most simple and basic distribution, we will begin with the uniform distribution.

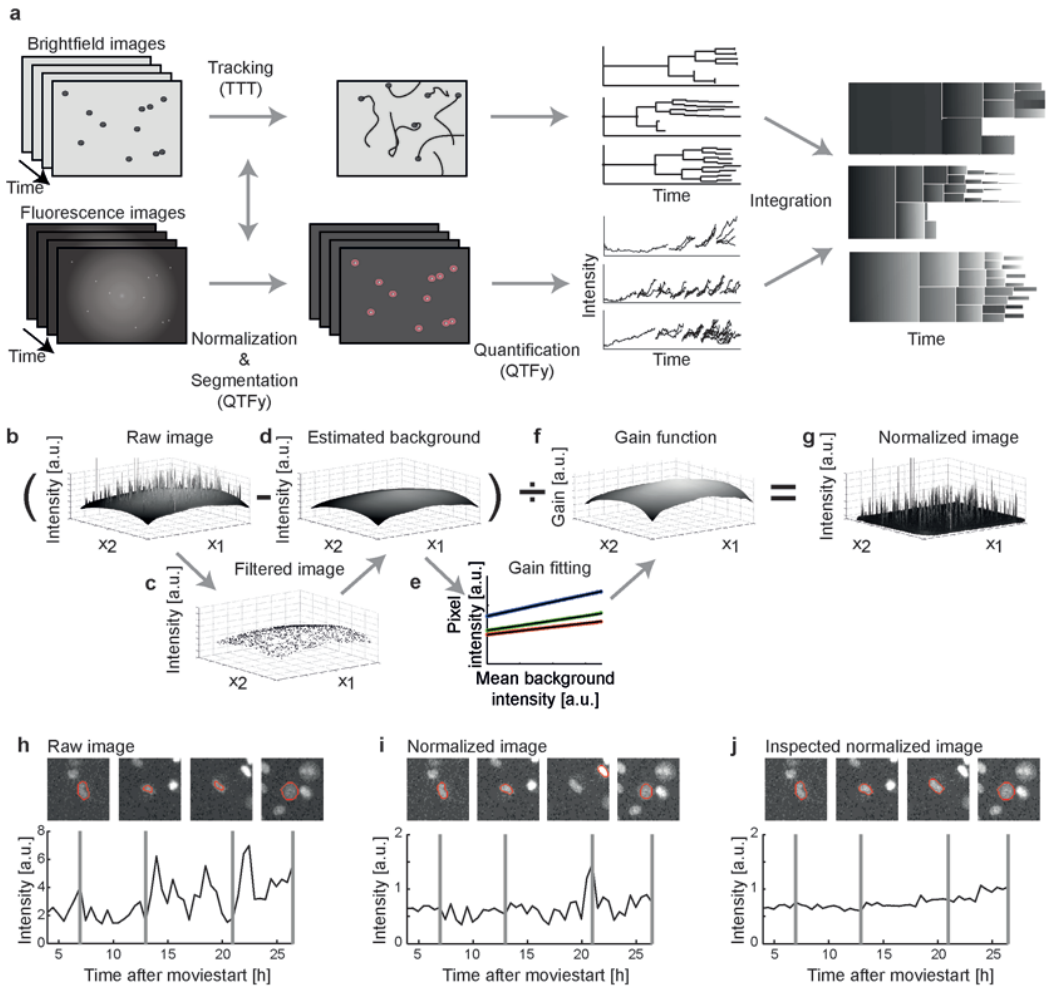


Figure 2.3: Work flow for time-lapse microscopy data collection, adopted from Schwarzfischer [2013]; Schwarzfischer *et al.* [2014]. The most essential parts of the work flow are the tracking of the individual cells and then the normalization, segmentation and quantification of the fluorescence intensity.

Example 2.1 (Uniform distribution). Let $\mathbb{1}_B$ be the **indicator function** on a non-empty measurable set $B \subset \mathbb{R}^n$, i.e.

$$\begin{aligned} \mathbb{1}_B : \mathbb{R}^n &\to \mathbb{R} \\ \boldsymbol{x} &\mapsto \begin{cases} 1, \text{ if } \boldsymbol{x} \in B \\ 0, \text{ otherwise.} \end{cases} \end{aligned}$$

Let now λ^n be the n-dimensional Lebesgue measure. If the density function

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{\lambda^n(B)} \mathbb{1}_B(\boldsymbol{x})$$

exists for $\boldsymbol{x} \in \mathbb{R}^n$ and a random vector $\boldsymbol{X} : \Omega \longrightarrow \mathbb{R}^n$, we call $\boldsymbol{X}$ **uniformly distributed** on B. We use the notation $\boldsymbol{X} \sim \mathcal{U}[B]$. Examples for this distribution can be seen in Figure 2.4(a).

Example 2.2 (Normal distribution). A random vector $\boldsymbol{X} : \Omega \longrightarrow \mathbb{R}^n$ is called **normally (or Gaussian) distributed**, if the density function $f_{\boldsymbol{X}}(\boldsymbol{x})$ exists for $\boldsymbol{x} \in \mathbb{R}^n$ and has the form

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

for a vector $\boldsymbol{\mu} \in \mathbb{R}^n$ and a positive semidefinite symmetric matrix $\boldsymbol{\Sigma} \in \mathrm{Mat}(n \times n, \mathbb{R})$. We use the notation $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We call $\boldsymbol{\mu}$ the **mean** and $\boldsymbol{\Sigma}$ the **covariance matrix** of the distribution. In the univariate case $n = 1$, the density function simplifies to

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

for the **standard deviation** $\sigma > 0$. We write $X \sim \mathcal{N}(\mu, \sigma^2)$. Later, we will also use the notation $\phi(x; \mu, \sigma)$ for the probability density function of the univariate normal distribution evaluated at x with mean μ and standard deviation σ. Examples for the univariate normal distribution can be seen in Figure 2.4(b).

Example 2.3 (Lognormal distribution). The univariate **lognormal distribution** for a random variable X is defined by its density function for $x \in \mathbb{R}$, $\mu \in \mathbb{R}$ and $\sigma > 0$, if it exists, by

$$f_X(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\log(x)-\mu)^2}{2\sigma^2}\right) \mathbb{1}_{(0,\infty)}(x),$$

where $\mathbb{1}_{(0,\infty)}(x)$ denotes the indicator function on $(0, \infty)$ as defined above. We write $X \sim \mathcal{LN}(\mu, \sigma^2)$. Figure 2.4(c) shows examples for this distribution.

Example 2.4 (Gamma distribution). A one dimensional random variable X is called **gamma distributed**, if its density $f_X(x)$ exists and is given by

$$f_X(x) = \frac{1}{\Gamma(k)\lambda^k} x^{k-1} \exp\left(-\frac{x}{\lambda}\right) \mathbb{1}_{(0,\infty)}(x)$$

for $x \in \mathbb{R}$, the **shape parameter** $k > 0$ and **scale parameter** $\lambda > 0$, where $\mathbb{1}_{(0,\infty)}(x)$ is again the indicator function on $(0, \infty)$ and

$$\Gamma(k) := \int_0^\infty t^{k-1} \exp(-t)\, \mathrm{d}t$$

is the gamma function. We write $X \sim \Gamma(k, \lambda)$. Later, we will also use the notation $\phi^\Gamma(x; k, \lambda)$ for the probability density function of the univariate gamma distribution evaluated at x with shape k and scale λ. The typical shapes of this distribution can also be seen in Figure 2.4(d).

Example 2.5 (Triangular distribution). The univariate **triangular distribution** for a random variable X is defined by its density function for $x \in \mathbb{R}$, $a \in \mathbb{R}, b \in \mathbb{R}$ and $c \in \mathbb{R}$ for $a < b < c$, if it exists, by

$$f_X(x) = \begin{cases} 0 & \text{for } x < a \\ \frac{2(x-a)}{(c-a)(b-a)} & \text{for } a \leq x \leq b \\ \frac{2(c-x)}{(c-a)(c-b)} & \text{for } b < x \leq c \\ 0 & \text{for } c < x \end{cases}$$

where a and c are the lower and upper bound, respectively, and b is the mode of the distribution. We write $X \sim \mathcal{T}(a, b, c)$. Figure 2.4(e) shows that the density of this distribution has indeed a triangular shape.

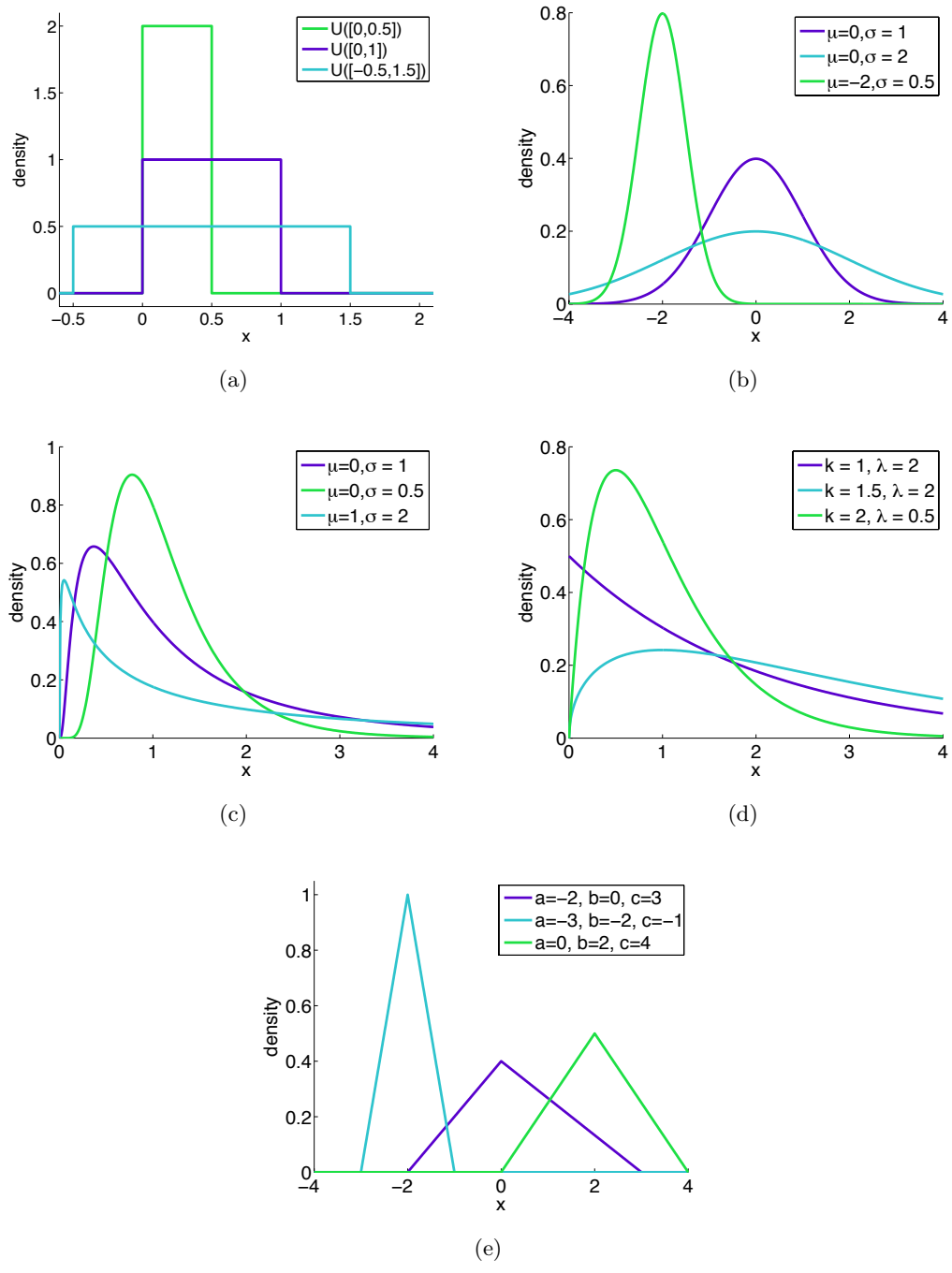


Figure 2.4: Univariate distributions. (a) The uniform distribution. (b) The normal distribution. (c) The lognormal distribution. (d) The gamma distribution. (e) The triangular distribution. (a-e) Shapes of the densities for different parameter values.

2.2.2 Markov chains

Markov chains yield the necessary theoretical background for the Markov chain Monte Carlo algorithms presented in Chapter 4. More detailed introductions can be found in e.g. in Meyn & Tweedie [1996]; Robert & Casella [2004]. Throughout this section, we will let $(\Omega, \mathcal{F}, P)$ denote a probability space. Furthermore, all random vectors are functions $\boldsymbol{X} : \Omega \longrightarrow \Omega'$ onto a measurable space $(\Omega', \mathcal{F}') \subseteq (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$. Without loss of generality let the density function for each random vector exist and be positive for any realization $x \in \Omega'$ considered here. Furthermore, we consider only state spaces $\Omega' \subseteq \mathbb{R}^n$ that are thus continuous.

First, there is of course a precise mathematical definition of a Markov chain:

Definition 2.1 (Markov chain)**.** Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\{\boldsymbol{X}^{(t)}\}_{t \in I}$ a stochastic process. $\{\boldsymbol{X}^{(t)}\}_{t \in I}$ with values in the state space Ω' is then a **Markov chain**, if $I = \mathbb{N}_0$ and for any measurable set $A \subseteq \Omega'$, any index $T \in I \setminus \{0, 1\}$ and any realization $\boldsymbol{x}^{(0)}, \ldots, \boldsymbol{x}^{(T)}$ of $\boldsymbol{X}^{(0)}, \ldots, \boldsymbol{X}^{(T)}$, the random vector $\boldsymbol{X}^{(T+1)}$ does not depend on $\boldsymbol{x}^{(0)}, \ldots, \boldsymbol{x}^{(T-1)}$. For the distributions, that means that

$$P_{\boldsymbol{X}^{(T+1)}|\boldsymbol{X}^{(0)} \otimes \cdots \otimes \boldsymbol{X}^{(T)}} \left(\boldsymbol{X}^{(T+1)} \in A | \boldsymbol{x}^{(0)}, \ldots, \boldsymbol{x}^{(T)} \right) = P_{\boldsymbol{X}^{(T+1)}|\boldsymbol{X}^{(T)}} \left(\boldsymbol{X}^{(T+1)} \in A | \boldsymbol{x}^{(T)} \right)$$

Since $I = \mathbb{N}_0$, a Markov chain is an entity with discrete indices, which attains a value from an in our case continuous state space for each index. The important thing about such a first-order Markov chain is that the value at index $T + 1$ only depends on the value at T, if $\boldsymbol{X}^{(T)}$ is known, thus it is in a sense memoryless. This is also called the Markov property.

Markov chains can have many desirable properties. We will now give only a brief mathematical introduction that we hope is easy to understand for those not already familiar with the concepts. For those interested, we refer to the literature for more detailed introductions, e.g. Robert & Casella [2004].

When the state of the system changes, this is called a transition. It is associated with a probability, which is thus called transition probability. If the state space Ω' were finite and discrete, we could write this transition probability as $p_{ij} = P\{\boldsymbol{X}^{(T)} = j | \boldsymbol{X}^{(T-1)} = i\}$ (Meyn & Tweedie [1996]). However we here deal with continuous state spaces, thus

we use the notion of a transition kernel, which will be introduced in a moment. But first we define:

Definition 2.2 (Time-homogeneous Markov chain)**.** Let $\{\boldsymbol{X}^{(t)}\}_{t\in\mathbb{N}_0}$ with values in Ω' be a Markov chain. We call $\{\boldsymbol{X}^{(t)}\}_{t\in\mathbb{N}_0}$ **time-homogeneous**, if for all all $s \in I$ and $(s+1)$-uplets $t_0 \leq t_1 \leq t_s$ the distributions of $\boldsymbol{X}^{(t_1-t_0)}, \ldots, \boldsymbol{X}^{(t_s-t_0)}$ given a realization $\boldsymbol{x}^{(t_0)}$ and $\boldsymbol{X}^{(t_1)}, \ldots, \boldsymbol{X}^{(t_s)}$ given a realization $\boldsymbol{x}^{(0)}$ are equal.

In other words, the joint distribution is time independent, time here refers to the chain index (t). In the following, we will always assume that **all Markov chains considered are time-homogeneous**.

The definition of Markov chains also makes a simplification of notation possible when considering the joint distribution $P_{\boldsymbol{X}^{(0)}\otimes\cdots\otimes\boldsymbol{X}^{(t)}}$ of the random vectors $\boldsymbol{X}^{(0)}, \ldots, \boldsymbol{X}^{(t)}$.

Definition 2.3 (Transition kernel)**.** Let $\{\boldsymbol{X}^{(t)}\}_{t\in\mathbb{N}_0}$ be a Markov chain on a probability space $(\Omega, \mathcal{F}, P)$ with values in Ω'. For a measurable set $A \in \Omega'$, the distribution

$$k(A|\boldsymbol{x}) := P_{\boldsymbol{X}^{(t+1)}|\boldsymbol{X}^{(t)}}\left(\boldsymbol{X}^{(t+1)} \in A|\boldsymbol{X}^{(t)} = \boldsymbol{x}\right) = \int_A P_{\boldsymbol{X}^{(t+1)}|\boldsymbol{X}^{(t)}}(\mathrm{d}\boldsymbol{y}|\boldsymbol{x})$$

is called (time homogeneous) **transition kernel** (or **transition probability**) **from** $\boldsymbol{x} \in \Omega'$ **to** $A \subseteq \Omega'$.

It can be seen from the definition already that the transition kernel is a time independent function

$$k : \mathcal{F}' \times \Omega' \longrightarrow [0, 1],$$

with the properties (see also Robert & Casella [2004])

(i) $k(\cdot|\boldsymbol{x})$ is a probability measure for all $\boldsymbol{x} \in \Omega'$

(ii) $k(A|\cdot)$ is measurable for all $A \in \mathcal{F}'$.

Moreover, the transition kernel can also be written down in a different form via a function $p : \Omega' \times \Omega' \longrightarrow [0, \infty)$ as

$$k(\mathrm{d}\boldsymbol{y}|\boldsymbol{x}) = p(\boldsymbol{y}|\boldsymbol{x})\,\mathrm{d}\boldsymbol{y} + r(\boldsymbol{x})\mathbb{1}_{\boldsymbol{x}}(\mathrm{d}\boldsymbol{y}) \tag{2.1}$$

As before, $\mathbb{1}_{\boldsymbol{x}}(\mathrm{d}\boldsymbol{y})$ is the indicator function. Per definition, it also follows that $p(\boldsymbol{x}|\boldsymbol{x}) = 0$ and that $r(\boldsymbol{x}) = 1 - \int_{\Omega'} p(\boldsymbol{y}|\boldsymbol{x})\,\mathrm{d}\boldsymbol{y}$.

The aim of most Markov chain Monte Carlo methods as presented in Chapter 4 is the inference of a distribution π by approximating it with samples from a Markov chain $\{\boldsymbol{X}^{(t)}\}_{t\in\mathbb{N}_0}$. Like any iterative method, the key question is that of convergence, i.e. making sure the Markov chain converges to π, irrespective of its starting point. For this, we have to define what is meant by convergence in this context, but first we will define several desirable properties of Markov chains that will be necessary for achieving said convergence.

Definition 2.4 (Invariant/stationary distribution)**.** A distribution π is called **invariant** or **stationary** for the transition kernel $k(\cdot|\cdot)$ if

$$\pi(A) = \int_{\Omega'} k(A|\boldsymbol{x})\pi(\mathrm{d}\boldsymbol{x}) \tag{2.2}$$

$$= \int_{\Omega'} k(A|\boldsymbol{x})\pi_d(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}, \qquad \forall A \in \mathcal{F}' \tag{2.3}$$

where π_d is the probability density function to π with respect to the Lebesgue measure.

Definition 2.5 (Reversible Markov chain)**.** We call a stationary Markov chain $\{\boldsymbol{X}^{(t)}\}_{t\in\mathbb{N}_0}$ **reversible** if for $A \in \mathcal{F}'$

$$P(\boldsymbol{X}^{(t+1)} \in A|\boldsymbol{X}^{(t+2)} = \boldsymbol{x}) = P(\boldsymbol{X}^{(t+1)} \in A|\boldsymbol{X}^{(t)} = \boldsymbol{x}). \tag{2.4}$$

True to its name, reversibility for a Markov chain means that the direction of the evolution does not influence the dynamics of the chain. An important sufficient condition for stationarity and reversibility is given by the following definition.

Definition 2.6 (Detailed balance condition)**.** Let $\{\boldsymbol{X}^{(t)}\}_{t\in\mathbb{N}_0}$ be a Markov chain with transition kernel $k(\mathrm{d}\boldsymbol{y}|\boldsymbol{x}) = p(\boldsymbol{y}|\boldsymbol{x})\,\mathrm{d}\boldsymbol{y} + r(\boldsymbol{x})\mathbb{1}_{\boldsymbol{x}}(\mathrm{d}\boldsymbol{y})$ as specified in Equation (2.1). The Markov chain is said to fulfill the **detailed balance condition**, if there exists a probability density function π_d such that

$$p(\boldsymbol{x}|\boldsymbol{y})\pi_d(\boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{x})\pi_d(\boldsymbol{x}). \tag{2.5}$$

With these three definitions, we have all that is necessary for the following theorem:

Theorem 2.1. *Suppose the detailed balance condition holds for a Markov chain with transition kernel k and density function π_d. Then*

(i) the associated distribution π is invariant with respect to k and

(ii) the Markov chain is reversible.

The proof can be found in Robert & Casella [2004].

While the detailed balance condition is sufficient, it is not necessary for the existence of a stationary distribution π. Yet it is easy to check and thus a very popular assumption for many MCMC algorithms, which are the main application for Markov chains in this thesis. Even if detailed balance is fulfilled, and we thus know that a reversible stationary distribution π exists, this distribution might still be non-unique. Since this outcome is not desirable, we include a new concept:

Definition 2.7 (Equilibrium distribution)**.** Suppose every Markov chain associated with the transition kernel k is converging to the same invariant distribution π, independent of the starting value $\boldsymbol{x}^{(0)} \in \Omega'$. Then we call π an **equilibrium distribution**.

For an m**-step transition kernel** $k^m(A|\boldsymbol{x}) = \int_{\Omega'} k^{m-1}(A|\boldsymbol{y})k(\mathrm{d}\boldsymbol{y}|\boldsymbol{x})$ for the transition from $\boldsymbol{x}$ to A in $m \in \mathbb{N}$ steps this means

$$\lim_{m\to\infty} k^m(A|\boldsymbol{x}^{(0)}) = \pi(A)$$

point-wise for π-almost all $\boldsymbol{x}^{(0)} \in \Omega'$ (Tierney [1994]). Of course, $k^1(A|\boldsymbol{x}) := k(A|\boldsymbol{x})$. To verify if π is an equilibrium distribution, we will need the concepts of irreducibility and recurrence.

Definition 2.8 (π-irreducible Markov chain)**.** Let $\{\boldsymbol{X}^{(t)}\}_{t\in\mathbb{N}_0}$ be a Markov chain with transition kernel k. It is called π**-irreducible** for a σ-finite π, if for any $\boldsymbol{x} \in \Omega'$ and $A \in \mathcal{F}'$ with $\pi(A) > 0$ there exists an $m \in \mathbb{N}$ such that

$$k^m(A|\boldsymbol{x}) > 0,$$

where $k^m(A|\boldsymbol{x})$ is the associated m-step transition kernel.

This means that any state in Ω' can be reached by the Markov chain from $\boldsymbol{x} \in \Omega'$ within a finite number of steps. If $m = 1$, the chain is called **strongly** π**-irreducible**.

Definition 2.9 (Periodic Markov chain)**.** Let a Markov chain $\{\boldsymbol{X}^{(t)}\}_{t\in\mathbb{N}_0}$ with transition kernel k be π-irreducible. The chain is called **periodic**, if for some integer

$s \geq 2$ there exists a sequence $(A_0, A_1, \ldots, A_{s-1})$ of pairwise disjoint nonempty subsets $A_i \in \mathcal{F}'$, such that for all $i = 0, \ldots, s-1$ and all $\boldsymbol{x} \in A_i$

$$k(A_j|\boldsymbol{x}) = 1 \qquad \text{for } j = i+1 \mod s.$$

If a chain is not periodic, we call it **aperiodic**. In layman's terms, aperiodic Markov chains do not cycle deterministically.

Definition 2.10 (Harris recurrent Markov chain)**.** Let $c_A^{(t)} := \left| \left\{ \boldsymbol{x}^{(s)} \in A | 0 \leq s \leq t \right\} \right|$ be the number of visits to some subset $A \in \mathcal{F}'$ up to index t, starting at some $\boldsymbol{x} \in \Omega'$. Let furthermore $P_{\boldsymbol{x}}(A)$ reflect the probability that $c_A^{(t)} \to \infty$ as $t \to \infty$. A Markov chain is **Harris recurrent**, if there exists an invariant distribution π, such that for every $A \in \mathcal{F}'$ with $\pi(A) > 0$

$$P_{\boldsymbol{x}}(A) = 1 \qquad \forall \boldsymbol{x} \in \Omega'.$$

For every π-irreducible Harris recurrent Markov chain, there exists an invariant measure ν on Ω', see Nummelin [2004]; Schmidl [2012]; Tierney [1994] for more details. This measure ν is unique up to a multiplicative constant. If $\nu(\Omega') < \infty$, we call the Markov chain **positive Harris recurrent**. The notion of Harris recurrence is slightly stronger than the more common notion of a **recurrent** Markov chain. However, we use it here as it makes the following two theorems on the convergence of Markov chains more universally valid. We will focus on these two results, more theoretical considerations and the proofs of the two theorems can be found elsewhere in the literature (Revuz [1984]; Schmidl [2012]; Sethuraman *et al.* [1992]; Tierney [1994]).

Theorem 2.2. *Let $\{\boldsymbol{X}^{(t)}\}_{t \in \mathbb{N}_0}$ be a π-irreducible, aperiodic and Harris recurrent Markov chain with transition kernel k and stationary distribution π. Then*

(i) the Markov chain is positive Harris recurrent

(ii) π is the unique equilibrium distribution

(iii) k is **ergodic** *for π, i.e. $\{\boldsymbol{X}^{(t)}\}_{t \in \mathbb{N}_0}$ converges regardless of its starting value $\boldsymbol{x}^{(0)} \in \Omega'$.*

This theorem yields a useful criterion for showing that an MCMC method is valid: it should have an existing stationary distribution π, as well as π-irreducibility, aperiodicity

and Harris recurrence. The following theorem can be derived as a consequence:

Theorem 2.3. *Let $\{\boldsymbol{X}^{(t)}\}_{t\in\mathbb{N}_0}$ be a π-irreducible, aperiodic and Harris recurrent Markov chain with stationary distribution π. Let furthermore $f : \Omega' \to \mathbb{R}$ be π-integrable with $\int_{\Omega'} |f(\boldsymbol{x})|\pi(\mathrm{d}\boldsymbol{x}) < \infty$. It then follows for the sample mean that for a realization $\{\boldsymbol{x}^{(t)}\}_{t\in\mathbb{N}_0}$*

$$\overline{f}_m = \frac{1}{m+1}\sum_{t=0}^{m} f(\boldsymbol{x}^{(t)}) \to \int_{\Omega'} f(\boldsymbol{x})\pi(\mathrm{d}\boldsymbol{x}) = \mathbb{E}_\pi[f(\Omega')] \qquad \textit{almost surely as } m \to \infty.$$

For a nice example on how to verify the required conditions in a discrete state space, see Schmidl [2012].

In probability theory, the "speed of convergence" is measured as the mixing time of a Markov chain, which can be seen as the time until the Markov chain is "close" to its steady state distribution. Time in this context means the discrete indices t of the chain. Closeness to the steady state distribution can be measured by an appropriate distribution distance, most commonly the total variation distance, see Levin *et al.* [2009] for a thorough introduction.

A practical measure for the quality of a MCMC algorithm is the inefficiency factor. This is the number $\hat{r}$ that samples in a stationary Markov chain realization $\{\boldsymbol{x}^{(t)}\}_{t\in 1,\dots,T}$ have to be apart in order to be considered independent. More commonly used is the effective sampling size (ESS) defined as $ESS = T/\hat{r}$, see also Schmidl [2012] and references therein.

2.2.3 Numerical integration

Numerical integration is a branch of numerical mathematics in which definite integrals are approximated, mostly in cases where they are analytically intractable. A thorough introduction is presented in Atkinson & Han [1985]. We will need these methods, since probability theory provides us with just these types of integrals. In the one-dimensional case, the aim of numerical integration, also called **quadrature**, is to compute the definite Riemann integral

$$I(f) = \int_a^b f(x)\,\mathrm{d}x \tag{2.6}$$

of a function $f : \mathbb{R} \to \mathbb{R}, x \mapsto f(x)$ on the interval $[a, b]$. For a continuous f and an antiderivative F of f, this would mean that $I(f) = F(b) - F(a)$, from the fundamental

theorem of calculus. When f is continuous, an antiderivative of course exists, however it is often not easily computable.

The basic idea of quadrature is now to replace f with a function as similar as possible to f, but whose integral is analytically tractable. We will concentrate on the two most simple approximations, as these are the ones needed later. The most simple approximation imaginable is to approximate f with a linear polynomial:

$$P_1(x) = \frac{(b-x)f(a) + (x-a)f(b)}{b-a}. \tag{2.7}$$

Integrating this yields the approximation

$$J_1(f) = \frac{b-a}{2}(f(a) + f(b)) \tag{2.8}$$

This is called the **trapezoidal rule** and is only a good approximation if the function is almost linear. It can be shown that the approximation error $R_1(f)$ is proportional to the second derivative of the function (if it exists):

$$R_1(f) = I(f) - J_1(f) = \frac{-(b-a)^3}{12} f''(\xi) \tag{2.9}$$

for some $\xi \in [a, b]$. This confirms that the approximation order of the trapezoidal rule is one, i.e. it is exact for linear polynomials, since their second derivative is zero. Of course this can be refined by subdividing the interval $[a, b]$ into smaller intervals and applying the trapezoidal rule to each.

The "next better" approximation is to approximate f with a quadratic polynomial instead of a linear one, interpolating f at a, b and $m = (a + b)/2$:

$$P_2(x) = \frac{(x-m)(x-b)}{(a-m)(a-b)} f(a) + \frac{(x-a)(x-b)}{(m-a)(m-b)} f(m) + \frac{(x-a)(x-m)}{(b-a)(b-m)} f(b) \tag{2.10}$$

After integration, this is **Simpson's rule**

$$J_2(f) = \frac{b-a}{6}\left(f(a) + 4f(m) + f(b)\right) \tag{2.11}$$

Also for this the approximation error can be computed. We set $h = \frac{b-a}{2}$ and find that

$$R_2(f) = I(f) - J_2(x) = -h^5 f^{(4)}(\xi^*)/90 \tag{2.12}$$

for some $\xi^* \in [a, b]$. As the fourth derivative is zero at all points for a cubic polynomial, Simpson's rule is exact for those. Thus the approximation order of Simpson's rule is three, two higher than the one for the trapezoidal rule.

2.3 Dynamical systems

Ordinary differential equations have become a popular tool for modeling dynamical systems. They are able to handle quantitative data in contrast to Boolean models and easier to deal with than stochastic differential equations and very useful for modeling biological processes as introduced in the beginning of this chapter.

2.3.1 ODE and DDE systems for biological processes

A lot of scientific effort has gone into the understanding and modeling of molecular processes. Here the interactions of chemical species $Z_1, \ldots, Z_{\mathfrak{D}}$ are studied. If we consider $[Z_d]$ molecules of species Z_d, the concentration of Z_d in the reaction volume Ω can be defined as $z_d = \frac{Z_d}{\Omega}$ (McNaught & Wilkinson [2000]). Furthermore, let $\mathbf{z} = (z_1, \ldots, z_{\mathfrak{D}})^\top$ denote the concentration vector of all $\mathfrak{D}$ involved species.

In most of our applications, we want to model the dynamics of reactions involving the chemical species over time. To this end, various approaches exist and have been proven useful. Which approach to apply depends strongly on the biological question at hand.

In this thesis, we solely consider systems fulfilling two assumptions (Gillespie [1992]). First, the system has to be in thermodynamic equilibrium, i.e. well-mixed. This means that we have uniform concentration in space. Furthermore, we require large molecule numbers $\gg 1$ for all involved chemical species. Then the chemical kinetics can be modeled by **ordinary differential equations** (Girolami [2008]; Kholodenko [2006]).

Generally we want to consider reactions $R_j, j = 1, \ldots, \mathfrak{J}$, where some chemical species are transformed to other chemical species with a specific rate. Each reaction is associated with a reaction flux vector $v_j(\mathbf{z})$, giving the instantaneous frequency with which reaction $R_j, j = 1, \ldots, \mathfrak{J}$ occurs given $\mathbf{z} = (z_1, \ldots, z_{\mathfrak{D}})^\top$. A general formulation of such a reaction R_j is

$$\underbrace{\sum_{d=1}^{\mathfrak{D}} \nu_{d,j} Z_d}_{\text{educts}} \overset{\kappa_{+j}}{\underset{\kappa_{-j}}{\rightleftharpoons}} \underbrace{\sum_{d=1}^{\mathfrak{D}} \eta_{d,j} Z_d}_{\text{products}} \tag{2.13}$$

Here, the $\nu_{d,j} \in \mathbb{N}_0$ and the $\eta_{d,j} \in \mathbb{N}_0$ are stoichiometric coefficients for $d = 1, \ldots, \mathfrak{D}, j = 1, \ldots, \mathfrak{J}$ and $\kappa = [\kappa_{+1}, \kappa_{-1}, \ldots, \kappa_{+\mathfrak{J}}, \kappa_{-\mathfrak{J}}] \in \mathbb{R}_+^{2\mathfrak{J}}$ are the reaction parameters or rate constants. We can now formulate the following three definition building on one another:

Definition 2.11 (Mass-action kinetics)**.** Chemical reaction kinetics of reaction R_j follow **mass-action kinetics**, if the flux satisfies the isotropic assumption:

$$v_j(\mathbf{z}) = \kappa_{+j} \prod_{d=1}^{\mathfrak{D}} z_d^{\nu_{d,j}} - \kappa_{-j} \prod_{d=1}^{\mathfrak{D}} z_d^{\eta_{d,j}}. \tag{2.14}$$

The law of mass action was first discussed by Guldberg & Waage [1899]. Basically, it postulates that the rate of a reaction is proportional to the probability of a collision of the reactants. This is turn is proportional to the concentration of reactants to the power of the respective stoichiometric coefficients, which represents the molecularity or number in which a reactant participates in reaction R_j.

Definition 2.12 (Stoichiometric matrix)**.** The entries $S_{d,j}$ of the **stoichiometric matrix** $S \in \mathbb{Z}^{\mathfrak{D}\times\mathfrak{J}}$ are given by

$$S_{d,j} = \eta_{d,j} - \nu_{d,j}. \tag{2.15}$$

Definition 2.13 (Reaction rate equation)**.** The biological system follows a **reaction rate equation**, if its kinetics are given by an ODE of the form

$$\dot{\mathbf{z}} = Sv(\mathbf{z}). \tag{2.16}$$

Example 2.6 (Dimerization). We consider a dimerization reaction

$$2Z_1 \xrightarrow{\xi} Z_2 \tag{2.17}$$

The flux of this single reaction is given by $v(\mathbf{z}) = \xi z_1^2$ for the concentrations z_1, z_2 of the two chemical species Z_1, Z_2. Furthermore, the stoichiometric matrix is given by $S = \begin{pmatrix} -2 \\ 1 \end{pmatrix}$. Assuming mass action kinetics, this leads to the two differential equations

$$\dot{z_1}(t) = -2\xi z_1^2(t) \tag{2.18}$$

$$\dot{z_2}(t) = \xi z_1^2(t) \tag{2.19}$$

These have a unique analytically tractable solution for $\mathbf{z}(0) = (z_{0,1}, z_{0,2})^\top$:

$$z_1(t) = \frac{1}{z_{0,1}^{-1} + 2\xi t} \tag{2.20}$$

$$z_2(t) = z_{0,2} + \frac{\xi t z_{0,1}}{z_{0,1}^{-1} + 2\xi t} \tag{2.21}$$

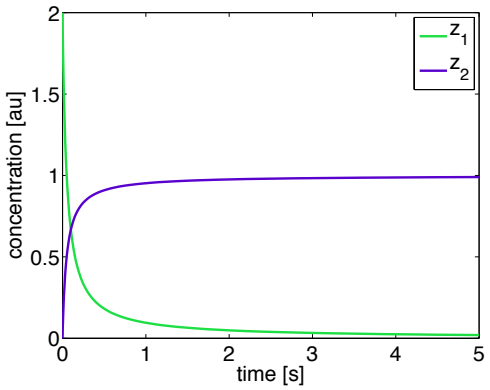


Figure 2.5: Dimerization: analytical solution. For the dimerization in Example 2.6, we show the analytical solutions of the ODE. Already after a few seconds, the concentration for z_1 (green) goes to zero, as it is consumed to produce z_2 (purple), which reaches a steady state of $z_2 = 1$.

A visualization of this can be seen in Figure 2.5. There we show the true analytical solution for a choice of $\xi = 5$ and $\mathbf{z}(0) = (z_{0,1}, z_{0,2})^\top = (2, 0)^\top$. It can be seen that the system nears a steady state where the concentrations don't change much any more already after a few seconds, as all Z_1 molecules are converted to Z_2 molecules.

As reaction rate equations are quite specific, we consider also the following more general basic model of the dynamics of a system:

$$\frac{\mathrm{d}\boldsymbol{z}(t)}{\mathrm{d}t} = g(t, \mathbf{z}(t), \boldsymbol{\xi}, \boldsymbol{u}(t)), \mathbf{z}(0) = \mathbf{z}_0(\boldsymbol{\xi}) \tag{2.22}$$

This of course includes reaction rate equations as defined above, but lends more flexibility. Here we consider time points t in the interval $[0, T]$. This equation relates the concentration vector $\boldsymbol{z}(t)$ to its time derivative via a $\boldsymbol{\xi}$-parameterized, usually nonlinear function g, which we assume to be Lipschitz-continuous. Usually, the explicit dependence of $\boldsymbol{z}(t)$ on the parameter vector $\boldsymbol{\xi} \in \mathbb{R}^{\mathfrak{d}_\xi}$ is dropped. This parameter vector can contain parameters of various types like the reaction rate constants $\kappa_{\pm j}, j = 1, \ldots, \mathfrak{J}$, initial values to Equation (2.22) or other types like scaling factors. This will also be seen in later chapters. The function $\boldsymbol{u}(t)$ represents an external input to the system, e.g. the concentration of a species that is not included in the system, with $\boldsymbol{u}(t) \in \mathbb{R}^{\mathfrak{d}_u}$.

Sometimes it is biochemically more plausible to include time delays into the dynamics of the systems, as some reactions might take a while, for example if an educt first has to be produced. If we want to include the possibility to include time delays for

some processes such as e.g. transcription, we can model this by **delay differential equations** (DDEs). We consider discrete and constant time delays $\tau_1, \ldots, \tau_\rho$. Then we can define a delay differential equation by

$$\frac{\mathrm{d}\boldsymbol{z}(t)}{\mathrm{d}t} = g(t, \mathbf{z}(t), \boldsymbol{\xi}, \boldsymbol{u}(t), \mathbf{z}(t-\tau_1), \ldots, \mathbf{z}(t-\tau_\rho)), t \geq t_0 \tag{2.23}$$

While the restriction to discrete constant time delays still leaves a large and useful class of DDEs, it is furthermore especially useful, since in this case we get a result for the existence: If g satisfies a Lipschitz condition in its dependent variables and is sufficiently smooth in all its variables, the numerical solution to Equation (2.23) exists (Shampine & Thompson [2001]). More mathematically, Baker *et al.* [1995] elaborate on the difficulties that can arise when solving DDEs with the most commonly applied method, the method of steps. The theory of DDEs is very rich and distinct from ODE theory, but for the sake of brevity, we refer interested readers to Smith [2011].

Both ordinary and delay differential equations can be seen as a **dynamical system**. We do not consider other types of dynamical systems, but take the term to mean either of the mentioned differential equation types. For our dynamical systems, it is necessary to assume that the parameters are constant, i.e. time-independent. Furthermore, an implicit assumption for our dynamical systems is that they are well-stirred and external influences such as temperature or osmotic pressure stay constant throughout the modeling process. These restrictions yield for ODEs a well-developed theory of existence, uniqueness and computational complexity. Incorporation of such aspects on the other hand would call for spatial modeling, e.g. in the form of partial differential equations, which are generally computationally more expensive to solve than ODEs. If a system does not fulfill the conditions for the thermodynamic limit, modeling with stochastic differential equations might be called for, e.g. when dealing with only a few molecules of each modeled species. Well-established examples for differential equations in systems biology include the modeling of mRNA synthesis (Goodwin [1963]) or the modeling of cell cycles in *Caulobacter crescentus* (Li *et al.* [2008]) or yeast (Chen *et al.* [2004]).

For the JAK/STAT pathway as mentioned in Section 2.1.2, it was shown in Raia *et al.* [2011] that the number of STAT5 and STAT6 molecules contained in human lymphoma cells is $\sim 2 \cdot 10^5$ each. Hence the assumptions for the thermodynamic limit for the JAK/STAT pathway is well justified. Also for the single-cell example in Chapter 6, it was shown that protein numbers are $> 10^3$ (Schwarzfischer [2013]). We can thus model these systems with ODEs and do not have to consider stochastic differential equations.

For most real-world examples, the system in Equation (2.22) is nonlinear and thus

often not analytically solvable. In order to be able to compare measurement data and the dynamical system, the differential equation thus has to be solved numerically on a computer. For ODEs, this can be done quite efficiently using for example MATLAB's `ode15s` (Shampine & Reichelt [1997]) or SUNDIAL's `CVODEs` (Serban & Hindmarsh [2005]) solvers. For delay differential equations, usually MATLAB's `dde23` (Shampine & Thompson [2001]) solver is the tool of choice. Otherwise, the linear chain trick (Smith [2011]) can be applied to transform delay differential equations to ordinary differential equations. This trick is used to transform the only DDE in this thesis, the one of Chapter 8, into an ODE, which can then more easily be solved. For all further considerations, we will thus restrict ourselves to ODE systems.

2.3.2 Multi-compartmental models

A special case of the first-order ordinary differential equation models introduced in the previous section are **multi-compartmental models**. Here we do not consider separate chemical species, but a finite set of mutually exclusive compartments. Each compartment is assumed to be a homogeneous entity, holding a group of objects unambiguously identifiable with the respective compartment, cf. Jacquez [1985]. Usually, we only consider the transfer of one chemical species of interest between these compartments, in contrast to the reaction rate equations presented previously. All of these compartments are assumed to be well-mixed and homogeneous with constant volume. For example, compartments might represent different parts of the body within which the concentration of the species of interest can be assumed to be equal. An example of this is also the JAK2/STAT5 pathway model in Figure 2.1, which considers the two compartments cytoplasma and nucleus.

Transitions of the chemical species between the compartments is governed by several assumptions:

- Instant homogeneous distribution of the species within the compartment,
- the net flow between the compartments then depends on the density of their objects,
- the volume of the compartment has to stay constant over time, since otherwise it is not sensible to look at the concentration of the species.

Interactions between the compartments are then governed by **transition equations**, which control the exchange of objects between compartments. Multi-compartmental

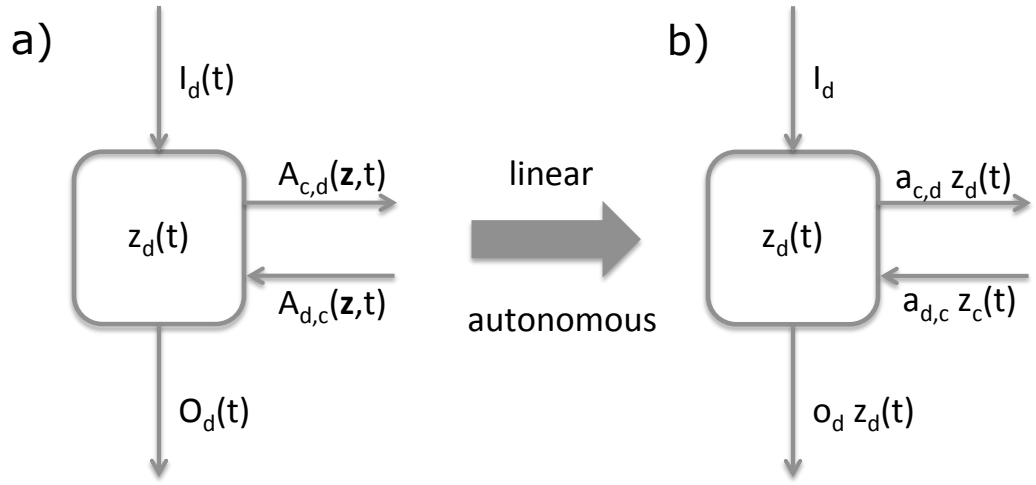


Figure 2.6: Multi-compartmental models. (a) Standard multi-compartmental model with non-negative transfer matrix A, input transfers I and output transfers O for an exemplary compartment with concentration $z_d(t)$. (b) The same multi-compartmental model under the assumption of linearity and autonomy.

models can also be seen as dynamical systems, since their transition equations can be represented by differential equations, in this case a system of first-order ordinary differential equations. In contrast to many biochemical reaction systems, multi-compartmental models are usually **closed systems**, meaning that mass is conserved and there is no external flow of objects into or out of the system.

Then a formal definition of a multi-compartmental system can be given as follows.

Definition 2.14 (Multi-compartmental model)**.** A $\mathfrak{D}$-dimensional **multi-compartmental model** is defined by a nonnegative transfer matrix $A \in \mathcal{C}^r(\mathbb{R}_+^{\mathfrak{D}} \times \mathbb{R}_+; Mat(\mathfrak{D} \times \mathfrak{D}; \mathbb{R}))$, input transfers $I \in \mathcal{C}^r(\mathbb{R}_+; \mathbb{R}_+^{\mathfrak{D}})$ and output transfers $O \in \mathcal{C}^r(\mathbb{R}_+^{\mathfrak{D}} \times \mathbb{R}_+; \mathbb{R}_+^{\mathfrak{D}})$ such that

1. $A_{c,d}(\mathbf{z}, t) \geq 0$, $I_d(t) \geq 0$ and $O_d(\mathbf{z}, t) \geq 0$ $\forall \mathbf{z} \in \mathbb{R}_+^{\mathfrak{D}}, t \in \mathbb{R}_+$ and $d = 1, \ldots, \mathfrak{D}$
2. If for $\mathbf{z} \in \mathbb{R}_+^{\mathfrak{D}}$ we have $z_d = 0$, then $A_{c,d}(\mathbf{z}, t) = 0$ and $O_d(\mathbf{z}, t) = 0$ $\forall c = 1, \ldots, \mathfrak{D}$ and $t \in \mathbb{R}_+$.

Here $A_{c,d}$ quantifies transfer or flow from compartment d to compartment c, while I_d, O_d are the inflow into and outflow of compartment z_d. This is also illustrated in Figure 2.6, with arrows indicating flow into and out of the compartment.

The dynamics of a multi-compartmental system are then given by

$$\frac{\mathrm{d}z_d(t)}{\mathrm{d}t} = \sum_{c \neq d} \left(-A_{c,d}(\mathbf{z},t) + A_{d,c}(\mathbf{z},t) + I_d(t) - O_d(\mathbf{z},t)\right) \tag{2.24}$$

With the assumption that the function $z_d \mapsto A_{c,d}(\mathbf{z},t)$ is $\mathcal{C}^r(\mathbb{R}_+ \times \mathbb{R}_+)$ with $r \geq 1$ and the second property of Definition 2.14, it is possible to rewrite a multi-compartmental system as $A_{c,d}(\mathbf{z},t) = z_d a_{c,d}(\mathbf{z},t)$ and $O_d(\mathbf{z},t) = z_d o_d(\mathbf{z},t)$ for all $\mathbf{z} \in \mathbb{R}_+^D, t \in \mathbb{R}_+$.

We can then rewrite the dynamics as

$$\frac{\mathrm{d}z_d(t)}{\mathrm{d}t} = -\left(o_d(\mathbf{z},t) + \sum_{c \neq d} a_{c,d}(\mathbf{z},t)\right) z_d(t) + \sum_{c \neq d} a_{d,c}(\mathbf{z},t) z_c(t) + I_d(t) \tag{2.25}$$

Definition 2.15 (Fractional transfer coefficients)**.** The $a_{c,d}(\mathbf{z},t)$, $o_d(\mathbf{z},t)$ are the **fractional transfer coefficients**. If they are constant in $\mathbf{z}$, the multi-compartmental system is called **linear**.

With $a_{d,d}(\mathbf{z},t) := -\left(o_d(\mathbf{z},t) + \sum_{c \neq d} a_{c,d}(\mathbf{z},t)\right)$, the dynamics can be rewritten to an even easier matrix notation for $\mathbf{a} = (a_{c,d})_{c=1,\ldots,\mathfrak{D},d=1,\ldots,\mathfrak{D}}$:

$$\frac{\mathrm{d}\mathbf{z}(t)}{\mathrm{d}t} = \mathbf{a}\mathbf{z} + I \tag{2.26}$$

As already mentioned, in this thesis we only consider closed systems, meaning that $I, O = 0$. Furthermore, the multi-compartmental model presented in Chapter 7 is also linear and autonomous, meaning that $a_{c,d}(\mathbf{z},t) = a_{c,d}\ \forall c = 1,\ldots,\mathfrak{D}, d = 1,\ldots,\mathfrak{D}$. An illustration of this can also be seen in Figure 2.6(b).

For a linear, autonomous and closed multi-compartmental system $\frac{\mathrm{d}\mathbf{z}(t)}{\mathrm{d}t} = \mathbf{a}\mathbf{z}, \mathbf{z}(0) = \mathbf{z}_0$, the analytical solution of the matrix differential equation is given by

$$\mathbf{z}(t) = \exp(\mathbf{a}t)\mathbf{z}_0, \tag{2.27}$$

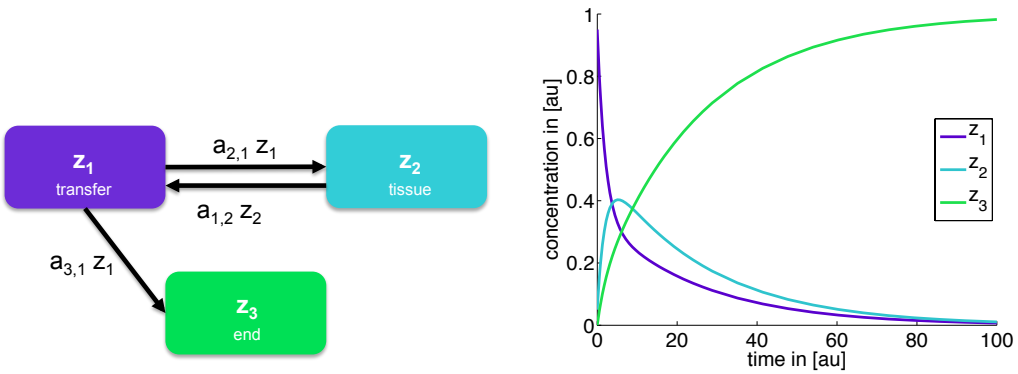
where $\exp(\mathbf{a}t)$ is a matrix exponential. In MATLAB, this can be computed by eigenvalue decomposition using the `eig` function, see also Appendix A.

Multi-compartmental models are often used when the whole body should be modeled, since then a certain level of abstraction is needed. They thus appear for example in pharmacokinetic models for the processing of drugs in the human body (Gelman *et al.* [1996a]; Shargel *et al.* [2005]), or in radiation science, where the processing of radioactive

substances in the human body is monitored (Greiter *et al.* [2011a,b]). An application example for this will be the processing of radioactive zirconium in the human body as presented in Chapter 7.

Example 2.7 (Small multi-compartmental model). We present an example for multi-compartmental models inspired by pharmacology. It models the dynamics of a chemical such as a medical drug. The model consists of three compartments: z_1, transfer (in this context this usually means the blood), z_2, the tissue into which the chemical distributes and from which it is transferred back to the transfer compartment and z_3, the end compartment where the chemical is excreted, e.g urine. The transitions between the compartments are then governed by a system of ODEs as introduced above. We assume a closed, linear and autonomous system. A visualization of the model and an example time course of the model can be seen in Figure 2.7. In equations, this yields

$$\begin{aligned}
\frac{\mathrm{d}z_1(t)}{\mathrm{d}t} &= -a_{2,1}z_1(t) - a_{3,1}z_1(t) + a_{1,2}z_2(t) \\
\frac{\mathrm{d}z_2(t)}{\mathrm{d}t} &= a_{2,1}z_1(t) - a_{1,2}z_2(t) \\
\frac{\mathrm{d}z_3(t)}{\mathrm{d}t} &= -a_{3,1}z_1(t)
\end{aligned}$$



(a) (b)

Figure 2.7: A multi-compartmental model. (a) The small multi-compartmental model made of the three compartments transfer, tissue and end. (b) Example time course for the multi-compartmental model with parameters $a_{2,1} = 0.25$, $a_{1,2} = 0.2$ and $a_{3,1} = 0.1$. Initial conditions were chosen as $z_1(0) = 0.95$, $z_2(0) = 0.05$ and $z_3(0) = 0$.

2.4 Observability and non-Bayesian parameter estimation for dynamical systems

In this section, we introduce the notion of an observable of a dynamical system, which then leads to the first, non-Bayesian parameter estimation procedure, the computation of a least-squares estimator. We furthermore present a bootstrapping procedure for assessing the goodness-of-fit.

2.4.1 Observables of a dynamical system

When dealing with ODE models, the most common task we are faced with in the context of systems biology is fitting the model to measurement data of the system. This means that the parameters of the model have to be determined or tuned to best fit the measured data, which typically contains measurement noise such that no perfect fit is possible. In a non-Bayesian estimation, this mostly means that optimization techniques are applied.

We already introduced a general formulation of a dynamical system in Equation (2.22), which we repeat here:

$$\frac{\mathrm{d}\boldsymbol{z}(t)}{\mathrm{d}t} = g(t, \mathbf{z}(t), \boldsymbol{\xi}, \boldsymbol{u}(t)), \mathbf{z}(0) = \mathbf{z}_0(\boldsymbol{\xi}).$$

It is parametrized by the vector $\boldsymbol{\xi} \in \mathbb{R}^{\mathfrak{d}_\xi}$, defining for example rate constants for the dynamical system. This formulation also includes the multi-compartmental systems introduced in the previous section. In biological systems, it is often not possible to measure all individual occurring components of $\mathbf{z}$. Instead they are only **partially observed**. Therefore we define:

Definition 2.16 (observable)**.** The l-th **observable** $\mathcal{Y}_l$ of the dynamical system in Equation (2.22) is

$$\mathcal{Y}_l(t, \boldsymbol{\xi}_o) = h_l(\mathbf{z}(t), \boldsymbol{\xi}_o), \tag{2.28}$$

where the function $h_l : \mathbb{R}^{\mathfrak{D}} \times \mathbb{R}^{\mathfrak{d}_o} \to \mathbb{R}$ is a link function relating the species of the ODE to the observable of the system at time t, possibly parametrized with parameters $\boldsymbol{\xi}_o \in \mathbb{R}^{\mathfrak{d}_o}$. In the easiest case, a link function is just a projection on one component of $\mathbf{z}$, but in real biological application it is often only possible to measure e.g. sums of species, their ratios or scaled versions, which can all be expressed by a link function

h_l, e.g. $h_l(\mathbf{z}(t), \boldsymbol{\xi}_o) = z_1(t) + z_2(t)$. It is also possible to consider an observable vector $\mathcal{Y} = \mathbf{h}(\mathbf{z}(t), \boldsymbol{\xi}_o)$, whose components are the $\mathcal{Y}_l$ as defined above.

We can now define what we actually mean by a "model":

Definition 2.17 (model)**.** A **model M** with model parameters $\boldsymbol{\psi} = (\boldsymbol{\xi}, \boldsymbol{\xi}_o)$ is a combination of

- **dynamics** $\frac{\mathrm{d}\boldsymbol{z}(t)}{\mathrm{d}t} = g(t, \mathbf{z}(t), \boldsymbol{\xi}, \boldsymbol{u}(t)), \mathbf{z}(0) = \mathbf{z}_0(\boldsymbol{\xi})$ and
- **observables** $\mathcal{Y}_l(t, \boldsymbol{\xi}_o) = h_l(\mathbf{z}(t), \boldsymbol{\xi}_o), l = 1, \ldots, L.$

We now assume that the system has L observables, typically we will have that $L < \mathfrak{D}$, meaning that there are fewer observables than species in the system. Nevertheless in all applications presented later in this thesis, the aim is to infer the parameters of the system from a set of given observations $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\} \in \mathbb{R}^{L \times N}$. The individual $\mathbf{y}_n \in \mathbb{R}^L, n = 1, \ldots, N$ usually correspond to measurements taken at time points $t_n \in [0, T]$. It is then assumed that the measurement data corresponds to the state of the observables of the system, overlaid with measurement noise.

It should thus satisfy

$$y_{n,l} = \mathcal{Y}_l(t_n, \boldsymbol{\xi}_o) + \epsilon_{n,l}, \quad l = 1, \ldots, L, n = 1, \ldots, N, \tag{2.29}$$

or

$$y_{n,l} = \mathcal{Y}_l(t_n, \boldsymbol{\xi}_o) \cdot \epsilon_{n,l}, \quad l = 1, \ldots, L, n = 1, \ldots, N, \tag{2.30}$$

for each data vector $\mathbf{y}_n$, in which the $\epsilon_{n,l}$'s are independent realizations of the assumed error distributions that have to be determined.

2.4.2 Least squares estimators

If we assume that $\epsilon_{n,l} \sim \mathcal{N}(0, \sigma_{n,l}^2)$ with $\sigma_{n,l}$ known, we can set up a cost function for fitting the parameters from the **least squares** approach. We find the arg min of the function

$$\chi^2(\boldsymbol{\psi}) = \sum_{n=1}^{N} \sum_{l=1}^{L} \frac{(y_{n,l} - \mathcal{Y}_l(t_n, \boldsymbol{\xi}_o)))^2}{\sigma_{n,l}^2}. \tag{2.31}$$

This function is bounded from below by zero. Still, for a nonlinear relationship between the parameters and the observables, there usually exists no closed form solution and

numerical schemes might face convergence issues. Furthermore, the set for which the minimum is attained can contain more than one point. In this case, the parameter vector $\boldsymbol{\psi} = (\boldsymbol{\xi}, \boldsymbol{\xi}_o)$ that minimizes the mean squared error does not have to be unique. To find the arg min, various optimization techniques can be applied, in the nonlinear case, this is not a trivial problem. Popular methods include nonlinear global (often stochastic) optimization algorithms such as simulated annealing (Kirkpatrick *et al.* [1983]), the genetic algorithm (Fraser & Burnell [1970]) or scatter search (Egea *et al.* [2007]; Rodriguez-Fernandez *et al.* [2006]). However, we have found that often local, gradient based optimization algorithms like trust-region or interior point (Byrd *et al.* [1999, 2000]; Coleman & Li [1996]) perform better when combined with a meaningful strategy for choosing initial points like Latin hypercube sampling (Iman [2008]). For a nice comparison in the context of systems biology, we recommend Raue *et al.* [2013b].

Example 2.8 (Dimerization continued). We can now return to Example 2.6. There we consider the dynamics of two species Z_1 and Z_2. If we now assume that we can observe only the concentration z_2, the product of the reaction, this can be defined as the observable (if we only have one observable, we can skip the index l):

$$\mathcal{Y}(t, \boldsymbol{\xi}_o) = h(\mathbf{z}(t), \boldsymbol{\xi}_o) = z_2(t). \tag{2.32}$$

If we have obtained some measurement data for this observable with known normally distributed noise of standard deviation $\sigma = 0.05$, we can then perform a least squares fit, yielding the fit in Figure 2.8.

The question of how to fit a model to measurement data is considered in more detail in the following chapter, as the method used in this thesis is Bayesian inference, the topic of said chapter. Another approach for assessing parameter uncertainty is also the profile likelihood approach, which will be introduced in Section 3.3.

2.4.3 Bootstrapping a goodness-of-fit statistic

In some cases, especially with newly proposed models, it is not clear initially if a model explains the data at all. In this case, we can use a bootstrap procedure for assessing a goodness-of-fit statistic, as introduced in e.g. Efron & Tibshirani [1993]; Stute *et al.* [1993]. We follow the overview at von Davier [1997].

Generally speaking, we consider a **goodness-of-fit** function S of the observed data $\mathbf{Y}$

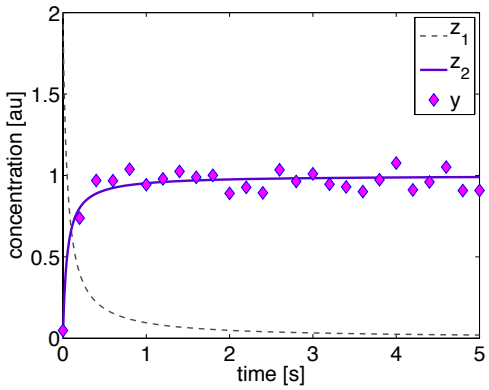


Figure 2.8: Dimerization: measurement data, observable and least squares fit. The observable $\mathcal{Y}(t, \boldsymbol{\xi}_o) = z_2(t)$ (solid line) and the measurement data $\mathbf{Y}$. The best parameter ξ was obtained by a least squares fit, such that the observable and the measurement data agree well. The concentration of z_1 (dashed) is in this case unobserved, but can of course be computed from the model.

and a vector $\boldsymbol{\psi}$ of parameters, $S(\mathbf{Y}, \boldsymbol{\psi})$. This can for example be the least squares cost function as introduced in Equation (2.31) or a likelihood function as introduced in the following chapter. Through optimization of this cost function, we obtain an optimal parameter value $\hat{\boldsymbol{\psi}}$ and thus a value $s^* = S(\mathbf{Y}, \hat{\boldsymbol{\psi}})$ for the goodness-of-fit statistic.

We now want to approximate the unknown distribution F_S of S. We do this by a bootstrap, i.e. by generating a sample of independent outcomes s_j for $j = 1, \ldots, J_{\mathbf{BS}}$ and thus construct an empirical distribution $\hat{F}_{s^*}$.

This is achieved by generating $J_{\mathbf{BS}}$ additional artificial data sets $\mathbf{Y}^j$ from the model with the estimated parameter vector $\hat{\boldsymbol{\psi}}$. Then parameters $\boldsymbol{\psi}^j$ are estimated based on the simulated dataset, e.g. by re-optimization. Each dataset then yields one value $s^j = S(\mathbf{Y}^j, \boldsymbol{\psi}^j)$.

If the model can in principle explain the original measurement data, the value s^* should not be significantly different from the bootstrap sample $s^j, j = 1, \ldots, J_{\mathbf{BS}}$. This can for example be tested by computing a z-score (Kendall & Stuart [1979]) for s^* by

$$z = \frac{s^* - \mu}{\sigma}, \tag{2.33}$$

in which μ and σ are the mean and standard deviation of the $s^j, j = 1, \ldots, J_{\mathbf{BS}}$, respectively.

It should be mentioned that performing a bootstrap for the goodness-of-fit statistic is

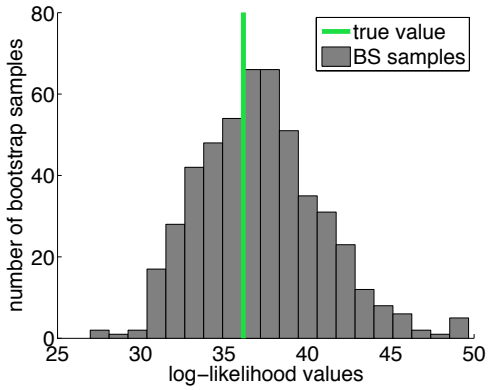


Figure 2.9: Dimerization: bootstrapping the goodness-of-fit. Histogram for the dimerization model's log-likelihood values based on 500 bootstrap (BS) samples, with a green line for the true maximum likelihood value.

usually associated with high computational costs, since for each generated data set, the model has to be fitted. Of course this also crucially depends on the desired number of bootstrap samples $J_{\mathbf{BS}}$.

Example 2.9 (Dimerization continued). We can now return again to Example 2.6. There we have already obtained a least squares fit. Similarly, we can also obtain a likelihood fit, as will be elaborated on further in the next chapter. We now draw $J_{\mathbf{BS}} = 500$ bootstrap samples for assessing the goodness-of-fit based on the log-likelihood. From these bootstrap samples, we find a z-score of -0.28, indicating that the model can indeed fit the data. This is not surprising, since the model was used to generate the data. A histogram of the bootstrapped values together with the true value can also be seen in Figure 2.9.

We have given all the preliminaries from mathematics and biology. We will now move on to Bayesian parameter estimation, also called Bayesian inference, which will be presented in the following chapter.

3

Bayesian inference for dynamical systems

In this chapter, we will give a very brief overview over the principles of Bayesian inference. We will concentrate on introducing the posterior and prior distribution and on the property of identifiability of the parameters of a distribution. The theory for Bayesian inference is very rich, as is the discussion culture between frequentist statisticians and Bayesian statisticians, however we feel that such ground-laying work is better looked up elsewhere and we encourage our readers to delve into the literature themselves (see e.g. Box & Tiao [2011]; O'Hagan *et al.* [2004]).

This chapter is based on some concepts introduced in Section 2.4, e.g. the definition of a model. We here concentrate on inference of the posterior distribution, the inference of the marginal likelihood for model selection will be the topic of Chapter 5.

3.1 Bayesian parameter inference of the posterior distribution

The aim of Bayesian inference in our context is primarily to infer the distribution of the parameter vector $\boldsymbol{\theta} \in \mathbb{R}^{\mathfrak{d}}$ for a given parametrized ODE model $\mathbf{M}$, based on the set of measurements $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$. By $\boldsymbol{\theta}$ we here mean the combination of all relevant parameters. This includes all parameters $\boldsymbol{\psi}$ associated with model $\mathbf{M}$ as defined in Definition 2.17 in the previous section, which might be more than the parameters $\boldsymbol{\xi}$ by

which solely the ODE is parametrized. Furthermore, as introduced in Equations (2.29) and (2.30), $\boldsymbol{\theta}$ often also includes error model parameters that have to be considered.

As presented in detail in Robert & Casella [2004], in a Bayesian view both the parameter vector and every data point are seen as realizations of random variables. This means for example that a parameter has a distribution in the Bayesian paradigm. For the data points, the realizations come from a density function $f(\cdot|\boldsymbol{\theta})$ conditioned on the parameter vector $\boldsymbol{\theta}$. Given this density, the **likelihood function** or simply likelihood for the data $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ is given by

$$p(\mathbf{Y}|\boldsymbol{\theta}) = p(\mathbf{y}_1, \ldots, \mathbf{y}_N|\boldsymbol{\theta}) = \prod_{n=1}^{N} f(\mathbf{y}_n|\boldsymbol{\theta}), \tag{3.1}$$

if the $\mathbf{y}_1, \ldots, \mathbf{y}_N$ are independent. If the assumed error distribution is Gaussian and for known parameters of this error distribution $\sigma_{n,l}^2$, the likelihood coincides with the least squared error of Equation (2.31).

Bayes' theorem now combines the likelihood with existing **prior** knowledge about the parameters $p(\boldsymbol{\theta})$, which is a distribution for the parameter, to yield the **posterior distribution** $p(\boldsymbol{\theta}|\mathbf{Y})$ of the parameter given the observed data:

$$p(\boldsymbol{\theta}|\mathbf{Y}) = \frac{p(\mathbf{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int_{\mathbb{R}^\mathfrak{d}} p(\mathbf{Y}|\tilde{\boldsymbol{\theta}})p(\tilde{\boldsymbol{\theta}})\,\mathrm{d}\tilde{\boldsymbol{\theta}}}. \tag{3.2}$$

The posterior distribution is the core of Bayesian inference. It corresponds to the probability distribution of the parameter $\boldsymbol{\theta}$ taking into account both the information provided by the measurement data $\mathbf{Y}$ as well as previously available information about the parameters in form of the prior $p(\boldsymbol{\theta})$. Though the first task of the integral in the denominator

$$p(\mathbf{Y}) = \int_{\mathbb{R}^\mathfrak{d}} p(\mathbf{Y}|\tilde{\boldsymbol{\theta}})p(\tilde{\boldsymbol{\theta}})\,\mathrm{d}\tilde{\boldsymbol{\theta}} \tag{3.3}$$

is to provide normalization in order for the posterior to be a true probability density that integrates to one, it is in itself an important quantity known as the **marginal likelihood** or **evidence**. This marginal likelihood is the key ingredient for Bayesian model selection, as it provides the probability of the data. However, the integral is often difficult to evaluate, since it is typically high dimensional and analytically intractable. Instead, sophisticated numerical or sampling based methods have to be applied. For more on this, see Chapter 5.

For inference of the posterior distribution, it is desirable to be able to skip the computation of the marginal likelihood. This is often done by Markov chain Monte Carlo

methods, see Chapter 4. These exploit the following fact. Since the marginal likelihood only depends on the data and not on the parameter vector, and the data is considered fixed, it is possible to base the inference on the relation

$$p(\boldsymbol{\theta}|\mathbf{Y}) \propto p(\mathbf{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{3.4}$$

This relation is one of the main reasons for the increasing popularity of Markov chain Monte Carlo methods in the last few years.

We now go back to the non-Bayesian parameter estimation in Section 2.4. There we defined a model to consist of the dynamical system

$$\frac{\mathrm{d}\boldsymbol{z}(t)}{\mathrm{d}t} = g(t, \mathbf{z}(t), \boldsymbol{\xi}, \boldsymbol{u}(t)), \mathbf{z}(0) = \mathbf{z}_0(\boldsymbol{\xi}),$$

parametrized by the vector $\boldsymbol{\xi}$ and of the observables $\mathcal{Y}_l(t, \boldsymbol{\xi}_o) = h_l(\mathbf{z}(t), \boldsymbol{\xi}_o)$ for $l = 1, \ldots, L$. To fit the observables of the model to the measurement data $\mathbf{Y}$, we have to make assumptions on the error model. If we assume additive measurement noise, we get

$$y_{n,l} = \mathcal{Y}_l(t_n, \boldsymbol{\xi}_o) + \epsilon_{n,l}, \quad l = 1, \ldots, L, n = 1, \ldots, N \tag{3.5}$$

where the $\epsilon_{n,l}$'s are independent realizations of the assumed error distributions. The functions h_l are again the link functions, relating the species of the ODE to the l-th observable of the system for each time point t_n. Then the parameter vector for the inference $\boldsymbol{\theta}$ corresponds to $\boldsymbol{\xi}$, $\boldsymbol{\xi}_o$ and all parameters of the error distribution, e.g. $\sigma^2_{n,l}$. Let $f^{(n,l)}$ be the probability density for $\epsilon_{n,l}$. Then the posterior can be written as

$$p(\boldsymbol{\theta}|\mathbf{Y}) \propto \underbrace{\prod_{n=1}^{N}\prod_{l=1}^{L} f^{(n,l)}\left(y_{n,l} - \mathcal{Y}_l(t_n, \boldsymbol{\xi}_o)|\boldsymbol{\theta}\right)}_{p(\mathbf{Y}|\boldsymbol{\theta})} p(\boldsymbol{\theta}) \tag{3.6}$$

The same is also possible with a multiplicative noise model. With the Bayesian approach it is thus straightforward to simultaneously infer both the parameters of the dynamical system as well as the noise parameters, as they all together yield $\boldsymbol{\theta}$. Inference of the posterior distribution can for example be done with the Markov chain Monte Carlo methods introduced in the following chapter.

The posterior $p(\boldsymbol{\theta}|\mathbf{Y})$ and the likelihood $p(\mathbf{Y}|\boldsymbol{\theta})$ in Equation (3.6) yield a key ingredient for many analyses of a model. For one, we define the parameter vector $\hat{\boldsymbol{\theta}}$ as the one where the likelihood attains its maximum value and call $\hat{\boldsymbol{\theta}}$ the **maximum likelihood estimate** (MLE). Note that the maximum likelihood estimate does not have to exist

and does not have to be unique, for example in case of a plateau in the likelihood. Correspondingly, the vector where the posterior distribution $p(\boldsymbol{\theta}|\mathbf{Y})$ achieves its maximal value is called the **maximum a posteriori estimate** (MAP) estimate, if it exists. It is often denoted by $\hat{\boldsymbol{\theta}}_{\text{MAP}}$. Obviously, the MAP and MLE are the same vector if the maximum is unique, the prior distribution is uniform and its support contains the MLE.

The Bayesian paradigm also leads to different uncertainty measures than frequentist inference. Analogously to confidence intervals, but with different interpretation, the Bayesian set-up yields **credible sets** or **intervals** for $\boldsymbol{\theta}$. For the desired **confidence level** α, the $100\%(1-\alpha)$ credible interval is given by the set

$$\text{CI} = \{\boldsymbol{\lambda}; p(\boldsymbol{\lambda}|\mathbf{Y}) \geq \zeta_\alpha\}, \tag{3.7}$$

where the coverage constraint ζ_α has to be determined such that

$$p(\boldsymbol{\lambda} \in \text{CI}|\mathbf{Y}) = 1 - \alpha. \tag{3.8}$$

This credible set is usually viewed component-by-component to yield credible sets for each individual component θ_s of $\boldsymbol{\theta}$. If the likelihood is "well-behaved", the credible set of each component is indeed one connected interval, but the credible set can also have a more irregular shape (Marin & Robert [2007]).

Example 3.1 (Dimerization continued). We reconsider the dimerization example from Section 2.3:

$$\dot{z_1}(t) = -2\xi z_1^2(t) \tag{3.9}$$

$$\dot{z_2}(t) = \xi z_1^2(t) \tag{3.10}$$

For a choice of $\xi = 5$ and $\mathbf{z}(0) = (z_{0,1}, z_{0,2})^\top = (2, 0)^\top$, we can generate artificial data from the observable $\mathcal{Y}(t) = z_2(t)$, with additive normally distributed noise with standard deviation $\sigma = 0.05$. We use $N = 26$ time points equally spaced between $t = 0s$ and $t = 5s$ and obtain $\mathbf{Y} = (y_1, \ldots, y_N)$. A visualization of the data was already given in Figure 2.8. We now want to infer the parameters $\boldsymbol{\theta} = (\xi, \sigma)$. The likelihood is given by

$$p(\mathbf{Y}|\boldsymbol{\theta}) = \prod_{n=1}^{N} \phi\left(y_n; \mathcal{Y}(t_n), \sigma\right), \tag{3.11}$$

in which $\phi(x; \mu, \sigma)$ is the probability density function of the univariate normal distribution evaluated at x with mean μ and standard deviation σ as introduced in Section

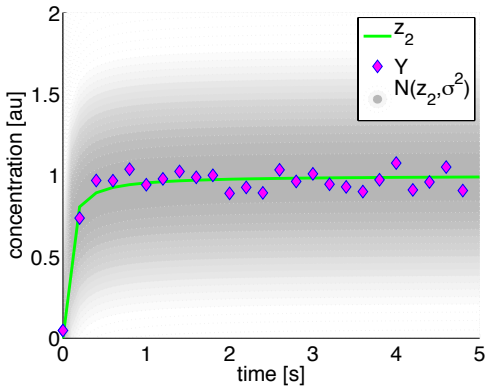


Figure 3.1: Dimerization: maximum likelihood estimate. For the dimerization introduced in Example 2.6, we obtain the maximum likelihood estimate of both the reaction rate ξ and the noise parameter σ. The model fit according to the inferred reaction rate ξ can be seen in green and shows good agreement with the measurement data (magenta diamonds). Grey shading shows the probability density of the noise distribution with the inferred parameter σ around the model fit.

2.2.1. Optimizing the likelihood yields an MLE of $\hat{\boldsymbol{\theta}} = (5.2356, 0.0602)^\top$, which is close to the true value with which the data was generated. The fit and the error distribution around the fit can be seen in Figure 3.1. Using MCMC as will be introduced in the following chapter, we can also derive credible intervals.

3.2 Choice of prior distribution

For an overview on the prior distribution already introduced in the previous section, we follow along the lines of Marin & Robert [2007]. As the name already suggests, the prior information summarizes the information available on the parameter $\boldsymbol{\theta}$ without knowing the data $\mathbf{Y}$. Ideally, this would be from previous experiments or based on independent datasets. However, in practice such information is often not readily available. Then other prior choices have to be made, for example based on practical grounds or rather in the form of noninformative priors.

If genuine and quantitative prior information exists, it can and must be used for the prior distribution. Otherwise, more generic solutions exist, like conjugate priors. Conjugate priors are chosen such that the prior and posterior distribution belong to the same parametric family. All prior distributions can be parametrized themselves with the so called hyperparameters, which could in turn have hyperpriors.

For most applications in this thesis, the appropriate choice of conjugate priors is not easy, as for example for a noise model based on the gamma distribution, the conjugate priors exist, but are not of standard form. Furthermore, such a choice is often a rather severe interference with the actual shape of the posterior. Thus instead of using conjugate priors, we rather choose the prior distributions such that their influence on the inference is weakened. Such prior distributions can be called **noninformative** or vague. Marin & Robert [2007] take noninformative priors to mean extensions of the uniform distribution. A standard choice for a noninformative prior would thus be $\theta_s \sim \mathcal{U}[B_s]$ for a non-empty measurable set B_s for each parameter θ_s contained in $\boldsymbol{\theta}$. In fact, it would be possible to choose any non-negative function $\pi : \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ with

$$\int_{\mathbb{R}^{\mathfrak{d}}} \pi(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} = c \tag{3.12}$$

for some constant $c \in (0, \infty]$ as long as the marginal likelihood is finite almost surely. In case $c = \infty$, we call the prior **improper** (Robert [2001]). Note that also an improper prior can lead to a proper posterior distribution, and a proper prior can still lead to an improper posterior.

Often, it is rather straightforward to introduce uniform prior distributions in our examples, since biological rate constants have to be positive and are thus bounded below by zero. Also upper boundaries are often easily chosen. In general, nothing is known about the dependence structure of the parameters, we usually assume independence and thus have $p(\boldsymbol{\theta}) = \prod_{s=1}^{\mathfrak{d}} p(\theta_s)$ for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$.

Asymptotically it can also be said that the influence of the prior distribution diminishes as the number of data points increases. Then for example also the MAP recovers asymptotic properties of the MLE.

3.3 Parameter identifiability

A related concept to Bayesian inference, yet different, is that of **profile posteriors** for **identifiability** analysis in a computational model (Raue *et al.* [2013a]; Vanlier *et al.* [2012]).

In layman's terms, identifiability analysis of a parameter means to analyze if a parameter can be determined at all with finite confidence bounds. Identifiability as such is thus a property of either the model on a global scale or the posterior on a local scale.

Precise and detailed mathematical definitions of these different views of identifiability can be found in Audoly *et al.* [1998, 2001]; Chis *et al.* [2011]; Little *et al.* [2010].

In a more measurement data oriented setting, we can distinguish **structural and practical non-identifiability**. Both can have several reasons (Raue *et al.* [2009]). Structural non-identifiability is independent of the measurement data and instead due to some underlying fundamental redundancy in the parametrization of the model. A structurally non-identifiable parameter can thus not be inferred at all, independent of the quality of the measurement data. A typical example is a setting where only the ratio of two parameters is determinable, but neither of the two individual parameters.

Practical non-identifiability depends on the amount and quality of the data: if the data is insufficient, it is often only possible to derive either an upper or lower confidence bound for the parameter, but not both.

Identifiability analysis is usually based on the maximum likelihood or, in our case, maximum a posteriori estimate $\hat{\boldsymbol{\theta}}$ of the parameters and can then be conducted with the profile posterior approach.

The profile posterior approach is analogous to the profile likelihood approach introduced by Raue *et al.* [2009]. The basic idea of the approach is to explore the parameter space for each parameter separately in direction of the least decrease of the posterior $p(\boldsymbol{\theta}|\mathbf{Y})$. This can be done by calculating the profile posterior

$$p_{pp}(\theta_s) = \max_{\theta_{j\neq s}} \left[p(\boldsymbol{\theta}|\mathbf{Y}) \right], \tag{3.13}$$

re-optimizing the posterior with respect to all parameters $\theta_{j\neq s}$, for each value of θ_s. This is an extension of the originally used **profile likelihood**

$$p_{pl}(\theta_s) = \max_{\theta_{j\neq s}} \left[p(\mathbf{Y}|\boldsymbol{\theta}) \right]. \tag{3.14}$$

The presence of local optima, i.e. multiple modes in the posterior (or likelihood), can be detected by repeated optimization runs from different starting values. If such local optima are detected, the profile calculation has to be initiated in each of the optima. Note that the calculation of the profiles for different parameters can be performed independently and simultaneously on different computer cores. For more details on the implementation, see Raue *et al.* [2009]. A generalization of this approach to model predictions by calculating prediction profiles was proposed in Kreutz *et al.* [2012]. All profile posteriors in this thesis were computed with MATLAB, based on code from either Hasenauer [2014] or Raue *et al.* [2013b].

If the prior distribution of the parameters is uniform, the profile posterior yields a scaled and bounded version of the profile likelihood. In the case of non-uniform priors, the comparison of profile likelihood and profile posterior can also be used to assess the information content of the data with respect to the parameters. Furthermore, this comparison reveals if identifiability is only enforced by the prior distribution. This comparison of the two profile types is thus related to a prior/posterior evaluation in a Bayesian setting.

The profile posterior or rather the profile likelihood allows for the calculation of confidence intervals which can be compared to the Bayesian credible intervals obtained from MCMC sampling. More specifically, we consider likelihood-based confidence intervals to a 95% confidence level. In contrast to the Bayesian view, this means that the true value of the parameter θ_s^* is expected to be inside the interval with 95% probability. Following Raue *et al.* [2013a], we can derive point-wise intervals by using a threshold of $\Delta_\alpha = Q(\chi_1^2, \alpha)$, which is the α quantile of the χ^2-distribution with 1 degree of freedom. The confidence intervals are then given by

$$\{\theta_s | -2\log(p_{pl}(\theta_s)/p(\mathbf{Y}|\hat{\boldsymbol{\theta}})) < \Delta_\alpha\}, \tag{3.15}$$

in which $p_{pl}(\theta_s)$ is the profile likelihood, see also Raue *et al.* [2013a], and $\hat{\boldsymbol{\theta}}$ is the MLE.

Since profile posteriors are computationally often less expensive, they form an optimal basis for any MCMC procedure for Bayesian inference and thus complement these for gaining thorough insight into the dynamical system at question. Then also the histogram of the marginalized samples can very well be compared with the posterior profiles, as will be shown in the applications part.

A completely flat profile posterior is a necessary, yet not sufficient condition for a structurally non-identifiable parameter. If the profile is flat, this does thus not automatically mean that the parameter is structurally non-identifiable, though this is often the case. A parameter is practically non-identifiable at confidence level α, if at least one boundary of the corresponding confidence interval is infinite. This usually indicates a profile that is flat in at least one direction. A nice visualization can be found in Raue [2013]. In this thesis, we focus on the detection of practical non-identifiabilities in our models.

Example 3.2 (Dimerization continued). We reconsider the dimerization example from Section 2.3:

$$\dot{z_1}(t) = -2\xi z_1^2(t) \tag{3.16}$$

$$\dot{z_2}(t) = \xi z_1^2(t) \tag{3.17}$$

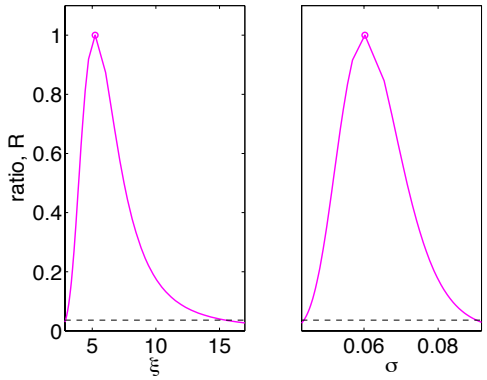


Figure 3.2: Dimerization: profile likelihoods. For the dimerization introduced in Example 2.6, we obtain the profile likelihoods of both the reaction rate ξ and the noise parameter σ (magenta lines). We scale the profile likelihood with the maximum likelihood value to obtain a maximum of 1 in the plot. Dashed lines are the point wise confidence interval thresholds for 95% confidence derived from Equation (3.15).

For a choice of $\xi = 5$ and $\mathbf{z}(0) = (z_{0,1}, z_{0,2})^\top = (2, 0)^\top$, we generated artificial data from the observable $\mathcal{Y}(t) = z_2(t)$, with additive normally distributed noise with standard deviation $\sigma = 0.05$. We use $N = 26$ time points equally spaced between $t = 0$ and $t = 5$ and obtain $\mathbf{Y} = (y_1, \ldots, y_N)$. A visualization of the data was already given in Figure 2.8. Optimizing the likelihood yielded an MLE of $\hat{\boldsymbol{\theta}} = (5.2356, 0.0602)^\top$, which is close to the true value with which the data was generated. The fit and the error distribution around the fit can be seen in Figure 3.1. We now computed the profile likelihood of the parameters and find that both parameters are identifiable at a 95% confidence level. We obtained 95% confidence intervals of $[3.3, 10.5]$ for ξ and $[0.047, 0.081]$ for σ. A scaled visualization of the profiles can also be seen in Figure 3.2. The example will be continued in Example 4.1 with the results of an MCMC algorithm for the model.

Bayesian inference as presented in this chapter is the basis for the Markov chain Monte Carlo and Bayesian model selection methods presented in the next two chapters. Identifiability analysis will be important for the analysis of the application examples in the second part of this thesis.

4

Markov chain Monte Carlo methods

In this chapter, we give a rather brief overview over the Markov chain Monte Carlo (MCMC) algorithms that are applied in this thesis. The focus here is rather to provide some hands-on experience of algorithms we have found useful and suitable for our problems from systems biology, and not so much to delve into the rich theory of MCMC. The difficulty lies in tuning the algorithms to optimize performance for sampling distributions arising from ODE models. Most algorithms cannot be applied "out of the box" but have to be tuned, mostly due to issues relating to the scalability of MCMC algorithms to higher-dimensional sampling spaces. We first present the respective algorithms, starting with the well-known Metropolis-Hastings algorithm. Finally we also show how to diagnose if the performance of the algorithms is satisfactory.

Generally speaking, MCMC algorithms are used for sampling from a probability distribution by constructing a Markov chain whose equilibrium distribution is the target probability distribution. The samples from the Markov chain then give an approximation of the target distribution. MCMC algorithms are most commonly applied for numerically calculating multi-dimensional integrals as will also be shown in Chapter 5 or for empirically estimating the statistical moments like mean or variance of the target distribution, for example in order to describe the posterior distribution in parameter inference problems. All presented algorithms are implemented in MATLAB.

This chapter introduces the new sampling algorithm Adaptive Metropolis Parallel Hierarchical Sampling, a variant of the established Parallel Hierarchical Sampling from

Rigat & Mira [2012]. It was already introduced in the following publication:

- **S. Hug***, A. Raue*, J. Hasenauer, J. Bachmann, U. Klingmüller, J. Timmer and F.J. Theis (2013). High-dimensional Bayesian parameter estimation: case study for a model of JAK2/STAT5 signaling. Mathematical Biosciences, 246(2), 293-304.

4.1 The Metropolis-Hastings algorithm

One of the most popular and widely applied MCMC algorithms is the **Metropolis-Hastings** (MH) algorithm as introduced by Metropolis *et al.* [1953] and Hastings [1970]. Its properties and theoretical background have been studied extensively (Beichl & Sullivan [2000]; Robert & Casella [2004]; Roberts *et al.* [1997]). It has the advantage of imposing only minimal requirements on the target density and is thus universally applicable. The basis is the target density $p(x)$ that we want to sample from, we often also refer to the **target distribution** for the distribution with density $p(x)$. We then need a conditional density $q(y|x)$, where the distribution with density $q(\cdot|x)$ is often called the **proposal distribution**, suggesting a move from the current state x to a new state y. The proposal has to satisfy only few conditions, e.g. symmetry ($q(y|x) = q(x|y)$). Furthermore, we require that the ratio $p(y)/q(y|x)$ is known up to a constant independent of x. This requirement is not severe for our usual posterior distributions that we want to sample from. The Metropolis-Hastings algorithm then generates a Markov chain as follows: Given a current state $\mathbf{X}^{(j)}$, draw a proposed next state $\mathbf{X}^* \sim q(\cdot|\mathbf{X}^{(j)})$. The move to this next state is accepted with the **Metropolis-Hastings acceptance probability**

$$\alpha(\mathbf{X}^{(j)}, \mathbf{X}^*) = min\left\{1, \frac{p(\mathbf{X}^*)}{p(\mathbf{X}^{(j)})} \frac{q(\mathbf{X}^{(j)}|\mathbf{X}^*)}{q(\mathbf{X}^*|\mathbf{X}^{(j)})}\right\} \tag{4.1}$$

If the move is accepted, we set $\mathbf{X}^{(j+1)} = \mathbf{X}^*$, otherwise $\mathbf{X}^{(j+1)} = \mathbf{X}^{(j)}$. Pseudo-code describing the method can be seen in Algorithm 1.

Theorem 4.1. *Let $\mathbf{X}^{(j)}$ be a Markov chain from the Metropolis-Hastings algorithm where moves are accepted according to Equation (4.1). For every proposal distribution $q(y|x)$ whose support includes the support of $p(x)$, it holds that*

- *the transition kernel of the chain satisfies the detailed balance condition with p;*
- *p is a stationary distribution of the chain.*

The proof can be found in Robert & Casella [2004].

An important criterion for the quality of the resulting Markov chain is its overall acceptance rate, the number of accepted moves divided by the number of total proposed moves. Since it has been shown that for a d-dimensional Gaussian target distribution, the optimal acceptance rate is 23% (Gelman *et al.* [1996b]; Roberts *et al.* [1997]), we aim for acceptance rates of $15 - 35\%$ in our applications by scaling the proposal distribution accordingly. In most applications, the proposal distribution will be a normal distribution centered at the current state. The covariance matrix of the normal distribution however has to be tuned or determined, often by trial and error. In easy cases, an identity matrix is sufficient, and the desired acceptance rate can be achieved by scaling the matrix accordingly. In more involved cases, it is often not easy to find a good proposal distribution. It is crucial to achieve good mixing of the chain. If the proposed steps are very small, the acceptance rate might be very high, but convergence is very slow. This means that a lot of samples are required until the sample distribution is close to the target distribution. On the other hand, if the proposed steps are very large, the acceptance rate usually suffers and the sampling gets "stuck". This leads to high autocorrelation in the chain and thus also bad convergence properties. In our experience, the MH algorithm needs some tuning by an expert in every application to yield good results.

Algorithm 1: The Metropolis-Hastings algorithm

input : Initial sample $\mathbf{X}^{(0)}$, sampling distribution $p(x)$, desired number of samples J, proposal distribution $q(y|x)$

output: Samples $\mathbf{X}^{(j)}, j = 0, \ldots, J$ from $p(x)$

for $j \leftarrow 0$ **to** J **do**
  Generate a proposal $\mathbf{X}^* \sim q(\cdot|\mathbf{X}^{(j)})$ and draw $u \sim \mathcal{U}[0,1]$;
  if $u \leq \alpha(\mathbf{X}^{(j)}, \mathbf{X}^*)$ **then**
    Set $\mathbf{X}^{(j+1)} = \mathbf{X}^*$;
  else
    $\mathbf{X}^{(j+1)} = \mathbf{X}^{(j)}$;
  end
end

4.2 The Adaptive Metropolis algorithm

The performance of regular Metropolis-Hastings algorithms strongly depends on finding an adequate proposal distribution, which is not always an easy task. For this reason, adaptive methods that refine the proposal distribution iteratively, have generated quite a bit of interest. One popular and easy to use adaptive method is the **Adaptive Metropolis** (AM) algorithm as introduced by Haario *et al.* [2001]. It is especially suitable for sampling higher-dimensional target distributions since its proposal function can be continuously adapted to guarantee efficient sampling of the high-dimensional space. This is a crucial factor for the convergence of the algorithm. In the AM algorithm, the proposal function is a multivariate normal distribution whose covariance matrix is updated with the information gained from the obtained samples by a recursion formula. Newly proposed samples are then accepted or rejected according to a standard Metropolis-Hastings acceptance scheme. This sampling process is strictly speaking non-Markovian as the samples depend on the past of the sampling procedure and not just on their immediate predecessor, however Haario et al. show that the algorithm has the correct ergodic properties and is thus a valid method for sampling from a target distribution. It is not straightforward to guarantee ergodicity in an adaptive sampling scheme, but special care has to be taken, which is a good reason for applying the AM, where ergodicity can be proven.

The AM scheme is based on a normal distribution as proposal with mean zero as usual and an adaptively updated covariance matrix. For this update, an index s_0 where the adaption starts is chosen and the covariance matrix of the proposal is set to

$$C_j = \begin{cases} C_0, & \text{if } j \leq s_0 \\ \gamma_{\mathfrak{d}} cov(\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(j)}) + \gamma_{\mathfrak{d}} \epsilon I_{\mathfrak{d}}, & \text{if } j > s_0 \end{cases} \tag{4.2}$$

Here, $\gamma_{\mathfrak{d}}$ is a scaling parameter depending on the dimension $\mathfrak{d}$ of the sampling space and $I_{\mathfrak{d}}$ is the $\mathfrak{d}$-dimensional identity matrix. The constant $\epsilon > 0$ may be chosen to be very small, it ensures that the covariance matrix does not become singular. The initial covariance matrix C_0 has to be strictly positive definite and should represent the best available prior knowledge. The empirical covariance matrix of the samples permits the derivation of the following recursion formula for the proposal covariance matrix:

$$C_{j+1} = \frac{j-1}{j} C_j + \frac{\gamma_{\mathfrak{d}}}{j} \left(j \bar{\mathbf{X}}_{j-1} \bar{\mathbf{X}}_{j-1}^\top - (j+1) \bar{\mathbf{X}}_j \bar{\mathbf{X}}_j^\top + \mathbf{X}^{(j)} \mathbf{X}^{(j)\top} + \epsilon I_{\mathfrak{d}} \right) \tag{4.3}$$

The computational cost for this recursion formula is moderate, since the mean $\bar{\mathbf{X}}_j =$

$\frac{1}{j}\sum_{s=1}^{j}\mathbf{X}^{(s)}$ also follows an obvious recursion formula. The canonical choice for the scaling parameter is $\gamma_{\mathfrak{d}} = (2.4)^2/\mathfrak{d}$, adopted from Gelman *et al.* [1996b].

Theorem 4.2. *The chain produced by the Adaptive Metropolis algorithm as described above is ergodic for p and thus provides samples from the target distribution.*

The proof of this theorem can be found in the original publication (Haario *et al.* [2001]). It is quite technical and not straightforward since the adaptive algorithm yields strictly speaking not a first order Markov chain, since the sampling depends on the previous samples.

The algorithm is furthermore implemented in such a way that it also adaptively tunes the acceptance probability to be within a desirable range. A basic pseudo-code implementation is provided in Algorithm 2.

Example 4.1 (Dimerization continued). We illustrate the Adaptive Metropolis algorithm by showing some sampling results, again from the dimerization example 2.6. We draw 10000 samples and use an appropriately scaled identity matrix as initial covariance matrix. We find an acceptance rate of 38%. Figures 4.1(a) and 4.1(b) show the output of the algorithm for the two parameters ξ and σ, respectively. Figures 4.1(c) and 4.1(d) show the corresponding histograms for ξ and σ, respectively. From the histogram, it is straightforward to derive sample-based 95% credible intervals for both parameters. We find $[3.6, 15.9]$ for ξ and $[0.049, 0.089]$ for σ. The upper boundary for ξ is larger than for the profile likelihood based confidence intervals derived in Example 3.2. This is due to the different natures of the cutoff.

4.3 Parallel Hierarchical Sampling

As will be seen in Chapter 8, single-chain sampling algorithms may perform insufficiently in high-dimensional settings. Thus sampling algorithms using multiple chains (Gelman & Rubin [1992]) can be used in order to achieve better mixing properties of the chain. This is often necessary because one single chain would have to be run for an impractically long time in order to be sure to have captured the entire mass of the posterior. This is particularly true for high-dimensional parameter spaces of over 100 parameters such as the one in Chapter 8.

Different varieties of multi-chain methods have been proposed in the literature such as

Algorithm 2: The Adaptive Metropolis algorithm

input : Initial sample $\mathbf{X}^{(0)}$, sampling distribution $p(x)$, desired number of samples J, initial proposal covariance matrix C_0, scaling parameter $\gamma_{\mathfrak{d}}$, small $\epsilon > 0$, index s_0 for starting the adaption

output: Samples $\mathbf{X}^{(j)}, j = 0, \ldots, J$ from $p(x)$

for $j \leftarrow 0$ **to** J **do**
  if $j \leq s_0$ **then**
    Generate a proposal $\mathbf{X}^* \sim \mathcal{N}(\mathbf{0}, C_0)$ *and draw* $u \sim \mathcal{U}[0,1]$*;*
    if $u \leq \alpha(\mathbf{X}^{(j)}, \mathbf{X}^*)$ **then**
      Set $\mathbf{X}^{(j+1)} = \mathbf{X}^*$*;*
    else
      $\mathbf{X}^{(j+1)} = \mathbf{X}^{(j)}$*;*
    end
    Set $C_{j+1} = C_0$*;*
  else
    Generate a proposal $\mathbf{X}^* \sim \mathcal{N}(\mathbf{0}, C_j)$ *and draw* $u \sim \mathcal{U}[0,1]$*;*
    if $u \leq \alpha(\mathbf{X}^{(j)}, \mathbf{X}^*)$ **then**
      Set $\mathbf{X}^{(j+1)} = \mathbf{X}^*$*;*
    else
      $\mathbf{X}^{(j+1)} = \mathbf{X}^{(j)}$*;*
    end
    Update $C_{j+1} = \frac{j-1}{j} C_j + \frac{\gamma_{\mathfrak{d}}}{j} \left(j \bar{\mathbf{X}}_{j-1} \bar{\mathbf{X}}_{j-1}^\top - (j+1) \bar{\mathbf{X}}_j \bar{\mathbf{X}}_j^\top + \mathbf{X}^{(j)} \mathbf{X}^{(j)\top} + \epsilon I_{\mathfrak{d}} \right)$*;*
  end
end

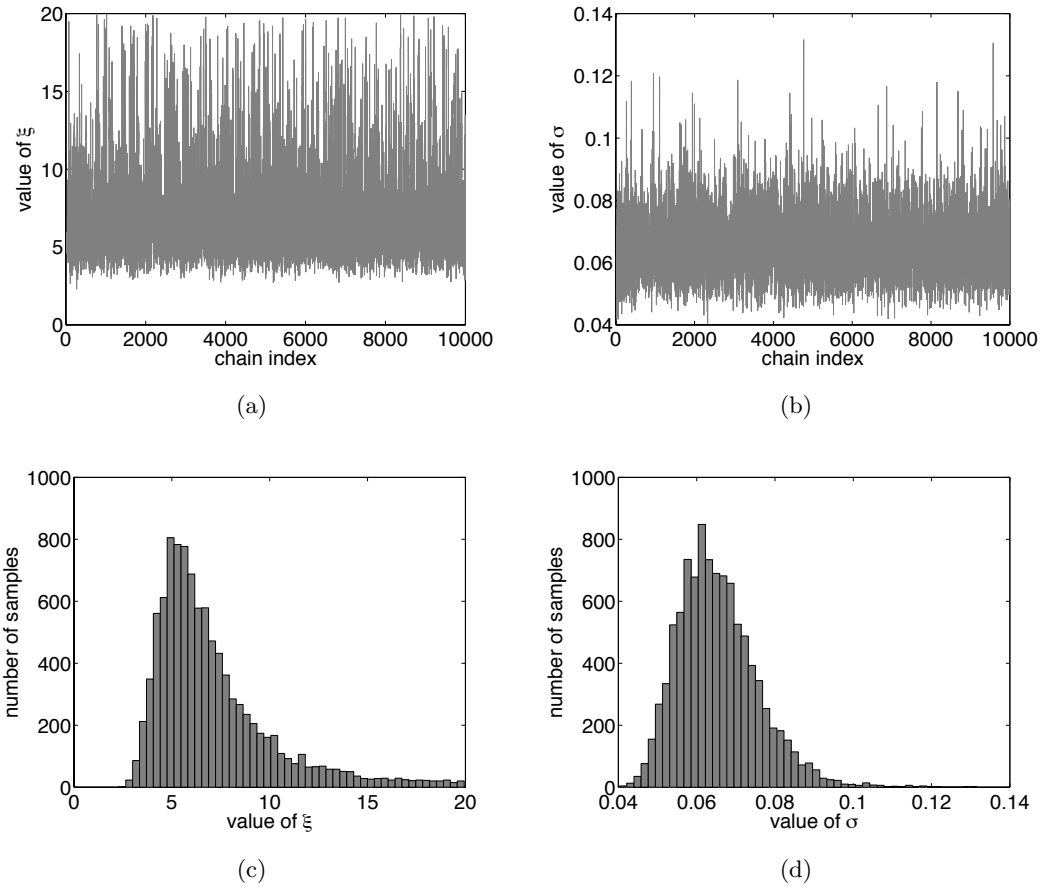
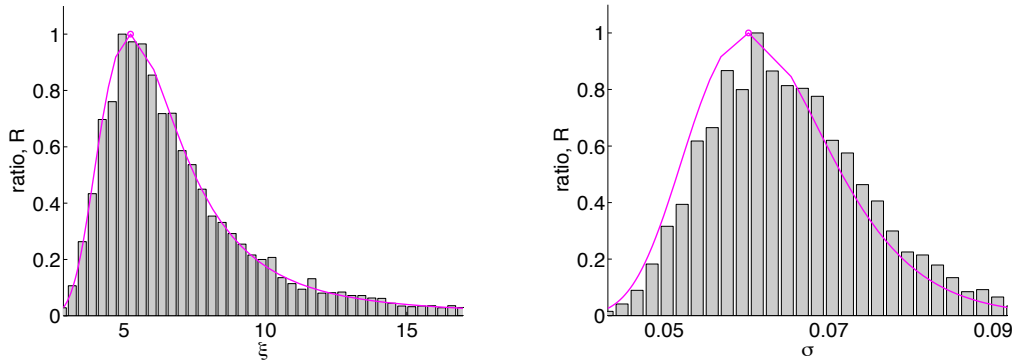


Figure 4.1: Chains and histograms. (a) The chain for parameter ξ. (b) The chain for parameter σ. (c) The histogram for parameter ξ. (d) The histogram for parameter σ.

parallel tempering (Neal [1996]), exchange Monte Carlo (Hukushima & Nemoto [1996]) or population-based reversible jump MCMC (Jasra *et al.* [2007]). While these methods are also advocated for closely related problems, we applied **Parallel Hierarchical Sampling** (PHS) from Rigat & Mira [2012]. Tempering approaches suffer from the fact that the finiteness of the chain-specific temperature dependent normalizing constant is difficult to check. Even if the tempered distributions are proper, Woodard *et al.* [2009] have shown that their modes tend to be narrow and mixing is numbed.

These difficulties are overcome in Parallel Hierarchical Sampling in the following way: several MCMC chains are run in parallel, each chain with any choice of single chain MCMC algorithm, all in all $\mathfrak{M}$ ones. One chain is selected to be the mother chain, while the others are auxiliary chains. At each iteration, one auxiliary chain is randomly



(a) (b)

Figure 4.2: Comparison of histograms and profiles. (a) Histogram and profile likelihood for ξ. (b) Histogram and profile likelihood for σ. (a-b) Magenta lines are the profile likelihoods, grey bars the histogram for the Adaptive Metropolis output. Both profile and histogram are scaled to obtain a maximum of 1. The agreement between both approaches for this simple example is especially good for ξ.

chosen and its state is swapped with the mother chain. This move is always accepted. All other auxiliary chains run a regular step in their chosen single chain sampling procedure. More formally, this can be seen in Algorithm 3. We use the notation $\mathbf{X}^{(j,\mathfrak{i})}$ for the j-th sample in the chain with index $\mathfrak{i}$.

Algorithm 3: Parallel Hierarchical Sampling

input : number of chains $\mathfrak{M}$, initial sample $\mathbf{X}^{(0,\mathfrak{i})}, \mathfrak{i} = 1, \ldots, \mathfrak{M}$, sampling distribution $p(x)$, desired number of samples J, individual proposals $q_{\mathfrak{i}}(\cdot|\cdot)$ for $\mathfrak{i} = 2, \ldots, \mathfrak{M}$, symmetric chain-swap proposal distribution $q'_{\mathfrak{s}}(\cdot|\cdot)$

output: Samples $\mathbf{X}^{(j,1)}, j = 0, \ldots, J$ from the mother chain for $p(x)$

for $j \leftarrow 0$ **to** J **do**

- Randomly select an index $\mathfrak{m}_{j+1} \in \{2, \ldots, \mathfrak{M}\}$ from the symmetric proposal distribution $q'_{\mathfrak{s}}(\mathfrak{m}_{j+1}|\mathfrak{m}_j)$ and swap: $\mathbf{X}^{(j+1,1)} = \mathbf{X}^{(j,\mathfrak{m}_{j+1})}$ and $\mathbf{X}^{(j+1,\mathfrak{m}_{j+1})} = \mathbf{X}^{(j,1)}$;
- Update chains $\mathfrak{i} = 2, \ldots, \mathfrak{m}_{j+1} - 1, \mathfrak{m}_{j+1} + 1, \ldots, \mathfrak{M}$ all targeting $p(x)$ according to their chain specific proposal distribution $q_{\mathfrak{i}}\left(\cdot|\mathbf{X}^{(j,\mathfrak{i})}\right)$;

end

As opposed to a parallel tempering approach, the chains all target the posterior distribution, but differ by their starting points and proposal distributions. This strategy fully exploits cross-chain swap transitions to maximize the mixing in the mother chain.

One significant advantage of this sampling scheme is that there is no need to find a single optimal proposal distribution. Every auxiliary chain can and should have a different proposal distribution. Usually it should prove beneficial to have a few chains which propose "large jumps" in the parameter space for quick parameter space traversals as well as a few chains with more local proposals for an adequate acceptance rate. The mother chain plays a prominent role, hence the term "hierarchical" in the name of the algorithm. The use of different proposal distributions in the individual auxiliary chains is what makes the algorithm especially suited to high-dimensional inference, since there finding a single optimal proposal scheme would be especially difficult.

Furthermore the PHS algorithm separates between local mixing through the within-chain updates of the auxiliary chains and global mixing through the swaps with the mother chain. Thus unlike for parallel tempering algorithms, the competition between local and global mixing is minimized.

Another important aspect in the context of this thesis is the fact that the PHS scheme does not require the knowledge of the marginal likelihood as normalization constant and is thus suited for inferring posterior distributions.

This can also directly be seen from the transition kernel, which takes for a Metropolis-Hastings update accepted according to Equation (4.1) in the individual auxiliary chains the form

$$\begin{aligned} & k_{\mathrm{PHS}}\left(\mathbf{X}^{(j)}, \mathbf{X}^{(j+1)}\right) \\ & = \sum_{\mathfrak{m}_{j+1}}^{\mathfrak{M}} q'_{\mathfrak{s}}(\mathfrak{m}_{j+1}|\mathfrak{m}_j) \prod_{\mathfrak{i}=2,\mathfrak{i}\neq\mathfrak{m}_{j+1}}^{\mathfrak{M}} q_{\mathfrak{i}}\left(\mathbf{X}^{(j+1,\mathfrak{i})}|\mathbf{X}^{(j,\mathfrak{i})}\right) \cdot \alpha_{\mathfrak{i}}\left(\mathbf{X}^{(j,\mathfrak{i})}, \mathbf{X}^{(j+1,\mathfrak{i})}\right) \end{aligned} \tag{4.4}$$

Theorem 4.3. *Let the auxiliary chains* $\{2, \ldots, \mathfrak{M}\}$ *be irreducible, aperiodic and reversible with respect to* $p(x)$ *and let* $q'_{\mathfrak{s}}(\cdot)$ *be a symmetric proposal distribution allowing for swaps between chain 1 and any of the other chains. Then the PHS joint transition kernel satisfies detailed balance with respect to the joint distribution* $\pi(\mathbf{X}^1, \ldots, \mathbf{X}^{\mathfrak{M}}) = \prod_{\mathfrak{i}=1}^{\mathfrak{M}} p(\mathbf{X}^{\mathfrak{i}})$.

Rigat & Mira provide a proof for this theorem, which can be directly derived from the form of the transition kernel. For this proof, it is only necessary that each of the single chain transition kernels has the correct ergodic properties.

4.4 Adaptive Metropolis Parallel Hierarchical Sampling

The parallel hierarchical sampling algorithm is designed to work with every available single chain MCMC algorithm. Since the AM algorithm is usually superior to the standard Metropolis-Hastings algorithm in our experience, we advocate the use of AM for the single chains. We call this novel variant **Adaptive Metropolis Parallel Hierarchical Sampling** (AMPHS).

Each auxiliary chain runs its own Adaptive Metropolis, such that the covariance matrix for the proposal distribution is adapted to each chain individually. This ensures a good sampling performance in each chain, while at the same time having excellent mixing properties for the mother chain.

Algorithm 4: Adaptive Metropolis Parallel Hierarchical Sampling

input : number of chains $\mathfrak{M}$, initial samples $\mathbf{X}^{(0,\mathfrak{i})}, \mathfrak{i} = 1, \ldots, \mathfrak{M}$, sampling distribution $p(x)$, desired number of samples J, initial proposal covariance matrices $C_0^{\mathfrak{i}}$, scaling parameters $\gamma_0^{\mathfrak{i}}$, small $\epsilon^{\mathfrak{i}} > 0$, indices $s_0^{\mathfrak{i}}$ for starting the adaption for $\mathfrak{i} = 2, \ldots, \mathfrak{M}$, symmetric chain-swap proposal distribution $q'_{\mathfrak{s}}(\cdot|\cdot)$

output: Samples $\mathbf{X}^{(j,1)}, j = 0, \ldots, J$ from the mother chain for $p(x)$

for $j \leftarrow 0$ **to** J **do**

- Randomly select an index $\mathfrak{m}_{j+1} \in \{2, \ldots, \mathfrak{M}\}$ from the symmetric proposal distribution $q'_{\mathfrak{s}}(\mathfrak{m}_{j+1}|\mathfrak{m}_j)$ and swap: $\mathbf{X}^{(j+1,1)} = \mathbf{X}^{(j,\mathfrak{m}_{j+1})}$ and $\mathbf{X}^{(j+1,\mathfrak{m}_{j+1})} = \mathbf{X}^{(j,1)}$;
- Update chains $\mathfrak{i} = 2, \ldots, \mathfrak{m}_{j+1} - 1, \mathfrak{m}_{j+1} + 1, \ldots, \mathfrak{M}$ all targeting $p(x)$ according to their chain specific proposal distribution as defined in Algorithm 2;

end

We chose this approach since it is especially beneficial when the posterior is multimodal or more generally non-standard shaped, as it increases mixing between the modes, even though that naturally comes at higher computational costs. Furthermore, the use of several different proposal settings is especially beneficial when it is difficult or impossible to analytically derive an optimal proposal scaling. As pointed out by Rigat & Mira, a single proposal kernel may for example not be optimal for exploring a distribution with both very narrow and very wide peaks.

Corollary 4.1. *The AMPHS transition kernel is ergodic for the joint distribution* $\pi(\mathbf{X}^1, \ldots, \mathbf{X}^{\mathfrak{M}}) = \prod_{\mathfrak{i}=1}^{\mathfrak{M}} p(\mathbf{X}^{\mathfrak{i}})$.

This can be seen directly from the shape of the transition kernel, the considerations for PHS and the fact that the single AM transition kernels are ergodic, as proven in Haario *et al.* [2001].

Example 4.2 (Dimerization continued). We again return to the dimerization example 2.6. We use AMPHS for sampling the posterior distribution of the parameters in contrast to Section 4.2, where we used the Adaptive Metropolis algorithm. We now use $\mathfrak{M} = 5$ chains, i.e. one mother chain and four auxiliary chains. We choose identity matrices as initial covariance matrices and two different scaling factors, one for chains two and three and one for chains four and five. We find that the mother chain of the AMPHS shows excellent mixing, see Figure 4.3. While the mixing in the auxiliary chains is not as good as for the mother chain, also sampling in the auxiliary chains is acceptable, see Figure 4.4. The auxiliary chains still all target the posterior distribution. Nevertheless, AMPHS uses only the samples from the mother chain, as these are the ones with the best properties.

We can now also compute the effective sampling size (ESS) of the output as introduced in Section 2.2.2 and Schmidl [2012] and compare with the ESS achieved for the AM. For the AM output in Section 4.2, we find an ESS of 2597, meaning that about every fourth sample of the 10000 can be considered independent.

For AMPHS, we find an ESS of 4235 for the mother chain, meaning that almost every second sample in this very simple setting can be considered approximately independent. This is an excellent value. For the four auxiliary chains, we find ESS's of 1019, 958, 873 and 982, which is also in an acceptable range. We see that for the AMPHS, we get 1.5 times as many independent samples as with the AM in this simple example. The downside is the $\mathfrak{M} = 5$ times higher computational cost caused by the multi-chain approach. This higher cost is nevertheless worth it for complex, high-dimensional target distributions, see e.g. Chapter 8.

4.5 Copula-based independence Metropolis-Hastings

We have recently also introduced a copula based Markov chain algorithm (Schmidl *et al.* [2013a,b]). Copulas are widely applied in finance and ecology (Min & Czado [2010]; Salvadori [2007]), but not yet in systems biology. Mathematically, they are a concept from probability theory for assessing and sampling from multivariate distributions by capturing dependence structures.

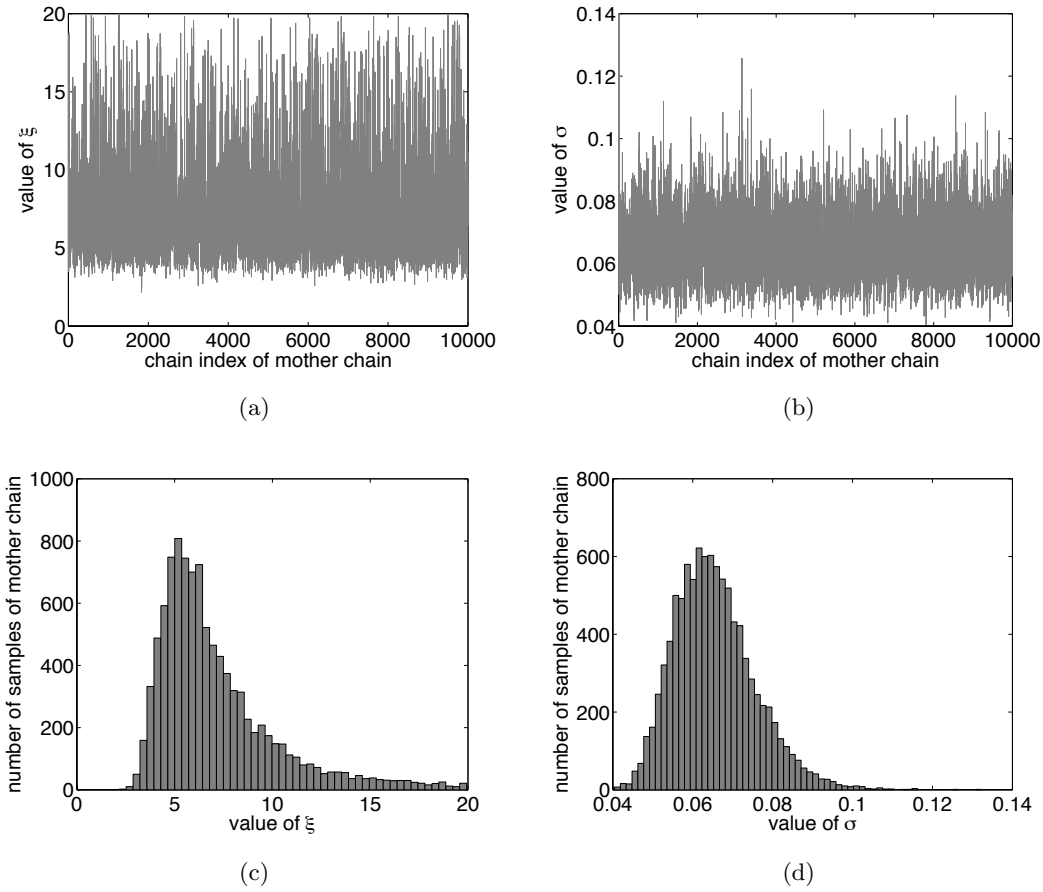


Figure 4.3: Chains and histograms for the mother chain of AMPHS. (a) The mother chain for parameter ξ. (b) The mother chain for parameter σ. (c) The histogram for parameter ξ. (d) The histogram for parameter σ.

We can decompose any absolutely continuous multivariate cumulative distribution function (cdf) $F(x_1, \ldots, x_\mathfrak{d})$ with marginal cdf's $F_i(x_i), i = 1, \ldots, \mathfrak{d}$, joint density function $f(x_1, \ldots, x_\mathfrak{d})$ and marginal density functions $f_i(x_i), i = 1, \ldots, \mathfrak{d}$ into

$$f(x_1, \ldots, x_\mathfrak{d}) = c\left(F_1(x_1), \ldots, F_d(x_\mathfrak{d})\right) \cdot f_1(x_1) \cdot \ldots \cdot f_\mathfrak{d}(x_\mathfrak{d}), \tag{4.5}$$

where $c(u_1, \ldots, u_\mathfrak{d})$ is a unique copula density function defined on $[0,1]^\mathfrak{d}$ with values in $[0,1]$ and uniformly distributed marginals on $[0,1]$. The existence of such a copula is guaranteed by Sklar's theorem (Nelsen [2006]). This copula function covers the full dependency structure of the variables. In other words, every joint distribution function can be decomposed into the marginal behavior of its individual variables and a function covering its dependency structure (Kurowicka & Joe [2011]). The MH pro-

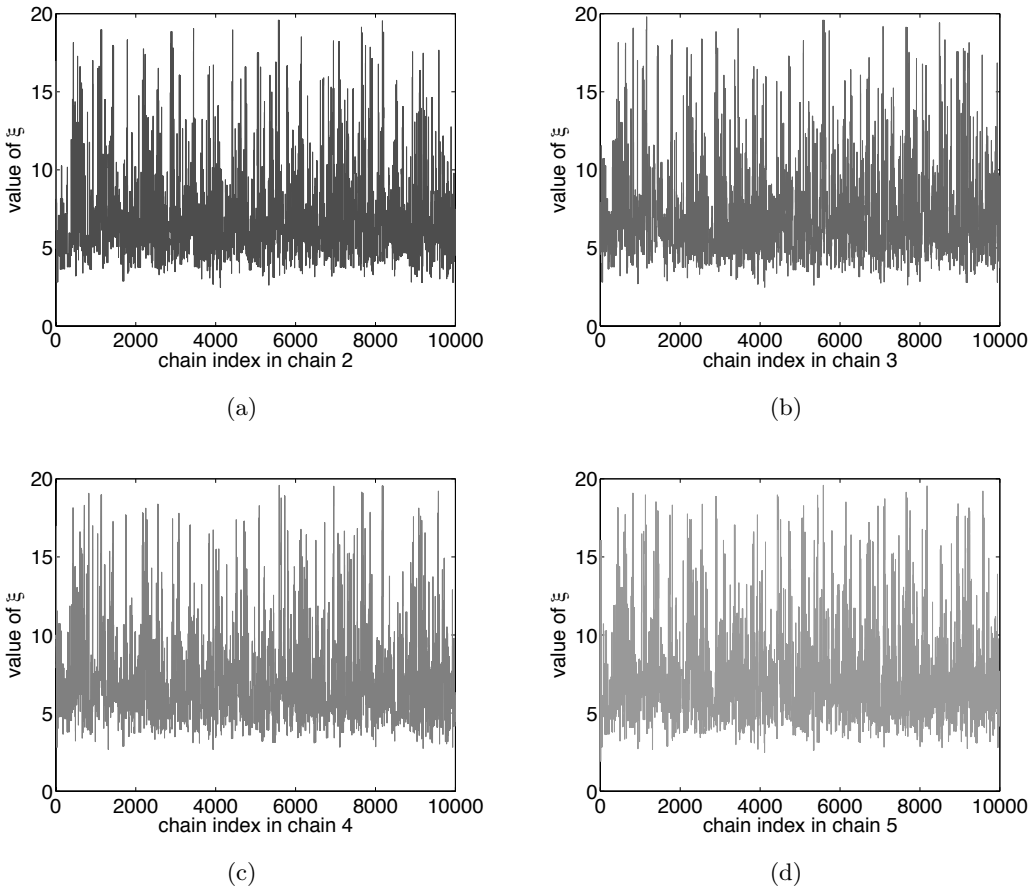


Figure 4.4: Auxiliary chains of AMPHS. (a) Auxiliary chain 2 for parameter ξ. (b) Auxiliary chain 3 for parameter ξ. (c) Auxiliary chain 4 for parameter ξ. (d) Auxiliary chain 5 for parameter ξ.

posal function then generates problem specific proposals with an according dependence structure drawn from the copula. The copula is a function on the $\mathfrak{d}$-dimensional unit cube, but it can be efficiently decomposed using so called vines into combinations of two-dimensional pair copula distributions, see e.g. Min & Czado [2010] for details. The copula corresponding to the target distribution that we want to sample has to be inferred. Inference in practice can be done on the basis of preruns from standard Metropolis-Hastings sampling runs. Schmidl *et al.* [2013a] provide details on how the copula can then be estimated and how samples are then generated.

The copula based MH approach is especially suited to deal with the dependence structure in a medium-dimensional sampling space of ca. 10-20 dimensions and allows for

high proposal acceptance rates at simultaneously high ESS's.

4.6 Convergence diagnostics

Sampling based algorithms are random by their nature. This results in the fact that convergence is very hard to prove. Markov chain theory has provided detailed convergence properties that sampling algorithms should fulfill, such as for example ergodicity.

In any case, convergence of the sample distribution to the target distribution is a theoretical consideration for an infinite number of samples. In practice, it is however only possible to look at a finite realization of the sampling algorithm. This means that we can obtain samples whose distribution converges to the target distribution, but the individual samples are in general never independently distributed and are identically distributed only in the limit.

Most convergence diagnostics thus aim for providing proof that the sample distribution has not yet converged in distribution to the target distribution. In recent years, there were quite a few reviews on the topic of different convergence statistics (Brooks & Roberts [1998]; Cowles & Carlin [1996]; Mengersen *et al.* [1999]).

In our applications, the target distribution for the sampling algorithms is usually the posterior distribution of parameters given measurement data and a model as introduced in Chapter 3. Unless stated otherwise, we started all sampling procedures in the maximum a posteriori estimates obtained prior to sampling by optimization. For the auxiliary chains in PHS and AMPHS, we sampled initial values randomly from the prior distribution, and then let an optimization algorithm run in order to start in a region with substantial posterior values. This overall strategy minimizes the influence of the starting point, which might in some cases be quite severe and should thus improve convergence to the target distribution.

As pointed out by Geyer [1992], thinning the chain would increase the variance, thus we usually use all samples from the Markov chain, except for a burn-in period. This burn-in period is often necessary in spite of starting in the MAP estimate, naturally especially when using adaptive sampling schemes. Furthermore, the MAP estimate might lie at the boundary of the parameter space, which also hinders convergence of the chain even though we start in a region of high posterior density. In real applications, it is often very useful to inspect the Markov chains visually. Bad mixing properties or a large burn-in period are often most easily detected that way. In the case of bad

mixing, the proposal distribution or the parameters of the sampling algorithm have to be tuned. The burn-in period is often chosen after inspecting the chain. It can then be verified by the Geweke test (Geweke [1992]).

The Geweke test on a realization of a one-dimensional Markov chain $\{\mathbf{X}^{(j)}\}_{j\in I}$ works by splitting the chain into two subsamples. In practice, it is most common to use the first 10% and the last 50% of samples. Failure of convergence can be detected if the mean of the two subsamples is very different. For verifying this, a z-score can be derived and from the z-score also a p-value. The z-score is calculated by

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\hat{\sigma_1} + \hat{\sigma}_2}}. \tag{4.6}$$

Here, $\bar{X}_1$ is the empirical mean of the first subsample and $\bar{X}_2$ the mean of the second subsample. Furthermore, $\hat{\sigma_1}$ and $\hat{\sigma}_2$ are the empirical standard deviations of the first and second subsample, respectively. This z-score can be used as the test statistic for calculating a p-value by

$$p_z = 2(1 - \Phi(z; 0, 1)), \tag{4.7}$$

where $\Phi(z; \mu, \sigma)$ is the cdf of the univariate normal distribution, evaluated at z with mean $\mu = 0$ and standard deviation $\sigma = 1$. For higher-dimensional Markov chains, the same procedure is applied to each dimension separately.

We chose to apply the Geweke test because of its easy applicability and low computational cost. If several Markov chains for the same target distribution are available, we propose to use the Gelman-Rubin statistic $\hat{R}$ (Brooks & Gelman [1998]; Gelman & Rubin [1992]). This statistic compares the variances between the different Markov chains with the variance within each chain. Again, if the chains are all stationary, the two variances should be very similar. Let thus L be the number of realizations $\{\mathbf{x}_l^{(j)}\}_{j\in J}$ of the Markov chain $\{\mathbf{X}^{(j)}\}_{j\in\mathbb{N}}$. We assume that these chains start at different initial values $\mathbf{x}_l^{(0)}, l = 1, \ldots, L$. Then, depending on the number of samples m, we define the between-chain variance B as

$$B(m) = \frac{J - m}{L - 1} \sum_{l=1}^{L} \left(\bar{\mathbf{x}}_l(m) - \bar{\mathbf{x}}(m)\right)^2. \tag{4.8}$$

For this, we need the mean within a chain

$$\bar{\mathbf{x}}_l(m) = \frac{1}{J - m} \sum_{j=m+1}^{J} \mathbf{x}_l^{(j)} \tag{4.9}$$

and the overall mean

$$\bar{\mathbf{x}}(m) = \frac{1}{L}\sum_{l=1}^{L} \bar{\mathbf{x}}_l(m). \tag{4.10}$$

Furthermore, we define the within-chain variance W for index of samples m as

$$W(m) = \frac{1}{L}\sum_{l=1}^{L} \hat{\sigma}_l^2(m). \tag{4.11}$$

This definition includes the empirical standard deviation in the chain

$$\hat{\sigma}_l^2(m) = \frac{1}{J-m-1}\sum_{j=m+1}^{J} \left(\mathbf{x}_l^{(j)} - \bar{\mathbf{x}}_l\right)^2. \tag{4.12}$$

The Gelman-Rubin statistic $\hat{R}$ now compares the two variances by

$$\hat{R}(m) = \sqrt{\frac{\hat{\sigma}(m)}{W(m)}}, \tag{4.13}$$

where

$$\hat{\sigma}(m) = \left(1 - \frac{1}{J-m}\right) W(m) + \frac{1}{J-m} B(m) \tag{4.14}$$

The statistic $\hat{R}$ is also called potential scale reduction factor. It can be interpreted as a convergence diagnostic. If it is large, this suggests one of two cases: either the estimate of the variance $\hat{\sigma}(m)$ can be further decreased with larger m (further simulations), or secondly that larger m will result in a larger $W(m)$, as the chains have not yet explored the target distribution fully. If the statistic $\hat{R}$ is however close to 1, it is generally accepted that the L Markov chains indeed are stationary. In practice, often a cut-off value of 1.2 for $\hat{R}$ is used. It is also possible, but not overly common to apply the Gelman-Rubin statistic to a single chain by splitting it into a number of subsamples of equal length. In general, we follow the recommendation of Geyer [1992] that one long Markov chain should be preferred over several shorter Markov chains with the overall same number of samples, as shorter chains might not have reached convergence during the sampling process. However, in some applications, a single Markov chain is too slow at exploring the parameter space, in which case we advocate the use of parallel hierarchical sampling.

Example 4.3 (Dimerization continued). To further illustrate the importance of assessing the convergence of the sampling, we provide three illustrative examples, again from the dimerization example 2.6. The first example shows a chain for the parameter ξ with very good properties, see Figure 4.5(a), this is the same figure as Figure 4.1(a), put again for

comparison. The Adaptive Metropolis algorithm here has an acceptance rate of 38%. This yields a p-value of 0.94 of Geweke's test. In Figure 4.5(b), we see a Markov chain from the Metropolis-Hastings algorithm with very low acceptance rate of 1.2%. This leads to not-so-good mixing and inefficient exploration of the parameter space. Note that nevertheless the p-value for Geweke's test is 0.98, which would indicate convergence if the acceptance rate would not indicate otherwise. In the third panel (Figure 4.5(c)), we see a Markov chain that is obviously not yet stationary, also indicated by a p-value of 0.0056 from Geweke's test. A high acceptance rate of 69% is due to a too small step size in the proposal distribution.

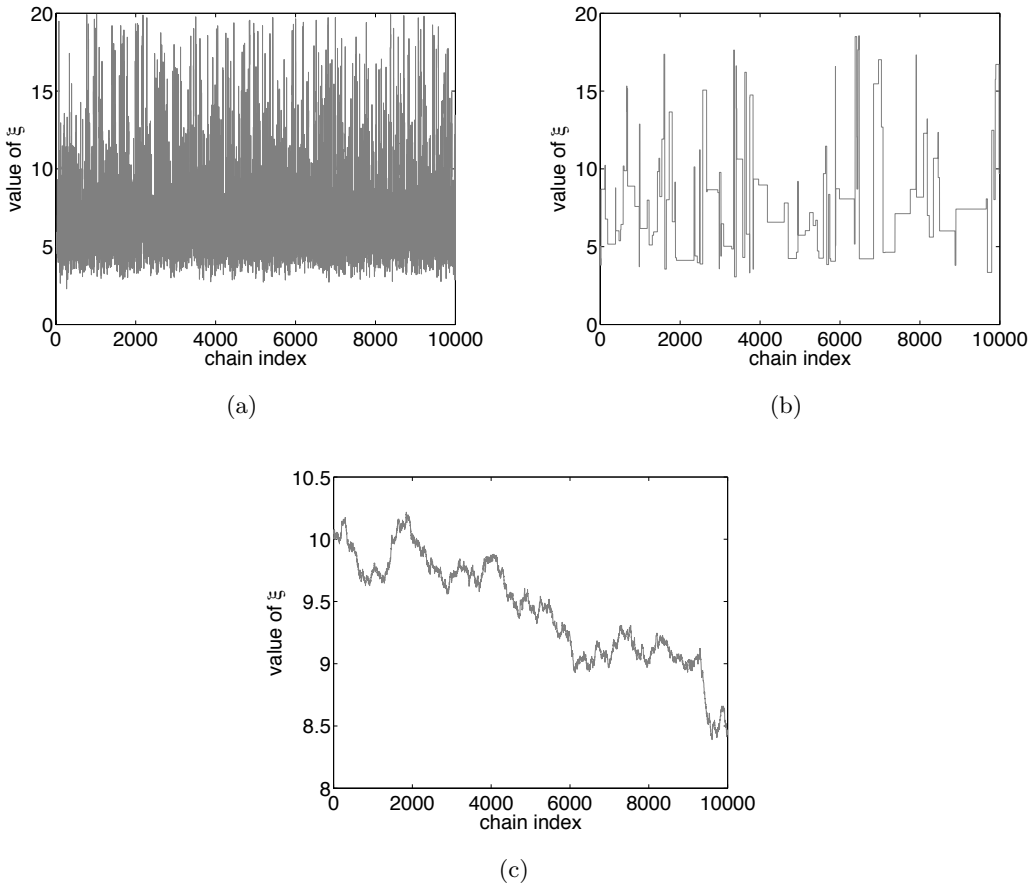


Figure 4.5: Markov chains. (a) A Markov chain with good mixing, good acceptance rate and high p-value for Geweke's test. (b) A Markov chain with not-so-good mixing, low acceptance rate, but high p-value for Geweke's test. (c) A Markov chain in the transient phase, with high acceptance rate and low p-value for Geweke's test.

In this chapter, we introduced well-known and novel MCMC algorithms and how to assess their convergence properties. All introduced algorithms will be applied in the applications part of this thesis. The fact that each application example uses a different sampling algorithm shows well that Bayesian methods always have to be tailored to their target.

5

Model selection methods

Make everything as simple as possible,
but not simpler.

Albert Einstein

In this chapter, we present established model selection methods and their newly developed improvements that can be used to choose the best available model from a finite candidate set, according to the selected method and on the basis of measurement data.

Among the presented indicators, the Bayes factor is often preferred, as it accounts for uncertainty of parameters and intrinsically prevents over-fitting. However, the Bayes factor is the ratio of two potentially high-dimensional integrals, the marginal likelihoods of the models, and thus not easy to compute. A sophisticated method for evaluating marginal likelihoods is thermodynamic integration, presented in Section 5.3.

In this chapter, we introduce a new adaptive variant for thermodynamic integration, which computes marginal likelihoods more efficiently and with controlled accuracy. For this, we apply the adaptive Simpson's rule. This is a novel contribution.

This chapter is in parts based on or even identical with the following two publications:

- **S. Hug**, M. Schwarzfischer, J. Hasenauer, C. Marr and F.J. Theis. An adaptive method for calculating Bayes factors using Simpson's rule, *in revision*
- **S. Hug**, D. Schmidl, W.B. Li, M.B. Greiter and F.J. Theis. Uncertainty in Biology: a computational modeling approach, chapter Bayesian model selection methods and their application to biological ODE systems, *in revision*

The need for model selection often arises naturally when modeling biological systems as the structure of the system itself is uncertain. For example, this could mean that it is not known if a reaction is actually present or not or if the steady state of the system is zero or non-zero. Such competing hypotheses can be formulated as individual ODE models. Typically, we then want to select the model from a candidate set that best fits the measurement data.

The models that we deal with in this thesis are always parametrized, i.e. the exact shape of the solution of a model depends on parameters. To check if a model fits the measured data, the agreement between model and data has to be optimized by adjusting the parameters. In our ordinary differential equation models, these parameters are for example the rate constants in the ODEs or initial conditions. The shape of the ODE solution will change depending on these parameters.

All model selection methods presented in the following take into account how good the fit of the model to the data is. Furthermore, they consider the number of parameters a model contains, or even the full parameter distributions. We now first present established methods before we come to our novel adaptive variant. Finally, all methods are evaluated on an analytically tractable example.

Also all methods presented in this chapter are implemented in MATLAB.

5.1 Likelihood based model selection methods

Quite a few rather easily accessible ways of doing model selection are based on the maximum likelihood estimates (MLEs) of the parameters for each model. As introduced in Chapter 3, the likelihood $p(\mathbf{Y}|\boldsymbol{\theta}^i, \mathbf{M}_i)$, now conditioned explicitly on a model $\mathbf{M}_i$, is a measure for the agreement between data $\mathbf{Y}$ and model $\mathbf{M}_i$ from a candidate set $\mathbf{M}_\nu, \nu = 1, \ldots, \mathcal{I}$ parametrized with parameters $\boldsymbol{\theta}^i$. The parameter vector which maximizes the likelihood is called the **maximum likelihood estimate**, often written as $\hat{\boldsymbol{\theta}}^i$. It can

be seen as the single best point estimate. To find the maximum likelihood estimate, optimization techniques have to be applied. In high-dimensional models, this task is far from trivial. We find that a strategy which combines a local optimizer with multiple restarts performs best in our applications. While the optimization is surely very important for the overall performance, it is not the special focus of our work, interested readers should refer e.g. to Raue *et al.* [2013b]. We now briefly present several methods that we applied to our specific problems. We focus on an overview in the style of Kirk *et al.* [2013], for more detailed mathematical derivations of the methods, we refer our readers to the provided references.

5.1.1 Akaike and Bayesian information criteria

Based on the MLE, several model selection criteria or tests have been proposed. Best known among them might be the Akaike Information Criterion (AIC) (Akaike [1973, 1974]). The theoretical basis comes from information theory and is based on the loss of information measured by the Kullback-Leibler divergence. The AIC is defined as

$$AIC(\hat{\boldsymbol{\theta}}^i) = -2\log p(\mathbf{Y}|\hat{\boldsymbol{\theta}}^i, \mathbf{M}_i) + 2\mathfrak{d}_i, \tag{5.1}$$

where $\mathfrak{d}_i$ is the number of independently adjusted parameters of model $\mathbf{M}_i$, meaning $\boldsymbol{\theta}^i \in \mathbb{R}^{\mathfrak{d}_i}$. The preferred model is the one with the minimal value for the AIC. The AIC weighs the goodness of fit, given by the log-likelihood value, with the associated number of parameters, preferring smaller models over large models. However, the AIC does not give information about the quality of the models in an absolute sense, even the best model might not fit the data at all, just better than the other models, see also Section 2.4.3. The value of the AIC does not directly give any information about a poor fit of all the models. The estimate is only valid asymptotically in the limit of large numbers of data points.

Somehow closely related is the Bayesian Information Criterion (BIC) (Schwarz [1978]). In contrast to the AIC, it also takes into account the number of data points on which the choice is based:

$$BIC(\hat{\boldsymbol{\theta}}^i) = -2\log p(\mathbf{Y}|\hat{\boldsymbol{\theta}}^i, \mathbf{M}_i) + \mathfrak{d}_i \log(N). \tag{5.2}$$

Here, N is the number of data points in $\mathbf{Y}$. Again, the model with the lowest BIC value should be chosen.

Neither of the two criteria gives an absolute measure of how much one model is "better" than another model.

Yang [2005] has pointed out several more interesting considerations: under the assumption that the exactly true model is not in the candidate set, the AIC is asymptotically optimal in selecting the model with the least mean square error, the BIC not. However, the AIC is not consistent while the BIC is.

5.1.2 The likelihood ratio test

While the AIC and BIC are rather closely related, a very different, yet also MLE based model choice method between two models is the likelihood ratio test (LRT) (Kirk *et al.* [2013]). This method requires the models to be nested, meaning that the smaller of the models needs to be a special case of the larger model. The LRT is a hypothesis test with the null hypothesis that the smaller model (without loss of generality from now on model $\mathbf{M_1}$) is the true model that generated the data versus the alternative hypothesis that the larger model $\mathbf{M_2}$ generated the data. As the models are nested, the ratio of the logarithms of the maximum likelihood values is approximately χ^2-distributed, with degrees of freedom $\mathfrak{d}_1$ and $\mathfrak{d}_2$ corresponding to the numbers of parameters in the two models:

$$-2\log\left(\frac{p(\mathbf{Y}|\hat{\boldsymbol{\theta}}^1,\mathbf{M_1})}{p(\mathbf{Y}|\hat{\boldsymbol{\theta}}^2,\mathbf{M_2})}\right) \sim \chi^2_{\mathfrak{d}_2-\mathfrak{d}_1}. \tag{5.3}$$

For two nested models, the larger model always explains the data at least as well as the smaller model, thus $\frac{p(\mathbf{Y}|\hat{\boldsymbol{\theta}}^1,\mathbf{M}_1)}{p(\mathbf{Y}|\hat{\boldsymbol{\theta}}^2,\mathbf{M}_2)} < 1$. With the LRT, it is possible to determine if the improvement is significant by deriving a p-value under the appropriate χ^2-distribution. Classical hypothesis testing then reveals if the null model can be rejected at the desired significance level.

5.2 Bayesian model selection

5.2.1 The Bayes factor

The BIC (or Schwarz criterion) can be seen as a rough approximation to the logarithm of a different model selection criterion, the marginal likelihood used for calculating a **Bayes factor** (BF) (Kass & Raftery [1995]). The Bayes factor is derived from Bayes' theorem. Here, the likelihood $p(\mathbf{Y}|\boldsymbol{\theta})$ is complemented with prior information $p(\boldsymbol{\theta})$ available for the parameters to yield the general posterior distribution $p(\boldsymbol{\theta}|\mathbf{Y})$ of the

parameters given the data:

$$p(\boldsymbol{\theta}|\mathbf{Y}) = \frac{p(\mathbf{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{Y})} \tag{5.4}$$

An important quantity for the purpose of model selection is actually the marginal likelihood $p(\mathbf{Y})$ in the denominator of the posterior distribution.

With Bayes' theorem once again, we get:

$$p(\mathbf{M}_i|\mathbf{Y}) = \frac{p(\mathbf{Y}|\mathbf{M}_i)p(\mathbf{M}_i)}{\sum_{\nu=1}^{\mathcal{J}} p(\mathbf{Y}|\mathbf{M}_\nu)p(\mathbf{M}_\nu)}, \tag{5.5}$$

which is to compute the marginal likelihood $p(\mathbf{Y}|\mathbf{M}_i)$ for the desired model $\mathbf{M}_i$ from the candidate set $\mathbf{M}_\nu, \nu = 1, \ldots, \mathcal{J}$. It is important to notice that the marginal likelihood is not straightforward to compute, since it is a usually high-dimensional, analytically intractable integral:

$$p(\mathbf{Y}|\mathbf{M}_i) = \int_{\mathbb{R}^{\vartheta_i}} p(\mathbf{Y}|\boldsymbol{\theta}^i, \mathbf{M}_i)p(\boldsymbol{\theta}^i|\mathbf{M}_i)\,\mathrm{d}\boldsymbol{\theta}^i \tag{5.6}$$

This integral has to be approximated, usually with sampling based approaches. Nevertheless, if we then want to compare two models $\mathbf{M_1}$ and $\mathbf{M_2}$, we can do so by computing the ratio of the two marginal likelihoods, the so-called **Bayes factor**

$$B_{12} = \frac{p(\mathbf{Y}|\mathbf{M_1})}{p(\mathbf{Y}|\mathbf{M_2})}, \tag{5.7}$$

where a value of B_{12} greater than 1 indicates a preference for model $\mathbf{M_1}$. Analogously, a value less than 1 indicates a preference for model $\mathbf{M_2}$.

Harold Jeffreys established a widely used interpretation of the Bayes factor in Jeffreys [1961]. It is based on a classification of the evidence in favor of model $\mathbf{M_1}$ in log_{10}-half-scale units as:

$\log_{10}(B_{12})$	B_{12}	Evidence in favor of model $\mathbf{M_1}$
0 - 0.5	1 - 3.2	Not worth more than a bare mention
0.5 - 1	3.2 -10	Substantial
1.0 - 1.5	10 - 32.6	Strong
1.5 - 2.0	32.6 - 100	Very strong
2.0 - ∞	100 - ∞	Decisive

This has become known as Jeffreys' scale of evidence. While it certainly can be challenged, it nevertheless is well established and widely used in the Bayesian community.

The Bayes factor offers certain advantages over the presented point-based model selection methods. First, in contrast the likelihood ratio test, it provides evidence for either of the models, since the Bayes factor in favor of model $\mathbf{M_2}$ can easily be interpreted by the same Jeffreys' scale by taking $B_{21} = 1/B_{12}$. Secondly, it works for non-nested models. Thirdly, point-based methods might not be appropriate in cases where the MLE is not representative for the whole distribution, e.g. for multimodal likelihoods or in the presence of non-identifiabilities as introduced in Section 3.3. Furthermore, by taking into consideration the whole parameter space, the Bayes factor is more efficient in preventing overfitting (Myung & Pitt [1997]) than the other introduced methods.

As already mentioned, the crux is that the marginal likelihood is computed by integrating over the whole parameter space, which is computationally costly and also often not straightforward. Because of this, standard methods for computing Bayes factors are mostly sampling based. This can be seen e.g. from the following relationship:

$$p(\mathbf{Y}|\mathbf{M}_i) = \int_{\mathbb{R}^{\mathfrak{d}_i}} p(\mathbf{Y}|\boldsymbol{\theta}^i, \mathbf{M}_i) p(\boldsymbol{\theta}^i|\mathbf{M}_i)\, \mathrm{d}\boldsymbol{\theta}^i = \mathbb{E}_{p(\boldsymbol{\theta}^i|\mathbf{M}_i)}\left[p(\mathbf{Y}|\boldsymbol{\theta}^i, \mathbf{M}_i)\right], \tag{5.8}$$

as the approximation of expectations is a strength of MCMC algorithms.

5.2.2 The prior arithmetic mean estimate

The easiest approach for sampling any of the marginal likelihoods, here now simply denoted $p(\mathbf{Y}|\mathbf{M})$ without an explicit model index, is the prior arithmetic mean. For this approach, a total of J samples $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(J)}$ are drawn from the prior distribution $p(\boldsymbol{\theta})$. From Equation (5.8) it can then be inferred that

$$p(\mathbf{Y}|\mathbf{M}) = \mathbb{E}_{p(\boldsymbol{\theta})}\left[p(\mathbf{Y}|\boldsymbol{\theta}, \mathbf{M})\right] \approx \frac{1}{J}\sum_{j=1}^{J} p(\mathbf{Y}|\boldsymbol{\theta}^{(j)}, \mathbf{M}) \tag{5.9}$$

The right hand side of this equation is known as the prior arithmetic mean estimate. The strong law of large numbers guarantees (almost surely) convergence as the sample number tends to infinity. However, in many practical applications, the prior does not contain too much information about the actual shape of the posterior. Then many samples might have very low likelihood values, thus a large number of samples might be needed for accurate results.

5.2.3 The posterior harmonic mean estimate

Slightly more involved is the approach by Newton & Raftery [1994] called the posterior harmonic mean. As the name already implies, for this approach samples are not drawn from the prior, but from the posterior distribution directly. Similarly to the prior arithmetic mean, we draw a total of J samples $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(J)}$ from the posterior distribution $p(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{M})$. This then yields the following marginal likelihood approximation:

$$p(\mathbf{Y}|\mathbf{M}) \approx \left(\frac{1}{J} \sum_{j=1}^{J} \frac{1}{p(\mathbf{Y}|\boldsymbol{\theta}^{(j)}, \mathbf{M})} \right)^{-1} \tag{5.10}$$

The derivation can for example be found in Schmidl [2012]. However, already Neal [2008] showed that this estimate suffers from severe issues. Newton & Raftery [1994] also proposed a weighted combination of the prior arithmetic mean estimator and posterior harmonic mean estimator called the stabilized harmonic mean estimator. This helps to reduce the issues of the individual estimators.

5.2.4 Chib's method

Also often mentioned is Chib's method, which is originally also a point-based estimate. In Chib & Jeliazkov [2001], Chib and Jeliazkov show how to apply the method to the output of a Metropolis-Hastings sampling algorithm. The basic idea is to rearrange Bayes's theorem:

$$\log p(\mathbf{Y}|\mathbf{M}) = \log p(\mathbf{Y}|\boldsymbol{\theta}^*, \mathbf{M}) + \log p(\boldsymbol{\theta}^*|\mathbf{M}) - \log p(\boldsymbol{\theta}^*|\mathbf{Y}, \mathbf{M}) \tag{5.11}$$

with a suitable $\boldsymbol{\theta}^*$, for example the maximum likelihood estimate. While this might yield an easily computable result, it might suffer from the same issues as other point-based estimates. Furthermore, the posterior probability $p(\boldsymbol{\theta}^*|\mathbf{Y}, \mathbf{M})$ is often not readily available, since sampling and optimization are mostly based on "likelihood times prior" instead of the posterior, ignoring the proportionality constant that is actually the marginal likelihood at question here. For estimating the posterior value at the chosen point estimate, Chib and Jeliazkov propose to use the output of a Metropolis-Hastings sampler $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(J)}$.

If $q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{Y})$ denotes the proposal density of the Metropolis-Hastings algorithm for the transition from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$, where the proposal density is allowed to depend on the data

$\mathbf{Y}$, and $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{Y})$ denotes the standard Metropolis-Hastings acceptance probability of a move, then it can be shown that

$$\hat{p}(\boldsymbol{\theta}^*|\mathbf{Y}, \mathbf{M}) = \frac{S^{-1}\sum_{s=1}^{S}\alpha(\boldsymbol{\theta}^{(s)}, \boldsymbol{\theta}^*|\mathbf{Y})q(\boldsymbol{\theta}^{(s)}, \boldsymbol{\theta}^*|\mathbf{Y})}{J^{-1}\sum_{j=1}^{J}\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{(j)}|\mathbf{Y})} \tag{5.12}$$

is a simulation-consistent estimate of the posterior value. Here, the $\boldsymbol{\theta}^{(s)}$ are samples drawn from the posterior, while the $\boldsymbol{\theta}^{(j)}$ are drawn from $q(\boldsymbol{\theta}^*, \cdot|\mathbf{Y})$ with $\boldsymbol{\theta}^*$ fixed. This can then be plugged into Equation (5.11) to yield an estimate for the marginal likelihood.

5.3 Thermodynamic integration for the computation of Bayes factors

Nowadays, the preferred method for calculating marginal likelihoods is often **thermodynamic integration** (TI). While it is computationally costlier than the other methods, it yields more robust and numerically stable results, since its variance is well controlled and sometimes significantly smaller than for the other introduced methods (Gelman & Meng [1998]). Thermodynamic integration in a statistical context is based on path sampling ideas (Gelman & Meng [1998]), and was then discussed for marginal likelihoods in the papers by Lartillot & Philippe [2006] and Friel & Pettitt [2008]. It has recently found increasing application in systems biology (Calderhead & Girolami [2009]; Eydgahi *et al.* [2013]; Xu *et al.* [2010]).

Central to the method is the **power posterior**, a variant of the usual posterior of the Bayesian setting,

$$p_\tau(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{M}) = \frac{1}{p_\tau(\mathbf{Y}|\mathbf{M})}p(\mathbf{Y}|\boldsymbol{\theta}, \mathbf{M})^\tau p(\boldsymbol{\theta}|\mathbf{M}) \tag{5.13}$$

where $\tau \in [0, 1]$ is a so-called temperature parameter and the denominator $p_\tau(\mathbf{Y}|\mathbf{M}) = \int_{\mathbb{R}^9} p(\mathbf{Y}|\boldsymbol{\theta}, \mathbf{M})^\tau p(\boldsymbol{\theta}|\mathbf{M})d\boldsymbol{\theta}$ is a normalization term necessary for making the power posterior a probability density. For $\tau = 0$, we get $p_{\tau=0}(\mathbf{Y}|\mathbf{M}) = 1$, since this is the prior integrated over $\boldsymbol{\theta}$ and thus simply 1. The power posterior is then equal to the prior $p(\boldsymbol{\theta}|\mathbf{M})$. For $\tau = 1$, we get $p_{\tau=1}(\mathbf{Y}|\mathbf{M}) = p(\mathbf{Y}|\mathbf{M})$, the marginal likelihood, and thus the power posterior is the regular posterior. Intuitively, a power posterior with a low value of τ corresponds to a distribution closer to the prior, which is often smooth and thus allows for more movement of the Markov chains through the parameter space. A higher value of τ corresponds to a distribution closer to the posterior,

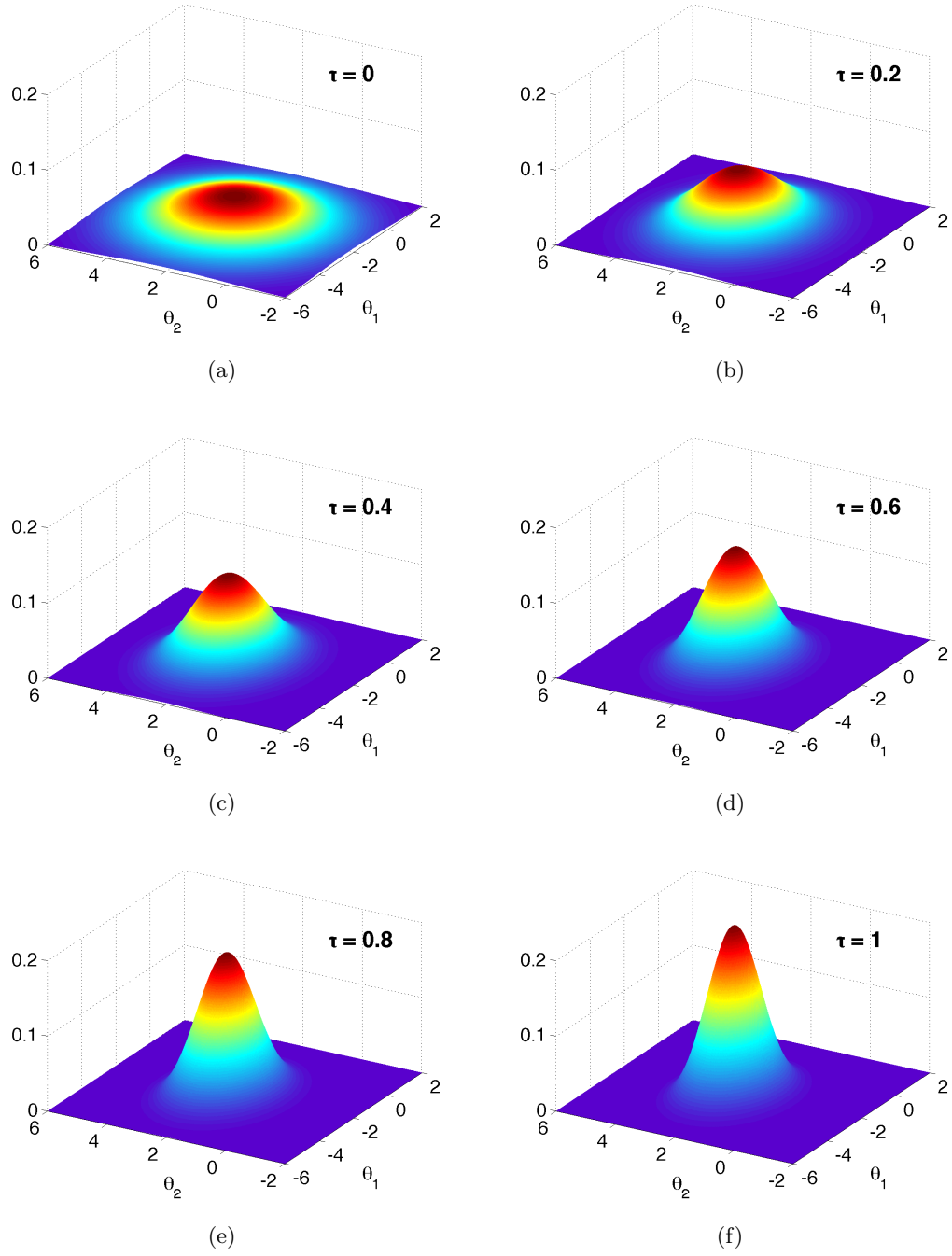


Figure 5.1: Power posteriors. Visualization of the smooth transition from prior to posterior through the power posterior. Shown is the power posterior of the two-parameter model $\mathbf{M_2}$ introduced in Section 5.4, for six different temperatures.

which might be e.g. spiky due to the influence of the likelihood. The power posterior in total thus corresponds to a smooth transition from the prior to the posterior, which can also be seen in Figure 5.1 with the example we will present later. We now derive an expression for the log marginal likelihood with respect to the power posterior which can be evaluated using MCMC methods. First we note that

$$\begin{aligned}
\frac{d}{d\tau}\log p_\tau(\mathbf{Y}|\mathbf{M}) &= \frac{d}{d\tau}\log\int_{\mathbb{R}^\mathfrak{d}} p(\mathbf{Y}|\boldsymbol{\theta},\mathbf{M})^\tau p(\boldsymbol{\theta}|\mathbf{M})d\boldsymbol{\theta} \\
&= \frac{1}{p_\tau(\mathbf{Y}|\mathbf{M})}\int_{\mathbb{R}^\mathfrak{d}} \frac{dp(\mathbf{Y}|\boldsymbol{\theta},\mathbf{M})^\tau}{d\tau} p(\boldsymbol{\theta}|\mathbf{M})d\boldsymbol{\theta} \\
&= \frac{1}{p_\tau(\mathbf{Y}|\mathbf{M})}\int_{\mathbb{R}^\mathfrak{d}} \frac{dp(\mathbf{Y}|\boldsymbol{\theta},\mathbf{M})^\tau}{d\tau}\frac{p(\mathbf{Y}|\boldsymbol{\theta},\mathbf{M})^\tau p(\boldsymbol{\theta}|\mathbf{M})}{p(\mathbf{Y}|\boldsymbol{\theta},\mathbf{M})^\tau} d\boldsymbol{\theta} \\
&= \int_{\mathbb{R}^\mathfrak{d}} \frac{d\log p(\mathbf{Y}|\boldsymbol{\theta},\mathbf{M})^\tau}{d\tau}\frac{p(\mathbf{Y}|\boldsymbol{\theta},\mathbf{M})^\tau p(\boldsymbol{\theta}|\mathbf{M})}{p_\tau(\mathbf{Y}|\mathbf{M})} d\boldsymbol{\theta} \\
&= \int_{\mathbb{R}^\mathfrak{d}} \log p(\mathbf{Y}|\boldsymbol{\theta},\mathbf{M})\frac{p(\mathbf{Y}|\boldsymbol{\theta},\mathbf{M})^\tau p(\boldsymbol{\theta}|\mathbf{M})}{p_\tau(\mathbf{Y}|\mathbf{M})} d\boldsymbol{\theta} \\
&= \mathbb{E}_{p_\tau}\{\log p(\mathbf{Y}|\boldsymbol{\theta},\mathbf{M})\}
\end{aligned} \tag{5.14}$$

Integrating both sides with respect to τ yields the **thermodynamic integral**,

$$\begin{aligned}
\int_0^1 \mathbb{E}_{p_\tau}\{\log p(\mathbf{Y}|\boldsymbol{\theta},\mathbf{M})\}d\tau &= \int_0^1 \frac{d}{d\tau}\log\ p_\tau(\mathbf{Y}|\mathbf{M}) \\
&= \log p_{\tau=1}(\mathbf{Y}|\mathbf{M}) - p_{\tau=0}(\mathbf{Y}|\mathbf{M}) \\
&= \log p(\mathbf{Y}|\mathbf{M})
\end{aligned} \tag{5.15}$$

The integrand on the left hand side of Equation (5.15) is also called the expected log deviance. The integral in Equation (5.15) can be solved numerically by choosing a discretization (or temperature schedule) $0 = \tau_0 < \tau_1 < ... < \tau_{K-1} < \tau_K = 1$, then the numerical approximation with the trapezoidal rule for quadrature is

$$\log p(\mathbf{Y}|\mathbf{M}) \approx \tfrac{1}{2}\sum_{k=0}^{K-1}(\tau_{k+1}-\tau_k)\big(\mathbb{E}_{p_{\tau_{k+1}}}\{\log p(\mathbf{Y}|\boldsymbol{\theta},\mathbf{M})\}+\mathbb{E}_{p_{\tau_k}}\{\log p(\mathbf{Y}|\boldsymbol{\theta},\mathbf{M})\}\big). \tag{5.16}$$

The expectation for a specific τ can be obtained by Monte Carlo estimates,

$$\mathbb{E}_{p_{\tau_k}}\{\log p(\mathbf{Y}|\boldsymbol{\theta},\mathbf{M})\} \approx \frac{1}{S}\sum_{s=1}^{S}\log p(\mathbf{Y}|\boldsymbol{\theta}^{(s)},\mathbf{M}), \tag{5.17}$$

where $\boldsymbol{\theta}^{(s)}$ denotes a sample drawn from $p_{\tau_k}(\boldsymbol{\theta}|\mathbf{Y},\mathbf{M})$. The individual τ_k are also called temperature rungs. In most applications, the chosen quadrature method is the trapezoidal rule. However, the trapezoidal method is not the most sophisticated quadrature method, as it is of very low approximation order. Secondly, the accuracy of the integration depends strongly on the location of the temperature rungs. Therefore, different scheduling procedures have been introduced.

5.3.1 Fixed schedule

Already in the first papers about thermodynamic integration, the temperature schedule, i.e. the discretization $0 = \tau_0 < \tau_1 < ... < \tau_{K-1} < \tau_K = 1$, has been discussed. It is immediately clear that the schedule has a significant influence on performance and numerical stability of the whole scheme. It is well known that the Bayes factor is sensitive to the choice of prior, therefore most schedules try to concentrate more temperature rungs near $0 = \tau_0$. Thus already Friel & Pettitt [2008] recommend a power law temperature schedule of the type $\tau_k = (x_k)^q$, where $x_k = k/K, k = 0, \ldots, K$ is an equal spacing of $K + 1$ points in the interval $[0, 1]$, and $q > 1$ is a constant. This leads to temperature rungs that are chosen with higher frequency close to $\tau = 0$. Calderhead & Girolami [2009] advocate the choice of $K = 30$ and $q = 5$. It has been shown that this is superior to uniform spacing.

5.3.2 Adaptive trapezoidal rule

Friel *et al.* [2013] have developed a first adaptive method for choosing the temperature rungs. It is based on an interesting connection between the derivative of the expectation that has to be calculated and the associated variance:

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\mathbb{E}_{p_\tau}\{\log p(\mathbf{Y}|\boldsymbol{\theta},\mathbf{M})\} = \mathbb{E}_{p_\tau}\{\log p(\mathbf{Y}|\boldsymbol{\theta},\mathbf{M})\}^2 - (\mathbb{E}_{p_\tau}\{\log p(\mathbf{Y}|\boldsymbol{\theta},\mathbf{M})\})^2 \tag{5.18}$$

$$= \mathrm{Var}_{p_\tau}\{\log p(\mathbf{Y}|\boldsymbol{\theta},\mathbf{M})\} \tag{5.19}$$

Here, $\mathrm{Var}_{p_\tau}\{\log p(\mathbf{Y}|\boldsymbol{\theta},\mathbf{M})\}$ is the variance of the log deviance at temperature τ. Based on this, Friel et al. suggest first estimating the curve at $\tau = 0$ and $\tau = 1$. By taking the tangents to the curve at these two points and finding their intersection, they find the next temperature rung. If thus f_k and V_k are the estimated function and the gradient at τ_k and the same respectively for τ_{k+1}, the new temperature τ^* is set to

$$\tau^* = \frac{f_{k+1} - f_k + \tau_k V_k - \tau_{k+1} V_{k+1}}{V_k - V_{k+1}}. \tag{5.20}$$

This new temperature rung τ^* can be outside the interval $[\tau_k, \tau_{k+1}]$ if there is an inflection in the interval. In this case, the authors propose to use a weighted average instead:

$$\tau^* = \tau_k + \frac{V_{k+1}}{V_k + V_{k+1}} (\tau_{k+1} - \tau_k) . \tag{5.21}$$

From the trapezoidal rule, error estimates are available, and a new τ^* provides us with the two new contributions. The following temperature rung is then placed in the subinterval with the larger contribution. In practice, due to Monte Carlo error in the sampling of the expectation, the function might not be strictly increasing, then the interval with the biggest error estimate is picked. This strategy is cheap and follows a reasonable idea, however, the placing will probably not be optimal.

5.3.3 Adaptive Simpson's rule

In complex model selection tasks, it is very desirable to control the number of temperature rungs and use only as many as necessary for achieving a predetermined error tolerance in the marginal likelihood. The just introduced adaptive trapezoidal algorithm determines the placing of the temperature rungs on the fly based on the already chosen rungs and their function values. This is however based on a number of rungs that has to be set beforehand.

We developed a different adaptive strategy for placing the temperature rungs. It is based on classical integral approximation theory from numerical mathematics, and uses Simpson's rule instead of the trapezoidal rule as with both previous strategies. Both quadrature methods were already introduced in Section 2.2.3.

Simpson's rule provides a natural extension to an adaptive approximation of an integral. Furthermore, Simpson's rule is of approximation order four, while the trapezoidal rule only has order two, thus we should gain two orders of accuracy. In practice, this is however hard to assess, since the function evaluations are tainted by Monte Carlo error and the analytical shape of the cost function is not available in all but the most simple cases.

In principle, we apply the standard adaptive Simpson's rule. It is based on the regular Simpson's rule and approximates the integral $J(f)$ of a function $f(x)$ on the interval $[a, b]$ as

$$J(f)[a, b] = \frac{b-a}{6} \left(f(a) + 4f(\frac{a+b}{2}) + f(b) \right) . \tag{5.22}$$

This is combined as Simpson's sum, which corresponds to subdividing the interval to $\left[a, \frac{a+b}{2}\right]$ and $\left[\frac{a+b}{2}, b\right]$, applying Simpson's rule to both and summing up. This yields

$$\hat{J}(f)[a,b] = J(f)\left[a, \frac{a+b}{2}\right] + J(f)\left[\frac{a+b}{2}, b\right]. \tag{5.23}$$

Furthermore, we have the following error estimates: for Simpson's rule, we get for $h = \frac{b-a}{2}$ that $J(f) = I(f) + h^5 f^{(4)}(\xi)/90$, where $I(f)$ is the exact analytical value of the integral and ξ some point in the interval $[a,b]$. For the Simpson sum, we get $\hat{J}(f) = I(f) + 1/16 \cdot (h^5 f^{(4)}(\xi^*)/90)$ for again some $\xi^* \in [a,b]$. Under the assumption that the fourth derivative is reasonably flat, $f^{(4)}(\xi) \doteq f^{(4)}(\xi^*)$, we can formulate the following iterative procedure called **adaptive Simpson's rule** (Lyness [1969]):

Algorithm 5: Adaptive Simpson's rule for recursively calculating a definite integral with controlled accuracy.

---

input : error tolerance `TOL`, function $f(x)$, interval borders a and b
output: value of the interval $I(f)[a,b]$

initialize $J := J(f)[a,b]$ and $\hat{J} := \hat{J}(f)[a,b]$;
Set

$$I(f)[a,b] := \begin{cases} \hat{J} & \text{if } |J - \hat{J}| < 15 \cdot \texttt{TOL} \\ I(f)\left[a, \frac{a+b}{2}\right] + I(f)\left[\frac{a+b}{2}, b\right] & \text{otherwise} \end{cases} \tag{5.24}$$

It is obvious from the proportionality of the error term to the fourth derivative $f^{(4)}(\xi)$ for $\xi \in [a,b]$ of the integrand that the adaptive Simpson's rule is exact for cubic functions. Until now, most applications use the trapezoidal rule, which is only exact for linear functions, since its error term is proportional to the second derivative $f''(\xi)$ for $\xi \in [a,b]$. Switching from the trapezoidal rule to Simpson's rule should thus lead to a gain of two orders of accuracy. We expect a much smaller error for the same number of function evaluation.

In contrast to classical quadrature problems, in our setting the integrand is a random variable and thus its evaluation is tainted by Monte Carlo errors. If this Monte Carlo error is large, it could cause convergence problems, therefore it should be controlled as close as possible. Still, a few heuristics controlling the performance of the algorithm and preventing infinite loops have to be introduced.

5.3.4 Power law scheduling for the adaptive Simpson's rule

As introduced in Section 5.3.1, a power law schedule like $\tau_k = (k/K)^5, k = 0, \ldots, K$ with $K + 1$ rungs performs well in standard non-adaptive thermodynamic integration (Calderhead & Girolami [2009]). This lead us to considering this also for the adaptive Simpson's rule. One easy way to apply this is through integration by substitution. To reproduce a power law rung placement, we choose a substitution function of the form $\lambda = \psi(\tau) = \tau^{1/q}$ with some exponent q, e.g. $q = 5$. Integration by substitution leads then to a thermodynamic integral of

$$\begin{aligned} &\int_0^1 \mathbb{E}_{p_\tau}\{\log p(\mathbf{Y}|\boldsymbol{\theta}, \mathbf{M})\} d\tau \\ =&\int_0^1 \mathbb{E}_{p_\lambda}\{\log p(\mathbf{Y}|\boldsymbol{\theta}, \mathbf{M})\} \cdot q\lambda^{q-1} d\lambda. \end{aligned} \tag{5.25}$$

To this transformed function, i.e. the integrand on the right hand side of Equation (5.25), we can then apply the adaptive Simpson's rule as described in the previous chapter. This might combine advantages of the adaptive refinement and advantages of temperature scheduling heuristics to achieve the requested tolerance with even fewer function evaluations and will be evaluated in Section 5.4.2.

5.4 An analytically tractable numerical example

In this section we evaluate all proposed methods, especially the newly proposed adaptive Simpson's rule for thermodynamic integration using an analytically tractable example based on normal distributions. Here the Bayes factor can be computed analytically and thus the error made by the presented approximations is accessible. It was already introduced in Schmidl [2012]. We have found this possible with a very simple model selection where we choose between the following two models:

- Model $\mathbf{M_1}$: a normal distribution with expected value μ and standard deviation σ, with N data points drawn.
- Model $\mathbf{M_2}$: two normal distributions with expected values μ_1 and $\mu_2 = -\mu_1$ and standard deviation σ, with N_1 data points drawn from the first normal distribution and N_2 drawn from the second.

We generate artificial data from model $\mathbf{M_2}$, so that the analytical Bayes factor and the results of thermodynamic integration should both point towards model $\mathbf{M_2}$ consider-

ably, since this model selection problem is rather simple.

We draw $N = 10$ samples from model $\mathbf{M_2}$ in the following way to obtain the data $\mathbf{Y} = (Y_1, ..., Y_N)$

- From the first normal distribution N_1 points $Y_1, ..., Y_{N_1} \sim \mathcal{N}(\mu_1, \sigma^2)$ with $N_1 = 3$ and $\mu_1 = -2$.
- From the second normal distribution N_2 points $Y_{N_1+1}, ..., Y_N \sim \mathcal{N}(\mu_2, \sigma^2)$ with $N_2 = N - N_1 = 7$ and $\mu_2 = 2$.

This is then compared to the hypothesis that the data comes from model $\mathbf{M_1}$ in such a fashion that

- $Y_1, ..., Y_N \sim \mathcal{N}(\mu, \sigma^2)$ with $N = 10$.

A visualization of the data generated from model $\mathbf{M_2}$ can be seen in Figure 7.6(a). To facilitate the computations, we choose a fixed σ, to which we assign the value $\sigma = 2$ or $\sigma = 1$ in our implementation. This leaves us with one free parameter, μ, for model $\mathbf{M_1}$ and two parameters μ_1 and μ_2 for model $\mathbf{M_2}$. For reasons of computational convenience, we choose the following prior distributions for our parameters:

- $\mu \sim \mathcal{N}(0, \sigma^2)$
- $\mu_1 \sim \mathcal{N}(-2, \sigma^2)$
- $\mu_2 \sim \mathcal{N}(+2, \sigma^2)$,

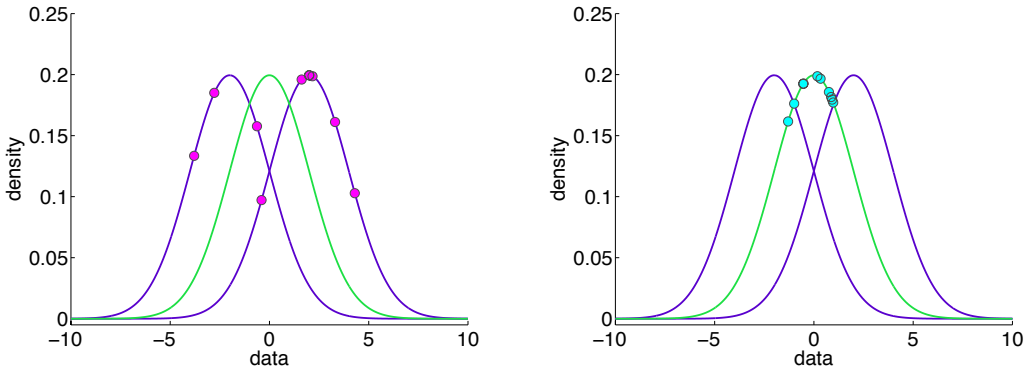
with μ_1 and μ_2 independent.

Of course it is also possible to generate data from the smaller model $\mathbf{M_1}$, see e.g. in Figure 7.6(b). The following calculations are valid irrespective of which model the data was generated from.

5.4.1 Analytical computation of the Bayes factor

The likelihood we obtain for both models is given by

$$p(\mathbf{Y}|\mu, \mathbf{M_1}) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{n=1}^{N}(Y_n - \mu)^2\right)\right) \tag{5.26}$$



(a) (b)

Figure 5.2: Generated artificial data. (a-b) The two models $\mathbf{M_1}$ (green) and $\mathbf{M_2}$ (purple). (a) Artificial data generated from model $\mathbf{M_2}$ (magenta dots). (b) Artificial data generated from model $\mathbf{M_1}$ (cyan dots).

and

$$p(\mathbf{Y}|\mu_1, \mu_2, \mathbf{M_2}) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{n=1}^{N_1}(Y_n - \mu_1)^2 + \sum_{n=N_1+1}^{N}(Y_n - \mu_2)^2\right)\right). \tag{5.27}$$

After some straightforward calculations we find that the posterior distributions for the parameters within the two models are

- $\mu \sim \mathcal{N}\left(\frac{1}{N+1}\sum_{n=1}^{N} Y_n, \frac{\sigma^2}{N+1}\right)$ for model $\mathbf{M_1}$
- $\begin{pmatrix}\mu_1 \\ \mu_2\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}\frac{1}{N_1+1}(-2+\sum_{n=1}^{N_1} Y_n) \\ \frac{1}{N_2+1}(+2+\sum_{n=N_1+1}^{N} Y_n)\end{pmatrix}, \begin{pmatrix}\frac{\sigma^2}{N_1+1} & 0 \\ 0 & \frac{\sigma^2}{N_2+1}\end{pmatrix}\right)$ for model $\mathbf{M_2}$.

In order to compute the Bayes factor, we need to compute the marginal likelihoods $p(\mathbf{Y}|\sigma, \mathbf{M_1})$ and $p(\mathbf{Y}|\sigma, \mathbf{M_2})$. We will begin with model $\mathbf{M_1}$ and thus $p(\mathbf{Y}|\sigma, \mathbf{M_1})$:

$$
\begin{aligned}
p(\mathbf{Y}|\sigma,\mathbf{M_1}) &= \int_{\mathbb{R}} p(\mathbf{Y}|\mu,\mathbf{M_1})p(\mu|\mathbf{M_1})\,\mathrm{d}\mu \\
&= \int_{\mathbb{R}} \left(\tfrac{1}{\sqrt{2\pi}\sigma}\right)^{N+1} \exp\left(-\tfrac{1}{2\sigma^2}\sum_{n=1}^{N}(Y_n-\mu)^2 - \tfrac{1}{2\sigma^2}\mu^2\right)\mathrm{d}\mu \\
&= \int_{\mathbb{R}} \left(\tfrac{1}{\sqrt{2\pi}\sigma}\right)^{N+1} \exp\left(-\tfrac{1}{2\sigma^2}\left(\sum_{n=1}^{N}Y_n^2 - 2\mu\sum_{n=1}^{N}Y_n + (N+1)\mu^2\right)\right)\mathrm{d}\mu \\
&= \int_{\mathbb{R}} \left(\tfrac{1}{\sqrt{2\pi}\sigma}\right)^{N+1} \exp\left(-\tfrac{N+1}{2\sigma^2}\left(\tfrac{1}{N+1}\sum_{n=1}^{N}Y_n^2 - 2\mu\tfrac{N}{N+1}\bar{\mathbf{Y}} + \mu^2\right)\right)\mathrm{d}\mu \\
&\qquad \cdot \exp\left(-\tfrac{N+1}{2\sigma^2}\left(\tfrac{N}{N+1}\bar{\mathbf{Y}} - \mu\right)^2\right)\mathrm{d}\mu \\
&= \left(\tfrac{1}{\sqrt{2\pi}\sigma}\right)^{N} \tfrac{1}{\sqrt{N+1}} \exp\left(-\tfrac{N+1}{2\sigma^2}\left(\tfrac{1}{N+1}\sum_{n=1}^{N}Y_n^2 - (\tfrac{N}{N+1}\bar{\mathbf{Y}})^2\right)\right)
\end{aligned}
$$

where $\bar{\mathbf{Y}} = \frac{1}{N}\sum_{n=1}^{N} Y_n$ is the sample mean. In a very similar fashion, we can also calculate $p(\mathbf{Y}|\sigma,\mathbf{M_2})$. For that, we introduce the notation $\bar{\mathbf{Y}}_1 = \frac{1}{N_1}\sum_{n=1}^{N_1} Y_n$ and $\bar{\mathbf{Y}}_2 = \frac{1}{N_2}\sum_{n=N_1+1}^{N} Y_n$:

$$
\begin{aligned}
p(\mathbf{Y}|\sigma,\mathbf{M_2}) &= \int_{\mathbb{R}}\int_{\mathbb{R}} p(\mathbf{Y}|\mu_1,\mu_2,\mathbf{M_2})p(\mu_1|\mathbf{M_2})p(\mu_2|\mathbf{M_2})\,\mathrm{d}\mu_1\,\mathrm{d}\mu_2 \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}} \left(\tfrac{1}{\sqrt{2\pi}\sigma}\right)^{N+2} \exp\Bigg(-\tfrac{1}{2\sigma^2}\Bigg(\sum_{n=1}^{N}Y_n^2 + 8 - 2(N_1\bar{\mathbf{Y}}_1 - 2)\mu_1 - 2(N_2\bar{\mathbf{Y}}_2 + 2)\mu_2 \\
&\qquad + (N_1+1)\mu_1^2 + (N_2+1)\mu_2^2\Bigg)\Bigg)\mathrm{d}\mu_1\,\mathrm{d}\mu_2 \\
&= \left(\tfrac{1}{\sqrt{2\pi}\sigma}\right)^{N} \tfrac{1}{\sqrt{N_1+1}\sqrt{N_2+1}} \exp\left(-\tfrac{1}{2\sigma^2}\left(\sum_{n=1}^{N}Y_n^2 + 8 - \frac{(N_1\bar{\mathbf{Y}}_1 - 2)^2}{N_1+1} - \frac{(N_2\bar{\mathbf{Y}}_2+2)^2}{N_2+1}\right)\right).
\end{aligned}
$$

Having obtained the marginal likelihoods, we can now compute the Bayes factor B_{21} in favor of model $\mathbf{M_2}$:

$$B_{21} = \frac{p(\mathbf{Y}|\sigma, \mathbf{M_2})}{p(\mathbf{Y}|\sigma, \mathbf{M_1})} \tag{5.28}$$

$$= \frac{\sqrt{N+1}}{\sqrt{N_1+1}\sqrt{N_2+1}} \exp\left(-\frac{1}{2\sigma^2}\left(\frac{(N\bar{\mathbf{Y}})^2}{N+1} - \frac{(N_1\bar{\mathbf{Y}}_1 - 2)^2}{N_1+1} - \frac{(N_2\bar{\mathbf{Y}}_2 + 2)^2}{N_2+1} + 8\right)\right) \tag{5.29}$$

Since B_{21} only depends on the data $\mathbf{Y}$ and the standard deviation σ, which we fixed, we can easily evaluate the Bayes factor in our implementation.

In this simple example, also the expected log deviance as a function of temperature is easily analytically tractable for model $\mathbf{M_1}$, meaning that the integrand in thermodynamic integration is known. After some cumbersome, but straightforward computations, which can be found in Schmidl [2012], we find

$$\mathbb{E}_{p_\tau}\{\log p(\mathbf{Y}|\mu, \mathbf{M_1})\}$$

$$= -\frac{1}{2\sigma^2}\left\{\left(\sum_{n=1}^{N}\left(Y_n^2 + 2\sigma^2\log(\sqrt{2\pi}\sigma)\right)\right) + \frac{N\sigma^2 - 2\bar{\mathbf{Y}}^2N^2\tau}{N\tau+1} + \frac{N^3\bar{\mathbf{Y}}^2\tau^2}{(N\tau+1)^2}\right\} \tag{5.30}$$

5.4.2 Comparison of methods

After the introduction of the model selection scenario, we now first want to focus on the comparison of all presented methods on this example. In the following two sections, we will then perform an in-depth analysis of the novel adaptive Simpson's rule for TI, first on the linear scale and later on the power law schedule introduced in Section 5.3.4.

All of the presented results were obtained from computations in MATLAB. For the optimization based criteria AIC, BIC and LRT, we used 10,000 runs of a local optimization routine in MATLAB. Starting values where drawn uniformly from the intervals $[-5, 5]$ for μ in $\mathbf{M_1}$ and from $[-5, 0]$ and $[0, 5]$ for μ_1 and μ_2 in $\mathbf{M_2}$, respectively, to find the maximum likelihood estimates. Since the posterior distributions of the parameters in the two models in the example are normal distributions, it is possible to compare the optimization results to the true values of the maximum a posteriori estimates. We find very good agreement, certainly due to the simplicity of the optimization task.

For the first three sampling based approaches (prior arithmetic mean, posterior harmonic mean and Chib's method), we drew 100,000 samples each from the required densities. For the prior, sampling was directly available. For the posterior, we sampled with the Adaptive Metropolis Sampler as introduced in Section 4.2. Since Chib's

method is in our opinion not tailored to accommodate adaptive sampling, we there chose a regular Metropolis-Hastings algorithm with a normal distribution as proposal distribution.

For the standard thermodynamic integration, we follow the recommendations of Calderhead & Girolami [2009] and first choose a power law temperature schedule of $\tau_k = (k/19)^5$ with $K = 20$ temperature steps as the example of a fixed schedule. For each temperature, we draw 5000 samples, yielding also a total of 100,000 samples. For the adaptive trapezoidal rule, we also used $K = 20$ temperature rungs with 5000 samples each. The adaptive Simpson's rule does not have a predetermined number of temperature rungs, but we also draw 5000 samples per temperature. As tolerance value `TOL` for this initial comparison, we choose a value of $\texttt{TOL} = 10^{-3}$. The effect of the tolerance will be examined in more detail in the following section.

All sampling algorithms were initialized at the maximum a posteriori estimates found for the AIC/BIC/LRT approaches, minimizing the influence of the starting point on the sampling.

For the sampling based approaches, we ran the sampling 30 times on the same data to correct for randomness. This takes less than an hour on a standard desktop computer. These 30 runs yield mean results for the Bayes factor and the standard error. The standard error does not provide a confidence interval for the value in this case, as it incorporates both deterministic and MCMC errors of the methods. Furthermore, the Bayes factor is the ratio of the two marginal likelihoods and thus the error is a combination of the two single errors on the estimated quantities. Nevertheless the Bayes factor is often the quantity of interest and the standard error provides an idea about the spread of attained values.

We first study the situation where the artificial data as generated from model $\mathbf{M_2}$. The AIC for model $\mathbf{M_1}$ is 49.58, while the AIC for the correct model $\mathbf{M_2}$ is 40.74, thus the AIC makes the correct choice in this simple example. For the BIC we find that the value for model $\mathbf{M_1}$ is 49.88, while the value for model $\mathbf{M_2}$ is 41.35, also indicating a preference for the correct model. The likelihood ratio test rejects the smaller model $\mathbf{M_1}$ with a p-value of 0.000956.

For the sampling based approaches, the results are given in Table 5.1. For the Bayes factor, the analytical computation based on our drawn data shows a true value of 139.23 for model $\mathbf{M_2}$ over $\mathbf{M_1}$, i.e. decisive preference for model $\mathbf{M_2}$. This is based on log marginal likelihoods of $\log p(\mathbf{Y}|\sigma, \mathbf{M_1}) = -25.0582$ and $\log p(\mathbf{Y}|\sigma, \mathbf{M_2}) = -20.12$. All

Method	**Mean B_{21}**	**s.e.**
TRUE	139.23	
Prior arithmetic mean	138.76	0.03
Posterior harmonic mean	149.37	1.85
Chib's method	55.76	0.01
Fixed schedule TI	140.66	0.13
Adaptive trapezoidal TI (Friel)	141.24	0.14
Adaptive Simpson's rule TI	140.62	0.10

Table 5.1: Bayes factors for model $\mathbf{M_2}$ versus model $\mathbf{M_1}$. Data was generated from model $\mathbf{M_2}$. Given is the mean over 30 runs as well as the corresponding standard error for all sampling bases methods.

sampling based approaches also find a preference for this model.

The posterior harmonic mean estimate overestimates the Bayes factor rather considerably, also the standard error is very large compared to the other estimation methods. Both indicates a rather bad approximation and reliability, which is in agreement with the general issues of this sampling method. Chib's method in our case performs worst. While the Bayes factor is still very strong in favor of model $\mathbf{M_2}$, it underestimates the true value by a factor of 2. This is mostly due to a systematic underestimation of both log marginal likelihoods, e.g. for model $\mathbf{M_2}$, all sampling results were < -21.4 while the true value is $\log p(\mathbf{Y}|\sigma, \mathbf{M_2}) = -20.12$. Note that this difference between true and sampled value is on a log scale, thus on the non-log scale of the Bayes factor, the discrepancy is more drastic. This seems to be a systematic issue, since the sampling passed Geweke's convergence criterion with all p-values larger than 0.98. Furthermore, the mean of the samples for model $\mathbf{M_2}$ for example can be compared to the analytical posterior distribution. We find that the sample means of -2.3007 and 2.1318 agree very well with the analytical values -2.2994 and 2.1260. Also the sample covariance matrix $[0.9915, -0.0053; -0.0053, 0.4974]$ agrees very well with the analytical one $[1, 0; 0, 0.5]$. We conclude that Chib's method seems to suffer from severe numerical issues and should thus only be used very carefully.

The prior arithmetic mean and the three variants of thermodynamic integration perform best. The good performance of the prior arithmetic mean is certainly due to the simplicity of the model selection problem, as well as the goodness of the prior. We would not advice to use this result as an indicator of good performance of the estimator

in larger model selection problems. Thermodynamic integration performs very well in our scenario and we expect it to also perform well in other applications.

The three different temperature schedules for the thermodynamic integration perform very similarly in this easy example. The adaptive Simpson needs between 17 and 29 function evaluations, on average over the thirty runs, it needs 19.3 evaluations for model $\mathbf{M_1}$ and 22.1 evaluations for model $\mathbf{M_2}$. This shows that in this simple example, the Simpson rule saves time for the simpler model due to its adaptivity, but invests a few extra function evaluations for a good accuracy in the larger model compared to the 20 that the other two thermodynamic integration methods can use.

As further verification of our model selection methods, we also considered the reverse scenario where the simpler model $\mathbf{M_1}$ is the true model used to generate the data. We generated 10 data points from the model using $\mu = 0$ as the true parameter and choose $\sigma = 1$ and again apply all mentioned model selection methods. The theoretical Bayes factor for our generated data is $B_{21} = 0.0126$. For ease of comparability to the preceding model selection, we will instead refer to $B_{12} = 1/B_{21} = 79.5958$. We used the same specifications as previously reported.

The AIC and BIC for model $\mathbf{M_1}$ are 26.83 and 27.14, the AIC and BIC for model $\mathbf{M_2}$ are 28.26 and 28.86, thus both information criteria make the correct choice. The p-value from the likelihood ratio test is 0.4468, thus model $\mathbf{M_1}$ can not be rejected at reasonable confidence levels.

For the sampling based approximations of the Bayes factor, we find the results in Table 5.2.

Method	**Mean B_{12}**	**s.e.**
TRUE	79.5958	
Prior arithmetic mean	79.2716	0.10
Posterior harmonic mean	28.7435	1.10
Chib's method	111.8052	0.04
Fixed schedule TI	84.1115	0.13
Adaptive trapezoidal TI (Friel)	91.3555	0.35
Adaptive Simpson's rule TI	79.1988	0.10

Table 5.2: Bayes factors for model $\mathbf{M_1}$ versus model $\mathbf{M_2}$. Data was generated from model $\mathbf{M_1}$. Given is the mean over 30 runs as well as the corresponding standard error for all sampling bases methods.

As before, the posterior harmonic mean and Chib's method perform worst. Due to the simplicity of the problem and the quality of prior information, the prior arithmetic mean again performs very well. For the three variants of thermodynamic integration, the situation is more interesting. The adaptive Simpson's rule needs an average 19.2 (between 17 and 25) function evaluations for model $\mathbf{M_1}$. For model $\mathbf{M_2}$, it uses 21 to 41 function evaluations, on average 29.5. This means that the adaptive Simpson uses more function evaluations in all runs than are available for the other temperature schedules. However, this pays off since we achieve significantly higher accuracy.

We believe the reason for the suboptimal performance of the adaptive trapezoidal rule lies in the placing of the rungs. In model $\mathbf{M_1}$, it places for example only 4 rungs between 0 and 0.2, but 8 rungs between 0.2 and 0.5, for model $\mathbf{M_2}$ the situation is similar. This does not seem to be optimal.

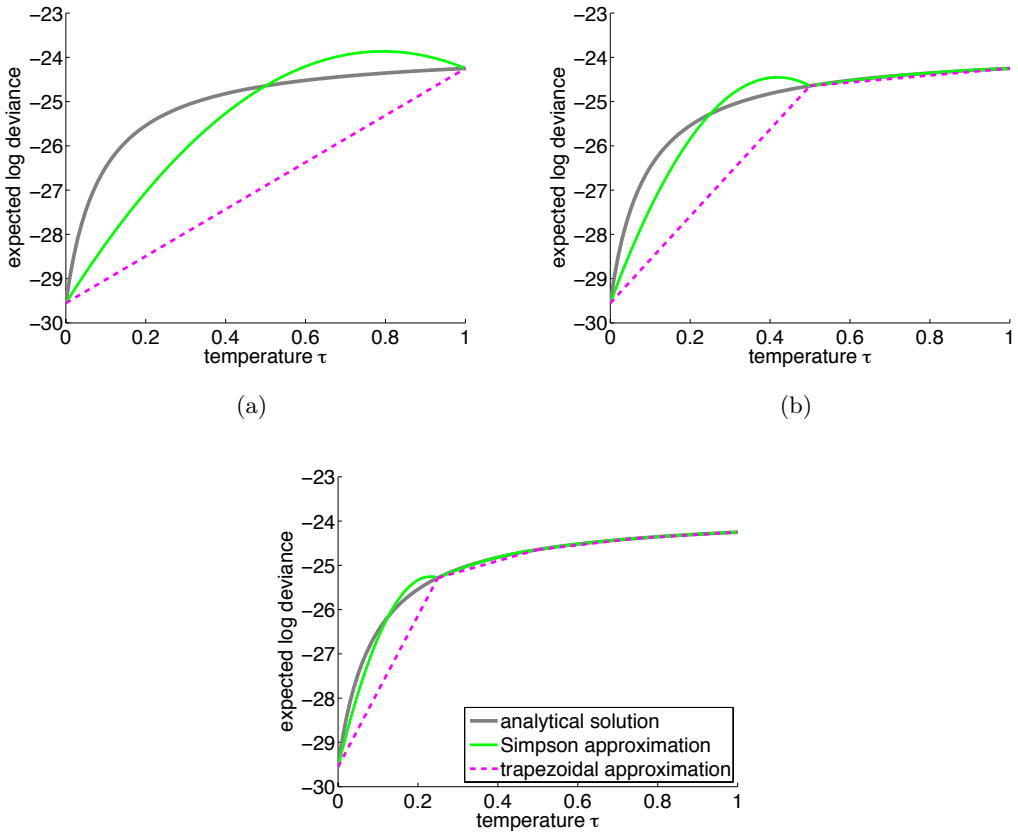
We conclude that in both presented scenarios, thermodynamic integration methods generally perform well and often superior to other methods. We will thus prefer this general method in the model selection problems presented in the Applications part of this thesis. While the prior arithmetic mean performed well, it is only an alternative if the prior information is strong.

One reason for the improved performance of Simpson's rule is that it approximates the integrand with a parabola, which fits the typical shape of the expected log deviance better than the straight lines from the trapezoidal approximation, when the same rungs are used. This results in a improved convergence speed, even if three instead of two function evaluations per subinterval have to be used. The quality of fit is also shown for the analytically available expected log deviance as a function of the temperature from model $\mathbf{M_1}$ in Figure 5.3. The formula for this expected log deviance is given in Equation (5.30).

5.4.3 Numerical results for the adaptive Simpson's rule

After the introduction of the model selection scenario and the comparison of methods, we now want to present an in-depth evaluation of the adaptive Simpson's rule on this example on the linear scale.

As already mentioned, the adaptive Simpson's rule does not have a predetermined number of temperature rungs. We introduce one heuristic that sets a cut off after four levels of refinement, meaning at most 65 function evaluations, i.e. 65 MCMC runs. If the



(a) (b) (c)

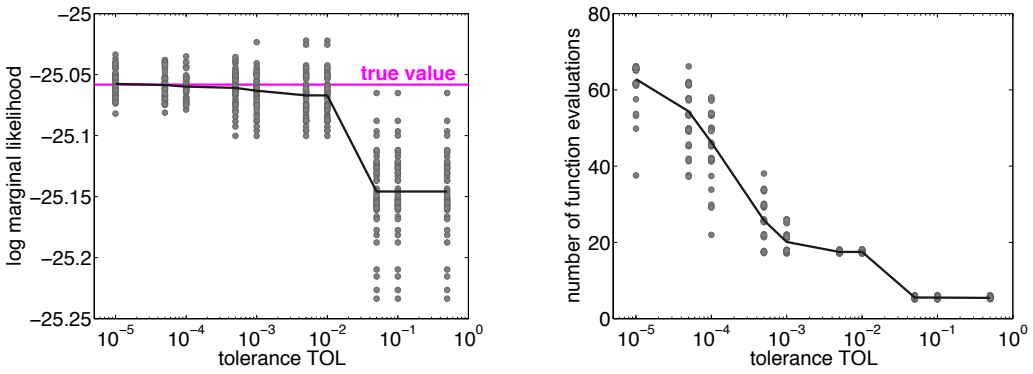
Figure 5.3: Analytical expected log deviances. The analytical expected log deviance of model $\mathbf{M_1}$ (grey), Simpson approximations (green) and trapezoidal approximations (magenta). (a) First approximation based on the whole interval. (b) Approximation based on the two intervals $[0, 0.5]$ and $[0.5, 1]$. (c) Approximation based on the four intervals $[0, 0.25]$,$[0.25, 0.5]$, $[0.5, 0.75]$ and $[0.75, 1]$. Clearly, the Simpson approximation is always better than the trapezoidal approximation.

prescribed tolerance is not reached with these 65 evaluations, we instead take the best available approximation. In this case, the integration error is larger than the tolerance and thus has to be treated accordingly. We believe this is a reasonable setting, since most thermodynamic integration implementations use about 30 function evaluations (Calderhead & Girolami [2009]). In our implementation, only the number of function evaluations necessary for the desired accuracy are used. Thus the number of function evaluations varies between five, the minimum number necessary for a Simpson's sum, and 65.

It was already noted in the previous section that the adaptive Simpson's rule performs well irrespective of which model the data was generated from. When model $\mathbf{M_2}$ is used to generate the data with a true value for the Bayes factor of 139.23 for model $\mathbf{M_2}$ over $\mathbf{M_1}$, thirty runs of the adaptive Simpson on the same data yields a result of 140.62 with a standard error of 0.10, i.e. a very good fit. Also when data is drawn from model $\mathbf{M_1}$, the smaller model, finding an analytical Bayes factor of $B_{21} = 0.0126$ or $B_{12} = 1/B_{21} = 79.5958$. In this scenario, thirty runs of the adaptive Simpson yield a result of $B_{12} = 79.1988$ with standard error 0.10, thus also a very good result.

We now focus on evaluating the effect of the single parameter controlling the performance of the adaptive Simpson's rule, the error tolerance `TOL`, on the number of required function evaluations. Based on model $\mathbf{M_1}$ in the situation where model $\mathbf{M_2}$ is true, we run the adaptive Simpson 50 times for varying tolerances. From this, we derive a mean number of function evaluations as well as its spread, see Figure 5.4(a) and Figure 5.4(b). The true value of the log marginal likelihood in this case is $\log p(\mathbf{Y}|\sigma, \mathbf{M_1}) = -25.0582$. As expected, the accuracy of the estimate gets higher with lower error tolerances. Also the required number of function evaluations rises with lower tolerances from the minimum five to the maximum 65.

We find that the tolerance chosen in the example of $\texttt{TOL} = 10^{-3}$ is a very good trade off between accuracy and the required number of function evaluations. It is not trivial to choose an appropriate tolerance for the adaptive Simpson's rule in a general setting, since the tolerance is an absolute error tolerance and has to be appropriate for the achieved function values. The tolerance chosen here corresponds to a relative error in the order of magnitude of 10^{-5}, when divided by the expected log deviance at $\tau = 1$. Thus, we suggest taking `TOL` in this order of magnitude for other examples as well.



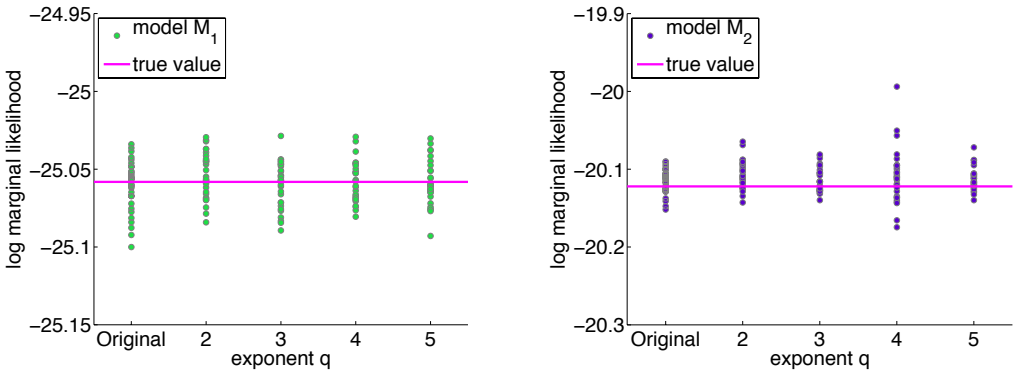
(a) (b)

Figure 5.4: Varying tolerances for the adaptive Simpson's rule. (a) The effect of the chosen tolerance `TOL` on the achieved integral value (the magenta line denotes the true value), based on 50 runs of model $\mathbf{M_1}$ for each chosen tolerance. (b) The required number of function evaluations rises with lower error tolerance. (a-b) The circles denote 50 runs for each chosen tolerance, the black line their mean.

5.4.4 Evaluation of the power law schedule for the adaptive Simpson's rule

As we hypothesized in Section 5.3.4 that the combination of the heuristic power law temperature schedules and adaptive Simpson's rule might improve the performance further, we analyzed this numerically. Therefore, we repeat the sampling from the previous section for the adaptive Simpson, for values of the exponent q of $2, 3, 4$ and 5 and compare this with the results from the previous section. We find that the power law schedule performs well for both models, irrespective of chosen exponent, see Figure 5.5. It has to be noted that we applied the same $\texttt{TOL} = 10^{-3}$ for all exponents, as it represents a common requirement for all integrations, even though this might be different for other convergence criteria.

We conclude that the power law scheduling does not immediately lead to significant improvement. If an adaptive scheme and a high accuracy is used, the benefit from the power law temperature scheduling seems to be negligible, in particular if the optimal power is not known. However, we still feel that for a fixed schedule version of thermodynamic integration like in Calderhead & Girolami [2009], a power law schedule should be used since it has been shown to perform well then.



(a) (b)

Figure 5.5: The effect of the power law schedule. The effect of the power law schedule for model $\mathbf{M_1}$ (Panel (a)) and model $\mathbf{M_2}$ (Panel (b)) for different values of the exponent, compared with the true values (magenta lines) and the results for the original sampling (left-most column of markers in each plot). It can be clearly seen that the power law schedule performs well for both models.

5.5 Conclusions

In this chapter, we have presented several more or less involved methods for doing (Bayesian) model selection, using methods introduced in the previous chapters like MCMC. When we are now faced with a model selection task on a more complex example than the one presented in this chapter, the example can be used as a guideline as to which method should be preferred. In a general setting with unknown prior quality, we advocate the use of thermodynamic integration methods over the prior arithmetic mean despite the good performance of the prior arithmetic mean. Thermodynamic integration is state-of-the-art. With the newly introduced adaptive Simpson's rule, it can be used efficiently and with controlled accuracy. This scheme adaptively determines the number of function evaluations that are necessary for achieving a required accuracy. Furthermore, it possesses a higher approximation order than the usually applied trapezoidal rule. We expect this to be important especially in high-dimensional problems, for example the one in Chapter 6. In more medium-sized applications like in Chapter 7, the standard thermodynamic integration can be expected to yield good results.

In general, thermodynamic integration can be combined with sequential sampling from the tempered distributions, or population-based MCMC (Calderhead & Girolami [2009]). This is of course also possible with the adaptive Simpson's rule and could in-

crease sampling efficiency, especially in examples that might be more difficult to sample than the examples presented in this thesis.

We have provided a nice, analytically tractable example for model selection, comparing normal distributions. Since there the expected log deviance is analytically assessable, it can be clearly seen that there is a gain in approximation order when using Simpson's rule instead of the trapezoidal rule. We propose that the example be used for evaluating new methods also in the future.

For the power law scheduling of the adaptive Simpson's rule, we find that the potentially best choice for the exponent q is not obvious from setting the error tolerance `TOL` alone. The computation of the ideal q might be an important step also for fixed schedule thermodynamic integration, on a way towards an optimal temperature schedule.

Part II

Applications

6

Model selection of models for single-cell dynamics

In this chapter, we apply model selection with the adaptive Simpson's rule for the thermodynamic integration to two simple models. These two compared models are low-dimensional in their individual parameter dimensions with three and four parameters, respectively. However, they represent single-cell time courses and should thus be fitted individually to a rather large number of individual data sets in parallel. This parallel single-cell inference is applied to measurement data from two different cell types and represents the second methodological challenge presented in this chapter.

The two models represent two possibilities for protein degradation in the presence of cycloheximide. In living cells, proteins are continuously produced and degraded. It is experimentally possible to block protein synthesis to observe degradation isolated by treating cells with the cycloheximide. This gives protein half-lives which are important prior knowledge for experiments or models where proteins are both produced and degraded (Eden *et al.* [2011]; Schwanhäusser *et al.* [2011]).

This chapter is based on and in part identical with the following two publications:

- **S. Hug**, M. Schwarzfischer, J. Hasenauer, C. Marr and F.J. Theis. An adaptive method for calculating Bayes factors using Simpson's rule, *in revision*
- M. Schwarzfischer, O. Hilsenbeck, B. Schauberger, **S. Hug**, A. Filipczyk, P.S. Hoppe, M. Strasser, F. Buggenthin, J.S. Feigelman, J. Krumsiek, D. Loeffler, K.D. Kokkaliaris, A.J.J. van den Berg, M. Endele, S. Hastreiter, C. Marr, F.J. Theis

and T. Schroeder. Single-cell quantification of cellular and molecular behavior in long-term time-lapse microscopy, *in preparation*

6.1 Biological setup

To assess the performance of the adaptive Simpson's rule for thermodynamic integration for more complex problems than presented in the previous chapter, we used the method to study time-lapse microscopy data of single cells, as introduced in Section 2.1.

Proteins produced by the cells decay with to-be-determined half-life. However, these protein half-lives are important for assessing all models based on protein expression in cells. To assess protein stability, cells are treated with cycloheximide to block protein translation (Halter *et al.* [2007]), allowing to observe the protein decay of fluorescently labeled proteins over time in individual cells. The protein stability following cycloheximide treatment can be assessed on the basis of measurement data for two very different cell types.

The first cell type was measured by Halter *et al.* [2007]. This data set provides automatically quantified single-cell time-lapse fluorescence microscopy data for fibroblast cells. A fibroblast is a connective tissue cell that secretes proteins which are important for the structural framework of cells. Genetically modified fibroblasts produce an enhanced, destabilized green fluorescent protein (GFP) reporter. All in all, over 500 cells were quantified. Of these, we randomly chose 200 as the basis for our inference. These 200 cells were automatically measured every 15 minutes for 12 hours, yielding 49 data points for each cell. Measurement data of ten representative cells is depicted in Figure 6.1.

The second data set was obtained through computer-assisted single-cell time-lapse microscopy by Schwarzfischer *et al.* [2014] as introduced in Section 2.1. Here, primary murine granulocyte/macrophage progenitor (GMP) cells are observed. The protein PU.1 is thought to play an important role in the decision making process of the cell (Scott *et al.* [1994]). In (genetically modified) GMPs, the protein PU.1 is produced in a fluorescent variant (PU.1eYFP) and can thus be observed through fluorescence microscopy (Kirstetter *et al.* [2006]). GMPs are observed in three replicate experiments with 45, 46 and 48 cells, respectively. A visualization of this data can also be seen in Figure 6.2. Already a first look at the data reveals that the data is rather heterogeneous. The data for some cells shows a much more pronounced decay than for other

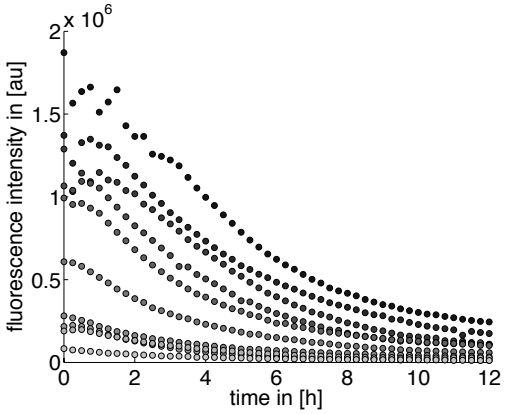


Figure 6.1: Fibroblast cell trajectories. Trajectories of 10 randomly chosen fibroblast cells, each gray scale indicates one cell. Cells are aligned such that cycloheximide was added at time $t = 0$. The fluorescence intensity decays with time, however it does not decay completely in the observed time frame. Cells were automatically imaged every 15 minutes for 12 hours.

cells. Furthermore, some cells can be observed only briefly after cycloheximide treatment, while others live up to 40 hours after treatment. This yields a variable amount of data points for each cell, between 11 and 77. It is noteworthy that for experiment 3, the observation was stopped after about 20 hours. We summarize the properties of our data in Table 6.1. The heterogeneity of the GMP data here might indicate a systematic heterogeneity in the GMP population, which is the topic of current research, as these are primary cells which might actually belong to several subgroups (Hoppe *et al.* [2014]).

Cell type	**# cells**	**# of data points**
Fibroblasts	200	49
GMP Replicate	**# cells**	**# of data points (mean, [min, max])**
Replicate 1	45	$25.3, [11, 66]$
Replicate 2	46	$35.6, [11, 77]$
Replicate 3	48	$23.1, [10, 40]$

Table 6.1: Overview of available measurement data. Number of cells and number of data points for each cell in all available data sets. For the fibroblasts, all cells have the same number of data points. For the GMPs, we give the mean, minimum and maximum number of data points measured for the individual cells.

Experiments like the presented ones, where the decay of the protein can be observed

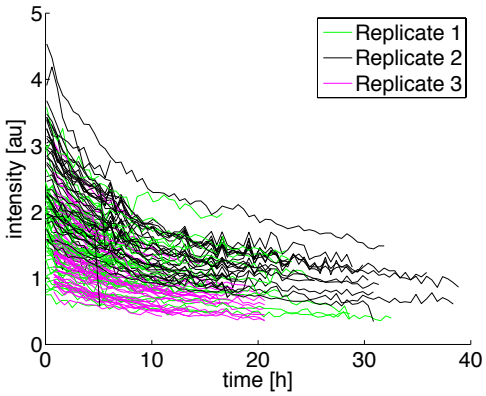


Figure 6.2: GMP cell trajectories. Fluorescence of PU.1 in GMPs for three replicates (color-coded). Cells are aligned such that cycloheximide was added at time $t = 0$. The fluorescence intensity decays over time, but does not decay completely in the observed time frame. Measurements for individual cells are taken as long as possible, yielding different observation times for each cell. Note the differences between replicates, e.g. in experiment 3, all measurements were stopped after about 20 hours.

isolated for one protein in single cells are an important basis for more complex models, where several proteins may interact with each other. A rigorous statistical evaluation as presented in the following sections cannot distinguish hidden biological mechanisms, but gives insights into the behavior of the system and is thus a very important cornerstone for further inference in even more complex systems.

6.2 Model selection task

The conducted experiments allow us to observe the protein decay over time in individual cells. By fitting a simple exponential decay model to the decreasing fluorescence intensity of every cell, Halter *et al.* [2007] infer protein half-lives in the fibroblasts. Interestingly, the original contribution here poses a model selection problem: it was shown that a simple exponential model fits rather well at early time points, however, at the end, the curves seem to be systematically too low to fit the data properly, see also Figure 6.1 for 10 exemplary single-cell data trajectories. This encourages the following model choice problem:

- The fluorescence intensity decays completely (model $\mathbf{M_1}$)
- There exists a non-zero steady state level for the fluorescence intensity (model

$\mathbf{M_2}$).

Both models can be formulated as ordinary differential equations (ODEs), which can be solved analytically.

Also for the GMPs in Figure 6.2, we observe that many of the single-cell time courses do not seem to decay to zero, in neither of the three replicates. Thus we are faced with the same model selection problem also for the GMPs, so that we can use the same models as for the fibroblasts.

If we let $z_{(i)}(t)$ be the time dependent fluorescence intensity according to model $\mathbf{M}_i$, the ODEs for the model can be written down as follows. The first model $\mathbf{M_1}$ for the fluorescence intensity $z_{(1)}(t)$ is $\dot{z}_{(1)}(t) = -\rho z_{(1)}(t), z_{(1)}(0) = \alpha$. Solving the ODE yields $z_{(1)}(t) = \alpha \exp(-\rho t)$. The ODE for the second model $\mathbf{M_2}$ is $\dot{z}_{(2)}(t) = -\rho(z_{(2)}(t) - \gamma), z_{(2)}(0) = \tilde{\alpha}$. Solving this leads to $z_{(2)}(t) = \alpha \exp(-\rho t) + \gamma$, where $\alpha = \tilde{\alpha} - \gamma$. In both models, ρ is the decay rate. We reformulate both models: the parameter $\beta = \frac{\log(2)}{\rho}$ is the half-life time of the proteins which are degraded. Knowing this half-life is of biological relevance. In model $\mathbf{M_2}$ we have an additional parameter γ corresponding to the steady state level of fluorescence intensity, as $z_{(2)}(t \to \infty) \to \gamma$. In model $\mathbf{M_1}$, the parameter α controls initial intensity. In model $\mathbf{M_2}$, the initial value is $z_{(2)}(0) = \tilde{\alpha} := \alpha + \gamma$, to facilitate the comparison of parameter values with model $\mathbf{M_1}$. As previously suggested for fluorescence intensities (Harper *et al.* [2011]), we assume multiplicative gamma distributed noise distributed according to $\Gamma(k, 1/k)$ for both models.

The analytical solutions for the two models then read:

- $z_{(1)}(t) = \alpha \exp\left(-\frac{log(2)t}{\beta}\right)$ for model $\mathbf{M_1}$,
- $z_{(2)}(t) = \alpha \exp\left(-\frac{log(2)t}{\beta}\right) + \gamma$ for model $\mathbf{M_2}$.

In the case at hand, we actually have a fully observed system, since the time courses from the models are directly observed, meaning that

$$\mathcal{Y}_{(i)}(t) = z_{(i)}(t), \ i = 1, 2 \tag{6.1}$$

for the observable $\mathcal{Y}_{(i)}(t)$ of model $\mathbf{M}_i$.

6.3 Single-cell inference

In this section, we formulate the likelihoods and posteriors for the chosen models. This includes specifying a noise model and prior distributions. Furthermore, we use a bootstrapping approach to verify that the models can in principle explain the measurement data.

6.3.1 Set-up of likelihood and posterior

For the fibroblasts, we randomly selected 200 cells from Halter *et al.* [2007]. It is apparent from the data in Figure 6.1 that the individual cells show different decay properties, e.g. initial intensities. To address this, we considered cell-to-cell variability and assume that all parameters can differ between individual cells. We thus use have a cell-specific likelihood in model $\mathbf{M}_i$

$$p(\mathbf{Y}_r|\boldsymbol{\theta}_r^i, \mathbf{M}_i) = \prod_{n=1}^{N_r} \phi^{\Gamma}\Big(Y_{r,n}; k_r, (\mathcal{Y}_{(i)}(t_{r,n})/k_r)\Big), \tag{6.2}$$

where r is the index of the cell, $\mathbf{Y}_r$ the data for cell r consisting of the individual data points $Y_{r,n}$ taken at $t_{r,n}$ for $n = 1, \ldots, N_r$, where N_r is the number of measurements for cell r. As introduced in Section 2.2.1, $\phi^{\Gamma}\Big(x; k, \lambda\Big)$ is the probability density function of the univariate gamma distribution evaluated at x with shape k and scale λ. The special form in Equation (6.2) corresponds to multiplicative noise of mean 1 as introduced in Equation (2.30). The noise parameter k_r is also specific for each cell r. The parameter vectors $\boldsymbol{\theta}_r^1 = (\alpha_r, \beta_r, k_r)$ or $\boldsymbol{\theta}_r^2 = (\alpha_r, \beta_r, \gamma_r, k_r)$ are the parameters for cell r within the respective model.

Thus the individual cells do not actually share any parameters, but are considered independent. This is preferable, since we want to make use of the availability of single-cell measurement data. We apply uniform priors of biologically reasonable upper and lower bounds for all parameters. More precisely, we assume prior distributions as shown in Table 6.2. The prior distributions for α and γ for fibroblasts and GMPs differ due to the different scales of fluorescence intensities in the two data sets as seen in Figures 6.1 and 6.2, since these two parameters have to match the scale of fluorescence intensity. This also means that α and γ have the same unit as the fluorescence intensity, in our case thus $[au]$, while β has the unit hours, $[h]$, and k does not have a unit. We assume no prior dependence between the individual parameters.

Cell type	Model	Parameter	Prior distribution	Unit of parameter
Fibroblasts	$\mathbf{M_1}, \mathbf{M_2}$	α	$\mathcal{U}[0, 3 \cdot 10^6]$	$[au]$
Fibroblasts	$\mathbf{M_1}, \mathbf{M_2}$	β	$\mathcal{U}[0.1, 150]$	$[h]$
Fibroblasts	$\mathbf{M_2}$	γ	$\mathcal{U}[0, 1 \cdot 10^5]$	$[au]$
Fibroblasts	$\mathbf{M_1}, \mathbf{M_2}$	k	$\mathcal{U}[0, 5000]$	-
GMPs	$\mathbf{M_1}, \mathbf{M_2}$	α	$\mathcal{U}[0, 10]$	$[au]$
GMPs	$\mathbf{M_1}, \mathbf{M_2}$	β	$\mathcal{U}[0.1, 150]$	$[h]$
GMPs	$\mathbf{M_2}$	γ	$\mathcal{U}[0, 10]$	$[au]$
GMPs	$\mathbf{M_1}, \mathbf{M_2}$	k	$\mathcal{U}[0, 5000]$	-

Table 6.2: Overview over used prior distributions.

Each of the resulting three- or four-dimensional single-cell posterior distributions $p(\boldsymbol{\theta}_r^i|\mathbf{Y}_r, \mathbf{M}_i)$ can be inferred by Adaptive Metropolis sampling from Section 4.2.

If we want to make general statements about protein stability after cycloheximide treatment, it is desirable to consider the information for all cells from one data set at the same time, because in this case, we get one answer based on the consensus of the single-cell analyses. Since the cells are considered independent, we can easily define a likelihood for the ensemble of cells by

$$p(\mathbf{Y}|\boldsymbol{\theta}^i, \mathbf{M}_i) = \prod_{r=1}^{R} p(\mathbf{Y}_r|\boldsymbol{\theta}_r^i, \mathbf{M}_i), \tag{6.3}$$

where $p(\mathbf{Y}_r|\boldsymbol{\theta}_r^i, \mathbf{M}_i)$ is the individual likelihood of each cell as defined in Equation (6.2), $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_R)$ is the collection of all the data in this data set, $\boldsymbol{\theta}^i = (\boldsymbol{\theta}_1^i, \ldots, \boldsymbol{\theta}_R^i)$ is the parameter vector obtained by concatenating all individual cell parameter vectors and R is the number of cells in the data set. Assuming the prior $p(\boldsymbol{\theta}^i|\mathbf{M}_i) = \prod_{r=1}^{R} p(\boldsymbol{\theta}_r^i|\mathbf{M}_i)$ leads then also the posterior distribution $p(\boldsymbol{\theta}^i|\mathbf{Y}, \mathbf{M}_i)$.

For the fibroblast data, we have $R = 200$ and thus 600 parameters are inferred for model $\mathbf{M_1}$ and 800 parameters for model $\mathbf{M_2}$. The ensemble of all single-cell sampling results is then the sampling result for $p(\boldsymbol{\theta}^i|\mathbf{Y}, \mathbf{M}_i)$. This can also be done accordingly for sampling from the power posterior.

For the GMPs, each of the three replicates is considered separately, but the cells within the replicate are considered together just as for the fibroblasts. Thus the likelihood

reads

$$p_\kappa(\mathbf{Y}_\kappa|\boldsymbol{\theta}^i, \mathbf{M}_i) = \prod_{r=1}^{R_\kappa} p(\mathbf{Y}_{\kappa,r}|\boldsymbol{\theta}^i_r, \mathbf{M}_i), \tag{6.4}$$

where $\kappa = 1, 2, 3$ is the index of the replicate and R_κ is the number of cells within this replicate. $\mathbf{Y}_\kappa = (\mathbf{Y}_{\kappa,1}, \ldots, \mathbf{Y}_{\kappa,R_\kappa})$ is the data of replicate κ. Also here, specifying the prior as $p(\boldsymbol{\theta}^i|\mathbf{M}_i) = \prod_{r=1}^{R} p(\boldsymbol{\theta}^i_r|\mathbf{M}_i)$ leads to the posterior distribution $p(\boldsymbol{\theta}^i|\mathbf{Y}, \mathbf{M}_i)$. We treat the three replicates separately, since already the raw data in Figure 6.2 showed differences between the replicates which we do not want to neglect. Coinciding results for the model selection in the three replicates are then a stronger argument than just one model choice.

6.3.2 Assessing the goodness-of-fit

Since at least model $\mathbf{M_2}$ has not yet been used on similar measurement data, a desirable sanity check for this model, but also for model $\mathbf{M_1}$, is the bootstrap of the goodness-of-fit as introduced in Section 2.4.3. We thus want to investigate if the chosen models can fit the individual single-cell data at all. We thus consider the MLE $\widehat{\boldsymbol{\theta}^i_r}$ for each cell obtained by multi-start local optimization. From this we obtain $s^*_{i,r} = \log p(\mathbf{Y}_r|\widehat{\boldsymbol{\theta}^i_r}, \mathbf{M}_i)$, the maximum value of the log-likelihood in model $\mathbf{M}_i$ for cell r as a goodness-of-fit value. For each cell, we then generate $J_{\mathbf{BS}} = 500$ artificial single-cell measurements from the model parametrized with $\widehat{\boldsymbol{\theta}^i_r}$. Re-optimization then yields the parameter vector best fitting this new data set, $\boldsymbol{\theta}^i_{r,j}$ with $j = 1, \ldots, J_{\mathbf{BS}}$. This then yields new log-likelihood values $s^j_{i,r}, j = 1, \ldots, J_{\mathbf{BS}}$. From these bootstrap values from the unknown distribution of the log-likelihood, we can then calculate the z-score of $s^*_{i,r}$ in the empirical distribution of the $s^j_{i,r}, j = 1, \ldots, J_{\mathbf{BS}}$ as introduced in Equation (2.33), indicating if model $\mathbf{M}_i$ is able to fit the data of cell r in principle. A low absolute value of the z-score shows that the true log-likelihood value is in the range of the bootstrap samples. For this the distribution of the $s^j_{i,r}, j = 1, \ldots, J_{\mathbf{BS}}$ should be unimodal, which has to be checked and is fulfilled in our case.

Due to the high computational costs for generating artificial data and re-optimizing the parameters for all cells in the data set, we restrict our analysis to Replicate 3 of the GMP data. We present our findings in Table 6.3. The distribution of all computed z-scores can also be seen in Figure 6.3.

As all absolute values of the z-scores are less than one, we conclude that the true value is well within the range of the bootstrapped values, which can also be seen in Figure

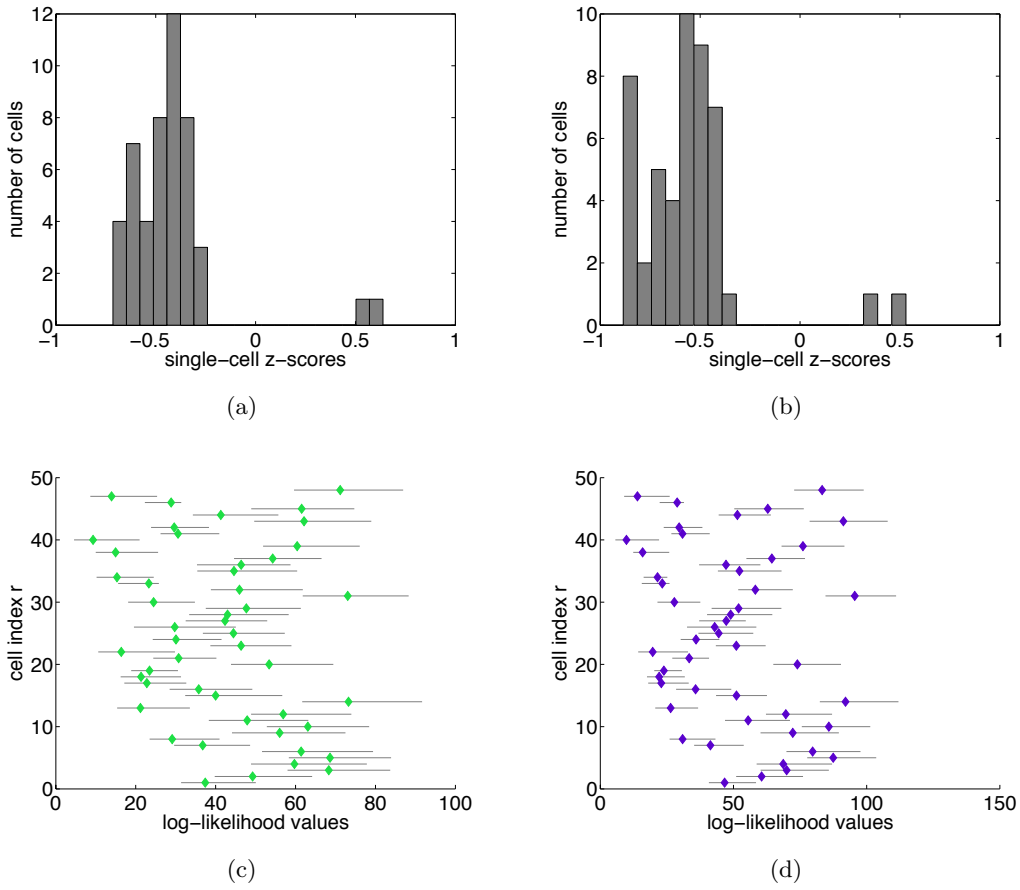


Figure 6.3: Single-cell z-scores from the bootstrap. (a) Z-scores for model $\mathbf{M_1}$ for the single cells computed individually based on 500 bootstrap samples each. (b) Z-scores for model $\mathbf{M_2}$ for the single cells computed individually based on 500 bootstrap samples each. (c) True log-likelihood values (diamonds) and range of bootstrap sample values (grey lines) for all 48 cells individually for model $\mathbf{M_1}$. (d) True log-likelihood values (diamonds) and range of bootstrap sample values (grey lines) for all 48 cells individually for model $\mathbf{M_2}$.

Model	mean value of z-score	min. and max. of z-score
$\mathbf{M_1}$	-0.43	$[-0.71, 0.64]$
$\mathbf{M_2}$	-0.56	$[-0.88, 0.53]$

Table 6.3: Overview of z-scores from the bootstrap for the goodness-of-fit. For all 48 cells in Replicate 3 of the GMP data, we generate 500 new datasets, refit them and compute log-likelihood values. From these 500 samples, we compute a z-score for the true log-likelihood value of the MLE for each cell. This table shows the mean of the absolute values of the z-score as well as its minimum and maximum values.

6.3(c) and (d) and thus both models can in principle fit the data observed for individual cells. This illustrates the need for thorough model selection. It should be noted that model $\mathbf{M_1}$ assumes a higher level of measurement noise for fitting the data.

We take a closer look at one cell, cell 39, whose individual z-scores are close to the mean z-scores in both models. We see in Figure 6.4 that the histogram of the bootstrapped log-likelihood values is unimodal and the true log-likelihood values of the MLEs are well contained in the range of samples. Thus both models can fit the data, which is also shown in the last panel of Figure 6.4 for this cell.

Furthermore, only two cells have a positive z-score, while 46 have a negative z-score for both models. We take a closer look at these two cells that have positive z-scores, cells 33 and 46. We find that the fits for model $\mathbf{M_1}$ and $\mathbf{M_2}$ are indistinguishable for both cells, as the MLE for γ in model $\mathbf{M_2}$ is very close to zero for both cells, see also Figure 6.5(a) and (b). Furthermore, both cells are observed only very shortly, with ten time points covering just over 5h. This leads to a situation where both models tend more to overfitting the data than underfitting as for the other 46 cells, as indicated by the histograms of the bootstrapped samples in Figure 6.5(c)-(f). However, the bulk of the data has a negative z-score, indicating that both models have a tendency to underfitting.

6.4 Parameter distributions and identifiability analysis

In this section, we infer the posterior distributions of both models and conduct an analysis of the resulting parameter distributions as well as identifiability analysis on the likelihood.

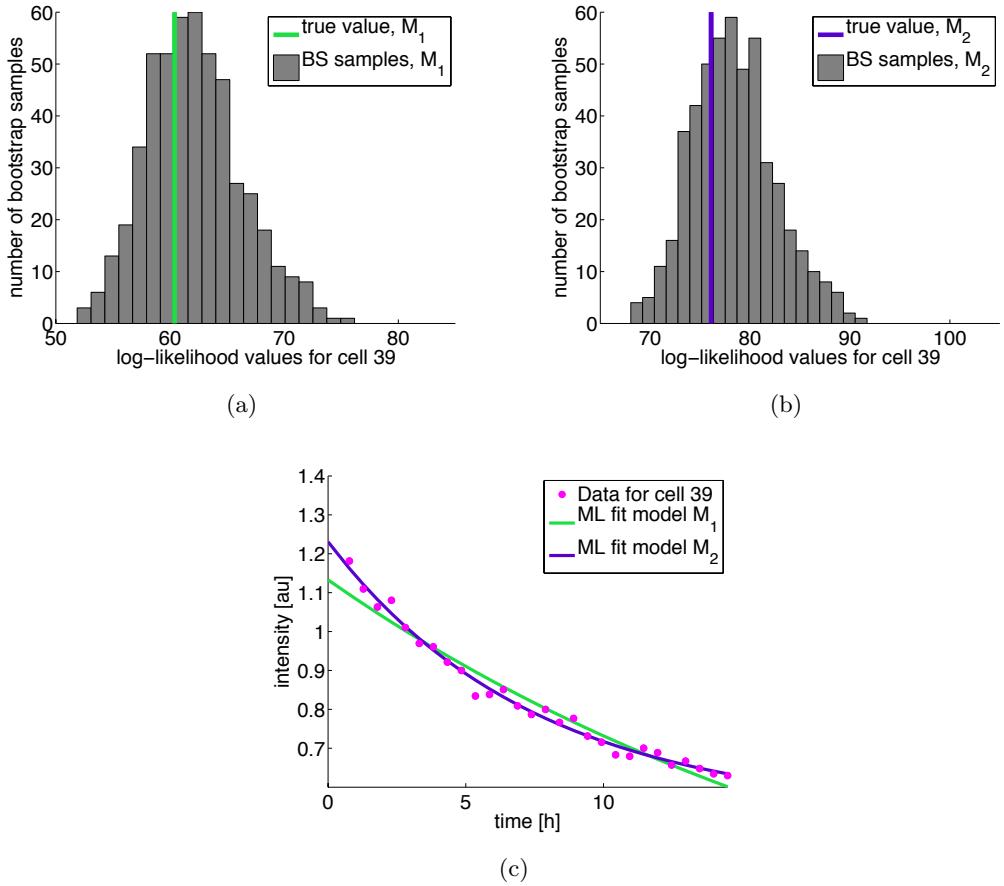


Figure 6.4: A single-cell example for the bootstrap. (a) Histogram for model $\mathbf{M_1}$ for cell 39 based on 500 bootstrap (BS) samples, with a green line for the true maximum likelihood (ML) value. (b) Histogram for model $\mathbf{M_2}$ for cell 39 based on 500 bootstrap (BS) samples, with a purple line for the true maximum likelihood (ML) value. (c) Maximum likelihood (ML) fits (solid lines) to cell 39 (magenta dots) for both models.

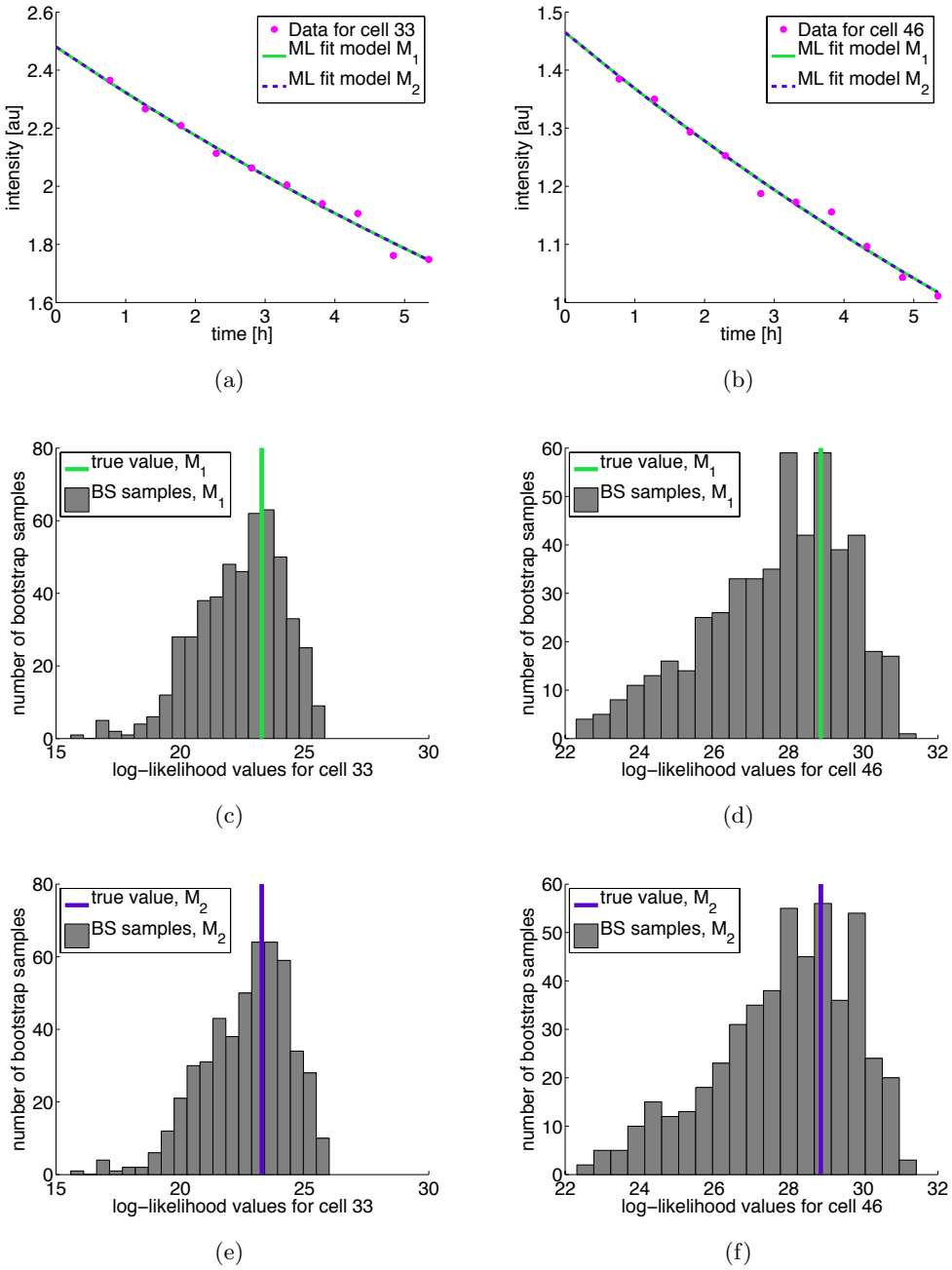


Figure 6.5: Single-cell examples for positive z-scores. (a) Maximum likelihood (ML) fits (solid lines) to cell 33 (magenta dots) for both models. (b) Maximum likelihood (ML) fits (solid lines) to cell 46 (magenta dots) for both models. (c) Histogram for model $\mathbf{M_1}$ for cell 33 based on 500 bootstrap (BS) samples, with a green line for the true maximum likelihood value. (d) Histogram for model $\mathbf{M_1}$ for cell 46 based on 500 bootstrap (BS) samples, with a green line for the true maximum likelihood value. (e) Histogram for model $\mathbf{M_2}$ for cell 33 based on 500 bootstrap (BS) samples, with a purple line for the true maximum likelihood value. (f) Histogram for model $\mathbf{M_2}$ for cell 46 based on 500 bootstrap (BS) samples, with a purple line for the true maximum likelihood value.

6.4.1 Parameter distributions

For the fibroblasts, a mean half-life of the protein of 2.8h was reported in the previous analysis (Halter *et al.* [2007]). This was however not based on a full Bayesian approach, but on a χ^2-fit for each cell. We have inferred a full distribution of half-lives for each cell, since this comes naturally from the posterior samples drawn for thermodynamic integration. For model $\mathbf{M_1}$, we find a mean half-life of cells of 3.6h, taken over all posterior samples for the half-lives, with a 90% credible interval of $[2.84\text{h}, 4.69\text{h}]$. For model $\mathbf{M_2}$ in contrast, we find a lower half-life than in the previous paper, namely 2.3h, with a 90% credible interval of $[1.49\text{h}, 3.30\text{h}]$.

For the GMPs, we find from model $\mathbf{M_1}$ protein half-lives of 25.00h, 18.29h and 13.60h, taken over all posterior samples in the replicates. We pool all posterior samples from the three replicates, i.e. 139 sample-based distributions of 50000 samples each. This is an unweighted sum of sample distributions. From this we find a mean half-life of 18.94h with a 90% credible interval of $[9.4\text{h}, 36.2\text{h}]$.

For model $\mathbf{M_2}$, we find mean half-lives of 6.02h, 7.17h and 9.98h, taken over all posterior samples in the three replicates. When pooling all posterior samples from the three replicates, we find a mean of 7.78h, with a 90% credible interval of $[2.1\text{h}, 19.5\text{h}]$. This agrees with the maximum likelihood fit of Schwarzfischer *et al.* [2014], where we reported 9.2h$\pm$7.7h (median $\pm$ standard deviation) and also with Nutt *et al.* [2005], who reported 5.5h, but from Western blot data. The large discrepancy between the credible intervals derived from the two models illustrates the need for model selection to clarify which results are a better representation of the measurement data.

Since the intervals for the GMPs are rather broad, we now inspect the distributions and the time courses they induce. We focus on model $\mathbf{M_2}$, the situation for model $\mathbf{M_1}$ is similar. Figure 6.6 shows two versions (truncated and untruncated) of the histogram of all available MCMC samples from model $\mathbf{M_2}$ for PU.1 half-life, showing a clear unimodal distribution, with a long tail to longer half-lives.

We can also compute for each single cell the empirical coefficient of variation c_v for posterior samples from model $\mathbf{M_2}$ of the half-life in this cell. The coefficient of variation is a normalized measure of dispersion of a distribution (Klipp *et al.* [2008]), defined by

$$c_v = \frac{\sigma}{\mu}. \tag{6.5}$$

Here μ is the mean of the distribution and σ its standard deviation. Since we here look at sample distributions, we use the sample mean and sample standard deviation.

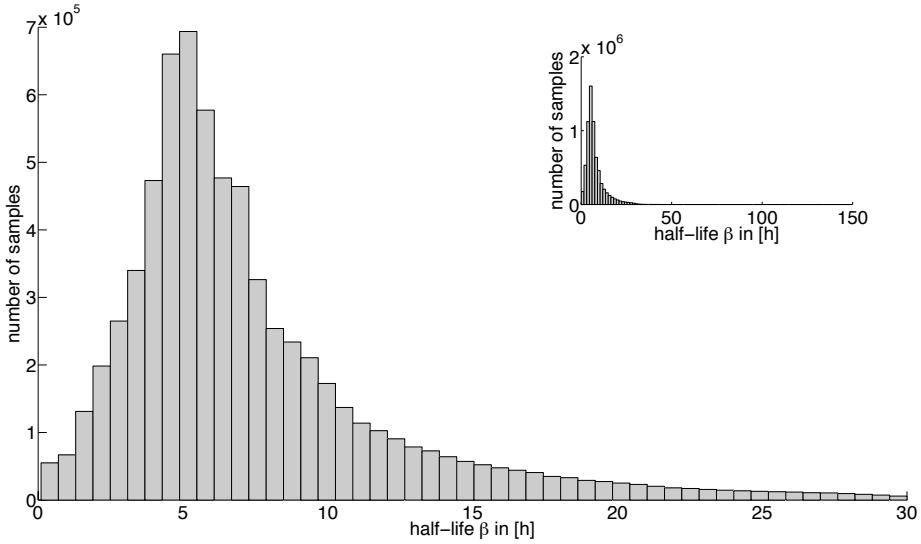
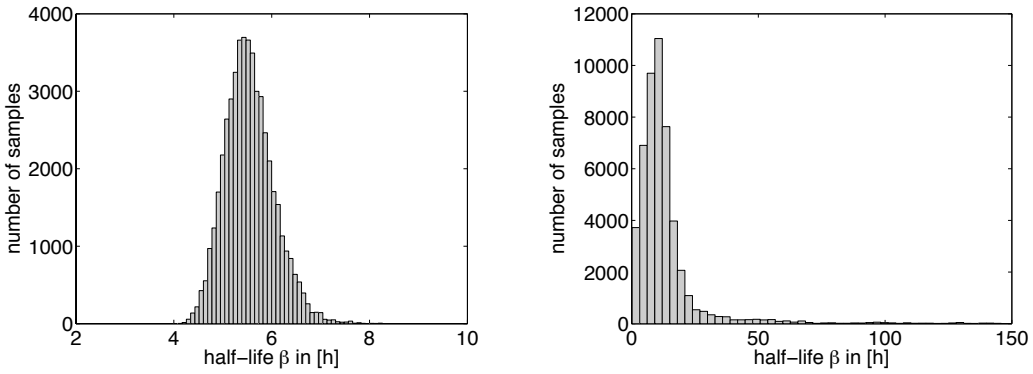


Figure 6.6: Histogram of all obtained MCMC samples from model M_2 for PU.1 half-life. This histogram is based on all available MCMC samples for PU.1 half-life, truncated at 30h for better visualization, showing a clear unimodal distribution. The small inset figure shows the untruncated histogram where it can be seen that small numbers of samples reach all the way to 150h, the boundary of the prior distribution.

We find values from 0.03 to 1.07, with a mean of 0.31, indicating that the distribution of half-lives within a cell is not too broad. We visualize this by looking at a cell with posterior samples for the half-life with a low c_v of 0.09, Figure 6.7(a), and a cell with posterior samples for the half-life with a high c_v of 1.07, Figure 6.7(b).



(a) (b)

Figure 6.7: Single-cell histograms of MCMC samples of model $\mathbf{M_2}$ for the half-life of PU.1. (a) Posterior samples for a cell with low parameter uncertainty ($c_v = 0.09$) for the half-life. (b) Posterior samples for a cell with high parameter uncertainty ($c_v = 1.07$) for the half-life. Note that in this histogram noticeably more samples for half-lives > 50h exist than in (a). (a-b) Both histograms are based on 50000 posterior samples each.

The posterior samples also confirm what the first look at the data in Figure 6.2 already hinted at: that the three replicates show somewhat different behavior. We can look at the histograms for the pooled MCMC samples for the half-life in each replicate separately, Figure 6.8 shows the situation for model $\mathbf{M_2}$. We immediately see that replicate 3 has a much narrower distribution than the other two replicates, with fewer samples in the tail towards longer half-lives. This can also be seen in the lower coefficient of variation from the pooled samples of replicate 3 of $c_v = 0.42$ compared with $c_v = 0.78$ for replicate 1 and $c_v = 0.96$ in replicate 2. This is certainly due to the fact that in replicate 3, observation stopped after about 20 hours. Table 6.4 shows that also the single-cell coefficients of variation and the credible intervals for the protein half-lives differ between the replicates, for both models. We thus treat each replicate separately in the coming model selection in the following section.

For the time courses we find that model $\mathbf{M_2}$ fits rather well, but the fit can look very different due to the highly different data. Figures 6.9(a) and 6.9(b) compare two fits for a cell where the posterior samples for the half-life have a high c_v and a cell where these

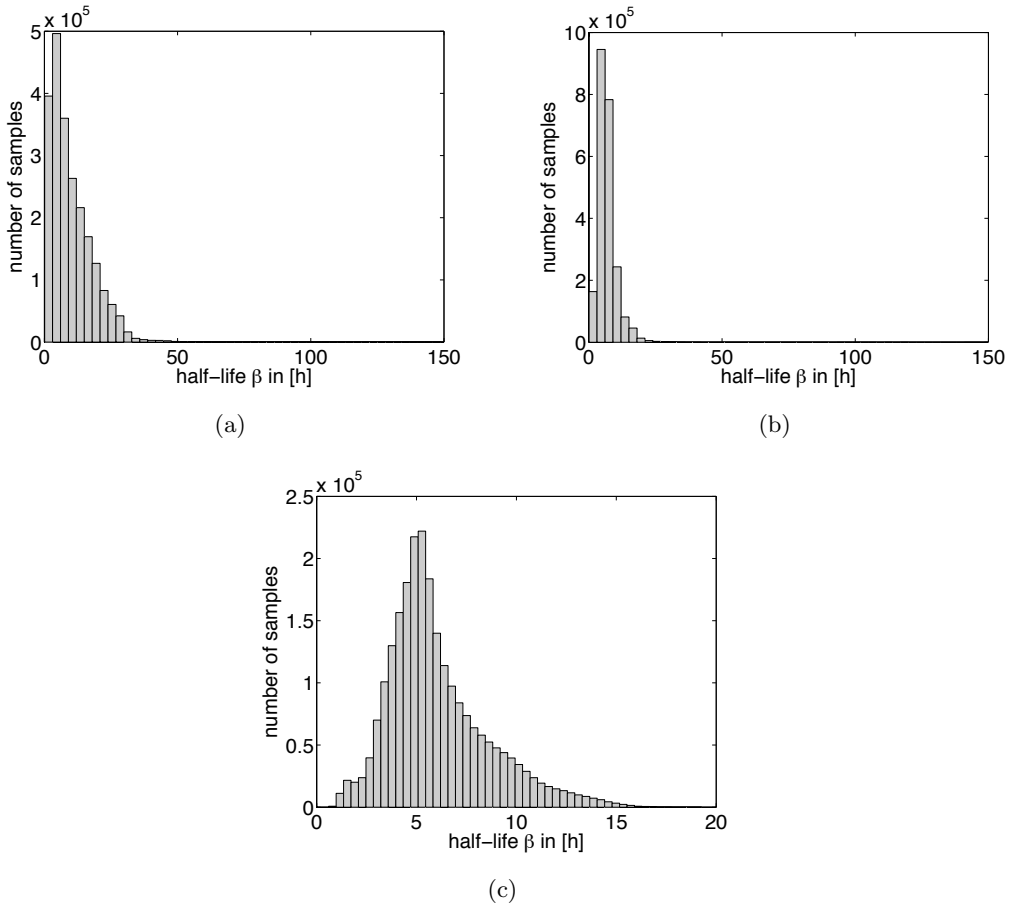


Figure 6.8: Histograms of MCMC samples from model M_2 for the half-life of PU.1 from the three replicates. (a) Pooled samples for the half-life from the first replicate. (b) Pooled samples for the half-life from the second replicate. (c) Pooled samples for the half-life from the third replicate. (a-c) Note that replicate 3 has a lower coefficient of variation c_v from the pooled samples of 0.42 compared with 0.78 for replicate 1 and 0.96 in replicate 2, which is also visible in the histogram.

Model	Replicate	c_v, mean, [min, max]	Half-life, mean & 90% CI
$\mathbf{M_1}$	Replicate 1	0.17, [0.03, 0.87]	25.0h, [10.8h, 49.9h]
$\mathbf{M_1}$	Replicate 2	0.11, [0.03, 1.19]	18.3h, [9.4h, 30.0h]
$\mathbf{M_1}$	Replicate 3	0.05, [0.02, 0.14]	13.6h, [9.0h, 19.4h]
$\mathbf{M_2}$	Replicate 1	0.46, [0.12, 1.07]	6.0h, [1.6h, 25.12h]
$\mathbf{M_2}$	Replicate 2	0.22, [0.04, 0.54]	7.2h, [2.6h, 13.6h]
$\mathbf{M_2}$	Replicate 3	0.23, [0.08, 0.76]	10.0h, [2.8h, 11.1h]

Table 6.4: Overview of differences between replicates for the GMPs. For all three replicates of the GMP data and both models, we give the mean coefficient of variation (c_v) and its minimum and maximum value as well as the mean and 90% credible interval (CI) for the protein half-life based on the posterior samples for the individual cells.

posterior samples have a low c_v. We note that the credible interval is frayed at the ends of the time line in the first case. Figures 6.9(c) and 6.9(d) compare two cells with very short mean half-life and very long mean half-life, based on the posterior samples. We note that the cell with short half-life of 0.36h has been observed only for ca. 5h, while the cell with long half-life of 26.0h has been observed for over 28h. Nevertheless, the model fit shows a good agreement with the measurement data for all the different cells.

6.4.2 Comparison of GMPs with other stem cells

The credible intervals for the GMPs are very broad, which is biologically surprising. We thus conducted several sanity checks by comparing with results for other stem cells.

In Schwarzfischer *et al.* [2014], we showed that the GMPs there showed a half-life of $9.2\text{h} \pm 7.7\text{h}$ (median $\pm$ standard deviation), as already mentioned. This median and standard deviation are based on single-cell maximum likelihood fits of model $\mathbf{M_2}$. In this contribution, several other types of hematopoietic stem cells were also analyzed with the same methods, yielding results for the half-life of PU.1 of $7.3\text{h} \pm 2.0\text{h}$ (HSCs), $7.9\text{h} \pm 4.7\text{h}$ (MPP cells) and $7.0\text{h} \pm 2.4\text{h}$ (Gatamid cells), for median $\pm$ standard deviation, respectively. This already shows that the GMP data shows a broader range of decay values than other cells. The high standard deviation of the GMPs is an indicator for high biological variability, since it is based on single-cell analyses. This indicates increased heterogeneity, as was also recently suggested in Hoppe *et al.* [2014].

The Gatamid cells of Schwarzfischer *et al.* [2014] also provide another interesting ver-

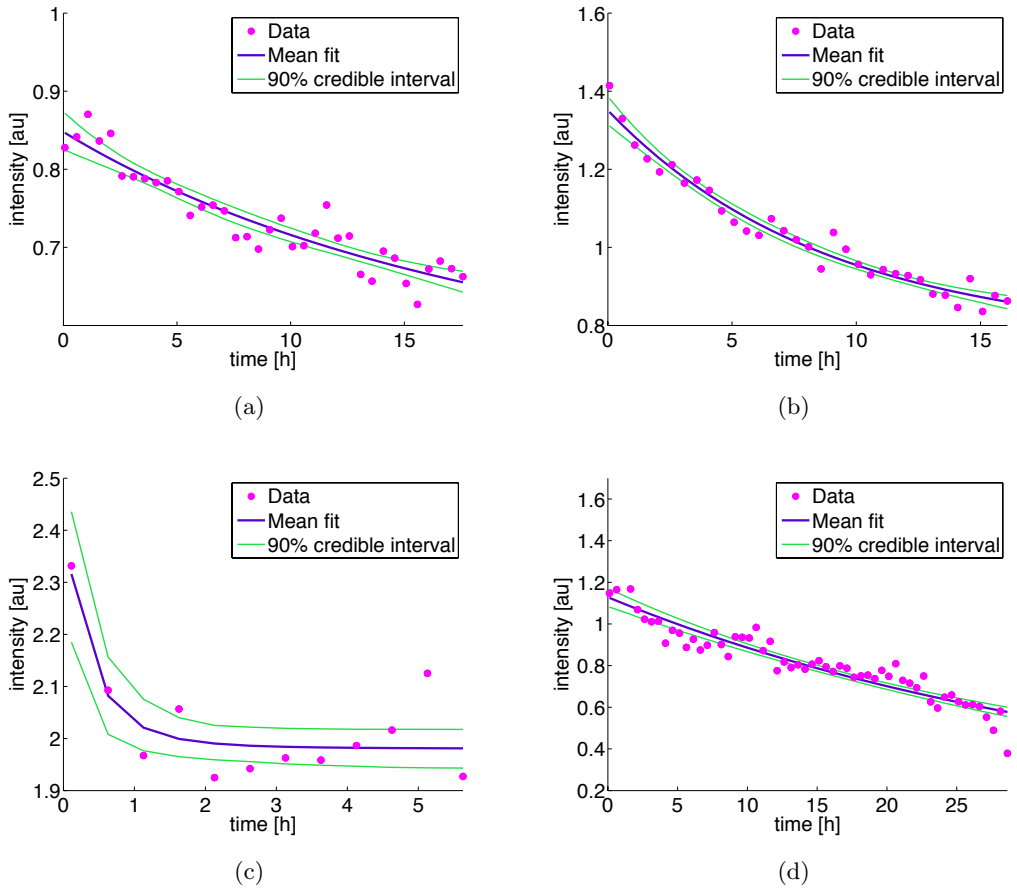


Figure 6.9: Time-courses of PU.1 intensity from model $\mathbf{M_2}$. (a) Data and fit for a cell with high coefficient of variation c_v. Note the broadening of the credible interval at both ends of the time line. (b) Data and fit for a cell with low coefficient of variation c_v, where the credible interval is quite narrow. (c) Data and fit for a cell with a very short half-life of 0.36h. Note that only few time points (12) are available. (d) Data and fit for a cell with a very long half-life of 26.0h. This cell has been measured for 57 time points. (a-d) All time courses come from model $\mathbf{M_2}$. Magenta dots are the measurement data, purple lines the mean fit, green lines the boundaries of the 90% credible intervals derived from the posterior samples of the respective cell. Note that the mean and credible interval represent only the uncertainty in the parameters, in contrast to measurement uncertainty (not shown).

ification of results. In these cells, not only PU.1 is fluorescently expressed, but also another important protein, namely Gata1. Both proteins have been given different colors of fluorescence and can thus be imaged in the same cells simultaneously, for details see Schwarzfischer *et al.* [2014]. If the broad distribution of half-lives for the GMPs was due to some damaging influence of the cycloheximide in these cells, we would expect to also see this effect in both the PU.1 and Gata1 intensities in the Gatamid cells. More precisely, we would expect the half-lives of PU.1 and Gata1 to be significantly correlated. We obtained the MLE of the half-lives from model $\mathbf{M_2}$ for each single cell. Fortunately though, we find that for all three available replicates of Gatamid cells, the Kendall's τ between the half-lives of PU.1 and Gata1 are rather small (0.31, -0.22 and 0.16). We thus conclude that the half-life distribution of the GMPs cannot be solely due to some undesirable side-effects of cycloheximide treatment. Still the model choice to be performed in the following section asks if cycloheximide treatment works perfectly or not, as this could be a biological reason for the fluorescence intensity to decay to zero or not, respectively.

We also find in Schwarzfischer *et al.* [2014] that no correlation can be found between cell cycle time at the time of cycloheximide treatment and protein half-life. This data is unfortunately only available for the protein Nanog in embryonic stem cells. However, we also find no significant correlation between initial intensity of the cells and their half-life in the GMPs, which can be taken as an indicator that also in GMPs, there is no correlation between cell cycle time and half-life.

6.4.3 Uncertainty analysis for credible intervals

To see if the credible intervals are dominated by outliers, we also did a bootstrap resampling of the credible intervals (Davison & Hinkley [1997]). We concentrate on model $\mathbf{M_2}$. For this, we sampled uniformly a 1000 times with replacement the indices of cells from all 139 cells contributing to the credible interval of $[2.1\text{h}, 19.5\text{h}]$, which is at question, and recomputed the credible intervals based on the MCMC samples for the half-lives of the chosen bootstrap cells. This yields 90% confidence intervals for the lower boundary of the original credible interval of $[1.51\text{h}, 2.46\text{h}]$ and $[17.86\text{h}, 23.00\text{h}]$ for the upper boundary of the original credible interval, computed according to Equation 5.6 in Davison & Hinkley [1997]. We thus see that both boundaries are well-contained in the respective bootstrapped confidence intervals. Also for our reported mean 7.78h, we find a bootstrap confidence interval of $[7.23\text{h}, 8.52\text{h}]$, which obviously contains 7.78h.

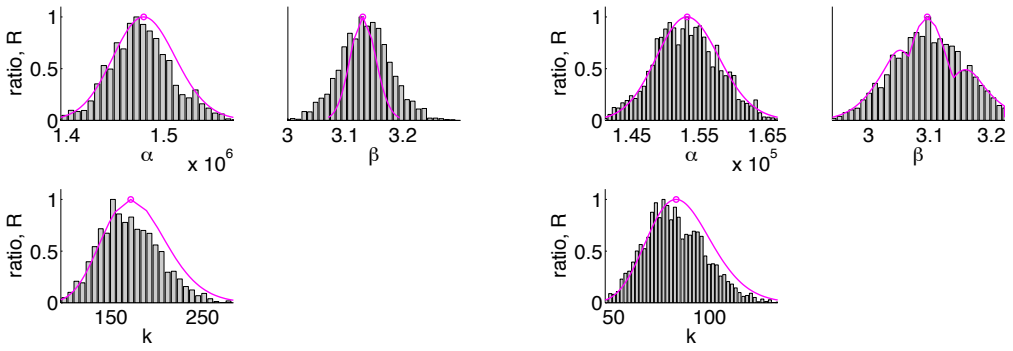
6.4.4 Identifiability analysis

The question of the credible intervals leads directly also to the question of identifiability, i.e. if the credible intervals or the corresponding confidence intervals are at all finite for each individual parameter. We performed in-depth identifiability analysis for the parameters of each cell individually for both cell types based on profile posteriors as introduced in Section 3.3 and compare the resulting profiles with the histograms of the obtained posterior samples.

For the fibroblasts, we find that in model $\mathbf{M_1}$ about a third of all cells show local optima in the profile for the half-life β. Still, the parameter is practically identifiable for a confidence level of 95%, since the confidence interval is finite. Thus, this does not present a serious issue for the inference process. Two typical examples of the resulting profiles can be seen in Figure 6.10, one where all parameters are identifiable and one where the half-life shows local optima. Furthermore, we see quite good agreement between the histogram of the posterior samples (marginalization) and the profiles (optimization) in most cases. For the identifiable cell without local optima in the profiles, we see a slightly broader distribution in the histogram than in the profile and for the cell with the local optima, we do not clearly see the local optima in the sampling results. Still the resulting confidence intervals/credible intervals agree very well.

For model $\mathbf{M_2}$, for about two thirds of cells, the parameter for the half-life possesses local optima, but all with finite confidence intervals for a confidence level of 95%. Furthermore, 13 of the 200 cells show a practical non-identifiability for the parameter γ, the steady state level, for which the profile posterior does not drop below the thresholds defined in Section 3.3 for a confidence level of 95% before the parameter reaches zero and thus the boundary of the uniform prior. All in all, we note that the biologically most interesting parameter, the half-life, is identifiable in both models for the fibroblasts, while it is not possible in a small number of cells to give a confidence interval for the fluorescence intensity steady state level.

For the GMPs, we find that all parameters of model $\mathbf{M_1}$ are identifiable for all cells in all three replicates for a confidence level of 95%. For model $\mathbf{M_2}$, the situation is different. Here, about half of the cells in each replicate show a practical non-identifiability for a confidence level of 95% for the parameter γ, the steady state level, for which the profile posterior does not drop below reasonable thresholds before the parameter reaches zero, again the boundary of the uniform prior. Figure 6.11 shows two typical examples, one for a cell where all parameters in model $\mathbf{M_2}$ are identifiable and one where the steady



(a) (b)

Figure 6.10: Typical results for identifiability analysis in fibroblasts. (a) A cell where all parameters are identifiable in model $\mathbf{M_1}$. (b) A cell where the half-life β shows local optima, but still has a finite confidence interval in model $\mathbf{M_1}$. (a-b) Magenta lines are the profile posteriors obtained through optimization, in grey the respective histograms of the posterior samples. The histograms have been scaled such that their maximum value is also one, as for the profiles.

state level's lower confidence bound is practically non-identifiable at a confidence level of 95%. Here the agreement between posterior samples and profile posteriors is even better than for the fibroblasts.

We can call a cell non-identifiable, if the steady state level γ is practically non-identifiable at a confidence level of 95% and identifiable otherwise. As can be seen from Figure 6.12, a rather clear separation between identifiable and non-identifiable cells can be seen when plotting the MAP values of the half-life versus the initial condition $\alpha + \gamma$, with only few outlier cells. Cells with low initial concentrations, i.e. with little protein that can decay and long half-lives tend to have practically non-identifiable steady state levels, which seems reasonable. The practical non-identifiability manifests as non-determinable lower bounds, meaning that for these cells it can not be identified if the protein in the cell decays completely ($\gamma = 0$) or not at a confidence level of 95%. We find similar behavior in all three replicates. Note that the initial condition $\alpha + \gamma$ of the ODE for model $\mathbf{M_2}$ is well-determined from the data. We use the MAP for illustrative purposes, of course the MAP does not have to be representative in the case of non-identifiability.

This shows that identifiability analysis can lead to new insights in settings with a large number of biologically meaningful parameters. In our case, identifiability analysis shows

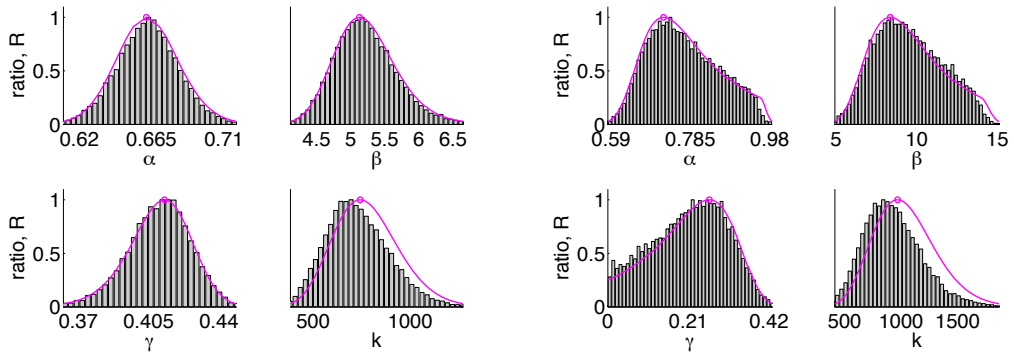


Figure 6.11: Typical results of identifiability analysis in GMPs. (a) A cell in which all parameters are identifiable in model $\mathbf{M_2}$. (b) A cell in which the steady state level γ in model $\mathbf{M_2}$ is practically non-identifiable. (a-b) Magenta lines are the profile posteriors obtained through optimization, in grey the respective histograms of the posterior samples. The histograms have been scaled such that their maximum value is also one, as for the profiles.

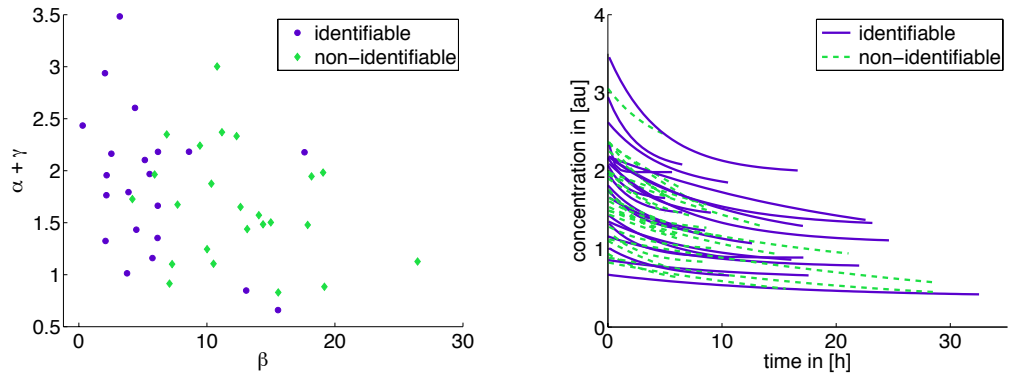


Figure 6.12: Identifiability of the single-cell parameters depends on single-cell properties. (a) Shown are the MAP values based on model $\mathbf{M_2}$ of replicate 1 of each individual cell, color coded according to identifiable (purple dots) or non-identifiable (green diamonds) for the value of half-life β versus the initial condition of the cell $\alpha + \gamma$. A rather clear separation is visible: The steady state level γ can usually not be identified if the half-life is large and the initial condition rather low. (b) The according model trajectories, also color-coded for identifiable (purple, continuous) or non-identifiable (green, dashed).

that the parameters of cells with low initial concentrations and long half-lives are harder to determine from measurement data and mostly lead to practically non-identifiable parameters. Combining posterior samples and profile posteriors as in Figures 6.10 and 6.11 can be very worthwhile, which will also be demonstrated in Chapter 8.

6.5 Model selection based on data for an ensemble of single cells

After thorough analysis of the posterior distributions, we can now come back to the original main question, which is whether the fluorescence intensity in the cells decays to zero or to a non-zero steady state level. In this section, we thus present the results of model selection through thermodynamic integration on all data sets, the fibroblasts and the three replicates of GMPs. For calculating the respective marginal likelihoods with thermodynamic integration, we applied the adaptive Simpson's rule as introduced in Chapter 5. One could ask why we do not resort to simpler model selection methods such as AIC and BIC. However, only the Bayes factor considers the whole distribution of parameters. As the previous section reveals that one parameter is practically non-identifiable in many cells, the whole distribution has to be considered for reliable inference. We thus compute Bayes factors.

For both cell types, we try to do inference as similarly as possible for comparability. We first discuss the fibroblasts, and then the GMPs.

6.5.1 Model selection for GFP decay in fibroblasts

For distinguishing the two models of GFP decay in the fibroblasts, we calculated the Bayes factor with thermodynamic integration with Simpson's adaptive rule. For the MCMC, we used the Adaptive Metropolis algorithm as introduced in Section 4.2. We drew 1000 samples with a thinning factor of 50 (i.e. 50,000 samples) after a burn-in of also 50,000 samples for each temperature and each parameter and verified convergence by the Geweke test (Geweke [1992]). Due to high computational cost (caused by the bulk of individually fitted data), the sampling was run only once.

As already mentioned in Section 5.4.3, some care should be taken when choosing a tolerance for the adaptive Simpson's rule. Thus we first obtain samples from the posterior distributions, which are later needed anyway and determine their expected log

deviance values. Here we find e.g. $\mathbb{E}_{p_{\tau=1}}\{\log p(\mathbf{Y}|\boldsymbol{\theta}^1, \mathbf{M}_1)\} = -1.05 \cdot 10^5$. We then choose $\texttt{TOL} = 1$, since this then corresponds to a relative approximation error of 10^{-5} which performed well on the analytical example.

We find that the adaptive Simpson's rule then uses 29 function evaluations for model $\mathbf{M_1}$ and 25 for model $\mathbf{M_2}$. As Figure 6.13 shows, the adaptive Simpson's rule places more rungs close to zero for both models, where the curvature of the expected log deviance is highest.

We also see that the marginal likelihood is sensitive to choice of prior. Uniform priors lead to problems when sampling from them for the value at $\tau = 0$. For our scenario, model $\mathbf{M_1}$ yields values of $\mathbb{E}_{p_{\tau=0}}\{\log p(\mathbf{Y}|\boldsymbol{\theta}^1, \mathbf{M_1})\} = -5.63 \cdot 10^{34}$ compared to $\mathbb{E}_{p_{\tau=1}}\{\log p(\mathbf{Y}|\boldsymbol{\theta}^1, \mathbf{M}_1)\} = -1.05 \cdot 10^5$, a difference of 29 orders of magnitude. For model $\mathbf{M_1}$ we therefore find that many samples from the prior possess a low log-likelihood value, which significantly increases the variance at $\tau = 0$. The results for model $\mathbf{M_2}$ are numerically not as drastic, but with the same tendency, $\mathbb{E}_{p_{\tau=0}}\{\log p(\mathbf{Y}|\boldsymbol{\theta}^2, \mathbf{M_2})\} = -4.51 \cdot 10^7$ versus $\mathbb{E}_{p_{\tau=1}}\{\log p(\mathbf{Y}|\boldsymbol{\theta}^2, \mathbf{M_2})\} = -9.43 \cdot 10^4$. The reason for the more stable behavior of model $\mathbf{M_2}$ is the offset parameter γ together with the fact that we have multiplicative gamma distributed noise, as both together prevents having to evaluate the probability density function of the univariate gamma distribution for a scale parameter close to zero in Equation (6.2), since this causes very negative values of the log-likelihood. Already for small non-zero values of τ, the situation improves, e.g. $\mathbb{E}_{p_{\tau=1/64}}\{\log p(\mathbf{Y}|\boldsymbol{\theta}^1, \mathbf{M_1})\} = -1.32 \cdot 10^5$, which is in the order of magnitude of the $\tau = 1$ value. Still, the error approximation for the adaptive Simpson's rule is larger than $\texttt{TOL}$ in the leftmost subintervals even after the last refinement due to the high difference in function values.

We find log marginal likelihood values of $\log p(\mathbf{Y}|\mathbf{M}_1) = -2.9 \cdot 10^{32}$ and $\log p(\mathbf{Y}|\mathbf{M}_2) = -3.30 \cdot 10^5$, where we see that the value of the log marginal likelihood for $\mathbf{M_2}$ is numerically negligible compared with the one for model $\mathbf{M_1}$ when forming a Bayes factor. Altogether, we find a Bayes factor of $B_{21} = \exp(2.9 \cdot 10^{32})$ from this. Furthermore, we see that also all expected log deviances except the one for $\tau = 0$, which is 10^{34}, are numerically negligible for the computation of $\log p(\mathbf{Y}|\mathbf{M}_1)$ as only these two values are in such high orders of magnitude.

To ensure that our conclusion for the Bayes factor is not influenced by the numerical problems with the prior, we looked at an approximation. We checked what happens when replacing the value at $\tau = 0$ with the value for the smallest non-zero τ, here

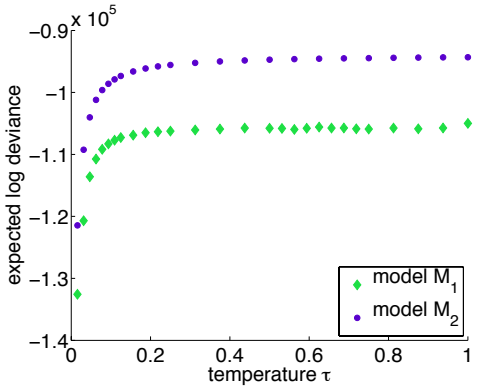


Figure 6.13: Expected log deviances for the fibroblasts. Function value of the expected log deviance over the temperature τ, in green diamonds for model $\mathbf{M_1}$, in purple dots for model $\mathbf{M_2}$. The function values for $\tau = 0$ were truncated for the plots. A clear tendency of higher function values for model $\mathbf{M_2}$ can be seen. Note that function evaluations for both functions are mostly monotonically increasing, as they should be, indicating that the MCMC error is on an acceptable level.

$\tau = 1/64$, for both models. Obviously, this introduces a bias, as $\{\log p(\mathbf{Y}|\mathbf{M}_1)\}_{app} > \log p(\mathbf{Y}|\mathbf{M}_1)$ for the approximated log marginal likelihood $\{\log p(\mathbf{Y}|\mathbf{M}_1)\}_{app}$, as the function within the thermodynamic integral is monotonically increasing. We do not propose to regard this approximation as a correct Bayes factor, but rather as an interpretation aid for the marginal likelihoods.

We thus compute $\hat{B}_{21} = \exp(\{\log p(\mathbf{Y}|\mathbf{M}_2)\}_{app} - \{\log p(\mathbf{Y}|\mathbf{M}_1)\}_{app})$. For this, we find that $\hat{B}_{21} = \exp(1.01 \cdot 10^4)$, based on log marginal likelihoods of $\{\log p(\mathbf{Y}|\mathbf{M}_1)\}_{app} = -9.65 \cdot 10^4$ and $\{\log p(\mathbf{Y}|\mathbf{M}_2)\}_{app} = -8.64 \cdot 10^4$. While this might not be a remedy for all model selection tasks, in this application, the approximate Bayes factor $\hat{B}_{21}$ is numerically more robust, since both approximate marginal likelihoods are now in the same order of magnitude and thus contribute equally to the final result. We conclude that overall the applied model selection method demonstrates a decisive preference for model $\mathbf{M_2}$.

6.5.2 Model selection for PU.1 decay in GMPs

For the decay of PU.1 in the GMPs, we looked at each replicate separately, since the data from these replicates is visibly different as seen in Figure 6.2. Also the analysis of the posterior samples revealed differences between the replicates, see Figure 6.8 and Table 6.4.

We calculated marginal likelihoods with thermodynamic integration with Simpson's adaptive rule. Again we used the Adaptive Metropolis algorithm for the MCMC from Section 4.2. We drew 50,000 samples with a burn-in of 50,000 without thinning for each temperature and each parameter and verified convergence by the Geweke test (Geweke [1992]). Even though the parameter number is here only maximally 192 (which are inferred separately in sets of three or four, as introduced before, for each cell), computational cost is still high and thus the sampling was run only once in each replicate.

As already presented in Section 6.4, we first obtain samples from the posterior distributions and determine their expected log deviance values. Here we find e.g. $\mathbb{E}_{p_{\tau=1}}\{\log p_{\kappa=1}(\mathbf{Y}_{\kappa=1}|\boldsymbol{\theta}^1, \mathbf{M}_1)\} = 1539.7$ for the first replicate. We then choose $\texttt{TOL} = 0.1$, since this then corresponds to a relative approximation error in the order of magnitude of 10^{-5} for all three replicates. This leads to 25 function evaluations being used in all three replicates for both models. This is comparable to the number used for the inference in the fibroblasts.

As for the fibroblasts, we also see that the marginal likelihood is sensitive to choice of prior in all three replicates. Uniform priors lead to problems when sampling from them for the value at $\tau = 0$. For the first replicate, model $\mathbf{M_1}$ yields values of $\mathbb{E}_{p_{\tau=0}}\{\log p_{\kappa=1}(\mathbf{Y}_{\kappa=1}|\boldsymbol{\theta}^1, \mathbf{M_1})\} = -2.01 \cdot 10^{82}$, which have to be compared to $\mathbb{E}_{p_{\tau=1}}\{\log p_{\kappa=1}(\mathbf{Y}_{\kappa=1}|\boldsymbol{\theta}^1, \mathbf{M}_1)\} = 1539.7$, a difference of 85 orders of magnitude, and similarly for the other two replicates.

This is caused for model $\mathbf{M_1}$ by many samples from the prior possessing a low log-likelihood value, which significantly increases the variance at $\tau = 0$. The results for model $\mathbf{M_2}$ are with the same tendency, for example $\mathbb{E}_{p_{\tau=0}}\{\log p_{\kappa=1}(\mathbf{Y}_{\kappa=1}|\boldsymbol{\theta}^2, \mathbf{M_2})\} = -3.28 \cdot 10^6$ and $\mathbb{E}_{p_{\tau=1}}\{\log p_{\kappa=1}(\mathbf{Y}_{\kappa=1}|\boldsymbol{\theta}^2, \mathbf{M_2})\} = 1681.0$ in the first replicate, and similarly for the other two replicates.

For small non-zero values of τ, we find $\mathbb{E}_{p_{\tau=1/64}}\{\log p_{\kappa=1}(\mathbf{Y}_{\kappa=1}|\boldsymbol{\theta}^1, \mathbf{M_1})\} = -1423$ in the first replicate, which is much closer to the $\tau = 1$ value. The situation of the expected log deviances as functions of temperature can also be seen in Figure 6.14.

Since the differences in order of magnitude are even larger than for the fibroblast data, we only show the approximate Bayes factors as introduced in the previous section. The results are shown in Table 6.5.

Also for the GMPs in all replicates, the model with a non-zero steady state level of fluorescence intensity is decisively favored. We conclude that this model, model $\mathbf{M_2}$,

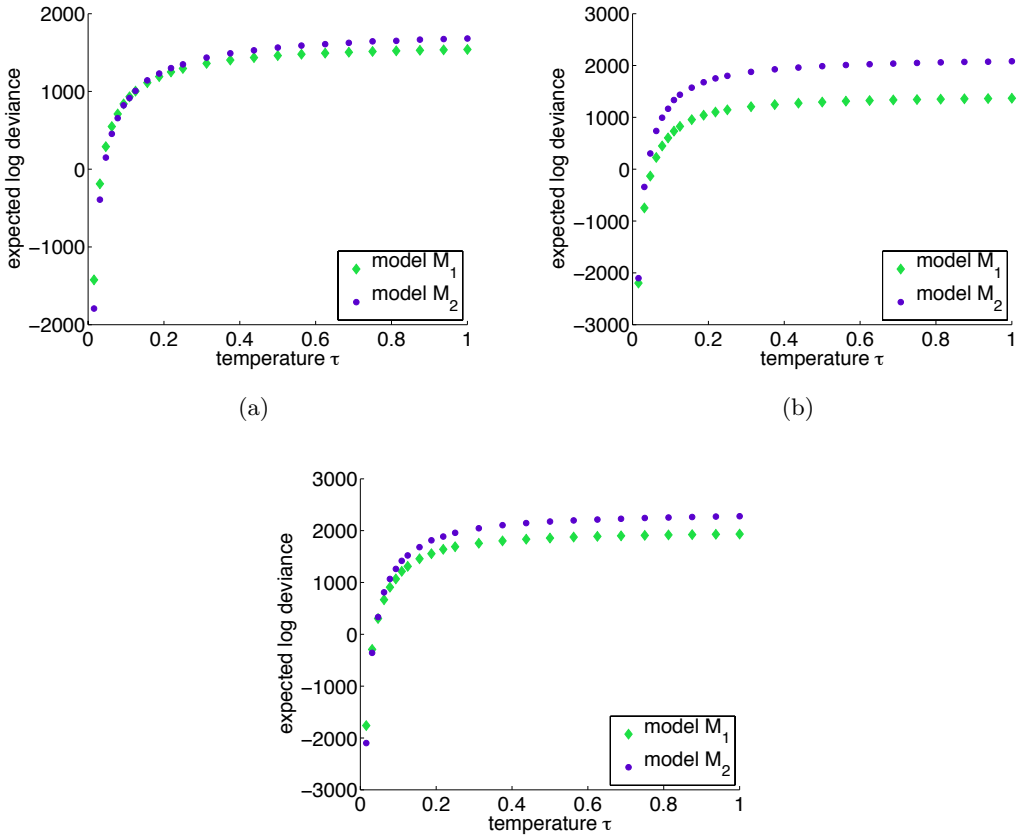


Figure 6.14: Expected log deviances for the GMPs. Expected log deviance as function of the temperature τ, in green diamonds for model $\mathbf{M_1}$, in purple dots for model $\mathbf{M_2}$. The function values for $\tau = 0$ were truncated for the plots. (a) Replicate 1, (b) replicate 2, (c) replicate 3 of the experiment for the decay of PU.1 in GMPs. A clear tendency of higher function values for model $\mathbf{M_2}$ can be seen. Note that the function evaluations for both functions are mostly monotonically increasing. This indicates that the MCMC errors are on an acceptable level.

Replicate κ	$\{\log p_\kappa(\mathbf{Y}_\kappa \vert \mathbf{M}_1)\}_{app}$	$\{\log p_\kappa(\mathbf{Y}_\kappa \vert \mathbf{M}_2)\}_{app}$	**app. BF in favor of** $\mathbf{M_2}$
Replicate 1	1245.03	1326.77	$3.15 \cdot 10^{35}$
Replicate 2	1083.97	1697.99	$4.60 \cdot 10^{266}$
Replicate 3	1584.52	1853.64	$7.48 \cdot 10^{116}$

Table 6.5: Overview over achieved approximate marginal likelihoods and Bayes factors.

is favored for both cell types examined here. This indicates that fluorescence intensity does not decay completely, but only to a certain level. The biological reasons for this should be more thoroughly examined in the future.

6.6 Conclusions

In this chapter, we presented inference results based on single-cell time-lapse microscopy data. We showed what insights can be gained from parameter distributions and identifiability analysis.

Model $\mathbf{M_2}$ is decisively favored for both cell types. This means that fluorescence intensity does not decay completely, but only to a certain level. The biological mechanism behind this mathematical result is yet unknown. The non-zero steady state might be due to a fraction of stable proteins which cannot decay, e.g. because they are bound to some other protein in the cell and thus shielded. Cycloheximide treatment might not work completely, but only to a certain percentage. A third possible biological reason might also be that cells that are stressed can become more autofluorescent. Further experiments are necessary to elucidate the underlying biological mechanism.

The single-cell applications presented in this chapter are a practical example for the influence of weak prior information on the computation of Bayes factors and the numerical problems this might pose, which do however not influence the validity of the methodology. An alternative approach for weak priors might be the approach of Behrens *et al.* [2012], which resorts to importance sampling for low temperature values, discarding poor samples.

We also see that e.g. in the fibroblast data set, the Bayes factor at $B_{21} = \exp(2.9 \cdot 10^{32})$ is much larger than the 100 that should be achieved to classify the selection process as decisive in Jeffrey's scale. This is also due to the large number of almost 10,000 data points available.

For the GMPs, we see that their shared sample-based posterior distribution is broad. While the width of this distribution is a combined effect of parameter uncertainty and biological variability, this is still an indication for a large heterogeneity of the cell population, which was also recently suggested in Hoppe *et al.* [2014]. Since the GMPs are primary cells and not a cell culture like the fibroblasts, it is clear that the GMPs are more heterogeneous than the fibroblasts, which is clearly visible in the pooled credible intervals for the protein half-lives from both cell types. Also Schwarzfischer *et al.* [2014]

showed that the GMPs are more heterogeneous than related cells in the hematopoietic differentiation tree. We see differences between the three different replicates of the experiment at hand, however all three replicates decisively favor model $\mathbf{M_2}$, such that the mechanism behind the decay to a non-zero steady state seems to be the same between all three replicates.

We have fitted all cells individually to answer if their intensity decays to zero or not. However, if the bootstrap of the goodness-of-fit is done for all cells in a replicate at the same time, we arrive at a z-score of < -3 for model $\mathbf{M_2}$. This indicates underfitting on the population level. While the Bayes factor still asserts that model $\mathbf{M_2}$ explains the data better than model $\mathbf{M_1}$, more involved models might be needed to take a more thorough look at the situation. One possible approach might be multi-experiment fitting (Maiwald & Timmer [2008]), where cells share some parameters. Also approaches where some parameters of the cells are considered to come from a shared distribution can be considered (Hasenauer *et al.* [2011, 2014]; Zechner *et al.* [2012]). Our analysis nevertheless is an important fist step for the analysis of single-cell data.

Protein half-lives are important for assessing all models based on protein expression in cells (Eden *et al.* [2011]; Schwanhäusser *et al.* [2011]). Thus experiments like the presented ones, where the decay of proteins can be observed isolated for one protein in single cells are an important basis for more complex models, where several proteins may interact with each other. A rigorous statistical evaluation as presented in this thesis is thus very important as cornerstone for further inference.

7

Model selection for the processing of zirconium in the human body

In this chapter, we present a medium sized model selection problem. The example compares a model with 12 parameters to one with 15 parameters on the basis of 16 data sets. Both models are linear ordinary equation systems representing multi-compartmental models as introduced in Chapter 2. They come from radiation protection where biokinetic ODE models are of crucial importance in dose estimation and further risk analysis for humans exposed to radioactive substances. More concretely, we examine the processing of zirconium in the human body after intake by ingestion. The models in question provide limiting values of detrimental effects and build the basis for applications in internal dosimetry, the prediction for radioactive zirconium retention in various organs as well as retrospective dosimetry.

This chapter is based on and in part identical with the following two publications:

- D. Schmidl*, **S. Hug***, W.B. Li, M.B. Greiter and F.J. Theis (2012). Bayesian model selection validates a biokinetic model for zirconium processing in humans. BMC Systems Biology, 6(1), 95.
- **S. Hug**, D. Schmidl, W.B. Li, M.B. Greiter and F.J. Theis. Uncertainty in Biology: a computational modeling approach, chapter Bayesian model selection methods and their application to biological ODE systems, *in revision*

The content of the first of these papers is also in part contained in another thesis (Schmidl [2012]), as this was a joint first author work. Dr. Schmidl is responsible for the copula-based sampling of the models and the insights that can be gained from the posterior distribution, while the author of this thesis performed the analysis of the thermodynamic integration and the identifiability analysis of the models. Furthermore, the author contributed the analysis of an additional model variant.

7.1 Experimental data and model setup

7.1.1 Experimental data

Mathematically, the models for zirconium processing are multi-compartmental models, corresponding to linear ODE systems as introduced in Section 2.3.2. In a multi-compartmental model for radiation science, all major human organs are represented as separate compartments representing a kinetically homogeneous amount of radionuclides (ICRP [1989]; Jacquez [1985]). Transfer between these compartments is governed by the law of mass balance and described by time-constant transfer rates, which are the parameters that have to be inferred to fit the model to the data. More precisely, the multi-compartmental models considered here are autonomous, linear and closed. However, determining the exact interaction mechanisms is a challenging task. In the present case, there exist two competing models as suggestions for these interaction mechanisms. For the first time, experimental data with measurements in humans from blood plasma and urine were now available, taken in vivo (Greiter *et al.* [2011a]). Applying thermodynamic integration for the computation of Bayes factors, we could establish dominance of one model over the other.

The first model is well established in the community and was put forward by the International Commission on Radiological Protection (ICRP) (ICRP [1989]). The transfer rates for this model were mostly derived from animal data and yield extensive prior information for our inference. The Helmholtz Zentrum München (HMGU) recently published another, physiologically more plausible biokinetic model (Greiter *et al.* [2011a]). It is the first model based on measurement data in humans, taken in 16 investigations from 12 healthy human subjects, meaning that four subjects were measured twice completely independently. In vivo measurements were taken from blood plasma and urine of up to 100 days after ingestion by application of the double tracer technique. More details on the measurement process as well as a global statistical uncertainty and sen-

sitivity analysis of this HMGU model can be found in the respective publications (Li *et al.* [2011a,b]).

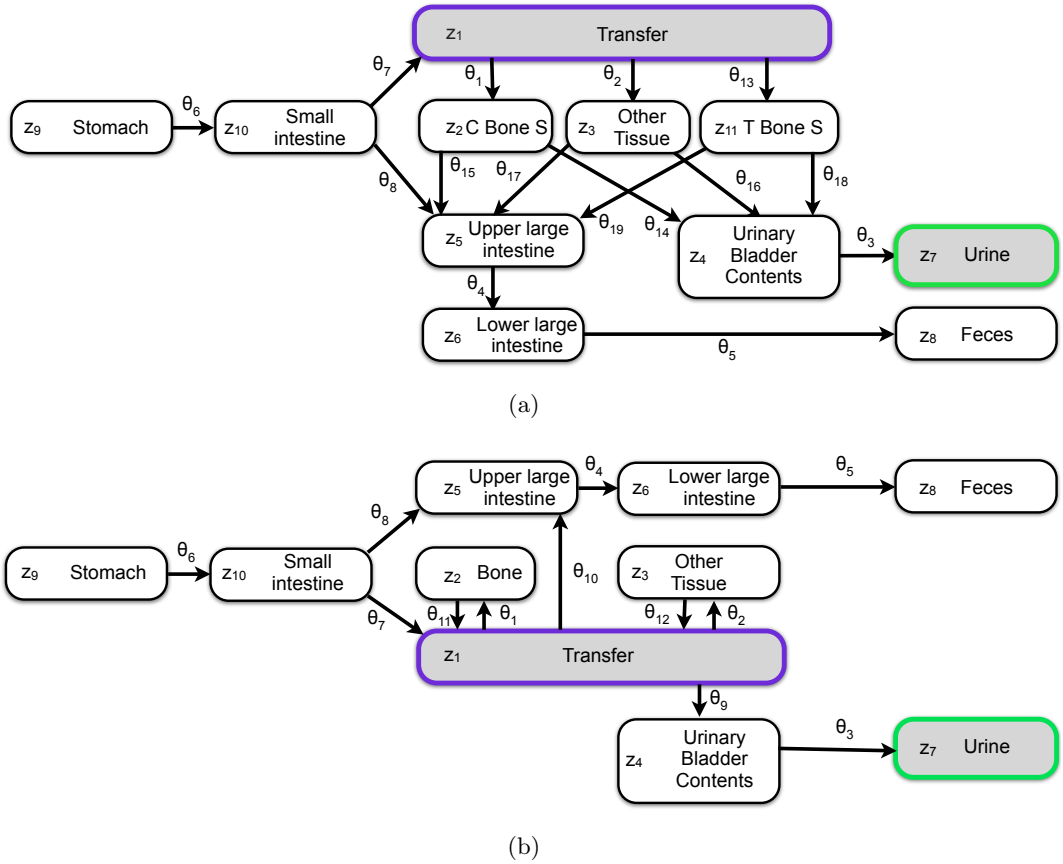


Figure 7.1: Models for zirconium processing. (a) ICRP model. This model contains eleven compartments $z_1,\ldots,z_{11}$ and 15 time-constant transfer rates $\theta_1,\ldots,\theta_8,\theta_{13},\ldots,\theta_{19}$. (b) HMGU model. This model contains ten compartments $z_1,\ldots,z_{10}$ and twelve transfer rates $\theta_1,\ldots,\theta_{12}$. (a-b) In both models zirconium enters the body in the stomach compartment z_9 and is processed through the system until it reaches either one of the two final compartments urine, z_7, or feces, z_8. The gray-shaded compartments z_1 and z_7 are corresponding to those where measurements are taken.

The first of the two compartmental models under examination here was recommended by the ICRP in ICRP [1975, 1989, 1993] (Figure 7.1(a)). The model contains eleven compartments, which are linked through 15 transfer rates. Since zirconium is ingested, it enters the body through the stomach compartment z_9 and is processed until it reaches one of the two final compartments urine, z_7, or feces, z_8. The HMGU model (Greiter *et al.* [2011a]) differs from this model: it contains only ten compartments and twelve

transfer rates, since the physiologically questionable distinction between the two bone compartments of the ICRP model, trabecular bone surface and cortical bone surface, was abolished in this model (Figure 7.1(b)). Furthermore, most mass transfers are now mediated by the transfer compartment representing the blood plasma instead of direct transfers, which is physiologically more plausible. To better represent that some rates are shared between the models, we use the following notation. We introduce a general parameter vector $\boldsymbol{\theta}$. Both models share eight transfer rates, which are thus denoted $\theta_1, \ldots, \theta_8$. The additional rates present in only one of the models then have a unique index $\theta_9, \ldots, \theta_{19}$, beginning with the HMGU model specific rates, since the HMGU model is smaller. Thus we assign the model designation $\mathbf{M_1}$ to the HMGU model and $\mathbf{M_2}$ to the ICRP models. This also means that the parameters for the model $\mathbf{M_1}$ are $\boldsymbol{\theta}^1 = (\theta_1, \ldots, \theta_{12})$ and for model $\mathbf{M_2}$ they are $\boldsymbol{\theta}^2 = (\theta_1, \ldots, \theta_8, \theta_{13}, \ldots, \theta_{19})$.

7.1.2 ODE model and model likelihood

Both multi-compartmental models correspond to systems of coupled linear first-order ordinary differential equations. For the time-dependent concentrations $\mathbf{z}(t)$, we have in model $\mathbf{M}_i$ as introduced in Section 2.3.2

$$\frac{\mathrm{d}\mathbf{z}(t)}{\mathrm{d}t} = \mathbf{a}^i(\boldsymbol{\theta}^i)\mathbf{z}. \tag{7.1}$$

The matrix $\mathbf{a}^i(\boldsymbol{\theta}^i)$ is specific for model $\mathbf{M}_i$ and contains $\boldsymbol{\theta}^i$ as entries. Initial concentrations are needed for a unique solution of the ODEs, thus we choose $z_9(0) = 100\%$ and $z_{d\neq 9}(0) = 0\%$, since we assume that all zirconium is in the stomach compartment at the beginning of the investigation. The detailed ODEs and thus the matrices $\mathbf{a}^i(\boldsymbol{\theta}^i), i = 1, 2$ can straightforwardly be derived and can be found in Appendix A.

Zirconium was measured in plasma and urine through the double tracer technique in 16 investigations (Greiter *et al.* [2011a,b]). The raw data tracer concentrations were then normalized to the respective investigation-specific tracer amount to yield 100% at $t = 0$ in the stomach compartment z_9. For the development of the model, the transfer compartment was taken to be identical with blood plasma, the measured concentrations were then expressed as % per kg plasma. Absolute concentrations were obtained by scaling with the total amount of plasma in the body (Alberts *et al.* [2002]). The measurements in urine correspond to an excretion rate in % per day.

This corresponds to two observables of the system, which are the same for both models.

The first observable is the concentration in the transfer compartment,

$$\mathcal{Y}_1(t) = h_1(\mathbf{z}(t)) = z_1(t). \tag{7.2}$$

The second observable is the excretion rate in the urine compartment,

$$\mathcal{Y}_2(t) = h_2(\mathbf{z}(t)) = \frac{\mathrm{d}z_7(t)}{\mathrm{d}t}. \tag{7.3}$$

The data for investigation r is given by the measurements in plasma and urine

$$\mathbf{Y}_r = (y_1^{r,1}, y_1^{r,2}, \ldots, y_1^{r,N_r^b}, \dot{y}_7^{r,1}, \dot{y}_7^{r,2}, \ldots, \dot{y}_7^{r,N_r^u}) \tag{7.4}$$

While y_1^{r,n_b} indicates the n_b-th measurement in plasma in investigation r, i.e. for observable $\mathcal{Y}_1(t)$, $\dot{y}_7^{r,n_u}$ designates the n_u-th measurement of the excretion rate in the urine compartment and thus for observable $\mathcal{Y}_2(t)$. There are $n_b = 1, \ldots, N_r^b$ measurements in plasma and $n_u = 1, \ldots, N_r^u$ measurements in urine for investigation r.

For each of the R investigations $r = 1, \ldots, R = 16$, we find the likelihood by assuming Gaussian noise on both observables for any of the two models $\mathbf{M_1}$ or $\mathbf{M_2}$ and the corresponding parameter vector $\boldsymbol{\theta}^i$, where the model index $i \in \{1, 2\}$.

The likelihood is then given for each investigation r and for model $\mathbf{M}_i$ by

$$p(\mathbf{Y}_r|\boldsymbol{\theta}^i, \mathbf{M}_i) = \underbrace{\prod_{\alpha=1}^{N_r^b} \phi\left(y_1^{r,\alpha}; \mathcal{Y}_1(t_\alpha), \sigma_r^b\right)}_{p^{(b)}(\mathbf{Y}_r|\boldsymbol{\theta}^i, \mathbf{M}_i)} \underbrace{\prod_{\beta=1}^{N_r^u} \phi\left(\dot{y}_7^{r,\beta}; \mathcal{Y}_2(t_\beta), \sigma_r^u\right)}_{p^{(u)}(\mathbf{Y}_r|\boldsymbol{\theta}^i, \mathbf{M}_i)}. \tag{7.5}$$

Here, $\phi(x; \mu, \sigma)$ is the probability density function of the univariate normal distribution evaluated at x with mean μ and standard deviation σ as introduced in Section 2.2.1. Also, $\mathcal{Y}_1(t_\alpha)$ is the observable for the transfer compartment z_1 at time point t_α, corresponding to the measurement at $y_1^{r,\alpha}$. Accordingly $\mathcal{Y}_2(t_\alpha)$ is the observable for the urinary excretion rate at time point t_β, corresponding to the measurement $\dot{y}_7^{r,\beta}$. The standard deviations of the normal distributions for plasma (observable $\mathcal{Y}_1$), σ_r^b, and for urine (observable $\mathcal{Y}_2$), σ_r^u, are fitted for each investigation separately by applying the simulated annealing algorithm (Kirkpatrick *et al.* [1983]) before starting the MCMC sampling process. This error model corresponds to the combined strength of all deviations from the "true" ODE solution, which include (possibly amongst others) measurement error as well as natural internal fluctuations not considered by an ODE approach. With this assumptions, both models are able to fit the data in principle, justifying our ODE approach with additive normally distributed noise, see also Figure 7.3.

It is however still important to account for biological variability between the individual investigations, for which we account by fitting different σ_r^b and σ_r^u for each investigation r and thus get investigation-specific likelihoods. This leads to individual credible intervals for each parameter in each investigation in the MCMC sampling procedure later on.

In contrast to this individual treatment, it also makes sense to consider the complete (i.e. concatenated) data. Then the likelihood is given by $p_{ALL}(\mathbf{Y}|\boldsymbol{\theta}^i, \mathbf{M}_i) = \prod_{r=1}^{16} p(\mathbf{Y}_r|\boldsymbol{\theta}^i, \mathbf{M}_i)$, with $\mathbf{Y} = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_{16}\}$ and fitting investigation independent σ^b and σ^u.

For the calculation of the likelihood, the ODE has to be solved depending on $\boldsymbol{\theta}^i$. As already mentioned in Section 2.3.2, in our special case the ODEs can be solved easily with the matrix exponential, see also Appendix A.

This now enables us to compute a Bayes factor for each investigation r

$$B_{12}^r = \frac{p(\mathbf{Y}_r|\mathbf{M}_1)}{p(\mathbf{Y}_r|\mathbf{M}_2)}. \tag{7.6}$$

as well as an overall Bayes factor

$$B_{12} = \frac{p_{ALL}(\mathbf{Y}|\mathbf{M}_1)}{p_{ALL}(\mathbf{Y}|\mathbf{M}_2)}. \tag{7.7}$$

7.2 Prior information and algorithmic setup

7.2.1 Prior information

Models for zirconium processing have been used for a few decades already, and quite a large number of animal studies has been held. From these, comprehensive prior information for both models can be curated. The priors for each single transfer rate are given as triangular, normal or lognormal distributions with known hyperparameters. Of the eight transfers present in both models, only θ_8 has different prior distributions in the ICRP and HMGU model. Each univariate prior distribution was truncated at zero. The prior information is naturally the same for each investigation, since it represents information from a large number of preceding examinations and is not specific to the present investigations.

7.2.2 Summary of prior distributions

Table 7.1 provides an overview of the prior distributions and distribution parameters used for parameter inference in the HMGU and ICRP model. The prior distributions for the ICRP model are directly derived from the recommendations of the ICRP and thus well-established over the years. The priors for the HMGU model are also in part derived from these recommendations, plus information gained from additional experiments based on injected zirconium doses (Li *et al.* [2011a]).

A keen reader might notice the difference in orders of magnitude for the σ's of the HMGU model and the ICRP model. However, these stem mostly from the fact that the HMGU model has lognormally distributed priors where the ICRP model has normally distributed ones, both models assume a coverage factor of 3 to represent 99.7% confidence intervals (ISO [1995]) for normal and lognormal distributions.

7.2.3 Algorithmic setup

In order to be able to do model selection via thermodynamic integration, we need to be able to sample from the model, investigation and temperature specific distribution. For this we use the copula-based independence/random walk Metropolis-Hastings approach (CIMH) (Schmidl *et al.* [2013a]) as introduced in Section 4.5. The necessary fitting of the copula distribution was done on the basis of preruns containing 1,000,000 unthinned samples each. These were generated for each investigation and model separately with a standard Metropolis-Hastings with a normal distribution proposal. The required back-and-forth conversion of the prerun samples and proposals was done with the according prior distributions of the models at hand.

We use simulated annealing to find the maximum a posteriori estimate and use this as starting point for the sampling, enabling us to skip the burn-in phase. For this application, we chose to apply thinning by the autocorrelation based effective sample size (ESS) (Neal [1993]). Though generally not necessary, we used this as an additional quality insurance. Our sampling algorithm CIMH is able to provide a high ESS at simultaneously high acceptance rates. From all required distributions we generated 30,000 proposals.

7. MODEL SELECTION FOR THE PROCESSING OF ZIRCONIUM IN THE HUMAN BODY

HMGU model

Par.	Compartments	Median (d^{-1})	99.7% CI	distr.	μ or a	σ or b	c
θ_1	TC → Bone	0.10	[0.013, 0.8]	$\mathcal{LN}(\mu,\sigma)$	-2.3026	0.6931	
θ_2	TC → Other	1.35	[0.17, 10.8]	$\mathcal{LN}(\mu,\sigma)$	0.3001	0.6931	
θ_3	UBC → Urine	12.0		$\mathcal{T}(a,b,c)$	6.0	8.0	24.0
θ_4	UpLI → LoLI	1.8		$\mathcal{T}(a,b,c)$	0.9	1.2	3.6
θ_5	LoLI → Feces	1.0		$\mathcal{T}(a,b,c)$	0.3	1.0	1.7
θ_6	Stomach → SI	24.0		$\mathcal{T}(a,b,c)$	12.0	16.0	48.0
θ_7	SI → TC	0.03	$[1.11 \cdot 10^{-3}, 0.81]$	$\mathcal{LN}(\mu,\sigma)$	-3.5066	1.0986	
θ_8	SI → UpLI	17.21	[0.64, 464.67]	$\mathcal{LN}(\mu,\sigma)$	2.8455	1.0986	
θ_9	TC → UBC	0.031	[0.0011, 0.8370]	$\mathcal{LN}(\mu,\sigma)$	-3.4738	1.0986	
θ_{10}	TC → UpLI	0.0062	[0.0002, 0.1674]	$\mathcal{LN}(\mu,\sigma)$	-5.0832	1.0986	
θ_{11}	Bone → TC	$6.93 \cdot 10^{-5}$	$[8.66 \cdot 10^{-6}, 5.55 \cdot 10^{-4}]$	$\mathcal{LN}(\mu,\sigma)$	-9.5769	0.6931	
θ_{12}	Other → TC	0.53	[0.066, 4.24]	$\mathcal{LN}(\mu,\sigma)$	-0.6349	0.6931	

ICRP model

Par.	Compartments	Median (d^{-1})	99.7% CI	distr.	μ or a	σ or b	c
θ_1	TC → CBS	0.69	[0.086, 5.52]	$\mathcal{LN}(\mu,\sigma)$	-0.3711	0.6931	
θ_2	TC → Other	1.39	[0.174, 11.12]	$\mathcal{LN}(\mu,\sigma)$	0.3293	0.6931	
θ_3	UBC → Urine	12		$\mathcal{T}(a,b,c)$	6	8	24
θ_4	UpLI → LoLI	1.8		$\mathcal{T}(a,b,c)$	0.9	1.2	3.6
θ_5	LoLI → Feces	1		$\mathcal{T}(a,b,c)$	0.3	1	1.7
θ_6	Stomach → SI	24		$\mathcal{T}(a,b,c)$	12	16	48
θ_7	SI → TC	0.06	[0.0075, 0.48]	$\mathcal{LN}(\mu,\sigma)$	-2.8134	0.6931	
θ_8	SI → UpLI	6		$\mathcal{T}(a,b,c)$	3	4	12
θ_{13}	TC → TBS	0.69	[0.086, 5.52]	$\mathcal{LN}(\mu,\sigma)$	-0.3711	0.6931	
θ_{14}	CBS → UBC	$5.78 \cdot 10^{-5}$	$[5.78 \cdot 10^{-6}, 1.1 \cdot 10^{-4}]$	$\mathcal{N}(\mu,\sigma)$	$5.78 \cdot 10^{-5}$	$1.73 \cdot 10^{-5}$	
θ_{15}	CBS → UpLI	$1.16 \cdot 10^{-5}$	$[1.16 \cdot 10^{-6}, 2.2 \cdot 10^{-5}]$	$\mathcal{N}(\mu,\sigma)$	$1.16 \cdot 10^{-5}$	$3.48 \cdot 10^{-6}$	
θ_{16}	Other → UBC	0.083	[0.0083, 0.158]	$\mathcal{N}(\mu,\sigma)$	0.083	0.025	
θ_{17}	Other → UpLI	0.0165	[0.00165, 0.0314]	$\mathcal{N}(\mu,\sigma)$	0.0165	0.00495	
θ_{18}	TBS → UBC	$5.78 \cdot 10^{-5}$	$[5.78 \cdot 10^{-6}, 1.1 \cdot 10^{-4}]$	$\mathcal{N}(\mu,\sigma)$	$5.78 \cdot 10^{-5}$	$1.73 \cdot 10^{-5}$	
θ_{19}	TBS → UpLI	$1.16 \cdot 10^{-5}$	$[1.16 \cdot 10^{-6}, 2.2 \cdot 10^{-5}]$	$\mathcal{N}(\mu,\sigma)$	$1.16 \cdot 10^{-5}$	$3.48 \cdot 10^{-6}$	

Table 7.1: Overview of priors. The tables are based on Li *et al.* [2011a], where the confidence intervals and the medians are given. From these the derivation of the parameters of the normal and lognormal distributions are straightforward. The abbreviation are: $\mathcal{N}(\mu,\sigma)$: normal distribution with mean μ and standard deviation σ, $\mathcal{LN}(\mu,\sigma)$: lognormal distribution with location parameter μ and scale parameter σ, $\mathcal{T}(a,b,c)$: triangular distribution with lower limit a, upper limit c, and mode b. Par. = parameter, CI = confidence interval, distr. = distribution type, TC= Transfer compartment; CBS = Cortical Bone Surface; Other = Other Tissues; UBC = Urine Bladder Contents; UpLi = Upper Large Intestine; LoLI = Lower Large Intestine; SI = Small Intestine; TBS= Trabecular Bone Surface.

7.3 Inference results

7.3.1 Parameters are investigation specific

Since the experimental data as basis for the model selection comes from 16 investigations, one can ask if the models should be compared based on the complete data, yielding one overall Bayes factor, or on each dataset separately, yielding 16 Bayes factors, which cannot be directly compared. When taking a closer look at the data (Figure 7.2), we see that all investigations exhibit a pulse-like time course in the plasma measurements, while the excretion rates in urine point more to an exponential decay behavior. Despite these shared characteristics, the actual zirconium tracer concentrations showed up to a 50-fold difference between maximum plasma concentrations, i.e. for investigation $r = 10$ (1.616%) and $r = 6$ (0.033%).

This already suggest that the investigations should be treated separately, since the differences in the concentrations propagate to differences in the transfer rates. To verify this, we did a pairwise comparison of the posterior samples marginal (corresponding to the temperature $\tau = 1$) by the Kolmogorov-Smirnov test. Since this test is univariate, we picked parameter θ_7 in the ICRP model as example, as it directly affects the observed concentrations in plasma (Li *et al.* [2011b]). Except for one pair, all obtained p-values were $< 6 \cdot 10^{-8}$. We take this as a strong indication that all investigations should be treated separately.

However, for many applications of the models, reference values for an average subject are needed. This is why we also included the Bayes factor for the complete data in our analysis. The differences between the overall Bayes factor and the investigation specific ones can also be the basis for the study of influence factors like gender or weight.

7.3.2 Analysis of sampling results

The obtained posterior samples yield credible intervals for the parameters at hand as well as a maximum a posteriori estimate based on the complete data, which can be used if single parameter values for an average subject are required. If we now propagate these posterior samples to the ODE solution, we see in Figure 7.3 that both models are in principle able to fit the measurement data, justifying our approach. While no rigorous model selection is possible merely from these fits, especially the plasma data already hints at a better suitability of the HMGU model. In the urine data the difference

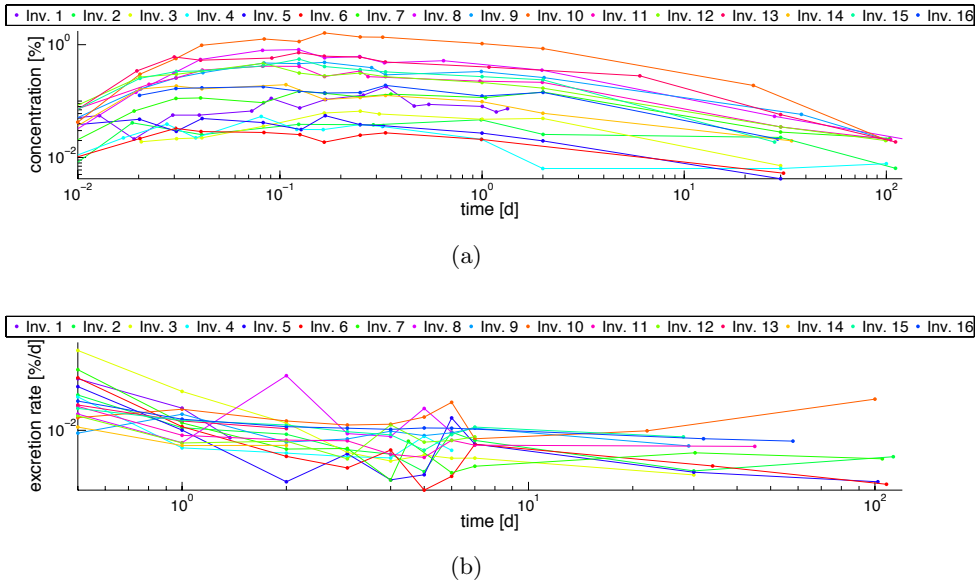


Figure 7.2: The experimental data. (a) Plasma concentrations and (b) excretion rate in urine for all 16 investigations on log-log-scale.

between the two models is not as pronounced. We want to point out that the credible intervals in Figure 7.3 represent only the uncertainty coming from the parameters, in contrast to the measurement uncertainties accounted for by the noise parameters σ_r^b and σ_r^u, which are not shown.

7.3.3 Identifiability analysis

We performed an identifiability analysis for the HMGU and ICRP model based on the according posterior distributions for each of the 16 investigations. The identifiability analysis was done as introduced in Section 3.3 using profile posteriors. For all investigations every model was clearly identifiable at a 95% confidence level as can be seen in Figure 7.4 and Figure 7.5. Of course this is only true when using the posterior distribution and not the data likelihood in a profile likelihood, since then alternative routes through the system could not be distinguished. The identifiability of the models induces a valid estimate of the maximum a posteriori estimate as well as the credible regions for the parameter estimates.

It can also be noted that some parameters have very similar profiles in all individuals, such as θ_4 and θ_5 in both models. Especially for these two parameters, this is due to

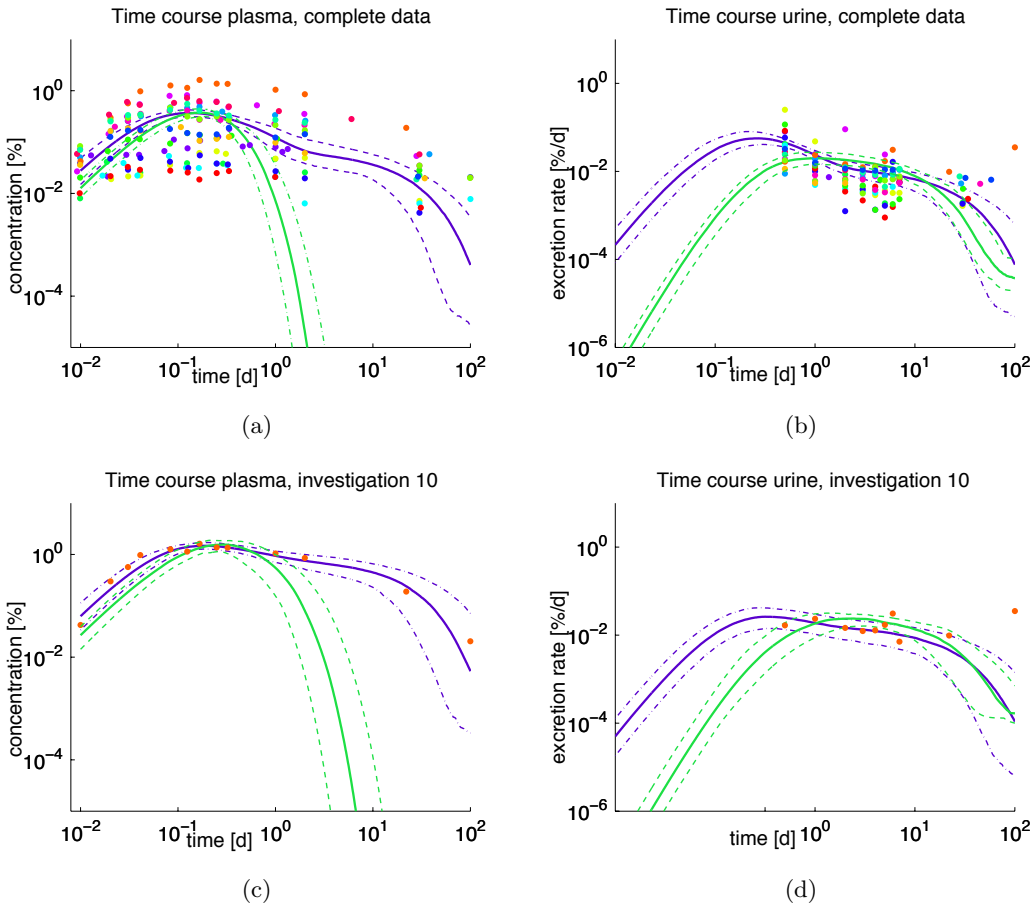


Figure 7.3: Posterior time courses. Sample median (solid line) and 90% point-wise credible interval (CI, borders as dashed lines) drawn from the time courses based on the $\tau = 1$ HMGU (purple) and ICRP (green) MCMC samples for (a) the complete plasma data, (b) the excretion rate in urine over time of the complete data, (c) as a single example the plasma data of investigation 10, and (d) the corresponding urinary excretion rate over time of investigation 10, all plotted on a log-log scale. The plasma plots were truncated at $1 \cdot 10^{-5}[\%]$ and urine plots at $1 \cdot 10^{-6}[\%/d]$. Note that the median and CI represent only the uncertainty in the parameters, in contrast to measurement uncertainty (not shown). Colored markers are the raw experimental data points. At each time point the median and the 90% credible interval were computed point-wisely over all MCMC based solutions.

the fact that they connect unobserved compartments and are only identifiable due to the stringent prior information and not due to the data likelihood.

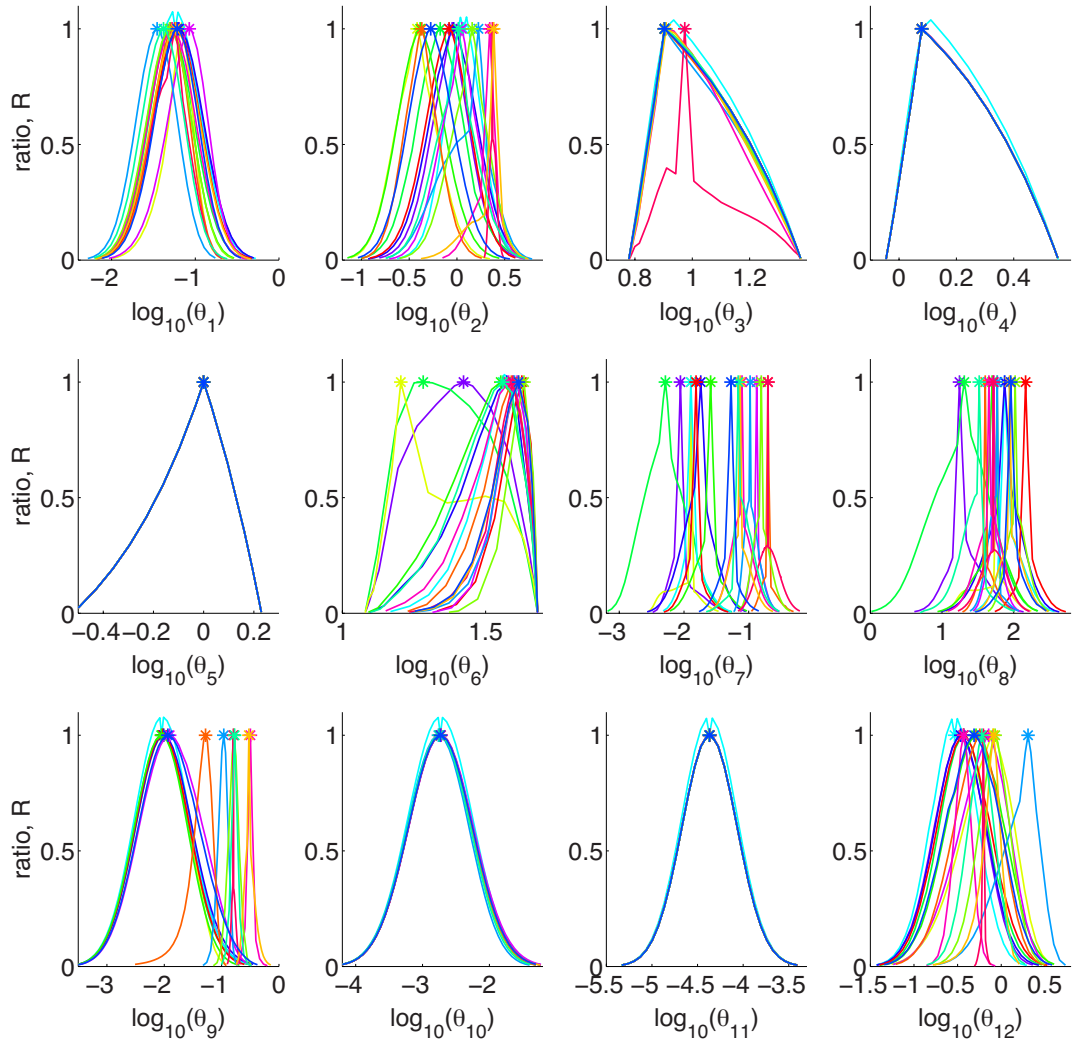


Figure 7.4: Identifiability for the HMGU model. Profile posteriors are color-coded for investigation. All parameters in all investigations are identifiable at a 95% confidence level. Where there is only one line visible for a parameter, all profiles are coinciding.

7.3.4 Bayesian model selection for the two proposed models

For the actual model selection, we now compared the HMGU and ICRP models based on both the complete data as well as the individual investigations, yielding 17 Bayes factors. This task is medium sized with respect to the parameter dimensions. Fur-

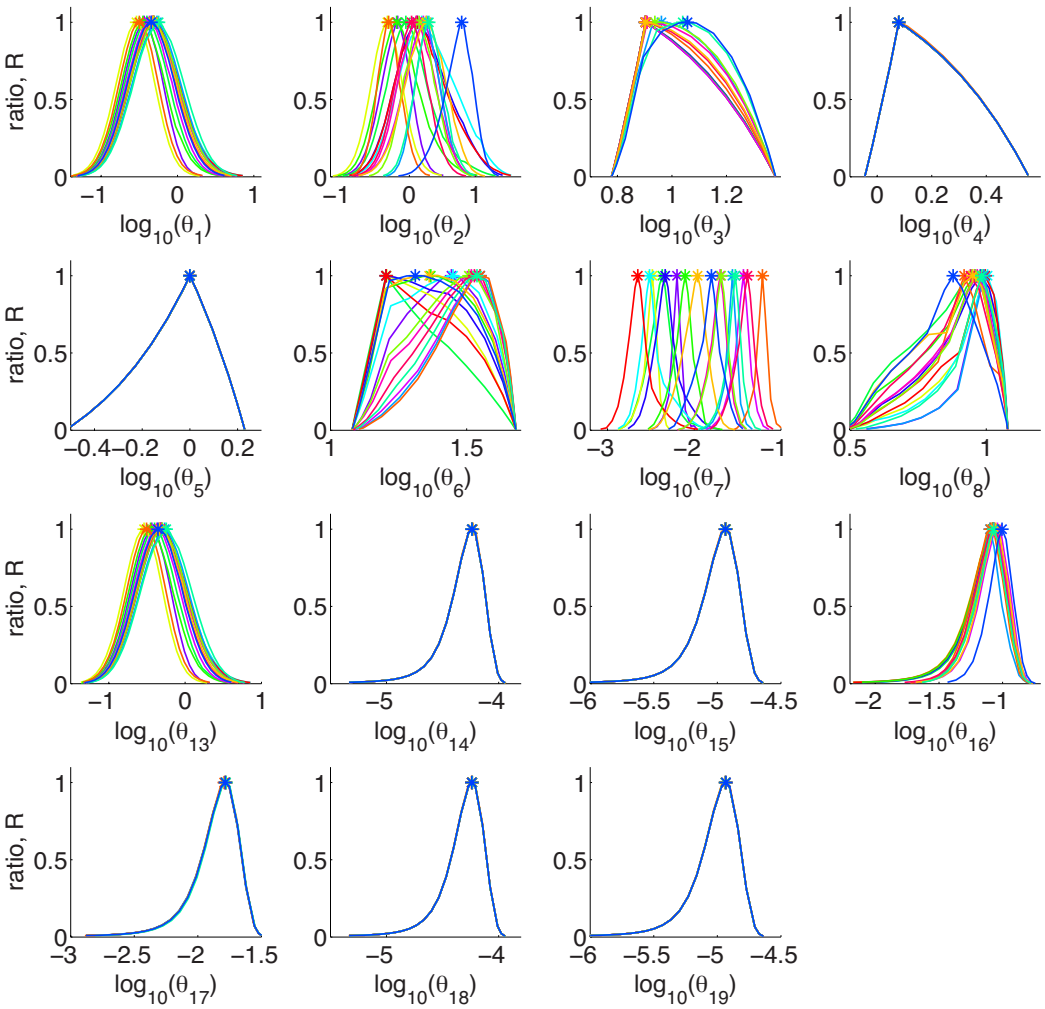


Figure 7.5: Identifiability for the ICRP model. Profile posteriors are color-coded for investigation. All parameters in all investigations are identifiable at a 95% confidence level. Where there is only one line visible for a parameter, all profiles are coinciding.

thermore, we are encouraged to consider all 17 data scenarios separately. For this reason, we abstain from using the adaptive Simpson's rule and apply the standard power law trapezoidal rule. This leads to all individual marginal likelihoods being computed on the same grid and thus better comparability. We use a schedule of $\tau_k = (k/29)^5, k = 0, 1, \ldots, 29$ with $K = 30$ rungs for computing the marginal likelihoods in all 17 scenarios. We found that all Bayes factors favor the HMGU model, 14 out of 17 even decisively (Table 7.2).

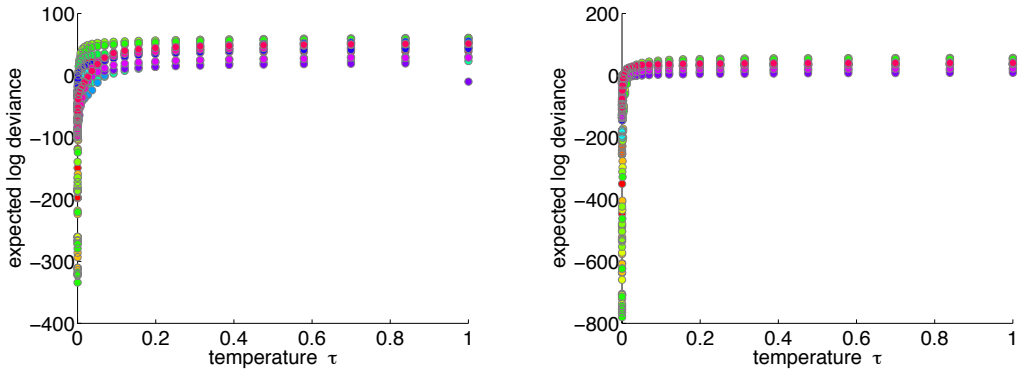
Table 7.2: Bayes factors for the HMGU versus the ICRP model (B_{12}^r) for the individual investigations as well as for the complete data (ALL,B_{12}) and the according Bayes factors for the blood plasma ($B_{12}^{b,r}$ and B_{12}^b) and urine ($B_{12}^{u,r}$ and B_{12}^u) data. Green color indicates a Bayes factor in favor of the HMGU model and red color a Bayes factor in favor of the ICRP model. The HMGU model is favored *substantially*, when the Bayes factor is > 3 and *decisively*, when it is > 100.

Inv. r	B_{12}^r & B_{12}	$B_{12}^{b,r}$ & B_{12}^b	$B_{12}^{u,r}$ & B_{12}^u
1	$7.17 \cdot 10^1$	$7.12 \cdot 10^1$	1.05
2	$1.15 \cdot 10^2$	$2.93 \cdot 10^2$	$3.94 \cdot 10^3$
3	$5.95 \cdot 10^4$	$5.23 \cdot 10^4$	1.34
4	$1.07 \cdot 10^3$	$2.64 \cdot 10^3$	$3.47 \cdot 10^1$
5	$2.19 \cdot 10^2$	$4.73 \cdot 10^2$	$1.34 \cdot 10^2$
6	$4.64 \cdot 10^3$	$3.93 \cdot 10^3$	$2.38 \cdot 10^3$
7	$2.18 \cdot 10^2$	$2.30 \cdot 10^2$	$1.34 \cdot 10^3$
8	$3.75 \cdot 10^1$	$1.28 \cdot 10^2$	0.22
9	$4.62 \cdot 10^2$	$2.32 \cdot 10^2$	0.18
10	$8.62 \cdot 10^2$	$1.16 \cdot 10^2$	0.20
11	$1.17 \cdot 10^5$	$1.81 \cdot 10^1$	$2.94 \cdot 10^3$
12	$1.78 \cdot 10^2$	5.48	$1.14 \cdot 10^1$
13	$7.19 \cdot 10^2$	$1.41 \cdot 10^1$	4.41
14	$3.58 \cdot 10^1$	7.43	9.77
15	$6.29 \cdot 10^3$	$2.17 \cdot 10^1$	$1.60 \cdot 10^2$
16	$6.22 \cdot 10^2$	$1.34 \cdot 10^1$	$1.20 \cdot 10^4$
ALL	$1.20 \cdot 10^{11}$	$3.43 \cdot 10^4$	$4.73 \cdot 10^7$

As the time courses already indicated that the HMGU model might be more suitable since it fits the plasma data better, we computed additional Bayes factors based on either only the plasma or urine data. This corresponds to considering either only "the plasma likelihood" $p^{(b)}(\mathbf{Y}_r|\boldsymbol{\theta}^i, \mathbf{M}_i)$ or "the urine likelihood" $p^{(u)}(\mathbf{Y}_r|\boldsymbol{\theta}^i, \mathbf{M}_i)$ from Equation (7.4), where $i = 1, 2$ and $r = 1, \ldots, 16$ and accordingly for the complete data. The Bayes factors support our theory that the plasma is fitted better by the HMGU model: all 17 Bayes factors based on the plasma data favored the HMGU model, in ten cases again decisively (Table 7.2, 3rd column). For the urine data, the situation is slightly more ambivalent, as here three investigations favor the ICRP model (Table

7.2, 4th column), but not decisively, while still eight Bayes factors favor the HMGU model decisively. All in all, we assert that all decisive Bayes factors are in favor of the HMGU model, meaning that the ICRP model was decisively rejected in the majority of cases. Thus we conclude that the HMGU model is superior to the ICRP model for representing zirconium processing in the human body, both on an individual level as well as for an average subject represented by the complete data.

We can now take again a closer look at the expected log deviances from the two models. It can clearly be seen in Figure 7.6 that the expected log deviance has a similar functional shape for all investigations in both models. However, function values differ due to differences in the quality of fit to the data. One can also note that the ICRP model reaches lower values of the expected log deviance at $\tau = 0$ than the HMGU model, which can be taken as a hint that the prior information for the HMGU model is of higher accuracy. All in all, the sampled expected log deviances show largely monotonic behavior, as can be expected from theory. This indicates that the sampling process is well-behaved.



(a) (b)

Figure 7.6: Model-specific expected log deviances for the zirconium models. (a) Expected log deviances depending on the temperature τ for all investigations for the HMGU model. (b) Expected log deviances for all investigations for the ICRP model. Individual investigations are color-coded. It can be seen that all show similar behavior with respect to function shape, if not with respect to function value.

A close look can be taken at the expected log deviances stemming from the combined data in Figure 7.7. A clear tendency can be seen that the HMGU model has higher expected log deviance almost everywhere in the interval. Accordingly the integral value for the log marginal likelihood of the HMGU model is 271.0 and for the ICRP model

we get a log marginal likelihood of 245.5. This is also the reason for the resulting large Bayes factor of $1.20 \cdot 10^{11} = \exp(25.5)$.

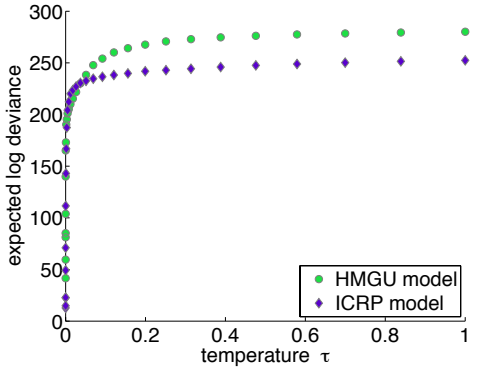


Figure 7.7: Expected log deviances for the zirconium models. Expected log deviances depending on the temperature τ for the combined data. Here the green circles represent the HMGU model, and purple diamonds the ICRP model.

If we look at the same plot for the 16 individual investigations (Figure 7.8), the difference can not be seen as easily. This is because the plots are on log scale, so that for a Bayes factor of 100 a difference in log marginal likelihoods of only $\log(100) = 4.6$ is required. Thus the integrals from the function evaluations in Figure 7.8 do not need to be very different. A clear visual example is nevertheless given by Investigation 11, where a strong tendency for higher values for the HMGU model can be seen, which results in a Bayes factor of $1.17 \cdot 10^5$.

7.3.5 Dismissing a more complex model variant

Inspired by the clear superiority of the HMGU model over the ICRP model with respect to the provided human data, one could think about further improving the HMGU model. This could correspond to even better physiological plausibility of the model. One such hypothesis would be to introduce a new compartment to the model, representing the urinary path. The resulting model can be seen in Figure 7.9 and will be designated $\mathbf{M}_3$. It has eleven compartments and 15 parameters like the ICRP model, but with a different model topology. The initial concentration in the new compartment z_{12} is assumed to be zero the same as for all compartments except the stomach compartment in the original HMGU model. From literature, prior distributions for the three new parameters can be derived, see Table 7.3.

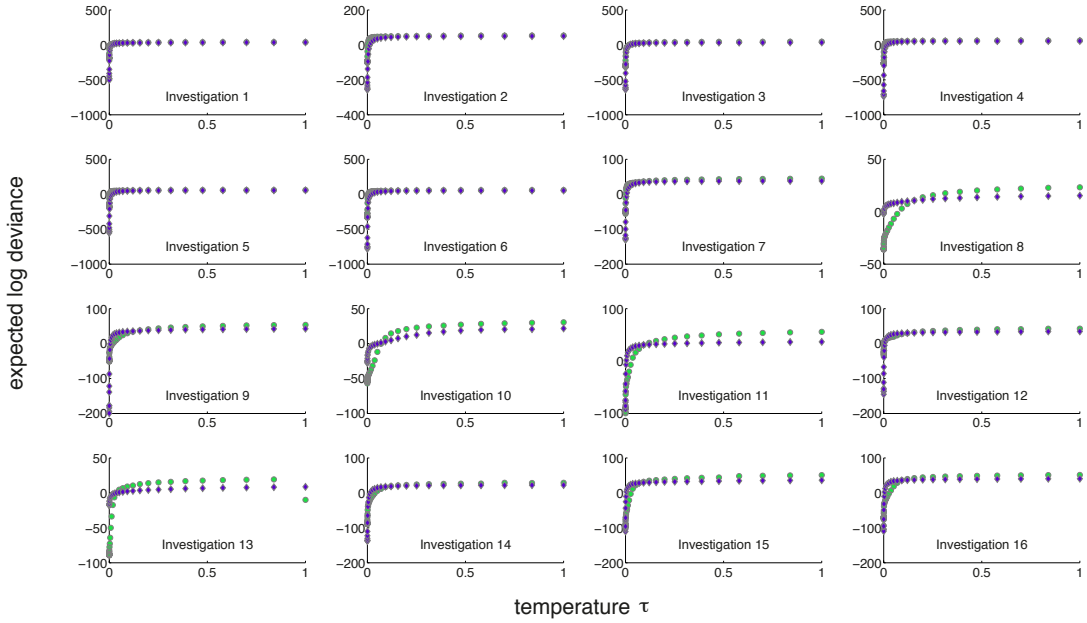


Figure 7.8: Investigation-specific expected log deviances for the zirconium models. Expected log deviances depending on the temperature τ for the all individual investigations. The green circles represent the HMGU model, and purple diamonds the ICRP model.

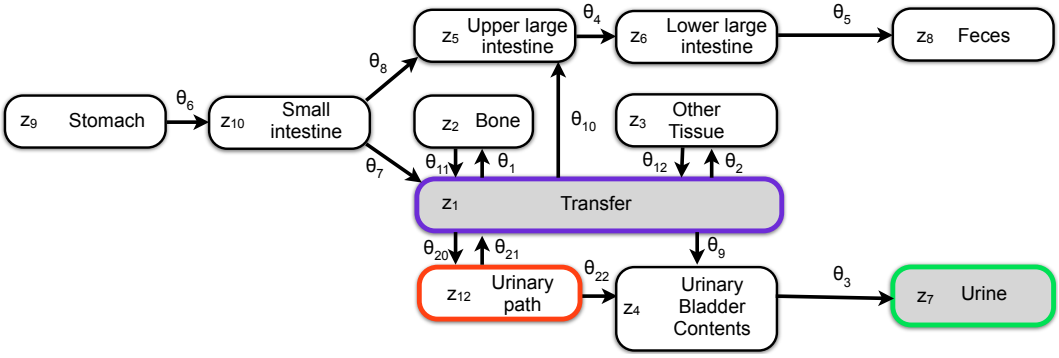


Figure 7.9: Extended HMGU model M_3 for zirconium processing. This model contains now eleven compartments $z_1,\ldots,z_{10},z_{12}$ and 15 transfer rates $\theta_1,\ldots,\theta_{12},\theta_{20},\theta_{21},\theta_{22}$. Zirconium enters the body in the stomach compartment z_9 and is processed through the system until it reaches either one of the two final compartments urine, z_7, or feces, z_8. The gray-shaded compartments z_1 and z_7 are corresponding to those where measurements are taken. The red compartment z_{12} represents the newly added urinary path compartment. It is connected to the transfer compartment and the urinary bladder contents compartment through three reactions with rates $\theta_{20},\theta_{21},\theta_{22}$.

Table 7.3: Additional prior information for the extended HMGU model

Par.	Compartments	Median (d^{-1})	distr.	μ	σ
θ_{20}	TC $\rightarrow$ UP	0.02	$\mathcal{LN}(\mu, \sigma)$	-3.6535	1.0986
θ_{21}	UP $\rightarrow$ TC	0.14	$\mathcal{LN}(\mu, \sigma)$	-1.9661	0.6931
θ_{22}	UP $\rightarrow$ UBC	1.25	$\mathcal{LN}(\mu, \sigma)$	0.2231	0.6931

Also for this model, marginal likelihoods and thus Bayes factors comparing to either of the other two models can be computed in the same fashion as before. We focus on the comparison with the original HMGU model to see if there is a quantitative improvement. The results can be seen in Table 7.4. We observe that there are only few decisive Bayes factors for this comparison, and all of them are in favor of the original HMGU model. Since the extended HMGU model was designed to better reproduce the urine data, especially the last column in Table 7.4, where we use only the urine data for computing Bayes factors, is interesting. Here we see that all but three Bayes factors actually favor the original model, two of them even decisively. We conclude that the extended HMGU model does not yield any gain for representing the measurement data. Therefore the original HMGU model should be preferred. This is also a good example that not always a more complex model is really better than a simpler one.

Table 7.4: Bayes factors for the HMGU versus the extended HMGU model (B^r_{13} and B_{13}) for the individual investigations as well as for the complete data (ALL) and the according Bayes factors for the blood plasma ($B^{b,r}_{13}$ and B^b_{13}) and urine ($B^{u,r}_{13}$ and B^u_{13}) data. Green color indicates a Bayes factor in favor of the original HMGU model and red color a Bayes factor in favor of the extended HMGU model.

Inv. r	B^r_{13} & B_{13}	$B^{b,r}_{13}$ & B^b_{13}	$B^{u,r}_{13}$ & B^u_{13}
1	1.1582	0.9758	1.1019
2	1.1557	1.0307	34.8195
3	0.9283	1.3681	1.3446
4	0.7688	1.6415	14.7689
5	1.4955	1.1345	55.4583
6	1.7465	1.0266	1030.8867
7	0.1836	0.7113	793.4081
8	0.8293	2.5451	0.5550
9	0.1030	0.1407	0.2851
10	0.2548	0.4999	0.7125
11	38.7192	0.7661	13.9183
12	1.6931	0.6516	2.8094
13	12.6123	2.4175	4.6281
14	2.9665	1.1759	2.5476
15	1.0935	2.3948	2.5774
16	274.8282	0.7980	2.4061
ALL	17.6047	0.9893	13.8294

7.4 Conclusions

In summary, we could show that the newer HMGU model was unequivocally superior with 14 of 17 Bayes factors being decisive when compared to the well-established ICRP model. Also, when restricting the data on plasma and urine measurements only, we found that the HMGU model was clearly favored.

We also tested a variant of the original HMGU model, an extended model with an additional compartment for the urinary path. However, the Bayes factors for the comparison of the original versus the extended model favor the original model.

Closer looks at the expected log deviances show that the integration for obtaining the log marginal likelihoods is well-behaved as the function values are mostly monotonically increasing as can be expected from theory.

We also conducted identifiability analysis for our two main models and find that all parameters of the posterior distribution are identifiable. The analysis can yield confidence intervals for the parameters if these are desired. Of course credible intervals can also be derived from the posterior samples obtained through the copula-based sampling in the thermodynamic integration and are given in Appendix A.

7. MODEL SELECTION FOR THE PROCESSING OF ZIRCONIUM IN THE HUMAN BODY

8

Inference in high dimensions: A signaling pathway example

In this chapter, we present a proof of principle that parameter inference through MCMC is possible even in dynamical systems with over 100 parameters. Usually MCMC is only applied in systems of circa ten parameters, even though Eydgahi *et al.* [2013] have recently presented a sampling of 78 parameters. Nevertheless, it is well known that scalability of MCMC algorithms is an issue, which makes our proof of principle even more important. We thus focus on MCMC sampling and not on model selection. Special care has to be taken to verify convergence, as convergence diagnostics like the Geweke test might be misleading. We show how this can be done with a multi-chain sampling approach in combination with identifiability analysis.

The dynamical system under consideration is the JAK2/STAT5 signaling pathway, which is important for erythropoiesis, the production of red blood cells. This system has 27 dynamic and initial condition parameters, however the system is blown up to 113 parameters by scaling and offset parameters that also have to be estimated, thus providing a very challenging example.

This chapter is based on and in part identical with the following publication:

- **S. Hug***, A. Raue*, J. Hasenauer, J. Bachmann, U. Klingmüller, J. Timmer and F.J. Theis (2013). High-dimensional Bayesian parameter estimation: case study for a model of JAK2/STAT5 signaling. Mathematical Biosciences, 246(2), 293-304.

The content of this paper is also to a small part contained in another thesis (Raue [2013]), as this was a joint first-author work. The contribution by the author of this thesis is the MCMC and its interpretation and evaluation, while Dr. Raue provided the profile posteriors.

8.1 Problem description

Here we examine a model for qualitative inference in the JAK/STAT pathway. Signal transduction models are often high-dimensional and possess a large number of unknown parameters, thus the uncertainty evaluation can pose severe computational challenges. In this chapter, we illustrate that a rigorous statistical assessment is feasible for nonlinear high-dimensional dynamical models with over 100 parameters. For this we consider Epo-induced JAK2/STAT5 signaling, a process which has been studied extensively in recent years (e.g. Aaronson & Horvath [2002]; Swameye *et al.* [2003]).

As introduced in Section 2.1.2, the hormone erythropoietin (Epo) regulates erythropoiesis, the production of red blood cells. Figure 8.1 shows again an illustration of the model. The quantitative link between the integral STAT5 response in the nucleus and survival of erythroid progenitor cells has recently been elucidated (Bachmann *et al.* [2011]). The broad dynamical range of Epo concentrations up to 1000-fold in vivo (Becker *et al.* [2010]) requires a stringent regulatory system. In Bachmann *et al.* [2011], it was shown that STAT5 responses are controlled by a dual feedback consisting of two inhibitory proteins, CIS and SOCS3. The two proteins adjust STAT5 phosphorylation levels over the entire range of Epo concentrations, where CIS regulates predominantly the lower concentrations range and SOCS3 the upper range. Model predictions showed that the absence (knock-out) of CIS resulted in an increase of STAT5 phosphorylation at low Epo concentrations, whereas the absence of SOCS3 caused an increase in the phosphorylation level at high Epo concentrations. This observation revealed division of labor by the two feedback proteins as the key property to control STAT5 responses.

8.2 Description of model and experimental data

The ODE system for the JAK2/STAT5 signaling pathway as in Figure 8.1 is described by 25 dynamical components and was solved by using the Data 2 Dynamics software (Raue *et al.* [2013b]). This software enables efficient simulation and optimization of

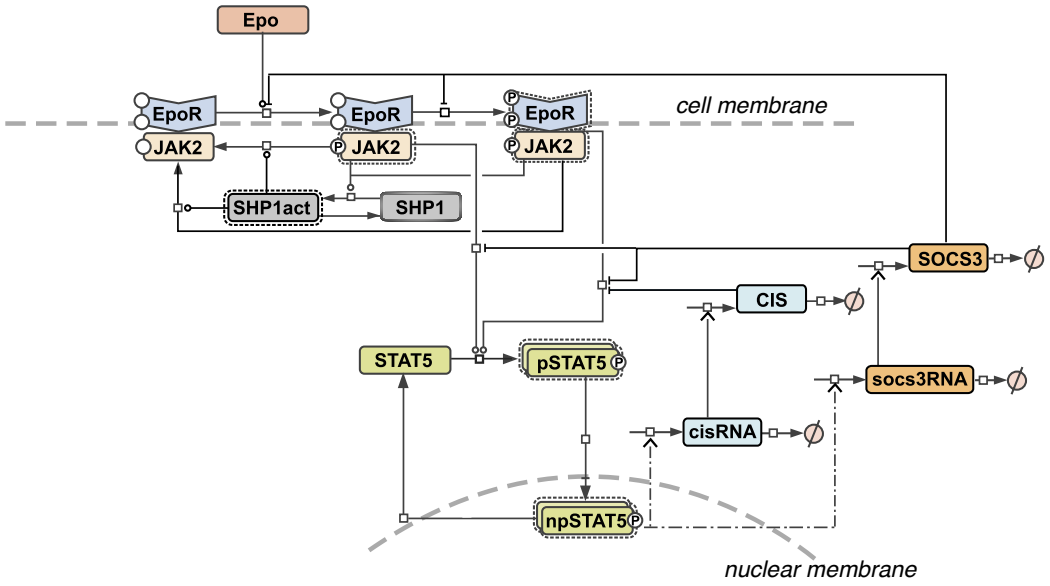


Figure 8.1: Dynamical model of the Epo induced JAK2/STAT5 signal transduction pathway, adopted from Bachmann *et al.* [2011]. The hormone Erythropoietin (Epo) binds to its membrane receptor (EpoR) and subsequently leads to receptor phosphorylation (pEpoR) and to phosphorylation of its associated Janus kinase (JAK2, pJAK2). Receptor phosphorylation is balanced by activation of a phosphatase (SHP1, SHP1act). Active EpoR/JAK2 complexes lead to phosphorylation of the Signal Transducer and Activator of Transcription (STAT5, pSTAT5) that transmits the signal to the nucleus (npSTAT5). In the nucleus, STAT5 leads to target gene expression that induces pro-survival signals and self-regulating negative feedbacks. In this case, two regulator proteins and their respective mRNAs are involved, Suppressor Of Cytokine Signaling (SOCS3) and the Cytokine-Inducible SH2-containing protein (CIS).

ODE models by solving the sensitivity equation together with the original ODE. This is done by using a CVODES solver (Serban & Hindmarsh [2005]) with an absolute and relative accuracy of 10^{-8}. The ODE equations and the solver were compiled into C-executable files for MATLAB. They can also be found in Appendix A. Note that the JAK2/STAT5 system is actually a DDE equation that is transformed to an ODE with the linear chain trick.

Experimental data is available for 24 different experimental conditions, corresponding to different observables, which was described in detail in the supplementary material for Bachmann *et al.* [2011]. As the evaluations of the ODE systems for different experimental conditions are independent, they could be parallelized for numerical efficiency. After all suitable transformations etc. described in more details in Bachmann *et al.* [2011], 115 unknown parameters remain, of these two more could be fixed to a scale. All in all, like for the analysis in Bachmann *et al.* [2011], 113 parameters are sampled by our approach. The experimental data for the dynamics of the system consists of 541 data points which are the basis of our inference. Optimization of the parameters can then be performed with MATLAB optimizers with user supplied derivatives from the sensitivities, the results of this can be found in Raue *et al.* [2013b].

The likelihood of the model is obtained by assuming normally distributed measurement noise on the logarithmically transformed model output and experimental data, since it is known that in immunoblotting experiments, which are predominant here, the measurement noise is lognormally distributed (Kreutz *et al.* [2007]). Also the parameters can be logarithmically transformed for computational convenience. From this point on, we will always assume that these transformations took place. The likelihood then takes the form already introduced in Equation (3.6).

For one of the parameters, the absolute concentration of the EpoR_JAK complex, we found a literature value that could be included as prior information for the concentration scale of the receptor complex into the sampling. For all other parameters, uniform priors in logarithmic parameter space were used. The range of these priors was determined already for the optimization done in Bachmann *et al.* [2011]. It is especially important for those of the parameters that were already shown there to be non-identifiable. The prior in this case prevents the sampler from going to infinity in these parameters which would be detrimental for the complete exploration of the parameter space.

8.3 MCMC for the JAK/STAT model

In this section, we present our results for the high-dimensional model of Epo-induced JAK2/STAT5 signaling. Using this application we illustrate problems which can arise when studying high-dimensional models and outline potential solutions. In this case study 113 parameter have to inferred, 27 parameters of interest which are the dynamical parameters and initial conditions and 86 nuisance parameters. The parameters of interest determine the model predictions, while the nuisance parameters have to be estimated to compare model observables to the experimental data.

8.3.1 Limitations of single-chain sampling

To evaluate the identifiability of the individual parameters we perform at first a profile posterior analysis as introduced in Section 3.3. The results are depicted in Figure 8.3 and Figure 8.4. Indeed, most parameters are well determined, but there are also a few which are practically non-identifiable, e.g. CISRNATurn and SOCS3Turn. A closer inspection of the profile posteriors reveals that two parameters, namely SOCS3RNADelay and SOCS3RNATurn, do exhibit a secondary mode, see in Figure 8.3. The higher mode of these is the MAP estimate found by optimization, while the secondary mode is close to the threshold that defines a 95% confidence region (Raue *et al.* [2009]). To take a closer look at this potential second mode, we used MCMC.

To sample the posterior distribution of the Epo-induced JAK2/STAT5 signaling pathway we first employed a single-chain method. In particular we started with Adaptive Metropolis (AM) sampling as introduced in Section 4.2. When initializing the AM at the MAP estimate, we found that the MCMC chain converged according to the commonly used Geweke's test after 500,000 samples (100,000 burn-in and 400,000 retained samples).

To validate the sampling result, we started a second AM chain in the secondary mode detected using the profile posterior method, as explained above. It turned out that also this second AM run seems to converge after 500,000 samples, according to Geweke's test. However, the sample distributions of the two runs differed severely. The difference was particularly pronounced in the parameters SOCS3RNADelay and SOCS3RNATurn for which we observed the bimodality in the posterior profiles. This indicated that the individual chains indeed sufficiently sample the modes in which they were started, but failed to cover the bimodality of the posterior, see Figure 8.2. We further confirmed

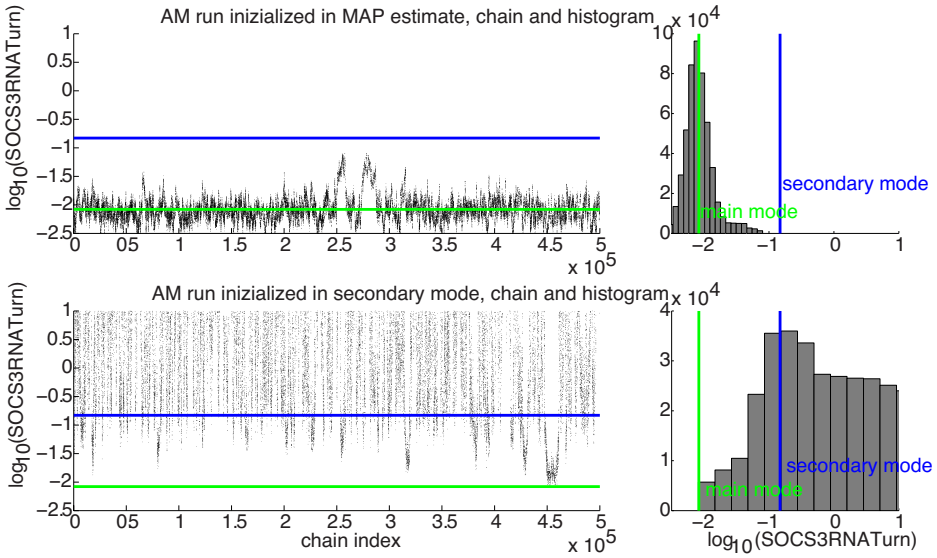


Figure 8.2: Single MCMC chains and profiles for SOCS3RNATurn. The top row displays the chain and histogram for the AM run started in the MAP estimate of exemplary parameter SOC3RNATurn, while the bottom row shows the AM run started in the secondary mode. Both chains show nice mixing, however both the chains and the histograms reveal the totally different marginal distributions.

the non-convergence with the Gelman-Rubin statistic. For the two aforementioned parameters, the values of $\hat{R}$ were 2.28 and 2.65 respectively, indicating that the two chains were not sampling from the same distribution.

To unravel the source of the convergence problems, we analyzed the distribution of the MCMC samples obtained when starting the chains in the two different modes. In particular we studied whether or not the MCMC samples from the two chains are non-overlapping. This would indicate that not only the posterior profiles are bimodal but also that the corresponding modes of the posterior distribution are separated.

While this was already visible from the marginalized one-dimensional samples, we quantified the overlap of the samples with support vector machines (SVMs) (Vapnik [1995]). The SVM allowed us to assign the samples to the two modes in the high-dimensional space and thus visualize them accordingly, which will be shown in Figure 8.4.

Our findings raised doubts concerning the sampling performances in high-dimensional parameter spaces. Although the single adaptive MCMC chains achieved a good sampling performance within the modes and although the modes were connected, the sam-

pling of the true posterior distribution was very inefficient. To improve upon this and to ensure good mixing of the chain, we applied the Adaptive Metropolis Parallel Hierarchical Sampling (AMPHS) as introduced in Section 4.4.

8.3.2 Multi-chain sampling

For the AMPHS we used 20 auxiliary chains, each with 500,000 samples. The AMPHS was initialized with the mother chain and ten auxiliary chains in the MAP estimate, while the other ten auxiliary chains were initialized in local optima, also some which are close to the secondary mode found in the profile posterior.

Furthermore, the AMPHS requires the specification of a starting covariance matrix. We found it sufficient to take an identity matrix in each auxiliary chain. Alternatively, one could run a short chain initialized with an identity matrix and then calculate an initial covariance matrix from these prerun samples. Since it is a special advantage of the Parallel Hierarchical Sampling scheme that different proposals can be used in each chain, we chose different scaling factors for the identity matrix ranging from 10^{-6} to 10^{-9}.

After visual inspection of the mother chain for the dynamical parameters, we set the burn-in period to be the first 100,000 samples, so that the further evaluation could be based on 400,000 samples. Convergence of the chain was again verified by Geweke's test. Although the test p-values were very good, that alone does not ensure convergence of the sampling procedure. However, convergence is supported by the good agreement of marginal distributions and the profile posterior, as will be shown in the following section. Note that the Gelman-Rubin statistic is not easily applicable to the outcome of a single run of AMPHS due to the specific structure of the chains.

By analyzing the mother chain we found that mixing is much enhanced in this algorithm, as clearly the mother chain mixes very well between the two modes. Figure 8.3 depicts the sampling results for the two parameters SOCS3RNADelay and SOCS3RNATurn, which were chosen here because of the bimodality expected from the profile posterior. The bottom and right panels clearly show that the chain mixed very well in the single dimensions. When looking at the samples in two dimensions, the middle panel indicates that they visit both modes, although the main mode obviously has more weight. AMPHS correctly estimated the weight assigned to each mode, see also Rigat & Mira [2012] for additional examples. This can be observed from the fact that the initialization was in a weighting of 50 % of the auxiliary chains near the main

mode and 50 % near the secondary mode. Using the SVM trained from the single chains (Section 8.3.1), we found that about 84% of the final samples in the mother chain belong to the main mode around the MAP estimate and 16% of samples are classified as belonging to the secondary mode. Hence, the masses of the modes seem to have a weight ratio of ca. 5 : 1. This was significantly different from the initial weighting of 1 : 1 and thus together with the convergence of the sampling indicated correct weighting of the modes. A more systematic evaluation of the weighting of the modes depending on the initialization of the chains should be the focus of future work.

To evaluate the AMPHS, we considered all chains, even though for all further analysis, we only used the samples from the mother chain. Each of the auxiliary chains had at the end an acceptance rate of about 6.5%, the mother chain had per definition an acceptance rate of 100%, since the swap with an auxiliary chain was always accepted. While 6.5% might sound suboptimal, we believe that for the AMPHS sampling scheme the acceptance rate is adequate, since the swaps with the mother chain perturb the adaption of the covariance matrix in the auxiliary chains. This is the price that has to be paid for the excellent mixing in the mother chain.

We did not thin the Markov chains, but saved all generated samples. Thus the computational cost for such a large system was quite heavy, the sampling run took about three days on a standard AMD Opteron 2.4 GHz multi-core machine using 5 cores. Higher degrees of parallelization are possible and would reduce the run time.

8.4 Comparison of sampling and profile posterior results

To compare the profile posterior and the AMPHS sampling results, Figure 8.4 shows the histograms of the individual parameters against the corresponding profile posteriors. For most of the parameters, we found an excellent agreement between the shape of the profile posterior and the marginalized samples, cf. exemplary parameters CISEqc or CISEqcOE in Figure 8.4. However, especially for SOCS3RNADelay, we saw a much more pronounced bimodality in the samples than in the profile posterior alone. The same, though not as clearly, held true for SOCS3RNATurn. The difference between sampling result and posterior profile arises from the fact that the height of the modes – as determined by the posterior profile (maximization) – does not necessarily correspond to the mass of the mode (marginalization) – as determined by the MCMC samples. Interestingly, for this example, the ratio between the masses of the modes was rather similar to the ratio of the maximum posterior probability density in the individual

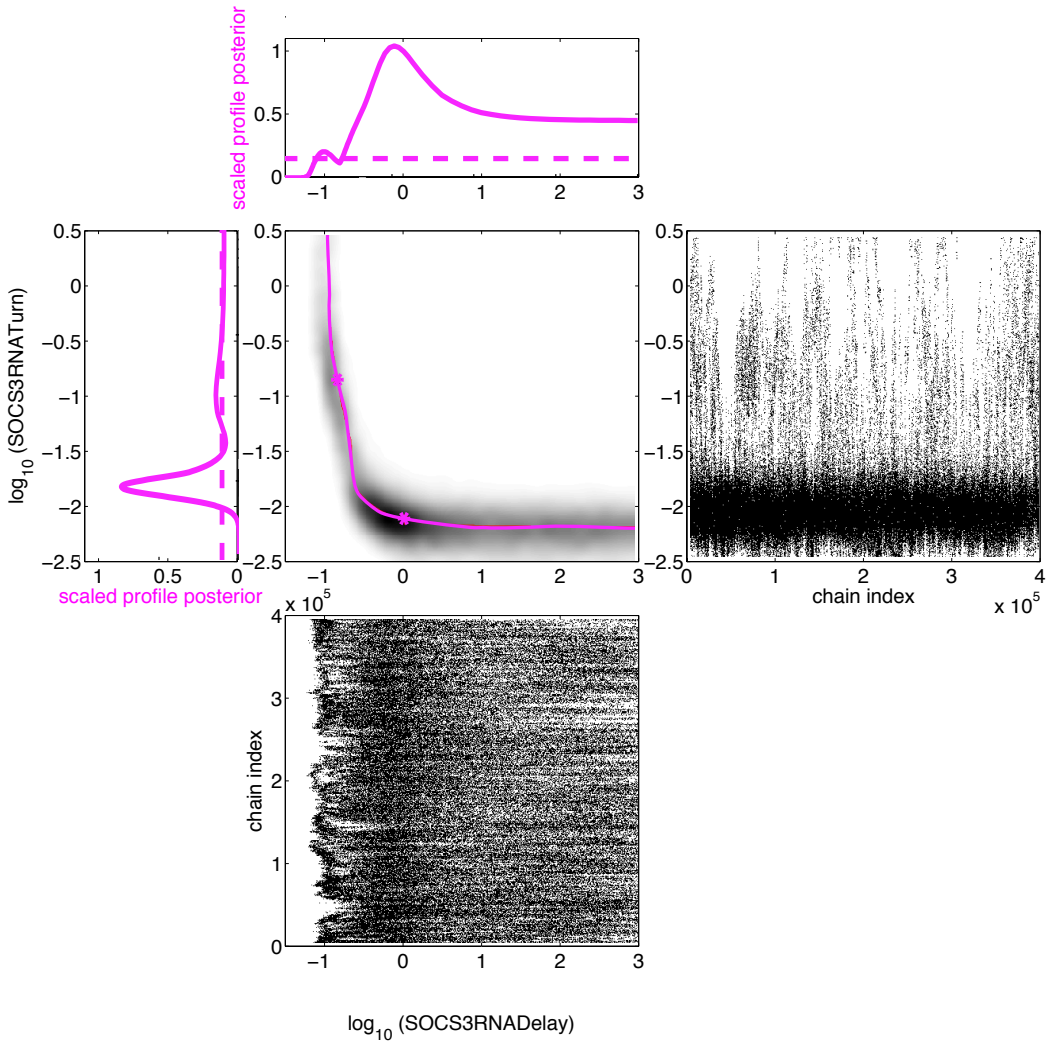


Figure 8.3: MCMC chains and profiles for SOCS3RNADelay and SOCS3RNATurn. The middle panel shows the estimated posterior density over the two parameters in grey. The magenta line is the two-dimensional profile, the magenta stars the two modes. The left and top panels show the one-dimensional profiles, while the bottom and right panels show the MCMC chain for the two parameters. All panels imply a separation of the two modes in the parameter space, although they are connected by a banana-shaped ridge of high posterior density.

modes, which is also roughly 5 : 1.

When taking a closer look at Figure 8.3, again of the two parameters SOCS3RNADelay and SOCS3RNATurn against each other, one can see that the region of high posterior density is not one with two clear modes with deep valleys in between, but rather a banana-shaped ridge with one global and one local maximum. This highly non-elliptical shape also explains the failure of the single chain AM runs to switch between the two modes adequately fast and often. Obviously, an elliptically-shaped normal distribution was not ideally suited as a proposal distribution for inferring a bimodal banana-shaped distribution. However, in combination with the AMPHS scheme, it was sufficiently efficient.

8.5 Model predictions of inhibitory effects

In Bachmann *et al.* [2011], it was already shown that SOCS3 and CIS act as a dual negative feedback on the level of nuclear phosphorylated STAT5, thus providing regulation over a broad range of Epo concentrations. The effect of SOCS3 is more pronounced for high Epo levels, while CIS primarily works as a negative feedback at low Epo levels. In addition to confidence intervals estimated from the profile posterior (see Bachmann *et al.* [2011]) the obtained MCMC samples now also allow computing the posterior density of the prediction, see Figure 8.5.

Regarding the previously observed modes in the parameter posterior distribution we found that the differences between the corresponding predictions for pSTAT5 are minor, see Figure 8.5. This can be explained by the fact that the two modes mainly differ in the parameters SOCS3RNADelay and SOCS3RNATurn which primarily influence SOCS3. As the effect of SOCS3 on pSTAT5 is indirect, the bimodality has negligible effect on the level of pSTAT5 predictions. The results confirm the role of the dual negative feedback. However, the advantage of the sampling-based approach only becomes fully apparent when considering SOCS3 itself.

When we analyzed the predicted dynamics for SOCS3 we found that these indeed depend on the mode, as shown in Figure 8.6. For the parameters in the main mode, the model predicted that after stimulation SOCS3 goes directly to a steady state. In contrast, for the parameters in the secondary mode we observed an overshoot. This overshoot was caused by the increased delay in the SOCS3 RNA export from the nucleus, SOCS3RNADelay. Based on this prediction it would be sufficient to have a

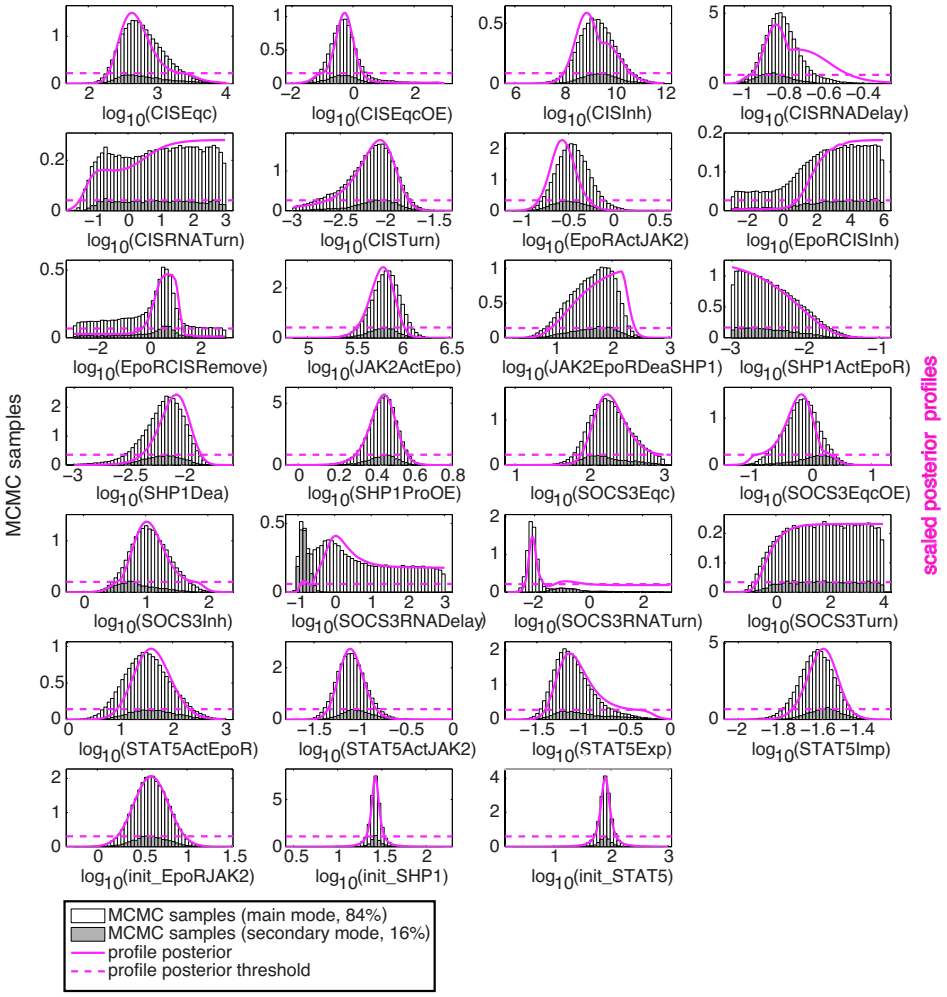


Figure 8.4: MCMC samples and profile posteriors for all 27 inferred dynamical and initial condition parameters. Shown are the histograms of the marginalized MCMC samples, color-coded for mode membership. The height is scaled such that the area of all bars in each histogram is one. In magenta: profile posterior pdf, scaled so as to minimize distance to histogram.

better resolved measurement of SOCS3 between 0 and 100min to distinguish between the two modes.

In summary, the obtained MCMC samples allow for a detailed evaluation of the model. Furthermore, new experiments that allow to further characterize the model and improve the explanatory power were designed.

8.6 Conclusions

Statistical inference for high-dimensional problems is a challenging issue. In this chapter, we provide a proof of concept that Bayesian inference in high-dimensional dynamical systems is feasible. MCMC sampling of over 100 parameters is nevertheless a challenging task. Special care has to be taken when checking and verifying the results.

In line with results obtained for smaller applications (Raue *et al.* [2013a]; Vanlier *et al.* [2012]), we advocate the combination of MCMC sampling with the profile posterior approach to ensure the robustness and reliability of the results.

We have shown that single-chain algorithms can run into severe problems in high-dimensional systems, which are furthermore not easily diagnosed from the MCMC run alone. In the single-chain case, the Geweke test could not detect that the chains get locked in local modes of the posterior. If single chains are run repeatedly from different starting points, the Gelman-Rubin statistics can detect non-convergence. However, this relies on a representative set of starting points that have to be determined beforehand. For the multi-chain approach, the selection of representative starting points is important to ensure convergence to the posterior distribution, i.e. correct weighting of the posterior modes, in acceptable time. Once reliable results of MCMC sampling are obtained, uncertainties can easily be projected on any model prediction including the high-dimensional correlation structure.

In high-dimensional systems it is important to have excellent mixing in the Markov chain to ensure exploration of the whole parameter space. Multi-chain approaches are clearly superior to single-chain sampling schemes from the Metropolis-Hastings family with respect to the mixing. Alternatively, other sampling schemes that provide good mixing could be employed, for example an independence walk, however it is unclear what kind of problems one could run into when using these. Additionally, the shape of the parameter space with dynamical and nuisance parameters could be exploited for a blocked update in the sampling.

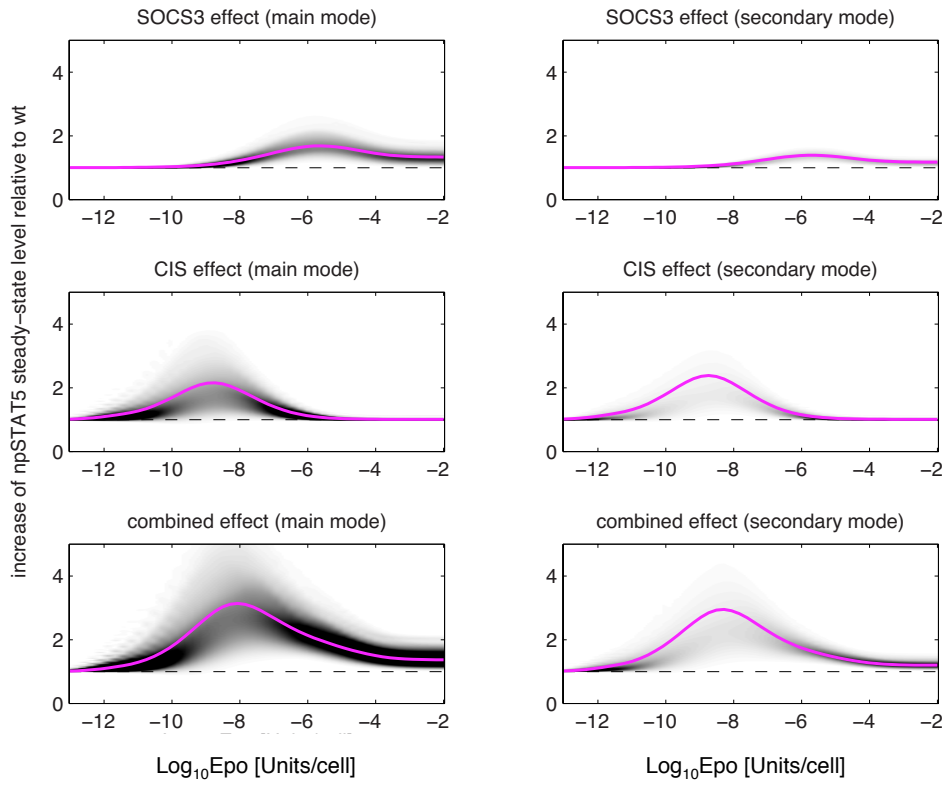


Figure 8.5: Uncertainty in prediction of the cellular response. Simulation of the steady-state level of phosphorylated STAT5 in the nucleus, with only one transcriptional negative regulator, CIS or SOCS3, being present and their combined effect. The increase of pSTAT5 steady-state levels was calculated relative to wild-type cells (black dashed line) in steady state. Grey shading indicates the density calculated from the posterior samples, the magenta line represents the solution belonging to the MAP estimate. Left column based on main mode samples, right column based on secondary mode samples.

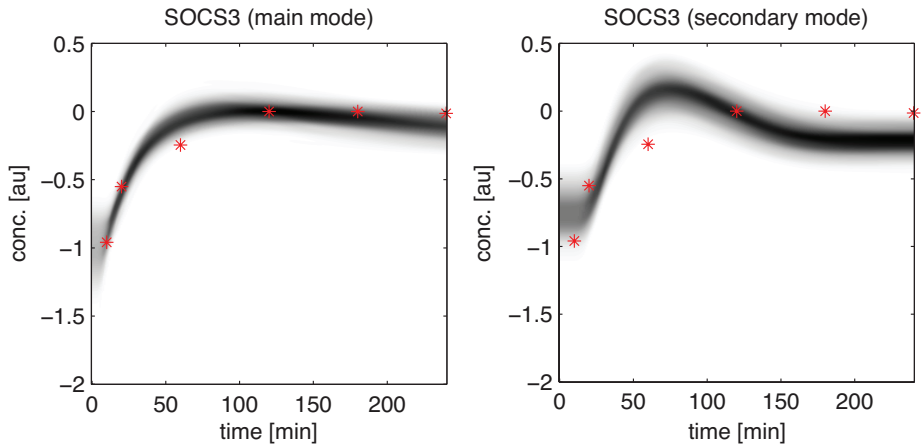


Figure 8.6: Differences in SOCS3 dynamics. Experimental data for SOCS3 (red stars) and density of the trajectories corresponding to the MCMC samples for main mode (left) and secondary mode (right). Grey shading indicates the density calculated from the posterior samples classified as belonging to the respective mode. For the parameters in the main mode, the model predicts that after stimulation SOCS3 goes directly to a steady state. In contrast, for the parameters in the secondary mode we observe an overshoot.

Using the sampling results we could show that the posterior distribution is bimodal and that the two modes correspond to alternative parameterizations of the model. Either the turnover rates of SOCS3 RNA can be high and the RNA export in the cytosol low, or vice versa. By inspection of the predictions corresponding to the individual modes it would be possible to verify one of the scenarios experimentally. The sampling results were used for the prediction of the inhibitory effect of SOCS3 and CIS for different Epo levels, as well as for the dynamics of SOCS3. The two modes of the posterior clearly manifest in the SOCS3 dynamics. This is due to the fact that the two parameters showing the bimodality most prominently, namely SOCS3RNADelay and SOCS3RNATurn, are directly linked to SOCS3.

9

Discussion and outlook

In this thesis, we presented novel Bayesian methods for the inference and model selection in ODE models as well as three challenging application examples from systems biology.

In this chapter, we give a short review of the methods and applications that were discussed, with a focus on the occurring issues, how to overcome them and what insights could be gained. Furthermore, we provide some cues for targets of further research.

9.1 Summary

For Markov chain Monte Carlo sampling of high-dimensional distributions, we introduced the Adaptive Metropolis Parallel Hierarchical Sampling based on the concept by Rigat & Mira [2012]. Exchanging multi-chain algorithms are clearly superior to single-chain algorithms for covering large sampling spaces. Since scalability of MCMC algorithms is a well-known and much discussed issue, it is especially worthwhile to have a proof of principle that MCMC can work reliably in over 100 dimensions.

We furthermore introduced an adaptive method for thermodynamic integration. Thermodynamic integration is nowadays usually the method of choice for computing marginal likelihoods. Its strength is the transformation of the marginal likelihood to a one dimensional integral over the so called temperature parameter.

The new method introduced in this thesis is based upon solving this one dimensional integral with Simpson's rule. This scheme adaptively determines the number of function

evaluations that are necessary for achieving a required accuracy. This is especially important since every function evaluation corresponds to a full MCMC run and is thus computationally costly. Furthermore, Simpson's rule possesses a higher approximation order than the usually applied trapezoidal rule.

We elaborated on an analytically tractable example for model selection, based on normal distributions. Knowing the expected log deviance is especially beneficial for comparing the different model selection indicators.

Additionally to the presented purely Bayesian methods, we have also discussed identifiability analysis with profile likelihoods or profile posteriors. In line with results obtained for smaller applications (Raue *et al.* [2013a]; Vanlier *et al.* [2012]), we advocate the combination of MCMC sampling with the profile posterior approach to ensure the robustness and reliability of the inference results, as shown in all three examples. Especially for the JAK/STAT pathway in Chapter 8, the combination of profiles and MCMC was worthwhile. Here agreement between the profile posteriors and the MCMC samples makes us more confident that both methods yield reliable results.

In the single-cell application of Chapter 6, we find that thermodynamic integration with the adaptive Simpson's rule works well. Nevertheless, the application is a practical example of the detrimental effects of weak prior information on the computation of Bayes factors and the numerical issues that arise.

Protein half-lives derived from the best model are important for assessing all experiments in which proteins are observed in living cells. The measurement data in our application represents the decay of the protein that can be observed isolated for one protein in single cells. This is an important basis for more complex models where several proteins may interact with each other. A rigorous statistical evaluation as presented in this contribution is an important cornerstone for further inference.

For the processing of zirconium in the human body in Chapter 7, we could show that the physiologically more plausible HMGU model indeed better represents the measurement data than the previously used ICRP model. However, a more complex variant of the HMGU model could be repudiated on the basis of Bayes factors. In this application, we showed identifiability analysis and Bayes factors for both the sixteen individual investigations as well as for the combined data.

To check the convergence of MCMC schemes in high-dimensional parameter spaces or multimodal problems in the JAK/STAT pathway or in general we used a combination of approaches. During our studies we found that the complementation of classical

convergence criteria (Brooks & Roberts [1998]) with information about the modes is beneficial. This information can be obtained using e.g. multi-start optimization algorithms. If a multi-chain sampler like the AMPHS is initialized distributed across all detected modes and the convergence diagnostics indicates convergence, it is more likely that all modes of the posterior distribution have been adequately sampled, compared to the single-chain approach. As multi-start optimization is often used to determine the MAP estimate (Bachmann *et al.* [2011]), all necessary information is readily available and does not require additional computational effort.

The sampling of high-dimensional posterior distributions is methodologically and computationally challenging. However, obtaining the full distribution of parameters is worth the effort, especially when the likelihood is not strictly unimodal such as the one in Chapter 8. Using the sampling results we could show that the posterior distribution is bimodal and that the two modes correspond to alternative parameterizations of the model.

9.2 Outlook

In general, we see three main targets for future research following up from this thesis:

1. Further improvements to thermodynamic integration,
2. MCMC in high-dimensional systems and
3. Inference for single-cell data.

We now elaborate on each of the three points.

This thesis presents the first application of the adaptive Simpson's rule for thermodynamic integration. The promising results are a motivation to study further sophisticated quadrature methods, like e.g. Romberg's method, Gaussian quadrature or Clenshaw-Curtis quadrature. However, for all of these methods, many function evaluations might be necessary. Thus alternative quadrature methods should be chosen very carefully.

The quadrature methods applied to the thermodynamic integral in this thesis only control the integration error, not the Monte Carlo error on the function evaluations themselves. As both the trapezoidal and the adaptive Simpson's rule are relatively simple, error propagation is straightforward. A simultaneous estimation of both quadrature

and Monte Carlo error should yield better estimates of the error and refine the integration strategy further.

In some applications, thermodynamic integration was combined with sequential sampling from the tempered distributions, or population-based MCMC (Calderhead & Girolami [2009]). This is of course also possible with the new adaptive Simpson's rule and might increase sampling efficiency, especially in examples that might be more difficult to sample than the examples presented in this thesis.

The analytical example in Chapter 5 also shows that determining a sensible exponent for a power law schedule is not straightforward and further research should go into this, since this could lead to schedules close to the optimal temperature schedule as described in Calderhead & Girolami [2009].

In this thesis, we have established a proof of principle that MCMC is possible in over 100 dimensions. This is highly relevant, since increasing biological knowledge leads to larger and larger models being proposed. One relevant application for high-dimensional inference are for example PBPK (Physiologically based pharmacokinetic) models, which are mechanistic models used for predicting the absorption, distribution, metabolism and excretion of a substance in the human body (Gelman *et al.* [1996a]; Shargel *et al.* [2005]). These models are often fitted to several patients' data simultaneously, creating high-dimensional problems. Since these models sometimes show specific characteristics such as oscillatory solutions, e.g. in the cardiovascular system, they require specific MCMC algorithms.

High-dimensional sampling spaces require specially tailored MCMC algorithms. Further possible improvements include the use of gradient information for multi-chain algorithms, or more generally using local curvature information for the sampling procedure. Also parallel hierarchical sampling in combination with copula information as in the CIMH algorithm might be a worthwhile research target.

A prominent example for high-dimensional systems are signaling pathways. Here an interesting target for future research is MCMC sampling results for the signaling pathway based on variable types of data, e.g. integrating Western blot data with data for cell morphology. This induces interesting dependence structures in the parameters which should be exploited for efficient sampling.

Thirdly, single-cell data poses an interesting target for method development. In this context especially ODE-constrained mixture models can be of interest (Hasenauer *et al.* [2014]. Also hierarchical models like in Woodcock *et al.* [2013] might be extended to

fluorescence intensity, where copy numbers of molecules are large.

Another interesting research project would be model selection between an ODE model for single-cell dynamics like presented in this thesis and a stochastic differential equation (SDE) model, where each cell can be regarded as one realization of the stochastic process described by the SDE. Such a model selection should be based on information theoretic approaches instead of Bayesian ones due to the very different natures of the two likelihoods, e.g. along the lines of Chehreghani *et al.* [2012].

All methods presented in this thesis are applicable to many dynamical systems, not only from systems biology. Especially in nonlinear and high-dimensional systems special care has to be taken when doing model inference or model selection. All methods in this thesis and their combinations aim at obtaining reliable and robust results. Overall, these are vital for the holistic understanding of biological processes.

Appendix A

Ordinary differential equations for the presented examples

A.1 Zirconium models

A.1.1 HMGU model

The model for biokinetics of zirconium put forward by the Helmholtz Zentrum München (HMGU) consists of ten compartments $z_1,\ldots,z_{10}$ and twelve reaction rates $\theta_1,...,\theta_{12}$ (Greiter *et al.* [2011a]). The extended HMGU model additionally contains the compartment z_{12} and the reaction rates θ_{20}, θ_{21} and θ_{22}. In either model zirconium enters the body in the stomach compartment z_9 and diffuses through the system until it reaches either one of the two final compartments urine, z_7, or feces, z_8.

Mathematically, the original HMGU model is described by the following system of

coupled ODEs:

$$
\begin{aligned}
\frac{\mathrm{d}z_1(t)}{\mathrm{d}t} &= (-\theta_1 - \theta_2 - \theta_9 - \theta_{10})\, z_1(t) + \theta_{11} z_2(t) + \theta_{12} z_3(t) + \theta_7 z_{10}(t) \\
\frac{\mathrm{d}z_2(t)}{\mathrm{d}t} &= \theta_1 z_1(t) - \theta_{11} z_2(t) \\
\frac{\mathrm{d}z_3(t)}{\mathrm{d}t} &= \theta_2 z_1(t) - \theta_{12} z_3(t) \\
\frac{\mathrm{d}z_4(t)}{\mathrm{d}t} &= \theta_9 z_1(t) - \theta_3 z_4(t) \\
\frac{\mathrm{d}z_5(t)}{\mathrm{d}t} &= \theta_{10} z_1(t) - \theta_4 z_5(t) + \theta_8 z_{10}(t) \\
\frac{\mathrm{d}z_6(t)}{\mathrm{d}t} &= \theta_4 z_5(t) - \theta_5 z_6(t) \\
\frac{\mathrm{d}z_7(t)}{\mathrm{d}t} &= \theta_3 z_4(t) \\
\frac{\mathrm{d}z_8(t)}{\mathrm{d}t} &= \theta_5 z_6(t) \\
\frac{\mathrm{d}z_9(t)}{\mathrm{d}t} &= -\theta_6 z_9(t) \\
\frac{\mathrm{d}z_{10}(t)}{\mathrm{d}t} &= \theta_6 z_9(t) + (-\theta_7 - \theta_8)\, z_{10}(t)
\end{aligned}
$$

with initial conditions

$$
\begin{aligned}
z_1(t=0) &= 0\% \\
z_2(t=0) &= 0\% \\
z_3(t=0) &= 0\% \\
z_4(t=0) &= 0\% \\
z_5(t=0) &= 0\% \\
z_6(t=0) &= 0\% \\
z_7(t=0) &= 0\% \\
z_8(t=0) &= 0\% \\
z_9(t=0) &= 100\% \\
z_{10}(t=0) &= 0\%.
\end{aligned}
$$

The extended HMGU model is identical to the original HMGU model except for the

following changes (the additional reaction rates and compartment are depicted in red):

$$\begin{aligned}
\frac{\mathrm{d}z_1(t)}{\mathrm{d}t} &= (-\theta_1 - \theta_2 - \theta_9 - \theta_{10} - \theta_{20})\, z_1(t) + \theta_{11} z_2(t) + \theta_{12} z_3(t) + \theta_7 z_{10}(t) + \theta_{21} z_{12}(t) \\
\frac{\mathrm{d}z_4(t)}{\mathrm{d}t} &= \theta_9 z_1(t) - \theta_3 z_4(t) + \theta_{22} z_{12}(t) \\
\frac{\mathrm{d}z_{12}(t)}{\mathrm{d}t} &= \theta_{20} z_1(t) + (-\theta_{21} - \theta_{22})\, z_{12}(t)
\end{aligned}$$

with initial conditions

$$\begin{aligned}
z_1(t=0) &= 0\% \\
z_2(t=0) &= 0\% \\
z_3(t=0) &= 0\% \\
z_4(t=0) &= 0\% \\
z_5(t=0) &= 0\% \\
z_6(t=0) &= 0\% \\
z_7(t=0) &= 0\% \\
z_8(t=0) &= 0\% \\
z_9(t=0) &= 100\% \\
z_{10}(t=0) &= 0\% \\
z_{12}(t=0) &= 0\%.
\end{aligned}$$

The initial conditions for compartments $z_1, \ldots, z_{10}$ coincide for both models.

A.1.2 ICRP model

The model for biokinetics of zirconium put forward by the International Commission on Radiological Protection (ICRP) is a compartmental model consisting of eleven compartments $z_1,\ldots,z_{11}$ and 15 reaction rates $\theta_1,\ldots,\theta_8,\theta_{13},\ldots,\theta_{19}$ (ICRP [1989], ICRP [1993]). Zirconium enters the body in the stomach compartment z_9 and diffuses through the system until it reaches either one of the two final compartments urine, z_7, or feces, z_8.

A. ORDINARY DIFFERENTIAL EQUATIONS FOR THE PRESENTED EXAMPLES

Mathematically, the model is described by the following system of eleven coupled ODEs:

$$
\begin{aligned}
\frac{\mathrm{d}z_1(t)}{\mathrm{d}t} &= (-\theta_1 - \theta_2 - \theta_{13})\, z_1(t) + \theta_7 z_{10}(t) \\
\frac{\mathrm{d}z_2(t)}{\mathrm{d}t} &= \theta_1 z_1(t) + (-\theta_{14} - \theta_{15})\, z_2(t) \\
\frac{\mathrm{d}z_3(t)}{\mathrm{d}t} &= \theta_2 z_1(t) + (-\theta_{16} - \theta_{17})\, z_3(t) \\
\frac{\mathrm{d}z_4(t)}{\mathrm{d}t} &= \theta_{14} z_2(t) + \theta_{16} z_3(t) - \theta_3 z_4(t) + \theta_{18} z_{11}(t) \\
\frac{\mathrm{d}z_5(t)}{\mathrm{d}t} &= \theta_{15} z_2(t) + \theta_{17} z_3(t) - \theta_4 z_5(t) + \theta_8 z_{10}(t) + \theta_{19} z_{11}(t) \\
\frac{\mathrm{d}z_6(t)}{\mathrm{d}t} &= \theta_4 z_5(t) - \theta_5 z_6(t) \\
\frac{\mathrm{d}z_7(t)}{\mathrm{d}t} &= \theta_3 z_4(t) \\
\frac{\mathrm{d}z_8(t)}{\mathrm{d}t} &= \theta_5 z_6(t) \\
\frac{\mathrm{d}z_9(t)}{\mathrm{d}t} &= -\theta_6 z_9(t) \\
\frac{\mathrm{d}z_{10}(t)}{\mathrm{d}t} &= \theta_6 z_9(t) + (-\theta_7 - \theta_8)\, z_{10}(t) \\
\frac{\mathrm{d}z_{11}(t)}{\mathrm{d}t} &= \theta_{13} z_1(t) + (-\theta_{18} - \theta_{19})\, z_{11}(t)
\end{aligned}
$$

with initial conditions

$$
\begin{aligned}
z_1(t=0) &= 0\% \\
z_2(t=0) &= 0\% \\
z_3(t=0) &= 0\% \\
z_4(t=0) &= 0\% \\
z_5(t=0) &= 0\% \\
z_6(t=0) &= 0\% \\
z_7(t=0) &= 0\% \\
z_8(t=0) &= 0\% \\
z_9(t=0) &= 100\% \\
z_{10}(t=0) &= 0\% \\
z_{11}(t=0) &= 0\%.
\end{aligned}
$$

A.1.3 Solution of the ODE systems

For the calculation of the likelihood $p(\mathbf{Y}_r|\boldsymbol{\theta}^i, \mathbf{M}_i)$, the ODE has to be solved based on the current parameters $\boldsymbol{\theta}^i$. Since the ODEs at question are of first order, they can be written as

$$\frac{d\mathbf{z}_{\boldsymbol{\theta}^i}(t)}{dt} = \mathbf{a}^i(\boldsymbol{\theta}^i) \cdot \mathbf{z}_{\boldsymbol{\theta}^i}(t), \tag{A.1}$$

where $\mathbf{z}_{\boldsymbol{\theta}^i}(t)$ is the vector of all the compartments of model $\mathbf{M}_i$ and the time independent matrix $\mathbf{a}^i(\boldsymbol{\theta}^i)$ represents all the interactions between these compartments, depending on the transfer rate values $\boldsymbol{\theta}^i$. The analytical solution is then given by

$$\mathbf{z}_{\boldsymbol{\theta}^i}(t) = e^{\mathbf{a}^i(\boldsymbol{\theta}^i)t} \cdot \mathbf{z}_{\boldsymbol{\theta}^i}(t=0). \tag{A.2}$$

The matrix exponential $e^{\mathbf{a}^i(\boldsymbol{\theta}^i)t}$ was computed by decomposing

$$\mathbf{a}^i(\boldsymbol{\theta}^i) = U(\boldsymbol{\theta}^i)\ \mathrm{diag}\left(d_1(\boldsymbol{\theta}^i), d_2(\boldsymbol{\theta}^i), \ldots, d_V(\boldsymbol{\theta}^i)\right)\ U(\boldsymbol{\theta}^i)^{-1} \tag{A.3}$$

such that

$$e^{\mathbf{a}^i(\boldsymbol{\theta}^i)t} = U(\boldsymbol{\theta}^i) \begin{pmatrix} e^{d_1(\boldsymbol{\theta}^i)t} & \cdots & \cdots & 0 \\ \vdots & e^{d_2(\boldsymbol{\theta}^i)t} & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & e^{d_V(\boldsymbol{\theta}^i)t} \end{pmatrix} U(\boldsymbol{\theta}^i)^{-1} \tag{A.4}$$

for the eigenvalues $d_1(\boldsymbol{\theta}^i), d_2(\boldsymbol{\theta}^i), \ldots, d_V(\boldsymbol{\theta}^i)$ of $\mathbf{a}^i(\boldsymbol{\theta}^i)$. For the HMGU model V=10, for the extended HMGU model and the ICRP model V=11. In our case, the eigenvalues and matrices $U(\boldsymbol{\theta}^i), U(\boldsymbol{\theta}^i)^{-1}$ were numerically approximated by MATLAB's `eig` function.

By using the `eig` function, we of course introduce numerical error to our solution, however we verified that it is of the same order as the error made by the `ode45` function, yet calculations are much faster. Also, if the analytical solution is available, it is clear that it should be preferred to a purely numerical solution.

A.1.4 Regions of highest posterior density

We derived the 95% credible intervals, i.e. the regions of highest posterior density from the posterior samples of the complete data run for the HMGU model, see table A.1. Furthermore, we also give the maximum a posteriori (MAP) estimate as the sample with the highest posterior value. Since these parameter values are derived from the concatenated data, they are valid for all investigations and thus represent the parameters of choice for an average subject, where no individual information or measurements are available.

A. ORDINARY DIFFERENTIAL EQUATIONS FOR THE PRESENTED EXAMPLES

Table A.1: Credible intervals for best parameters - from the posterior samples of the complete data for the HMGU model

Param.	θ_1	θ_2	θ_3	θ_4
95% CI	[0.03,0.42]	[0.63,2.99]	[7.14,20.91]	[1.03,3.18]
MAP	0.08	1.48	9.54	1.28
Param.	θ_5	θ_6	θ_7	θ_8
95% CI	[0.47,1.55]	[17.57,45.15]	[0.10,0.61]	[19.58,134.48]
MAP	1.03	37.43	0.19	41.86
Param.	θ_9	θ_{10}	θ_{11}	θ_{12}
95% CI	[0.12,0.28]	$[6.75 \cdot 10^{-4}, 0.06]$	$[1.86 \cdot 10^{-5}, 2.57 \cdot 10^{-4}]$	[0.14,0.82]
MAP	0.20	0.0028	$3.57 \cdot 10^{-5}$	0.27

A.2 Equations of the JAK2/STAT5 model

The rate equations of the reactions are

$$
\begin{aligned}
v_1 &= \frac{[\text{Epo}] \cdot [\text{EpoRJAK2}] \cdot \text{JAK2ActEpo}}{[\text{SOCS3}] \cdot \text{SOCS3Inh} + 1} \\
v_2 &= [\text{EpoRpJAK2}] \cdot \text{JAK2EpoRDeaSHP1} \cdot [\text{SHP1Act}] \\
v_3 &= \frac{[\text{EpoRpJAK2}] \cdot \text{EpoRActJAK2}}{[\text{SOCS3}] \cdot \text{SOCS3Inh} + 1} \\
v_4 &= \frac{3 \cdot [\text{EpoRpJAK2}] \cdot \text{EpoRActJAK2}}{(\text{EpoRCISInh} \cdot [\text{EpoRJAK2\_CIS}] + 1) \cdot ([\text{SOCS3}] \cdot \text{SOCS3Inh} + 1)} \\
v_5 &= \frac{3 \cdot \text{EpoRActJAK2} \cdot [\text{p1EpoRpJAK2}]}{(\text{EpoRCISInh} \cdot [\text{EpoRJAK2\_CIS}] + 1) \cdot ([\text{SOCS3}] \cdot \text{SOCS3Inh} + 1)} \\
v_6 &= \frac{\text{EpoRActJAK2} \cdot [\text{p2EpoRpJAK2}]}{[\text{SOCS3}] \cdot \text{SOCS3Inh} + 1} \\
v_7 &= \text{JAK2EpoRDeaSHP1} \cdot [\text{SHP1Act}] \cdot [\text{p1EpoRpJAK2}] \\
v_8 &= \text{JAK2EpoRDeaSHP1} \cdot [\text{SHP1Act}] \cdot [\text{p2EpoRpJAK2}] \\
v_9 &= \text{JAK2EpoRDeaSHP1} \cdot [\text{SHP1Act}] \cdot [\text{p12EpoRpJAK2}] \\
v_{10} &= [\text{EpoRJAK2\_CIS}] \cdot \text{EpoRCISRemove} \cdot ([\text{p12EpoRpJAK2}] + [\text{p1EpoRpJAK2}]) \\
v_{11} &= [\text{SHP1}] \cdot \text{SHP1ActEpoR} \cdot ([\text{EpoRpJAK2}] + [\text{p12EpoRpJAK2}] + \\
&\quad + [\text{p1EpoRpJAK2}] + [\text{p2EpoRpJAK2}]) \\
v_{12} &= \text{SHP1Dea} \cdot [\text{SHP1Act}] \\
v_{13} &= \frac{[\text{STAT5}] \cdot \text{STAT5ActJAK2}}{[\text{SOCS3}] \cdot \text{SOCS3Inh} + 1} \cdot ([\text{EpoRpJAK2}] + [\text{p12EpoRpJAK2}] + \\
&\quad + [\text{p1EpoRpJAK2}] + [\text{p2EpoRpJAK2}]) \\
v_{14} &= \frac{[\text{STAT5}] \cdot \text{STAT5ActEpoR} \cdot ([\text{p12EpoRpJAK2}] + [\text{p1EpoRpJAK2}])^2}{([\text{CIS}] \cdot \text{CISInh} + 1) \cdot ([\text{SOCS3}] \cdot \text{SOCS3Inh} + 1)} \\
v_{15} &= \text{STAT5Imp} \cdot [\text{pSTAT5}] \\
v_{16} &= \text{STAT5Exp} \cdot [\text{npSTAT5}] \\
v_{17} &= -\text{CISRNAEqc} \cdot \text{CISRNATurn} \cdot [\text{npSTAT5}] \\
v_{18} &= [\text{CISnRNA1}] \cdot \text{CISRNADelay} \\
v_{19} &= [\text{CISnRNA2}] \cdot \text{CISRNADelay} \\
v_{20} &= [\text{CISnRNA3}] \cdot \text{CISRNADelay} \\
v_{21} &= [\text{CISnRNA4}] \cdot \text{CISRNADelay} \\
v_{22} &= [\text{CISnRNA5}] \cdot \text{CISRNADelay} \\
v_{23} &= [\text{CISRNA}] \cdot \text{CISRNATurn} \\
v_{24} &= [\text{CISRNA}] \cdot \text{CISEqc} \cdot \text{CISTurn} \\
v_{25} &= [\text{CIS}] \cdot \text{CISTurn} \\
v_{26} &= -\text{SOCS3RNAEqc} \cdot \text{SOCS3RNATurn} \cdot [\text{npSTAT5}] \\
v_{27} &= [\text{SOCS3nRNA1}] \cdot \text{SOCS3RNADelay} \\
v_{28} &= [\text{SOCS3nRNA2}] \cdot \text{SOCS3RNADelay}
\end{aligned}
$$

$$
\begin{aligned}
v_{29} &= [\text{SOCS3nRNA3}] \cdot \text{SOCS3RNADelay} \\
v_{30} &= [\text{SOCS3nRNA4}] \cdot \text{SOCS3RNADelay} \\
v_{31} &= [\text{SOCS3nRNA5}] \cdot \text{SOCS3RNADelay} \\
v_{32} &= [\text{SOCS3RNA}] \cdot \text{SOCS3RNATurn} \\
v_{33} &= [\text{SOCS3RNA}] \cdot \text{SOCS3Eqc} \cdot \text{SOCS3Turn} \\
v_{34} &= [\text{SOCS3}] \cdot \text{SOCS3Turn}
\end{aligned}
$$

Reactions v_{18} to v_{22} and v_{27} to v_{31} account for a delay that summarize the processing steps of the mRNA by a linear chain of reactions (MacDonald [1976]) with common rate constant CISRNADelay and SOCS3RNADelay, respectively. The ODE systems is composed out of the rate equations by

$$
\begin{aligned}
\text{d[EpoRJAK2]}/\text{d}t &= -v_1 + v_2 + v_7 + v_8 + v_9 \\
\text{d[EpoRpJAK2]}/\text{d}t &= +v_1 - v_2 - v_3 - v_4 \\
\text{d[p1EpoRpJAK2]}/\text{d}t &= +v_3 - v_5 - v_7 \\
\text{d[p2EpoRpJAK2]}/\text{d}t &= +v_4 - v_6 - v_8 \\
\text{d[p12EpoRpJAK2]}/\text{d}t &= +v_5 + v_6 - v_9 \\
\text{d[EpoRJAK2\_CIS]}/\text{d}t &= -v_{10} \\
\text{d[SHP1]}/\text{d}t &= -v_{11} + v_{12} \\
\text{d[SHP1Act]}/\text{d}t &= +v_{11} - v_{12} \\
\text{d[STAT5]}/\text{d}t &= -v_{13} - v_{14} + v_{16} \cdot \frac{0.275}{0.4} \\
\text{d[pSTAT5]}/\text{d}t &= +v_{13} + v_{14} - v_{15} \\
\text{d[npSTAT5]}/\text{d}t &= +v_{15} \cdot \frac{0.4}{0.275} - v_{16} \\
\text{d[CISnRNA1]}/\text{d}t &= +v_{17} - v_{18} \\
\text{d[CISnRNA2]}/\text{d}t &= +v_{18} - v_{19} \\
\text{d[CISnRNA3]}/\text{d}t &= +v_{19} - v_{20} \\
\text{d[CISnRNA4]}/\text{d}t &= +v_{20} - v_{21} \\
\text{d[CISnRNA5]}/\text{d}t &= +v_{21} - v_{22} \\
\text{d[CISRNA]}/\text{d}t &= +v_{22} \cdot \frac{0.275}{0.4} - v_{23} \\
\text{d[CIS]}/\text{d}t &= +v_{24} - v_{25} \\
\text{d[SOCS3nRNA1]}/\text{d}t &= +v_{26} - v_{27} \\
\text{d[SOCS3nRNA2]}/\text{d}t &= +v_{27} - v_{28} \\
\text{d[SOCS3nRNA3]}/\text{d}t &= +v_{28} - v_{29} \\
\text{d[SOCS3nRNA4]}/\text{d}t &= +v_{29} - v_{30} \\
\text{d[SOCS3nRNA5]}/\text{d}t &= +v_{30} - v_{31} \\
\text{d[SOCS3RNA]}/\text{d}t &= +v_{31} \cdot \frac{0.275}{0.4} - v_{32} \\
\text{d[SOCS3]}/\text{d}t &= +v_{33} - v_{34}.
\end{aligned}
$$

The volume factors vol_cyt = 0.4 pl and vol_nuc = 0.275 pl account for transitions between different compartments and are determined experimentally. The species npSTAT5, CISnRNA1–5 and SOCS3nRNA1–5 are located in the nuclear compartment, the remaining species in the cytoplasmatic compartment.

The initial condition are set to zero except for

$$\begin{aligned} [\text{EpoRJAK2}](0) &= \text{init\_EpoRJAK2} \\ [\text{SHP1}](0) &= \text{init\_SHP1} \\ [\text{STAT5}](0) &= \text{init\_STAT5}. \end{aligned}$$

The maximum likelihood estimate of the parameters can be found in the supplementary material of Bachmann *et al.* [2011].

A. ORDINARY DIFFERENTIAL EQUATIONS FOR THE PRESENTED EXAMPLES

References

Aaronson, D. & Horvath, C. (2002). A Road map for those who don't know JAK-STAT. *Science*, **296**, 1653 – 1655. 15, 154

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B.N. Petrov & F. Csaki, eds., *Second International Symposium on Information Theory*, 267–281, Akademiai Kiado, Budapest, Hungary. 75

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723. 75

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. & Walter, P. (2002). *Molecular Biology of the Cell*, vol. 4. Garland Science. 14, 136

Aldridge, B.B., Burke, J.M., Lauffenburger, D.A. & Sorger, P.K. (2006). Physicochemical modelling of cell signalling pathways. *Nature cell biology*, **8**, 1195–1203. 3

Aldridge, B.B., Gaudet, S., Lauffenburger, D.A. & Sorger, P.K. (2011). Lyapunov exponents and phase diagrams reveal multi-factorial control over trail-induced apoptosis. *Molecular Systems Biology*, **7**. 3

Artavanis-Tsakonas, S., Rand, M.D. & Lake, R.J. (1999). Notch signaling: cell fate control and signal integration in development. *Science*, **284**, 770–776. 15

Atkinson, K.E. & Han, W. (1985). *Elementary numerical analysis*. Wiley New York. 30

Audoly, S., D'Angio, L., Saccomani, M. & Cobelli, C. (1998). Global identifiability of linear compartmental models - a computer algebra algorithm. *IEEE Transactions on Biomedical Engineering*, **45**, 36–47. 51

Audoly, S., Bellu, G., D'Angio, L., Saccomani, M.P. & Cobelli, C. (2001). Global identifiability of nonlinear models of biological systems. *IEEE Transactions on Biomedical Engineering*, **48**, 55–65. 51

Bachmann, J., Raue, A., Schilling, M., Böhm, M., Kreutz, C., Kaschek, D., Busch, H., Gretz, N., Lehmann, W., Timmer, J. & Klingmüller, U. (2011). Division of

labor by dual feedback regulators controls JAK2/STAT5 signaling over broad ligand range. *Molecular Systems Biology*, **7**. 15, 16, 154, 155, 156, 162, 169, 181

Baker, C.T., Paul, C.A. & Willé, D. (1995). Issues in the numerical solution of evolutionary delay differential equations. *Advances in Computational Mathematics*, **3**, 171–196. 35

Battogtokh, D., Asch, D., Case, M., Arnold, J. & Schüttler, H.B. (2002). An ensemble method for identifying regulatory circuits with special reference to the qa gene cluster of Neurospora crassa. *Proceedings of the National Academy of Sciences*, **99**, 16904–16909. 4

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions*, **53**, 370–418. 4

Becker, A.J., McCulloch, E.A. & Till, J.E. (1963). Cytological demonstration of the clonal nature of spleen colonies derived from transplanted mouse marrow cells. *Nature*. 17

Becker, V., Schilling, M., Bachmann, J., Baumann, U., Raue, A., Maiwald, T., Timmer, J. & Klingmüller, U. (2010). Covering a broad dynamic range: information processing at the erythropoietin receptor. *Science*, **328**, 1404–1408. 17, 154

Behrens, G., Friel, N. & Hurn, M. (2012). Tuning tempered transitions. *Statistics and Computing*, **22**, 65–78. 130

Beichl, I. & Sullivan, F. (2000). The Metropolis algorithm. *Computing in Science & Engineering*, **2**, 65–69. 56

Box, G.E. & Draper, N.R. (1987). *Empirical model-building and response surfaces*. John Wiley & Sons. 2

Box, G.E. & Tiao, G.C. (2011). *Bayesian inference in statistical analysis*, vol. 40. John Wiley & Sons. 45

Brooks, S. & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 434–455. 69

Brooks, S.P. & Roberts, G.O. (1998). Convergence assessment techniques for Markov chain Monte Carlo. *Statistics and Computing*, **8**, 319–335. 68, 169

Brown, K. & Sethna, J. (2003). Statistical mechanical approaches to models with many poorly known parameters. *Physical Review E*, **68**, 021904. 4

Byrd, R.H., Hribar, M.E. & Nocedal, J. (1999). An interior point algorithm for large-scale nonlinear programming. *SIAM Journal on Optimization*, **9**, 877–900. 42

Byrd, R.H., Gilbert, J.C. & Nocedal, J. (2000). A trust region method based on interior point techniques for nonlinear programming. *Mathematical Programming*, **89**, 149–185. 42

Calderhead, B. & Girolami, M. (2009). Estimating Bayes factors via thermodynamic integration and population MCMC. *Computational Statistics & Data Analysis*, **53**, 4028–4045. 80, 83, 86, 91, 96, 97, 98, 170

Chehreghani, M.H., Busetto, A.G. & Buhmann, J.M. (2012). Information Theoretic Model Validation for Spectral Clustering. **XX**. 171

Chen, K., Calzone, L., Csikasz-Nagy, A., Cross, F., Novak, B. & Tyson, J. (2004). Integrative analysis of cell cycle control in budding yeast. *Molecular Biology of the Cell*, **15**, 3841–3862. 35

Chib, S. & Jeliazkov, I. (2001). Marginal likelihood from the Metropolis–Hastings output. *Journal of the American Statistical Association*, **96**, 270–281. 79

Chis, O.T., Banga, J. & Balsa-Canto, E. (2011). Structural identifiability of systems biology models: A critical comparison of methods. *PLoS ONE*, **6**, e27755. 51

Choudhuri, S. (2003). The path from nuclein to human genome: A brief history of DNA with a note on human genome sequencing and its impact on future research in biology. *Bulletin of Science, Technology & Society*, **23**, 360–367. 14

Cobelli, C. & DiStefano, J. (1980). Parameter and structural identifiability concepts and ambiguities: a critical review and analysis. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, **239**, R7–R24. 5

Coleman, T.F. & Li, Y. (1996). An interior trust region approach for nonlinear minimization subject to bounds. *SIAM Journal on optimization*, **6**, 418–445. 42

Collins, F.S., Morgan, M. & Patrinos, A. (2003). The human genome project: lessons from large-scale biology. *Science*, **300**, 286–290. 14

Cowles, M.K. & Carlin, B.P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, **91**, 883–904. 68

Crick, F. (1970). Central dogma of molecular biology. *Nature*, **227**, 561–563. 14

Davison, A.C. & Hinkley, D.V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press, Cambridge. 121

Eden, E., Geva-Zatorsky, N., Issaeva, I., Cohen, A., Dekel, E., Danon, T., Cohen, L., Mayo, A. & Alon, U. (2011). Proteome half-life dynamics in living human cells. *Science*, **331**, 764–768. 103, 131

Efron, B. & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York. 42

Egea, J.A., Rodríguez-Fernández, M., Banga, J.R. & Martí, R. (2007). Scatter search for chemical and bio-process optimization. *Journal of Global Optimization*, **37**, 481–503. 42

REFERENCES

Engl, H.W., Flamm, C., Kügler, P., Lu, J., Müller, S. & Schuster, P. (2009). Inverse problems in systems biology. *Inverse Problems*, **25**, 123014. 4

Eydgahi, H., Chen, W.W., Muhlich, J.L., Vitkup, D., Tsitsiklis, J.N. & Sorger, P.K. (2013). Properties of cell death models calibrated and compared using Bayesian approaches. *Molecular Systems Biology*, **9**. 80, 153

Foster, S.D., Oram, S.H., Wilson, N.K. & Göttgens, B. (2009). From genes to cells to tissues - modelling the haematopoietic system. *Molecular Biosystems*, **5**, 1413–1420. 19

Fraser, A. & Burnell, D. (1970). *Computer models in genetics*. McGraw-Hill. 42

Friel, N. & Pettitt, A.N. (2008). Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**, 589–607. 80, 83

Friel, N., Hurn, M. & Wyse, J. (2013). Improving power posterior estimation of statistical evidence. *Statistics and Computing*, 1–15. 83

Frixione, E. (2000). Recurring views on the structure and function of the cytoskeleton: a 300-year epic. *Cell motility and the Cytoskeleton*, **46**, 73–94. 14

Gelman, A. & Meng, X. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, 163–185. 80

Gelman, A. & Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 457–472. 59, 69

Gelman, A., Bois, F. & Jiang, J. (1996a). Physiological pharmacokinetic analysis using population modeling and informative prior distributions. *Journal of the American Statistical Association*, **91**, 1400–1412. 38, 170

Gelman, A., Roberts, G. & Gilks, W. (1996b). Efficient Metropolis jumping rules. *Bayesian Statistics*, **5**, 599–608. 57, 59

Geweke, J. (1992). *Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments*. Oxford University Press. 69, 125, 128

Geyer, C. (1992). Practical Markov chain Monte Carlo. *Statistical Science*, **7**, 473–483. 68, 70

Gillespie, D. (1992). A rigorous derivation of the chemical master equation. *Physica A: Statistical Mechanics and its Applications*, **188**, 404–425. 32

Girolami, M. (2008). Bayesian inference for differential equations. *Theoretical Computer Science*, **408**, 4–16. 32

Goodwin, B. (1963). *Temporal organization in cells*. Academic Press New York. 35

Graf, T. & Enver, T. (2009). Forcing cells to change lineages. *Nature*, **462**, 587–594. 17

Greiter, M., Giussani, A., Höllriegl, V., Li, W. & Oeh, U. (2011a). Human biokinetic data and a new compartmental model of zirconium – a tracer study with enriched stable isotopes. *Science of the Total Environment*, **409**, 3701–3710. 39, 134, 135, 136, 173

Greiter, M., Höllriegl, V. & Oeh, U. (2011b). Method development for thermal ionization mass spectrometry in the frame of a biokinetic tracer study with enriched stable isotopes of zirconium. *International Journal of Mass Spectrometry*, **304**, 1–8. 39, 136

Grimmett, G. & Stirzaker, D. (2001). *Probability and Random Processes*. Probability and Random Processes, OUP Oxford. 20

Guldberg, C. & Waage, P. (1899). Experiments concerning chemical affinity. *German translation by Abegg in Ostwald's Klassiker der Exacten Wissenschaften*, 10–125. 33

Gutenkunst, R.N., Waterfall, J.J., Casey, F.P., Brown, K.S., Myers, C.R. & Sethna, J.P. (2007). Universally sloppy parameter sensitivities in systems biology models. *PLoS Computational Biology*, **3**, e189. 5

Haario, H., Saksman, E. & Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli*, 223–242. 58, 59, 65

Halter, M., Tona, A., Bhadriraju, K., Plant, A.L. & Elliott, J.T. (2007). Automated live cell imaging of green fluorescent protein degradation in individual fibroblasts. *Cytometry Part A*, **71**, 827–834. 19, 104, 106, 108, 115

Harper, C.V., Finkenstädt, B., Woodcock, D.J., Friedrichsen, S., Semprini, S., Ashall, L., Spiller, D.G., Mullins, J.J., Rand, D.A., Davis, J.R. & White, M.R. (2011). Dynamic analysis of stochastic transcription cycles. *PLoS Biology*, **9**, e1000607. 107

Hartwell, L.H. & Weinert, T.A. (1989). Checkpoints: controls that ensure the order of cell cycle events. *Science*, **246**, 629–634. 14

Hasenauer, J. (2014). A MATLAB package for the efficient calculation of profile likelihoods and profile posteriors. 51

Hasenauer, J., Waldherr, S., Doszczak, M., Radde, N., Scheurich, P. & Allgöwer, F. (2011). Identification of models of heterogeneous cell populations from population snapshot data. *BMC bioinformatics*, **12**, 125. 131

Hasenauer, J., Hasenauer, C., Hucho, T. & Theis, F.J. (2014). ODE constrained mixture modelling: A method for unraveling subpopulation structures and dynamics. *In revision*. 131, 170

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109. 4, 56

REFERENCES

Hodgkin, A.L. & Huxley, A.F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of physiology*, **117**, 500. 2

Hoppe, P.S., Schwarzfischer, M., Loeffler, D., Kokkaliaris, K.D., Hilsenbeck, O., Moritz, N., Endele, M., Filipczyk, A., Rieger, M.A., Marr, C., Strasser, M., Schauberger, B., Burtscher, I., Ermakova, O., Buerger, A., Lickert, H., Nerlov, C., Theis, F.J. & Schroeder, T. (2014). Random PU.1 / Gata1 protein ratios do not induce early myeloid lineage choice. *submitted*. 105, 119, 130

Horbelt, W., Timmer, J. & Voss, H. (2002). Parameter estimation in nonlinear delayed feedback systems from noisy data. *Physics Letters A*, **299**, 513–521. 4

Hug, S. & Theis, F.J. (2012). Bayesian inference of latent causes in gene regulatory dynamics. In *Latent Variable Analysis and Signal Separation*, 520–527, Springer.

Hug, S., Raue, A., Hasenauer, J., Bachmann, J., Klingmüller, U., Timmer, J. & Theis, F.J. (2013). High-dimensional Bayesian parameter estimation: Case study for a model of JAK2/STAT5 signaling. *Mathematical Biosciences*.

Hug, S., Schmidl, D., Li, W., Greiter, M.B. & Theis, F.J. (2014a). Uncertainty in Biology: a computational modeling approach, chapter Bayesian model selection methods and their application to biological ODE systems. *in revision*.

Hug, S., Schwarzfischer, M., Hasenauer, J., Marr, C. & Theis, F.J. (2014b). An adaptive method for calculating Bayes factors using Simpson's rule. *in revision*.

Hukushima, K. & Nemoto, K. (1996). Exchange Monte Carlo method and application to spin glass simulations. *Journal of the Physical Society of Japan*, **65**, 1604–1608. 61

ICRP (1975). Report on the task group on reference man. ICRP publication 23. *Annals of the ICRP*. 135

ICRP (1989). *Age-dependent Doses to Members of the Public from Intake of Radionuclides (Part 1). ICRP Publication 56*. Annals of the ICRP **20**(2), Pergamon Press. 134, 135, 175

ICRP (1993). *Age-dependent Doses to Members of the Public from Intake of Radionuclides (Part 2: Ingestion dose coefficients). ICRP Publication 67*. Annals of the ICRP **23**(3–4), Pergamon Press. 135, 175

Igaz, P., Toth, S. & Falus, A. (2001). Biological and clinical significance of the JAK-STAT pathway; lessons from knockout mice. *Inflammation Research*, **50**, 435–441. 15

Iman, R.L. (2008). *Latin hypercube sampling*. Wiley Online Library. 42

ISO (1995). Guide to the expression of uncertainty in measurement. Tech. rep. 139

Jacquez, J. (1985). *Compartmental analysis in biology and medicine*. University of Michigan Press Ann Arbor, MI. 36, 134

Jasra, A., Stephens, D.A. & Holmes, C.C. (2007). Population-based reversible jump Markov chain Monte Carlo. *Biometrika*, **94**, 787–807. 61

Jeffreys, H. (1961). *Theory of probability*. Clarendon Press. 77

Kass, R. & Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, 773–795. 76

Kendall, M. & Stuart, A. (1979). *The Advanced Theory of Statistics*, vol. 2. Charles Griffin & Company, London, 4th edn. 43

Kholodenko, B.N. (2006). Cell-signalling dynamics in time and space. *Nature reviews Molecular cell biology*, **7**, 165–176. 32

Kim, M.S., Pinto, S.M., Getnet, D., Nirujogi, R.S., Manda, S.S., Chaerkady, R., Madugundu, A.K., Kelkar, D.S., Isserlin, R., Jain, S., Thomas, J.K., Muthusamy, B., Leal-Rojas, P., Kumar, P., Sahasrabuddhe, N.a., Balakrishnan, L., Advani, J., George, B., Renuse, S., Selvan, L.D.N., Patil, A.H., Nanjappa, V., Radhakrishnan, A., Prasad, S., Subbannayya, T., Raju, R., Kumar, M., Sreenivasamurthy, S.K., Marimuthu, A., Sathe, G.J., Chavan, S., Datta, K.K., Subbannayya, Y., Sahu, A., Yelamanchi, S.D., Jayaram, S., Rajagopalan, P., Sharma, J., Murthy, K.R., Syed, N., Goel, R., Khan, A.a., Ahmad, S., Dey, G., Mudgal, K., Chatterjee, A., Huang, T.C., Zhong, J., Wu, X., Shaw, P.G., Freed, D., Zahari, M.S., Mukherjee, K.K., Shankar, S., Mahadevan, A., Lam, H., Mitchell, C.J., Shankar, S.K., Satishchandra, P., Schroeder, J.T., Sirdeshmukh, R., Maitra, A., Leach, S.D., Drake, C.G., Halushka, M.K., Prasad, T.S.K., Hruban, R.H., Kerr, C.L., Bader, G.D., Iacobuzio-Donahue, C.a., Gowda, H. & Pandey, A. (2014). A draft map of the human proteome. *Nature*, **509**, 575–581. 14

Kirk, P., Thorne, T. & Stumpf, M.P. (2013). Model selection in systems and synthetic biology. *Current Opinion in Biotechnology*, 1–8. 3, 75, 76

Kirkpatrick, S., Gelatt, C.D. & Vecchi, M.P. (1983). Optimization by simulated annealing. *Science*, **220**, 671–680. 42, 137

Kirstetter, P., Anderson, K., Porse, B.T., Jacobsen, S.E.W. & Nerlov, C. (2006). Activation of the canonical wnt pathway leads to loss of hematopoietic stem cell repopulation and multilineage differentiation block. *Nature immunology*, **7**, 1048–1056. 104

Kitano, H. (2002a). Computational systems biology. *Nature*, **420**, 206–210. 2

Kitano, H. (2002b). Systems biology: a brief overview. *Science*, **295**, 1662–1664. 2, 3

Klipp, E., Herwig, R., Kowald, A., Wierling, C. & Lehrach, H. (2008). *Systems biology in practice: concepts, implementation and application*. John Wiley & Sons. 115

REFERENCES

Kollmann, M., Løvdok, L., Bartholomé, K., Timmer, J. & Sourjik, V. (2005). Design principles of a bacterial signalling network. *Nature*, **438**, 504–507. 3

Komorowski, M., Costa, M.J., Rand, D.A. & Stumpf, M.P. (2011). Sensitivity, robustness, and identifiability in stochastic chemical kinetics models. *Proceedings of the National Academy of Sciences*, **108**, 8645–8650. 5

Kowarsch, A. (2011). *The impact of microRNAs on signaling pathways: From general perspectives to a computational model of the JAK-STAT pathway*. Ph.D. thesis, Technische Universität München, Germany. 15

Kreutz, C., Rodriguez, M.M.B., Maiwald, T., Seidl, M., Blum, H.E., Mohr, L. & Timmer, J. (2007). An error model for protein quantification. *Bioinformatics*, **23**, 2747–2753. 156

Kreutz, C., Raue, A. & Timmer, J. (2012). Likelihood based observability analysis and confidence intervals for predictions of dynamic models. *BMC Systems Biology*, **6**, 120. 51

Kurowicka, D. & Joe, H. (2011). *Dependence Modeling - Handbook on Vine Copulae*. World Scientific Publishing Co. 66

Lartillot, N. & Philippe, H. (2006). Computing Bayes factors using thermodynamic integration. *Systematic Biology*, **55**, 195–207. 80

Latchman, D.S. (1997). Transcription factors: an overview. *The International Journal of Biochemistry & Cell Biology*, **29**, 1305–1312. 14

Lawrence, N., Girolami, M., Rattray, M. & Sanguinetti, G. (2010). *Learning and Inference in Computational Systems Biology*. The MIT Press. 2, 3, 4

Levin, D.A., Peres, Y. & Wilmer, E.L. (2009). *Markov chains and mixing times*. American Mathematical Society. 30

Li, S., Brazhnik, P., Sobral, B. & Tyson, J. (2008). A quantitative study of the division cycle of Caulobacter crescentus stalked cells. *PLoS Computational Biology*, **4**, e9. 35

Li, W., Greiter, M., Oeh, U. & Hoeschen, C. (2011a). Reliability of a new biokinetic model of zirconium in internal dosimetry. Part I, Parameter uncertainty analysis. *Health Physics*, **101**, 660–676. 135, 139, 140

Li, W., Greiter, M., Oeh, U. & Hoeschen, C. (2011b). Reliability of a new biokinetic model of zirconium in internal dosimetry. Part II, Parameter sensitivity analysis. *Health Physics*, **101**, 676–692. 135, 141

Little, M.P., Heidenreich, W.F. & Li, G. (2010). Parameter identifiability and redundancy: theoretical considerations. *PloS ONE*, **5**, e8915. 5, 51

Ljung, L. (1999). *System Identification - Theory For the User, 2nd ed*. PTR Prentice Hall, Upper Saddle River, N.J. 3

Lodish, H., Berk, A., Kaiser, C.A., Krieger, M., Bretscher, A., Ploegh, H., Amon, A. & Scott, M.P. (2012). *Molecular cell biology*. W. H. Freeman, 7th edn. 14, 15

Lyness, J.N. (1969). Notes on the adaptive Simpson quadrature routine. *Journal of the ACM*, **16**, 483–495. 85

MacDonald, N. (1976). Time delay in simple chemostat models. *Biotechnology and Bioengineering*, **18**, 805–812. 180

Maiwald, T. & Timmer, J. (2008). Dynamical modeling and multi-experiment fitting with potterswheel. *Bioinformatics*, **24**, 2037–2043. 131

Marin, J. & Robert, C. (2007). *Bayesian core: a practical approach to computational Bayesian statistics*. Springer. 48, 49, 50

McNaught, A.D. & Wilkinson, A. (2000). IUPAC compendium of chemical terminology. 32

Mengersen, K.L., Robert, C.P. & Guihenneuc-Jouyaux, C. (1999). MCMC convergence diagnostics: a "reviewww". *Bayesian Statistics*, **6**, 415–440. 68

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, **21**, 1087–1092. 4, 56

Meyn, S. & Tweedie, R. (1996). *Markov chains and stochastic stability*. Springer. 25

Min, A. & Czado, C. (2010). Bayesian inference for multivariate copulas using pair-copula constructions. *Journal of Financial Econometrics*, **8**, 511–546. 65, 67

Myung, I. & Pitt, M. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, **4**, 79–95. 78

Neal, R. (1993). Probabilistic inference using Markov chain Monte Carlo methods. Tech. Rep. CRG-TR-93-1, University of Toronto. Department of Computer Science. 139

Neal, R. (1996). Sampling from multimodal distributions using tempered transitions. *Statistics and Computing*, **6**, 353–366. 61

Neal, R. (2008). The Harmonic Mean of the Likelihood: Worst Monte Carlo Method Ever, `http://radfordneal.wordpress.com/2008/08/17/the-harmonic-mean-of-the-likelihood-worst-monte-carlo-method-ever`. 79

Nelsen, R. (2006). *An Introduction to Copulas*. Springer. 66

REFERENCES

Nerlov, C., Querfurth, E., Kulessa, H. & Graf, T. (2000). GATA-1 interacts with the myeloid PU. 1 transcription factor and represses PU. 1-dependent transcription. *Blood*, **95**, 2543–2551. 19

Newton, M. & Raftery, A. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **56**, 3–48. 79

Nummelin, E. (2004). *General irreducible Markov chains and non-negative operators*, vol. 83. Cambridge University Press. 29

Nutt, S.L., Metcalf, D., D'Amico, A., Polli, M. & Wu, L. (2005). Dynamic regulation of PU.1 expression in multipotent hematopoietic progenitors. *The Journal of experimental medicine*, **201**, 221–231. 115

O'Hagan, A., Forster, J. & Kendall, M.G. (2004). *Bayesian inference*. Arnold London. 45

Olby, R.C. (1974). *The path to the double helix: the discovery of DNA*. Courier Dover Publications. 14

Orkin, S.H. & Zon, L.I. (2008). Hematopoiesis: an evolving paradigm for stem cell biology. *Cell*, **132**, 631–644. 17

Pearson, H. (2006). Genetics: what is a gene? *Nature*, **441**, 398–401. 14

Raia, V., Schilling, M., Böhm, M., Hahn, B., Kowarsch, A., Raue, A., Sticht, C., Bohl, S., Saile, M., Möller, P., Gretz, N., Timmer, J., Theis, F., Lehmann, W.D., Lichter, P. & U., K. (2011). Dynamic mathematical modeling of IL13-induced signaling in Hodgkin and primary mediastinal B-cell lymphoma allows prediction of therapeutic targets. *Cancer Research*, **71**, 693–704. 35

Raue, A. (2013). *Quantitative dynamic modeling Theory and Application to Signal Transduction in the Erythropoietic System*. Ph.D. thesis, Albert-Ludwigs-Universität Freiburg, Germany. 9, 52, 154

Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U. & Timmer, J. (2009). Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, **25**, 1923–1929. 5, 51, 157

Raue, A., Kreutz, C., Theis, F.J. & Timmer, J. (2013a). Joining forces of Bayesian and frequentist methodology: a study for inference in the presence of non-identifiability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **371**, 20110544. 50, 52, 164, 168

Raue, A., Schilling, M., Bachmann, J., Matteson, A., Schelke, M., Kaschek, D., Hug, S., Kreutz, C., Harms, B.D., Theis, F.J., Klingmüller, U. & Timmer, J. (2013b). Lessons learned from quantitative dynamical modeling in systems biology. *PLoS ONE*, **8**, e74335. 42, 51, 75, 154, 156

Revuz, D. (1984). *Markov chains*, vol. 11. North Holland. 29

Rieger, M. & Schroeder, T. (2007). Hämatopoetische Stammzellen. *BIOspektrum-Heidelberg*, **13**, 254. 17

Rigat, F. & Mira, A. (2012). Parallel hierarchical sampling: A general-purpose interacting Markov chains Monte Carlo algorithm. *Computational Statistics & Data Analysis*, **56**, 1450 – 1467. 56, 61, 159, 167

Robert, C. & Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer. 25, 26, 28, 46, 56, 57

Robert, C.P. (2001). The Bayesian choice: From decision-theoretic foundations to computational implementation. 50

Roberts, G., Gelman, A. & Gilks, W. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, **7**, 110–120. 56, 57

Rodriguez-Fernandez, M., Egea, J.A. & Banga, J.R. (2006). Novel metaheuristic for parameter estimation in nonlinear dynamic biological systems. *BMC bioinformatics*, **7**, 483. 42

Rossi, L., Lin, K.K., Boles, N.C., Yang, L., King, K.Y., Jeong, M., Mayle, A. & Goodell, M.A. (2012). Less is more: Unveiling the functional core of hematopoietic stem cells through knockout mice. *Cell Stem Cell*, **11**, 302 – 317. 18

Salvadori, G. (2007). *Extremes in nature: an approach using copulas*. Springer, New York. 65

Sanguinetti, G., Lawrence, N.D. & Rattray, M. (2006). Probabilistic inference of transcription factor concentrations and gene-specific regulatory activities. *Bioinformatics*, **22**, 2775–2781. 4

Schmidl, D. (2012). *Bayesian model inference in dynamic biological systems using Markov Chain Monte Carlo methods*. Ph.D. thesis, Technische Universität München, Germany. 8, 29, 30, 65, 79, 86, 90, 134

Schmidl, D., Hug, S., Li, W., Greiter, M.B. & Theis, F.J. (2012). Bayesian model selection validates a biokinetic model for zirconium processing in humans. *BMC Systems Biology*, **6**.

Schmidl, D., Czado, C., Hug, S. & Theis, F. (2013a). A Vine-copula Based Adaptive MCMC Sampler for Efficient Inference of Dynamical Systems. *Bayesian Analysis*, **8**, 1–22. 65, 67, 139

Schmidl, D., Czado, C., Hug, S. & Theis, F. (2013b). Rejoinder for: A Vine-copula Based Adaptive MCMC Sampler for Efficient Inference of Dynamical Systems. *Bayesian Analysis*, **8**, 33–42. 65

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W. & Selbach, M. (2011). Global quantification of mammalian gene expression control. *Nature*, **473**, 337–342. 103, 131

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*. 75

Schwarzfischer, M. (2013). *Quantification and analysis of single-cell protein dynamics in stem cells using time-lapse microscopy*. Ph.D. thesis, Technische Universität München, Germany. 18, 21, 35

Schwarzfischer, M., Hilsenbeck, O., Schauberger, B., Hug, S., Filipczyk, A., Hoppe, P., Strasser, M., Buggenthin, F., Feigelman, J., Krumsiek, J., Loeffler, D., Kokkaliaris, K., van den Berg, A., Endele, M., Hastreiter, S., Marr, C., Theis, F. & Schroeder, T. (2014). Single-cell quantification of cellular and molecular behavior in long-term time-lapse microscopy. *in preparation*. 19, 21, 104, 115, 119, 121, 130

Scott, E.W., Simon, M.C., Anastasi, J. & Singh, H. (1994). Requirement of transcription factor PU.1 in the development of multiple hematopoietic lineages. *Science*, **265**, 1573–1577. 104

Serban, R. & Hindmarsh, A. (2005). CVODES: the Sensitivity-Enabled ODE Solver in SUNDIALS. In *Proceedings of IDETC/CIE*, vol. 24. 36, 156

Sethuraman, J., Athreya, K. & Ddoss, H. (1992). A Proof of convergence of the Markov Chain simulation Method. Tech. rep. 29

Shampine, L. & Reichelt, M. (1997). The MATLAB ODE suite. *SIAM Journal on Scientific Computing*, **18**, 1–22. 36

Shampine, L. & Thompson, S. (2001). Solving DDEs in MATLAB. *Applied Numerical Mathematics*, **37**, 441–458. 35, 36

Shargel, L., Andrew, B. & Wu-Pong, S. (2005). *Applied biopharmaceutics & pharmacokinetics*. Appleton & Lange Reviews/McGraw-Hill, Medical Pub. Division. 38, 170

Shimomura, O., Johnson, F.H. & Saiga, Y. (1962). Extraction, purification and properties of aequorin, a bioluminescent protein from the luminous hydromedusan, Aequorea. *Journal of cellular and comparative physiology*, **59**, 223–239. 19

Smith, H.L. (2011). *An introduction to delay differential equations with applications to the life sciences*. Springer. 35, 36

Stute, W., Manteiga, W.G. & Quindimil, M.P. (1993). Bootstrap based goodness-of-fit-tests. *Metrika*, **40**, 243–256. 42

Swameye, I., Müller, T., Timmer, J.t., Sandra, O. & Klingmüller, U. (2003). Identification of nucleocytoplasmic cycling as a remote sensor in cellular signaling by databased modeling. *Proceedings of the National Academy of Sciences*, **100**, 1028–1033. 3, 15, 154

Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics*, 1701–1728. 28, 29

Turing, A. (1952). The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society Series B*, **237**, 37–42. 2

Vanlier, J., Tiemann, C., Hilbers, P. & van Riel, N. (2012). An integrated strategy for prediction uncertainty analysis. *Bioinformatics*. 50, 164, 168

Vapnik, V. (1995). *The nature of statistical learning theory*. Springer. 158

Vehlow, C., Hasenauer, J., Kramer, A., Raue, A., Hug, S., Timmer, J., Radde, N., Theis, F.J. & Weiskopf, D. (2013). iVUN: interactive visualization of uncertain biochemical reaction networks. *BMC Bioinformatics*, **14**, S2.

Villaverde, A.F. & Banga, J.R. (2014). Reverse engineering and identification in systems biology: strategies, perspectives and challenges. *Journal of The Royal Society Interface*, **11**. 4

von Davier, M. (1997). Bootstrapping goodness-of-fit statistics for sparse categorical data: Results of a Monte Carlo study. *Methods of Psychological Research*, **2**, 29–48. 42

von Foerster, H. (1959). Some remarks on changing populations. *The kinetics of cellular proliferation*, **382**. 2

Welch, B.L. (1947). The generalization of 'student's' problem when several different population variances are involved. *Biometrika*, 28–35. 2

Wiener, N. (1948). *Cybernetics: Control and communication in the animal and the machine*. Wiley New York. 2

Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A.M., Lieberenz, M., Savitski, M.M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H., Mathieson, T., Lemeer, S., Schnatbaum, K., Reimer, U., Wenschuh, H., Mollenhauer, M., Slotta-Huspenina, J., Boese, J.H., Bantscheff, M., Gerstmair, A., Faerber, F. & Kuster, B. (2014). Mass-spectrometry-based draft of the human proteome. *Nature*, **509**, 582–587. 14

REFERENCES

Wilkinson, D. (2006). *Stochastic Modelling for Systems Biology*. Chapman & Hall/CRC. 4

Woodard, D.B., Schmidler, S.C. & Huber, M. (2009). Conditions for rapid mixing of parallel and simulated tempering on multimodal distributions. *The Annals of Applied Probability*, 617–640. 61

Woodcock, D.J., Vance, K.W., Komorowski, M., Koentges, G., Finkenstädt, B. & Rand, D.A. (2013). A hierarchical model of transcriptional dynamics allows robust estimation of transcription rates in populations of single cells with variable gene copy number. *Bioinformatics*, **29**, 1519–25. 170

Xu, T.R., Vyshemirsky, V., Gormand, A., von Kriegsheim, A., Girolami, M., Baillie, G.S., Ketley, D., Dunlop, A.J., Milligan, G., Houslay, M.D. & Kolch, W. (2010). Inferring signaling pathway topologies from multiple perturbation measurements of specific biochemical species. *Science Signaling*, **3**, ra20. 80

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika*, **92**, 937–950. 75

Zechner, C., Ruess, J., Krenn, P., Pelet, S., Peter, M., Lygeros, J. & Koeppl, H. (2012). Moment-based inference predicts bimodality in transient gene expression. *Proceedings of the National Academy of Sciences*, **109**, 8340–8345. 131

Zhang, P., Behre, G., Pan, J., Iwama, A., Wara-Aswapati, N., Radomska, H.S., Auron, P.E., Tenen, D.G. & Sun, Z. (1999). Negative cross-talk between hematopoietic regulators: GATA proteins repress PU. 1. *Proceedings of the National Academy of Sciences*, **96**, 8705–8710. 19

Zhang, P., Zhang, X., Iwama, A., Yu, C., Smith, K.A., Mueller, B.U., Narravula, S., Torbett, B.E., Orkin, S.H. & Tenen, D.G. (2000). PU. 1 inhibits GATA-1 function and erythroid differentiation by blocking GATA-1 DNA binding. *Blood*, **96**, 2641–2648. 19