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

TECHNISCHE UNIVERSITÄT MÜNCHEN

Bayesian model inference in dynamic biological systems using Markov Chain Monte Carlo methods

Daniel Schmidl

February 2012

TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl M12 (Biomathematik)

"Bayesian model inference in dynamic biological systems using Markov Chain Monte Carlo methods"

Daniel Schmidl

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.

Vorsitzender:

Univ.-Prof. Dr. Oliver Junge

Prüfer der Dissertation:

1. Univ.-Prof. Dr. Dr. Fabian J. Theis
2. Univ.-Prof. Claudia Czado, Ph.D.
3. Univ.-Prof. Dr. Achim Tresch, Universität zu Köln (nur schriftliche Beurteilung)

Die Dissertation wurde am 14.03.2012 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 21.08.2012 angenommen.

To my parents.

Acknowledgments

Working on Monte Carlo methods I was able to gain quite some expertise in the field of educated guessing. This is why I can almost surely say that for a few people these are the most interesting pages of the entire work. Here is where you find out who incited me to write this thesis and who supported, pushed and accompanied me over the last few years. I am deeply grateful that all of them made this a very valuable time of my life.

The work in this thesis has been supervised by Prof. Dr. Dr. Fabian Theis. First and foremost, I would like to thank him for putting trust in me and giving me the opportunity to work almost independently on the various projects. Not only was he able to push me scientifically, but also in various running competitions. I must admit that there might still be some potential left on my side. Thanks for the chance to take part in building up an extraordinary interdisciplinary group on computational modeling in biology.

I am also very grateful to Prof. Dr. Claudia Czado, who never tired in explaining details on copula distributions and (grapeless) vines. It was a pleasure to work with her.

Thank you Prof. Dr. Achim Tresch for digging yourself through this thesis and evaluation this work.

The entire CMB group: Thanks for many valuable discussions, but moreover for putting fun into science. It was really nice to work with every single one of you. Thank you Jan for almost reinventing computer science in order to solve some of my computer problems and Dom and Florian for teaching me basics on biology and cows. A special thank to Dominik for always taking time to discuss mathematical problems and for making our conference trips a fun time. Andreas, thank you for teaching me how to get things done in a quick and diligent way. Our time in Boston and New York was legendary! I want to express my gratitude and appreciation to Sabine for putting a lot of effort into the zirconium processing project. It was really nice to work with you! Moreover, thank you for organizing an endless number of group

events, for carefully proof-reading this thesis and for introducing me to pink colored corrections.

Eike and Ulf, thank you for your support on vine estimation. I am grateful to the whole statistics group at the Technische Universität München. Teaching with you taught me a lot as well.

Andi, Cathi, Kerstin, Markus, and Tom, my dear companions, I wish anyone could have friends like you! Knowing you is without exaggeration one of the greatest gifts in life and I am truly looking forward to all the fun times to come.

Nina, talking to you always makes my day! Your eagerness along with your open-hearted nature is unmatched. Ironically, we still have to work on your understanding of the word moaning. Thank you for always having an open ear. I can say you are in many ways a pure source of motivation.

Richard, I really enjoyed going through PhD with you, showing me that the grass is not greener on the other side. Thank you for always listening carefully to a friend's needs! Your jolly company is true inspiration.

Finally, a very special thank goes to my parents! Although not exactly familiar with my scientific way of life, their lifelong encouragement and unconditional support always kept me on track and taught me what being true to yourself really means. Words cannot convey my gratitude!

Abstract

Dynamical systems are a valuable tool for exploring the regulatory organization of living organisms on a molecular level. Here, profound knowledge about underlying reaction rate parameters is essential for biological predictions and hypotheses. Bayesian methods provide a sophisticated approach to infer these parameters including statistical dependencies and uncertainties. Moreover, they are able to determine an appropriate model structure by model selection. As the posterior distributions involved are generally analytically intractable, Markov Chain Monte Carlo (MCMC) methods are used for inference. However, owing to posterior complexity, these methods are often inefficiently sampling from the according parameter spaces. In this thesis we develop a Copula based Independence/random walk Metropolis-Hastings (CIMH) sampling scheme for efficient model inference of differential equation based dynamical systems. The concept exploits a vine copula decomposition of the estimated posterior distribution in order to generate proposals that are distributed according to an approximation of the true posterior distribution, which yields high acceptance rates. The basic CIMH algorithm is furthermore extended to an Adaptive MCMC scheme (ACIMH) to speed up convergence in complex systems. We thoroughly compare CIMH and ACIMH to existing methods on various examples. Furthermore, in applications from the field of systems biology, we examine the mechanism of nuclear phosphorylated STAT3 dimer import in the JAK1-STAT3 signaling pathway. Here, ACIMH infers a pathway model based on mouse hepatocyte data. In addition, using CIMH, we analyze a biokinetic compartment model for zirconium processing in the human body that can readily be used in radiation protection. Transfer rates (including credible intervals) for an average individual are provided and form the basis for analyses considering retrospective dosimetry and bone retention of zirconium.

Zusammenfassung

Dynamische Systeme sind nützliche Hilfsmittel zur Erforschung der molekularen Funktionsweise lebender Organismen. Fundiertes Wissen über Reaktionsparameter ist hierbei entscheidend für biologische Modellvorhersagen und Hypothesen. Bayesianische Verfahren bieten einen differenzierten Ansatz zur Schätzung dieser Parameter. Darüber hinaus ermöglichen sie die Inferenz einer geeigneten Modellstruktur durch Modellselektion. Da die zugehörigen Posteriori-Verteilungen i.Allg. analytisch unlösbar sind, verwendet man approximative Markov-Chain-Monte-Carlo-Verfahren (MCMC-Verfahren). Im Falle komplexer Posteriori-Verteilungen generieren diese jedoch häufig ineffiziente Stichprobenvorschläge. Wir entwickeln in dieser Arbeit einen Metropolis-Hastings-Algorithmus (CIMH) zur effizienten Modellinferenz von dynamischen Systemen, welche auf Differentialgleichungen basieren. Das Konzept nutzt die Vine-Copula-Zerlegung einer approximativen Posteriori-Verteilung zur Erzeugung von Stichprobenvorschlägen, welche ähnlich der wahren Posteriori-Verteilung verteilt sind. Der CIMH-Algorithmus wird anschließend zu einem adaptiven Verfahren erweitert (ACIMH). Dies beschleunigt die Konvergenz in komplexen Systemen. Wir vergleichen CIMH und ACIMH mit etablierten Verfahren anhand verschiedener Beispiele. Darüber hinaus wenden wir die Algorithmen auf Fragestellungen aus dem Bereich der Systembiologie, wie etwa den Ablauf des STAT3-Dimertransports in den Zellkern für den JAK1-STAT3-Signalweg, an. Hier benutzen wir ACIMH, um ein auf Maushepatozyten basierendes, mathematisches Signalwegmodell zu inferieren. Zudem verwenden wir CIMH zur Analyse biokinetischer Modelle für die Verarbeitung von Zirkonium (Zr) im menschlichen Körper. Übertragungsraten (mit Kredibilitätsintervallen) für eine Durchschnittsperson werden berechnet und bilden die Grundlage für retrospektive Dosimetrie und die Analyse von Zr-Einlagerungen in Knochen.

Contents

List of Figures

List of Tables

LIST OF TABLES

1

Introduction

Dynamical systems are used in various fields of science. Frequently modeled by parametrized ordinary or delay differential equations they find application in physics, engineering, computational biology, and many more. In systems biology differential equation driven dynamical systems are commonly used to study the evolution and maintenance of cellular functionality over time. Here, scarce and noisy data for the mostly large and complex models hamper the unraveling of molecular interaction mechanisms. However, determining the underlying reaction rate parameters is crucial for profound model based predictions. Extensive research has therefore been done on the inference of these systems. The issue is typically addressed by initial value or maximum likelihood approaches (Horbelt *et al.* [2002]), which yield the supposedly most probable systematic parameter values. However, as the data and models are generally imprecise, determining single best values is often inadequate. In the last few years fully statistical Bayesian approaches were considered for parameter estimation in differential equation systems (Brown & Sethna [2003]; Lawrence *et al.* [2010]; Wilkinson [2006]). These Bayesian methods provide a nice way of combining the parameters of interest with the underlying data and a priori information by means of a posterior distribution, even when dealing with very complex models or partially unobserved quantities. Moreover they can readily be applied for model selection purposes and provide an overall model inference scheme that naturally corrects for overfitting – an issue classical model selection approaches are subject to.

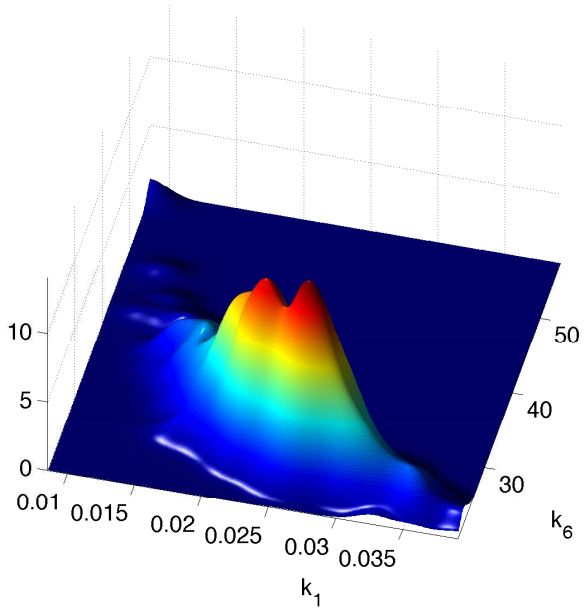


Figure 1.1: Posterior distribution of the JAK1-STAT3 model (7.1) introduced in Chapter 7 marginalized on the k_1 and k_6 dimension.

As analytic inference of the posterior distributions quickly becomes intractable, Markov Chain Monte Carlo (MCMC) methods are nowadays widely used to tackle the problem (Brooks [1998]; Gamerman & Lopes [2006]). Despite of their computationally costly nature MCMC methods have of late drawn major interest in the scientific community. One of the most successful and influential (Beichl & Sullivan [2000]; Wilkinson [2007]) algorithms for Markov chain sampling was developed by Metropolis and Hastings (Hastings [1970]; Metropolis *et al.* [1953]). It can draw samples from any probability distribution, given that a function proportional to the according probability density function is available. Throughout the sampling process a random walk proposal function based on the current Markov chain sample is used to generate a possible subsequent Markov chain candidate.

Fine-tuning the Metropolis-Hastings (MH) algorithm for performing efficient posterior inference is nevertheless a daunting task: It is easy to see that an efficient MH proposal function should ideally be (at least locally) very similar to the actual posterior distribution of interest. Figure 1.1 shows the "banana"-shaped posterior distribution

of the JAK1-STAT3 model (7.1) introduced in Chapter 7 marginalized on two dimensions (i.e. for the parameters k_1 and k_6). Such spiky, complex distributions are hard to mimic using a globally constant random walk proposal scheme as applied in the classical MH algorithm. More generally, strong parameter dependencies and complex sampling spaces limit MH algorithms to conservative parameter update schemes (Ramsay *et al.* [2007]). In other words, vast traversals in the parameter space call for huge amounts of MCMC iterations in situations with complex sampling spaces. Towards this end, a variety of algorithms based on techniques from mathematical optimization theory have been developed by Duane *et al.* [1987], Girolami & Calderhead [2011], Roberts & Stramer [2002], or Ter Braak & Vrugt [2008] in order to improve the MCMC sampling efficiency. These approaches propose Markov chain candidates using local information about the posterior manifold structure and with this attain higher proposal acceptance rates compared to the classical random walk MH algorithm. Another *Ansatz* that has of late drawn a lot of interest in the MCMC community is based on the successive adaption of the MH proposal function in order to amplify the sampling efficiency. Various algorithms were proposed by Haario *et al.* [1999], Haario *et al.* [2001], Roberts & Rosenthal [2007], Holden *et al.* [2009], or Müller & Sbalzarini [2010]. This can speed up the inference process severely as pointed out by Rosenthal [2011], and Gilks *et al.* [1998].

In this thesis we extend the classical MH algorithm by a novel estimated MH proposal function which generates Markov chain candidates almost independent of the current Markov chain state while taking into account the full estimated parameter dependency structure. This results in a proposal function that is ideally close to the actual posterior distribution and is therefore able to efficiently sample the often times complex distributions of the dynamical systems mentioned above. The sampling scheme is based on mimicking the posterior distribution by means of a D-vine copula decomposition. Updating the proposal copula during the sampling process leads to an adaptive sampling scheme. We employ this approach for the inference of two biological systems: first of all we analyze a model of the JAK1-STAT3 signaling pathway in order to investigate the effectiveness of tyrosine-phosphorylated STAT3 homodimers working as transcription factors for gene regulation. Secondly, we infer a biokinetic model for zirconium

processing in the human body, which can readily be used to derive limiting values of detrimental effects in radiation protection.

Overview of this thesis

In Chapter 2 we first introduce some basic notions and notations from probability theory used throughout this thesis. Then the concept of vine copula decomposition of probability densities is outlined. In addition, we shortly summarize the theory of random processes and Markov chains in particular. We focus on the important aspects regarding Markov Chain Monte Carlo methods. Moreover, a short introduction on biological signaling pathways and dynamical systems is given.

Chapter 3 considers the concept of Bayesian model inference. Here, Bayes' theorem yields posterior distributions that are proportional to the likelihood in combination with prior distributions for a series of given observations. More precisely, Bayesian model inference comprises two main aspects: On the one hand we introduce Bayes factors, which are capable of inferring the best model structure for a given set of parametrized models, i.e. we deduce the model $\mathcal{M}$ with the highest probability that the observations were generated according to $\mathcal{M}$. For this model selection task various approaches, such as the prior arithmetic mean estimate, the posterior harmonic mean estimate, and thermodynamic integration are presented. On the other hand the second strain of Bayesian model inference is treated: Given a specific parametrized model $\mathcal{M}$, we summarize the concept of posterior parameter inference.

Drawing samples from a posterior distribution is essential for Bayesian inference. As the posteriors are generally non-standard distributions, Chapter 4 addresses the concept of Markov Chain Monte Carlo (MCMC) sampling. Interestingly, both Bayesian model inference aspects can be simultaneously covered by MCMC sampling. We introduce the basic version of the Metropolis-Hastings sampling scheme (Hastings [1970]; Metropolis et al. [1953]) and discuss dependency and convergence diagnostics of the generated Markov chains. The latter find use in all applications throughout this thesis. Finally, extensions to optimization problems via simulated annealing (Kirkpatrick *et al.* [1983]) and the direct application to the model selection issue via reversible jump MCMC (Green [1995]) are given.

Due to the aforementioned complex posterior surfaces MCMC approaches often struggle with sampling efficiency. Especially the inference of parametrized differential equations likes to trap MCMC samplers between high proposal rejection rates and strong autocorrelation structures – both leading to a low number of independent samples drawn over time. In Chapter 5 we review two current approaches addressing the issue of improving the Metropolis-Hastings proposal function. Considered are a successive proposal function adaption scheme (Müller & Sbalzarini [2010]) and the simplified Riemann Manifold Metropolis adjusted Langevin algorithm (Girolami & Calderhead [2011]), which exploits the geometric posterior parameter structure for proposal generation.

In Chapter 6 we develop a novel vine copula based (adaptive) MCMC approach for efficient parameter inference in complex dynamic systems. Although copulas are well established in various fields of science, we are the first to exploit this concept for fine tuning the Metropolis-Hastings algorithm. Copulas are capable of handling asymmetric dependency structures (Aas *et al.* [2009]; Kurowicka & Cooke [2006a]; Kurowicka & Joe [2011]) – a characteristic also inherent to most posterior densities subject to MCMC sampling. The basic version of our novel hybrid Copula based Independence/random walk Metropolis-Hastings algorithm (CIMH) is presented and subsequently extend to an adaptive sampling scheme (ACIMH). This is the first major contribution of this thesis. The chapter is based on Schmidl *et al.* [2012a] and in part even identical.

In Chapter 7 we apply ACIMH to infer a model of the JAK1-STAT3 signaling pathway. Using thermodynamic integration we analyze the effect of direct tyrosine-phosphorylated STAT3 dimer import into the nucleus as compared to a model considering tyrosine-serine-phosphorylated STAT3 dimer import only. The estimated maximum a posteriori rate constants of the JAK1-STAT3 pathway are provided.

We conclude with the second major contribution of this thesis by analyzing a model for zirconium processing in the human body in Chapter 8. Again, using thermodynamic integration, this time in combination with CIMH, we compare a biokinetic model recently put forward by Greiter *et al.* [2011] to the model established by the International Commission on Radiological Protection. The latter is currently used for radiation protection. The former turns out to be superior based on *in vivo* plasma and urine data of 16 investigations in humans. Transfer rates (including credible intervals) for an average

individual are given. We also provide an estimation of initially ingested amounts of zirconium for *ex post* measurements, which is crucial for determining detrimental effects at occupational exposure. Furthermore, zirconium retention in bone is analyzed. The chapter is based on Schmidl *et al.* [2012b] and in part even identical.

Main scientific contributions

The main scientific contributions of this thesis are (i) the development of a novel vine copula based (adaptive) MCMC approach for efficient parameter inference in complex dynamic systems and (ii) the in-depth analysis of a model for zirconium processing in the human body. The latter includes the estimation of initially ingested amounts of zirconium for *ex post* measurements and zirconium retention in bone. These contributions are contained in the following two manuscripts:

- D. Schmidl, C. Czado, and F.J. Theis. A vine-copula based adaptive MCMC sampler for efficient inference of dynamical systems. *Under revision at Bayesian Analysis*.
- D. Schmidl, S. Hug, W.B. Li, M.B. Greiter, and F.J. Theis. Bayesian model selection validates a biokinetic model for zirconium processing in humans. *BMC Systems Biology*, 6(95), 2012.

Further scientific contributions

The following publications are not contained in this thesis. They present the results of various collaborations and projects in computational systems biology.

- D.M. Wittmann, D. Schmidl, F. Blöchl, and F.J. Theis. Reconstruction of graphs based on random walks. *Journal of Theoretical Computer Science*, 410 (38-40), 2009.
- A. Ruepp, A. Kowarsch, D. Schmidl, F. Buggenthin, B. Brauner, I. Dunger, G. Fobo, G. Frishman, C. Montrone, and F.J. Theis. PhenomiR: a knowledgebase for microRNA expression in diseases and biological processes. *Genome Biology* 11(1):R6, 2010.

- A. Kowarsch, C. Marr, D. Schmidl, A. Ruepp, and F.J. Theis. Tissue-specific target analysis of disease-associated microRNAs in human signaling pathways. *PLoS one*, 5(6), 2010.
- A. Kowarsch, D. Schmidl, S. Braun, S. Bohl, R. Merkle, U. Klingmüller, and F.J. Theis. MicroRNA-mediated regulation has an impact on the dynamic behavior of the JAK-STAT pathway. *Manuscript in preparation.*

2

Prerequisites

The following chapter holds the basic notions and notations from probability theory, the theory of Markov chains, copula constructions, and dynamical systems. These will be used in subsequent chapters.

2.1 Basics on probability theory

We start with the introduction of basics on probability theory, mostly following Chapter 1.3 of Theis [2002] with some extensions. For proofs see e.g. the book of Klenke [2008]. Throughout this thesis we denote vectors/matrices by bold letters, while non bold letters with subscript indices denote vector/matrix elements. Markov chains are displayed as sets, such as $\{\boldsymbol{X}^{(t)}\}_{t\in I}$ for some index set I, where the superscript (t) denotes the t^{th} element.

Definition 2.1 (Probability space)**.** A *probability space* is a triplet $(\Omega, \mathcal{F}, P)$ consisting of a non-empty set Ω, a sigma-algebra $\mathcal{F}$ on Ω and a *probability measure* P on $\mathcal{F}$ with $P(\Omega) = 1$.

The sets $A \in \mathcal{F}$ are called *events*, while $P(A)$ is called the *probability of the event* A. By definition we have $P(A) \in [0, 1]$. An event $A \in \mathcal{F}$ is said to occur *almost surely*, if the probability of A to not occur is zero.

Definition 2.2 (Random variable/random vector)**.** Suppose $(\Omega, \mathcal{F}, P)$ is a probability space and $(E', \mathcal{E}')$ a measurable space. Suppose furthermore $X : \Omega \longrightarrow E'$ is a measurable mapping. Then X is called a *random variable with values in* E'. An n-dimensional *random vector* with values in E is a vector-valued function $\boldsymbol{X} = (X_1, \ldots, X_n)^\top : \Omega \longrightarrow E$ into a measurable space $(E, \mathcal{E})$ whose components X_i are random variables on $(\Omega, \mathcal{F}, P)$.

For each $\omega \in \Omega$ we call $\boldsymbol{x} = \boldsymbol{X}(\omega) \in E$ a *realization* of $\boldsymbol{X}$. As we allow $n = 1$, we will further on speak of random vectors only. In our applications $(E, \mathcal{E}) = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, where $\mathcal{B}(\mathbb{R}^n)$ denotes the Borel sigma algebra on $\mathbb{R}^n$. In this case we call $\boldsymbol{X}$ a *real-valued random vector* or simply *random vector*. For $A \in \mathcal{E}$ we write $\boldsymbol{X}(P)(A) := P(\boldsymbol{X}^{-1}(A)) := P(\boldsymbol{X} \in A) := P_{\boldsymbol{X}}(A)$. This is a probability measure $\boldsymbol{X}(P) : \mathcal{E} \longrightarrow [0, 1]$, $A \longmapsto P_{\boldsymbol{X}}(A)$, since $P_{\boldsymbol{X}}(E) = P(\Omega) = 1$. The function $\boldsymbol{X}(P)$ is called *distribution of* $\boldsymbol{X}$ *with respect to* P. We write $\boldsymbol{X} \sim P_{\boldsymbol{X}}$ and call $\boldsymbol{X}$ to be $P_{\boldsymbol{X}}$ distributed.

For a finite sequence $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(T)}$ of $T \in \mathbb{N}$ random vectors on $(\Omega, \mathcal{F}, P)$ with values in $(E, \mathcal{E})$ we define the function $\boldsymbol{X}^{(1)} \otimes \ldots \otimes \boldsymbol{X}^{(T)}$ by

$$\begin{aligned} \boldsymbol{X}^{(1)} \otimes \ldots \otimes \boldsymbol{X}^{(T)} : \Omega &\longrightarrow E \times \ldots \times E \\ \omega &\longmapsto (\boldsymbol{X}^{(1)}(\omega)^\top, \ldots, \boldsymbol{X}^{(T)}(\omega)^\top)^\top. \end{aligned}$$

The function $\boldsymbol{X}^{(1)} \otimes \ldots \otimes \boldsymbol{X}^{(T)}$ also constitutes a random vector called *product random vector of* $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(T)}$. The according distribution $P_{\boldsymbol{X}^{(1)} \otimes \ldots \otimes \boldsymbol{X}^{(T)}}$ is called *joint (product) distribution of* $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(T)}$.

Definition 2.3 (Distribution function)**.** For a real-valued random vector $\boldsymbol{X} : \Omega \longrightarrow \mathbb{R}^n$ on $(\Omega, \mathcal{F}, P)$ the function

$$\begin{aligned} F_{\boldsymbol{X}} : \mathbb{R}^n &\longrightarrow [0, 1] \\ (x_1, \ldots, x_n)^\top &\longmapsto P_{\boldsymbol{X}}((-\infty, x_1]) \times \ldots \times (-\infty, x_n]) \end{aligned}$$

is called *distribution function of* $\boldsymbol{X}$ *with respect to* P, or likewise, *cumulative distribution function (cdf) of* $\boldsymbol{X}$.

The opposite direction also holds true, which allows us to identify a real-valued random vector with its distribution:

Proposition 2.1. *For each distribution function F there exists a real-valued random vector $\boldsymbol{X}$ with $F_{\boldsymbol{X}} = F$.*

Definition 2.4 (Density function)**.** If $f : \mathbb{R}^n \longrightarrow [0, \infty)$ is a non-negative Lebesgue-integrable function, such that the distribution function $F_{\boldsymbol{X}} : \mathbb{R}^n \longrightarrow [0, 1]$ with respect to P can be written as

$$F_{\boldsymbol{X}}(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \ldots \int_{-\infty}^{x_n} f(u_1, \ldots, u_n) \, \mathrm{d}u_n \ldots \mathrm{d}u_1$$

then f is called *density function with respect to P*, or likewise, *probability density function (pdf) of $\boldsymbol{X}$*. We also write $f_{\boldsymbol{X}}$ for f.

Clearly, if a distribution function F is sufficiently differentiable at a point $(x_1, \ldots, x_n)^\top \in \mathbb{R}^n$, the according probability density function is given by

$$f(x_1, \ldots, x_n) = \frac{\partial^n}{\partial x_1 \ldots \partial x_n} F(x_1, \ldots, x_n). \tag{2.1}$$

In this situation we also write $\boldsymbol{X} \sim f(\boldsymbol{x})$ instead of $X \sim P_{\boldsymbol{X}}$.

Following are two very important examples of distributions: (i) the uniform distribution will play a key role when dealing with copula densities as their margins are uniformly distributed random variates and (ii) the normal distribution, which is a common proposal density for the prominent Metropolis-Hastings algorithm defined in Chapter 4.1. A selection of important univariate density functions is given in Appendix A.

Example 2.1 (Uniform distribution). Suppose $A \subset \mathbb{R}^n$ is a non-empty measurable set and $\mathbb{1}_A$ the *indicator function on A*, this is

$$\begin{aligned} \mathbb{1}_A : \mathbb{R}^n &\longrightarrow \mathbb{R} \\ \boldsymbol{x} &\longrightarrow \begin{cases} 1, & \text{if } \boldsymbol{x} \in A \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

A random vector $\boldsymbol{X} : \Omega \longrightarrow \mathbb{R}^n$ is called *uniformly distributed on A*, if for the n-dimensional Lebesgue measure λ^n the density function

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

exists. We then write $\boldsymbol{X} \sim \mathcal{U}[A]$.

Example 2.2 (Normal distribution). A random vector $\boldsymbol{X} : \Omega \longrightarrow \mathbb{R}^n$ is said to be *normally distributed* (or *Gaussian*), if there exists a vector $\boldsymbol{\mu} \in \mathbb{R}^n$ along with a positive-semidefinite symmetric matrix $\boldsymbol{\Sigma} \in \mathrm{Mat}(n \times n, \mathbb{R})$, such that the density function

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

exists. We then write $\boldsymbol{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The quantities $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are called *mean* and *covariance matrix of* $\mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, respectively. An example of a bivariate normal density function can be seen in Figure 2.1(b).

Heading towards Markov chains, we need the notions of *marginal* and *conditional distribution functions*. While the former averages over a subset of random variables of $\boldsymbol{X}$, the latter fixes this subset to a specific value and in some sense cuts through the graph of the joint probability distribution function along this value. More precisely: Suppose we are given a random vector $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^\top : \Omega \longrightarrow \mathbb{R}^n$ together with its density function $f_{\boldsymbol{X}} : \mathbb{R}^n \longrightarrow \mathbb{R}$. Setting $\boldsymbol{Y} := (X_{n_1} \ldots, X_{n_k})^\top$ for a non-empty subset $\{n_1, \ldots, n_k\} \subsetneq \{1, \ldots, n\}$, then for $\{n'_1, \ldots, n'_l\} := \{1, \ldots, n\} \setminus \{n_1, \ldots, n_k\}$, the vector $\boldsymbol{Y}$ is a random vector with distribution function

$$F_{\boldsymbol{Y}}(x_{n_1} \ldots, x_{n_k}) = \int_{-\infty}^{x_{n_1}} \cdots \int_{-\infty}^{x_{n_k}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\boldsymbol{X}}(u_1, \ldots, u_n)\, \mathrm{d}u_{n'_l} \ldots \mathrm{d}u_{n'_1}\, \mathrm{d}u_{n_k} \ldots \mathrm{d}u_{n_1}.$$

The function $F_{\boldsymbol{Y}}$ is said to be the *marginal distribution function* of $\boldsymbol{Y}$. Assuming (after variable permutation) without loss of generality that $\{n_1, \ldots, n_k\} = \{1, \ldots, k\}$ and defining the complement of $\boldsymbol{Y}$ with respect to $\boldsymbol{X}$ as $\boldsymbol{Z} = (X_{k+1}, \ldots, X_n)^\top$ then the *marginal density function* of $\boldsymbol{Y}$ is given by

$$f_{\boldsymbol{Y}}(x_1 \ldots, x_k) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\boldsymbol{X}}(x_1, \ldots, x_k, u_{k+1}, \ldots, u_n)\, \mathrm{d}u_n \ldots \mathrm{d}u_{k+1}.$$

We can now turn to *conditional distribution functions*.

Definition 2.5 (Conditional distribution function)**.** Using the definitions from above the *conditional distribution function of* $\boldsymbol{Y} = (X_1, \ldots, X_k)^\top$ *given* $\boldsymbol{Z} = (x_{k+1}, \ldots, x_n)^\top$ for some $x_{k+1}, \ldots, x_n \in \mathbb{R}$ is for $f_{\boldsymbol{Z}}(x_{k+1} \ldots, x_n) > 0$ given by

$$F_{\boldsymbol{Y}|\boldsymbol{Z}}(x_1, \ldots, x_k | x_{k+1}, \ldots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_k} \frac{f_{\boldsymbol{X}}(u_1, \ldots, u_k, x_{k+1}, \ldots, x_n)}{f_{\boldsymbol{Z}}(x_{k+1} \ldots, x_n)}\, \mathrm{d}u_k \ldots \mathrm{d}u_1.$$

Definition 2.6 (Conditional density function)**.** The *conditional density function of* $\boldsymbol{Y}$ *given* $\boldsymbol{Z}$ is for $f_{\boldsymbol{Z}}(x_{k+1} \ldots, x_n) > 0$ given by

$$f_{\boldsymbol{Y}|\boldsymbol{Z}}(x_1, \ldots, x_k | x_{k+1}, \ldots, x_n) = \frac{f_{\boldsymbol{X}}(x_1, \ldots, x_n)}{f_{\boldsymbol{Z}}(x_{k+1} \ldots, x_n)}.$$

Suppose $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$ and A is the open n-dimensional cube with lower limits $\boldsymbol{a}$ and upper limits $\boldsymbol{b}$, i.e $A = (a_1, b_1) \times \ldots \times (a_n, b_n)$. Then $P(\boldsymbol{X} \in A)$ computes to

$$P(\boldsymbol{X} \in A) = \int_{a_1}^{b_1} \ldots \int_{a_n}^{b_n} f(u_1, \ldots, u_n) \, \mathrm{d}u_n, \ldots, \mathrm{d}u_1.$$

With this the notion of a conditional distribution function generalizes in the sense of the Borel measure in the usual way for all measurable sets $A \subset \mathbb{R}^n$ that are non-cubic (such as unions of n-dimensional cubes). This gives a sound definition for the probability $P(\boldsymbol{Y} \in B | x_{n'_1}, \ldots, x_{n'_l})$ for any set $B \in \mathcal{B}(\mathbb{R}^k)$. For these we also write $P(\boldsymbol{Y} \in B | X_{n'_1} = x_{n'_1}, \ldots, X_{n'_l} = x_{n'_l})$ or $P(B | x_{n'_1}, \ldots, x_{n'_l})$.

In order to simplify notation later on we also write for $\boldsymbol{Y} = (X_{n_1} \ldots, X_{n_k})^\top$ and $\boldsymbol{Z} = (X_{n'_1} \ldots, X_{n'_l})^\top$

$$F_{n_1 \ldots n_k}, \; f_{n_1 \ldots n_k}, \; F_{n_1 \ldots n_k | n'_1 \ldots n'_l}, \text{ and } f_{n_1 \ldots n_k | n'_1 \ldots n'_l}$$

instead of

$$F_{\boldsymbol{Y}}, \; f_{\boldsymbol{Y}} \, , F_{\boldsymbol{Y}|\boldsymbol{Z}}, \text{ and } f_{\boldsymbol{Y}|\boldsymbol{Z}},$$

respectively.

Definition 2.7 (Independent random vector)**.** Let $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(T)}$ be a finite sequence of random vectors on some probability space $(\Omega, \mathcal{F}, P)$. Let furthermore $F_{\boldsymbol{X}^{(1)} \otimes \ldots \otimes \boldsymbol{X}^{(T)}}(\cdot)$ be the distribution function of $\boldsymbol{X}^{(1)} \otimes \ldots \otimes \boldsymbol{X}^{(T)}$. Then $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(T)}$ are called *independent*, if

$$F_{\boldsymbol{X}^{(1)} \otimes \ldots \otimes \boldsymbol{X}^{(T)}}(\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(T)}) = \prod_{t=1}^{T} F_{\boldsymbol{X}^{(t)}}(\boldsymbol{x}^{(t)}).$$

In Markov Chain Monte Carlo methods a sequence of realizations of random vectors is drawn in order to approximate a distribution $P_{\boldsymbol{X}}$ for some random vector $\boldsymbol{X}$. The law of large numbers will lay the basis for this procedure. We therefore define:

Definition 2.8 (Expectation)**.** For an integrable random vector $\boldsymbol{X} : \Omega \longrightarrow \mathbb{R}^n$ on a probability space $(\Omega, \mathcal{F}, P)$

$$\mathbb{E}_{P_{\boldsymbol{X}}}[\boldsymbol{X}] = \int_{\mathbb{R}^n} \boldsymbol{x}\, P_{\boldsymbol{X}}$$

is referred to as the *expectation* or *mean of* $\boldsymbol{X}$.

In case there exists a density function $f(\boldsymbol{x})$ corresponding to $P_{\boldsymbol{X}}$, we also write $\mathbb{E}_{f(\boldsymbol{x})}[\boldsymbol{X}]$ instead of $\mathbb{E}_{P_{\boldsymbol{X}}}[\boldsymbol{X}]$, or, where clear without ambiguity, simply $\mathbb{E}[\boldsymbol{X}]$. The expectation behaves nicely with respect to the product of independent random vectors:

Proposition 2.2. *For a sequence of independent integrable random vectors* $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(T)}$ *the component-wise product* $\prod_{j=1}^{T} \boldsymbol{X}^{(j)}$ *is integrable and*

$$\mathbb{E}\left[\prod_{j=1}^{T} \boldsymbol{X}^{(j)}\right] = \prod_{j=1}^{T} \mathbb{E}[\boldsymbol{X}^{(j)}].$$

We denote a series of realizations $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(T)}$ from the very same random vector $\boldsymbol{X}$ to be *independent identically distributed* (i.i.d.) of the distribution $P_{\boldsymbol{X}}$. This leads us to:

Theorem 2.1 (Strong law of large numbers)**.** *For an i.i.d. sequence* $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(T)}$ *of realizations of* $\boldsymbol{X}$

$$\lim_{T\to\infty} \frac{1}{T} \sum_{i=1}^{T} (\boldsymbol{x}^{(i)} - \mathbb{E}[\boldsymbol{X}]) = \lim_{T\to\infty} \left(\frac{1}{T} \sum_{i=1}^{T} \boldsymbol{x}^{(i)}\right) - \mathbb{E}[\boldsymbol{X}] = \boldsymbol{0} \in \mathbb{R}^n$$

almost surely.

This shows that the *expected mean*

$$\bar{\boldsymbol{x}} = \frac{1}{T} \sum_{i=1}^{T} \boldsymbol{x}^{(i)}$$

of a set of i.i.d. realizations $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(T)}$ truly converges to the expectation of $\boldsymbol{X}$. The approximation of $\mathbb{E}[\boldsymbol{X}]$ by $\bar{\boldsymbol{x}}$ gets better as T increases.

For a random vector $\boldsymbol{X}$ an unbiased approximation for the *variance*

$$\mathrm{Var}[\boldsymbol{X}] = \mathbb{E}\left[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])^2\right],$$

if existent, is given by the *expected variance*

$$s_{\boldsymbol{x}} = \frac{1}{T-1}\sum_{i=1}^{T}(\boldsymbol{x}^{(i)} - \bar{\boldsymbol{x}})^2,$$

where the square is taken component-wise.

For two random vectors $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ with existing variance the *covariance* $\mathrm{cov}[\boldsymbol{X}, \boldsymbol{Y}]$ is defined by the matrix

$$\mathrm{cov}[\boldsymbol{X}, \boldsymbol{Y}] = (\mathrm{cov}_{i,j})_{i,j=1,\ldots,n} \qquad \text{with} \qquad \mathrm{cov}_{i,j} = \mathbb{E}[(X_i - \mathbb{E}[X_i])(Y_j - \mathbb{E}[Y_j])].$$

Moreover, *Pearson's correlation matrix* of $\boldsymbol{X}$ and $\boldsymbol{Y}$ is given by

$$\mathrm{corr}[\boldsymbol{X}, \boldsymbol{Y}] = (\mathrm{corr}_{i,j})_{i,j=1,\ldots,n} \qquad \text{with} \qquad \mathrm{corr}_{i,j} = \frac{\mathrm{cov}[X_i, X_j]}{\sqrt{\mathrm{Var}[X_i]} \cdot \sqrt{\mathrm{Var}[Y_j]}}.$$

Pearson's correlation matrix contains the linear dependence between $\boldsymbol{X}$ and $\boldsymbol{Y}$. Since the dependence between $\boldsymbol{X}$ and $\boldsymbol{Y}$ can also be nonlinear, we also define a rank-based dependence measure: For two random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$, *Kendall's* τ is given by the component-wise difference of the probability of concordance and the probability of discordance of $\boldsymbol{X}$ and $\boldsymbol{Y}$, this is

$$\tau[\boldsymbol{X}, \boldsymbol{Y}] = (\tau_{i,j})_{i,j=1,\ldots,n}$$
$$\text{with} \qquad \tau_{i,j} = P((X_i - X_i')(Y_j - Y_j') \geq 0) - P((X_i - X_i')(Y_j - Y_j') < 0)$$

for the joint probability P of $\boldsymbol{X}$ and $\boldsymbol{Y}$ and two independent identically distributed copies $\boldsymbol{X}'$ of $\boldsymbol{X}$ and $\boldsymbol{Y}'$ of $\boldsymbol{Y}$.

Lastly, in order to compare the distance between two distributions $P_{\boldsymbol{X}}$ and $P_{\boldsymbol{Y}}$ for two random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ with values in $(E, \mathcal{E})$, we define the *total variation norm* as

$$\|P_{\boldsymbol{X}} - P_{\boldsymbol{Y}}\|_{TV} = \sup_{A \in \mathcal{E}} |P_{\boldsymbol{X}}(A) - P_{\boldsymbol{Y}}(A)|.$$

2.2 Copula distributions

We now turn to the task of characterizing a continuous n-variate distribution function $F(\boldsymbol{x})$ with given margins by its according univariate pdf's $f(x_i)$, $i = 1, \ldots, n$, and a second n-variate distribution function $C(\boldsymbol{u})$ with uniformly distributed margins on

$[0,1]$, called *copula* (see below for a rigorous definition). This concept allows for efficient modeling of multivariate distributions with asymmetric tail dependencies. It splits up $F(\boldsymbol{x})$ into one part containing the dependency structure and another one containing all marginal informations on the according random vector $\boldsymbol{X}$. Copulas have been successfully applied in the fields of economics, finance, or geology. For a thorough introduction on the theory (including proofs) and applications see for example Joe [1997] or Nelsen [2006].

Definition 2.9 (Copula)**.** A function $C : [0,1]^n \longrightarrow [0,1]$ is called *n-dimensional copula*, if the following properties hold:

(i) $C(\boldsymbol{u}) = 0$ for all $\boldsymbol{u} = (u_1, \ldots, u_n)^\top \in [0,1]^n$ with $u_i = 0$ for some $i \in \{1, \ldots, n\}$.

(ii) $C(\boldsymbol{u}) = u_i$ for all $\boldsymbol{u} = (u_1, \ldots, u_n)^\top \in [0,1]^n$ with $u_j = 1$ for all $i \neq j$.

(iii) $C(\boldsymbol{u})$ is *n-increasing*, i.e. for all cubes $A = \times_{i=1}^n [a_i, b_i] \subseteq [0,1]^n$ the volume of A with respect to C is non-negative:

$$V_C(A) := \sum_{\boldsymbol{u} \in \times_{i=1}^n \{a_i, b_i\}} \operatorname{sgn}(\boldsymbol{u}) C(\boldsymbol{u}) \geq 0.$$

Here, sgn is a sign function with value one, if $u_i = a_i$ for an even number of i's and minus one for an odd number of i's.

The following fundamental theorem by Sklar [1959] uniquely links multivariate cdf's on $\overline{\mathbb{R}}^n$ with its univariate margins by means of a copula. Here, $\overline{\mathbb{R}}^n$ denotes the n-dimensional cartesian product of $\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, \infty\}$.

Theorem 2.2 (Sklar)**.** *Suppose F is an n-dimensional distribution function with continuous univariate margins $F_1, \ldots, F_n$. Then there exists a unique copula C, such that for all $\boldsymbol{x} = (x_1, \ldots, x_n)^\top \in \overline{\mathbb{R}}^n$*

$$F(\boldsymbol{x}) = C(F_1(x_1), \ldots, F_n(x_n)). \tag{2.2}$$

Conversely, for any copula C and univariate distribution functions $F_1, \ldots, F_n$ the function F defined by Equation (2.2) is a multivariate distribution function with margins $F_1, \ldots, F_n$.

The relation (2.2) furthermore defines the density of a copula C: Suppose the necessary derivatives exist, then the chain rule yields

$$f(\boldsymbol{x}) = \frac{\partial^n C(F_1(x_1), \ldots, F_n(x_n))}{\partial x_1 \ldots \partial x_n} = \frac{\partial^n C(F_1(x_1), \ldots, F_n(x_n))}{\partial F_1(x_1) \ldots \partial F_n(x_n)} \prod_{i=1}^{n} f_i(x_i) \tag{2.3}$$

at some point $\boldsymbol{x} = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n$. As suggested by Equation (2.1) we set

$$c(F_1(x_1), \ldots, F_n(x_n)) := \frac{\partial^n C(F_1(x_1), \ldots, F_n(x_n))}{\partial F_1(x_1) \ldots \partial F_n(x_n)}. \tag{2.4}$$

Hence, every probability density function f can be decomposed as product

$$f(\boldsymbol{x}) = c(F_1(x_1), \ldots, F_n(x_n)) \cdot f_1(x_1) \cdot \ldots \cdot f_n(x_n). \tag{2.5}$$

for *the* corresponding copula density c to f. For strictly positive marginal distribution functions $f_1, \ldots, f_n$

$$c(F_1(x_1), \ldots, F_n(x_n)) = \frac{f(\boldsymbol{x})}{f_1(x_1) \cdot \ldots \cdot f_n(x_n)}.$$

The copula density thus contains the dependency structure, while the marginal information is "divided" out.

Sklar's theorem also yields a natural recipe for the construction of copulas based on distribution functions F with invertible margins $F_1, \ldots, F_n$: we set for $\boldsymbol{u} = (u_1, \ldots, u_n)^\top \in [0,1]^n$

$$C(\boldsymbol{u}) = F(F_1^{-1}(u_1), \ldots, F_n^{-1}(u_n)). \tag{2.6}$$

This method is called *inversion method* (Nelsen [2006]) and lays the basis for sampling from copula distributions. In order to familiarize the reader with the concept of copulas we consider two very prominent classes of multivariate copula densities: these are for one Archimedean and for the other elliptical copulas.

Theorem 2.3 (Archimedean copula)**.** *Suppose $\varphi : [0,1] \longrightarrow [0,\infty]$ is a continuous strictly decreasing function with $\varphi(0) = \infty$ and $\varphi(1) = 0$. Suppose furthermore that the inverse φ^{-1} of φ is completely monotonic, i.e. φ^{-1} is continuous and satisfies $(-1)^k \frac{\mathrm{d}^k}{\mathrm{d}x^k}\varphi^{-1}(x) \geq 0$ for all $x \in (0,\infty)$ and all $k \in \mathbb{N}_0$. Then*

$$C(\boldsymbol{u}) = \varphi^{-1}(\varphi(u_1) + \ldots + \varphi(u_n))$$

is a copula.

The copula C defined by Theorem 2.3 is called *Archimedean copula*. The function φ is said to be the *generator* of C. A proof of Theorem 2.3 is given in Nelsen [2006].

Example 2.3 (Independence copula). The *independence copula* is defined as

$$C_I : [0,1]^n \longrightarrow [0,1]$$
$$(u_1, \ldots, u_n) \longmapsto \prod_{i=1}^{n} u_i.$$

It is an Archimedean copula with generator $\varphi(x) = -\log(x)$. A plot of a bivariate independence copula density function is shown in 2.1(a). The term "independence" originates from the corresponding copula density function c_I: For any multivariate continuous distribution function F with according density function f Equation (2.4) yields

$$c_I(F_1(x_1), \ldots, F_n(x_n)) = \frac{\partial^n C_I(F_1(x_1), \ldots, F_n(x_n))}{\partial F_1(x_1) \ldots \partial F_n(x_n)} = 1.$$

Equation (2.5) therefore gives

$$f(\boldsymbol{x}) = \prod_{i=1}^{n} f_i(x_i).$$

This means the corresponding random variables $X_1, \ldots, X_n$ are independent.

The inversion method of Equation (2.6) yields the class of *elliptical copulas* for *elliptical density functions* $f(\boldsymbol{x})$, this is

$$f(\boldsymbol{x}) = \frac{c_n}{\sqrt{\det(\boldsymbol{\Sigma})}} g((\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}))$$

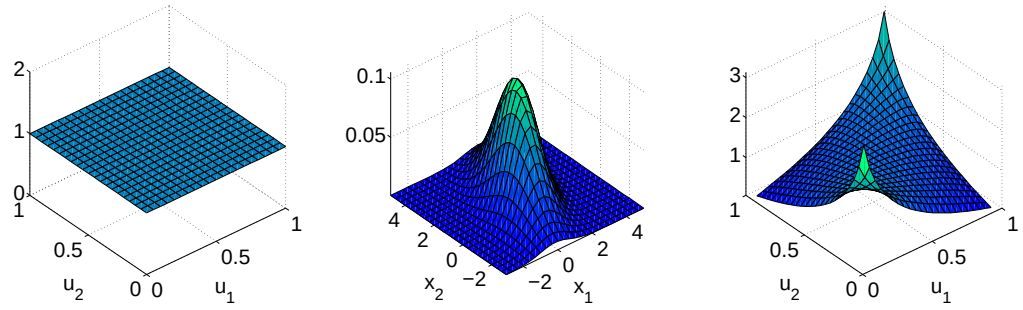
for some constant $c_n \in \mathbb{R}$, a univariate function g, a mean vector $\boldsymbol{\mu} \in \mathbb{R}^n$, and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$.

Example 2.4 (Gaussian copula). Suppose $f(\boldsymbol{x})$ is the n-variate Gaussian density function from Example 2.2 with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. For the corresponding correlation matrix $\boldsymbol{S}$ the Gaussian copula is given by

$$C(\boldsymbol{u}) = \Phi_{\boldsymbol{S}}(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_n)),$$

where Φ^{-1} is the inverse cdf of the univariate standard normal distribution $\mathcal{N}(0,1)$ and $\Phi_{\boldsymbol{S}}$ the n-variate normal cdf with covariance matrix $\boldsymbol{S}$. The density is given by

$$c(\boldsymbol{u}) = \frac{1}{\sqrt{\det(\boldsymbol{S})}} \exp\left(-\frac{1}{2} \boldsymbol{x}^\top (\boldsymbol{S}^{-1} - \boldsymbol{I}_n) \boldsymbol{x}\right)$$



(a) (b) (c)

Figure 2.1: (a) Independence copula density function. (b) Correlated Gaussian density function with correlation parameter $\rho = 0.5$. (c) Gaussian copula density function for the correlated Gaussian density from (b).

with $\boldsymbol{x} := (\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_n))^\top$ for $\boldsymbol{u} = (u_1, \ldots, u_n)^\top \in [0,1]^n$ and the n-dimensional identity matrix $\boldsymbol{I}_n$ (Arbenz [2011]). A plot of a bivariate Gaussian density with

$$\boldsymbol{\mu} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & \sqrt{3/4} \\ \sqrt{3/4} & 3 \end{pmatrix}$$

is shown in Figure 2.1(b). Here, the according correlation parameter $\rho = 0.5$. The corresponding copula density function is depicted in Figure 2.1(c).

Although there are multivariate copulas that are neither Archimedean, nor elliptic, the task of fitting a parametrized[1] multivariate copula to some vector of observations can be daunting. In the following we introduce a more flexible approach for fitting parametrized copulas to a given vector of observations.

2.2.1 Pair copula decomposition

The class of classical multivariate copulas has been considerably extended by Joe [1996]. Joe showed that a decomposition involving only bivariate copula densities and marginal densities provides a valid multivariate density. We follow Aas *et al.* [2009] for the introduction of these pair copula decompositions.

[1] We call a copula parametrized, if it depends on some parameter vector $\boldsymbol{\theta}$ i.e. $C(\cdot) = C(\cdot|\boldsymbol{\theta})$.

Suppose $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ is a random vector with distribution function $F(x_1, \ldots, x_n)$ and probability density function $f(x_1, \ldots, x_n)$. Then, except for permutation of the variables, we are given the unique decomposition

$$f(x_1, \ldots, x_n) = f(x_n) \cdot f(x_{n-1}|x_n) \cdot f(x_{n-2}|x_{n-1}, x_n) \cdot \ldots \cdot f(x_1|x_2, \ldots, x_n), \quad (2.7)$$

for the respective conditioned distribution functions $F(\cdot|\cdot)$ and density functions $f(\cdot|\cdot)$.

We iteratively derive a pair copula decomposition of f starting at $n = 2$: Using Sklar's theorem (Theorem 2.2), there exists a unique bivariate copula density function $c_{1,2}$, such that

$$f(x_1, x_2) = c_{1,2}(F_1(x_1), F_2(x_2)) \cdot f_1(x_1) \cdot f_2(x_2).$$

Thus, it follows from Equation (2.7) that, using $c_{1,2}$, the conditional density $f(x_1|x_2)$ can be written as

$$f(x_1|x_2) = c_{1,2}(F_1(x_1), F_2(x_2)) \cdot f_1(x_1). \quad (2.8)$$

For $n = 3$ we consider the conditional density function $f(x_1, x_2|x_3)$ of $f(x_1, x_2, x_3)$. Again, by Sklar's theorem there exists a copula density function $c_{1,2|3}$, such that

$$f(x_1|x_2, x_3) = c_{1,3|2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2)|x_2) \cdot f(x_1|x_2). \quad (2.9)$$

We already see that this decomposition is not unique as there also exists a copula density function $c_{1,2|3}$ with

$$f(x_1|x_2, x_3) = c_{1,2|3}(F_{1|3}(x_1|x_3), F_{2|3}(x_2|x_3)|x_3) \cdot f(x_1|x_3). \quad (2.10)$$

Nevertheless, using Equation (2.9) in combination with (2.8)

$$f(x_1|x_2, x_3) = c_{1,3|2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2)|x_2) \cdot c_{1,2}(F_1(x_1), F_2(x_2)) \cdot f_1(x_1).$$

For $n > 3$ we now see that the decomposition of the conditional density function $f(x_t|x_{t+1}, \ldots, x_n)$ is for $t \leq n - 2$ given by

$$f(x_t|\boldsymbol{v}) = c_{t,j|\mathcal{D}_{-j}}(F(x_t|\boldsymbol{v}_{-j}), F(v_j|\boldsymbol{v}_{-j})|\boldsymbol{v}_{-j}) \cdot f(x_t|\boldsymbol{v}_{-j}), \quad (2.11)$$

where $j \in \mathcal{D}$ for $\mathcal{D} := \{1, \ldots, t-1\}$ with $\mathcal{D}_{-j} := \mathcal{D} \setminus j$, and $\boldsymbol{v}_{-j}$ denotes the vector $\boldsymbol{v} = (x_1, \ldots, x_{t-1})^\top$ with missing j^{th} component. This allows us to decompose f into

the product of a series of bivariate copula density and marginal density functions using Equation (2.7).

In general the conditional pair copula densities in (2.11) depend on the conditioning values $\boldsymbol{v}_{-j}$. However, we will assume the restriction that the $c_{t,j|\mathcal{D}_j}(\cdot,\cdot|\boldsymbol{v}_{-j})$ do not depend on $\boldsymbol{v}_{-j}$. This means that the decomposition (2.11) captures the dependency on the conditioning values solely through the arguments $F(x_t|\boldsymbol{v}_{-j})$ and $F(v_j|\boldsymbol{v}_{-j})$. Hobæk Haff *et al.* [2010] showed that this restriction is not severe. In the Gaussian and multivariate Student-t-case the conditional pair copula densities are independent of the conditioning values.

Aas *et al.* [2009] were the first to consider standard estimation methods for parameters of vine copulas such as stepwise and maximum likelihood estimation (MLE). Since we have an explicit expression for the joint density, the likelihood is easily derived (see e.g. Aas *et al.* [2009]). These expressions however involve conditional cdf's: Joe [1996] showed that the conditional distribution functions corresponding to Equation (2.11) can be computed by

$$F(x_t|\boldsymbol{v}) = \frac{\partial\, C_{t,j|\mathcal{D}_{-j}}(F(x_t|\boldsymbol{v}_{-j}), F(v_j|\boldsymbol{v}_{-j}))}{\partial F(v_j|\boldsymbol{v}_{-j})}, \tag{2.12}$$

where $C_{t,j|\mathcal{D}_{-j}}$ is the copula distribution function corresponding to $c_{t,j|\mathcal{D}_{-j}}$. Therefore, the required conditional cdf's can be computed recursively. Equation (2.12) is furthermore used for sampling from copulas as shown in Aas *et al.* [2009].

For bivariate Gaussian copulas as building blocks and Gaussian marginals the resulting joint density is multivariate Gaussian. Here, the bivariate copula parameters correspond to partial correlations, which can be chosen arbitrarily between -1 and 1 and still induce a positive definite correlation matrix (Joe [1996]). Similarly, if one uses bivariate t-copulas together with a restriction on the degree of freedom parameters for different numbers of conditioning variables, a multivariate t-distribution emerges.

Example 2.5 (Trivariate copula decomposition). Suppose $\boldsymbol{X} = (X_1, X_2, X_3)^\top$ is a normally distributed random vector with standard normally distributed marginals $f_1(x_1)$, $f_2(x_2)$, $f_3(x_3)$ and pdf $f(x_1, x_2, x_3)$. Then there exist bivariate Gaussian copula densities $c_{1,2}$, $c_{2,3}$, and $c_{1,3|2}$ with respective correlation parameters ρ_{12}, ρ_{23} and $\rho_{13|2}$ such

that

$$\begin{aligned} f(x_1, x_2, x_3) =& c_{1,3|2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2)) \\ & \cdot c_{1,2}(F_1(x_1), F_2(x_2)) \cdot c_{2,3}(F_2(x_2), F_3(x_3)) \cdot \prod_{i=1}^{3} f_i(x_i). \end{aligned} \tag{2.13}$$

Conversely (c.f. Aas *et al.* [2009]), for given correlation parameters ρ_{12}, ρ_{23} and $\rho_{13|2}$, Equation (2.13) defines a trivariate normal distribution with correlation matrix

$$S = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{pmatrix}.$$

Here, ρ_{13} is given by

$$\rho_{13} = \rho_{13|2}\sqrt{(1-\rho_{12}^2)(1-\rho_{23}^2)} + \rho_{12}\rho_{23}.$$

In summary, pair copula decomposition constitutes a natural and flexible tool for modeling complex multivariate distribution functions. Fundamental properties, such as asymmetric or tail dependencies are broken down to the direct interactions of pairs of random variables (Joe *et al.* [2010]). A list of important bivariate copulas is given in Appendix B.

2.2.2 Vines

We already saw from Equations (2.9) and (2.10) that the pairwise decomposition of a density function f into the product of bivariate copulas and marginal distributions is by no means unique. Bedford & Cooke [2001, 2002] came up with a graphical representation to classify general pair copula models, called *regular vines* or *R-vines*. Essentially, they are represented by a collection of linked trees. In our applications we will focus on a subclass of these *R-vines*, called *D-vines*. This fixes the decomposition structure to some extent and by that reduces the number of possible decompositions severely. However, before we can characterize D-vines, we need some definitions from graph theory:

An *(undirected) graph* is a pair $\mathcal{G} = (V, E)$, where for $k, k' \in \mathbb{N}$, $V = \{v_1, \ldots, v_k\}$ is a set of *nodes* (also called *vertices*) and $E = \{e_1, \ldots, e_{k'}\}$ is a set of *edges*, such that

$e = (v, v') \in E$ for some $v, v' \in V$. The order of v and v' in e is of no importance here, i.e. we identify $e = (v, v') = (v,' v)$. A *path* $P = (V^*, E^*)$ in a graph $\mathcal{G} = (V, E)$ is a graph with $V^* = \{v_1^*, \dots, v_l^*\} \subseteq V$ and $E^* = \{(v_1^*, v_2^*), (v_2^*, v_3^*), \dots, (v*_{l-1}, v_l^*)\} \subseteq E$. If $v_l^* = v_1^*$, then $\mathcal{G}$ is said to be a *cycle*. A graph $\mathcal{G}$ is called *acyclic*, if it does not contain cycles. An acyclic graph $T = (V, E)$ is called a *tree*.

Definition 2.10 (Regular vine)**.** A *regular vine* on $n \in \mathbb{N}$ nodes is a collection of $(n-1)$ trees $\mathcal{V} = (T_1, \dots, T_{n-1})$ such that:

(i) $T_1 = (V_1, E_1)$ has the set of nodes $V_1 = \{v_1, \dots, v_n\}$.

(ii) For $i = 2, \dots, n-1$ the set of nodes of $T_i = (V_i, E_i)$ is given by $V_i = E_{i-1}$.

(iii) For $i = 2, \dots, n-1$ every element $(v_i, v_i') \in E_i$ consists of two elements (v_{i-1}, v_{i-1}') and $(w_{i-1}, w_{i-1}') \in E_{i-1}$ where exactly one of the v's coincides with one of the w's.

Loosely speaking, condition (ii) says that every edge e in a tree T_i becomes a node in the subsequent tree T_{i+1}, while condition (iii) states that two adjacent nodes in tree T_i are connected to a common node in T_{i-1}. In the following the nodes of tree $T_1 = (V_1, E_1)$ will correspond to some random vector $\boldsymbol{X} = (X_1, \dots, X_n)^\top$, i.e. $v_i = X_i$ for all $v_i \in V_1$. On the other hand, we depict the edges of tree T_i by an ordered labeling $kl|\mathcal{D}$ with $k < l$ for $v_k = X_k$ and $v_l = X_l$. Here, $\mathcal{D}$ is the set of conditioning variables, which is ordered increasingly as well. This results in a unique representation for every R-vine. The labeling corresponds directly to the set of varying and conditioning variables k, l and $\mathcal{D}$ of the copula $c_{k,l|\mathcal{D}}$. An example of a graphical representation of a six-dimensional R-vine is shown in Figure 2.2(a). The class of R-vines is sufficient to represent pair copula decompositions. However, this is no longer true for higher dimensional copulas as their graphical representation can contain loops (see e.g. Diestel [2000] for a definition).

Since the notion of R-vines does not impose much structure on the number of resulting decompositions, we later on restrict ourselves to *D-vines*.

Definition 2.11 (D-vine)**.** A regular vine is called *D-vine*, if the degree of each node v in T_1 is at most two, i.e. v is contained in at most two edges of E_1.

For completeness we need to point out another very prominent class of vines: We call a regular vine a *canonical vine* or *C-vine*, if there exists one node v in T_1, which is directly connected to all other nodes. Such a graph is also often called a *star*. Note that for $i = 2, \dots, n-1$ all trees T_i in a C-vine are stars as well. An example of a graphical representation of a six-dimensional C-vine and a six-dimensional D-vine is depicted in Figure 2.2(b) and 2.2(c), respectively. While the number of different R-vines on n nodes computes to $\binom{n}{2} \cdot (n-2)! \cdot 2^{\binom{n-2}{2}}$ (Morales-Nápoles *et al.* [2010]), the number of possible C- or D-vines on n nodes is given by $\frac{n!}{2}$ (Aas *et al.* [2009]). For $n = 6$ we can build 23.040 different R-vines and 320 different C- or D-vines. Hence, the notion of C- and D-vines clearly imposes some structure on the pair copula decomposition. A comprehensive introduction to vines is for instance given in Kurowicka & Cooke [2006b].

For distinct indices $i, j, i_1, \dots, i_k \in \{1, \dots, n\}$ with $k \leq (n-2)$, $i < j$ and $i_1 < \dots < i_k$ we abbreviate

$$c_{i,j|i_1,\dots,i_k} := c_{i,j|i_1,\dots,i_k}(F(x_i|x_{i_1}, \dots, x_{i_k}), F(x_j|x_{i_1}, \dots, x_{i_k})).$$

Using the variable order of Equation (2.7) we get the D-vine decomposition

$$f(x_1, \dots, x_n) = \left[\prod_{j=1}^{n-1} \prod_{i=1}^{n-j} c_{i,i+j|i+1,\dots,i+j-1}\right] \cdot \left[\prod_{k=1}^{n} f_k(x_k)\right]. \tag{2.14}$$

For example, the joint density function of a five dimensional D-vine is given by

$$f(x_1, \dots, x_5) = \left[\prod_{k=1}^{5} f_k(x_k)\right] \cdot c_{1,2} \cdot c_{2,3} \cdot c_{3,4} \cdot c_{4,5} \cdot c_{1,3|2} \cdot c_{2,4|3} \cdot c_{3,5|4} \cdot c_{1,4|2,3} \cdot c_{2,5|3,4} \cdot c_{1,5|2,3,4}.$$

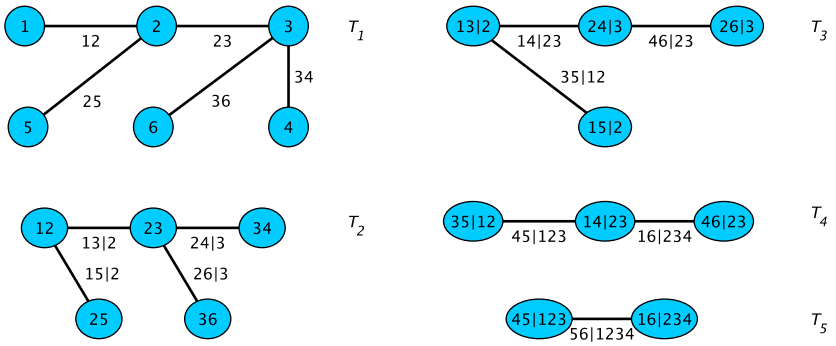
For a C-vine we on the other hand have the decomposition

$$f(x_1, \dots, x_n) = \left[\prod_{j=1}^{n-1} \prod_{i=1}^{n-j} c_{j,i+j|1,\dots,j-1}\right] \cdot \left[\prod_{k=1}^{n} f_k(x_k)\right]. \tag{2.15}$$

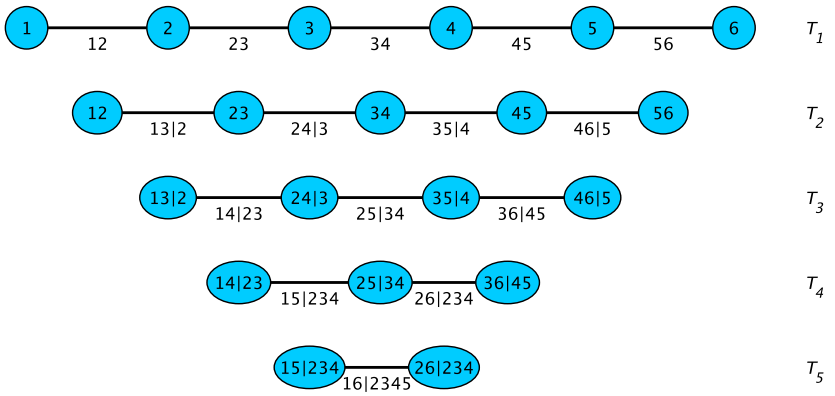
And the joint density function of a five dimensional C-vine is given by

$$f(x_1, \dots, x_5) = \left[\prod_{k=1}^{5} f_k(x_k)\right] \cdot c_{1,2} \cdot c_{1,3} \cdot c_{1,4} \cdot c_{1,5} \cdot c_{2,3|1} \cdot c_{2,4|1} \cdot c_{2,5|1} \cdot c_{3,4|1,2} \cdot c_{3,5|1,2} \cdot c_{4,5|1,2,3}.$$

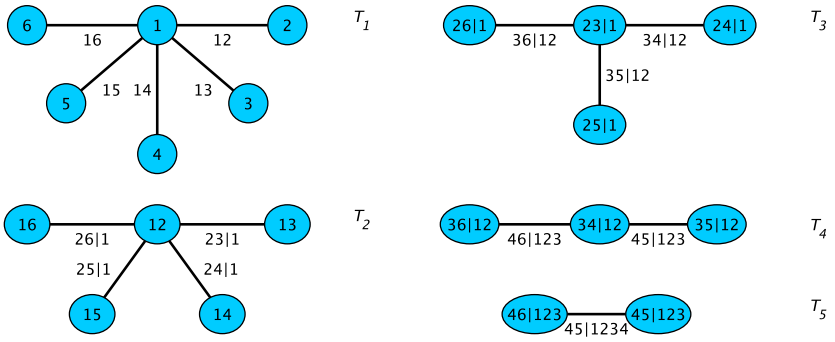
Note that the decompositions (2.14) and (2.15) of the joint density consist of pair copula densities $c_{i,j|i_1,\dots,i_k}(\cdot,\cdot)$ evaluated at conditonal distribution functions $F(x_i|x_{i_1}, \dots x_{i_k})$ and $F(x_j|x_{i_1}, \dots, x_{i_k})$ for specified indices $i, j, i_1, \dots, i_k$ and marginal densities f_k.



(a)



(b)



(c)

Figure 2.2: (a) Example of an R-vine on six nodes. (b) Example of a D-vine on six nodes. (c) Example of a C-vine on six nodes.

2.3 Markov chains

In the following we introduce the concept of Markov chains and address some of their properties. We do not go into full detail, but rather focus on the important aspects regarding Markov Chain Monte Carlo (MCMC) methods. The reader may consult Wilkinson [2006] for an easily readable and Meyn *et al.* [1996] for a thorough introduction on the topic. We here take an approach similar to Robert & Casella [2004] and Tierney [1994]. Loosely speaking a Markov chain is a series of random vectors in which every element depends on its very last predecessor only. This can be seen as a generalization of a first-order autoregressive process allowing for non-linear dependency functions and non-Gaussian noise. Before we proceed to the definition of Markov chains, we introduce the more general concept of random vectors evolving on some arbitrary non-empty index set I. In later applications $I \subseteq \mathbb{N}_0$ indexes a series of model parameters $\boldsymbol{\xi}^{(t)} \in \mathbb{R}^n$. Throughout this chapter $(\Omega, \mathcal{F}, P)$ denotes a probability space, and all random vectors are functions $\boldsymbol{X} : \Omega \longrightarrow E$ onto a measurable space $(E, \mathcal{E}) \subseteq (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$. Without loss of generality we furthermore assume that the density function for each random vector exists and that it is positive for any realization $\boldsymbol{x} \in E$ mentioned.

Definition 2.12 (Stochastic/random process)**.** Given a probability space $(\Omega, \mathcal{F}, P)$ together with the measurable space $(E, \mathcal{E})$, a *stochastic* (or *random*) *process* $\{\boldsymbol{X}^{(t)}\}_{t \in I}$ on some index set I is a function

$$\begin{aligned} \boldsymbol{X} : \Omega \times I &\longrightarrow E \\ (\boldsymbol{\omega}, t) &\longmapsto \boldsymbol{X}^{(t)}(\boldsymbol{\omega}), \end{aligned}$$

such that the functions

$$\begin{aligned} \boldsymbol{X}^{(t)} : \Omega &\longrightarrow E \\ \boldsymbol{\omega} &\longmapsto \boldsymbol{X}^{(t)}(\boldsymbol{\omega}), \end{aligned}$$

are $(\mathcal{F}, \mathcal{E})$-measurable, i.e. $\boldsymbol{X}^{(t)} : \Omega \longrightarrow E$ are random vectors living on the same probability space. If $I \subseteq \mathbb{R}$, we call $\{\boldsymbol{X}^{(t)}\}_{t \in I}$ a *time-continuous random process* and if $I \subseteq \mathbb{Z}$ (naturally including $I = \mathbb{N}$) a *time-discrete random process.*

For a finite subset $I^f \subseteq I$ the set $\{\boldsymbol{x}^{(t)}\}_{t\in I^f}$ is a *realization* or *sample path* of $\{\boldsymbol{X}^{(t)}\}_{t\in I}$, if for all $t \in I^f$, $\boldsymbol{x}^{(t)}$ is a realization of $\boldsymbol{X}^{(t)}$. In our applications the realizations $\{\boldsymbol{x}^{(t)}\}_{t\in I^f}$ consist of samples $\boldsymbol{x}^{(t)}$ of a posterior distribution dependent on some vector of observations.

Example 2.6 (Time-discrete random process.). Suppose $I = \mathbb{N}$ and $\{X^{(t)}\}_{t\in\mathbb{N}}$ are independent univariate random vectors with

$$P(X^{(t)} = +1) = P(X^{(t)} = -1) = \frac{1}{2}.$$

Setting $X^{(0)} = 0$ and $Y^{(k)} = \sum_{t=1}^{k} X^{(t)}$, the process $Y := \{Y^{(k)}\}_{k\in\mathbb{N}}$ defines a time-discrete (and even space-discrete) random process. A realization of $\{Y^{(k)}\}_{k\in\mathbb{N}}$ can be seen in Figure 2.3(a). For standard normally distributed $X^{(t)}$'s we get a time-discrete random process on a continuous sample space. A realization is given in Figure 2.3(b).

Example 2.7 (Time-continuous random process). A random process $\{\boldsymbol{W}^{(t)}\}_{t\in\mathbb{R}_0^+}$ with $\boldsymbol{W}^{(t)} = (W_1^{(t)}, \ldots, W_n^{(t)})^\top$ and

(a) $\boldsymbol{W}^{(0)} = \boldsymbol{0}$

(b) $\boldsymbol{W}^{(t)}$ is almost surely continuous and

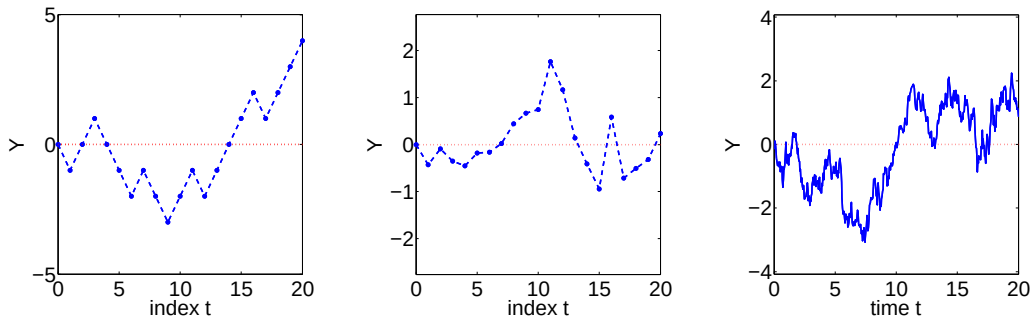
(c) $\boldsymbol{W}^{(t)}$ has independent increments with $W_j^{(t)} - W_j^{(s)} \sim \mathcal{N}(0, t-s)$ (for $j = 1, \ldots, n$ and $0 \leq s < t$)

is called a *Wiener process* or *standard Brownian motion*. This is a time- and space-continuous random process. A one dimensional realization is given in Figure 2.3(c).

By prerequisite (c) above each random vector $\boldsymbol{W}^{(t)}$ in a Wiener process depends on the history of all preceding random vectors $\boldsymbol{W}^{(s)}$ for $0 \leq s < t$. On the contrary, a random process for which the transition probability between different states in the state-space only depends on the current state is termed a *Markov process*. Mathematically this means that for the joint distributions $P_{\boldsymbol{X}^{(0)}\otimes\ldots\otimes\boldsymbol{X}^{(t+1)}}$ of $\boldsymbol{X}^{(0)}, \ldots, \boldsymbol{X}^{(t+1)}$ and $P_{\boldsymbol{X}^{(t)}\otimes\boldsymbol{X}^{(t+1)}}$ of $\boldsymbol{X}^{(t)}$ and $\boldsymbol{X}^{(t+1)}$ we have

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

for any measurable $A \subseteq E$ and $\boldsymbol{x}^{(0)}, \ldots, \boldsymbol{x}^{(t)} \in E$.



(a) (b) (c)

Figure 2.3: (a) Realization of a time-discrete and space-discrete random process. (b) Realization of a time-discrete and space-continuous random process. (c) Realization of a Wiener process.

Hence, while ignoring historic events, a Markov process predicts future events solely based on the current state. Note that the transition probabilities between two states $\boldsymbol{X}^{(t)}$ and $\boldsymbol{X}^{(t+1)}$ can depend on $t \in I$. As our realizations of Markov chain elements $\boldsymbol{x}^{(t)}$ later on are always drawn from the very same distribution for all $t \in I$, we define:

Definition 2.13 (Stationary process)**.** Let $I = \mathbb{R}/\mathbb{Z}/\mathbb{N}$; let furthermore $\{\boldsymbol{X}^{(t)}\}_{t\in I}$ be a stochastic process on I. We call $\{\boldsymbol{X}^{(t)}\}_{t\in I}$ *stationary*, if for all $\{t_1, \ldots, t_k\} \subseteq I$ and all $\tau \in I$ the joint distributions of

$$\boldsymbol{X}^{(t_{1+\tau})}, \ldots, \boldsymbol{X}^{(t_{k+\tau})} \quad \text{and} \quad \boldsymbol{X}^{(t_1)}, \ldots, \boldsymbol{X}^{(t_k)}$$

are equal, i.e.

$$P_{\boldsymbol{X}^{(t_{1+\tau})}\otimes\ldots\otimes\boldsymbol{X}^{(t_{k+\tau})}}(\boldsymbol{X}^{(t_{1+\tau})}, \ldots, \boldsymbol{X}^{(t_{k+\tau})}) = P_{\boldsymbol{X}^{(t_1)}\otimes\ldots\otimes\boldsymbol{X}^{(t_k)}}(\boldsymbol{X}^{(t_1)}, \ldots, \boldsymbol{X}^{(t_k)}).$$

This means that a shift in time has no effect on the joint statistics of any order. In other words, the joint distribution is time independent. Markov processes $\{\boldsymbol{X}^{(i)}\}_{i\in I}$ on the continuous index set $I = \mathbb{R}$ are commonly termed *diffusion processes*. Diffusion processes are an important modeling technique when dealing with single entity processes, such as molecule or protein interactions. Here, random effects often determine the modeling results and therefore call for a stochastic approach (see Dargatz [2010] or Øksendal [2003] for nice introductions). These are modeled by stochastic differential

equations of the form

$$\frac{\mathrm{d}\boldsymbol{X}^{(t)}}{\mathrm{d}t} = \boldsymbol{\mu}(\boldsymbol{X}^{(t)}, t) + \boldsymbol{\Sigma}(\boldsymbol{X}^{(t)}, t)\boldsymbol{\eta}^{(t)}, \tag{2.16}$$

where $\boldsymbol{\mu} : E \times I \longrightarrow \mathbb{R}^n$ and $\boldsymbol{\Sigma} : E \times I \longrightarrow \mathbb{R}^{n\times m}$ are jointly measurable functions for the n-dimensional random vectors $\boldsymbol{X}^{(t)}$ and m-dimensional Gaussian white noise $\boldsymbol{\eta} : I \longrightarrow \mathbb{R}^m$, such that $\boldsymbol{\eta}^{(t)} \overset{i.i.d.}{\sim} \mathcal{N}_m(0,1)\ \forall t \in I$. While the function $\boldsymbol{\mu}$ determines the so-called systematic (non-stochastic) drift, the $\boldsymbol{\Sigma}$-term regulates the strength of the diffusion. Equation (2.16) is also commonly written as

$$\mathrm{d}\boldsymbol{X}^{(t)} = \boldsymbol{\mu}(\boldsymbol{X}^{(t)}, t)\mathrm{d}t + \boldsymbol{\Sigma}(\boldsymbol{X}^{(t)}, t)\mathrm{d}\boldsymbol{W}^{(t)},$$

where $\mathrm{d}\boldsymbol{W}^{(t)} = \boldsymbol{\eta}^{(t)}\mathrm{d}t$ is the notation for standard Brownian motion (compare e.g. Dargatz [2010]). For a thorough introduction to Markov processes on continuous index sets the reader may be referred to Meyn *et al.* [1996]. In this work all measurements are taken at fixed time points. Therefore we will restrict ourselves to time-discrete Markov processes, called *Markov chains*.

Definition 2.14 (Markov chain)**.** Let $(\Omega, \mathcal{F}, P)$ be a probability space. A stochastic process $\{\boldsymbol{X}^{(t)}\}_{t\in I}$ with values in E is called a *Markov chain*, if the index set $I = \mathbb{N}_0$ and for any measurable set $A \subseteq E$, any $T \in I \setminus \{0, 1\}$, 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 $\boldsymbol{x}^{(0)}, \ldots, \boldsymbol{x}^{(T-1)}$, that is

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

Furthermore, a stationary Markov chain is also called (*time*) *homogeneous*. From this point on, we consider all Markov chains to be time homogeneous. In the next step, we want to simplify the notation of the joint distribution $P_{\boldsymbol{X}^{(0)}\otimes\ldots\otimes\boldsymbol{X}^{(t)}}$ for the random vectors $\boldsymbol{X}^{(0)}, \ldots, \boldsymbol{X}^{(t)}$: When dealing with Markov chains the joint probability $P_{\boldsymbol{X}^{(0)}\otimes\ldots\otimes\boldsymbol{X}^{(t+1)}}$ is often seen as transition probability from a state $\boldsymbol{X}^{(t)} = \boldsymbol{x}$ to some $A \subseteq E$. Put differently, it contains the distribution for $\boldsymbol{X}^{(t+1)}$ given $\boldsymbol{X}^{(t)} = \boldsymbol{x}$.

Definition 2.15 (Transition kernel)**.** For a Markov chain $\{\boldsymbol{X}^{(t)}\}_{t\in I}$ on a probability

space $(\Omega, \mathcal{F}, P)$ with values in E and a measurable set $A \subseteq E$ the distribution

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

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

The transition kernel is a time independent function

$$k : \mathcal{E} \times E \longrightarrow [0, 1],$$

where (c.f. Robert & Casella [2004])

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

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

Technically, it can be expressed via a function $p : E \times E \longrightarrow [0, \infty)$ as

$$k(\mathrm{d}\boldsymbol{y}|\boldsymbol{x}) = p(\boldsymbol{y}|\boldsymbol{x})\mathrm{d}\boldsymbol{y} + r(\boldsymbol{x})\mathbb{1}_{\boldsymbol{x}}(\mathrm{d}\boldsymbol{y}), \tag{2.17}$$

where $\mathbb{1}_{\boldsymbol{x}}(\mathrm{d}\boldsymbol{y})$ is the indicator function, $p(\boldsymbol{x}|\boldsymbol{x}) = 0$ and $r(\boldsymbol{x}) = 1 - \int_E p(\boldsymbol{y}|\boldsymbol{x})\mathrm{d}\boldsymbol{y}$ (c.f. Tierney [1994]). Here, the function p governs the transition from $\boldsymbol{x}$ to $\boldsymbol{y}$ while $r(\boldsymbol{x})$ holds the probability for $\boldsymbol{X}^{(t+1)}$ to remain at $\boldsymbol{X}^{(t)} = \boldsymbol{x}$. We will see later that in fact $r(\boldsymbol{x})$ needs to be positive for Markov Chain Monte Carlo methods.

Proposition 2.3. *Given an initial value* $\boldsymbol{x}^{(0)} \in E$*, the transition kernel* k *fully determines the respective Markov chain.*

Proof. For all $\boldsymbol{x}^{(0)} \in E$ and $A_i \in \mathcal{E}$

$$\begin{aligned} P_{\boldsymbol{X}^{(1)}|\boldsymbol{X}^{(0)}}(A_1|\boldsymbol{x}^{(0)}) &= k(A_1|\boldsymbol{x}^{(0)}) \\ P_{\boldsymbol{X}^{(1)}\otimes\boldsymbol{X}^{(2)}|\boldsymbol{X}^{(0)}}((\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}) \in A_1 \times A_2|\boldsymbol{x}^{(0)}) &= \int_{A_1} k(A_2|\boldsymbol{y}_1)k(\mathrm{d}\boldsymbol{y}_1|\boldsymbol{x}^{(0)}) \\ \vdots \qquad\qquad & \qquad\qquad \vdots \\ P_{\boldsymbol{X}^{(1)}\otimes\ldots\otimes\boldsymbol{X}^{(t)}|\boldsymbol{X}^{(0)}}((\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(t)}) \in A_1 \times \ldots \times A_t|\boldsymbol{x}^{(0)}) &= \int_{A_1} \ldots \int_{A_{t-1}} k(A_t|\boldsymbol{y}_{t-1}) \\ &\quad \times k(\mathrm{d}\boldsymbol{y}_{t-1}|\boldsymbol{y}_{t-2}) \ldots k(\mathrm{d}\boldsymbol{y}_1|\boldsymbol{x}^{(0)}). \end{aligned}$$

□

This allows us to not specify the different (conditional) probability distributions explicitly, but simply denote the distribution on any set $\boldsymbol{X}^{(t_1)}, \ldots, \boldsymbol{X}^{(t_s)}$ for $t_1, \ldots, t_s \subset I$ by P. For reasons of readability we will in the following make use of this notation. Note that the joint distribution $P = P_{\boldsymbol{X}^{(t_1)} \otimes \ldots \otimes \boldsymbol{X}^{(t_s)}}$ for $\boldsymbol{X}^{(t_1)}, \ldots, \boldsymbol{X}^{(t_s)}$ should not be confused with the probability measure P on the probability space $(\Omega, \mathcal{F}, P)$.

The primary goal of Markov Chain Monte Carlo (MCMC) methods lies in the inference of a distribution π by means of a Markov chain $\{\boldsymbol{X}^{(t)}\}_{t \in I}$. Towards this end we need to make sure that the chain essentially converges towards π, regardless of where it begins. The following definitions form the basis for proper MCMC methods: We call a distribution π *invariant* or *stationary* for the transition kernel $k(\cdot|\cdot)$, if

$$\begin{aligned} \pi(A) &= \int_E k(A|\boldsymbol{x}) \pi(\mathrm{d}\boldsymbol{x}) \\ &= \int_E k(A|\boldsymbol{x}) \pi_d(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}, \qquad \forall A \in \mathcal{E} \end{aligned} \tag{2.18}$$

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

A stationary Markov chain is *reversible*, if for $A \in \mathcal{E}$,

$$P(\boldsymbol{X}^{(t+1)} \in A | \boldsymbol{X}^{(t+2)} = \boldsymbol{x}) = P(\boldsymbol{X}^{(t+1)} \in A | \boldsymbol{X}^{(t)} = \boldsymbol{x}). \tag{2.19}$$

Reversibility essentially states that the direction of the evolution on I does not influence the dynamics of the chain. A sufficient condition for invariance and reversibility is given by:

Definition 2.16 (Detailed balance condition)**.** 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 introduced in (2.17) satisfies the *detailed balance condition*, if there exists a probability density function π_d, such that

$$p(\boldsymbol{x}|\boldsymbol{y})\pi_d(\boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{x})\pi_d(\boldsymbol{x}). \tag{2.20}$$

Theorem 2.4. *If 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.

Proof. Part (i) follows directly by checking Equation (2.18):

$$\begin{aligned}\int_E \pi_d(\boldsymbol{x})k(A|\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} &= \int_E\int_A \pi_d(\boldsymbol{x})p(\boldsymbol{y}|\boldsymbol{x})\,\mathrm{d}\boldsymbol{y}\,\mathrm{d}\boldsymbol{x} + \int_A \pi_d(\boldsymbol{x})r(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\\ &= \int_A\int_E \pi_d(\boldsymbol{x})p(\boldsymbol{y}|\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y} + \int_A \pi_d(\boldsymbol{x})r(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\\ &= \int_A \pi_d(\boldsymbol{y})\,(1-r(\boldsymbol{y})) + \pi_d(\boldsymbol{y})r(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}\\ &= \int_A \pi(\mathrm{d}\boldsymbol{y}).\end{aligned}$$

Part (ii): As π is invariant with respect to k it follows that if $\boldsymbol{X}^{(0)} \sim \pi$, then $\boldsymbol{X}^{(t)} \sim \pi\ \forall t \in I$. Together with Bayes' theorem this yields

$$\begin{aligned}p(\boldsymbol{y}|\boldsymbol{x}) + r(\boldsymbol{x})\mathbb{1}_{\boldsymbol{x}}(\boldsymbol{y}) &= \frac{(p(\boldsymbol{x}|\boldsymbol{y}) + r(\boldsymbol{y})\mathbb{1}_{\boldsymbol{y}}(\boldsymbol{x}))\cdot \pi_d(\boldsymbol{y})}{\pi_d(\boldsymbol{x})}\\ &= \frac{(p(\boldsymbol{y}|\boldsymbol{x}) + r(\boldsymbol{y})\mathbb{1}_{\boldsymbol{y}}(\boldsymbol{x}))\cdot \pi_d(\boldsymbol{x})}{\pi_d(\boldsymbol{x})}\\ &= p(\boldsymbol{y}|\boldsymbol{x}) + r(\boldsymbol{y})\mathbb{1}_{\boldsymbol{y}}(\boldsymbol{x}).\end{aligned}$$

Hence, for $A \in \mathcal{E}$

$$\begin{aligned}P(\boldsymbol{X}^{(t+1)} \in A|\boldsymbol{X}^{(t)} = \boldsymbol{x}) &= \int_A p(\boldsymbol{y}|\boldsymbol{x}) + r(\boldsymbol{x})\mathbb{1}_{\boldsymbol{y}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{y}\\ &= \int_A p(\boldsymbol{y}|\boldsymbol{x})\,\mathrm{d}\boldsymbol{y}\\ &= \int_A p(\boldsymbol{x}|\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}\\ &= P(\boldsymbol{X}^{(t+1)} \in A|\boldsymbol{X}^{(t+2)} = \boldsymbol{x}).\end{aligned}$$

□

We have to point out that the detailed balance condition is sufficient but not necessary for the existence of an invariant distribution π. However, its simplicity makes it easy to check and it is therefore frequently assumed in most MCMC algorithms. Up to now we laid grounds for the existence of a reversible invariant distribution π. This invariant distribution might nonetheless be non-unique. If every Markov chain governed by the transition kernel k is converging to the same invariant distribution π, independent of the starting value $\boldsymbol{x}^{(0)} \in E$, we call π an *equilibrium distribution*. In terms of the

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

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

for almost all $\boldsymbol{x}^{(0)} \in E$. Naturally, $k^1(A|\boldsymbol{x}) := k(A|\boldsymbol{x})$. We need some more definitions to characterize equilibrium distributions. Luckily these are easy to prove for most practical applications.

A Markov chain $\{\boldsymbol{X}^{(t)}\}_{t\in I}$ with transition kernel k is called π-*irreducible* for a σ-finite π, if for any $\boldsymbol{x} \in E$ and $A \in \mathcal{E}$ with $\pi(A) > 0$ there exists an $m \in \mathbb{N}$ such that

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

Here, $k^m(A|\boldsymbol{x})$ denotes the associated *m-step transition* kernel. This means the Markov chain can get from any state $\boldsymbol{x} \in E$ to any other state in E within a finite number of steps. If $m = 1$, we call the chain *strongly* π-*irreducible*.

A π-irreducible Markov chain with transition kernel k is *periodic*, if for some integer $s \geq 2$ there exists a sequence $(A_0, A_1, \ldots, A_{s-1})$ of pairwise disjoint non-empty subsets $A_i \in \mathcal{E}$, such that for all $i = 0, \ldots, s-1$ and all $\boldsymbol{x} \in A_i$

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

We call the chain *aperiodic*, if it is not periodic. Frankly spoken, aperiodic Markov chains do not contain deterministic cycles.

Suppose $P_{\boldsymbol{x}}(A)$ reflects the probability that, starting at $\boldsymbol{x} \in E$, we obtain for the number $c_A^{(t)} := \mathrm{card}\{\boldsymbol{x}^{(s)} \in A | 0 \leq s \leq t\}$ of visits to some subset $A \in \mathcal{E}$ up to t, that $c_A^{(t)} \to \infty$ for $t \to \infty$. A Markov chain is *Harris recurrent*, if there exists an invariant distribution π, such that for every $A \in \mathcal{E}$ with $\pi(A) > 0$

$$P_{\boldsymbol{x}}(A) = 1 \qquad \text{for all } \boldsymbol{x} \in E.$$

As pointed out in Tierney [1994], it follows from Corollary 5.2 of Nummelin [2004] that there exists an invariant measure ν on E for every π-irreducible Harris recurrent Markov chain. The measure ν is unique up to a multiplicative constant. The chain is called *positive Harris recurrent*, if $\nu(E) < \infty$. We want to point out that there exists

also the somewhat weaker notion of a *recurrent* – as opposed to Harris recurrent – Markov chain. However, as all MCMC methods introduced in Chapter 4 start at some arbitrary random vector $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$, we have to make sure that the algorithms converge to a unique invariant distribution independent of $\boldsymbol{x}^{(0)}$. This is shown in Theorem 2.5 for Harris recurrent chains claiming some weak assumptions. It summarizes the necessary and sufficient conditions for the convergence of a Markov chain to an equilibrium distribution (Tierney [1994]). For a thorough proof see Sethuraman *et al.* [1992]. In the case of recurrent chains the theorem only holds for almost all $\boldsymbol{x}^{(0)} \in E$.

Theorem 2.5. *Suppose $\{\boldsymbol{X}^{(t)}\}_{t\in I}$ is a π-irreducible, aperiodic and Harris recurrent Markov chain with transition kernel k and invariant distribution π. Then*

(i) k is positive Harris recurrent,

(ii) π is the (unique) equilibrium distribution and

(iii) k is ergodic for π, i.e. $\{\boldsymbol{X}^{(t)}\}_{t\in I}$ converges regardless of its starting value $\boldsymbol{x}^{(0)} \in E$, i.e. for every $\boldsymbol{x} \in E$ and every $A \in \mathcal{E}$

$$\|k^m(A|\boldsymbol{x}) - \pi(A)\|_{TV} \to 0 \qquad \text{for } m \to \infty.$$

Hence, we later on only need to test for the existence of an invariant distribution π, along with π-irreducibility, aperiodicity, and Harris recurrence in order to establish a valid MCMC method. A corollary to Theorem 3.6 in Chapter 4 of Revuz [1984] in combination with Corollary 1 of Tierney [1994] describes the limiting behavior of averages. It states a law of large numbers and can be derived from the Chacon-Ornstein or ergodic theorem (c.f. Tierney [1994]).

Theorem 2.6. *Suppose $\{\boldsymbol{X}^{(t)}\}_{t\in I}$ is a positive Harris recurrent aperiodic Markov chain with invariant distribution π. Suppose furthermore $f : E \longrightarrow \mathbb{R}$ is π-integrable with $\int_E |f(\boldsymbol{x})|\pi(\mathrm{d}\boldsymbol{x}) < \infty$. Then for a realization $\{\boldsymbol{x}^{(t)}\}_{t\in I}$ the sample mean*

$$\bar{f}_m = \frac{1}{m+1}\sum_{t=0}^{m} f(\boldsymbol{x}^{(t)}) \longrightarrow \int_E f(\boldsymbol{x})\pi(\mathrm{d}\boldsymbol{x}) = \mathbb{E}_\pi\left[f(E)\right] \qquad \text{almost surely as } m \to \infty.$$

Example 2.8 (Discrete state space). For a discrete state space $E = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_S\}$ the transition kernel $k(\boldsymbol{X}^{(t+1)} = \{\boldsymbol{x}_j\}|\boldsymbol{X}^{(t)} = \boldsymbol{x}_i) := k(\boldsymbol{x}_j|\boldsymbol{x}_i) := k_{i,j}$ can be written in

matrix notation as

$$\boldsymbol{K} = \begin{pmatrix} k(\boldsymbol{x}_1|\boldsymbol{x}_1) & \dots & k(\boldsymbol{x}_S|\boldsymbol{x}_1) \\ \vdots & \ddots & \vdots \\ k(\boldsymbol{x}_1|\boldsymbol{x}_S) & \dots & k(\boldsymbol{x}_S|\boldsymbol{x}_S) \end{pmatrix}.$$

The matrix $\boldsymbol{K}$ is right stochastic, i.e. $\sum_{j=1}^{S} k_{i,j} = 1$. Every starting distribution can be written as vector $\boldsymbol{\pi}^{(0)} \in [0,1]^S$. Suppose the Markov chain starts at $\boldsymbol{x}^{(0)} = \boldsymbol{x}_j$ for some $j \in \{1, \dots, S\}$, then $\pi_j^{(0)} = 1$ while $\pi_i^{(0)} = 0$ for $i \neq j$. The probability for moving to state $\boldsymbol{x}_s$ at $t = 1$ is given by

$$\boldsymbol{\pi}^{(1)} = \boldsymbol{\pi}^{(0)} \boldsymbol{K}.$$

Applying the m-step transition kernel $\boldsymbol{K}^m =: (k_{i,j}^{[m]})_{i,j=1,\dots,S}$, i.e. $k_{i,j}^{[m]} = P(\boldsymbol{X}^{(m+t)} = \boldsymbol{x}_j | \boldsymbol{X}^{(t)} = \boldsymbol{x}_i)$ we iteratively obtain

$$\boldsymbol{\pi}^{(t)} = \boldsymbol{\pi}^{(s)} \boldsymbol{K}^{t-s} \tag{2.21}$$

for $1 \leq s < t$. Equation (2.21) is known as the discrete Chapman-Kolmogorov equation. A continuous version is given in Lemma 4.1. It essentially states that a move from state $\boldsymbol{x}^{(t)}$ to $\boldsymbol{x}^{(t+2)}$ passes through any of the states $\boldsymbol{x}_1, \dots, \boldsymbol{x}_S$ with the respective probability. The Markov chain $\{\boldsymbol{X}^{(t)}\}_{t \in I}$ is $\boldsymbol{\pi}$-invariant, if $\boldsymbol{\pi} = \boldsymbol{\pi}\boldsymbol{K}$. Thus, $\boldsymbol{\pi}$ needs to be a left-eigenvector for the eigenvalue 1 of $\boldsymbol{K}$. For $\tilde{k}_{i,j} = P(\boldsymbol{X}^{(t+1)} = \boldsymbol{x}_j | \boldsymbol{X}^{(t+2)} = \boldsymbol{x}_i)$ and a stationary distribution $\boldsymbol{\pi} = (\pi_1, \dots, \pi_S)^\top$ Bayes' theorem yields

$$\tilde{k}_{i,j} = \frac{\tilde{k}_{j,i} \cdot \pi_j}{\pi_i} = \frac{\pi_j}{\pi_i} k_{j,i}. \tag{2.22}$$

Hence, the detailed balance condition emerges naturally by Equation (2.22) for a reversible chain in the discrete case. The Markov chain is irreducible, if every state can be reached from any other state within a certain number of steps, i.e if for all $i, j \in \{1, \dots, S\}$ there exists a natural number m_{ij} with $k_{i,j}^{[m_{ij}]} > 0$. It is aperiodic, if for all $m > 0$ $k_{i,i}^{[m]} \neq 1$ for all $i \in \{1, \dots, S\}$. For characterizing recurrence we finally need the *hitting times* $T_i = \inf\{t \geq 1 | \boldsymbol{X}^{(t)} = \boldsymbol{x}_i \text{ given } \boldsymbol{X}^{(0)} = \boldsymbol{x}_i\}$ for $i \in \{1, \dots, S\}$. The chain is recurrent, if the probability $Pr(T_i < \infty) = 1 \ \forall i$. It is positive (Harris) recurrent, if the expected value of the hitting time is finite.

2.4 A short introduction on molecular biology

It is an amazing fact that almost every cell in a living organism contains a full blueprint for the development and functionality of the entire organism. This information is stored

in the so-called Deoxyribonucleic acid (DNA) within the cell nucleus. A *gene* is a short sequence of this DNA strain holding information for the construction of a *protein*. While there are approximately 20,000-25,000 genes in the human genome (i.e. the entity of all genes in an organism), plants often times endow more than twice this number. The pure amount of genes is therefore no indicator for the complexity of a life form. An organism is in fact regulated by a complex network of protein interactions.

For a gene to code for a protein the processes of *transcription* and *translation* need to take place. In transcription RNA Polymerase enzymes (proteins) along with various *transcription factor* proteins identify a specific gene on the DNA strain. This gene then gets synthesized and transformed into Ribonucleic acid (RNA), the basis for protein construction. There are various types of RNA molecules that play an important role in the actual protein building process, which itself takes place in the cytoplasm during the translation step: Mitochondrial ribosomes consisting *inter alia* of ribosomal RNAs (rRNAs) convert the information stored on messenger RNAs (mRNAs) into a protein. This protein consist of different amino acids that are transported to the ribosomes by transfer RNAs (tRNAs). Moreover, short RNA sequences called silencing RNAs (siRNAs) and micro RNAs (miRNAs) control the protein coding mechanism. In addition, there are also non-coding RNAs (ncRNAs). Although these do not contain information for the translation process, it is believed that they control processes, such as gene regulation. Much of their functionality has however not been inferred yet.

Proteins, protein complexes, or peptides (i.e. short amino acid sequences generally built from larger precursor proteins) regulate the majority of cellular processes. These comprise amongst others structural proteins used for all rigid components of the cell (such as cell membranes), transcription factors that control the transcription process, enzymes responsible for regulating the metabolism and the *activation* of proteins by transferring phosphate groups (such as kinases), or growth factors (such as cytokines or hormones) that trigger proliferation and cellular growth. For a thorough introduction to cellular design and functionality the reader may be referred to Alberts *et al.* [2002].

Although the *Human Genome project* deciphered the entire human genome in 2003, we are yet far from fully understanding its mechanistic interplay. The dynamics of all of the underlying mechanisms are highly complex. They can be and are classified and studied as dynamical systems (see Chapter 2.5).

2.4.1 Signaling pathways

We saw above that proteins are essential entities for the functionality of an organism, but how does a cell know which proteins to express when? This is mostly regulated by cellular signaling. On the molecular level these signals are again mainly mediated by proteins, small peptides, single amino acids, or lipids. Specifically, intercellular signals are either transmitted by direct cell-to-cell exchange of molecules or by the secretion of molecules from the signaling cell; these in turn provoke a reaction on the surface receptor proteins of a receiving cell. In the latter mechanism a signaling protein binds to an extracellular receptor and induces a series of biochemical reactions that transport the information through the cell membrane. Within the receiving cell a receptor associated kinase or kinase domain is thereupon activated. This induces the activation of diverse intracellular proteins or other signaling molecules which finally transmit the signal to a specific cellular compartment, such as the nucleus. Generally, the alteration of various proteins by a series of phosphorylation and dephosphorylation steps is responsible for signal maintenance (Kowarsch [2011]). In the nucleus the transmitted signal controls processes like transcription. These are in turn responsible for cell growth, differentiation, apoptosis, or protein synthesis to name just a few. The mechanism of transmitting an extracellular signal to a cellular compartment for the induction of a specific response is called *cellular signaling pathway* or simply *signaling pathway.* In summary, cellular signaling pathways are processing and transmitting intercellular signals in order to control cellular processes.

2.4.2 The JAK-STAT pathway

An important representative for signaling pathways in mammals is the so-called JAK-STAT pathway (JAK stands for Janus Kinase, and STAT for Signal Transducer and Activator of Transcription). It is utilized by more than 50 different cytokines, hormones, and other growth factors and plays a key role in gene regulation (Subramaniam *et al.* [2001]). Scientifically it is therefore of major interest. Malfunctioning results in diseases like leukemia or bronchial asthma (Igaz *et al.* [2001]). In the JAK-STAT pathway a cellular transmembrane receptor is triggered by different molecules of the cytokine or growth factor families. Examples include the epidermal growth factor (EGF),

erythropoietin (EPO), interferones (INFα, INFβ, INFγ), and Interleukin-6 (IL-6). Cytoplasmic STAT proteins are inactive in unstimulated cells. Upon receptor activation the receptor-bound JAK proteins catalyze auto-phosphorylation and build tyrosine residues for tyrosine-phosphorylation of STAT proteins between STAT Src-homology 2 (SH2) domains and the tyrosine residues (Aaronson & Horvath [2002]). The tyrosine-activated STAT proteins homo- and heterodimerize and get rapidly transported into the nucleus subsequent to a possible second serine phosphorylation step of the dimer (Wen *et al.* [1995]). In the nucleus the activated STAT dimer dramatically upregulates the transcription rate of the target promoter. After the transcription process inactive, i.e. unphosphorylated, STAT is released back into the cytoplasm. Four evolutionarily conserved JAK proteins (JAK1, JAK2, JAK3, TYK2) and seven STAT coding genes with corresponding proteins (STAT1, STAT2, STAT3, STAT4, STAT5A, STAT5B, STAT6) are known for mammals (Aaronson & Horvath [2002]). These however only work in specific combinations coordinated by SH2-phosphotyrosine interactions.

In Chapter 6.3.4, we use a delay differential equation model of the JAK2-STAT5 pathway in order to evaluate the performance of the copula based Metropolis-Hastings algorithms introduced in this thesis. In Chapter 7 model inference helps to address the question, whether tyrosine phosphorylated STAT3 can work as transcription factor in the JAK1-STAT3 pathway.

2.5 Dynamical systems in molecular biology

Understanding the mechanisms of cellular functionality is a key challenge in the field of systems biology. In recent years much effort has gone into the inference of gene regulatory, metabolic and signaling networks, which *inter alia* govern gene expression, cellular communication, or intra cellular molecular transfer (De Jong [2002]; Palsson [2006]). All of these processes may be modeled by a system of *biochemical reactions* of the form

$$\begin{aligned} R_1 &: a_{11}A_{11} + \ldots + a_{1r_1}A_{1r_1} \longrightarrow b_{11}B_{11} + \ldots + b_{1s_1}B_{1s_1} \\ &\vdots \\ R_m &: a_{m1}A_{m1} + \ldots + a_{mr_m}A_{mr_m} \longrightarrow b_{m1}B_{m1} + \ldots + b_{ms_m}B_{ms_m} \end{aligned} \tag{2.23}$$

where the *reactants* $A_{i1}, \ldots, A_{ir_i}$ are transformed into the *products* $B_{i1}, \ldots, B_{is_i}$. The natural numbers $a_{i1}, \ldots, a_{ir_i}$ and $b_{i1}, \ldots, b_{is_i}$ hold the number of reactants and products involved in the reactions (2.23). Conventionally for $i = 1, \ldots, m$ the greatest common divisor of these quantities is equal to one. Biochemical reactions solely control cellular activity. Generally, each reaction R_i obeys the *law of mass action*, which states that the probability of R_i to occur is proportional to the product of the concentrations of all reactants.

Example 2.9 (Elementary biochemical reactions). Wilkinson [2006] analyzes a simple auto-regulatory gene network in prokaryotes: A *protein* P is coded for by a *gene* g, i.e. g is transcribed into the *transcript* r, which is subsequently translated into the protein P. After translation the protein P builds a *protein complex* P_2 consisting of two copies of P. This *homodimer* P_2 finally inhibits the transcription of gene g. The network is based on the interaction of the following biochemical reactions:

$$\begin{aligned}
\text{transcription:} \quad & g \longrightarrow g + r \\
\text{mRNA translation:} \quad & r \longrightarrow r + P \\
\text{repression:} \quad & g + P_2 \longleftrightarrow g \cdot P_2 \\
\text{dimerization:} \quad & 2P \longleftrightarrow P_2 \\
\text{mRNA degradation:} \quad & r \longrightarrow \emptyset \\
\text{protein degradation:} \quad & P \longrightarrow \emptyset
\end{aligned}$$

Here, products connected by a dot represent a gene-protein complex. The empty set $\emptyset$ indicates that the reactant on the left hand side of the reaction is degraded. Reactions with a double sided arrow are *reversible*, which means that the right hand side of the equation can act as reactant producing the left hand side as product as well.

There are various approaches for modeling the dynamics of the reactions (2.23) over time. Gillespie [2007] presented a nice review on the topic. We shortly summarize its key aspects in the following.

Biochemical reactions occur, when a molecule transforms itself to another isomeric form or two or more molecules form a molecular complex. While the first scenario is governed by quantum mechanics, the latter depends on the chance of these molecules to come within a certain distance to each other. The dynamics of molecular systems thus exhibit some stochasticity. Let us assume here that the system is well-stirred and has a constant volume and temperature throughout the modeling process. These

assumption guarantee that the positions and velocities of the individual molecules have no effect on the system's dynamics. As bottom line we want to estimate the state vector $\boldsymbol{X}^{(t)} = (X_1^{(t)}, \ldots, X_d^{(t)})^\top$ based on some initial configuration $\boldsymbol{X}^{(t_0)}$, where $X_i^{(t)}$ denotes the number of molecules of species i at time point t and d is the number of species present. Each reaction R_j ($j \in \{1, \ldots, m\}$) is then characterized by

(i) a *state change vector* $\boldsymbol{v}_j = (v_{1,j}, \ldots v_{d,j})^\top$ and

(ii) a *propensity function* $a_j(\cdot)$.

The elements $v_{i,j}$ hold the change in the copy number of species i, if reaction R_j occurs, i.e. reaction R_j updates the current configuration $\boldsymbol{X}^{(t)} = \boldsymbol{x}$ to $\boldsymbol{x} + \boldsymbol{v}_j$. On the other hand, $a_j(\boldsymbol{x})\mathrm{d}t$ gives the probability that reaction R_j occurs in the infinitesimal time interval $[t, t + \mathrm{d}t)$ while the system is in the configuration $\boldsymbol{x}$. For instance, let us consider a unimolecular reaction $R_j : A_{j,i} \longrightarrow B$, where $A_{j,i}$ represents a molecule of species i (reactant) and B is some product. Then, due to the laws of physics, there exists a *rate constants* (also called *reaction rate*) k_j, such that the probability for any of the molecules of type $A_{j,i}$ to react in the infinitesimal time interval $[t, t + \mathrm{d}t)$ is given by $k_j\mathrm{d}t$. Hence, the propensity function reads $a_j(\boldsymbol{x}) = k_j x_i$ ($i \in \{1, \ldots, d\}$) for the configuration $\boldsymbol{X}^{(t)} = \boldsymbol{x} = (x_1, \ldots, x_d)^\top$. Similarly, for a bimolecular reaction $R_{j'} : A_{j',i_1} + A_{j',i_2} \longrightarrow B'$ there exists a rate constant $k_{j'}$, such that $a_{j'}(\boldsymbol{x}) = k_{j'} x_{i_1} x_{i_2}$ ($i_1 \neq i_2, i_1, i_2 \in \{1, \ldots, d\}$). In the case $i_1 = i_2$ we instead have $a_{j'}(\boldsymbol{x}) = \frac{1}{2} k_{j'} x_{i_1}(x_{i_1} - 1)$. Since all biochemical reactions can be built using a combination of uni- and bimolecular reactions, we do not consider higher order reactions for now.

The propensity functions $a_j(\boldsymbol{x})$ ($j = 1, \ldots, m$) allow to describe the evolution over time of the probability $P(\boldsymbol{x}, t|\boldsymbol{x}_0, t_0)$ that the system is at configuration $\boldsymbol{x}$ at time point t, given it started in $\boldsymbol{x}_0$ at t_0. The result (see e.g. Gillespie [1992] for a derivation) is the so-called *chemical master equation*

$$\frac{P(\boldsymbol{x}, t|\boldsymbol{x}_0, t_0)}{\mathrm{d}t} = \sum_{j=1}^{m} \left(a_j(\boldsymbol{x} - \boldsymbol{v}_j) P(\boldsymbol{x} - \boldsymbol{v}_j, t|\boldsymbol{x}_0, t_0) - a_j(\boldsymbol{x}) P(\boldsymbol{x}, t|\boldsymbol{x}_0, t_0)\right). \tag{2.24}$$

Equation (2.24) completely determines $P(\boldsymbol{x}, t|\boldsymbol{x}_0, t_0)$. Unfortunately it can only be solved analytically in the most simplest scenarios and even a computational approximation is often times too costly in larger systems.

Algorithm 1: The stochastic simulation algorithm.

Input: System configuration $\boldsymbol{x}_0$ at time t_0, final simulation time $t_f > t_0$, state change vectors $\boldsymbol{v}_1 \ldots, \boldsymbol{v}_m$, and propensity functions $a_1(\cdot), \ldots, a_m(\cdot)$.

Output: Realization of the process $\boldsymbol{X}^{(t)}$ as finite series of pairs $\{(\boldsymbol{x}_0, t_0), (\boldsymbol{x}_1, t_1), (\boldsymbol{x}_2, t_2), \ldots\}$.

Initialize $k \leftarrow 0$
while $t_k < t_f$ **do**
  for $i \leftarrow 1$ **to** m **do**
    $b_i \leftarrow a_i(\boldsymbol{x}_k)$
  Set $b_0 \leftarrow \sum_{i=1}^{m} b_i$.
  Sample $r \sim \mathcal{U}[0,1]$ and set $\tau \leftarrow -b_0^{-1} \log(r)$.
  Sample $s \sim \mathcal{U}[0,1]$ and set $j \leftarrow$ smallest l such that $\sum_{l'=1}^{l} b_{l'} > s b_0$.
  Update $k \leftarrow k + 1$.
  Set $t_k \leftarrow t_{k-1} + \tau$ and $\boldsymbol{x}_k \leftarrow \boldsymbol{x}_{k-1} + \boldsymbol{v}_j$.

Instead of solving the chemical master equation explicitly, we can also try to simulate a realization of $\boldsymbol{X}^{(t)}$ over time. The key to this approach is to consider the probability $P(\tau, j|\boldsymbol{x}, t)\mathrm{d}\tau$ of the reaction R_j to occur as next reaction in the infinitesimal time interval $[t + \tau, t + \tau + \mathrm{d}\tau)$, given that the system is in configuration $\boldsymbol{X}^{(t)} = \boldsymbol{x}$ at time point t. Here, according to Gillespie [1992], we have

$$P(\tau, j|\boldsymbol{x}, t) = a_j(\boldsymbol{x}) \exp\left(- \left(\sum_{j'=1}^{m} a_{j'}(\boldsymbol{x}) \right) \tau \right). \tag{2.25}$$

This forms the basis for the *stochastic simulation algorithm* depicted in Algorithm 1. Extensions to Algorithm 1 e.g. for the case of large m and d were given by Gibson & Bruck [2000] or Cao *et al.* [2004].

We now want to approximate the sometimes computationally expensive outcome of the stochastic simulation algorithm by means of a stochastic differential equation. For this we assume that for some $\tau > 0$ all propensity functions $a_1(\cdot), \ldots, a_m(\cdot)$ are close to constant on the time interval $[t, t + \tau)$. Then the number of reactions R_j within $[t, t + \tau)$ is Poisson distributed with mean $a_j(\boldsymbol{x})\tau$. We obtain the discrete update rule

$$\boldsymbol{X}(t + \tau) \approx \boldsymbol{x} + \sum_{j=1}^{m} \eta_j \boldsymbol{v}_j$$

for the configuration $\boldsymbol{X}^{(t)} = \boldsymbol{x}$ and m independent Poisson distributed random variables η_j, $j = 1, \ldots, m$, with according means $a_j(\boldsymbol{x})\tau$. Assuming furthermore that for all

$j = 1, \ldots, m$, $a_j(\boldsymbol{x})\tau \gg 1$, we can approximate the Poisson distributions using the m normal distributions $\mathcal{N}(a_j(\boldsymbol{x})\tau, a_j(\boldsymbol{x})\tau)$ with means $a_j(\boldsymbol{x})\tau$ and variances $a_j(\boldsymbol{x})\tau$ $(j = 1, \ldots, m)$. For small τ's this leads to the *chemical Langevin equation*

$$\frac{\mathrm{d}\boldsymbol{X}^{(t)}}{\mathrm{d}t} = \sum_{j=1}^{m} a_j(\boldsymbol{X}^{(t)})\boldsymbol{v}_j + \sum_{j=1}^{m} \sqrt{a_j(\boldsymbol{X}^{(t)})}\boldsymbol{\eta}_j^{(t)}, \tag{2.26}$$

where $\boldsymbol{\eta}_j^{(t)}$ denotes some independent Gaussian white noise process (Dargatz [2010]). For a thorough introduction to stochastic differential equations see for instance Øksendal [2003]. Equation (2.26) can now be used to speed up the simulation of $\boldsymbol{X}^{(t)}$ over time.

In the thermodynamic limit, i.e. if the copy number of species i and the volume simultaneously approach infinity while their quotient stays constant, the second term on the right hand side of Equation (2.26) becomes negligible compared to the first term (Gillespie [1992]). Hence, if the substances involved in the biochemical reactions have numerous copy numbers and the cellular volume is large compared to the sizes of its molecules chemical kinetics can be modeled by sets of ordinary differential equations. Allowing additionally time delays for transcription, translation or diffusion processes (c.f. De Jong [2002]) we end up with a set of delay differential equations. In both cases we are dealing with continuous vectors of concentrations $\boldsymbol{x}(t) = (x_1(t), \ldots, x_d(t))^\top \geq \boldsymbol{0} \in \mathbb{R}^d$ of biochemical substances $x_1(t), \ldots, x_d(t)$ within a given time interval $[0, T]$. These substances can be proteins, RNA's, small molecules, and the like. As seen above, the rate of change for their concentrations is given by a set of differential equations

$$\frac{\mathrm{d}\boldsymbol{x}(t)}{\mathrm{d}t} = \boldsymbol{g}_{\boldsymbol{\xi}}(x_1(t), \ldots, x_d(t), x_1(t-\tau_1), \ldots, x_d(t-\tau_d), \boldsymbol{u}(t), t), \tag{2.27}$$

linking the solution $\boldsymbol{x}(t)$ via a $\boldsymbol{\xi}$-parametrized (nonlinear) Lipschitz-continuous function $\boldsymbol{g}_{\boldsymbol{\xi}} : \mathbb{R}_+^{2d} \times \mathbb{R}^{k+1} \longrightarrow \mathbb{R}^{2d}$ to the derivative of $\boldsymbol{x}(t)$ with respect to time t. For readability we generally omit the dependence of $\boldsymbol{x}(t)$ on the parameter vector $\boldsymbol{\xi} \in \mathbb{R}^n$. The latter can contain reaction rates, initial values to (2.27), but also noise parameters of the measurements or further constants as will become clear later. While $\boldsymbol{u}(t) \in \mathbb{R}^k$ represents an input vector of externally-supplied energy or input signals, the constants $\tau_1, \ldots, \tau_d$ denote discrete time delays. For $\tau_1 = \ldots = \tau_d = 0$ we call (2.27) a system of (nonlinear) *ordinary differential equations* (ODE's). Otherwise (2.27) is denoted as a system of (nonlinear) *delay differential equations* (DDE's). We also write

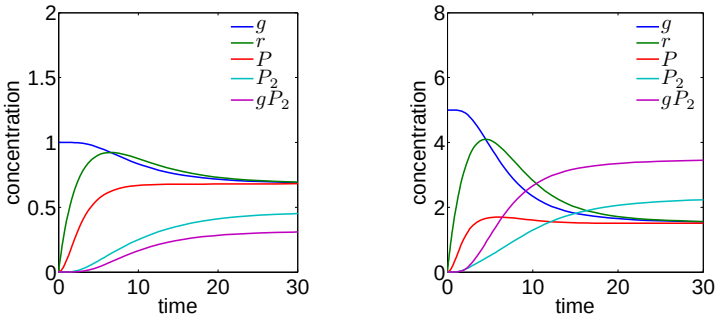
$$\frac{\mathrm{d}\boldsymbol{x}(t)}{\mathrm{d}t} = \boldsymbol{g}_{\boldsymbol{\xi}}(\boldsymbol{x}(t), \boldsymbol{\tau}, \boldsymbol{u}(t), t), \tag{2.28}$$

for the system (2.27), where $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_d)^\top$ denotes the vector of time delays. Examples for systems of differential equations in computational biology are the modeling of mRNA synthesis (Goodwin [1963]) or the modeling of cell cycles in *Caulobacter crescentus* (Li *et al.* [2008]) and yeast (Chen *et al.* [2004]).

Remark 2.1. We have to point out that there are other approaches for modeling the dynamics of cellular processes, too. For instance, especially in large systems we are often given no particular information about the number of reactants and products involved, i.e. we only have qualitative information such as "a gene is expressed at time point t". This gives a qualitative view on the system's behavior which is frequently modeled by a so-called *Boolean network*. The dynamics are given by simple logic-driven recombinations of binary ON/OFF states for the substances involved (c.f. Bornholdt [2008]; Kauffman [1969]; Thomas [1991] for an introduction). As Boolean networks do not contain information about molecular copy-numbers or concentrations they only provide a rather crude view on the system's evolution. This modeling technique might be essential if the number of species is very large. However, in general it is avoided in smaller systems.

In the following we will only deal with ordinary and delay differential equation models, which is what we call a *dynamical system* in this thesis. From the derivation of these systems above we saw that a large abundance of entities is necessary throughout the modeling process in order to justify the approach. This also implies that rate constants are constant at all times. We therefore consider the model parameters to be time-independent. We furthermore assume that the system is well-stirred and external influences such as the temperature or the osmotic pressure are constant throughout the modeling process. This allows us to ignore any spacial constraints that would e.g. call for models involving hard-to-handle partial differential equations. Raia *et al.* [2011] showed that the number of STAT5 and STAT6 molecules contained in human lymphoma cells (L1236) is $\sim 2 \cdot 10^5$ each. For the JAK-STAT pathways of Chapters 6.3.4 and 7 the assumption of being close to the thermodynamic limit is hence well justified and we can in fact approximate the dynamics by applying a differential equation model.

Example 2.10 (ODE representation of biochemical reactions). According to the law of



(a) (b)

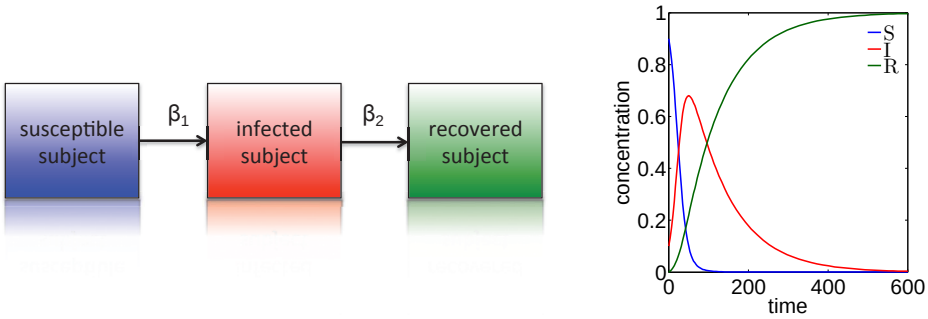
Figure 2.4: Time courses for Equation (2.29). (a) All rate constants were set to 0.5. The initial concentrations were $g(0) = 1$ and $r(0) = P(0) = P_2(0) = gP_2(0) = 0$. (b) Again, all rate constants were set to 0.5. The initial concentrations were $g(0) = 5$ and $r(0) = P(0) = P_2(0) = gP_2(0) = 0$. Both systems are close to individual steady states at time point $t = 30$.

mass action the corresponding ODE-system to Example 2.9 is given by

$$\begin{aligned}
\frac{\mathrm{d}g(t)}{\mathrm{d}t} &= k_1 gP_2(t) - k_2 g(t) P_2(t) \\
\frac{\mathrm{d}r(t)}{\mathrm{d}t} &= k_3 g(t) - k_4 r(t) \\
\frac{\mathrm{d}P(t)}{\mathrm{d}t} &= k_5 r(t) + k_6 P_2(t) - k_7 P(t)^2 - k_8 P(t) \\
\frac{\mathrm{d}P_2(t)}{\mathrm{d}t} &= \frac{1}{2} k_7 P(t)^2 + k_1 gP_2(t) - \frac{1}{2} k_6 P_2(t) - k_2 g(t) P_2(t) \\
\frac{\mathrm{d}gP_2(t)}{\mathrm{d}t} &= k_2 g(t) P_2(t) - k_1 gP_2(t)
\end{aligned} \tag{2.29}$$

for the concentrations $g(t), r(t), P(t), P_2(t), gP_2(t)$ of g, r, P, P_2, gP_2 at time point t and some non-negative reaction rates $k_1, \ldots, k_8$. The corresponding parameter vector is given by $\boldsymbol{\xi} = (k_1, \ldots, k_8)^\top \in \mathbb{R}^8_+$. Note that the dimerization process needs two proteins P to form one dimer P_2. This is reflected in the second to last equation by multiplying the rate constants k_6 and k_7 by one half. There is no basal production rate for any of the elements. Various examples for possible dynamics of Equation (2.29) are shown in Figure 2.4.

The solution of (nonlinear) ordinary differential equations can be numerically approximated using e.g. Matlab's `ode15s` (Shampine & Reichelt [1997]), or SUNDIALS' `CVODEs`



(a) (b)

Figure 2.5: (a) The SIR model. (b) Time course for the SIR model with transfer rates $\beta_1 = 0.1$, $\beta_2 = 0.01$ and initial conditions $s(0) = 0.9$, $i(0) = 0.1$, and $r(0) = 0$.

(Serban & Hindmarsh [2005]) solvers. For delay differential equations Matlab's `dde23` solver (Shampine & Thompson [2001]) can be used.

2.5.1 Compartment models

A *compartment model* consists of a finite set of mutually exclusive *compartments*. Each compartment holds a group of objects unambiguously identifiable with the respective compartment (Jacquez [1985]). The interaction of compartments is governed by *transition equations*, which control the exchange of objects between compartments. The net flow between compartments is based on the density of its objects. All compartments are assumed to be well-mixed and homogeneous with constant volume. This implies that all objects distribute instantly after transition. In this thesis compartment models are also seen as dynamical systems since transition equations are defined by a system of differential equations. In contrast to biochemical reactions we consider *closed systems* only, i.e. there is no external flow of objects into or out of the system.

Example 2.11 (SIR model). Probably the most prominent example for a compartment model stems from epidemiology. It estimates the spread of an epidemic, such as measles, in large populations (Anderson & May [1992]). A simple version contains the three compartments S (susceptible subjects), I (infectious subjects) and R (recovered subjects), which are governed by the system of ODE's:

$$\begin{aligned}\frac{\mathrm{d}s(t)}{\mathrm{d}t} &= -\beta_1 \cdot i(t) \cdot s(t)\\ \frac{\mathrm{d}i(t)}{\mathrm{d}t} &= \beta_1 \cdot i(t) \cdot s(t) - \beta_2 \cdot i(t)\\ \frac{\mathrm{d}r(t)}{\mathrm{d}t} &= \beta_2 \cdot i(t)\end{aligned}$$

where β_1 and β_2 are transfer rates and $s(t), i(t), r(t)$ the concentrations corresponding to the compartments S, I and R at time t. The hazard of an individual for an infection depends on the concentration of infected individuals and the transfer rate β_1. The chance for recovery on the other hand is solely controlled by β_2. A time course of the model is shown in Figure 2.5.

2.5.2 Parameter estimation in dynamical systems

Parameter inference of differential equation systems is a prominent topic in the field of computational systems biology. Despite the arrival of new, high-throughput measurement techniques, compared to model complexity most systems suffer from very low observation numbers and noisy measurements. Moreover, as biological organisms need to be able to quickly adjust to various environmental conditions, we hence expect these models to be somewhat insensitive to parameter variations. Some may even show two ranges of functionality (Kaplan *et al.* [2008]).

Given the general dynamical system

$$\frac{\mathrm{d}\boldsymbol{x}(t)}{\mathrm{d}t} = \boldsymbol{g}_{\boldsymbol{\xi}}(\boldsymbol{x}(t), \boldsymbol{\tau}, \boldsymbol{u}(t), t), \tag{2.30}$$

the vector $\boldsymbol{\xi} \in \mathbb{R}^n$ in (2.30) holds the parameters defining $\boldsymbol{g}_{\boldsymbol{\xi}}$. As mentioned above, it can contain rate constants, initial values to (2.30) or other constants e.g. necessary for the link functions defined in the following. In practical applications $\boldsymbol{\xi}$ needs to be inferred from a set of given observations $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m\}$, where $\boldsymbol{y}_i$ was observed at the time points $t_i \in [0, T]$ for $i = 1, \ldots, m$. It is assumed that there exists a parameter vector $\boldsymbol{\xi}$ such that the simulation of the differential equation trajectory of (2.30) contains the true dynamics of the particular biological system. More precisely, the observation $\boldsymbol{y}_i = (y_{i,1}, \ldots, y_{i,l_i})^\top$ with $l_i \in \mathbb{N}$ is supposed to satisfy the equation

$$y_{i,j} = h_{\boldsymbol{\xi}}^{(i,j)}(\boldsymbol{x}(t_i)) + \varepsilon_{i,j}, \qquad j = 1, \ldots, l_i \tag{2.31}$$

for the realizations $\varepsilon_{i,j}$ of some independent normally distributed random variable with mean zero and unknown variance. Note that the l_i's can vary between observations as we do not require the measurements to be of the same dimension at every time point t_i. The functions $h_{\boldsymbol{\xi}}^{(i,j)} : \mathbb{R}^d \longrightarrow \mathbb{R}$ denote $\boldsymbol{\xi}$-dependent *link functions*, where $h_{\boldsymbol{\xi}}^{(i,j)}$ can correspond to a simple projection on the j^{th} component of $\boldsymbol{x}(t_i)$, to sums $x_{j_1}(t_i)+\ldots+x_{j_r}(t_i)$ $(j_1,\ldots,j_r \in \{1,\ldots,l_i\})$ thereof, or their rescaled versions $s_1 \cdot x_j(t_i)$, or $s_2 \cdot (x_{j_1}(t_i)+\ldots+x_{j_r}(t_i))$, where s_1 and s_2 are unknown scaling constants contained in $\boldsymbol{\xi}$. These link functions arise since technical limitations frequently prevent to observe each concentration x_i individually.

On the basis of Equation (2.31) the parameter vector $\boldsymbol{\xi}$ of 2.30 is generally estimated by minimizing the squared error loss function

$$\chi^2(\boldsymbol{\xi}) = \sum_{i=1}^{m} \sum_{j=1}^{l_i} \frac{\left(y_{i,j} - h_{\boldsymbol{\xi}}^{(i,j)}(\boldsymbol{x}(t_i))\right)^2}{\sigma_{i,j}^2}, \tag{2.32}$$

with respect to $\boldsymbol{\xi}$ (see Horbelt *et al.* [2002] or Maiwald & Timmer [2008]), where $\sigma_{i,j}^2$ denote known measurement errors. Very promising approaches for the minimization process include global nonlinear optimization methods, such as the simulated annealing algorithm (see Černỳ [1985]; Kirkpatrick *et al.* [1983] or Chapter 4.6), the genetic algorithm (Fraser & Burnell [1970]), or coupled local minimizers (Suykens & Vandewalle [2002]; Suykens *et al.* [2002]). These techniques have shown to work well in practice as they try to avoid getting trapped in local minima. Nevertheless, deterministic methods, such as steepest decent algorithms (Fletcher [1987]) started various times at different initial $\boldsymbol{\xi}$-values can also be applied.

2.5.3 Parameter identifiability in dynamical systems

Technical limitations generally prevent experimentalists from individually measuring every substance involved in a biological process. This means that we are dealing with incomplete data when modeling a particular system. Moreover, there might even be a considerable amount of noise on the measurements. Therefore, we need to raise the question, whether a model – or rather its parameters – can at all be identified based on some noisy, incomplete observations $\boldsymbol{y} := \{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_m\}$?

An approach to address this issue for dynamical systems has been proposed by Raue *et al.* [2009] (see also Murphy & Van der Vaart [2000] and Venzon & Moolgavkar [1988]). The authors consider sets of the form

$$CR_\alpha := \{\boldsymbol{\xi} | \chi^2(\boldsymbol{\xi}) - \chi^2(\hat{\boldsymbol{\xi}}) < \Delta_\alpha\},$$

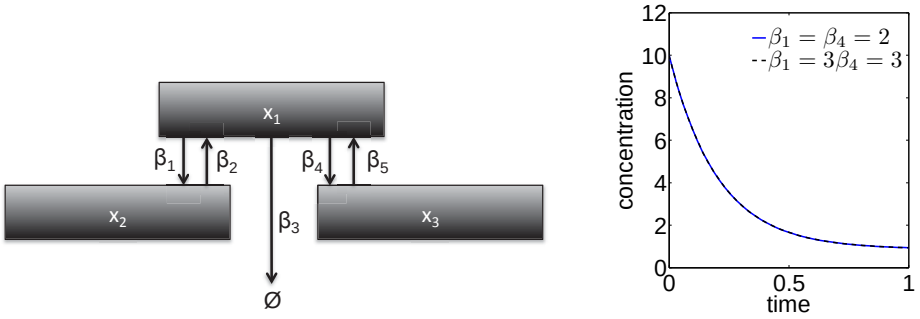
where $\chi^2(\boldsymbol{\xi})$ is the squared error loss function of Equation (2.32), $\hat{\boldsymbol{\xi}}$ is the according estimated argmin, i.e. $\chi^2(\boldsymbol{\xi}) > \chi^2(\hat{\boldsymbol{\xi}})$ for all defined $\boldsymbol{\xi}$'s, and Δ_α is the α-quantile of the χ^2-distribution with one degree of freedom (for details see Press *et al.* [1986]). Meeker & Escobar [1995] showed that the borders of CR_α represent confidence regions for $\hat{\boldsymbol{\xi}}$ in linear systems.

A parameter ξ_i of $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)^\top$ is then said to be *identifiable*, if the confidence interval $[l_i, u_i]$ (of the estimate $\hat{\xi}_i$) defined by $l_i := \min\{\xi_i | \exists\, \boldsymbol{\xi}, s.t.\ \chi^2(\boldsymbol{\xi}) - \chi^2(\hat{\boldsymbol{\xi}}) < \Delta_\alpha\}$ and $u_i := \max\{\xi_i | \exists\, \boldsymbol{\xi}, s.t.\ \chi^2(\boldsymbol{\xi}) - \chi^2(\hat{\boldsymbol{\xi}}) < \Delta_\alpha\}$ (if existent and $-\infty/+\infty$ otherwise), is finite. Moreover, Raue *et al.* [2009] propose even finer notions of identifiability: They call a system *structurally identifiable*, if $\chi^2(\boldsymbol{\xi})$ possesses a unique minimum. Furthermore, a system is named *practically identifiable*, if it possesses a unique minimum and none of the confidence intervals $[l_i, u_i]$ ($i \in \{1, \ldots, n\}$) has infinite size. Frankly spoken, structural identifiability issues arise, if there exists a functional relationship between individual model parameters. Practical identifiability issues, on the other hand, are caused by too noisy measurements; although the measurements allow for a minimum of $\chi^2(\boldsymbol{\xi})$ at $\hat{\boldsymbol{\xi}}$, the $\chi^2(\boldsymbol{\xi})$ function is too flat around $\hat{\boldsymbol{\xi}}$ to consider the estimate significant. The following example gives an instance of a structurally non-identifiable model:

Example 2.12 (Structural non-identifiability). Let us consider the compartment model inspired by the models of Chapter 8:

$$\begin{aligned}
\frac{\mathrm{d}x_1(t)}{\mathrm{d}t} &= -\beta_1 x_1(t) - \beta_3 x_1(t) - \beta_4 x_1(t) + \beta_2 x_2(t) + \beta_5 x_3(t) \\
\frac{\mathrm{d}x_2(t)}{\mathrm{d}t} &= \beta_1 x_1(t) - \beta_2 x_2(t) \\
\frac{\mathrm{d}x_3(t)}{\mathrm{d}t} &= \beta_4 x_1(t) - \beta_5 x_3(t)
\end{aligned} \tag{2.33}$$

where (without loss of generality) $\beta_1, \ldots, \beta_5$ are positive rate constants controlling the flow between the compartments x_1, x_2, and x_3. The rate β_3 corresponds to the degradation of the elements in x_1. Suppose $x_1(0) = 10$ and $x_2(0) = x_3(0) = 0$ in



(a) (b)

Figure 2.6: (a) Schematic representation of model (2.33). (b) Time course for compartment x_1 of the model in (a) for $x_1(0) = 10$ and $x_2(0) = x_3(0) = 0$. The blue line corresponds to $\beta_1 = \beta_4 = 2, \beta_2 = \beta_3 = \beta_5 = 0.5$, the dashed black line to $\beta_1 = 3, \beta_4 = 1, \beta_2 = \beta_3 = \beta_5 = 0.5$.

arbitrary units. For the vector of concentrations $\boldsymbol{x}(t) = (x_1(t), x_2(t), x_3(t))^\top$ the linear ODE (2.33) has the solution

$$\boldsymbol{x}(t) = \exp(\boldsymbol{A}\cdot t)\cdot \begin{pmatrix} 10 \\ 0 \\ 0 \end{pmatrix} \qquad \text{for the matrix} \qquad \boldsymbol{A} = \begin{pmatrix} -\beta_1 - \beta_3 - \beta_4 & \beta_2 & \beta_5 \\ \beta_1 & -\beta_2 & 0 \\ \beta_4 & 0 & -\beta_5 \end{pmatrix}.$$

Setting $\beta_2 = \beta_5$ the characteristic polynomial in λ is

$$\chi(\lambda) = (\beta_2 + \lambda)^2(c + \beta_3) - (\beta_2 + \lambda)\beta_2 c,$$

where $c = \beta_1 + \beta_4 > 0$. This shows that the parameters β_1 and β_4 are not identifiable, since for any $\beta_1 \neq c$ we get the same solution to (2.33) setting $\beta_4 = c - \beta_1$. A schematic representation of (2.33) including a time course can be found in Figure 2.6.

For checking the identifiability of a particular model the Matlab based PottersWheel software (Maiwald & Timmer [2008]) can be used. Here, the finiteness of the confidence intervals $[l_i, u_i]$ is estimated via the so-called *profile likelihood* function

$$\chi^2_{PL}(\xi_i^0) := \min_{\mathcal{A}} \chi^2(\boldsymbol{\xi})$$

for $\mathcal{A} = \{\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)^\top | \xi_i = \xi_i^0\}$. The profile likelihood checks whether for $i = 1, \ldots, n$ the bounds l_i and u_i are finite by proceeding into the direction of the least increase of $\chi^2(\boldsymbol{\xi})$ starting at the estimated argmin $\hat{\boldsymbol{\xi}}$ of $\chi^2(\boldsymbol{\xi})$.

Various other methods for parameter identifiability analysis were proposed by Hengl *et al.* [2007]; Lecourtier *et al.* [1987]; Ljung & Glad [1994] or D. *et al.* [2003]. All of these approaches do however not test for practical identifiability.

In summary, before performing parameter estimation in dynamic systems, an identifiability analysis has to be performed in order to ensure the validity of parameter estimates.

3

Bayesian model inference

We now turn to the concept of Bayesian model inference. For the application in dynamical systems this comprises two aspects: On the one hand, given a specific parametrized model $\mathcal{M} = \mathcal{M}(\boldsymbol{\xi})$, we want to infer the parameter values $\boldsymbol{\xi} \in \mathbb{R}^n$ and their corresponding uncertainties based on a series of observations $\boldsymbol{y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m\}$. This can be seen as extension to the parameter estimation approach described in Chapter 2.5.2. On the other hand, given a set of parametrized models $\mathcal{M}_1, \ldots, \mathcal{M}_k$ with according parameter vectors $\boldsymbol{\xi}_1 \in \mathbb{R}^{n_1}, \ldots, \boldsymbol{\xi}_k \in \mathbb{R}^{n_k}$, we want to infer the best model structure $\mathcal{M}_i$ resembling the observations $\boldsymbol{y}$, i.e. we want to deduce the model $\mathcal{M}_i$ with the highest probability that the observations $\boldsymbol{y}$ were generated by $\mathcal{M}_i$. Interestingly, both aspects can be covered by the technique of Markov Chain Monte Carlo (MCMC) sampling addressed in Chapter 4.

3.1 Bayesian parameter inference

First, we address the issue of inferring the parameter vector $\boldsymbol{\xi} \in \mathbb{R}^n$ for a given parametrized model $\mathcal{M}$ based on observations $\boldsymbol{y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m\}$. Taking a frequentistic approach, $\boldsymbol{y}$ is considered the outcome of one of infinitely many repetitions of the same experiment based on $\boldsymbol{\xi}$. Conversely, in the Bayesian paradigm $\boldsymbol{\xi}$ is considered as a realization of a random parameter vector (Robert & Casella [2004]). Every data point $\boldsymbol{y}_i$ is here seen as realization of a random vector $\boldsymbol{X} \sim f(\boldsymbol{x}|\boldsymbol{\xi})$ for a density function

$f(\cdot|\boldsymbol{\xi})$ conditioned on the parameter vector $\boldsymbol{\xi}$. Given $f(\cdot|\boldsymbol{\xi})$ we define the *likelihood function* (or simply *likelihood*) for $\boldsymbol{y} = \{\boldsymbol{y}_1, \dots, \boldsymbol{y}_m\}$ by

$$\begin{aligned}\mathcal{L}(\boldsymbol{\xi}|\boldsymbol{y}) &= \mathcal{L}(\boldsymbol{\xi}|\boldsymbol{y}_1, \dots, \boldsymbol{y}_m) \\ &= \prod_{i=1}^{m} f(\boldsymbol{y}_i|\boldsymbol{\xi}).\end{aligned}$$

This is a function of the parameter vector $\boldsymbol{\xi}$. In combination with a *prior* density function $\pi(\boldsymbol{\xi})$, containing information about $\boldsymbol{\xi}$ before observing $\boldsymbol{y}$, *Bayes' theorem* gives rise to a probability distribution function

$$\pi(\boldsymbol{\xi}|\boldsymbol{y}) = \frac{\mathcal{L}(\boldsymbol{\xi}|\boldsymbol{y})\pi(\boldsymbol{\xi})}{\int_{\mathbb{R}^n} \mathcal{L}(\boldsymbol{\xi}|\boldsymbol{y})\pi(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi}} \tag{3.1}$$

(see for instance Berger [1985] or Bernardo *et al.* [1994] for the foundations of this approach). The term $\pi(\boldsymbol{\xi}|\boldsymbol{y})$ in (3.1) is called *posterior distribution of* $\boldsymbol{\xi}$ and represents the core of Bayesian inference. It contains the distribution of $\boldsymbol{\xi}$ while taking into account both the data $\boldsymbol{y}$ itself as well as prior knowledge about the parameters (by means of $\pi(\boldsymbol{\xi})$). Note that we abusively use $\boldsymbol{\xi}$ to denote a realization of the random parameter vector and the variable of the density function (3.1). The integral in the denominator

$$\pi(\boldsymbol{y}) = \int_{\mathbb{R}^n} \mathcal{L}(\boldsymbol{\xi}|\boldsymbol{y})\pi(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi} \tag{3.2}$$

is called *marginal likelihood* (or *model evidence*). As we will see in Chapter 3.4, $\pi(\boldsymbol{y})$ is subject to model inference, but generally analytically intractable. In high dimensions numerical approximations of $\pi(\boldsymbol{y})$ are difficult and mostly inaccurate. Sophisticated approaches such as the ones introduced in Section 3.4.3 have to be applied. The inference of the posterior distribution of (3.1) is often done by MCMC sampling (Chapter 4). These methods exploit the fact that $\pi(\boldsymbol{y})$ does not depend on the parameter vector $\boldsymbol{\xi}$: As the vector of observations $\boldsymbol{y}$ is fixed, the inference is solely based on the relation

$$\pi(\boldsymbol{\xi}|\boldsymbol{y}) \propto \mathcal{L}(\boldsymbol{\xi}|\boldsymbol{y})\pi(\boldsymbol{\xi}). \tag{3.3}$$

Analogously to Chapter 2.5.2 we want to infer the parameter distribution of a parametrized differential equation of the form

$$\frac{\mathrm{d}\boldsymbol{x}(t)}{\mathrm{d}t} = \boldsymbol{g}_{\boldsymbol{\xi}}(\boldsymbol{x}(t), \boldsymbol{\tau}, \boldsymbol{u}(t), t), \tag{3.4}$$

based on noisy observations $\boldsymbol{y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m\}$ at the time points $t_1, \ldots, t_m$. For Bayesian parameter estimation we assume that the observations contain independent measurement errors, i.e. for $\boldsymbol{y}_i = (y_{i,1}, \ldots, y_{i,l_i})^\top$ we have

$$y_{i,j} = h_{\boldsymbol{\xi}}^{(i,j)}(\boldsymbol{x}(t_i)) + \varepsilon_{i,j}, \qquad j = 1, \ldots, l_i \tag{3.5}$$

where $\varepsilon_{i,j}$ are realizations distributed according to some time-independent noise model density functions $f^{(i,j)}(\cdot|\boldsymbol{\xi})$. Here, $h_{\boldsymbol{\xi}}^{(i,j)}$ again denote link functions introduced in Equation (2.31). We hence infer the distribution

$$\pi(\boldsymbol{\xi}|\boldsymbol{y}) \propto \prod_{i=1}^{m} \prod_{j=1}^{l_i} f^{(i,j)}(y_{i,j} - h_{\boldsymbol{\xi}}^{(i,j)}(\boldsymbol{x}(t_i))|\boldsymbol{\xi})\pi(\boldsymbol{\xi})$$

for some prior $\pi(\boldsymbol{\xi})$ and the solution $\boldsymbol{x}(t)$ of Equation (3.4) corresponding to the parameter vector $\boldsymbol{\xi}$. The vector $\boldsymbol{\xi}$ can contain rate constants, time delays, or initial conditions of the differential equation, but also parameters of the applied noise model, such as the standard deviation in case of a Gaussian error model.

We define the *maximum likelihood estimator* (MLE) for a given likelihood function $\mathcal{L}(\boldsymbol{\xi}|\boldsymbol{y})$ as the parameter vector $\hat{\boldsymbol{\xi}}_{ML} \in \mathbb{R}^n$ at which $\mathcal{L}(\boldsymbol{\xi}|\boldsymbol{y})$ attains its maximum, if existent. Furthermore, the *maximum a posteriori estimate* (MAP) for a posterior function $\pi(\boldsymbol{\xi}|\boldsymbol{y})$ is the value $\hat{\boldsymbol{\xi}}_{MAP} \in \mathbb{R}^n$ at which $\pi(\boldsymbol{\xi}|\boldsymbol{y})$ attains its maximum, if existent. Clearly, both estimators coincide, if the prior density $\pi(\boldsymbol{\xi})$ is uniform and its support contains $\hat{\boldsymbol{\xi}}_{ML}$. It is easy to see that for Gaussian noise model density functions $f^{(i,j)}(\cdot|\boldsymbol{\xi})$ with known measurement errors $\sigma^2_{(i,j)}$, the estimate $\hat{\boldsymbol{\xi}}_{ML}$ coincides with the minimization result of Equation (2.32). Note however that contrary to the classical parameter estimation approach unknown measurement errors $\sigma^2_{(i,j)}$ can be estimated simultaneously in the Bayesian paradigm, this is, $\boldsymbol{\xi}$ can contain the $\sigma^2_{(i,j)}$'s.

Analogously to confidence intervals, we can compute *credible intervals* for each component ξ_j of $\boldsymbol{\xi}$: For $j \in \{1, \ldots, n\}$ and *confidence level* α the $100\%(1-\alpha)$ *credible interval* for ξ_j is the interval $I_j = [I_j^{low}, I_j^{up}]$ with lower bound

$$I_j^{low} = \Pi_j^{-1}(\alpha/2|\boldsymbol{y}) \qquad \text{and upper bound} \qquad I_j^{up} = \Pi_j^{-1}(1-\alpha/2|\boldsymbol{y}),$$

where $\Pi_j^{-1}(\cdot|\boldsymbol{y})$ is the univariate *quantile function* corresponding to the marginal posterior density function $\pi(\xi_j|\boldsymbol{y})$ for ξ_j, i.e. $\Pi_j^{-1}(\cdot|\boldsymbol{y})$ is the inverse of the corresponding

posterior distribution function $\Pi(\xi_j|\boldsymbol{y})$. Clearly,

$$\int_{I_j} \pi(\xi_j|\boldsymbol{y})\,\mathrm{d}\xi_j = 1-\alpha. \tag{3.6}$$

As in classical approaches $\alpha = 0.1$ and $\alpha = 0.05$ are popular, although arbitrary, confidence levels. We may point out that there exist generic cases for which the j^{th} component ($j \in \{1,\ldots,n\}$) of the MLE or MAP is not contained in the credible interval I_j. However, in practical applications this is generally not an issue.

While in the frequentistic paradigm confidence intervals hold the probability for the "true" parameter value to be contained within the confidence interval of a parameter $\boldsymbol{\xi}$, in the Bayesian language credible intervals hold the probability that $\boldsymbol{\xi}$ itself is contained in this interval. Credible intervals are thus considered one of the strongest points of Bayesian inference.

3.2 Prior distributions

A crucial issue in Bayesian statistics is the correct choice of prior distributions. Whenever we are given information about the model or observations $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_m$, it should (and must) be contained in the inference process. However, since priors can have considerable influence on the obtained results, the topic has to be treated with care. A very popular choice for prior distributions constitutes the class of *conjugate priors*: When the likelihood is of the *exponential family*, this is, the likelihood has the form

$$\mathcal{L}(\boldsymbol{\xi}|\boldsymbol{y}) = h(\boldsymbol{y})\exp(\boldsymbol{\xi}\cdot\boldsymbol{R}(\boldsymbol{y}) - \Psi(\boldsymbol{\xi}))$$

with functions $h:\mathbb{R}^l \longrightarrow \mathbb{R}$, $\boldsymbol{R}:\mathbb{R}^l \longrightarrow \mathbb{R}^n$, $\Psi:\mathbb{R}^n \longrightarrow \mathbb{R}$ – the product $\boldsymbol{\xi}\cdot\boldsymbol{R}(\boldsymbol{y})$ is to be understood as scalar product in $\mathbb{R}^n$ – then for $\boldsymbol{\zeta}\in\mathbb{R}^n$, $\lambda > 0$ the *conjugate prior* is defined as

$$\pi(\boldsymbol{\xi}|\boldsymbol{\zeta},\lambda) = k(\boldsymbol{y})\exp(\boldsymbol{\zeta}\cdot\boldsymbol{\xi} - \lambda\Psi(\boldsymbol{\xi}))$$

where $k:\mathbb{R}^l \longrightarrow \mathbb{R}$ is a function dependent on $\boldsymbol{y}$ only (Marin & Robert [2007]). Since the data $\boldsymbol{y}$ is fixed, k can also be considered as constant. The prior parameters $\boldsymbol{\zeta}$ and λ are called *hyperparameters*, as are all parameters influencing the prior distribution. For a conjugate prior the posterior can be written as

$$\pi(\boldsymbol{\xi}|\boldsymbol{y}) = \pi(\boldsymbol{\xi}|\boldsymbol{\zeta}+\boldsymbol{R}(\boldsymbol{y}),\lambda+1).$$

In this case the posterior distribution just "updates" the prior parameters. Robert & Casella [2004] pointed out that using conjugate priors for computational reasons might introduce effects unrelated to reality on the inference process. Hence, in this thesis we frequently use *non-informative priors*[1]

$$\pi(\boldsymbol{\xi}) \propto \mathbb{1}_S(\boldsymbol{\xi})$$

on some support set $S \subseteq \mathbb{R}^n$. Although the choice of a finite support S already introduces prior information on $\boldsymbol{\xi}$, the restriction is usually very mild: Rate constants of biological systems are generally considered to be non-negative and setting an upper bound can also be easily done without considerably influencing the posterior distribution.

Interestingly, we can choose any non-negative function $\pi : \mathbb{R}^n \longrightarrow \mathbb{R}$ with

$$\int_{\mathbb{R}^n} \pi(\boldsymbol{\xi}) \, \mathrm{d}\boldsymbol{\xi} = c$$

for a constant $c \in (0, \infty]$ as prior distribution, as long as the marginal likelihood fulfills

$$\pi(\boldsymbol{y}) = \int_{\mathbb{R}^n} \mathcal{L}(\boldsymbol{\xi}|\boldsymbol{y})\pi(\boldsymbol{\xi}) \, \mathrm{d}\boldsymbol{\xi} < \infty \tag{3.7}$$

almost surely with respect to $\boldsymbol{y}$ (Marin & Robert [2007]). This becomes clear as c is independent of $\boldsymbol{\xi}$, recalling that we are only interested in the proportionality (3.3). In case $c \neq 1$, we call $\pi(\boldsymbol{\xi})$ *improper* and *proper* otherwise.

In summary, while the use of non-informative priors might be necessary for unknown parameter values, there needs to be a sufficient amount of observations in order to gain valuable inference results. The term "sufficient amount" is generally hard to specify and whenever possible the modelers expertise, information from related experiments or the literature should guide the choice in prior selection. Moreover, the use of very distinct priors can help weighting down the influence of the data in the posterior distribution (Bernardo *et al.* [1994]), a situation possibly aspired for error prone observations. In general we make use of independent prior distributions, i.e. $\pi(\boldsymbol{\xi}) = \prod_{i=1}^n \pi(\xi_i)$, for the marginal density functions $\pi(\boldsymbol{x}_i)$ of ξ_i. This, however, is not a necessity!

[1]The notion of non-informative priors is only vaguely defined in the literature. Essentially, the influence of the prior distribution needs to be as small as possible (Marin & Robert [2007]). Therefore, although there are other prior types with little influence on the posterior, we stick to the definition via the indicator function.

3.3 Bayesian parameter identifiability

The MAP of high dimensional dynamical systems is oftentimes the center of interest when performing parameter inference. However, especially for non-informative priors $\pi(\boldsymbol{\xi}) = \mathbb{1}_S(\boldsymbol{\xi})$ this might inherit some issues as seen in the structural non-identifiability Example 2.12. While for uni- or bivariate posterior distributions a graphical visualization of the posterior density can help to identify these issues this approach is not possible for higher dimensional systems. However, analogously to Meeker & Escobar [1995], exchanging the squared error loss function (2.32) by twice the negative logarithm $-2\log(\pi(\boldsymbol{\xi}|\boldsymbol{y}))$ of the posterior density function $\pi(\boldsymbol{\xi}|\boldsymbol{y})$ naturally extends the identifiability analysis of Chapter 2.5.3 to the Bayesian paradigm: We call a dynamical system *identifiable with respect to the MAP* or simply *identifiable*, if for $i = 1, \dots, n$ the confidence intervals $[l_i, u_i]$ (of the i^{th} component of the estimate $\hat{\boldsymbol{\xi}}_{MAP}$) defined by $l_i := \min\{\xi_i | \exists\ \boldsymbol{\xi}, s.t.\ 2\log(\pi(\hat{\boldsymbol{\xi}}_{MAP}|\boldsymbol{y})) - 2\log(\pi(\boldsymbol{\xi}|\boldsymbol{y})) < \Delta_\alpha\}$ and $u_i := \max\{\xi_i | \exists\ \boldsymbol{\xi}, s.t.\ 2\log(\pi(\hat{\boldsymbol{\xi}}_{MAP}|\boldsymbol{y})) - 2\log(\pi(\boldsymbol{\xi}|\boldsymbol{y})) < \Delta_\alpha\}$ (if existent and $-\infty/+\infty$ otherwise), is finite. With this, structural and practical identifiability can also be defined as in Chapter 2.5.3. Note that the posterior distribution only needs to be known up to a scaling constant as scaling constants cancel out in the differences of the definitions of the l_i's and u_i's. For Gaussian noise functions the Bayesian definition of identifiability coincides with the definition in Chapter 2.5.3.

3.4 Bayes factors

We now turn to the issue of Bayesian model selection. Generally, statistical modeling of any kind of data crucially depends on the modeler's expertise in model construction. Inferring parameter values based on the wrong model can have severe effects on the outcome of the model's predictions. Towards this end the task of model selection has ever since been a very important step in the process of data analysis. A well established approach to address this problem is the likelihood ratio test. For general applications it is based on the ratio of the maximal likelihood value of a model with restricted parameter space and the maximal likelihood value of the same model on the full parameter space (Casella & Berger [2001]). Since the latter always performs

at least as good as the former one, the test analyzes how much more likely the full model compared to the restricted one is with respect to the data. However, the need to restrict the parameter space limits this approach to nested models only, i.e. the simpler model can be transformed into the more complex one by the introduction of additional parameters. A more sophisticated test for non-nested models was established by Vuong (Vuong [1989]). As the likelihood ratio test, the Vuong test is also based on the likelihood ratio of two competing models, but additional corrects by a Kulback-Leibler information criterion driven term. In any case, all of the approaches above suffer from two major drawbacks: (i) they are based on two single maximally likely values – one for each model – not taking into account any kind of uncertainty in the parameters and, moreover, (ii) due to the maximality restriction the test statistic is defined on the probability of extreme events, anticipating that the data produces events at least as extreme as the tested ones.

The Bayesian way of model selection – the second strain in Bayesian model inference – naturally circumvents these problems by taking into account full parameter distributions rather than single maximal parameter values. In the following, we introduce the concept of Bayesian model selection including a practical approach for computing the so-called *Bayes factor*, a statistic used for pairwise model comparison. A nice comprehensive discussion about this topic is given in Kass & Raftery [1995].

As we have seen in Section 3.1, the posterior distribution $\pi(\boldsymbol{\xi}|\boldsymbol{y})$ contains the full structural dependency of all parameters involved in the model. In Bayesian model selection we simply extend $\pi(\boldsymbol{\xi}|\boldsymbol{y})$ by an additional *model parameter*, this is, given a set of (possibly non-nested) models $\mathcal{M}_1, \ldots, \mathcal{M}_k$ with according parameter vectors $\boldsymbol{\xi}_1 \in \mathbb{R}^{n_1}, \ldots, \boldsymbol{\xi}_k \in \mathbb{R}^{n_k}$ we consider the distribution defined by

$$\pi(\boldsymbol{\xi}_i, \mathcal{M}_i|\boldsymbol{y}) = \frac{\mathcal{L}_i(\boldsymbol{\xi}_i|\boldsymbol{y})\pi(\boldsymbol{\xi}_i, \mathcal{M}_i)}{\sum_{j=1}^{k} \int_{\mathbb{R}^{n_j}} \mathcal{L}_j(\boldsymbol{\xi}_j|\boldsymbol{y})\pi(\boldsymbol{\xi}_j, \mathcal{M}_j)\,\mathrm{d}\boldsymbol{\xi}_j} \tag{3.8}$$

where $\mathcal{L}_i(\boldsymbol{\xi}_i|\boldsymbol{y})$ denotes the likelihood function and $\pi(\boldsymbol{\xi}_i, \mathcal{M}_i)$ the joint prior density for model $\mathcal{M}_i$ and parameter vector $\boldsymbol{\xi}_i$. Clearly, conditioning Equation (3.8) on model $\mathcal{M}_i$ gives the relation (3.3), where the proportionality is to be understood with respect to the data and model.

In the case of two models $\mathcal{M}_1$ and $\mathcal{M}_2$, we – similarly to classical hypothesis testing – get the *posterior odds ratio* via integration over the corresponding parameter spaces:

$$\begin{aligned}\frac{\pi(\mathcal{M}_1|\boldsymbol{y})}{\pi(\mathcal{M}_2|\boldsymbol{y})} &= \frac{\int_{\mathbb{R}^{n_1}} \pi(\boldsymbol{\xi}_1, \mathcal{M}_1|\boldsymbol{y})\,\mathrm{d}\boldsymbol{\xi}_1}{\int_{\mathbb{R}^{n_2}} \pi(\boldsymbol{\xi}_2, \mathcal{M}_2|\boldsymbol{y})\,\mathrm{d}\boldsymbol{\xi}_2} \\ &= \left(\frac{\pi(\mathcal{M}_1)\ \int_{\mathbb{R}^{n_1}} \mathcal{L}_1(\boldsymbol{\xi}_1|\boldsymbol{y})\pi(\boldsymbol{\xi}_1|\mathcal{M}_1)\,\mathrm{d}\boldsymbol{\xi}_1}{\sum_{j=1}^{2}\pi(\mathcal{M}_j)\ \int_{\mathbb{R}^{n_j}} \mathcal{L}_j(\boldsymbol{\xi}_j|\boldsymbol{y})\pi(\boldsymbol{\xi}_j|\mathcal{M}_j)\,\mathrm{d}\boldsymbol{\xi}_j}\right) \\ &\qquad \cdot \left(\frac{\pi(\mathcal{M}_2)\ \int_{\mathbb{R}^{n_2}} \mathcal{L}_2(\boldsymbol{\xi}_2|\boldsymbol{y})\pi(\boldsymbol{\xi}_2|\mathcal{M}_2)\,\mathrm{d}\boldsymbol{\xi}_2}{\sum_{j=1}^{2}\pi(\mathcal{M}_j)\ \int_{\mathbb{R}^{n_j}} \mathcal{L}_j(\boldsymbol{\xi}_j|\boldsymbol{y})\pi(\boldsymbol{\xi}_j|\mathcal{M}_j)\,\mathrm{d}\boldsymbol{\xi}_j}\right)^{-1} \\ &= \frac{\pi(\boldsymbol{y}|\mathcal{M}_1)}{\pi(\boldsymbol{y}|\mathcal{M}_2)}\frac{\pi(\mathcal{M}_1)}{\pi(\mathcal{M}_2)},\end{aligned} \tag{3.9}$$

where

$$\pi(\boldsymbol{y}|\mathcal{M}_j) = \int_{\mathbb{R}^{n_j}} \mathcal{L}_j(\boldsymbol{\xi}_j|\boldsymbol{y})\pi(\boldsymbol{\xi}_j|\mathcal{M}_j)\,\mathrm{d}\boldsymbol{\xi}_j = \mathbb{E}_{\pi(\boldsymbol{\xi}_j|\mathcal{M}_j)}[\mathcal{L}_j(\boldsymbol{\xi}_j|\boldsymbol{y})] \tag{3.10}$$

$(j = 1, 2)$ is the model evidence of model $\mathcal{M}_j$ defined in Equation (3.2). The expression

$$B_{12} := \frac{\pi(\boldsymbol{y}|\mathcal{M}_1)}{\pi(\boldsymbol{y}|\mathcal{M}_2)} \tag{3.11}$$

from Equation (3.9) is called *Bayes factor* of model $\mathcal{M}_1$ versus $\mathcal{M}_2$. We have to emphasize that contrary to Bayesian parameter inference the Bayes factor can not handle improper prior distributions. This can be seen as follows: Assume (without loss of generality) we are given the improper prior distributions $\pi(\boldsymbol{\xi}_1|\mathcal{M}_1)$ for model $\mathcal{M}_1$. Let $\pi(\boldsymbol{y}|\mathcal{M}_1)$ be the marginal distribution corresponding to model $\mathcal{M}_1$ and $c > 0$ an arbitrary constant. Then $\pi^*(\boldsymbol{\xi}_1|\mathcal{M}_1) = c \cdot \pi(\boldsymbol{\xi}_1|\mathcal{M}_1)$ is a valid prior distribution and

$$\pi^*(\boldsymbol{y}|\mathcal{M}_1) = \int_{\mathbb{R}^{n_1}} \mathcal{L}_1(\boldsymbol{\xi}_1|\boldsymbol{y})\pi^*(\boldsymbol{\xi}_1|\mathcal{M}_j)\,\mathrm{d}\boldsymbol{\xi}_1 = c \cdot \pi(\boldsymbol{y}|\mathcal{M}_1).$$

Hence, Bayes factors are not well defined for improper prior distributions as they can be varied by arbitrary constants.

Typically we do not favor any model *a priori*. The *prior odds ratio* $\pi(\mathcal{M}_1)/\pi(\mathcal{M}_2)$ is then simply one and the Bayes factor coincides with the ratio of the posterior probabilities of model $\mathcal{M}_1$ and $\mathcal{M}_2$. As Kass & Raftery [1995] pointed out, it is possible for nested models to avoid specifying the prior density functions $\pi(\boldsymbol{\xi}_i, \mathcal{M}_i)$ from Equation (3.8) using the so-called *Schwarz criterion*

$$SC = \log(\pi(\boldsymbol{y}|\hat{\boldsymbol{\xi}}_1, \mathcal{M}_1)) - \log(\pi(\boldsymbol{y}|\hat{\boldsymbol{\xi}}_2, \mathcal{M}_2)) - \frac{1}{2}(n_1 - n_2)\log(N)$$

where log is the natural logarithm, N the number of samples used to compute SC, $\hat{\boldsymbol{\xi}}_j$ the MLE of model j, and n_j the dimension of $\hat{\boldsymbol{\xi}}_j$, respectively. For $N \to \infty$

$$\frac{SC - \log(B_{12})}{\log(B_{12})} \to 0.$$

Hence, $\exp(SC)$ is an approximation of the Bayes factor B_{12}. Although, the approximation is not very accurate even for large sample sizes N, a famous statistic derived from the Schwarz criterion is the *Bayesian information criterion* (BIC). It is given by $BIC = -2SC$. Harold Jeffreys established a widely used interpretation of the Bayes factor in Jeffreys [1961]. He suggested to classify the evidence in favor of model $\mathcal{M}_1$ by $\log_{10}$-half-scale units as:

$\log_{10}(B_{12})$	B_{12}	Evidence in favor of model $\mathcal{M}_1$
0 – 0.5	1 – 3.2	Not worth more than a bare mention
0.5 – 1.0	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 grouping is known as *Jeffreys' scale of evidence*. Certainly some applications challenge this classification (see e.g. Evett [1991]). Nevertheless, Jeffreys' scale of evidence is well established and widely used in the Bayesian community.

The Bayes factor has several advantages compared to classical odds ratio tests: Due to its construction it naturally holds both the evidence in favor of model $\mathcal{M}_1$ and the evidence in favor of model $\mathcal{M}_2$. The latter is simply given by

$$B_{21} = \frac{\pi(\boldsymbol{y}|\mathcal{M}_2)}{\pi(\boldsymbol{y}|\mathcal{M}_1)} = \frac{1}{B_{12}}$$

as we did not make any assumptions about the parameter space of model $\mathcal{M}_1$ and $\mathcal{M}_2$, i.e. the Bayes factor is able to handle non-nested models. Furthermore, for k models $\mathcal{M}_1, \ldots, \mathcal{M}_k$ with uniform model prior densities $\pi(\mathcal{M}_j)$ $(j = 1, \ldots, k)$ and Bayes factors

$$B_{1j} = \frac{\pi(\boldsymbol{y}|\mathcal{M}_1)}{\pi(\boldsymbol{y}|\mathcal{M}_j)}$$

we have the posterior probability that an observation $\boldsymbol{y}$ originates from model $\mathcal{M}_i$ given by (c.f. Kass & Raftery [1995])

$$\pi(\mathcal{M}_i|\boldsymbol{y}) = \frac{B_{i1}}{1 + \sum_{j=2}^{k} B_{j1}}.$$

A huge advantage of Bayes factors is their ability to naturally correct for overfitting issues (Lodewyckx *et al.* [2011]; Myung & Pitt [1997]; Pitt *et al.* [2002]). Since we marginalize rather than maximize with respect to the according parameter spaces, the Bayes factor can compensate for areas that extraordinarily pander a specific model (see Spiegelhalter & Smith [1982] as well as Jefferys & Berger [1992] and references therein).

Although there will never be any "certainty" we picked the "true" model, the process of model selection is still an important issue and draws lots of interest. Naturally, there has been extensive research on the computation of the marginal likelihood (see e.g. Kass & Raftery [1995]; Lartillot & Philippe [2006]; Newton & Raftery [1994] or Friel & Pettitt [2008]) in settings where no analytical solution is tractable. In the following we give a brief overview about the most prominent approaches. All of them are based on sampling – i.e. generating a number of realizations of– the parameter vector $\boldsymbol{\xi}_j$ of model $\mathcal{M}_j$.

3.4.1 The prior arithmetic mean estimate

The simplest approach to approximate a marginal likelihood $\pi(\boldsymbol{y})$ – for reasons of clarity we drop the model dependency for now – is based on drawing a total of T samples $\boldsymbol{\xi}^{(1)}, \ldots, \boldsymbol{\xi}^{(T)} \overset{i.i.d.}{\sim} \pi(\boldsymbol{\xi})$ from the prior distribution $\pi(\boldsymbol{\xi})$. Equation (3.10) then suggests

$$\pi(\boldsymbol{y}) = \mathbb{E}_{\pi(\boldsymbol{\xi})}[\mathcal{L}(\boldsymbol{\xi}|\boldsymbol{y})] \approx \frac{1}{T}\sum_{j=1}^{T} \mathcal{L}(\boldsymbol{\xi}^{(j)}|\boldsymbol{y}), \tag{3.12}$$

where $\mathcal{L}(\boldsymbol{\xi}|\boldsymbol{y})$ denotes the likelihood function. The right hand side of Equation (3.12) is known as the *prior arithmetic mean estimate* (Lartillot & Philippe [2006]). When the number of samples tend to infinity the strong law of large numbers (almost surely) guarantees convergence. However, in practical applications with complex and often times spiky posterior distributions the prior arithmetic mean estimate can be very inefficient as many samples might fall in regions with comparatively low likelihood values (Gamerman & Lopes [2006]). A large number of samples is needed for accurate results. Especially in high-dimensional systems this issue aggravates. Thus, the application of the prior arithmetic mean estimate can afford high computational power in order to obtain acceptable results (Lewis [1994]).

3.4.2 The posterior harmonic mean estimate

Newton & Raftery [1994] suggested an alternative to the prior arithmetic mean estimator: Instead of sampling from the prior distribution, they proposed to sample from the posterior directly. Explicitly, for the samples $\boldsymbol{\xi}^{(1)}, \ldots, \boldsymbol{\xi}^{(T)} \overset{i.i.d.}{\sim} \pi(\boldsymbol{\xi}|\boldsymbol{y})$ from the posterior distribution $\pi(\boldsymbol{\xi}|\boldsymbol{y})$, the *posterior harmonic mean estimate* is defined via the right hand side of

$$\pi(\boldsymbol{y}) \approx \left(\frac{1}{T} \sum_{j=1}^{T} \frac{1}{\mathcal{L}(\boldsymbol{\xi}^{(j)}|\boldsymbol{y})} \right)^{-1} \tag{3.13}$$

where $\mathcal{L}(\boldsymbol{\xi}|\boldsymbol{y})$ is the likelihood with respect to the observations $\boldsymbol{y}$. To see the relation in (3.13), we recall that the samples are drawn from the posterior distribution. Furthermore, the expectation of the inverse of the likelihood function with respect to the posterior distribution is given by

$$\begin{aligned}
\mathbb{E}_{\pi(\boldsymbol{\xi}|\boldsymbol{y})} \left[\frac{1}{\mathcal{L}(\boldsymbol{\xi}|\boldsymbol{y})} \right] &= \int_{\mathbb{R}^n} \frac{1}{\mathcal{L}(\boldsymbol{\xi}|\boldsymbol{y})} \pi(\boldsymbol{\xi}|\boldsymbol{y}) \, \mathrm{d}\boldsymbol{\xi} \\
&= \int_{\mathbb{R}^n} \frac{1}{\mathcal{L}(\boldsymbol{\xi}|\boldsymbol{y})} \frac{\mathcal{L}(\boldsymbol{\xi}|\boldsymbol{y})\pi(\boldsymbol{\xi})}{\pi(\boldsymbol{y})} \, \mathrm{d}\boldsymbol{\xi} \\
&= \frac{1}{\pi(\boldsymbol{y})} \int_{\mathbb{R}^n} \pi(\boldsymbol{\xi}) \, \mathrm{d}\boldsymbol{\xi} \\
&= \frac{1}{\pi(\boldsymbol{y})},
\end{aligned}$$

which leads directly to the approximation of Equation (3.13) (remember that all prior distributions should be proper). The strong law of large numbers guarantees almost sure convergence. Intuitively, this approach circumvents the aforementioned problem of spiky likelihood functions as the samples are generated by means of this spiky distribution itself. However, the posterior harmonic mean estimate suffers severe variance issues, which can be seen from the following simple example given by Neal [2008].

Example 3.1 (Neal [2008]). Suppose we are given the single data point $y \sim \mathcal{N}(\xi, \sigma_1^2)$ with a posterior distribution $\pi(\xi|y)$ for a parameter $\xi \in \mathbb{R}$. We want to infer the marginal likelihood $\pi(y)$ taking the prior distribution $\xi \sim \mathcal{N}(0, \sigma_2^2)$ and assuming that the variances σ_1^2 and σ_2^2 are known. This is a conjugate prior as we will see below. The posterior density function is then for

$$\sigma := \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} \right)^{-\frac{1}{2}}$$

given by

$$\begin{aligned}\pi(\xi|y) &= \frac{1}{\pi(y)}\frac{1}{\sqrt{2\pi}\sigma_1}\exp\left(-\frac{1}{2\sigma_1^2}(\xi-y)^2\right)\cdot\frac{1}{\sqrt{2\pi}\sigma_2}\exp\left(-\frac{1}{2\sigma_2^2}\xi^2\right)\\ &= \frac{1}{\pi(y)}\frac{1}{2\pi\sigma_1\sigma_2}\exp\left(-\frac{1}{2}\left(\frac{\xi^2}{\sigma_2^2}+\frac{\xi^2}{\sigma_1^2}-2\xi\frac{y}{\sigma_1^2}+\frac{y^2}{\sigma_1^2}\right)\right)\\ &= \frac{Z(y)}{\pi(y)}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}\left(\xi-\frac{y\sigma^2}{\sigma_1^2}\right)^2\right)\end{aligned} \tag{3.14}$$

where

$$\begin{aligned}Z(y) &= \frac{\sigma}{\sqrt{2\pi}\sigma_1\sigma_2}\exp\left(-\frac{1}{2}\left(\frac{y^2}{\sigma_1^2}-\frac{y^2\sigma^2}{\sigma_1^4}\right)^2\right)\\ &= \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}}\exp\left(-\frac{1}{2(\sigma_1^2+\sigma_2^2)}y^2\right).\end{aligned} \tag{3.15}$$

The calculation yields two things: (i) Equation (3.14) shows that the posterior distribution for ξ given y is $\mathcal{N}\left(\frac{y\sigma^2}{\sigma_1^2},\sigma^2\right)$, which allows us to sample directly from the posterior; (ii) Integrating both sides of Equation (3.14) over $\mathbb{R}$ yields $Z(y)=\pi(y)$. According to Equation (3.15) the true marginal likelihood for y can be computed via the probability density function corresponding to $\mathcal{N}\left(0,\sigma_1^2+\sigma_2^2,\right)$.

Now, as we have inferred all relevant terms, let us assume

$$y=1,\ \sigma_1^2=1 \text{ and } \sigma_2^2=100.$$

For $T=10^6$ random samples from the posterior $\mathcal{N}\left(\frac{100}{101},\frac{100}{101}\right)$ the estimated mean (including one standard error) for the marginal likelihood based on R =1,000 runs in Matlab is $0.0946\pm3.22\cdot10^{-7}$. The harmonic mean constantly overestimates the "true" value of $\pi(y=1)=0.0395$ more than twice. With increasing σ_2^2 this gap increases. The corresponding prior arithmetic mean estimate for $T=10^6$ samples from the prior $\mathcal{N}(0,101)$ is $0.0395\pm1.06\cdot10^{-11}$ and approximates the "true" value extraordinarily well.

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.

3.4.3 Thermodynamic integration

Contrary to the two approaches described above we employ a method based on path sampling (Gelman & Meng [1998]) in this thesis: The principle of *thermodynamic integration* for marginal likelihood estimation was introduced by Lartillot & Philippe [2006] and Friel & Pettitt [2008] and applied to problems from systems biology by Calderhead & Girolami [2009]. These methods are founded on the integral representation of the natural logarithm of the marginal likelihood by means of the *power posterior*

$$\pi_t(\boldsymbol{\xi}|\boldsymbol{y}) = \frac{\mathcal{L}(\boldsymbol{\xi}|\boldsymbol{y})^t \pi(\boldsymbol{\xi})}{\pi_t(\boldsymbol{y})}, \tag{3.16}$$

where $t \in [0,1]$ and for the prior $\pi(\boldsymbol{\xi})$

$$\pi_t(\boldsymbol{y}) = \int_{\mathbb{R}^n} \mathcal{L}(\boldsymbol{\xi}|\boldsymbol{y})^t \pi(\boldsymbol{\xi}) \, \mathrm{d}\boldsymbol{\xi}. \tag{3.17}$$

Note that the power posterior $\pi_t(\boldsymbol{\xi}|\boldsymbol{y})$ is truly a probability density function. For $t = 0$ and $t = 1$ the marginal $\pi_0(\boldsymbol{y})$ is equal to one and $\pi_1(\boldsymbol{y})$ is the marginal likelihood; for any other t, $\pi_t(\boldsymbol{y})$ is weighing the influence of the data $\boldsymbol{y}$ on the posterior $\pi_t(\boldsymbol{\xi}|\boldsymbol{y})$ via the influence of the likelihoods $\mathcal{L}(\boldsymbol{\xi}|\boldsymbol{y})^t$. Taking the derivative of the logarithm of $\pi_t(\boldsymbol{y})$ with respect to t yields

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} \log(\pi_t(\boldsymbol{y})) &= \frac{1}{\pi_t(\boldsymbol{y})} \frac{\mathrm{d}}{\mathrm{d}t} \pi_t(\boldsymbol{y}) \\
&= \frac{1}{\pi_t(\boldsymbol{y})} \int_{\mathbb{R}^n} \frac{\mathrm{d}}{\mathrm{d}t} \mathcal{L}(\boldsymbol{\xi}|\boldsymbol{y})^t \pi(\boldsymbol{\xi}) \, \mathrm{d}\boldsymbol{\xi} \\
&= \int_{\mathbb{R}^n} \log(\mathcal{L}(\boldsymbol{\xi}|\boldsymbol{y})) \frac{\mathcal{L}(\boldsymbol{\xi}|\boldsymbol{y})^t \pi(\boldsymbol{\xi})}{\pi_t(\boldsymbol{y})} \, \mathrm{d}\boldsymbol{\xi} \\
&= \mathbb{E}_{\pi_t(\boldsymbol{\xi}|\boldsymbol{y})}[\log(\mathcal{L}(\boldsymbol{\xi}|\boldsymbol{y}))].
\end{aligned}$$

The *thermodynamic integral* is then given by integration of t on $[0,1]$ as

$$\log(\pi(\boldsymbol{y})) = \int_0^1 \mathbb{E}_{\pi_t(\boldsymbol{\xi}|\boldsymbol{y})}[\log(\mathcal{L}(\boldsymbol{\xi}|\boldsymbol{y}))] \, \mathrm{d}t. \tag{3.18}$$

This approach tackles the problem of spiky likelihoods by considering a path (from 0 to 1) that gradually puts more and more weight on the likelihood function. An illustrative example for the two-dimensional posterior distribution of Figure 1.1 of the introduction (depicting the k_1-k_6 marginalized posterior distribution of the JAK1-STAT3 model (7.1)) is shown in Figure 3.1. Unfortunately, Equation (3.18) can only

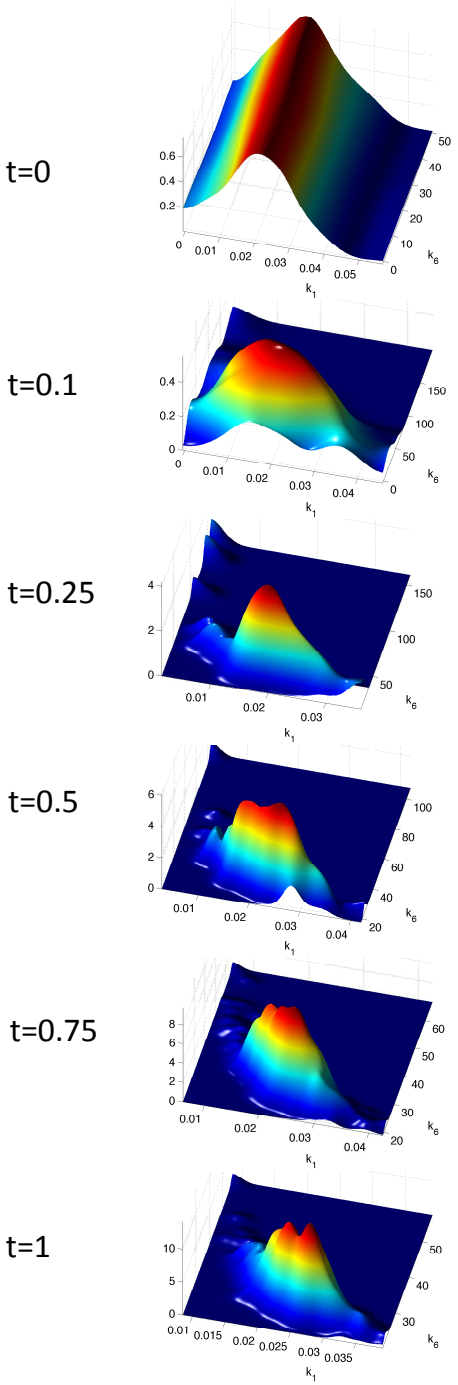


Figure 3.1: Example path for thermodynamic integration based on the posterior distribution of the JAK1-STAT3 model (7.1) introduced in Chapter 7 marginalized on the k_1 and k_6 dimension.

be solved analytically in the most simplest situations. In general $\log(\pi(\boldsymbol{y}))$ is obtained by numerical integration using a finite number of evaluations between $t = 0$ and $t = 1$ applying the trapezium rule (Friel & Pettitt [2008]). Using the discretization $0 = t_0 < t_1 < \ldots < t_T = 1$ Calderhead & Girolami [2009] have shown that

$$\begin{aligned}\log(\pi(\boldsymbol{y})) =& \frac{1}{2}\sum_{i=0}^{T-1}(t_{i+1}-t_i)\left(\mathbb{E}_{\pi_{t_{i+1}}(\boldsymbol{\xi}|\boldsymbol{y})}[\log(\mathcal{L}(\boldsymbol{\xi}|\boldsymbol{y}))] + \mathbb{E}_{\pi_{t_i}(\boldsymbol{\xi}|\boldsymbol{y})}[\log(\mathcal{L}(\boldsymbol{\xi}|\boldsymbol{y}))]\right) \\ &+ \frac{1}{2}\sum_{i=0}^{T-1}\left(KL(\pi_{t_i}(\boldsymbol{\xi}|\boldsymbol{y})||\pi_{t_{i+1}}(\boldsymbol{\xi}|\boldsymbol{y})) - KL(\pi_{t_{i+1}}(\boldsymbol{\xi}|\boldsymbol{y})||\pi_{t_i}(\boldsymbol{\xi}|\boldsymbol{y}))\right)\end{aligned} \tag{3.19}$$

where

$$KL(\pi_{t_i}(\boldsymbol{\xi}|\boldsymbol{y})||\pi_{t_{i+1}}(\boldsymbol{\xi}|\boldsymbol{y})) = \int_{\mathbb{R}^n}\pi_{t_i}(\boldsymbol{\xi}|\boldsymbol{y})\log\left(\frac{\pi_{t_i}(\boldsymbol{\xi}|\boldsymbol{y})}{\pi_{t_{i+1}}(\boldsymbol{\xi}|\boldsymbol{y})}\right)\mathrm{d}\boldsymbol{\xi}$$

is the *Kullback-Leibler divergence* between $\pi_{t_i}(\boldsymbol{\xi}|\boldsymbol{y})$ and $\pi_{t_{i+1}}(\boldsymbol{\xi}|\boldsymbol{y})$ (Kullback & Leibler [1951]). It is non-negative and can be seen as measure for the asymmetric difference between the two distributions as in the limit $KL(\pi_{t_i}(\boldsymbol{\xi}|\boldsymbol{y})||\pi_{t_{i+1}}(\boldsymbol{\xi}|\boldsymbol{y})) \to 0$ whenever $t_i \to t_{i+1}$. Equation (3.19) can easily be obtained based on the relation

$$\frac{\mathcal{L}(\boldsymbol{\xi}|\boldsymbol{y})^{t_{i+1}}}{\mathcal{L}(\boldsymbol{\xi}|\boldsymbol{y})^{t_i}}\pi_{t_i}(\boldsymbol{\xi}|\boldsymbol{y}) = \frac{\mathcal{L}(\boldsymbol{\xi}|\boldsymbol{y})^{t_{i+1}}\pi(\boldsymbol{\xi})}{\mathcal{L}(\boldsymbol{\xi}|\boldsymbol{y})^{t_i}\pi(\boldsymbol{\xi})}\cdot\frac{\mathcal{L}(\boldsymbol{\xi}|\boldsymbol{y})^{t_i}\pi(\boldsymbol{\xi})}{\pi_{t_i}(\boldsymbol{y})} = \frac{\pi_{t_{i+1}}(\boldsymbol{y})}{\pi_{t_i}(\boldsymbol{y})}\pi_{t_{i+1}}(\boldsymbol{\xi}|\boldsymbol{y}). \tag{3.20}$$

With $t \in [0,1]$ we get

$$\begin{aligned}\log(\pi(\boldsymbol{y})) = \log(\pi_1(\boldsymbol{y})) - \log(\pi_0(\boldsymbol{y})) &= \sum_{i=0}^{T-1}\log\left(\frac{\pi_{t_{i+1}}(\boldsymbol{y})}{\pi_{t_i}(\boldsymbol{y})}\right) \\ &= \sum_{i=0}^{T-1}\left[\int_{\mathbb{R}^n}\log\left(\frac{\pi_{t_{i+1}}(\boldsymbol{y})}{\pi_{t_i}(\boldsymbol{y})}\pi_{t_{i+1}}(\boldsymbol{\xi}|\boldsymbol{y})\right)\pi_t(\boldsymbol{\xi}|\boldsymbol{y})\,\mathrm{d}\boldsymbol{\xi} - \int_{\mathbb{R}^n}\log(\pi_{t_{i+1}}(\boldsymbol{\xi}|\boldsymbol{y}))\pi_t(\boldsymbol{\xi}|\boldsymbol{y})\,\mathrm{d}\boldsymbol{\xi}\right] \\ &= \sum_{i=0}^{T-1}\left[\int_{\mathbb{R}^n}\log\left(\frac{\mathcal{L}(\boldsymbol{\xi}|\boldsymbol{y})^{t_{i+1}}}{\mathcal{L}(\boldsymbol{\xi}|\boldsymbol{y})^{t_i}}\pi_{t_i}(\boldsymbol{\xi}|\boldsymbol{y})\right)\pi_t(\boldsymbol{\xi}|\boldsymbol{y})\,\mathrm{d}\boldsymbol{\xi} - \int_{\mathbb{R}^n}\log(\pi_{t_{i+1}}(\boldsymbol{\xi}|\boldsymbol{y}))\pi_t(\boldsymbol{\xi}|\boldsymbol{y})\,\mathrm{d}\boldsymbol{\xi}\right] \\ &= \sum_{i=0}^{T-1}\left[\int_{\mathbb{R}^n}\log\left(\frac{\mathcal{L}(\boldsymbol{\xi}|\boldsymbol{y})^{t_{i+1}}}{\mathcal{L}(\boldsymbol{\xi}|\boldsymbol{y})^{t_i}}\right)\pi_t(\boldsymbol{\xi}|\boldsymbol{y})\,\mathrm{d}\boldsymbol{\xi} + \int_{\mathbb{R}^n}\log\left(\frac{\pi_{t_i}(\boldsymbol{\xi}|\boldsymbol{y})}{\pi_{t_{i+1}}(\boldsymbol{\xi}|\boldsymbol{y})}\right)\pi_t(\boldsymbol{\xi}|\boldsymbol{y})\,\mathrm{d}\boldsymbol{\xi}\right].\end{aligned}$$

Applying the trapezium rule for $t = t_i$ and $t = t_{i+1}$ yields Equation (3.19). Naturally, this leads to the approximation

$$\log\pi((\boldsymbol{y})) \approx \sum_{i=0}^{T-1}\frac{1}{2}(t_{i+1}-t_i)\left(\mathbb{E}_{\pi_{t_{i+1}}(\boldsymbol{\xi}|\boldsymbol{y})}[\log(\mathcal{L}(\boldsymbol{\xi}|\boldsymbol{y}))] + \mathbb{E}_{\pi_{t_i}(\boldsymbol{\xi}|\boldsymbol{y})}[\log(\mathcal{L}(\boldsymbol{\xi}|\boldsymbol{y}))]\right), \tag{3.21}$$

as the second expression involving the Kullback-Leibler divergence introduces small errors for $\pi_{t_i+1} \approx \pi_{t_i}$ only. This approximation coincides with the one in Friel & Pettitt [2008]. The Kullback-Leibler term can be seen as bias for discretely approximating the integral in (3.18). Taking the expectation in (3.21) at each evaluation index t_i is adding to the numerical stability of the marginal likelihood estimate. Nevertheless, for estimating the expectation $\mathbb{E}_{\pi_t(\boldsymbol{\xi}|\boldsymbol{y})}[\log(\mathcal{L}(\boldsymbol{\xi}|\boldsymbol{y}))]$ at an arbitrary index t, a series of samples drawn from the power posterior $\pi_t(\boldsymbol{\xi}|\boldsymbol{y})$ is necessary. This makes thermodynamic integration computationally expensive. Hence, the discretization schedule of the unit interval is of great relevance in order to quickly obtain numerically stable estimations for the marginal likelihood: Friel & Pettitt [2008] propose a power law like division of the unit interval via

$$t_i = (i/T)^c \tag{3.22}$$

$(i = 0, \ldots, T)$ where $T \in \mathbb{N}$ and $c > 0$. Calderhead & Girolami [2009] then show that this scheme also minimizes the Kullback-Leibler bias for linear regression models in the approximation of the logarithm of the marginal likelihood. The application to dynamical systems yielded good results in Calderhead & Girolami [2009] which is why we mostly use their proposals of $T = 30$ and $c = 5$ in this thesis. Although thermodynamic integration is computationally more expensive than the methods in Chapters 3.4.1 and 3.4.2, it performs well on most statistical models including differential equations based systems (Calderhead & Girolami [2009]).

3.4.4 Example: A Gaussian mixture model

We now apply the techniques introduced above to a two component Gaussian mixture model. This rather simple setting is analytically tractable and we are able to depict and practically verify the different concepts. Suppose we are given observations $\boldsymbol{y} = \{y_1, \ldots, y_m\}$ with $y_1, \ldots, y_m \in \mathbb{R}$, such that $y_1, \ldots, y_{m_1} \overset{i.i.d.}{\sim} \mathcal{N}(\mu_1, \sigma^2)$ and $y_{m_1+1}, \ldots, y_m \overset{i.i.d.}{\sim} \mathcal{N}(\mu_2, \sigma^2)$. For known variance σ^2 our two competing models are defined by

(i) a model with $\mu := \mu_1 = \mu_2$, designated $\mathcal{M}_1$. This essentially is a single univariate normal distribution model with $y_1, \ldots, y_m \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$ for the mean $\mu = \mu_1 = \mu_2$ and standard deviation σ.

(ii) a model with possibly $\mu_1 \neq \mu_2$, designated $\mathcal{M}_2$,

This leaves us with one free parameter μ for model $\mathcal{M}_1$ and two free parameters μ_1 and μ_2 for model $\mathcal{M}_2$. The prior distributions are chosen to be conjugate with $\mu = \mu_1 = \mu_2 \sim \mathcal{N}(0, \sigma^2)$ in $\mathcal{M}_1$ and $\mu_1 \sim \mathcal{N}(2, \sigma^2)$ and $\mu_2 \sim \mathcal{N}(-2, \sigma^2)$ in $\mathcal{M}_2$. The likelihood for both models is

$$\mathcal{L}(\mu_1, \mu_2 | \boldsymbol{y}) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^m \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{m_1}(y_i - \mu_1)^2 + \sum_{j=m_1+1}^{m}(y_j - \mu_2)^2\right)\right)$$

yielding after some simple calculations (Appendix C.1 and C.2) the posterior distributions

(i) $\mathcal{N}\left(\frac{\sum_{i=1}^m y_i}{m+1}, \frac{\sigma^2}{m+1}\right)$ for model $\mathcal{M}_1$,

(ii) $\mathcal{N}_2\left(\begin{pmatrix} \frac{2+\sum_{i=1}^{m_1} y_i}{m_1+1} \\ \frac{-2+\sum_{j=m_1+1}^{m} y_j}{m-m_1+1} \end{pmatrix}, \begin{pmatrix} \frac{\sigma^2}{m_1+1} & 0 \\ 0 & \frac{\sigma^2}{m-m_1+1} \end{pmatrix}\right)$ for model $\mathcal{M}_2$.

In order to compute the Bayes factor, we need the marginal likelihoods $\pi(\boldsymbol{y}|\mathcal{M}_1)$ and $\pi(\boldsymbol{y}|\mathcal{M}_2)$. As computed in Appendix C.1 for $\bar{y} = \frac{1}{m}\sum_{i=1}^m y_i$:

$$\pi(\boldsymbol{y}|\mathcal{M}_1) = \frac{1}{\sqrt{m+1}}\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^m \exp\left(\frac{1}{2\sigma^2}\left(\frac{m^2\bar{y}^2}{m+1} - \sum_{i=1}^m y_i^2\right)\right).$$

Similarly, (c.f. Appendix C.2) we have for $\bar{y}_1 = \frac{1}{m_1}\sum_{i=1}^{m_1} y_i$ and $\bar{y}_2 = \frac{1}{m_2}\sum_{j=m_1+1}^{m} y_j$, where $m_2 = m - m_1$

$$\begin{aligned}\pi(\boldsymbol{y}|\mathcal{M}_2) = \frac{1}{\sqrt{(m_1+1)(m_2+1)}}&\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^m \\ &\cdot \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^m y_i^2 + 8 - \frac{(m_1\bar{y}_1+2)^2}{m_1+1} - \frac{(m_2\bar{y}_2-2)^2}{m_2+1}\right)\right).\end{aligned}$$

The Bayes factor B_{21} for model $\mathcal{M}_2$ versus model $\mathcal{M}_1$ is hence

$$\begin{aligned}B_{21} &= \frac{\pi(\boldsymbol{y}|\mathcal{M}_2)}{\pi(\boldsymbol{y}|\mathcal{M}_1)} \\ &= \frac{\sqrt{m+1}}{\sqrt{m_1+1}\sqrt{m_2+1}} \exp\left(-\frac{1}{2\sigma^2}\left(8 + \frac{(m\bar{y})^2}{m+1} - \frac{(m_1\bar{y}_1+2)^2}{m_1+1} - \frac{(m_2\bar{y}_2-2)^2}{m_2+1}\right)\right).\end{aligned} \tag{3.23}$$

Recall that the Bayes factor for $\mathcal{M}_1$ versus $\mathcal{M}_2$ is simply B_{12}^{-1}. For the use in thermodynamic integration we exemplary compute the expected value of the log-likelihood

$\log \mathcal{L}(\mu|\boldsymbol{y})$ of model $\mathcal{M}_1$ under the power posterior $\pi_t(\mu|\boldsymbol{y}, \mathcal{M}_1)$. As shown in Appendix C.3 we have

$$\begin{aligned}\mathbb{E}_{\pi_t(\mu|\boldsymbol{y},\mathcal{M}_1)}\left(\log \mathcal{L}(\mu|\boldsymbol{y})\right) &= \int_{\mathbb{R}} \log \mathcal{L}(\mu|\boldsymbol{y}) \frac{\mathcal{L}(\mu,|\boldsymbol{y})^t \pi(\mu)}{\int_{\mathbb{R}} \mathcal{L}(\mu,|\boldsymbol{y})^t \pi(\mu)\,\mathrm{d}\mu}\,\mathrm{d}\mu \\ &= -\frac{1}{2\sigma^2}\left\{ \left(\sum_{i=1}^{m} \left(y_i^2 + 2\sigma^2 \log(\sqrt{2\pi}\sigma) \right) \right) \right. \\ &\qquad \left. + \frac{m\sigma^2 - 2\bar{y}^2 m^2 t}{mt+1} + \frac{m^3 \bar{y}^2 t^2}{(mt+1)^2} \right\}.\end{aligned}$$

The computations for model $\mathcal{M}_2$ are very similar, but a lot more cumbersome. They hold no more insights and we leave this to the considerate reader.

Now, setting $\sigma^2 = 1$ and sampling ten observations

$$y_1, \ldots, y_5 \overset{i.i.d.}{\sim} \mathcal{N}(1, 1^2) \qquad \text{and} \qquad y_6, \ldots, y_{10} \overset{i.i.d.}{\sim} \mathcal{N}(-1, 1^2)$$

the prior arithmetic mean estimate (including one standard error) of the Bayes factor B_{21} based on ten runs with 100,000 samples per model each was 77.68 ± 0.11. This is close to the true value of 77.47 computed via Equation (3.23). On the other hand, the corresponding posterior harmonic mean estimate (211.50 ± 0.08) overestimated the true value quite strongly. Using the power law division $t_i = (i/T)^c$ with $c = 5$ and $T = 25$, we also computed B_{21} applying thermodynamic integration. At each t_i, we used 4,000 samples to estimate $\mathbb{E}_{\pi_t(\cdot|\boldsymbol{y},\cdot)}\left(\log \mathcal{L}(\cdot|\boldsymbol{y})\right)$, again resulting in a total of 100,000 samples per model for the approximation of B_{21}. Thermodynamic integration yielded the closest result of 77.34 ± 0.16. The average expected value of $\mathbb{E}_{\pi_t(\mu|\boldsymbol{y},\mathcal{M}_1)}\left(\log \mathcal{L}(\mu|\boldsymbol{y})\right)$, -25.40, was very closely approximated by -24.77 ± 10^{-2}.

4

Markov Chain Monte Carlo (MCMC) methods

We have seen in the previous chapter that sampling from some posterior distribution $\pi(\boldsymbol{\xi}|\boldsymbol{y})$ lies at the core of Bayesian inference. Up to this point, we were able to directly draw from $\pi(\boldsymbol{\xi}|\boldsymbol{y})$ using e.g. conjugate prior distributions. However, in most applications $\pi(\boldsymbol{\xi}|\boldsymbol{y})$ is not a standard sampling distribution and we have to turn to more advanced techniques. A solution to this problem is given by Markov Chain Monte Carlo (MCMC) methods: As the name implies, MCMC methods attempt to generate a Markov chain directly drawing from some complex posterior distribution. With the advent of MCMC methods Bayesian inference has skyrocketed in various fields of science and is likely to continue spreading in the future. One of the most successful and influential (Beichl & Sullivan [2000]; Wilkinson [2006]) algorithms was developed by Metropolis and Hastings (Hastings [1970]; Metropolis *et al.* [1953]). In the following we first introduce the basic version of the *Metropolis-Hasting* (MH) *algorithm*. Subsequently, Chapter 4.5 introduces its direct application to model selection. Interestingly, the MH algorithm can also be applied for optimization problems via *simulated annealing* (Chapter 4.6). Finally, we address the issues of dependency and convergence diagnostics of the Markov chains generated in Chapters 4.2 to 4.4.

4.1 The Metropolis-Hastings (MH) algorithm

Historically the Metropolis-Hastings algorithm was introduced by Metropolis and Hastings in Metropolis *et al.* [1953] and Hastings [1970] for integration of complex functions by random sampling: The integral

$$\int_a^b h(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi}$$

for some function $h : \mathbb{R}^n \longrightarrow \mathbb{R}$ can be computed by decomposing $h(\boldsymbol{\xi})$ into the product $f(\boldsymbol{\xi})\pi(\boldsymbol{\xi})$, where $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is a function defined over (a, b) and $\pi(\boldsymbol{\xi})$ a probability density function on (a, b). The integral is then

$$\int_a^b h(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi} = \int_a^b f(\boldsymbol{\xi})\pi(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi} = \mathbb{E}_{\pi(\boldsymbol{\xi})}[h(\boldsymbol{\xi})].$$

Now, given a number of samples $\boldsymbol{\xi}^{(0)}, \ldots, \boldsymbol{\xi}^{(T)}$ of $\pi(\boldsymbol{\xi})$, *Monte Carlo integration* approximates

$$\int_a^b h(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi} = \mathbb{E}_{\pi(\boldsymbol{\xi})}[h(\boldsymbol{\xi})] \approx \frac{1}{T}\sum_{i=1}^{T} f(\boldsymbol{\xi}^{(i)}).$$

We already applied this principle for the prior harmonic mean estimate (see Equation (3.13)), where f was the likelihood function. For sampling from the density function $\pi(\boldsymbol{\xi})$, Metropolis and Hastings used the algorithm depicted in Algorithm 2. Here, the realization $\{\boldsymbol{\xi}^{(j)}\}_{j=0,\ldots,T}$ of a Markov chain $\{\boldsymbol{X}^{(t)}\}_{t\in\mathbb{N}_0}$ is sampled as follows: In each iteration the algorithm generates a proposal $\boldsymbol{\xi}^p$ according to some *transition density* function $q(\boldsymbol{\xi}|\boldsymbol{\xi}^{(j)})$ that (possibly) depends on the current element $\boldsymbol{\xi}^{(j)}$ of the Markov chain. It is accepted with the *Metropolis-Hastings acceptance probability*

$$\alpha(\boldsymbol{\xi}^p|\boldsymbol{\xi}^{(j)}) = \min\left\{\frac{\pi(\boldsymbol{\xi}^p)q(\boldsymbol{\xi}^{(j)}|\boldsymbol{\xi}^p)}{\pi(\boldsymbol{\xi}^{(j)})q(\boldsymbol{\xi}^p|\boldsymbol{\xi}^{(j)})}, 1\right\}. \tag{4.1}$$

If $\boldsymbol{\xi}^p$ is accepted, the Markov chain element $\boldsymbol{\xi}^{(j+1)}$ is defined by $\boldsymbol{\xi}^{(j+1)} = \boldsymbol{\xi}^p$ and by $\boldsymbol{\xi}^{(j+1)} = \boldsymbol{\xi}^{(j)}$ otherwise. The transition density function $q(\boldsymbol{\xi}|\boldsymbol{\xi}')$ is also called *proposal function* and, like the density function $\pi(\boldsymbol{\xi})$, only needs to be explicitly available up to a multiplicative constant independent of $\boldsymbol{\xi}'$; this is due to the fact that these constants cancel out in the Metropolis-Hastings acceptance probability. We call the percentage of accepted MH steps the *acceptance rate* of a realization $\boldsymbol{\xi}^{(0)}, \ldots, \boldsymbol{\xi}^{(T)}$.

Algorithm 2: The Metropolis-Hastings algorithm

Input: Initial value $\boldsymbol{\xi}^{(0)} \in \mathbb{R}^n$, (transition) density function $q(\boldsymbol{\xi}|\boldsymbol{\xi}')$ on $\mathbb{R}^n$ such that $q(\boldsymbol{\xi}|\boldsymbol{\xi}^{(0)})$ exists, arbitrary target density function $\pi : \mathbb{R}^n \longrightarrow \mathbb{R}$, and chain length $T \in \mathbb{N}$.

Output: Markov chain realization $\{\boldsymbol{\xi}^{(j)}\}_{j=0,\dots,T}$.

for $j \leftarrow 0$ **to** $T-1$ **do**

Generate proposal $\boldsymbol{\xi}^p \sim q(\boldsymbol{\xi}|\boldsymbol{\xi}^{(j)})$

Set

$$\boldsymbol{\xi}^{(j+1)} \leftarrow \begin{cases} \boldsymbol{\xi}^p & \text{with probability } \alpha(\boldsymbol{\xi}^p|\boldsymbol{\xi}^{(j)}), \\ \boldsymbol{\xi}^{(j)} & \text{with probability } 1-\alpha(\boldsymbol{\xi}^p|\boldsymbol{\xi}^{(j)}), \end{cases}$$

where

$$\alpha(\boldsymbol{\xi}^p|\boldsymbol{\xi}^{(j)}) = \min\left\{\frac{\pi(\boldsymbol{\xi}^p)q(\boldsymbol{\xi}^{(j)}|\boldsymbol{\xi}^p)}{\pi(\boldsymbol{\xi}^{(j)})q(\boldsymbol{\xi}^p|\boldsymbol{\xi}^{(j)})}, 1\right\}.$$

In Bayesian model inference we use the MH algorithm for sampling from complex posterior distributions $\pi(\boldsymbol{\xi}|\boldsymbol{y})$. Since the algorithm is quite general, we impose the *regularity condition* $q(\boldsymbol{\xi}|\boldsymbol{\xi}') > 0$ for all $\boldsymbol{\xi}, \boldsymbol{\xi}'$ in the support of $\pi(\boldsymbol{\xi}|\boldsymbol{y})$ in order to avoid convergence issues. This means the proposal function $q(\boldsymbol{\xi}|\boldsymbol{\xi}')$ allows to proceed from any point to any other point on the support of $\pi(\boldsymbol{\xi}|\boldsymbol{y})$.

Theorem 4.1 (Convergence of the MH algorithm)**.** *Let $\{\boldsymbol{X}^{(t)}\}_{t\in\mathbb{N}_0}$ be a Markov chain governed by Algorithm 2 with target density function $\pi(\boldsymbol{\xi})$. Suppose the proposal function $q(\boldsymbol{\xi}|\boldsymbol{\xi}')$ fulfills the regularity condition. Then π is the equilibrium distribution of $\{\boldsymbol{X}^{(t)}\}_{t\in\mathbb{N}_0}$.*

Proof. According to Theorem 2.5 we have to show that $\{\boldsymbol{X}^{(t)}\}_{t\in\mathbb{N}_0}$ is (i) π-irreducible, (ii) Harris recurrent and (iii) aperiodic with (iv) invariant distribution π.

(i) Due to the regularity condition for q, (i) is naturally satisfied.

(ii) According to Lemma 7.3 in Robert & Casella [2004] π-irreducibility of $\{\boldsymbol{X}^{(t)}\}_{t\in\mathbb{N}_0}$ also implies Harris recurrence (a proof involves further theory on bounded *harmonic functions* and *tail events*, which, however, is beyond the scope of this thesis. Nevertheless, the interested reader may be referred to Robert & Casella [2004] or Nummelin [2004]).

(iii) The aperiodicity condition follows as for each $t \in \mathbb{N}$, $\boldsymbol{X}^{(t)} = \boldsymbol{X}^{(t+1)}$ with positive probability, this is, the Markov chain can stay in place in every step.

(iv) For proving invariance, we show that the kernel corresponding to $\{\boldsymbol{X}^{(t)}\}_{t\in\mathbb{N}_0}$ fulfills the detailed balance condition: The transition kernel density function associated with $\{\boldsymbol{X}^{(t)}\}_{t\in\mathbb{N}_0}$ is given by

$$k(\boldsymbol{\xi}'|\boldsymbol{\xi}) = \alpha(\boldsymbol{\xi}'|\boldsymbol{\xi})q(\boldsymbol{\xi}'|\boldsymbol{\xi}) + r(\boldsymbol{\xi})\mathbb{1}_{\boldsymbol{\xi}}(\boldsymbol{\xi}'),$$

where for the the support of S of π, $r(\boldsymbol{\xi}) = 1 - \int_S q(\boldsymbol{\xi}'|\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi}'$, as introduced in Chapter 2.3. Clearly,

$$r(\boldsymbol{\xi})\mathbb{1}_{\boldsymbol{\xi}}(\boldsymbol{\xi}')p(\boldsymbol{\xi}) = r(\boldsymbol{\xi}')\mathbb{1}_{\boldsymbol{\xi}'}(\boldsymbol{\xi})p(\boldsymbol{\xi}')$$

for any two realizations $\boldsymbol{\xi}$ and $\boldsymbol{\xi}'$ of $\boldsymbol{X}^{(t)}$ and $\boldsymbol{X}^{(t+1)}$ ($t \in \mathbb{N}_0$). Furthermore, without loss of generality let $\alpha(\boldsymbol{\xi}'|\boldsymbol{\xi}) < 1$ (while the case $\alpha(\boldsymbol{\xi}'|\boldsymbol{\xi}) = 1$ is trivial, exchanged roles of $\boldsymbol{\xi}$ and $\boldsymbol{\xi}'$ yield the case $\alpha(\boldsymbol{\xi}|\boldsymbol{\xi}') < 1$). Then

$$\alpha(\boldsymbol{\xi}'|\boldsymbol{\xi}) = \frac{\pi(\boldsymbol{\xi}')q(\boldsymbol{\xi}|\boldsymbol{\xi}')}{\pi(\boldsymbol{\xi})q(\boldsymbol{\xi}'|\boldsymbol{\xi})} \qquad \text{and} \qquad \alpha(\boldsymbol{\xi}|\boldsymbol{\xi}') = 1.$$

Hence,

$$\begin{aligned}\alpha(\boldsymbol{\xi}'|\boldsymbol{\xi})q(\boldsymbol{\xi}'|\boldsymbol{\xi})\pi(\boldsymbol{\xi}) &= \frac{\pi(\boldsymbol{\xi}')q(\boldsymbol{\xi}|\boldsymbol{\xi}')}{\pi(\boldsymbol{\xi})q(\boldsymbol{\xi}'|\boldsymbol{\xi})}q(\boldsymbol{\xi}'|\boldsymbol{\xi})\pi(\boldsymbol{\xi}) \\ &= q(\boldsymbol{\xi}|\boldsymbol{\xi}')\pi(\boldsymbol{\xi}') \\ &= \alpha(\boldsymbol{\xi}|\boldsymbol{\xi}')q(\boldsymbol{\xi}|\boldsymbol{\xi}')\pi(\boldsymbol{\xi}')\end{aligned}$$

and $k(\boldsymbol{\xi}'|\boldsymbol{\xi})$ satisfies the detailed balance condition. Theorem 2.4 now provides that π is an invariant distribution for the Markov chain, which finalizes the proof.

□

A very popular choice for the proposal density $q(\boldsymbol{\xi}|\boldsymbol{\xi}')$ is the n-dimensional normal distribution $\mathcal{N}_n(\boldsymbol{\xi}', \boldsymbol{\Sigma})$ with mean $\boldsymbol{\xi}'$ and covariance matrix $\boldsymbol{\Sigma}$. In each MCMC step a new proposal is generated based on the current sample $\boldsymbol{\xi}^{(c)}$ as $\boldsymbol{\xi}^p = \boldsymbol{\xi}^{(c)} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{\Sigma})$. We refer to this scheme as *Random Walk Metropolis-Hastings* (RWMH) *algorithm*. In case there is no knowledge about the covariance structure of the parameters, $\boldsymbol{\Sigma} = k_{RW}\boldsymbol{I}_n$ is typically chosen, where k_{RW} is the *step-size tuning* or *scaling parameter* and $\boldsymbol{I}_n$ the n-dimensional identity matrix. If the proposal function in each iteration is independent of the current sample, i.e. if $q(\boldsymbol{\xi}|\boldsymbol{\xi}') = q(\boldsymbol{\xi})$, the sampling scheme

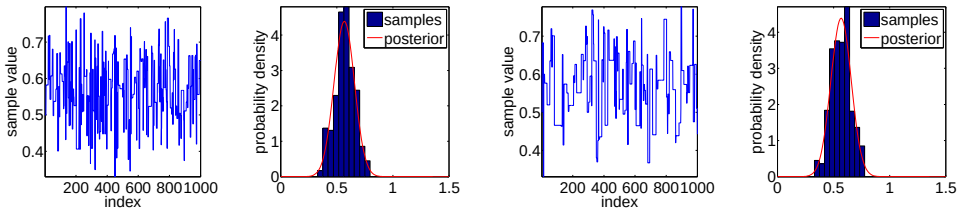


Figure 4.1: (a) Realization of a Markov chain generated by the random walk proposal function q_1; (b) corresponding histogram for the samples in (a) and "true" posterior density. (c) Realization of a Markov chain generated by the independence chain proposal function q_2; (d) corresponding histogram for the samples in (c) and "true" posterior density.

is called an *Independence chain Metropolis-Hastings* (IMH) *algorithm*. There are generally no limitations regarding the choice of q as long as it is positive on the support of the posterior distribution $\pi(\boldsymbol{\xi}|\boldsymbol{y})$. Its choice however is very crucial for the performance of the algorithm: An efficient MH algorithm shows high acceptance rates for the proposed samples at simultaneously low autocorrelation between generated Markov chain elements (see Chapter 4.3). Especially in high dimensions this is hard to attain, because small update step sizes result in high acceptance rates, but also in highly correlated Markov chain samples and vice versa.

Example 4.1 (Mean of a normal distribution). Suppose we want to generate the realization of a Markov chain from the posterior distribution of the one-component model $\mathcal{M}_1$ of Chapter 3.4.4 using the MH algorithm with (i) a random walk proposal function q_1 and (ii) an independence proposal function q_2. Given the i.i.d. observations $y_1, \ldots, y_{10} \sim \mathcal{N}(\mu = 0, 1^2)$ and the prior distribution $\mathcal{N}(0, 1^2)$ for the parameter μ the posterior distribution is then given by $\mathcal{N}\left(\frac{\sum_{i=1}^{10} y_i}{11}, \frac{1}{11}\right)$ (Chapter 3.4.4). Starting at $\boldsymbol{\xi}^{(0)} = 0.5$ we generate the proposal $\boldsymbol{\xi}^p$ based on the current Markov chain sample $\boldsymbol{\xi}^{(c)}$ as

$$\boldsymbol{\xi}^p \sim \mathcal{N}(\boldsymbol{\xi}^{(c)}, 0.5^2) \text{ in case of } q_1 \qquad \text{and} \qquad \boldsymbol{\xi}^p \sim \mathcal{N}(1, 1^2) \text{ in case of } q_2.$$

Figures 4.1(a) and 4.1(c) hold 1,000 realizations for q_1 and q_2, respectively. The according histograms of these realizations as well as the "true" posterior distribution and depicted in Figures 4.1(b) and 4.1(d). Note that our samples are no independent realizations of the posterior distribution as the Markov chain inherits some intrinsic dependency.

4.2 Independent identically distributed samples from a Markov chain

From a general point of view, we try to draw i.i.d. samples from a density π by means of a Markov chain $\{\boldsymbol{X}^{(t)}\}_{t\in\mathbb{N}_0}$. At first sight, this seems to be a conflicting goal as any realization of $\{\boldsymbol{X}^{(t)}\}_{t\in\mathbb{N}_0}$ can at best only approximate π and generating independent samples based on a dependent process seems unintuitive. However, in order to address the dependency issue we can make use of the limited memory of the Markov chain in combination with the so-called *Chapman-Kolmogorov equations* (see Papoulis *et al.* [1965] for a proof):

Lemma 4.1 (Chapman-Kolmogorov equations)**.** *Suppose $\{\boldsymbol{X}^{(t)}\}_{t\in\mathbb{N}_0}$ is a Markov chain on a probability space $(\Omega, \mathcal{F}, P)$ with values in a measurable space $(E, \mathcal{E})$. Suppose further $A \in \mathcal{E}$, $\boldsymbol{x} \in E$ and $k^m(A|\boldsymbol{x})$ denotes the m-step transition kernel corresponding to $\{\boldsymbol{X}^{(t)}\}_{t\in\mathbb{N}_0}$. It holds for every $(m_1, m_2) \in \mathbb{N}^2$*

$$k^{m_1+m_2}(A|\boldsymbol{x}) = \int_E k^{m_2}(A|\boldsymbol{y})k^{m_1}(\mathrm{d}\boldsymbol{y}|\boldsymbol{x}). \tag{4.2}$$

This means in order to get from $\boldsymbol{x}$ to A in $m_1 + m_2$ steps, we need to pass through some $\boldsymbol{y}$ on the m_2^{th} step. Thus, for the realization of a Markov chain $\{\boldsymbol{x}^{(j)}\}_{j\in\{0,\dots,T\}}$ and any two natural numbers i', i with $0 < i \leq T$

$$k^{i'}(A|\boldsymbol{x}^{(i)}) \leq \int_E k^{i'}(A|\boldsymbol{y})k(\mathrm{d}\boldsymbol{y}|\boldsymbol{x}^{(i-1)}) = k^{i'+1}(A|\boldsymbol{x}^{(i-1)})$$

The chain in some sense "loses" its memory on the states visited a long time back in the past. Therefore, taking every i^{th} sample only, provides more or less independent samples from the distribution π. How to infer i is derived in Chapter 4.3 with the help of the *autocorrelation function*, a statistic commonly used for analyzing time series data in the field of signal processing.

As far as the convergence of a Markov chain to a limiting distribution π is concerned, we need to keep in mind that computational methods are always discrete due to calculation accuracy. Thus, all we can hope for solving a continuous problem using a computational approach is a good approximation of the problem at hand. Recalling that we are dealing with dependent Markov chains, we nevertheless want to start the sampling procedure

in an area where π possesses a considerable amount of mass. This can e.g. be achieved using the *simulated annealing* algorithm introduced in Chapter 4.6 or the convergence statistics of Chapter 4.4.

4.3 A measure for independence

We now turn to the problem of determining the right amount of thinning in order to resolve the dependency issue of a Markov chain. For the rest of this chapter, we assume that the first and second moment of the posterior distribution $\pi(\cdot|\cdot)$ exist. To introduce the basic concept of thinning, we for now assume that the elements of the following random processes are univariate.

Definition 4.1 (Autocorrelation and autocovariance)**.** Let $\{X^{(t)}\}_{t\in\mathbb{N}_0}$ be a Markov chain. The *autocovariance* is defined as function

$$\gamma : \mathbb{N}_0 \times \mathbb{N}_0 \longrightarrow \mathbb{R},\ (s,t) \mapsto \mathbb{E}[(X^{(t)} - \mu_t)(X^{(s)} - \mu_s)],$$

where μ_t and μ_s denote the means of $X^{(t)}$ and $X^{(s)}$. If furthermore $\{X^{(t)}\}_{t\in\mathbb{N}_0}$ is non-Dirac, i.e. for all t the random vector $X^{(t)}$ is not constant (we exclude random varibles with Dirac distribution and therefore standard deviation 0), the *autocorrelation* is defined as

$$\rho : \mathbb{N}_0 \times \mathbb{N}_0 \longrightarrow \mathbb{R},\ (s,t) \mapsto \frac{\mathbb{E}[(X^{(t)} - \mu_t)(X^{(s)} - \mu_s)]}{\sigma_t \sigma_s},$$

where σ_t and σ_s are the respective standard deviations of $X^{(t)}$ and $X^{(s)}$.

The autocovariance takes on values in $[-\infty, +\infty]$ and the autocorrelation in $[-1, 1]$, similar to the covariance and correlation statistics. Our goal is to generate samples from the probability density function $\pi(\cdot|\cdot)$ based on a Markov chain $\{X^{(t)}\}_{t\in\mathbb{N}_0}$. Here, $\{X^{(t)}\}_{t\in\mathbb{N}_0}$ needs to be a stationary stochastic process. This means $X^{(t)}$ is independent of the index t and

$$\mu := \mathbb{E}[X^{(t)}] = \mu_t = \mu_s = \mathbb{E}[X^{(s)}]$$

for all $t, s \in \mathbb{N}_0$. In this setting, we somewhat abuse the notation of the autocorrelation and autocovariance above and write

$$\begin{aligned}\gamma(t,s) = \mathbb{E}[(X^{(t)} - \mu_t)(X^{(s)} - \mu_s)] &= \mathbb{E}[(X^{(t)} - \mu)(X^{(s)} - \mu)] \\ &= \mathbb{E}[(X^{(t)} - \mu)(X^{(t+\tau)} - \mu)] = \gamma(\tau),\end{aligned}$$

and also

$$\begin{aligned}\rho(t,s) = \frac{\mathbb{E}[(X^{(t)} - \mu_t)(X^{(s)} - \mu_s)]}{\sigma_t \sigma_s} &= \frac{\mathbb{E}[(X^{(t)} - \mu)(X^{(s)} - \mu)]}{\sigma^2} \\ &= \frac{\mathbb{E}[(X^{(t)} - \mu)(X^{(t+\tau)} - \mu)]}{\sigma^2} = \rho(\tau),\end{aligned}$$

for $\tau = s - t$ and $\sigma^2 = \sigma_t^2 = \sigma_s^2$.

The autocorrelation describes the correlation between all possible pairs $(X^{(t)}, X^{(s)})$ in the Markov chain. This definition also includes negative lags $\tau < 0$, which will not be of further interest to us, since γ and ρ are clearly symmetric, this is,

$$\gamma(\tau) = \gamma(-\tau) \quad \text{and} \quad \rho(\tau) = \rho(-\tau).$$

We now determine the thinning rate r for a realization $\{\xi^{(t)}\}_{t=1,\dots,T}$ of the Markov chain. More precisely, we want to infer the parameter r, such that the chain $\{\xi^{(t')}\}_{t' \in J}$ with $J = \{0, r, 2r, \dots, \lfloor \frac{T}{r} \rfloor \cdot r\}$ has a very low autocorrelation value for all lags $\tau \neq 0$ (compare Neal [1993]). Note that $\rho(\tau = 0) = 1$. For estimating r only very few approaches exist. For one, visual inspection of the sample autocorrelation function of the Markov chain $\{\boldsymbol{\xi}^{(t)}\}_t$ can help to determine the size of r. A sharp decrease in all dimensions should here be obtained within the first few lags after thinning. A widely used estimate for r, the so-called *INEFficiency Factor* (INEFF), is also based on autocorrelation functions (Bartlett [1966]; Kass [1993]; Kass *et al.* [1998]):

Lemma 4.2 (Inefficiency factor (INEFF))**.** *For the realization of a stationary Markov chain $\{\xi^{(t)}\}_{t \in \{1,\dots,T\}}$ and the autocorrelation function $\rho(\tau)$ we can consider samples that are at least*

$$1 + 2 \sum_{\tau=1}^{T} \left(1 - \frac{\tau}{T+1}\right) \rho(\tau)$$

indices apart as independent.

Proof. The idea is to monitor the effect of thinning on the expected variance of $\xi := \{\xi^{(t)}\}_t$. For the expected mean

$$\bar{\xi} = \frac{1}{T+1}\sum_{t=0}^{T}\xi^{(t)}, \tag{4.3}$$

the law of large numbers guarantees $\bar{\xi} \longrightarrow \mathbb{E}[\xi]$ almost surely for $T \to \infty$. Now, if the elements of the chain were pairwise independent, we would get the well-known result

$$\mathrm{Var}[\bar{\xi}] = \mathrm{Var}\left[\frac{1}{T+1}\sum_{t=0}^{T}\xi^{(t)}\right] = \frac{1}{(T+1)^2}\sum_{t=0}^{T}\mathrm{Var}\left[\xi^{(t)}\right] = \frac{\sigma^2}{T+1}.$$

Due to the autocorrelation in the chain the variance of the estimator $\bar{\xi}$ is given by

$$\begin{aligned}
\mathrm{Var}[\bar{\xi}] &= \mathbb{E}[(\bar{\xi} - \mathbb{E}[\xi])^2] \\
&\overset{(4.3)}{=} \mathbb{E}\left[\left(\frac{1}{T+1}\sum_{t=0}^{T}(\xi^{(t)} - \mathbb{E}[\xi])\right)^2\right] \\
&= \frac{1}{(T+1)^2}\sum_{t,t'=0}^{T}\mathbb{E}\left[(\xi^{(t)} - \mathbb{E}[\xi])(\xi^{(t')} - \mathbb{E}[\xi])\right] \\
&= \frac{1}{(T+1)^2}\sum_{t,t'=0}^{T}\gamma(t'-t) \\
&= \frac{1}{(T+1)^2}\left(\sum_{-T\le\tau\le T}(\gamma(\tau)\,(T+1-|\tau|)) + (T+1)\gamma(0)\right) \\
&= \frac{1}{T+1}\sum_{\substack{-T\,\le\,\tau\,\le\,T \\ \tau\,\neq\,0}}\left(1 - \frac{|\tau|}{T+1}\right)\gamma(\tau) + \frac{\sigma^2}{T+1} \\
&= \frac{\sigma^2}{T+1}\cdot\left(1 + 2\sum_{\tau=1}^{T}\left(1 - \frac{\tau}{T+1}\right)\rho(\tau)\right),
\end{aligned}$$

which proofs the claim. □

Generally, the autocorrelation function is not known explicitly and we have to use the *sample autocorrelation function*

$$\hat{\rho}(\tau) = \frac{1}{(T+1-\tau)\hat{\sigma}^2}\sum_{j=0}^{T-\tau}(\xi^{(j)} - \bar{\xi})(\xi^{(j+\tau)} - \bar{\xi})$$

as estimator for $\rho(\tau)$ instead. Here, $\hat{\sigma}^2 = \frac{1}{T}\sum_{j=0}^{T}(\xi^{(j)} - \bar{\xi})^2$ denotes the expected variance. For an n-dimensional Markov chain we compute the sample autocorrelation function $\hat{\rho}_i(\tau)$ along each dimension i and estimate the INEFF by

$$\hat{r} := \max_{i=1,\dots,n} \left\{ \left(1 + 2 \sum_{\tau=1}^{T} \left(1 - \frac{\tau}{T+1} \right) \hat{\rho}_i(\tau) \right) \right\}.$$

A related, frequently used statistic is the *Effective Sampling Size* (ESS) defined by $ESS = T/\hat{r}$.

4.4 Convergence to the stationary distribution

Another important issue for sampling from a Markov chain is its starting value $\boldsymbol{\xi}^{(0)}$. Choosing $\boldsymbol{\xi}^{(0)}$ arbitrarily can lead to convergence issues as can be seen in the following simple example:

Example 4.2 (Sampling from a one dimensional normal distribution). Say we want to generate samples from the univariate standard normal distribution $\mathcal{N}(0, 1^2)$ using a simple random walk Metropolis-Hastings sampler. We compare the convergence of the chain to $\mathcal{N}(0, 1^2)$ taking $\boldsymbol{\xi}^{(0)} = 0.1$ and $\boldsymbol{\xi}^{(0)} = 20$ as initial values. In both cases we tune the step-size of the Metropolis-Hastings algorithm to 2.4, which is optimal for convergence as shown by Gelman *et al.* [1996], i.e. the Metropolis-Hastings update function is given by $\boldsymbol{\xi}^{p} \sim \mathbb{N}(x^{(c)}, 2.4^2)$. Our quality measure is to approximate the true mean $\mu = 0$ and the standard deviation $\sigma = 1$ within an $[-0.05, 0.05]$ and $[0.95, 1.05]$ error bound, respectively. Each of the 100 runs is thinned by computing the INEFF. While starting at $\boldsymbol{\xi}^{(0)} = 0.1$ needs in average 137 steps to converge, starting at $\boldsymbol{\xi}^{(0)} = 20$ takes 7,257 steps. This is more than 50 times as much as for $\boldsymbol{\xi}^{(0)} = 0.1$. The reason for this behavior is the sometime slow proceeding of the Markov chain towards regions with high mass in the probability mass function.

In the following we introduce *convergence statistics* in order to remedy this effect. These can be split into two classes: While the first class is based on multiple chains, i.e. a set of Markov chains is run in parallel, generally starting at different initial values, the second class is based on a single chain. Both determine the number of samples to be removed before we can consider the Markov chain to be stationary. This limit is called *burn-in period* of the Markov chain.

We first consider one of the most prominent single chain methods: The *Geweke test* (Geweke [1992]) splits the realization of a Markov chain $\{\boldsymbol{\xi}^{(j)}\}_{j=0,\dots,T}$ into two distinct subsamples. It generally takes the first 10% and last 50% of the samples. For a stationary Markov chain the mean of these parts should be approximately equal. Therefore, in order to remove the burn-in period of $\{\boldsymbol{\xi}^{(j)}\}_{j=0,\dots,T}$ a *z-score* can be used, i.e. for given z_0 (generally $z_0 = 2$ is chosen) we monitor the number of samples m, such that for $a_m, b_m \in \mathbb{N}$ with $a_m + b_m + m \leq T$

$$\left| \frac{\bar{\boldsymbol{\xi}}_{a_m} - \bar{\boldsymbol{\xi}}_{b_m}}{\sqrt{\hat{\sigma}_{a_m} + \hat{\sigma}_{b_m}}} \right| < z_0$$

where $|\cdot|$ denotes the absolute value, $\hat{\sigma}.$ the according expected variances, and $\bar{\boldsymbol{\xi}}_{a_m} = \frac{1}{a_m} \sum_{j=m+1}^{m+a_m} \boldsymbol{\xi}^{(j)}$ and $\bar{\boldsymbol{\xi}}_{b_m} = \frac{1}{b_m} \sum_{j=m+T-b_m+1}^{m+T} \boldsymbol{\xi}^{(j)}$. Here, the natural number a_m is the index corresponding to the first 10% of the reduced Markov chain $\{\boldsymbol{\xi}^{(j)}\}_{j=m,\dots,T}$, b_m the index corresponding to the last 50% of this chain.

Let us now turn to the most prominent multiple chains method: The *Gelman-Rubin statistic* $\hat{R}$ (Brooks & Gelman [1998]; Gelman & Rubin [1992]) compares the variances within each chain to the variance between the chains. For a stationary Markov chain, these two statistics should coincide. More precisely, suppose we are given L realizations $\{\boldsymbol{\xi}_l^{(j)}\}_{j=0,\dots,T}$ of the Markov chain $\{\boldsymbol{X}^{(t)}\}_{t\in\mathbb{N}}$ starting at different initial values $\boldsymbol{\xi}_l^{(0)}$ $(l = 1, \dots, L)$. We monitor the number of samples m, such that for the between-chain variance B

$$B(m) = \frac{T-m}{L-1} \sum_{l=1}^{L} (\bar{\boldsymbol{\xi}}_l(m) - \bar{\boldsymbol{\xi}}(m))^2$$

where

$$\bar{\boldsymbol{\xi}}_l(m) = \frac{1}{T-m} \sum_{j=m+1}^{T} \boldsymbol{\xi}_l^{(j)} \qquad \text{and} \qquad \bar{\boldsymbol{\xi}}(m) = \frac{1}{L} \sum_{l=1}^{L} \bar{\boldsymbol{\xi}}_l(m)$$

and the within-chain variance W

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

where

$$\hat{\sigma}_l^2(m) = \frac{1}{T-m-1} \sum_{j=m+1}^{L} (\boldsymbol{\xi}_l^{(j)} - \bar{\boldsymbol{\xi}}_l(m))^2$$

the Gelman-Rubin statistic

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

is close to one (generally a limit of 1.2 is chosen), where

$$\hat{\sigma}(m) = (1 - 1/(T-m))W(m) + B(m)/(T-m).$$

The Gelman-Rubin statistic can also be computed using a single realization $\{\boldsymbol{\xi}^{(j)}\}_{j=0,\ldots,T}$ only. For this, we split apart $\{\boldsymbol{\xi}^{(j)}\}_{j=0,\ldots,T}$ into a number of subsamples of equal length. The subsamples are then considered as parallel chains and the Gelman-Rubin statistic is computed as described above. This raises the question, whether a single chain or multiple chains starting at different initial values should be used for posterior inference? In this thesis we will follow Geyer [1992] who argues that a single longer chain is superior to multiple chain approaches as various smaller chains may not reach convergence during the sampling process. We therefore do not consider multi-chain MCMC methods any further.

4.5 Reversible jump MCMC

We have seen in Chapter 3.4 how to apply Bayesian methods for the purpose of model selection using the posterior odds ratio $\frac{\pi(\mathcal{M}_1|\boldsymbol{y})}{\pi(\mathcal{M}_2|\boldsymbol{y})}$ for two competing models $\mathcal{M}_1$ and $\mathcal{M}_2$ based on the observations $\boldsymbol{y}$. Given a series of models $\mathcal{M}_1, \ldots, \mathcal{M}_k$, pairwise application of the posterior odds ratio can be used in order to infer the best model for $\boldsymbol{y}$. For each $\mathcal{M}_i$ the MH algorithm from Chapter 4.1 may be applied for inference of the according posterior distribution. The sampling is thereby restricted on the specific parameter space Ξ_i of $\mathcal{M}_i$. Naturally, there has been extensive research in order to simultaneously cover the issue of model selection and model inference (see e.g. Geyer & Møller [1994]; Ripley [1977]). This is of importance, if the number of models is denumerable, as could be the case for Gaussian mixture models (compare Chapter 3.4.4) with an unknown number of components, or for the application in *model averaging* (see e.g. Hastie *et al.* [2009] Chapter 8.8). However, a general formalization has only rather recently been presented by Green [1995]:

Suppose we are given a series of models $\{\mathcal{M}_i\}_{i\in I}$ on some discrete index set I with according proper parameter prior distributions $\pi(\boldsymbol{\xi}_i|\mathcal{M}_i)$ on the parameter space Ξ_i for model $\mathcal{M}_i$ and a model prior distribution $\pi(\mathcal{M}_i)$ holding the probability distribution for model $\mathcal{M}_i$. Then the joint prior for the parameter $\boldsymbol{\xi}_i$ in model $\mathcal{M}_i$ on the parameter space

$$\Xi = \bigcup_i \{i\} \times \Xi_i$$

is given by

$$\pi(\boldsymbol{\xi}_i, \mathcal{M}_i) = \pi(\boldsymbol{\xi}_i|\mathcal{M}_i)\pi(\mathcal{M}_i)$$

with respective Lebesgue measure induced by $\pi(\boldsymbol{\xi}_i|\mathcal{M}_i)$. For the jump between two parameter spaces Ξ_i and Ξ_j, Green's approach is based on the bijective identification of artificial extensions of these spaces. More precisely, we have to define a series of artificial sets $\{U_{ij}\}_{ij\in I\times I}$ together with deterministic bijections

$$\begin{aligned} T_{ij} : \Xi_i \times U_{ij} &\longrightarrow \Xi_j \times U_{ji} \\ (\boldsymbol{\xi}_i, \boldsymbol{u}_{ij}) &\longmapsto (\boldsymbol{\xi}_j, \boldsymbol{u}_{ji}). \end{aligned}$$

This is known as *dimension matching condition*. Note that $T_{ij} \neq T_{ji}^{-1}$ is possible, albeit not common. In general, a move from $\boldsymbol{\xi}_i \in \Xi_i$ to some $\boldsymbol{\xi}_j \in \Xi_j$ is proposed by sampling $u_i \sim g_i(\boldsymbol{u}_i)$ according to some proper proposal function $g_i(\cdot)$ and setting

$$(\boldsymbol{\xi}_j, \boldsymbol{u}_j) = T_{ij}(\boldsymbol{\xi}_i, \boldsymbol{u}_i).$$

The reverse move is given by sampling $u_j \sim g_j(u_j)$ for some proper proposal function $g_j(\cdot)$ such that $(\boldsymbol{\xi}_i, \boldsymbol{u}_i) = T_{ij}^{-1}(\boldsymbol{\xi}_j, \boldsymbol{u}_j)$. Green [1995] showed that the according acceptance probability for the move from $\mathcal{M}_i$ to $\mathcal{M}_j$ is then

$$\alpha(j, \boldsymbol{\xi}_j|i, \boldsymbol{\xi}_i) = \min\left\{ \frac{\mathcal{L}_j(\boldsymbol{\xi}_j|\boldsymbol{y})\pi(\boldsymbol{\xi}_j, j)p_{ji}g_j(\boldsymbol{u}_j)}{\mathcal{L}_i(\boldsymbol{\xi}_i|\boldsymbol{y})\pi(\boldsymbol{\xi}_i, i)p_{ij}g_i(\boldsymbol{u}_i)} \left|\frac{\partial T_{ij}(\boldsymbol{\xi}_i, \boldsymbol{u}_i)}{\partial(\boldsymbol{\xi}_i, \boldsymbol{u}_i)}\right|, 1 \right\},$$

where $\mathcal{L}_k(\boldsymbol{\xi}_k|\boldsymbol{y})$ is the likelihood function for model $\mathcal{M}_k$ given the observations $\boldsymbol{y}$, $\left|\frac{\partial T_{ij}(\boldsymbol{\xi}_i, \boldsymbol{u}_i)}{\partial(\boldsymbol{\xi}_i, \boldsymbol{u}_i)}\right|$ is the Jacobian of the transformation T_{ij} at $(\boldsymbol{\xi}_i, \boldsymbol{u}_{ij})$, p_{kl} is the probability for jumping from $\mathcal{M}_k$ to $\mathcal{M}_l$, and g_k is the density function of $\boldsymbol{u}_k$. As in the classical MH algorithm the posterior distribution needs to be known up to a multiplicative constant. However, all prior distributions $\pi(\boldsymbol{\xi}_i|\mathcal{M}_i)$ need to be normalized up to the same multiplicative constant. The according sampling scheme is called *reversible jump MCMC*

(RJMCMC) algorithm. It explores the joint model/parameter space posterior distribution $\pi(\boldsymbol{\xi}_i, \mathcal{M}_i|\boldsymbol{y})$ based on the observations $\boldsymbol{y}$ and can therefore be used for simultaneous model selection and inference. For a finite set of models $\{\mathcal{M}_i\}_{i=1,\dots,k}$ RJMCMC draws samples from the posterior distribution

$$\pi(\boldsymbol{\xi}_i, \mathcal{M}_i|\boldsymbol{y}) = \frac{\mathcal{L}_i(\boldsymbol{\xi}_i|\boldsymbol{y})\pi(\boldsymbol{\xi}_i|\mathcal{M}_i)\pi(\mathcal{M}_i)}{\pi(\boldsymbol{y})}$$

with normalizing constant

$$\pi(\boldsymbol{y}) = \sum_{j=1}^{k} \int_{\Xi_j} \mathcal{L}_j(\boldsymbol{\xi}_j|\boldsymbol{y})\pi(\boldsymbol{\xi}_j|\mathcal{M}_j)\pi(\mathcal{M}_j)\,\mathrm{d}\boldsymbol{\xi}_j.$$

The relation

$$\begin{aligned}\pi(\mathcal{M}_i|\boldsymbol{y}) &= \int_{\Xi_i} \pi(\boldsymbol{\xi}_i, \mathcal{M}_i|\boldsymbol{y})\,\mathrm{d}\boldsymbol{\xi}_i \\ &= \frac{1}{\pi(\boldsymbol{y})}\pi(\mathcal{M}_i) \int_{\Xi_i} \mathcal{L}_i(\boldsymbol{\xi}_i|\boldsymbol{y})\pi(\boldsymbol{\xi}_i|\mathcal{M}_i)\,\mathrm{d}\boldsymbol{\xi}_i\end{aligned}$$

naturally provides the Bayes factor by RJMCMC via

$$B_{ij} = \frac{\pi(\mathcal{M}_i|\boldsymbol{y})\pi(\mathcal{M}_j)}{\pi(\mathcal{M}_i|\boldsymbol{y})\pi(\mathcal{M}_i)} = \frac{\int_{\Xi_i} \mathcal{L}_i(\boldsymbol{\xi}_i|\boldsymbol{y})\pi(\boldsymbol{\xi}_i|\mathcal{M}_i)\,\mathrm{d}\boldsymbol{\xi}_i}{\int_{\Xi_j} \mathcal{L}_j(\boldsymbol{\xi}_j|\boldsymbol{y})\pi(\boldsymbol{\xi}_j|\mathcal{M}_j)\,\mathrm{d}\boldsymbol{\xi}_j} \tag{4.4}$$

which can be approximated using the parameter samples of the models $\mathcal{M}_i$ and $\mathcal{M}_j$. Hence, there is a one-to-one relation between Bayes factors and RJMCMC with respect to Bayesian model selection. However, for a denumerable number of models RJMCMC is an essential Bayesian modeling tool. Using a uniform model prior $\pi(\mathcal{M}_i) = \pi(\mathcal{M}_j)$ for all $i, j \in \{1, \dots, k\}$, Equation (4.4) can simply be approximated by the quotient $\frac{n_i}{n_j}$ of instances n_k the Markov chain visited model $\mathcal{M}_k$ (Bartolucci *et al.* [2006]).

Example 4.3 (RJMCMC for the Gaussian mixture model). We again turn to the Gaussian mixture example from Chapter 3.4.4 reusing the notation from above. The parameter space is given by

$$\Xi = \{1\} \times \mathbb{R} \cup \{2\} \times \mathbb{R}^2.$$

As proposed by Robert & Casella [2004] we define the bijections

$$T_{12}(\mu, u) = (\mu_1, \mu_2) := (\mu - u, \mu + u) \;\text{ and }\; T_{21}(\mu_1, \mu_2) = (\mu, u) = \left(\frac{\mu_2 + \mu_1}{2}, \frac{\mu_2 - \mu_1}{2}\right),$$

where $u \sim \mathcal{N}(0, 1^2)$. The Jacobians are then

$$\left|\frac{\partial T_{12}(\mu, u)}{\partial(\mu, u)}\right| = 2 \qquad \text{and} \qquad \left|\frac{\partial T_{21}(\mu_1, \mu_2)}{\partial(\mu_1, \mu_2)}\right| = \frac{1}{2}.$$

We set the jump probabilities $p_{12} = p_{21} = 1$, i.e. we propose a model jump in each RJMCMC step. With the model prior distribution $\rho(1) = \rho(2) = \frac{1}{2}$, the acceptance probabilities are given by

$$\alpha(2, \mu_1, \mu_2|1, \mu, u) = \min\left\{\frac{\mathcal{L}_2(\mu_2, \mu_1|\boldsymbol{y})\varphi_2(\mu_1)\varphi_{-2}(\mu_2)}{\mathcal{L}_1(\mu|\boldsymbol{y})\varphi_0(\mu)\varphi_0(u)} 2, 1\right\}$$

and

$$\alpha(1, \mu, u|2, \mu_1, \mu_2) = \min\left\{\frac{\mathcal{L}_1(\mu|\boldsymbol{y})\varphi_0(\mu)\varphi_0(u)}{\mathcal{L}_2(\mu_1, \mu_2|\boldsymbol{y})\varphi_2(\mu_1)\varphi_{-2}(\mu_2)} \frac{1}{2}, 1\right\},$$

where $\varphi_0(\cdot)$, $\varphi_2(\cdot)$, and $\varphi_{-2}(\cdot)$ are the probability density functions corresponding to $\mathcal{N}(0, 1^2)$, $\mathcal{N}(2, 1^2)$ and $\mathcal{N}(-2, 1^2)$, respectively. Reusing the observations $\boldsymbol{y}$ from the Gaussian mixture example of Chapter 3.4.4 and running the RJMCMC algorithm for ten runs on N =100,000 proposals each, the Bayes factor was computed as 63.21 ± 11.27 (including one standard error). Although this is close to the "true" value of 77.47, thermodynamic integration yielded better results!

4.6 The simulated annealing algorithm

For completeness we want to point out that the MH algorithm can also be directly applied to global optimization problems: Consider a real valued function $h : E \longrightarrow \mathbb{R}$ on a finite set E – for infinite sets E convergence is very problematic in practical applications – along with the minimization problem

$$\min_{\boldsymbol{\xi} \in E} h(\boldsymbol{\xi}). \tag{4.5}$$

Minimization is not a restriction here as we can maximize h by minimizing the function $-h$. Generally, h is a function with various local minima making deterministic minimization algorithms inapplicable. Now, for $T \in (0, 1]$ we can use Algorithm 2 to sample from the stationary distribution

$$\pi(\boldsymbol{\xi}) \propto \exp(-h(\boldsymbol{\xi})/T).$$

The MH acceptance probability then computes to

$$\alpha(\boldsymbol{\xi}|\boldsymbol{\xi}^{(c)}) = \min\{\exp((h(\boldsymbol{\xi}^{(c)}) - h(\boldsymbol{\xi}))/T), 1\}$$

for some proposal function $q(\boldsymbol{\xi}|\boldsymbol{\xi}^{(c)})$ where $\boldsymbol{\xi}^{(c)}$ is the current Markov chain element. Naturally, $\boldsymbol{\xi}$ is accepted, if $h(\boldsymbol{\xi}) < h(\boldsymbol{\xi}^{(c)})$. However, there is a chance (inversely proportional to T) for proposals $\boldsymbol{\xi}$ to get accepted even if $h(\boldsymbol{\xi}^{(c)}) < h(\boldsymbol{\xi})$. The Markov chain can hence escape local minima. Setting up a *cooling schedule*, i.e. defining a series of "temperatures" $T_1 = 1 > T_2 > \ldots T_k > 0$, the iterative application of the MH algorithm for these T_i's solves the minimization problem (4.5). The according algorithm is called *simulated annealing algorithm* and was introduced by Kirkpatrick *et al.* [1983]. Although the theory of time-homogeneous Markov chains introduced in Chapter 2.3 does not provide the necessary convergence properties (c.f. Robert & Casella [2004]), convergence can nevertheless be shown for slow cooling schedules (see e.g. Mitra *et al.* [1986]).

5

Extensions to the Metropolis-Hastings algorithm

In the last few years various extensions to the classical random walk MH algorithm have been proposed. Examples include adaptive MCMC, population MCMC, hybrid Monte Carlo and tempering methods (see Liu [2008] for an overview). All of these approaches try to generate MH proposals that have a high chance of getting accepted by the MH acceptance probability (4.1). This increases the overall sampling efficiency. In the current chapter we introduce two prominent MH extensions. The first exploits the geometric posterior parameter structure by generating proposals guided by the Jacobian matrix. The second successively improves the algorithmic parameters of the transition density function during the sampling process. Chapter 5.3 also introduces some notation needed for the definition of the copula based Metropolis-Hastings algorithms introduced in Chapter 6.

5.1 Simplified Riemann Manifold Metropolis Adjusted Langevin Algorithm (SMALA)

An extension to the classical MH algorithm was derived from diffusion theory (Grenander & Miller [1994]): For the stationary target distribution $\pi(\cdot)$ the *Langevin diffusion*

is defined by the stochastic differential equation

$$\mathrm{d}\boldsymbol{X}^{(t)} = \frac{1}{2}\nabla_{\boldsymbol{X}^{(t)}} \log(\pi(\boldsymbol{X}^{(t)}))\mathrm{d}t + \mathrm{d}\boldsymbol{W}^{(t)}, \tag{5.1}$$

where $\nabla_{\boldsymbol{X}^{(t)}}$ is the nabla operator with respect to $\boldsymbol{X}^{(t)}$ and $\boldsymbol{W}^{(t)}$ denotes n-dimensional standard Brownian motion. The stationary distribution of Equation (5.1) is given by $\pi(\cdot)$. A first order Euler discretization of (5.1) then leads to the *Metropolis Adjusted Langevin Algorithm* (MALA) proposal function $q(\boldsymbol{\xi}|\boldsymbol{\xi}')$ (used in an MH algorithm) with

$$\boldsymbol{\xi} = \boldsymbol{\xi}' + \frac{\varepsilon^2}{2}\nabla_{\boldsymbol{\xi}'} \log(\pi(\boldsymbol{\xi}')) + \varepsilon\boldsymbol{\eta} \tag{5.2}$$

for $\boldsymbol{\eta} \sim \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{I}_n)$. The scaling parameter ε can be used to fine tune the algorithmic step size, which opens up the possibility to control MH acceptance rates. Since the nabla operator $\nabla_{\boldsymbol{\xi}}$ only considers the directional derivatives with respect to the parameter vector $\boldsymbol{\xi}$, strong parameter correlations can disturb the efficacy of MALA. Roberts & Stramer [2002] suggested to circumvent this issue using a preconditioning matrix $\boldsymbol{G}$ such that

$$\boldsymbol{\xi} = \boldsymbol{\xi}' + \frac{\varepsilon^2}{2}\boldsymbol{G}\nabla_{\boldsymbol{\xi}'} \log(\pi(\boldsymbol{\xi}')) + \varepsilon\sqrt{\boldsymbol{G}}\boldsymbol{\eta},$$

where $\sqrt{\boldsymbol{G}}$ is the square root of $\boldsymbol{G}$ obtained by eigen- or Cholesky decomposition. However, finding an appropriate preconditioning matrix $\boldsymbol{G}$ needs structural information about the target distribution. Regarding Bayesian inference Girolami & Calderhead [2011] were the first to consider a parameter dependent preconditioning matrix $\boldsymbol{G}(\boldsymbol{\xi})$ as *geometric tensor* that takes into account the local geometric structure of the joint data and parameter distribution $\pi(\boldsymbol{y}, \boldsymbol{\xi})$. The concept exploits the distance between two $\boldsymbol{\xi}$-parametrized distributions $\pi(\boldsymbol{y}|\boldsymbol{\xi})$ and $\pi(\boldsymbol{y}|\boldsymbol{\xi} + \delta\boldsymbol{\xi})$ given by the quadratic form $\delta\boldsymbol{\xi}^\top \boldsymbol{G}(\boldsymbol{\xi})\delta\boldsymbol{\xi}$ for some positive definite metric tensor $\boldsymbol{G}(\boldsymbol{\xi})$ (Rao [1945]). Rao noted that $\boldsymbol{G}(\boldsymbol{\xi})$ by definition yields the metric of a Riemann manifold. Given that $\log(\pi(\boldsymbol{y}|\boldsymbol{\xi}))$ is twice absolutely continuously differentiable with respect to $\boldsymbol{\xi}$ (as is the case in almost all practical applications, Girolami & Calderhead [2011] finally suggested to take $\boldsymbol{G}(\boldsymbol{\xi})$

to be the expected *Fisher information matrix* minus the Hessian of the log-prior, i.e.

$$\begin{aligned}\boldsymbol{G}(\boldsymbol{\xi}) &= -\mathbb{E}_{\pi(\boldsymbol{y}|\boldsymbol{\xi})}\left[\frac{\partial^2}{\partial \boldsymbol{\xi}^2}\log(\pi(\boldsymbol{y},\boldsymbol{\xi}))\right] \\ &= -\mathbb{E}_{\pi(\boldsymbol{y}|\boldsymbol{\xi})}\left[\frac{\partial^2}{\partial \boldsymbol{\xi}^2}\log(\pi(\boldsymbol{y}|\boldsymbol{\xi})\pi(\boldsymbol{\xi}))\right] \\ &= \underbrace{-\mathbb{E}_{\pi(\boldsymbol{y}|\boldsymbol{\xi})}\left[\frac{\partial^2}{\partial \boldsymbol{\xi}^2}\log(\pi(\boldsymbol{y}|\boldsymbol{\xi}))\right]}_{\text{expected Fisher information matrix}} - \underbrace{\frac{\partial^2}{\partial \boldsymbol{\xi}^2}\log(\pi(\boldsymbol{\xi}))}_{\text{Hessian of log-prior}} \\ &= \operatorname{cov}\left[\left(\frac{\partial}{\partial \boldsymbol{\xi}}\log(\pi(\boldsymbol{y}|\boldsymbol{\xi}))\right)^\top, \left(\frac{\partial}{\partial \boldsymbol{\xi}}\log(\pi(\boldsymbol{y}|\boldsymbol{\xi}))\right)^\top\right] - \frac{\partial^2}{\partial \boldsymbol{\xi}^2}\log(\pi(\boldsymbol{\xi})).\end{aligned}$$

The last equation follows assuming Fisher regularity for the target distribution (Schervish [1995], Prop. 2.84). Generally, a Langevin diffusion with invariant distribution $\pi(\cdot)$ can be defined on a Riemann manifold via its metric tensor $\boldsymbol{G}(\cdot)$ by

$$\mathrm{d}\boldsymbol{X}^{(t)} = \frac{1}{2}\tilde{\nabla}_{\boldsymbol{X}^{(t)}}\log(\pi(\boldsymbol{X}^{(t)}))\mathrm{d}t + \mathrm{d}\tilde{\boldsymbol{W}}^{(t)}, \tag{5.3}$$

where

$$\tilde{\nabla}_{\boldsymbol{X}^{(t)}}\log(\pi(\boldsymbol{X}^{(t)})) = \boldsymbol{G}^{-1}(\boldsymbol{X}^{(t)})\nabla_{\boldsymbol{X}^{(t)}}\log(\pi(\boldsymbol{X}^{(t)}))$$

for the natural gradient $\nabla_{\boldsymbol{X}^{(t)}}$ on $\mathbb{R}^n$ (Amari & Nagaoka [2007]). On the other hand, the Brownian motion on the manifold is given by

$$\begin{aligned}\mathrm{d}\tilde{\boldsymbol{W}}_i^{(t)} = \det(\boldsymbol{G}(\boldsymbol{X}^{(t)}))^{-\frac{1}{2}}\sum_{j=1}^{n}\frac{\partial}{\partial \boldsymbol{X}^{(t)}}\left(\boldsymbol{G}(\boldsymbol{X}^{(t)})_{i,j}\det(\boldsymbol{G}(\boldsymbol{X}^{(t)}))^{\frac{1}{2}}\right)\mathrm{d}t \\ + \left(\sqrt{\boldsymbol{G}^{-1}(\boldsymbol{X}^{(t)})}\mathrm{d}\boldsymbol{W}^{(t)}\right)_i,\end{aligned} \tag{5.4}$$

$i = 1, \ldots, n$ (Chung [1982]). Expanding the differential in Equation (5.4) in combination with a first order Euler discretization of Equation (5.3) finally yields the *Riemann Manifold MALA* (MMALA) proposal function $q(\boldsymbol{\xi}|\boldsymbol{\xi}')$ via

$$\begin{aligned}\boldsymbol{\xi}_i =& \boldsymbol{\xi}_i' + \frac{\varepsilon^2}{2}\left(\boldsymbol{G}^{-1}(\boldsymbol{\xi}')\nabla_{\boldsymbol{\xi}'}\log\pi(\boldsymbol{y}|\boldsymbol{\xi}')\right)_i - \varepsilon^2\sum_{j=1}^{n}\left(\boldsymbol{G}^{-1}(\boldsymbol{\xi}')\frac{\partial \boldsymbol{G}(\boldsymbol{\xi}')}{\partial \boldsymbol{\xi}_j'}\boldsymbol{G}^{-1}(\boldsymbol{\xi}')\right)_{i,j} \\ &+ \frac{\varepsilon^2}{2}\sum_{j=1}^{n}\left(\boldsymbol{G}^{-1}(\boldsymbol{\xi}')\right)_{i,j}\operatorname{tr}\left(\boldsymbol{G}^{-1}(\boldsymbol{\xi}')\frac{\partial \boldsymbol{G}(\boldsymbol{\xi}')}{\partial \boldsymbol{\xi}_j'}\right) + \left(\varepsilon\sqrt{\boldsymbol{G}^{-1}(\boldsymbol{\xi}')}\boldsymbol{\eta}\right)_i\end{aligned} \tag{5.5}$$

for $i = 1, \ldots, n$ and $\boldsymbol{\eta} \sim \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{I}_n)$. In summary, MMALA exploits the local Riemann manifold structure of the parameter space at stake in order to efficiently explore the

target distribution $\pi(\boldsymbol{\xi}|\boldsymbol{y})$. This becomes clear as the MH proposal function proposes moves with respect to the Riemannian metric defined by $\boldsymbol{G}(\boldsymbol{\xi})$ rather than with respect to the standard Euclidian distance on $\mathbb{R}^n$. Sampling results in Girolami & Calderhead [2011] showed that simplifying the proposal (5.5) by assuming a constant curvature throughout the manifold, i.e. $\frac{\partial^2}{\partial \boldsymbol{\xi}^2} \log(\pi(\boldsymbol{y}, \boldsymbol{\xi})) = const.$ and hence $\frac{\partial}{\partial \boldsymbol{\xi}} \boldsymbol{G}(\boldsymbol{\xi}) = 0$, drastically decreases the computational demand for proposal generation. Although the effective sample size decreases, the number of i.i.d. samples drawn per second increases. This holds true especially for the dynamic systems considered in the paper. The MH sampling scheme assuming constant manifold curvature is called *Simplified MMALA* (SMALA) algorithm. The according proposal function simplifies to

$$\boldsymbol{\xi} = \boldsymbol{\xi}' + \frac{\varepsilon^2}{2} \boldsymbol{G}^{-1}(\boldsymbol{\xi}') \nabla_{\boldsymbol{\xi}'} \log \pi(\boldsymbol{y}|\boldsymbol{\xi}') + \varepsilon \sqrt{\boldsymbol{G}^{-1}(\boldsymbol{\xi}')} \boldsymbol{\eta}. \tag{5.6}$$

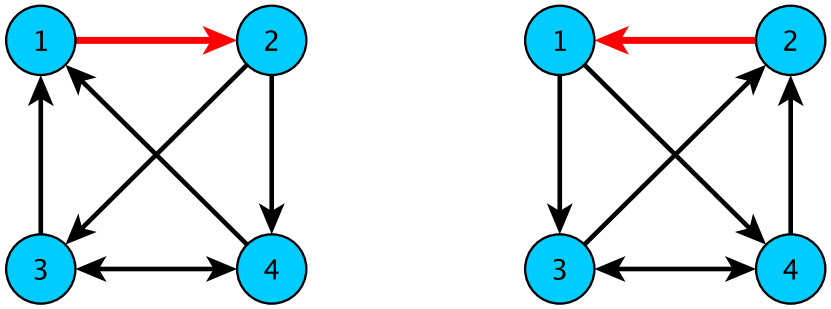
for $\boldsymbol{\eta} \sim \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{I}_n)$. Clearly, even though the manifold curvature is generally not constant, the proposal scheme (5.6) defines a valid MCMC sampler in combination with the MH acceptance rule. Here, the proposals are somewhat generated in the direction of highest local improvement with respect to the posterior distribution. Compared to random walk proposals this can improve the chance of acceptance.

5.2 Adaptive MCMC

Another very tempting approach to attain high proposal acceptance rates is the fine tuning of algorithmic parameters, such as the scaling parameter k_{RW} of the RWMH algorithm, during the sampling process. This algorithmic parameter adaption is generally done based on preceding Markov chain realizations. A major pitfall of this attempt is, that the stochastic process $\{\boldsymbol{X}^{(t)}\}_{t \in \mathbb{N}_0}$ is no longer Markovian, i.e.

$$P(\boldsymbol{X}^{(t)}|\boldsymbol{X}^{(t-1)}, \ldots, \boldsymbol{X}^{(0)}) \neq P(\boldsymbol{X}^{(t)}|\boldsymbol{X}^{(t-1)}).$$

Thus, the convergence, or rather the ergodicity of the MCMC sampler at hand is no longer guaranteed (Rosenthal [2011]). More generally spoken, the application of different Markov chain transition kernels in the very same inference process of an arbitrary probability distribution can destroy the ergodicity constraint, even if all of these transition kernels have the same equilibrium distribution. This can be seen in the following simple (discrete) example inspired by Roberts & Rosenthal [2007]:



(a) (b)

Figure 5.1: Graphical representation of the discrete state space models. Depicted are the transition probabilities for (a) the transition kernel $k_1(\cdot|\cdot)$ and (b) the transition kernel $k_2(\cdot|\cdot)$. Here, black arrows correspond to a transition probability of $\frac{1}{2}$ and red arrows to a transition probability of 1.

Example 5.1. Suppose $E = \{1, 2, 3, 4\}$ is a discrete state space and we want to sample from the uniform distribution $\pi(\cdot)$ with

$$\pi(1) = \pi(2) = \pi(3) = \pi(4) = \frac{1}{4}.$$

Suppose furthermore we are given the irreducible and aperiodic transition kernels $k_1(\cdot|\cdot)$ and $k_2(\cdot|\cdot)$ with

$$\begin{aligned} k_1(2|1) = k_2(1|2) &= 1, \\ k_1(3|2) = k_1(4|2) = k_1(1|3) = k_1(4|3) = k_1(1|4) = k_1(3|4) &= \frac{1}{2}, \\ k_2(3|1) = k_2(4|1) = k_2(2|3) = k_2(4|3) = k_2(2|4) = k_2(3|4) &= \frac{1}{2} \end{aligned}$$

(see Figure 5.1). Following Example 2.8 $\pi(\cdot)$ is the unique stationary distribution for k_1 and k_2. Now, iteratively applying k_1 and k_2 starting at $\xi^{(0)} = \{1\}$ yields a Markov chain that exclusively visits the states $\{1\}$ and $\{2\}$. Hence, the chain does not converge to $\pi(\cdot)$.

Adaptive MCMC methods preserving ergodicity have been proposed amongst others by Gilks *et al.* [1998] or Holden *et al.* [2009]. Gilks allows updating the proposal function whenever the Markov chain reaches a set $A \subset E$ of the state space $(E, \mathcal{E})$, such that $\pi(A) > 0$ and $\boldsymbol{X}^{(t+1)}, \boldsymbol{X}^{(t+2)}, \ldots$ is conditionally independent of $\boldsymbol{X}^{(0)}, \ldots, \boldsymbol{X}^{(t)}$ given $\boldsymbol{X}^{(t)} \in A$. The set A is called a *proper atom* of the Markov chain and whenever the

chain enters A, it is said to *regenerate*. For continuous state spaces $(E, \mathcal{E})$ proper atoms might not exist. Although a technique due to Nummelin [2004] allows to construct regenerating Markov chains on some augmented state space also for continuous state spaces, the practical application of regenerating Markov chains remains an engineering art form. Holden and coworkers, on the other hand propose an adaptive independence sampling scheme that takes into account almost the full history of the proposed samples. They show convergence as long as the proposal distribution satisfies the strong Doeblin condition that we introduce in Remark 6.1.

Theoretical results with respect to convergence of adaptive MCMC samplers have been derived by Roberts & Rosenthal [2007]. They all require that the transition kernels used are uniformly bounded. Furthermore, the amount of modification of the algorithmic parameters has to either diminish as $t \to \infty$ – this is e.g. the case for the *Adaptive Metropolis algorithm* of Haario *et al.* [2001] introduced in the following – or the adaption process is applied with diminishing probability $P_A(t)$, i.e. $P_A(t) \to 0$ for $t \to \infty$.

5.3 Metropolis Gaussian Adaption algorithm (M-GaA)

Haario *et al.* [1999] introduced a rather simple but efficient adaptive Monte Carlo sampler called *Adaptive Proposal* (AP) algorithm. The basic idea is to apply a RWMH algorithm using a multivariate Gaussian proposal function whose covariance matrix is continuously updated based on a fixed number of previously accepted Markov chain samples. With this the MCMC process (locally) adapts to the target distribution and provides an effective proposal scheme. As the covariance matrix can be updated sequentially (Haario *et al.* [2001]), the additional computational cost is by far outweighed by the improvement in efficiency. Although the AP algorithm has empirically proven to outperform the classical (non-adaptive) Metropolis-Hastings algorithm, rigorous proofs of convergence are yet still missing. In fact, Haario *et al.* [1999] pointed out that AP algorithm might be slightly biased. Nevertheless, the paper also shows that the difference between the AP sampled and true limiting distribution is very small for practical applications.

The AP algorithm was extended by Müller & Sbalzarini [2010] such that the adapted covariance matrix of the RWMH proposal function maximizes the entropy of the target

distribution $\pi(\cdot)$ under the constraint that the proposed MCMC samples are accepted with a predefined theoretical MH acceptance rate $\alpha_0 \in (0,1)$. In contrast to the classical AP algorithm the covariance matrix is based on all previously accepted Markov chain samples. This sampling scheme was named *Metropolis Gaussian Adaption* (M-GaA) algorithm. Without going into detail we shortly review the basic sampling procedure applying the *strategic* algorithmic parameter setting of Müller & Sbalzarini [2010]: Starting at some initial value $\boldsymbol{\xi}^{(0)} \in \mathbb{R}^n$, the empirical proposal covariance matrix of step j is decomposed as

$$\boldsymbol{\Sigma}^{(j)} = r^2 \sqrt{\boldsymbol{\Sigma}^{(j)}} \sqrt{\boldsymbol{\Sigma}^{(j)}}^{\top}$$

where $\sqrt{\boldsymbol{\Sigma}^{(j)}}$ denotes the normalized square root of $\boldsymbol{\Sigma}^{(j)}$ found by eigen- or Cholesky decomposition, i.e. $\det\left(\sqrt{\boldsymbol{\Sigma}^{(j)}}\right) = 1$. The parameter r is of no importance for the rest of the proposal scheme. The algorithm uses the n-dimensional identity matrix $\boldsymbol{I}_n$ as proposal covariance matrix $\boldsymbol{\Sigma}^{(i)}$ as long as $\boldsymbol{\xi}^{(i)} = \boldsymbol{\xi}^{(0)}$ for $i \in \mathbb{N}$. An MCMC proposal is then generated based on the current sample $\boldsymbol{\xi}^{(j)}$ via

$$\boldsymbol{\xi}^{p} = \boldsymbol{\xi}^{(j)} + r^{(j)} \sqrt{\boldsymbol{\Sigma}^{(j)}} \boldsymbol{\eta}^{(j)}$$

for some $\boldsymbol{\eta}^{(j)} \sim \mathcal{N}_n(\mathbf{0}, \boldsymbol{I}_n)$ and some step size parameter $r^{(j)}$ defined below. Due to the symmetry in the proposal function we accept $\boldsymbol{\xi}^{p}$ according to the Metropolis-Hastings acceptance probability

$$\alpha(\boldsymbol{\xi}^{p} | \boldsymbol{\xi}^{(j)}) = \min\left\{ \frac{\pi(\boldsymbol{\xi}^{p})}{\pi(\boldsymbol{\xi}^{(j)})}, 1 \right\}$$

with respect to the target density $\pi(\cdot)$. The proposal covariance matrix $\boldsymbol{\Sigma}^{(j)}$ is updated based on the Markov chain sample $\boldsymbol{\xi}^{(j+1)}$ via

$$\boldsymbol{\Sigma}^{(j+1)} = (1-s)\, \boldsymbol{\xi}^{(j+1)} + s(\boldsymbol{\xi}^{(j+1)} - \boldsymbol{\xi}^{(j)}) \cdot (\boldsymbol{\xi}^{(j+1)} - \boldsymbol{\xi}^{(j)})^{\top}$$

for some algorithmic parameter s. Müller & Sbalzarini [2010] suggested to choose $s = \frac{\log(n+1)}{(n+1)^2} < 1$ as s directly controls the influence of the n^2 values of $\boldsymbol{\Sigma}^{(j)}$. Finally, the step size parameter $r^{(j)}$ was suggested to be updated as

$$r^{(j+1)} = \begin{cases} (1 + s\,(1 - \alpha_0)) \cdot r^{(j)} & \text{if } \boldsymbol{\xi}^{p} \text{ was accepted and} \\ (1 - s\alpha_0) \cdot r^{(j)} & \text{otherwise,} \end{cases}$$

where α_0 denotes the predefined acceptance rate of the M-GaA algorithm. Note that $r^{(j+1)}$ is reduced in case $\boldsymbol{\xi}^p$ was rejected and increased otherwise. We will see in Chapter 6.3.4 that the actual acceptance rate of M-GaA can fail to meet α_0 in complex applications.

Remark 5.1 (Adaptive Metropolis algorithm). An adaptive MCMC scheme very similar to the AP and M-GaA algorithms was introduced in Haario *et al.* [2001]: The *Adaptive Metropolis* (AM) algorithm updates the multivariate Gaussian RWMH proposal function with covariance matrix $\boldsymbol{\Sigma}$ based on all previously accepted Markov chain samples. The explicit covariance adaption at the j^{th} MCMC step is given by

$$\boldsymbol{\Sigma}^{(j)} = \begin{cases} \boldsymbol{\Sigma}^{(0)} & \text{for } j \leq j_0 \\ s_d \text{Cov}(\boldsymbol{\xi}^{(0)}, \dots \boldsymbol{\xi}^{(j-1)}) + s_d \varepsilon \boldsymbol{I}_n & \text{for } j > j_0. \end{cases}$$

where $\boldsymbol{\Sigma}^{(0)}$ is some initial covariance matrix applied up to step j_0 with $\boldsymbol{\xi}^{(i)} \neq \boldsymbol{\xi}^{(i-1)}$ for some $i \leq j_0$. $\text{Cov}(\boldsymbol{\xi}^{(0)}, \dots \boldsymbol{\xi}^{(j-1)})$ denotes the empirical covariance matrix based on the Markov chain realizations $\boldsymbol{\xi}^{(0)}, \dots \boldsymbol{\xi}^{(j-1)}$ and $\boldsymbol{I}_n$ the n-dimensional identity matrix. The parameter s_d is a predefined scaling constant and $\varepsilon > 0$ an auxiliary constant introduced to avoid singularities in the covariance matrices $\boldsymbol{\Sigma}^{(j)}$. Haario *et al.* [2001] provided a proof of convergence for the AM algorithm for bounded target distributions $\pi(\cdot)$ on a bounded support $S \subset \mathbb{R}^n$, this is, there exists a constant $M \in \mathbb{R}$ such that $\pi(\boldsymbol{x}) \leq M$ for all $\boldsymbol{x} \in S$ and $\pi(\boldsymbol{X}) \equiv 0$ for all $\boldsymbol{x} \in \mathbb{R}^n \setminus S$.

While the SMALA and M-GaA algorithms try to locally improve the MH transition kernel, we took a global approach in order to elevate the MH sampling efficiency. Here, an approximation of the posterior distribution generates samples that are distributed similar to the true posterior. The concept is introduced in the next chapter.

6

Improving the Metropolis-Hastings algorithm using copulas

We have seen in Figure 1.1 that even in two dimensions posterior complexity can be quite severe. For an independent posterior parameter distribution $\pi(\xi_1, \ldots, \xi_n|\boldsymbol{y}) = \pi(\xi_1|\boldsymbol{y}) \cdot \ldots \cdot \pi(\xi_n|\boldsymbol{y})$ the inference process can be focused on the distributions $\pi(\xi_i|\boldsymbol{y})$, $i = 1, \ldots, n$, individually, avoiding the so-called *curse of dimensionality* (Hastie *et al.* [2009]) that aggravates the inference of the joint distribution $\pi(\xi_1, \ldots, \xi_n|\boldsymbol{y})$. Especially model inference of parametrized differential equations likes to trap Markov Chain Monte Carlo samplers between high proposal rejection rates and strong autocorrelation structures within the sampled Markov chains – both leading to a low number of independent samples drawn over time. A crucial issue for the efficacy of an MH algorithm is the choice of the proposal function. Clearly, the optimal proposal function is given by the actual posterior density, which transforms the Metropolis acceptance probability to

$$\alpha(\boldsymbol{\xi}^p|\boldsymbol{\xi}^{(c)}) = \min\left\{\frac{\pi(\boldsymbol{\xi}^p|\boldsymbol{y})\pi(\boldsymbol{\xi}^{(c)}|\boldsymbol{y})}{\pi(\boldsymbol{\xi}^{(c)}|\boldsymbol{y})\pi(\boldsymbol{\xi}^p|\boldsymbol{y})}, 1\right\} = 1.$$

However, if direct sampling from the posterior is possible, the MH algorithm becomes uncalled-for. The best we can hope for is to use a proposal function that approximates the posterior density as close as possible. We addressed this issue by developing

a novel (adaptive) MCMC approach for efficient parameter inference in highly dependent systems: The *Copula based Independence MH algorithm* (CIMH) and its extension, the *Adaptive Copula based Independence MH algorithm* (ACIMH), exploit the concept of a vine copula decomposition of the posterior distribution densities. They allow to generate problem specific proposals for a hybrid independence chain/random walk Metropolis-Hastings sampler. The key advantage of this approach is a reduced autocorrelation structure in the sampled Markov chain and with this an increased number of independent samples drawn over time. In ACIMH all copula densities are updated during the sampling procedure for fine-tuning.

The performance of our method(s) is assessed on three small scale examples and finally evaluated on a DDE model for the JAK2-STAT5 signaling pathway fitted to time-resolved western blot data. In the first three examples we compare our copula based sampler to a random walk MH algorithm (RWMH), an independence chain sampler (IMH) and a second order moment based random walk MH algorithm (CovRWMH), as introduced in Chapter 4.1 and further specified below. Due to the simplicity of the systems we do not consider the SMALA and M-GaA algorithms here. However, they additionally come into play for performance evaluation on the complex JAK2-STAT5 pathway.

6.1 Copula based Independence MH algorithm (CIMH)

We now turn to the definition of CIMH. As mentioned above, it is based on a vine copula proposal function similar to the posterior density. However, since the copula may be based on insufficient data, we extend this proposal function by two additional transition functions, the first of which is a random walk density and the second a heavy-tailed independence density. The latter is essential to safeguard convergence. Overall, we end up with a hybrid copula based independence/random walk proposal function. The sampling scheme consists of four steps: (i) a prerun, (ii) a uniformization step of the prerun samples, (iii) a D-vine copula decomposition of the dependent prerun samples, and (iv) the generation of a Markov chain by means of the hybrid copula based independence chain/random walk sampler. Throughout, we assume that the sampling space is a Borel measurable subset of $\mathbb{R}^n$.

6.1.1 The basic copula MH sampling procedure

(i) *Prerun:* Our goal is to efficiently sample an independent Markov chain realization $\{\boldsymbol{\xi}^{(j)}\}_{j=0,\dots,T}$ with $\boldsymbol{\xi}^{(j)} \in \mathbb{R}^n$ from the posterior distribution $\pi(\boldsymbol{\xi}|\boldsymbol{y})$ based on the observations $\boldsymbol{y}$. For this, we first generate an initial Markov chain $\{\breve{\boldsymbol{\xi}}^{(j)}\}_{j=0,\dots,T'}$, the so-called *prerun samples*, using e.g. RWMH or any other sampling algorithm. The chain length $T'+1$ should ideally be large enough for the prerun samples to sufficiently cover the support of $\pi(\boldsymbol{\xi}|\boldsymbol{y})$. Determining T' can either be left to the modeler's experience or monitored utilizing convergence statistics such as the Gelman-Rubin statistic introduced in Chapter 4.4. Although too small values of T' have a negative effect on the performance of CIMH, to our experience we can generally choose T' a lot smaller than the chain length $T+1$ of the final Markov chain $\{\boldsymbol{\xi}^{(j)}\}_{j=0,\dots,T}$. The prerun samples $\{\breve{\boldsymbol{\xi}}^{(j)}\}_{j=0,\dots,T'}$ form the basis for the copula proposal function. Note that we do not demand independence in the realization $\{\breve{\boldsymbol{\xi}}^{(j)}\}_j$.

(ii) *Uniformization:* Based on $\{\breve{\boldsymbol{\xi}}^{(j)}\}_j$, we fit a $\boldsymbol{\theta}$-parametrized D-vine copula $c_{1,\dots,n}(\boldsymbol{u}|\boldsymbol{\theta})$ in step (iii). As seen in Chapter 2.2, copulas are defined on the n-dimensional unit cube $[0,1]^n$. Hence, each prerun sample $\breve{\boldsymbol{\xi}}^{(j)}$ needs to be transformed to $[0,1]^n$. Depending on the shape of the histograms of the n sample marginals $\breve{\boldsymbol{\xi}}_i := (\breve{\xi}_i^{(1)}, \dots, \breve{\xi}_i^{(T')})^\top$, we fit for $i = 1, \dots, n$ $\boldsymbol{\gamma}_i$-parametrized continuous cdf's $G_i(\xi|\boldsymbol{\gamma}_i)$ to the respective sample marginal. Clearly, the support of the marginal posterior density function $\pi(\xi_i|\boldsymbol{y})$ needs to be covered by the support of $G_i(\xi|\boldsymbol{\gamma}_i)$. This is not a limitation as the support of $\pi(\xi_i|\boldsymbol{y})$ is controlled by the support of the prior distributions and the claim can easily be satisfied. Each $\breve{\boldsymbol{\xi}}^{(j)}$ is then transformed to $\breve{\boldsymbol{u}}^{(j)} := (G_1(\breve{\xi}_1^{(j)}|\hat{\boldsymbol{\gamma}}_1), \dots, G_n(\breve{\xi}_n^{(j)}|\hat{\boldsymbol{\gamma}}_n))^\top \in [0,1]^n$ based on the estimates $\hat{\boldsymbol{\gamma}}_i$ of $\boldsymbol{\gamma}_i$. In the following we refer to $\{\breve{\boldsymbol{u}}^{(j)}\}_{j=0,\dots,T'}$ as *copula data*. Let us consider a simple example: Say, for instance, $n = 2$ and the sample marginals of $\{\breve{\boldsymbol{\xi}}^{(j)}\}_j$ are normally distributed. Based on the estimated means $\hat{\mu}_1, \hat{\mu}_2$ and variances $\hat{\sigma}_1^2, \hat{\sigma}_2^2$ of $\{\breve{\boldsymbol{\xi}}^{(j)}\}_j$ we transform

$$\breve{u}_1^{(j)} = \Phi\left(\frac{\breve{\xi}_1^{(j)} - \hat{\mu}_1}{\hat{\sigma}_1}\right) \qquad \text{and} \qquad \breve{u}_2^{(j)} = \Phi\left(\frac{\breve{\xi}_2^{(j)} - \hat{\mu}_2}{\hat{\sigma}_2}\right),$$

where $\Phi(\cdot)$ is the cdf of a standard normal random variable. Step (ii) does not change the dependency structure inherent to the prerun samples $\{\breve{\boldsymbol{\xi}}^{(j)}\}_j$, which is exclusively modeled by the D-vine copula (see Aas *et al.* [2009]). This implies that the estimated

Kendall's τ's for $\{\breve{\boldsymbol{\xi}}^{(j)}\}_j$ are identical to the estimated Kendall's τ's for $\{\breve{\boldsymbol{u}}^{(j)}\}_j$. The procedure works for more general vine structures, such as R-vines, as well, but for the application to dynamical systems a sequential dependency structure seems appropriate.

(iii) *Copula decomposition:* The user defined D-vine structure fixes the variable order for the D-vine copula, i.e. a permutation function $\iota : \{1, \dots, n\} \longrightarrow \{1, \dots, n\}, i \mapsto \iota(i)$ rearranges each sample $\breve{\boldsymbol{u}}^{(j)} = (\breve{u}_1^{(j)}, \dots, \breve{u}_n^{(j)})^\top$ to $\tilde{\boldsymbol{u}}^{(j)} := (\breve{u}_{\iota(1)}^{(j)}, \dots, \breve{u}_{\iota(n)}^{(j)})^\top$. Defining ι such that for $i = 1, \dots, (n-1)$ the pairs $(\iota(i), \iota(i)+1)$ cover the highest pairwise absolute dependency works well in our simulations. For determining the dependency structure, estimated pairwise Kendall's τ's for $\{\breve{\boldsymbol{u}}^{(j)}\}_j$ can be used. Based on $\{\tilde{\boldsymbol{u}}^{(j)}\}_j$ we then fit[1] a $\boldsymbol{\theta}$-parametrized D-vine copula density function

$$c_{1,\dots,n}(\boldsymbol{u}|\boldsymbol{\theta}) = \prod_{j=1}^{n-1} \prod_{i=1}^{n-j} c_{j,j+i|j+1,\dots,j+i-1}(F(u_j|u_{j+1,\dots,j+i-1}, \boldsymbol{\theta}), F(u_{j+i}|u_{j+1,\dots,j+i-1}, \boldsymbol{\theta})|\boldsymbol{\theta}), \tag{6.1}$$

where $F(u_\ell|\boldsymbol{u}_{\mathcal{D}}, \boldsymbol{\theta})$ is the $\boldsymbol{\theta}$-parameterized conditional cdf of u_ℓ given $\boldsymbol{U}_{\mathcal{D}} = \boldsymbol{u}_{\mathcal{D}}$ and $\boldsymbol{u}_{\mathcal{D}}$ is a set of $[0,1]$ valued variables. Here, the order of the variables in (6.1) corresponds to the permutation ι chosen above. In our notation the parameter $\boldsymbol{\theta} = \{\boldsymbol{\theta}_{i,j+i|(j+1),\dots,(j+i-1)}\}$ for $j = 1, \dots, (n-1)$ and $i = 1, \dots, (n-j)$ contains the copula parameters and types. However, all bivariate copulas $c_{j,j+i|(j+1),\dots,(j+i-1)}$ depend only on $\boldsymbol{\theta}_{i,j+i|(j+1),\dots,(j+i-1)}$.

(iv) *Generation of the Markov chain:* The copula proposal function is defined as follows: For generating n-dimensional copula proposals $\tilde{\boldsymbol{\xi}} \in \mathbb{R}^n$, we sample $\tilde{\boldsymbol{u}} \sim c_{1,\dots,n}(\boldsymbol{u}|\hat{\boldsymbol{\theta}})$ from the estimated copula $c_{1,\dots,n}(\boldsymbol{u}|\hat{\boldsymbol{\theta}})$. The sample $\tilde{\boldsymbol{u}}$ is then transformed by $\tilde{\xi}_i := G^{-1}_{\iota^{-1}(i)}(\tilde{u}_{\iota^{-1}(i)}|\hat{\boldsymbol{\gamma}}_{\iota^{-1}(i)})$ to yield $\tilde{\boldsymbol{\xi}} = (\tilde{\xi}_1, \dots, \tilde{\xi}_n)^\top$. In the setting of the example above, say, we choose ι to be the identity function. The corresponding samples $\tilde{\boldsymbol{\xi}}$ on $\mathbb{R}^2$ are then for $i = 1, 2$ given by

$$\tilde{\xi}_i = G^{-1}_{\iota^{-1}(i)}(\tilde{u}_{\iota^{-1}(i)}|\hat{\mu}_{\iota^{-1}(i)}, \hat{\sigma}^2_{\iota^{-1}(i)}) = \Phi^{-1}(\tilde{u}_{\iota^{-1}(i)})\hat{\sigma}_{\iota^{-1}(i)} + \hat{\mu}_{\iota^{-1}(i)} = \Phi^{-1}(\tilde{u}_i)\hat{\sigma}_i + \hat{\mu}_i.$$

Thus, all copula proposals $\tilde{\boldsymbol{\xi}}$ are generated according to the joint proposal function

$$q_1(\boldsymbol{\xi}|\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\theta}}) := c_{1,\dots,n}(G_1(\xi_1|\hat{\boldsymbol{\gamma}}_1), \dots, G_n(\xi_n|\hat{\boldsymbol{\gamma}}_n)|\hat{\boldsymbol{\theta}}) \cdot \prod_{i=1}^{n} g_i(\xi_i|\hat{\boldsymbol{\gamma}}_i) \tag{6.2}$$

[1] See Remark 6.2.

where $g_i(\xi|\hat{\boldsymbol{\gamma}}_i)$ are the density functions corresponding to $G_i(\xi|\hat{\boldsymbol{\gamma}}_i)$. Now, let $q_2(\boldsymbol{\xi}|\boldsymbol{\xi}')$ be a random Metropolis-Hastings proposal function of choice and $q_3(\boldsymbol{\xi})$ a (compared to the posterior density $\pi(\boldsymbol{\xi}|\boldsymbol{y})$) heavy-tailed independence proposal function with $q_2(\boldsymbol{\xi}|\boldsymbol{\xi}') > 0$ and $q_3(\boldsymbol{\xi}) > 0$ on the support of the prior distribution. Let furthermore $\boldsymbol{\xi}^{(0)} \in \mathbb{R}^n$ be an initial sample. For fixed constants $r_1 \in [0,1)$ and $r_2 \in [0,1)$ with $r_1 + r_2 < 1$, we define the *copula based hybrid independence/random walk proposal function for CIMH* via the density function

$$q^{cop}(\boldsymbol{\xi}|\boldsymbol{\xi}', \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\theta}}) := r_1 q_1(\boldsymbol{\xi}|\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\theta}}) + r_2 q_2(\boldsymbol{\xi}|\boldsymbol{\xi}') + (1 - r_1 - r_2) q_3(\boldsymbol{\xi}). \tag{6.3}$$

Pseudo-code for CIMH is depicted in Algorithm 3. For readability, we write $q^{cop}(\boldsymbol{\xi}|\boldsymbol{\xi}')$ for $q^{cop}(\boldsymbol{\xi}|\boldsymbol{\xi}', \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\theta}})$. With this, the Metropolis-Hastings acceptance probability is given by

$$\alpha^{cop}(\boldsymbol{\xi}|\boldsymbol{\xi}') = \min\left\{ \frac{\pi(\boldsymbol{\xi}|\boldsymbol{y}) q^{cop}(\boldsymbol{\xi}'|\boldsymbol{\xi})}{\pi(\boldsymbol{\xi}'|\boldsymbol{y}) q^{cop}(\boldsymbol{\xi}|\boldsymbol{\xi}')}, 1 \right\}. \tag{6.4}$$

We need to point out that although $q^{cop}(\boldsymbol{\xi}|\boldsymbol{\xi}')$ is independent of the current Markov chain sample $\boldsymbol{\xi}'$ for $r_2 = 0$, the acceptance probability nevertheless depends on $\boldsymbol{\xi}'$. Hence, the Markov chain generally inherits some autocorrelation structure even for $r_2 = 0$. The constants r_1 and r_2 are generally chosen such that $r_1 + r_2$ is close to one in order to "waste" as few samples as possible.

Proposition 6.1 (Convergence of CIMH)**.** *The CIMH sampling scheme converges to the posterior equilibrium distribution.*

Proof. The proof is identical to the one of Theorem 4.1 recalling that for $i = 1, \ldots, n$ the support of the marginal posterior density function $\pi(\xi_i|\boldsymbol{y})$ is covered by the support of the transformation functions $G_i(\xi|\boldsymbol{\gamma}_i)$. Hence, with $q_2(\boldsymbol{\xi}|\boldsymbol{\xi}') > 0$ and $q_3(\boldsymbol{\xi}) > 0$ on the support of the prior distribution, $q^{cop}(\boldsymbol{\xi}|\boldsymbol{\xi}') > 0$ on the support of $\pi(\boldsymbol{\xi}|\boldsymbol{y})$, which yields the regularity condition of Theorem 4.1. □

Remark 6.1 (Strong Doeblin condition). The heavy-tailed independent proposal function $q_3(\boldsymbol{\xi})$ guarantees that the proposal distribution $q^{cop}(\boldsymbol{\xi}|\boldsymbol{\xi}')$ has uniformly heavier tails than the posterior distribution $\pi(\boldsymbol{\xi}|\boldsymbol{y})$. Hence, the *strong Doeblin condition*[1] holds, i.e. there exists an integer $s > 0$ and a constant $a_s \in (0,1]$ such that

$$(q^{cop})^s(\boldsymbol{\xi}, \boldsymbol{\xi}') \geq a_s \pi(\boldsymbol{\xi}|\boldsymbol{y}) \qquad \text{for all } \boldsymbol{\xi}, \boldsymbol{\xi}' \in \mathbb{R}^n. \tag{6.5}$$

[1] See e.g. Holden [2000] for details.

Algorithm 3: The CIMH algorithm

(i) **Input**: RWMH prerun samples $\{\breve{\boldsymbol{\xi}}^{(j)}\}_{j=0,\dots,T'}$ with $\breve{\boldsymbol{\xi}}^{(j)} = (\breve{\xi}_1^{(j)}, \dots, \breve{\xi}_n^{(j)})^\top$, sampling parameters r_1, and r_2, variable permutation function ι, chain length T, starting value $\boldsymbol{\xi}_0$, and transition densities q_2 and q_3.

Output: Markov chain $\{\boldsymbol{\xi}^{(j)}\}_{j=0,\dots,T}$.

Set $\boldsymbol{\xi}^{(0)} \leftarrow \boldsymbol{\xi}_0$

(ii) **for** $i \leftarrow 1$ **to** n **do**

  Fit $\hat{\boldsymbol{\gamma}}_i$ for parametrized cdf $G_i(\cdot|\boldsymbol{\gamma}_i)$ based on $\{\breve{\xi}_i^{(k)}\}_{k=0,\dots,T'}$

  for $k \leftarrow 0$ **to** $T' + j$ **do**

    Set $\breve{u}_i^{(k)} \leftarrow G_i(\breve{\xi}_i^{(k)}|\hat{\boldsymbol{\gamma}}_i)$

(iii) Fit $\hat{\boldsymbol{\theta}}$ for D-vine copula $c_{1,\dots,n}(u_1, \dots, u_n|\boldsymbol{\theta})$ based on $\{(\breve{u}_{\iota(1)}^{(k)}, \dots, \breve{u}_{\iota(n)}^{(k)})^\top\}_{k=0,\dots,T'}$

for $j \leftarrow 1$ **to** T **do**

(iv) Sample $r \sim \mathcal{U}[0,1]$

  if $r \leq r_1$ **then**

    Generate $(\tilde{u}_1, \dots, \tilde{u}_n)^\top$ with density $c_{1,\dots,n}(u_1, \dots, u_n|\hat{\boldsymbol{\theta}})$

    for $i \leftarrow 1$ **to** n **do**

      Set $\tilde{\xi}_i \leftarrow G_{\iota(i)}^{-1}(\tilde{u}_{\iota^{-1}(i)}|\hat{\boldsymbol{\gamma}}_{\iota(i)})$

    Define $\tilde{\boldsymbol{\xi}} = (\tilde{\xi}_1, \dots, \tilde{\xi}_n)^\top$

  else if $r_1 < r \leq r_1 + r_2$ **then**

    Sample $\tilde{\boldsymbol{\xi}} \sim q_2(\boldsymbol{\xi}|\boldsymbol{\xi}^{(j-1)})$

  else if $r > r_1 + r_2$ **then**

    Sample $\tilde{\boldsymbol{\xi}} \sim q_3(\boldsymbol{\xi})$

  Set

$$\boldsymbol{\xi}^{(j)} \leftarrow \begin{cases} \tilde{\boldsymbol{\xi}} & \text{with probability } \alpha^{cop}(\tilde{\boldsymbol{\xi}}|\boldsymbol{\xi}^{(j-1)}) \\ \boldsymbol{\xi}^{(j-1)} & \text{with probability } 1 - \alpha^{cop}(\tilde{\boldsymbol{\xi}}|\boldsymbol{\xi}^{(j-1)}) \end{cases}$$

As pointed out by Holden *et al.* [2009], if the strong Doeblin condition does not hold for some states in the sampling space, the MCMC algorithm will tend to undersample these areas. This is not crucial, if further inference does not depend on tail behavior, but may become problematic for questions of extreme value theory.

Remark 6.2 (Pairwise copula estimation). The estimation of the copula decomposition parameters of step (iii) for a given parameter order ι can be done as follows: Since the number of parameters grows quadratically in the dimension n, it is useful to consider a stepwise estimation approach, where we estimate the parameters from pair copulas

with no conditioning to the ones with $n-2$ conditioning variables. In an initial step, we estimate the parameters corresponding to the pair copulas with no conditioning. For the copula parameters with a single conditioning value, we transform the data with the appropriate conditional cdf's using the estimated parameters to determine pseudo realizations needed in the pair copulas with a single conditioning variable. We proceed as before until all parameters have been estimated. These so-called sequential estimates have shown to be consistent and asymptotically normally distributed (Hobæk Haff [2010]). They are then used as starting values for numerically determining the maximum likelihood estimates. When several bivariate copula families for a pair copula term are available, the family is chosen according to *Akaike's information criterion* (AIC). Brechmann [2010] shows that AIC performs well with regard to several alternatives. The R package *CDVine* of Brechmann & Schepsmeier [2011] can be used to fit the according D-vines.

Remark 6.3 (Likelihood based copula parameter estimation). Likelihood based copula parameter estimation was first proposed by Aas *et al.* [2009] and current developments in this active area can be found in Czado [2010] and Kurowicka & Joe [2011]. Bayesian analyses of D-vines using MCMC are also available (see Min & Czado [2010]). Furthermore, Bayesian model selection methods are implemented using indicator variables by Smith *et al.* [2010] and using reversible jump MCMC by Min & Czado [2011], respectively.

6.1.2 CIMH as adaptive sampling scheme

Interestingly, CIMH can also be seen as an *adaptive MCMC sampler*. Before we elaborate this, we need to introduce some more notation. We follow Roberts & Rosenthal [2007]: As usual, let $\boldsymbol{X}^{(t)}$ be a real valued random variable for $t \in \mathbb{N}_0$. Moreover, let $\Gamma^{(t)}$ be a $\mathcal{K}$-valued random variable representing the choice of the transition kernel for updating $\boldsymbol{X}^{(t)}$ to $\boldsymbol{X}^{(t+1)}$. The random variable $(\boldsymbol{X}^{(t)}, \Gamma^{(t)})$ generates a filtration

$$\mathcal{G}_t = \sigma(\boldsymbol{X}^{(0)}, \ldots, \boldsymbol{X}^{(t)}, \Gamma^{(0)}, \ldots, \Gamma^{(t)})$$

i.e. $\mathcal{G}_t$ is the smallest σ-algebra with respect to which $(\boldsymbol{X}^{(t)}, \Gamma^{(t)})$ is measurable for all $s < t$. Clearly, $\mathcal{G}_s \subseteq \mathcal{G}_t$ for all $s < t$. The Markov chain transition kernel is then

$$k_\gamma(A|\boldsymbol{\xi}) = P(\boldsymbol{X}^{(t+1)} \in A | \boldsymbol{X}^{(t)} = \boldsymbol{\xi}, \Gamma^{(t)} = \gamma, \mathcal{G}_{t-1})$$

for some $\boldsymbol{\xi} \in \mathbb{R}^n$, $A \subseteq \mathbb{R}^n$, and $\gamma \in \mathcal{K}$.

Definition 6.1 (Independent adaption)**.** An adaptive MCMC algorithm is called *independent adaption algorithm*, if for all $t \in \mathbb{N}$ the random variable $\Gamma^{(t)}$ is independent of $\boldsymbol{X}^{(0)}, \ldots, \boldsymbol{X}^{(t)}$.

Now, let us interpret CIMH as an independent adaption algorithm: Suppose $\{k_1, k_2, k_3\}$ is the set of Markov chain kernels induced[1] by the proposal functions q_1, q_2 and q_3 of CIMH, respectively. Then for $i = 1, \ldots, 3$, $k_i(\boldsymbol{\xi}|\boldsymbol{\xi}')$ is ergodic for the posterior distribution $\pi(\boldsymbol{\xi}|\boldsymbol{y})$, that is, each $k_i(\boldsymbol{\xi}|\boldsymbol{\xi}')$ converges to the equilibrium distribution $\pi(\boldsymbol{\xi}|\boldsymbol{y})$. Nevertheless, the speed of convergence can differ severely. Instead of using the proposal function q^{cop} introduced above, we can in each MCMC step j sample the proposal function q_i, $i \in \{1, 2, 3\}$, according to

$$P(\Gamma^{(j)} = 1) = r_1, \quad P(\Gamma^{(j)} = 2) = r_2, \quad \text{and} \quad P(\Gamma^{(j)} = 3) = 1 - r_1 - r2.$$

The Metropolis-Hastings acceptance probability is then computed with respect to q_i of step j instead of q^{cop}, i.e.

$$\alpha(\boldsymbol{\xi}^p|\boldsymbol{\xi}^{(j)}) = \min\left\{\frac{\pi(\boldsymbol{\xi}^p|\boldsymbol{y})q_i(\boldsymbol{\xi}^{(j)}|\boldsymbol{\xi}^p)}{\pi(\boldsymbol{\xi}^{(j)}|\boldsymbol{y})q_i(\boldsymbol{\xi}^p|\boldsymbol{\xi}^{(j)})}, 1\right\}.$$

This somewhat reduces the computational cost and speeds up the inference process as the proposed sample $\boldsymbol{\xi}^p \sim q_i$ of step j does not need to be evaluated by the proposal density functions q_j for $j \neq i$ in the Metropolis-Hastings acceptance probability. We call this sampling scheme *independent adaption CIMH*.

Proposition 6.2 (Convergence of CIMH as independent adaption algorithm)**.** *The independent adaption CIMH sampling scheme converges to the posterior equilibrium distribution.*

Proof. As $q_i(\cdot|\cdot) > 0$ for $i = 1, 2, 3$ on the support of the prior distributions, the corresponding Markov chain $\{\boldsymbol{X}^{(t)}\}_{t\in\mathbb{N}_0}$ of the independent adaption CIMH is $\pi(\cdot|\boldsymbol{y})$-irreducible, and Harris recurrent for the posterior $\pi(\cdot|\boldsymbol{y})$. Again, since for each $t \in \mathbb{N}$, $\boldsymbol{X}^{(t)} = \boldsymbol{X}^{(t+1)}$ with positive probability the aperiodicity condition holds. We are left to prove invariance of the posterior distribution, i.e.

$$\int_{\boldsymbol{\xi}\in\mathbb{R}^n} P(\boldsymbol{X}^{(t+1)} \in A|\boldsymbol{X}^{(t)} = \boldsymbol{\xi}, \mathcal{G}_{t-1})\pi(\mathrm{d}\boldsymbol{\xi}|\boldsymbol{y}) = \pi(A|\boldsymbol{y}) \qquad \text{for all } A \subseteq \mathbb{R}^n.$$

[1] See e.g. the proof of Theorem 4.1.

Since $\Gamma^{(t)}$ is independent of $\boldsymbol{X}^{(t')}$ for all $t, t' \in \mathbb{N}_0$ we have

$$\begin{aligned}
&\int_{\boldsymbol{\xi}\in\mathbb{R}^n} P(\boldsymbol{X}^{(t+1)} \in A | \boldsymbol{X}^{(t)} = \boldsymbol{\xi}, \mathcal{G}_{t-1}) \pi(\mathrm{d}\boldsymbol{\xi}|\boldsymbol{y}) \\
&= \int_{\boldsymbol{\xi}\in\mathbb{R}^n} \sum_{\gamma=1}^{3} P(\boldsymbol{X}^{(t+1)} \in A | \boldsymbol{X}^{(t)} = \boldsymbol{\xi}, \Gamma^{(t)} = \gamma, \mathcal{G}_{t-1}) P(\Gamma^{(t)} = \gamma | \boldsymbol{X}^{(t)} = \boldsymbol{\xi}, \mathcal{G}_{t-1}) \pi(\mathrm{d}\boldsymbol{\xi}|\boldsymbol{y}) \\
&= \int_{\boldsymbol{\xi}\in\mathbb{R}^n} \sum_{\gamma=1}^{3} P(\Gamma^{(t)} = \gamma | \mathcal{G}_{t-1}) k_\gamma(A|\boldsymbol{\xi}) \pi(\mathrm{d}\boldsymbol{\xi}|\boldsymbol{y}) = \pi(A|\boldsymbol{y}) \sum_{\gamma=1}^{3} P(\Gamma^{(t)} = \gamma | \mathcal{G}_{t-1}) = \pi(A|\boldsymbol{y}).
\end{aligned}$$

□

For more details on adaptive MCMC sampling we refer the reader to Fearnhead [2008].

6.2 Adaptive Copula based Independence MH algorithm (ACIMH)

Based on short preruns $\{\breve{\boldsymbol{\xi}}^{(j)}\}_{j=1,\dots T'}$, it is sometimes difficult to guarantee sufficient sampling from the posterior's marginals' tails in order to fit an efficient proposal copula. To avoid setting $r_1 + r_2 \ll 1$ and thus generating rather ineffective proposals, we propose an extension of the basic CIMH algorithm by sequentially updating the copula functions based on preceding Markov chain samples. This changes the proposal function q^{cop} during the sampling process and leads to a limited adaption scheme: For integers $R, S > 0$ we set the copula update probability for the j^{th} MCMC step, $P_u(j)$, to

$$P_u(j) = \begin{cases} 1, & \text{if } j \bmod R \equiv 0 \text{ and } j < R \cdot S \text{ and} \\ 0, & \text{otherwise.} \end{cases} \tag{6.6}$$

This is, the estimated copula parameters $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\theta}}$ become dependent on the proposal step j, resulting in a step dependent proposal function $q^{cop}(\boldsymbol{\xi}|\boldsymbol{\xi}^{(j)}, \hat{\boldsymbol{\gamma}}^{(j)}, \hat{\boldsymbol{\theta}}^{(j)})$, where $\hat{\boldsymbol{\gamma}}^{(j)}$ and $\hat{\boldsymbol{\theta}}^{(j)}$ are updated based on the concatenated prerun samples and the samples generated up to step j according to (6.6). We refer to the *hybrid Adaptive Copula update Independence chain/random walk MH algorithm* as ACIMH. The pseudo code for ACIMH is shown in Algorithm 4.

Proposition 6.3 (Convergence of ACIMH)**.** *The ACIMH sampling scheme converges to the posterior equilibrium distribution.*

Proof. The update probability (6.6) initiates a total of S update steps (including the initial copula estimation after the prerun). Hence, $\mathcal{K} = \{1, \ldots, S\}$. Setting

$$d(\boldsymbol{\xi}, \gamma, s) = \|k_\gamma^s(\cdot|\boldsymbol{\xi}) - \pi(\cdot|\boldsymbol{y})\|_{TV}$$

for the posterior distribution $\pi(\cdot|\boldsymbol{y})$, $\boldsymbol{\xi} \in \mathbb{R}^n$, $\gamma \in \{1, \ldots, S\}$, and the s step transition kernel

$$k_\gamma^s(A|\boldsymbol{\xi}) = P(\boldsymbol{X}^{(t+s+1)} \in A|\boldsymbol{X}^{(t)} = \boldsymbol{\xi}, \Gamma^{(t)} = \gamma, \mathcal{G}_{t-1}), \qquad (A \subseteq \mathbb{R}^n),$$

we have $\lim_{s\to\infty} d(\boldsymbol{\xi}, \gamma, s) = 0$ independent of $\boldsymbol{\xi} \in \mathbb{R}^n$. After $R' = (S-1) \cdot R$ Markov chain steps $\Gamma^{(R'+s)} = \Gamma^{(R')}$ for all $s \geq 0$. For any realization $\boldsymbol{\xi}^{(0)}, \ldots, \boldsymbol{\xi}^{R'}$ and $\gamma^{(0)}, \ldots, \gamma^{R'}$ of the Markov chain generated by ACIMH, $\lim_{s\to\infty} d(\boldsymbol{\xi}^{(R')}, \gamma^{(R')}, s) = 0$ independent of $\boldsymbol{\xi}^{(0)}, \ldots, \boldsymbol{\xi}^{R'}$ and $\gamma^{(0)}, \ldots, \gamma^{R'}$. Therefore, for all $A \subseteq \mathbb{R}^n$ and any $\boldsymbol{\xi}^{(R')}, \boldsymbol{\xi}^{(0)} \in \mathbb{R}^n$

$$0 = \lim_{s\to\infty} |k_\gamma^s(A|\boldsymbol{\xi}) - \pi(A|\boldsymbol{y})| = \lim_{s\to\infty} |P(\boldsymbol{X}^{(s+1)} \in A|\boldsymbol{X}^{(0)} = \boldsymbol{\xi}^{(0)}, \Gamma^{(0)} = \gamma^{(0)}) - \pi(A|\boldsymbol{y})|.$$

□

6.3 Performance of CIMH and ACIMH

For benchmarking CIMH and ACIMH, the algorithms were tested on four examples: First, we draw samples from a strongly correlated two dimensional normal distribution. This is a simple proof-of-concept example of an analytically tractable system. Subsequently, we turn to dynamic systems defined by differential equations (DEs). More precisely, examples 2 and 3 examine the performance for ordinary nonlinear parameter dependencies and parameter distributions with non-symmetric tail dependencies. Finally, we apply our samplers to a delay differential equation (DDE) model of the JAK2-STAT5 signaling pathway as published by Swameye *et al.* [2003]. Here, a sophisticated proposal generation is crucial as there exists no closed form solution of the DDE system, calling for a computationally very expensive numerical solution for every evaluation of the likelihood. Moreover, the seven parameters involved show high dependency, which additionally complicates the posterior inference.

We evaluated the following performance indices: ($\mathcal{I}_1$) the quotient of acceptance rate and INEFF. This was motivated by the antagonistic behavior of high acceptance rates

Algorithm 4: The ACIMH algorithm

(i) **Input**: RWMH prerun samples $\{\breve{\boldsymbol{\xi}}^{(j)}\}_{j=0,\dots,T'}$ with $\breve{\boldsymbol{\xi}}^{(j)} = (\breve{\xi}_1^{(j)}, \dots, \breve{\xi}_n^{(j)})^\top$, update and sampling parameters R, S, r_1, and r_2, variable permutation function ι, chain length T, starting value $\boldsymbol{\xi}_0$, and transition densities q_2 and q_3.

Output: Markov chain $\{\boldsymbol{\xi}^{(j)}\}_{j=0,\dots,T}$.

Initialize $s \leftarrow 0$
Set $\boldsymbol{\xi}^{(0)} \leftarrow \boldsymbol{\xi}_0$
for $j \leftarrow 0$ **to** T **do**
  if j *mod* $R \equiv 0$ *and* $j < R \cdot S$ **then**
    Update $s \leftarrow s + 1$
(ii) **for** $i \leftarrow 1$ **to** n **do**
      Fit $\hat{\boldsymbol{\gamma}}_i^{(s)}$ for parametrized cdf $G_i(\cdot|\boldsymbol{\gamma}_i)$ based on $\{\breve{\xi}_i^{(k)}\}_{k=0,\dots,T'+j}$
      for $k \leftarrow 0$ **to** $T' + j$ **do**
        Set $\breve{u}_i^{(k)} \leftarrow G_i(\breve{\xi}_i^{(k)}|\hat{\boldsymbol{\gamma}}_i^{(s)})$
(iii) Fit $\hat{\boldsymbol{\theta}}^{(s)}$ for D-vine copula $c_{1,\dots,n}(u_1, \dots, u_n|\boldsymbol{\theta})$ based on $\{(\breve{u}_{\iota(1)}^{(k)}, \dots, \breve{u}_{\iota(n)}^{(k)})^\top\}_{k=0,\dots,T'+j}$
(iv) Sample $r \sim \mathcal{U}[0,1]$
  if $j > 0$ **then**
    if $r \leq r_1$ **then**
      Generate $(\tilde{u}_1, \dots, \tilde{u}_n)^\top$ with density $c_{1,\dots,n}(u_1, \dots, u_n|\hat{\boldsymbol{\theta}}^{(s)})$
      for $i \leftarrow 1$ **to** n **do**
        Set $\tilde{\xi}_i \leftarrow G_{\iota(i)}^{-1}(\tilde{u}_{\iota^{-1}(i)}|\hat{\boldsymbol{\gamma}}_{\iota(i)}^{(s)})$
      Define $\tilde{\boldsymbol{\xi}} = (\tilde{\xi}_1, \dots, \tilde{\xi}_n)^\top$
    else if $r_1 < r \leq r_1 + r_2$ **then**
      Sample $\tilde{\boldsymbol{\xi}} \sim q_2(\boldsymbol{\xi}|\boldsymbol{\xi}^{(j-1)})$
    else if $r > r_1 + r_2$ **then**
      Sample $\tilde{\boldsymbol{\xi}} \sim q_3(\boldsymbol{\xi})$
    Set

$$\boldsymbol{\xi}^{(j)} \leftarrow \begin{cases} \tilde{\boldsymbol{\xi}} & \text{with probability } \alpha^{cop}(\tilde{\boldsymbol{\xi}}|\boldsymbol{\xi}^{(j-1)}) \\ \boldsymbol{\xi}^{(j-1)} & \text{with probability } 1 - \alpha^{cop}(\tilde{\boldsymbol{\xi}}|\boldsymbol{\xi}^{(j-1)}) \end{cases}$$

    and $\breve{\boldsymbol{\xi}}^{(T'+j)} \leftarrow \boldsymbol{\xi}^{(j)}$

versus high INEFF's as the Markov chain converges slowly for small proposal variances and the MH algorithm conversely rejects a large amount of its proposed moves for too high variances (see Roberts *et al.* [1997], Liu [2008], or Girolami & Calderhead [2011]).

Clearly $(\mathcal{I}_1) \in [0,1]$, with higher values being superior. Furthermore, we monitor the number $(\mathcal{I}_2)$ of i.i.d. samples generated per second. As all algorithms were implemented in Matlab using the same underlying MH code, $(\mathcal{I}_2)$ is well justified. Time here denotes the CPU-time on a two Six-Core Opteron 2427 (2.2 GHz) machine.

Using $(\mathcal{I}_1)$ and $(\mathcal{I}_2)$, the performance of CIMH and ACIMH was compared to (i) a regular RWMH, (ii) an IMH, and (iii) a random walk MH algorithm with a covariance based proposal function (CovRWMH) in the first three examples. The JAK2-STAT5 pathway was additionally evaluated by means of the (iv) SMALA and (v) M-GaA algorithms. In each example jointly updating all parameters of RWMH at once outperformed single parameter updates with respect to the acceptance rates. Therefore, a joint update scheme was used for all algorithms. Since there is generally little to no knowledge about the underlying parameter dependency, we applied an uncorrelated normal update function in RWMH, i.e. a new proposal $\boldsymbol{\xi}^p$ was generated based on the current sample $\boldsymbol{\xi}^{(c)}$ by $\boldsymbol{\xi}^p = \boldsymbol{\xi}^{(c)} + \boldsymbol{\varepsilon}$ for $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{RW})$ with $\boldsymbol{\Sigma}_{RW}$ defined as follows: We determined the maximum a posteriori estimate for all n parameters using a simulated annealing algorithm. Denoting these estimates by s_i, the i^{th} diagonal element of $\boldsymbol{\Sigma}_{RW}$ was set to $k_{RW} \cdot s_i$, where k_{RW} was adjusted in each example to yield an acceptance rate of approximately 23% as suggested in Roberts *et al.* [1997] – our exact limits were set to 10% and 36%. This approach tries to compensate the sometimes large differences in parameter magnitude and therefore sensitivity. The unthinned Markov chains sampled by RWMH were directly taken as prerun samples for IMH, CovRWMH, CIMH, and ACIMH. The sampling times for RWMH were added to the respective sampling times of IMH, CovRWMH, CIMH, and ACIMH.

Two major issues of RWMH are $(\mathcal{P}_1)$ a rather strong autocorrelation between subsequent MCMC iterations and $(\mathcal{P}_2)$ its lack of incorporating any information about the limiting distribution when proposing new samples. To address $(\mathcal{P}_1)$ we set up the IMH whose proposals are generated independently of the current state of the Markov chain: as for the copula based algorithms we fit one-dimensional parametrized cdf's $G_i(\xi|\boldsymbol{\gamma}_i)$ to each of the n empirical marginal parameter distributions sampled in the prerun. In fact, these were identical for IMH, CIMH, and ACIMH. The IMH proposals $\tilde{\xi}_i^{(j)}$ were jointly generated by sampling $n \cdot (T+1)$ independent samples $u_i^{(j)} \sim \mathcal{U}[0,1]$ $(i = 1, \ldots, n,$ $j = 0, \ldots, T)$, which are subsequently transformed to $\tilde{\xi}_i^{(j)} = G_i^{-1}(u_i^{(j)}|\hat{\boldsymbol{\gamma}}_i)$. In other

words, IMH generates proposals assuming an independent parameter structure. The MH acceptance probability is given by

$$\alpha = \min\left\{\frac{\pi(\boldsymbol{\xi}^p|\boldsymbol{y}) \cdot \prod_i g_i(\xi_i^{(c)}|\hat{\boldsymbol{\gamma}}_i)}{\pi(\boldsymbol{\xi}^{(c)}|\boldsymbol{y}) \cdot \prod_i g_i(\xi_i^p|\hat{\boldsymbol{\gamma}}_i)}, 1\right\},$$

where $g_i(\xi|\hat{\boldsymbol{\gamma}}_i)$ denotes the pdf to $G_i(\xi|\hat{\boldsymbol{\gamma}}_i)$. This somewhat reduces the autocorrelation compared to the Markov chain of the RWMH since all proposals are independent of any foregoing MCMC elements. Nevertheless, the IMH Markov chain can still be strongly autocorrelated and needs to be thinned. Note that with independence in $\pi(\boldsymbol{\xi}|\boldsymbol{y}) = \prod_i \pi(\xi_i|\boldsymbol{y})$ and $g_i(\xi_i|\hat{\boldsymbol{\gamma}}_i) = \pi(\xi_i|\boldsymbol{y})$ for all i, the MH acceptance probability collapses to $\alpha = 1$, allowing the IMH to generate an independent sample in each MCMC step.

Rather than directly reducing the autocorrelation in the Markov chain, CovRWMH exploits the expected covariance matrix $\hat{\boldsymbol{C}}$ of the prerun and addresses $(\mathcal{P}_2)$: The regular RWMH proposal function is changed to $\boldsymbol{\xi}^p = \boldsymbol{\xi}^{(c)} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, k_{CovRW} \cdot \hat{\boldsymbol{C}})$. Simulations show that the acceptance rate of CovRWMH outperforms the acceptance rate of RWMH for $k_{CovRW} = k_{RW}$. Thus, fine tuning k_{CovRW} to yield an approximate acceptance rate of 23% generally decreases the autocorrelation in the CovRWMH Markov chain by increasing the increments $\boldsymbol{\xi}_i^p - \boldsymbol{\xi}_i^{(c)}$ compared to the regular RWMH.

While IMH and CovRWMH can in some sense be seen as antagonistic approaches with respect to $(\mathcal{P}_1)$ and $(\mathcal{P}_2)$, CIMH and ACIMH address both issues at once. For performance assessment CIMH and ACIMH were applied as introduced in Chapter 6.1 and 6.2. Throughout, the Metropolis-Hastings proposal function q_2 was taken to be identical with the one of the CovRWMH, reusing the tuning parameter k_{CovRW}; q_3 and the proposal probabilities r_1 and r_2 were adjusted individually (see Chapter 6.3.1 - 6.3.4). For thorough performance evaluation, the first three examples were each run 100 times for 50,000 MCMC iterations, the last one 10 times for 50,000 MCMC iterations. In all examples the copula update parameters for ACIMH were set to R =10,000 and S =4. While the copulas were fitted on 1,000 prerun samples in the first three examples, we used 3,000 samples for the JAK2-STAT5 inference, owing to the complexity of the system. The time for the prerun was added to the sampling times of IMH, CovRWMH, CIMH, and ACIMH, respectively. Note that we do not need to generate independent samples from the prerun's MCMC chain since

fitting the copula to dependent data only slightly effects the efficiency of the proposal distribution. This also holds for fitting $G_i(\xi|\boldsymbol{\gamma}_i)$ and the estimation of the covariance matrix in the CovRWMH proposal function. Nevertheless, insufficient coverage of the sampling space by the prerun samples decreases the performance of all samplers. For copula fitting and copula sample generation the *CDVine* R-package (Brechmann & Schepsmeier [2011]) was used. Here, sampling from $c_{1,\dots,n}(\boldsymbol{u}|\hat{\boldsymbol{\theta}})$ is done sequentially as proposed by Aas *et al.* [2009]. For the copula type of each pair copula term 32 types were available: Implemented are the independence copula, Gaussian copula ($\mathcal{N}$), Student-t copula (S), Clayton copula (C), Gumbel copula (G), Frank copula (F), Joe copula (J), $BB1$ copula, $BB6$ copula, $BB7$ copula, $BB8$ copula, as well as the corresponding $90°, 180°$, and $270°$ rotated versions of $S, C, G, F, J, BB1, BB6, BB7$, and $BB8$ (see Appendix B for details). Copula type and the corresponding parameter(s) for each copula term in (6.1) are estimated by a sequential likelihood based approach described in Dißmann *et al.* [2011]. Finally, the SMALA and M-GaA algorithms were directly applied as introduced in Chapter 5.1 and 5.3. In both cases the respective free algorithmic parameters ε and α_0 were tuned to yield a Metropolis-Hastings acceptance rate between $10\% - 36\%$ and of 23%, respectively. The SMALA geometric tensor $\boldsymbol{G}(\cdot)$ for the JAK2-STAT5 DDE system is computed in Appendix E.

6.3.1 Sampling from a strongly correlated 2-dim. normal distribution

In our first example we want to draw samples from a strongly correlated bivariate normal distribution $\mathcal{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with respective mean and covariance matrix

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} 1 & 0.95 \cdot \sqrt{3} \\ 0.95 \cdot \sqrt{3} & 3 \end{pmatrix}.$$

Here, $\rho = 0.95$. The MH acceptance probability for this example is given by

$$\alpha = \min\left\{ \frac{\Phi_2\left(\boldsymbol{\xi}^p \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \cdot q(\boldsymbol{\xi}^{(c)}|\boldsymbol{\xi}^p)}{\Phi_2\left(\boldsymbol{\xi}^{(c)} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \cdot q(\boldsymbol{\xi}^p|\boldsymbol{\xi}^{(c)})}, 1 \right\},$$

where $\Phi_2\left(\boldsymbol{\xi} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ is the density function of $\mathcal{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $q(\boldsymbol{\xi}|\boldsymbol{\xi}')$ the proposal function.

We chose this example as it is both illustrative as well as analytically tractable. Canonically, the cdf's of $\mathcal{N}(0,1)$ and $\mathcal{N}(0,3)$ were used to transform the prerun samples to

$[0,1]^2$. The independence proposal density q_3 was taken to be a bivariate student-t distribution with location parameter $(0,1)^\top$ and identity scale matrix. Furthermore, we set $r_1 = 0.99$ and $r_2 = 0$. Table 6.1 (Lower table) nicely shows that all samplers approximate the two dimensional normal distribution with negligible errors. This can also be seen from the thinned Markov chain samples depicted in Figure 6.1. The samples are based on the (unthinned) Markov chains from Figure 6.2. The chains exhibit a better mixing behavior when generated by a copula based algorithm compared to the non-copula based ones. Although the sampling times for IMH and CovRWMH were about twice as long as for RWMH (Table 6.1 (Upper table)), IMH is comparable with RWMH and CovRWMH even outperforms RWMH with respect to $(\mathcal{I}_1)$ and $(\mathcal{I}_2)$ (Figure 6.3(a) and 6.3(b)). The most efficient of all algorithms turned out to be CIMH. Caused by the additional time needed for copula refitting, ACIMH in average produced slightly less independent samples per second than CovRWMH. We have to point out that $(\mathcal{I}_1)$ is very close to one for CIMH and ACIMH. This means that in almost every MCMC iteration an independent sample was generated (see also Figure 6.4 for the autocorrelation functions after thinning by INEFF). At first sight this might almost seem too good a result, but clearly, due to the simplicity of the problem, the copula was fit almost perfectly (Figure 6.3(c)) leading to an independent proposal function $q^{cop}(\boldsymbol{\xi}|\boldsymbol{\xi}') = q^{cop}(\boldsymbol{\xi})$ that is very close to the true sampling distribution $\mathcal{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. This pushes the MH acceptance probability close to one. We will see in later examples that this index generally attains high values. It can also be seen as combined goodness-of-fit index for the fitted marginal cdf's and vine copula decomposition.

The copula families for ACIMH did not change in any of the 100 runs, meaning that the dependency structure was already well covered by the preruns. The bivariate copula $c_{1,2}(u_1, u_2|\boldsymbol{\theta})$ of the first of the 100 runs was fit to be Gaussian with an estimated parameter value of $\hat{\boldsymbol{\theta}} = 0.953$. This is very close to the actual correlation value of $\rho = 0.95$. The corresponding Kendall's τ for the copula parameters was estimated to be $\hat{\tau}_m = 0.805$, which coincides with the Kendall's τ estimated for the prerun (Figure 6.3(c)). All other runs showed similar outcomes. Note that the time needed for fitting a single copula is almost identical to the time needed for the complete RWMH run (c.f. Table 6.1 (Upper table)). All MCMC runs were started at the origin.

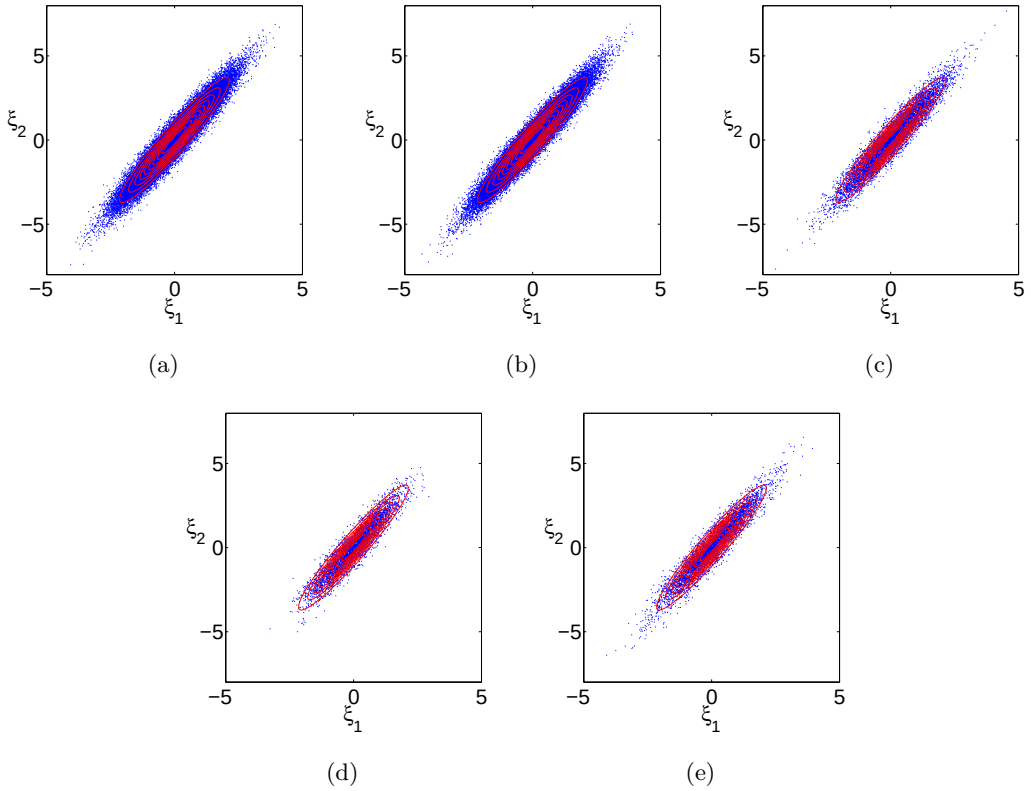


Figure 6.1: Thinned Markov chain samples of the first run of the (a) ACIMH, (b) CIMH, (c) CovRWMH, (d) IMH, and (e) RWMH. The red lines display the $p \cdot 10\%$ quantiles of the normal distribution for $p = 1, \ldots, 9$.

Thinning was performed according to the INEFF and caused a nice decrease in the autocorrelation functions as depicted in Figure 6.4. Interestingly, the INEFF slightly underestimated the thinning rate for the last three algorithms, which can be seen from the rather slow decreases in the autocorrelation functions. A nice analogy between CovRWMH and the copula based algorithms is given by the fact that all three were using the same Gaussian copula. However, while CovRWMH was applying it for locally proposing new samples, the ($\mathcal{I}_1$) index was quite low compared to CIMH and ACIMH. On the other hand, CovRWMH was taking less than half the time of the copula algorithms resulting in a very good performance on ($\mathcal{I}_2$).

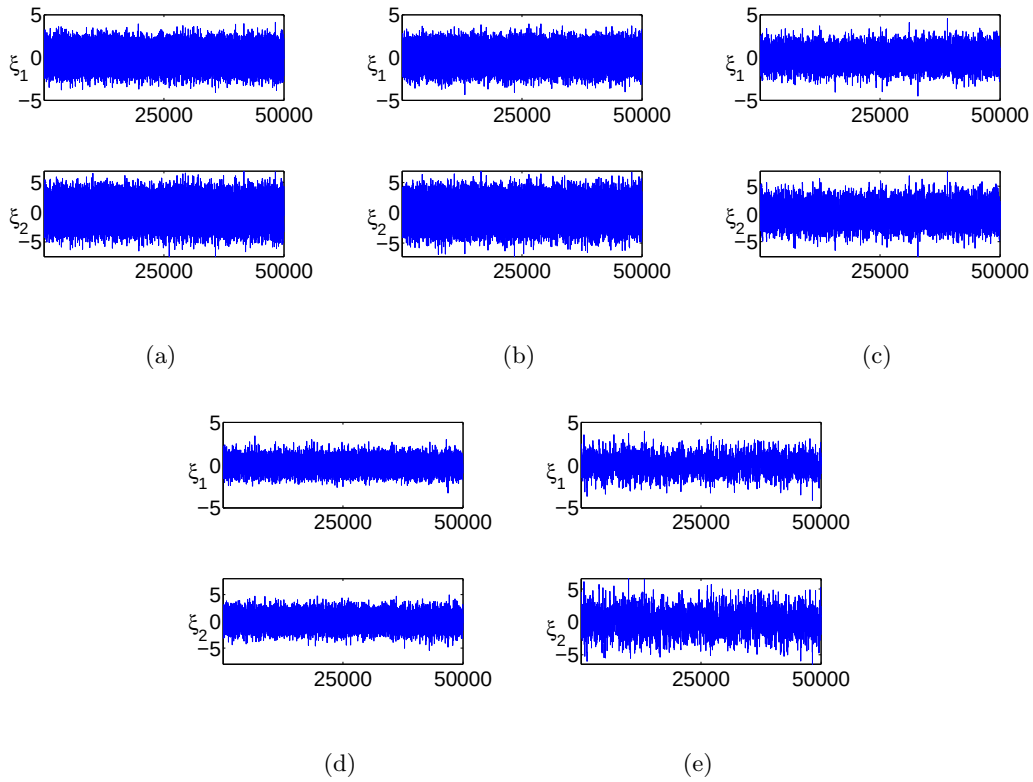


Figure 6.2: Unthinned Markov chains of the first run of the (a) ACIMH, (b) CIMH, (c) CovRWMH, (d) IMH, and (e) RWMH. While the x-axis holds the step number, the y-axis displays the parameter value. The copula based samplers show a slightly better mixing behavior compared to the non-copula algorithms.

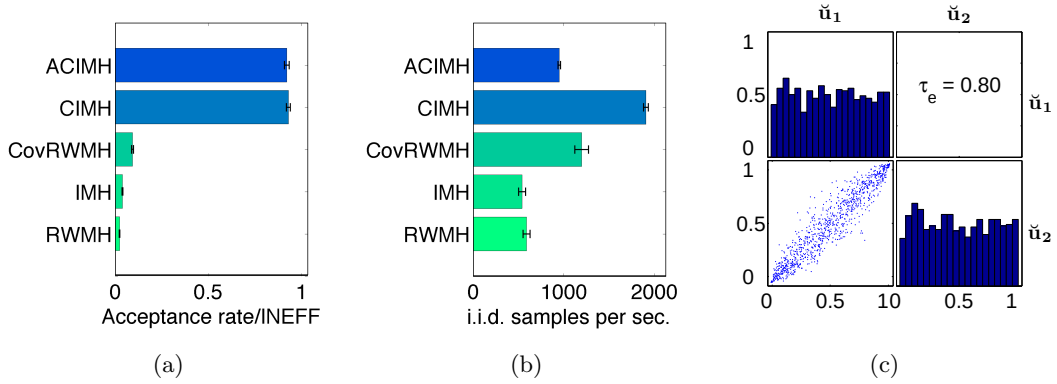
6.3.2 Performance on a steady state model with nonlinear parameter dependency

We will now consider posterior inference in dynamic systems. We first take a look at a simple, but completely unidentifiable[1] steady state example: Consider the system

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = 0 \qquad \text{with} \qquad x(0) := x_0 \qquad \text{unknown.} \tag{6.7}$$

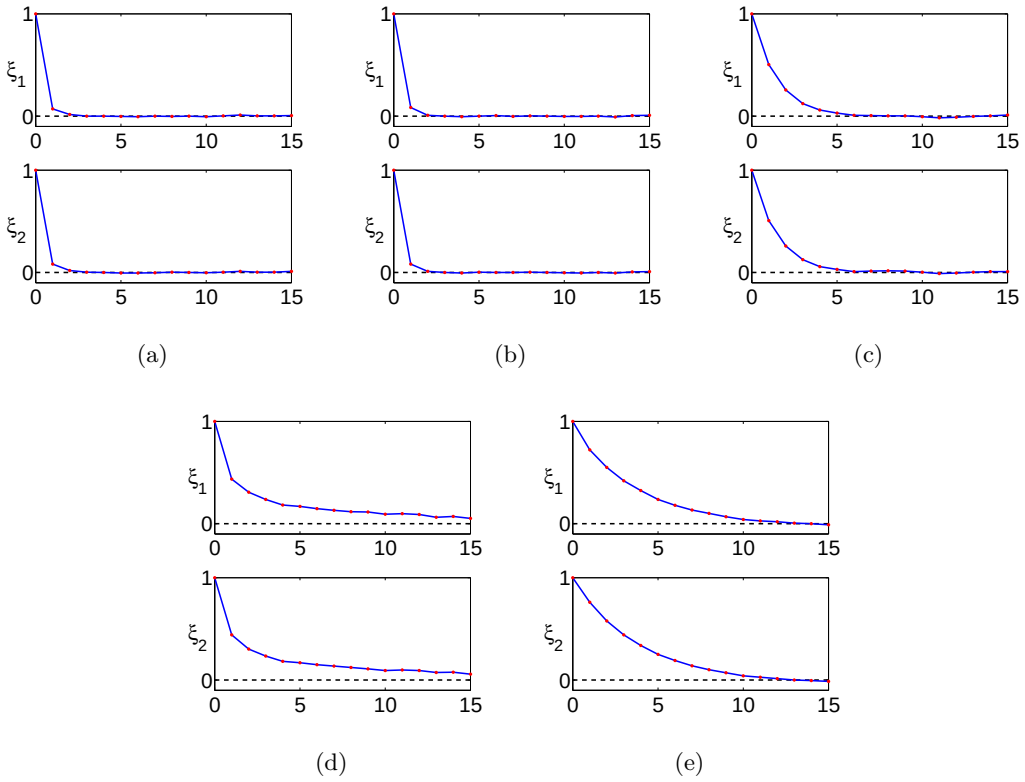
We are interested in the behavior of $y(t) = kx(t)$, where $x(t)$ is the solution to (6.7). The parameter k is introduced since biological experiments very often only yield relative concentrations. Again, k is unknown. At the time points $t = t_i = i$ $(i = 1, \ldots, 5)$ we

[1]Note that MCMC algorithms are able to deal with unidentifiable systems as such. The inference of meaningful reaction rates is however not possible in this scenario.



(a) (b) (c)

Figure 6.3: Results for the 2-dim. normal distribution. Figure (a): Quotient of acceptance rate versus INEFF ($\mathcal{I}_1$). Figure (b): Number of i.i.d. samples drawn per second ($\mathcal{I}_2$). Error bars show the estimated standard errors based on 100 runs. Figure (c): Marginal copula data (c.f. Chapter 6.1.1 (ii)) to fit the CIMH/ACIMH copula of run one – for uniformization the cdf's of $\mathcal{N}(0,1)$ and $\mathcal{N}(0,3)$ were applied. The diagonal holds the histograms of the MCMC sample marginals and τ_e is the corresponding empirical Kendall's τ.



(a) (b) (c)

(d) (e)

Figure 6.4: Autocorrelation functions of the first run of the (a) ACIMH, (b) CIMH, (c) CovRWMH, (d) IMH, and (e) RWMH. All plots range over 15 lags on the x-axis. The INEFF is slightly underestimated in the last three algorithms.

Sampler	i.i.d. samples	INEFF	AR (%)	time (sec.)
ACIMH	48250.0± 641.1	1.1±0.03	95.3±0.6	51.1± 0.3
CIMH	48750.0± 547.6	1.1±0.02	95.2±1.1	25.8± 0.2
CovRWMH	21051.5±1318.3	3.6±0.30	21.9±0.4	$17.6\pm4 \cdot 10^{-3}$
IMH	7821.3± 566.4	14.9±1.70	24.3±0.4	$14.6\pm7 \cdot 10^{-3}$
RWMH	4836.9± 316.2	17.9±1.70	23.5±0.2	$8.3\pm2 \cdot 10^{-3}$

Sampler	$\hat{\mu}_1 - \mu_1$ ($\times 10^{-3}$)	$\hat{\mu}_2 - \mu_2$ ($\times 10^{-3}$)	$\hat{\sigma}_1^2 - \sigma_1^2$ ($\times 10^{-3}$)	$\hat{\sigma}_2^2 - \sigma_2^2$ ($\times 10^{-3}$)	$\hat{\rho} - \rho$ ($\times 10^{-5}$)
ACIMH	0.40±0.46	0.27±0.77	−0.06±0.77	−1.35± 2.17	−692.84±0.04
CIMH	−0.33±0.50	−0.46±0.86	−0.86±0.62	−1.54± 1.68	−692.90±0.05
CovRWMH	−1.91±1.37	−3.80±2.41	0.48±2.17	2.44± 6.51	−692.85±0.13
IMH	1.74±3.19	3.82±5.57	−4.92±7.25	−13.88±20.96	−693.69±0.38
RWMH	4.12±3.07	6.94±5.41	1.52±3.67	5.04±11.92	−692.79±3.07

Table 6.1: Two-dimensional normal example. Upper table: Depicted are the average number of i.i.d. samples per run (ESS), INEFF's, acceptance rates (AR), and sampling times based on 100 runs including estimated standard errors. All samplers ran for 50,000 MCMC proposals. Lower table: Residual differences between the average posterior mean, standard deviation, and correlation coefficient estimates and the true parameter values based on 100 runs including estimated standard errors. The true values are $\mu_1 = 0$, $\mu_2 = 0$, $\sigma_1^2 = 1$, $\sigma_2^2 = 3$, and $\rho = 0.95$.

measure noisy observations y_i of $y(t)$, which are assumed to follow $y_i \sim \mathcal{N}(kx_i, 0.1^2)$, with $x_i = x(t_i)$. For simplicity we chose $k = x_0 = 1$ for toy data generation (see Figure 6.5(a)). We thus have $x(t) = kx_0 = 1$. Although there is clearly no way to determine any of the parameters k or $x_0 = 1/k$ here, models such as the one of the JAK2-STAT5 pathway introduced by Swameye *et al.* [2003] suffer a similar kind of practical unidentifiability (see Section 6.3.4). Nevertheless, the system provides a nice benchmark for the performance of our algorithms on nonlinear parameter dependencies. Since no prior knowledge other than $x_0 > 0$ and $k > 0$ is available we assume independent uniform prior distributions $x_0 \sim \mathcal{U}[0, 2.5]$ and $k \sim \mathcal{U}[0, 2.5]$. The system's posterior distribution is then given by

$$\begin{aligned} \pi(x_0, k|\boldsymbol{y}) &= \frac{1}{Z} \prod_{i=1}^{5} \Phi(y_i|kx_i, 0.1^2) \mathbf{1}_{[0,2.5]}(x_0) \mathbf{1}_{[0,2.5]}(k) \\ \text{with normalizing constant} \qquad Z &= \int_0^{2.5} \int_0^{2.5} \prod_{i=1}^{5} \Phi(y_i|kx_i, 0.1^2) \, \mathrm{d}x_0 \, \mathrm{d}k. \end{aligned} \tag{6.8}$$

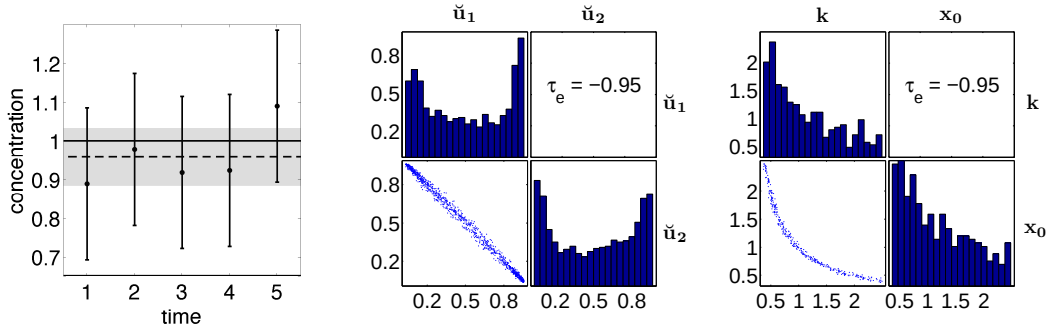
Here, $\Phi(\cdot|\mu,\sigma^2)$ denotes the density function of a univariate normal random variable with mean μ and variance σ^2 and $\mathbf{1}_{[a,b]}$ the density function corresponding to $\mathcal{U}[a,b]$. Note that $x_i = x(t_i)$ depends on the initial condition x_0.

Sampler	i.i.d. samples	INEFF	AR (%)	time (sec.)
ACIMH	35446.3±1511.5	1.8± 0.1	70.40±0.05	43.948± 1.3
CIMH	35224.9±1501.6	1.8± 0.1	70.20±0.09	23.043± 0.3
CovRWMH	1594.4± 134.5	56.4± 4.5	10.48±0.04	$13.847\pm6\cdot10^{-3}$
IMH	1576.6± 128.9	58.9± 5.5	5.43±0.02	$11.129\pm6\cdot10^{-3}$
RWMH	335.8± 23.5	241.5±18.9	21.14±0.03	$6.532\pm3\cdot10^{-3}$

	ACIMH	CIMH	CovRWMH	IMH	RWMH
$\mathbb{E}[k\|\boldsymbol{y}]$	1.129±0.001	1.128±0.001	1.125±0.003	1.129±0.003	1.124±0.007
$\mathbb{E}[x_0\|\boldsymbol{y}]$	1.129±0.001	1.130±0.001	1.132±0.003	1.128±0.003	1.135±0.008

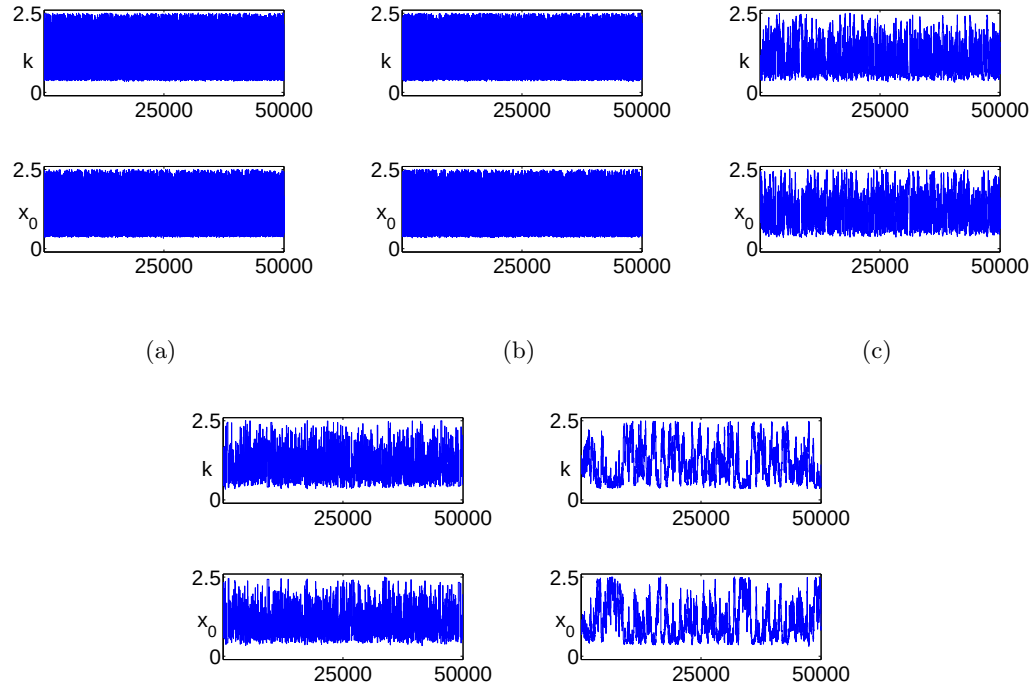
Table 6.2: Steady state model. Upper table: Depicted are the average number of i.i.d. samples per run (ESS), INEFF's, acceptance rates (AR), and sampling times based on 100 runs including standard errors. All samplers ran for 50,000 MCMC proposals. Lower table: Average posterior mean estimates for k and x_0 based on MCMC samples. Standard errors are estimated based on 100 runs. The numerical estimate is 1.129.

We chose the independence proposal density q_3 to be uniform on $[0,2.5]^2$, which coincides with the joint prior distribution. Again, $r_1 = 0.99$ and $r_2 = 0$. The prerun samples $\breve{\boldsymbol{\xi}}^{(j)} = (k^{(j)}, x_0^{(j)})^\top$ $(j = 1,\ldots,50{,}000)$ shown in Figure 6.5(c) indicate that the sample marginals for k and x_0 are close to being lognormally distributed. Hence, we used two lognormal distributions $g_i(\cdot|\mu_i,\sigma_i)$ $(i = 1,2)$ for the uniformization step. Although for $i = 1,2$ and fitted parameters $\hat{\mu}_i, \hat{\sigma}_i$ for μ_i, σ_i the transformed samples $u^{\breve{(j)}}{}_i = g_i(\xi^{\breve{(j)}}{}_i|\hat{\mu}_i,\hat{\sigma}_i)$ $(j = 1,\ldots,50{,}000)$ are not exactly uniformly distributed (c.f. histograms in Figure 6.5(b)), the CIMH and ACIMH proposals were accepted with a probability of about 70% (Table 6.2 (Upper table)). The approximation of the copula proposal function to the posterior of equation (6.8) was nevertheless very good as the index ($\mathbb{J}_1$) had values around 0.5 for CIMH and ACIMH (Figure 6.7(a)). This can also be seen from the excellent mixing behavior of the Markov chains for the copula based sampling schemes (Figure 6.6). Moreover, these algorithms drew an independent sample of the posterior in approximately every second MCMC iteration. We also tested uniformization by exponential and Gamma distributions ending up with comparable results (not shown). While the performance of CovRWMH with respect to



(a) (b) (c)

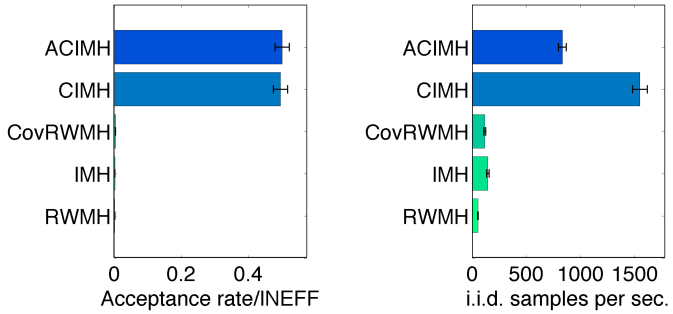
Figure 6.5: (a) Toy data used for sampling the steady state model. Depicted are the true underlying concentration $x(t)$ (solid line), the posterior median solution (dashed line) as well as the according 95% credible interval (shaded area) of the thinned first ACIMH run. The dots depict noisy data y_i including 95% confidence intervals. (b) Copula data (c.f. Section 6.1.1 (ii)) of the first run used to fit the CIMH copula. For uniformization of k and x_0 the cdf's of $\mathcal{LN}(\mu_1$= -0.03, σ_1=0.54) and $\mathcal{LN}(\mu_2$=0.01, σ_2=0.54) were applied. The diagonal displays the histograms of the MCMC sample marginals and τ_e the respective empirical Kendall's τ. (c) Thinned MCMC samples of the first RWMH run.



(a) (b) (c)

(d) (e)

Figure 6.6: Unthinned Markov chains of the first run of the (a) ACIMH, (b) CIMH, (c) CovRWMH, (d) IMH, and (e) RWMH. While the x-axis holds the step number, the y-axis displays the parameter value. The copula based samplers show a superior mixing.



(a) (b)

Figure 6.7: Results for the steady state model. Figure (a): Quotient of acceptance rate and INEFF ($\mathcal{J}_1$). Figure (b): Number of independent samples drawn per second ($\mathcal{J}_2$). Error bars show the estimated standard errors based on 100 runs.

($\mathcal{J}_2$) was similar to the performance of ACIMH in the first example, it here ran into problems generating adequate proposals due to the nonlinear parameter dependency. The estimated Pearson correlation coefficient used by CovRWMH lay around $\hat{\rho} = -0.9$, which is close to the rank based estimated Kendall's τ, given as $\tau_e = -0.95$ (Figure 6.5(b)). Nevertheless, ACIMH outperformed CovRWMH as well as RWMH and IMH more than 5-fold with respect to ($\mathcal{J}_2$). All ACIMH copulas were fitted to be of Gaussian type with a correlation parameter of $\hat{\boldsymbol{\theta}} = -0.99$ and did not change throughout the sampling process. Saving the time for copula refitting, CIMH even outperformed the non copula based algorithms more than 10-fold (Figure 6.7(b)). The MCMC samples' based solutions to equation (6.7) nicely approximate the data: Figure 6.5(a) depicts the posterior median solution with corresponding 95% credible interval for the thinned first ACIMH run, i.e. at time point t equation (6.7) was solved numerically for all ACIMH MCMC samples (after thinning); subsequently the pointwise median over all solutions – called *posterior median solution* – as well as the 95% credible interval were computed. Although in general neither the posterior median solution, nor the upper or lower boundary of the 95% credible interval need to be a solution to a differential equation system, the steady state property in this example guarantees that all three are in fact solutions to (6.7).

For analytically verifying the sampling results, we numerically evaluated the expected

posterior means of k and x_0. The means are given by

$$\mathbb{E}[x_0|\boldsymbol{y}] = \mathbb{E}[k|\boldsymbol{y}] = \frac{\int_0^{2.5} k \int_0^{2.5} \prod_{i=1}^{5} \Phi(y_i|kx_i, 0.1^2)\mathrm{d}x_0\mathrm{d}k}{\int_0^{2.5} \int_0^{2.5} \prod_{i=1}^{5} \Phi(y_i|kx_i, 0.1^2)\mathrm{d}x_0\mathrm{d}k} = 1.129. \tag{6.9}$$

As Table 6.2 (Lower table) shows, all samplers closely approximated (6.9).

6.3.3 Performance on a small compartment model

Our last toy example is motivated by a model for the biokinetic behavior of zirconium (Zr) in the human body (see Chapter 8). Compartmentalizing major organs, Li *et al.* [2011a,b] analyzed the circulation of Zr in the human body. The paper compares transfer rates of two competing compartment models with respect to sensitivity and predictability in order to establish a new model for radiation risk analysis. Both models are structurally identical as far as the interaction of the compartments "Small intestine" and "Transfer" is concerned, which is what our toy model is based on: After ingestion Zr passes through the "Small intestine". Subsequently it is either excreted directly or passes through the "Transfer" compartment as depicted in Figure 6.8(a). Since taking accurate measurements of Zr in the "Small intestine" compartment is technically not possible, we chose to generate data for the "Transfer" compartment only. The differential equations underlying the data are

$$\frac{\mathrm{d}x_1(t)}{\mathrm{d}t} = -k_2 x_1(t) - k_3 x_1(t) \qquad \text{and} \qquad \frac{\mathrm{d}x_2(t)}{\mathrm{d}t} = k_2 x_1(t) - k_1 x_2(t) \tag{6.10}$$

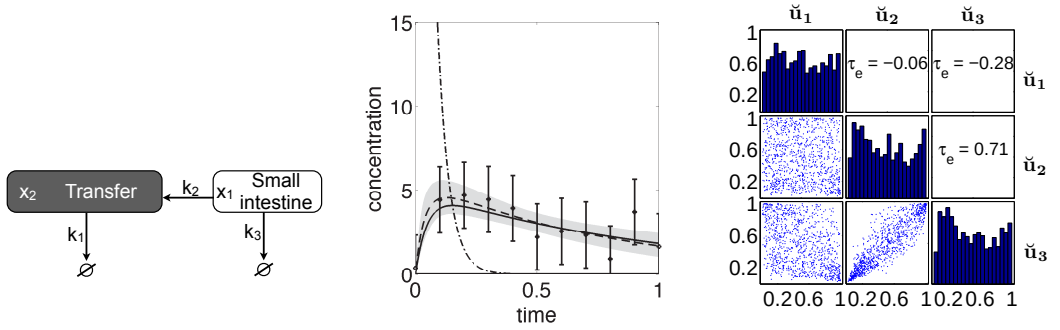
with $x_1(0) = 100$ and $x_2(0) = 0$ in arbitrary units, making our model similar to the ones proposed in Li *et al.* [2011a,b]. Note that the concentrations $x_i(t)$ depend on k_1, k_2, and k_3. However, for readability the dependency on these parameters is omitted. Our data was generated for $k_1 = 1$, $k_2 = 1$, and $k_3 = 20$ at the time points $t_i = 0, 0.1, 0.2, \ldots, 1.0$ as $y_i = x_2(t_i) + \varepsilon_i$ with $\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, 1^2)$ for $i = 1, \ldots, 11$. We assume the initial concentrations $x_1(0) = 100$ and $x_2(0) = 0$ to be known. Under independent prior distributions $k_1 \sim \mathcal{N}_{[0,1000]}(1, 1^2), k_2 \sim \mathcal{N}_{[0,1000]}(1, 1^2)$, and $k_3 \sim \mathcal{N}_{[0,1000]}(20, 20^2)$, where $\mathcal{N}_{[a,b]}(\mu, \sigma^2)$ denotes the $[a, b]$-truncated normal distribution, the posterior is proportional to

$$\pi(k_1 \ldots, k_3|y_1, \ldots, y_{11}) \propto \prod_{i=1}^{11} \Phi(y_i|x_2(t_i), 1^2) \prod_{j=1}^{2} \Phi_{[0,1000]}(k_j|1, 1^2)\Phi_{[0,1000]}(k_3|20, 20^2)$$

for the truncated normal density function $\Phi_{[a,b]}(\cdot|\mu,\sigma^2)$ corresponding to $\mathcal{N}_{[a,b]}(\mu,\sigma^2)$. The data is depicted in Figure 6.8(b).

As independence proposal density q_3 we chose a uniform distribution on $[0,1000]^3$ and set $r_1 = 0.99$ and $r_2 = 0$. There is an interesting dependency structure between the parameters k_1, k_2, and k_3 inherent to the system. While k_2 and k_3 show strong positive but non-symmetric dependency, k_1 is almost independent of k_2 and k_3 (compare Figure 6.8(c) and recall that the dependency structure between the k_i's and the $\breve{u}_i$'s is identical). This can be explained as follows: An increase in k_3 at constant k_2 results in a decrease of the Zr concentration in the "Transfer" compartment. In order to compensate this effect k_2 needs to be increased simultaneously, resulting in a positive dependence between k_2 and k_3. The degradation rate k_1 on the other hand is pairwise almost independent from k_2, and k_3 since it only depends on the concentration $x_2(t)$, which itself depends directly on the data $(y_1,\ldots,y_{11})^\top$.

After uniformization of the prerun samples by fitted normal distributions the pairwise scatterplots in Figure 6.8(c) indicate some lower tail dependence between the parameters $\breve{u}_2$ and $\breve{u}_3$ (corresponding to k_2 and k_3). Using the notation of Chapter 2.2 the proposal copula for the first of the 100 runs was decomposed as $c_{1,2,3}(u_{1,2,3}|\boldsymbol{\theta}) = c_{1,2}(u_{1,2}|\boldsymbol{\theta}) \cdot c_{2,3}(u_{2,3}|\boldsymbol{\theta}) \cdot c_{1,3|2}(F(u_1|u_2,\boldsymbol{\theta}), F(u_3|u_2,\boldsymbol{\theta})|\boldsymbol{\theta})$ with the following estimated pair copulas: (i) $\hat{c}_{1,2}$ a 90° rotated Clayton copula with estimated parameter $\hat{\theta}_{1,2} = -0.19$ and corresponding estimated Kendall's $\hat{\tau} = -0.08$, (ii) $\hat{c}_{2,3}$ a 180° rotated BB6 copula with estimated parameters $\hat{\theta}_{2,3} = (1.24, 3.07)$ and corresponding estimated Kendall's $\hat{\tau} = 0.71$, and (iii) $\hat{c}_{1,3|2}$ a Gaussian copula with estimated parameter $\hat{\theta}_{1,3|2} = -0.79$ and corresponding estimated Kendall's $\hat{\tau} = -0.58$. Note that the estimated Kendall's τ's of the $\hat{c}_{1,2}$ and $\hat{c}_{2,3}$ copula nicely coincided with the estimated Kendall's τ's of the copula sample (Figure 6.8(c)). This indicates a good parameter dependency coverage by the proposal copula, at least on the unconditioned copula decomposition level. All other runs yielded similar decompositions. The order ι was chosen to be the identity function. Lower tail dependence of $\breve{u}_2$ and $\breve{u}_3$ was covered by the rotated 180° BB6 copula $\hat{c}_{2,3}$. Interestingly, the estimated pairwise dependency between k_1 and k_3 ($\tau_e = -0,28$, Figure 6.8(c)) more than doubles when conditioning on k_2, as the estimated Kendall's τ corresponding to $\hat{\theta}_{1,3|2}$ is given by $\hat{\tau} = -0.58$. Loosely speaking Zr is to be excreted within a fixed time period. Moreover, Figure 6.8(c) gives

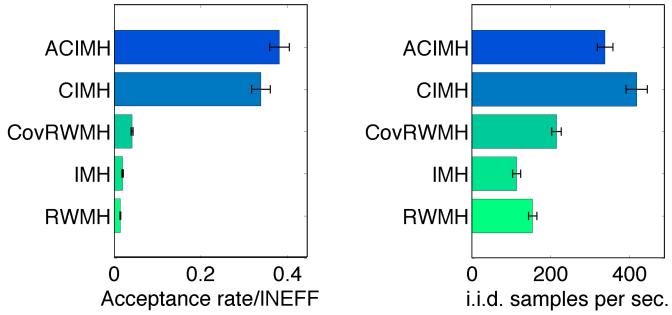


(a) (b) (c)

Figure 6.8: (a) Schematic representation of the small compartment model. The concentration of the shaded "Transfer" compartment is measured at the eleven time points $t_i = 0, 0.1, \ldots, 1.0$. The rates k_1 and k_3 lead to unobserved downstream compartments and are therefore considered as degradation rates. (b) Toy data used for sampling the small compartment model. Depicted are the true underlying concentration $x_2(t)$ of the "Transfer" compartment (solid line), the posterior median solution (dashed line) as well as the according 95% credible interval (shaded area) of the thinned first ACIMH run. The dots depict noisy data y_i including 95% confidence intervals and the unobserved concentration $x_1(t)$ of the "Small intestine" compartment is shown as dashed-dotted line. (c) Copula data (c.f. Section 6.1.1 (ii)) of the first run used to fit the CIMH copula. For uniformization of k_1, k_2, and k_3 the cdf's of $\mathcal{N}(1.33, 0.55^2)$, $\mathcal{N}(29.40, 13.03^2)$, and $\mathcal{N}(1.29, 0.47^2)$ were applied. The diagonal displays the histograms of the MCMC sample marginals and τ_e the respective empirical Kendall's τ.

a hint at the order for arranging the sequence of the copula variables. Here, $\breve{u}_2$ and $\breve{u}_3$ show a significant dependency and were hence modeled as a direct pair within the pair copula decomposition.

The fine-tuning parameter k_{covRW} in CovRWMH could be set to a relatively high value guaranteeing large jumps in the parameter space and therefore low INEFF's (Table 6.3 (Upper table)). CovRWMH hence outperformed RWMH and IMH on both ($\mathcal{J}_1$) and ($\mathcal{J}_2$) as can be seen from Figure 6.9(a) and 6.9(b). The dependency between k_1, k_2, and k_3 is significant enough to cause IMH to perform even worse than RWMH on ($\mathcal{J}_2$). Although taking in average more than 1.3 times as long as any other sampler, the copula based algorithms nicely detected the prerun dependency structure and yielded the best results. The INEFF slightly decreased when updating the copulas. This means that the copula structure is – although only slightly – recursively adjusted to better



(a) (b)

Figure 6.9: Results for the compartment model. Figure (a): Quotient of acceptance rate and INEFF. Figure (b): Number of independent samples drawn per second. Error bars show the estimated standard errors based on 100 runs.

fit the true underlying dependency structure of k_1, k_2, and k_3. However, compared to CIMH the additional time for re-fitting the copula lowered the efficiency with respect to ($\mathcal{I}_2$). For the inference of the marginal MAP estimates, we applied a kernel density estimator to the respective sampled Markov chains. The posterior mean and mode estimates including 90% credible intervals are given in Table 6.3 (Lower table). All predicted modes slightly overestimated the true values $k_1 = k_2 = 1$ and $k_3 = 20$.

6.3.4 Performance on a JAK2-STAT5 signaling pathway model

In this section, we apply our sampling schemes to a DDE model of the JAK2-STAT5 signaling pathway[1]. Here, STAT5 denotes either one of the STAT5A or STAT5B proteins. The system is based on a number of phosphorylation and dephosphorylation steps within a complex protein interaction network. In case of the JAK2-STAT5 pathway the Erythropoietin (Epo) hormone first binds to the transmembrane receptor phosphorylating Janus Kinase 2 (JAK2). Monomeric Signal Transducer and Activator of Transcription 5 (STAT5) is thereafter tyrosine phosphorylated by the JAK2/receptor complex. After a dimerization step the phosphorylated STAT5 homodimer enters the nucleus and binds to the promoter region of its target gene. It dephosphorylates in the

[1]For a thorough introduction to the JAK-STAT pathway see Chapter 2.4.2.

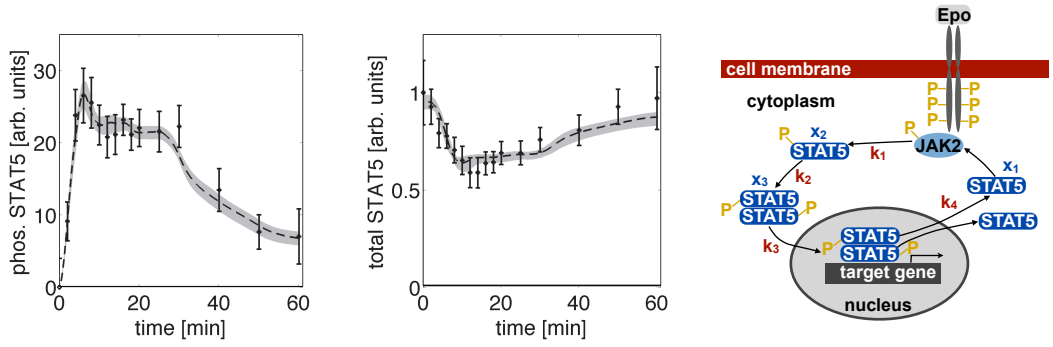
Sampler	i.i.d. samples	INEFF	AR (%)	time (sec.)
ACIMH	25255.7±1509.4	3.9±0.9	75.07±0.12	75.10±0.28
CIMH	22414.5±1434.1	5.1±1.0	75.07±0.11	54.20±0.31
CovMH	8756.8± 491.8	7.9±0.5	23.38±0.06	40.94±0.02
IMH	4532.2± 406.9	23.5±2.9	20.79±0.04	40.14±0.02
RWMH	2963.5± 200.5	27.6±2.4	23.46±0.03	19.33±0.01

Sampler	ACIMH	CIMH	CovRWMH	IMH	RWMH
$\mathbb{E}[k_1\|\boldsymbol{y}]$	1.25	1.24	1.25	1.25	1.25
$M[k_1\|\boldsymbol{y}]$	1.17	1.16	1.17	1.20	1.19
$CI[k_1\|\boldsymbol{y}]$	(0.74;1.79)	(0.74;1.79)	(0.74;1.79)	(0.74;1.78)	(0.74;1.79)
$\mathbb{E}[k_2\|\boldsymbol{y}]$	1.54	1.54	1.54	1.53	1.53
$M[k_2\|\boldsymbol{y}]$	1.33	1.35	1.34	1.31	1.29
$CI[k_2\|\boldsymbol{y}]$	(0.86;2.31)	(0.86;2.31)	(0.86;2.31)	(0.86;2.30)	(0.85;2.30)
$\mathbb{E}[k_3\|\boldsymbol{y}]$	27.30	27.29	27.30	27.24	27.21
$M[k_3\|\boldsymbol{y}]$	23.30	23.73	21.97	22.13	23.87
$CI[k_3\|\boldsymbol{y}]$	(14.09;42.07)	(14.08;42.08)	(14.05;42.10)	(14.24;41.83)	(14.01;41.92)

Table 6.3: Small compartment model. Upper table: Depicted are the average number of i.i.d. samples per run (ESS), INEFF's, acceptance rates (AR), and sampling times based on 100 runs including estimated standard errors. All samplers ran for 50,000 MCMC proposals. Lower table: Estimated marginal posterior means $\mathbb{E}[\cdot|\boldsymbol{y}]$, modes $M[\cdot|\boldsymbol{y}]$ (MAP estimates), and 90% posterior quantile based credible intervals $CI[\cdot|\boldsymbol{y}]$ for k_1, k_2, and k_3 for the concatenated data of 100 runs.

process and gets subsequently exported to the cytoplasm (Aaronson & Horvath [2002] and Hou *et al.* [2002]). A schematic representation can be seen in Figure 6.10(c).

We want to point out that although posterior parameter estimates are given in Table 6.4 the focus of this section lies clearly on the performance evaluation of CIMH and ACIMH on a complex dynamical system, rather than on novel biological insights. Due to the complexity of the system caused by high parameter dependencies MCMC sampling is a daunting task in this scenario. We therefore evaluated the performance of SMALA and M-GaA (introduced in Chapter 5.1 and Chapter 5.3) in addition to the RWMH, IMH, and CovRWMH algorithms.



(a) (b) (c)

Figure 6.10: (a) Time courses for the numerical solution of phosphorylated STAT5 in the cytoplasm ($y_1(t)$). Depicted are the posterior median solution (dashed line) as well as the according 95% credible interval (shaded area) of the thinned first ACIMH run. The dots represent given measurements y_i including 95% confidence intervals. (b) Similarly to (a), the measurements, the median (dashed line), and the 95% credible interval (shaded area) for the numerical solution of $y_2(t)$. (c) Schematic representation of the JAK2-STAT5 pathway: Erythropoietin (Epo) binds to the transmembrane receptor. Monomeric STAT5 (x_1) is tyrosine phosphorylated (x_2) by the activated JAK2/receptor complex in the cytoplasm. After dimerizing the phosphorylated STAT5 homodimer, (x_3) enters the nucleus and binds to the promoter target gene region. It is then dephosphorylated and released to the cytoplasm.

Our analysis is based on the data and mass-action DDE model of Swameye *et al.* [2003]:

$$\begin{aligned}
\frac{\mathrm{d}x_1(t)}{\mathrm{d}t} &= -k_1 x_1(t) Epo(t) + 2k_4 x_3(t+\tau) \\
\frac{\mathrm{d}x_2(t)}{\mathrm{d}t} &= -k_2 x_2^2(t) + k_1 x_1(t) Epo(t) \\
\frac{\mathrm{d}x_3(t)}{\mathrm{d}t} &= -k_3 x_3(t) + \frac{1}{2} k_2 x_2^2(t) \\
\frac{\mathrm{d}x_4(t)}{\mathrm{d}t} &= -k_4 x_3(t+\tau) + k_3 x_3(t),
\end{aligned} \tag{6.11}$$

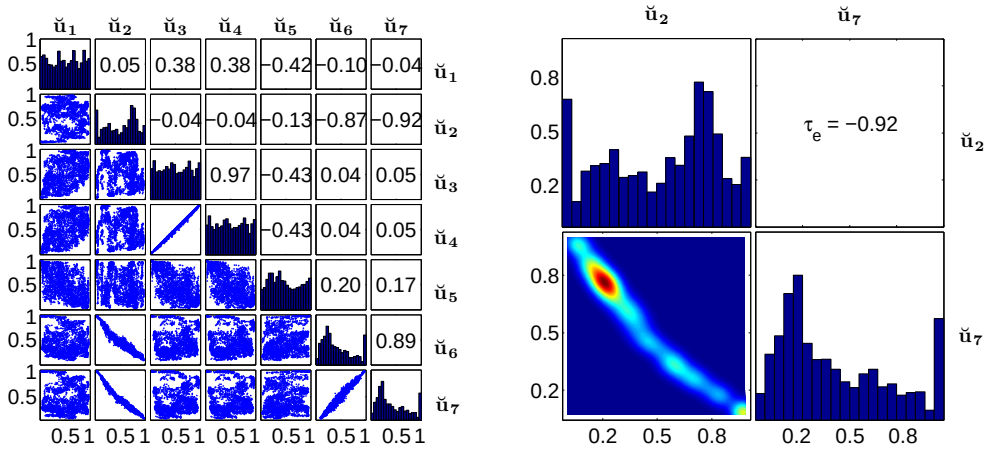
with $x_1(0) = 1$ and $x_2(0) = x_3(0) = x_4(0) = 0$, where $Epo(t)$ denotes the time-dependent Epo stimulation function and τ the time lag between STAT5 entering the nucleus and dephosphorylated cytoplasmic release. Furthermore, $x_1(t)$, $x_2(t)$, $x_3(t)$ are the concentrations of unphosphorlated, tyrosine-phosphorylated, and dimerized STAT5, respectively, while $x_4(t)$ is the concentration of STAT5 in the nucleus. Note that the system inherits a dimerization step as introduced in Example 2.10. Due to the law of mass conservation, we need to claim $k_3 \geq k_4$. The data we used for inference was provided

by J.Timmer at http://webber.physik.uni-freiburg.de/∼jeti/PNAS_Swameye_Data/. It contains (including 95% confidence intervals) the amount of phosphorylated STAT5, $y_1^\varepsilon(t_i) = k_5(x_2(t_i) + 2x_3(t_i) + \varepsilon_1(t_i))$ and the total concentration of cytoplasmic STAT5, $y_2^\varepsilon(t_i) = k_6(x_1(t_i) + x_2(t_i) + 2x_3(t_i) + \varepsilon_2(t_i))$, at 16 time points $t_1, \ldots, t_{16}$ (in minutes) in the interval $[0, 60]$. Here, k_5 and k_6 are introduced since all measurements are relative. The errors $\varepsilon_j(t_i)$ are measurement errors included in the data, which are assumed to be $\mathcal{N}(0, \sigma_{i,j}^2)$ distributed where $\sigma_{i,j}^2$ was estimated from various experiments. We performed a Kolmogorov-Smirnov test (Davison [2003]) on normality to ensure that the combined measurement/model error of the likelihood can not possibly be non-Gaussian. For this, the residuals between the data points and a simulated annealing MLE based time course were considered. The null-hypothesis of non-Gaussian noise could not be rejected on an $\alpha = 5\%$ significance level. All seven parameters $\boldsymbol{\xi} = (k_1, k_2, k_3, k_4, \tau, k_5, k_6)^\top$ are time-independent. Again, for readability the dependence of the solutions $x_i(t)$ to (6.11) on $\boldsymbol{\xi}$ is omitted. A picture of the data can be seen in Figure 6.10(a) and 6.10(b). Similarly to Swameye *et al.* [2003] we reparametrized the DDE system (see Appendix D) in order to resolve structural parameter identifiability issues. A discussion on the structural parameter identifiability issues of the particular system can be found in Timmer *et al.* [2004] and Raue *et al.* [2009]. Due to the lack of knowledge we chose the independent prior distributions $k_1, k_2, k_4, \tau, k_5, k_6 \overset{i.i.d.}{\sim} \mathcal{U}[0, 50]$ and $k_3 \sim \mathcal{U}[k_4, 50]$. The lower limit 0 was canonically introduced by the non-negativity constraint for reaction rates. For $\boldsymbol{y} := \{y_1^\varepsilon(t_1), \ldots, y_1^\varepsilon(t_{16}), y_2^\varepsilon(t_1), \ldots, y_2^\varepsilon(t_{16})\}$ and the prior $\pi(\boldsymbol{\xi})$ this leads to the posterior

$$\pi(\boldsymbol{\xi}|\boldsymbol{y}) \propto \prod_{i=1}^{16} \Phi(y_1^\varepsilon(t_i)|y_1(t_i), \sigma_{i,1}^2) \cdot \Phi(y_2^\varepsilon(t_i)|y_2(t_i), \sigma_{i,2}^2) \cdot \pi(\boldsymbol{\xi}).$$

Here, $y_1(t) = k_5(x_2(t) + 2x_3(t))$ and $y_2(t) = k_6(x_1(t) + x_2(t) + 2x_3(t))$ for the solutions $x_1(t)$, $x_2(t)$, and $x_3(t)$ of the DDE. Since there is no analytical solution to (6.11), we applied Matlab's `dde23` solver to numerically derive $x_i(t)$ $(i = 1, 2, 3)$ in case of the RWMH, IMH, CovRWMH, M-GaA, CIMH, and ACIMH algorithms. SMALA used Matlab's `ode15s` solver for the geometric tensor derived in Appendix E. As `dde23` and `ode15s` are quite time consuming, generating good proposals is essential for efficient sampling from the highly dependent seven dimensional parameter distribution.

We started the inference by choosing the independence proposal density q_3 to be uniform on $[0, 50]^7$ and setting $r_1 = 0.7$ and $r_2 = 0.25$. The outcome of a simulated annealing run was taken as starting value for RWMH, making a correction for a burn-in phase obsolete. A look at the copula data from Figure 6.11(a) reveals that fitting standard pair copulas to the data – at least on the unconditioned level – is rather involved: the density plot of the $(\breve{u}_2, \breve{u}_7)$-pair of the first run (Figure 6.11(b)), for instance, has a non-standard bent ridge shape with a very dense region at high $\breve{u}_7$ and low $\breve{u}_2$ values. The fitting issue results in rather low acceptance rates for the copula based algorithms (Table 6.5). Nevertheless, both copula algorithms had again comparatively better INEFF's and generated far more independent samples than RWMH, IMH, CovRWMH, M-GaA, or SMALA (Figure 6.12(a) and 6.12(b) and Table 6.5). Except for the M-GaA algorithm, ACIMH outperformed all non-copula based sampling schemes more than 2.5-fold with respect to $(\mathcal{I}_2)$. The prerun samples were transformed to $[0, 1]^7$ using fitted normal densities for the margins of k_1, k_2, k_3, k_4, τ, and k_5 and a fitted lognormal density for the margin of k_6. Owing to the complexity of the system, we used 3,000 samples to fit all copulas involved. By sequential adjustment of the proposal function during the sampling process ACIMH could increase $(\mathcal{I}_1)$ and $(\mathcal{I}_2)$ compared to CIMH. The average number of pair copula family updates in every ACIMH run was 48%, i.e. almost every second pair copula was fitted to have different copula types compared to the fit before. The order of the variables ι was chosen to best capture strong pairwise dependencies in the samples. More precisely we set $\iota(1) = 3, \iota(2) = 1, \iota(3) = 2, \iota(4) = 4, \iota(5) = 5, \iota(6) = 6, \iota(7) = 7$. SMALA had in average rather high INEFF's and was even more time consuming than the copula based samplers, i.e. the time needed for the computation of the geometric tensor could not be compensated by vast traversals through the parameter space. On the other hand, the second order moments based CovRWMH and M-GaA algorithms yielded already good sampling performances with respect to $(\mathcal{I}_2)$ as Figure 6.12(b) shows. Clearly, since the covariance matrix $\hat{\boldsymbol{C}}$ is based on a total of 3,000 prerun samples it covers the second order moment of the posterior in average better than the initial covariance matrices of the M-GaA proposal function. CovRWMH therefore outperforms M-GaA with respect to $(\mathcal{I}_1)$. Conversely, the prerun-time needed to tune $\hat{\boldsymbol{C}}$ slows down CovRWMH and gives M-GaA an advantage with respect to $(\mathcal{I}_2)$. As mentioned earlier the actual acceptance rate of M-GaA drops to $14.86\% \pm 4.79\%$ (Table 6.5) and failed to meet the predefined acceptance rate of



(a) (b)

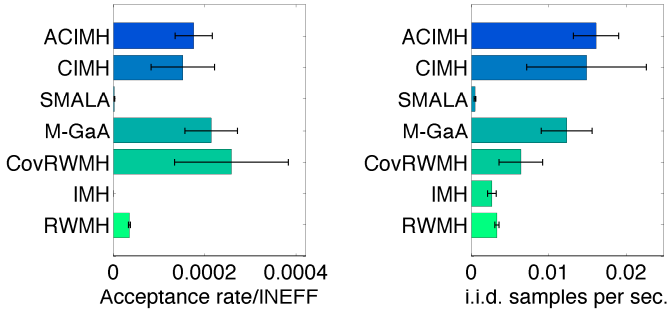
Figure 6.11: Copula data (c.f. Section 6.1.1 (ii)) of the first run used to fit the CIMH and initial ACIMH copula on 3,000 of the 50,000 MCMC iterations. For uniformization normal distributions were chosen for k_1, k_2, k_3, k_4, τ, and k_5 and a lognormal distribution for k_6. The diagonal displays the histograms of the sample marginals and the numbers in the upper right triangle the estimated Kendall's τ's. (b) Density plot of the $(\breve{u}_2, \breve{u}_7)$ copula data pair corresponding to (k_2, k_6) of the first run. Red areas depict higher, blue areas lower density values.

$\alpha_0 = 23\%$. It is nevertheless higher than the ones of IMH, SMALA, CIMH, or ACIMH. In summary, although CIMH and ACIMH somewhat struggle to cover the complex posterior distribution, which is indicated by the low $(\mathcal{I}_1)$ index (Figure 6.12(a)), they nevertheless yield a superior sampling performance with respect to $(\mathcal{I}_2)$ compared to all other algorithms considered (Figure 6.12(b)).

Our Bayesian MCMC approach revealed a strong indeterminacy with respect to the parameters of the system. There e.g. exist strong dependencies between k_2 and k_6 and k_5 and k_6 (compare pairs $(\breve{u}_2, \breve{u}_7)$ and $(\breve{u}_6, \breve{u}_7)$ of Figure 6.11(a)). Table 6.4 shows the marginal posterior means, modes (MAP estimates), and 90% posterior quantile based credible intervals for the concatenated data of 10 thinned runs for the respective algorithms RWMH, IMH, CovRWMH, CIMH, and ACIMH. The MAP estimates of the time τ a STAT5 molecule remains in the nucleus is $\approx$ 4 minutes. This means that the cytoplasmic release turns out to be a bit faster than the value of $\approx$ 6.4 minutes

Sampler	ACIMH	CIMH	SMALA	M-GaA	CovRWMH	IMH	RWMH
$\mathbb{E}[k_1\|\boldsymbol{y}]$	0.03	0.03	0.03	0.03	0.03	0.03	0.03
$M[k_1\|\boldsymbol{y}]$	0.03	0.03	0.04	0.03	0.03	0.03	0.03
$CI[k_1\|\boldsymbol{y}]$	(0.03;0.04)	(0.03;0.04)	(0.02; 0.05)	(0.03; 0.04)	(0.03;0.04)	(0.02;0.04)	(0.02;0.03)
$\mathbb{E}[k_2\|\boldsymbol{y}]$	2.46	3.11	2.37	1.03	2.41	2.81	2.56
$M[k_2\|\boldsymbol{y}]$	1.31	2.14	2.36	0.86	2.16	2.02	2.15
$CI[k_2\|\boldsymbol{y}]$	(1.11; 4.43)	(1.32; 6.12)	(2.33; 2.42)	(0.76; 1.34)	(1.38;3.57)	(1.76;4.10)	(1.52;3.73)
$\mathbb{E}[k_3\|\boldsymbol{y}]$	0.17	0.16	0.11	0.24	0.17	0.15	0.15
$M[k_3\|\boldsymbol{y}]$	0.15	0.14	0.11	0.20	0.15	0.15	0.14
$CI[k_3\|\boldsymbol{y}]$	(0.13;0.23)	(0.13;0.20)	(0.10; 0.11)	(0.16; 0.34)	(0.13;0.22)	(0.12;0.19)	(0.11;0.19)
$\mathbb{E}[k_4\|\boldsymbol{y}]$	0.17	0.16	0.11	0.24	0.17	0.15	0.15
$M[k_4\|\boldsymbol{y}]$	0.15	0.14	0.11	0.21	0.15	0.13	0.14
$CI[k_4\|\boldsymbol{y}]$	(0.13;0.23)	(0.13;0.20)	(0.10; 0.11)	(0.16; 0.34)	(0.12;0.22)	(0.12;0.18)	(0.11;0.18)
$\mathbb{E}[\tau\|\boldsymbol{y}]$	3.97	4.18	5.56	3.61	4.06	4.31	4.52
$M[\tau\|\boldsymbol{y}]$	3.82	4.23	5.59	3.79	3.86	4.28	4.03
$CI[\tau\|\boldsymbol{y}]$	(3.12; 4.86)	(3.43; 4.94)	(5.15; 6.08)	(2.83; 4.36)	(3.08;5.02)	(3.51;5.25)	(3.47;6.40)
$\mathbb{E}[k_5\|\boldsymbol{y}]$	35.90	36.06	34.11	35.93	36.10	35.44	36.08
$M[k_5\|\boldsymbol{y}]$	35.63	36.02	33.82	35.29	36.57	35.43	36.78
$CI[k_5\|\boldsymbol{y}]$	(33.90; 37.93)	(34.19; 37.97)	(33.32; 35.26)	(33.71; 38.27)	(34.24;37.93)	(32.42;38.36)	(34.08;37.68)
$\mathbb{E}[k_6\|\boldsymbol{y}]$	0.95	0.95	0.40	0.0.94	0.95	0.94	0.96
$M[k_6\|\boldsymbol{y}]$	0.95	0.95	0.36	0.93	0.96	0.94	0.94
$CI[k_6\|\boldsymbol{y}]$	(0.92;0.98)	(0.92;0.98)	(0.34; 0.46)	(0.91; 0.97)	(0.92;0.98)	(0.91;0.99)	(0.92;1.00)

Table 6.4: JAK2-STAT5 pathway model. Estimated marginal posterior means $\mathbb{E}[\cdot|\boldsymbol{y}]$, modes $M[\cdot|\boldsymbol{y}]$ (MAP estimates), and 90% posterior quantile based credible intervals $CI[\cdot|\boldsymbol{y}]$ for the parameters $k_1, k_2, k_3, k_4, \tau, k_5$ and k_6 for the concatenated data of 10 runs.



(a) (b)

Figure 6.12: Results for the JAK2-STAT5 model. Figure (a): Quotient of acceptance rate and INEFF. Figure (b): Number of independent samples drawn per second. Error bars show the estimated standard errors based on 10 runs.

Sampler	i.i.d. samp.	INEFF	AR (%)	time (sec.)
ACIMH	105.9± 19.5	658.4± 128.8	8.38± 1.38	6635.7± 197.7
CIMH	93.0± 48.1	1859.1± 591.1	8.03± 1.17	6339.7± 195.8
SMALA	47.0± 26.9	3242.1± 968.4	13.92± 4.96	7175.3±2211.3
M-GaA	40.3± 19.1	2922.3±1023.4	14.86± 4.79	7041.9±2346.3
CovRWMH	35.7± 15.6	3573.4± 887.0	29.04± 3.40	5574.8± 34.8
IMH	10.9± 2.4	5366.2± 597.6	$0.04\pm4\cdot10^{-3}$	4105.8± 40.3
RWMH	7.6± 0.6	6406.9± 463.3	21.47± 0.60	2328.9± 26.8

Table 6.5: JAK2-STAT5 pathway model. Depicted are the average number of i.i.d. samples per run (ESS), INEFF's, acceptance rates (AR), and sampling times based on 10 runs including estimated standard errors. All samplers ran for 50,000 MCMC proposals.

computed by Swameye *et al.* [2003]. Nevertheless $\tau \approx 4$ minutes is contained in their confidence interval of $(3.8, 6.9)$ minutes. All other results coincide well. Overall, the system represents a very challenging example for MH sampling schemes and is thus a good benchmark for performance evaluation.

6.3.5 Robustness with respect to the choice of the pair copula decomposition and cdf's for prerun sample transformation

Two very crucial factors for the performance of CIMH and ACIMH are the goodness-of-fit of (i) the pair copula decomposition and (ii) the cdf's for the transformation

of the prerun samples. We already inferred the effect of applying oversimplified copula decompositions: as the independence copula is defined by $C : [0,1]^n \longrightarrow [0,1]$, $(u_1, \ldots, u_n) \mapsto \prod_{i=1}^{n} u_i$, the IMH is essentially a CIMH algorithm with an independence copula based proposal function. The JAK2-STAT5 example showed that we can run into serious problems to fit an appropriate decomposition. More involved techniques such as fitting mixtures of pair copulas as well as non or semi parametric copula density estimation and sample generation (Hu [2006]) are therefore needed in future applications. As Kim *et al.* [2007] showed, the latter can in some cases considerably improve the robustness against misspecification of the marginal distribution types for $G_i(\xi|\boldsymbol{\gamma}_i)$. To assess the misspecification effect, we resampled the example of Chapter 6.3.2 over 10 runs using fitted lognormal distribution functions on the one hand and fitted normal distribution functions on the other hand. The amount of independent samples per second dramatically dropped by a factor of 9 when applying lognormal distributions. The index ($\jmath_1$) even decreased by a factor of 45. The issue of modeling the dependency structure therefore needs to be considered carefully. As mentioned above, ($\jmath_1$) can be taken as index for the goodness-of-fit of the copula proposal function. A value close to 1 guarantees an efficient sampling performance.

6.4 Conclusions on CIMH and ACIMH

In summary, we saw that the vine copula based sampling schemes CIMH and ACIMH showed superior performance compared to RWMH, IMH, and CovRWMH in every example. They outperformed the MCMC algorithms SMALA and M-GaA on complex systems, such as the JAK2-STAT5 pathway. Especially in the first three examples the copula based approaches covered the dependency structure of the posterior very well, which resulted in high sampling efficiencies. Avoiding computationally costly copula updates CIMH is doing best on these simple systems. However, in the complex JAK2-STAT5 case these updates could improve the performance of ACIMH by fine-tuning the transition function. The copula data of the JAK2-STAT5 posterior indicated that non-standard distributions for marginalization might improve the sampling efficiency even more; a topic that should be addressed in further research. Nevertheless, CIMH and ACIMH are yet promising concepts for the inference of dynamical systems.

7

Model inference of the JAK1-STAT3 pathway

We already used a mathematical model of the JAK2-STAT5 pathway in Chapter 6.3.4 to show that ACIMH is capable of efficiently inferring complex dynamical systems of this type. In the current chapter we focus on the question, whether tyrosine-phosphorylated STAT3 dimers can directly work as transcription factors in the JAK1-STAT3 pathway[1]? From a biological point of view this is of relevance since IL-6 stimulation controls the spreading of microbes in inflamed cells and endorses cell regeneration after injury as could be shown in mouse hepatocytes (Bonizzi & Karin [2004]).

Similarly to the JAK2-STAT5 case, STAT3 is activated, i.e. phosphorylated, by JAK1 upon association with the IL-6 stimulated glycoprotein 130 (gp130) transmembrane receptor. To reach full transcription factor activity the STAT3 dimer can also be serine phosphorylated (see Wen *et al.* [1995]). The latter is however not necessary for the STAT3 dimer to work as transcription factor in the nucleus. In the following we therefore infer the effect of tyrosine-phosphorylated STAT3 dimer transcription factors.

[1] For a thorough introduction to the JAK-STAT pathway see Chapter 2.4.2.

7.1 Experimental JAK1-STAT3 data

Our data is based on primary mouse hepatocytes (liver cells) stimulated with 1nm IL-6. Measured were the protein concentrations of total cytoplasmic STAT3, tyrosine-phosphorylated gp130, tyrosine phosphorylated STAT3, as well as tyrosine-serine-phosphorylated STAT3 dimers on a 90 minute time scale using quantitative immunoblotting as described in Schilling *et al.* [2005] and Bohl [2009]. The data was kindly provided by Prof. Dr. U. Klingmüller from the Deutsches Krebsforschungszentrum (DKFZ) in Heidelberg, Germany. Unfortunately, the strength of measurement errors is unknown. According to Bohl [2009], phosphorylated gp130 can be directly identified with phosphorylated JAK1 proteins. All measurements were normalized using calibrator or normalizer proteins and therefore contain arbitrary units.

7.2 Mathematical models for the JAK1-STAT3 pathway

For inference of the JAK1-STAT3 pathway we adapted the JAK2-STAT5 mass action DDE model of Chapter 6.3.4 by including an additional serine phosphorylation step of the STAT3 dimer. Applying the linear chain trick to convert the defining DDE system into an ODE system (Appendix E, Equation (E.3)), the model reads

$$\begin{aligned}
\frac{\mathrm{d}x_1(t)}{\mathrm{d}t} &= -k_1 x_1(t) pgp130(t) + 2k_5 x_7(t) \\
\frac{\mathrm{d}x_2(t)}{\mathrm{d}t} &= -k_2 x_2^2(t) + k_1 x_1(t) pgp130(t) \\
\frac{\mathrm{d}x_3(t)}{\mathrm{d}t} &= -k_3 x_3(t) + \frac{1}{2} k_2 x_2^2(t) \\
\frac{\mathrm{d}x_4(t)}{\mathrm{d}t} &= -k_4 x_4(t) + k_3 x_3(t) \\
\frac{\mathrm{d}x_5(t)}{\mathrm{d}t} &= -k_5 x_7(t) + k_4 x_4(t) \\
\frac{\mathrm{d}x_6(t)}{\mathrm{d}t} &= \frac{2}{\tau}(x_4(t) - x_6(t)) \\
\frac{\mathrm{d}x_7(t)}{\mathrm{d}t} &= \frac{2}{\tau}(x_6(t) - x_7(t)),
\end{aligned} \tag{7.1}$$

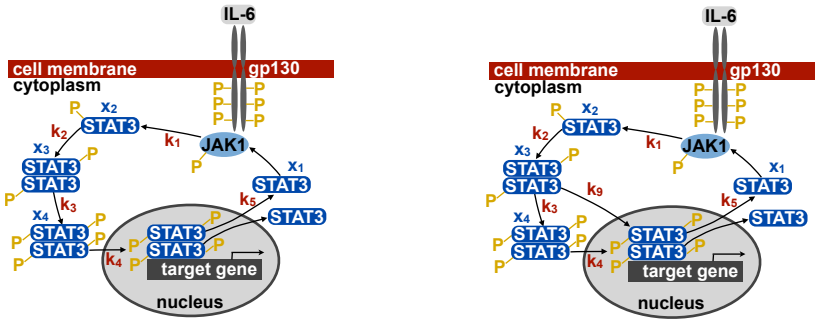
where $x_1(0) = 1$ and $x_2(0) \ldots = x_7(0) = 0$ in arbitrary units. The concentrations $x_5(t)$, $x_6(t)$, $x_7(t)$ are auxiliary and represent the concentration of STAT3 in the nu-

cleus. However, with respect to inference they are of no further interest. $pgp130(t)$ denotes the time-dependent tyrosine-phosphorylated gp130 stimulation function and τ the time lag between STAT3 entering the nucleus and dephosphorylated cytoplasmic release. Furthermore, $x_1(t)$, $x_2(t)$, $x_3(t)$, $x_4(t)$ are the concentrations of unphosphorylated, tyrosine-phosphorylated, tyrosine-phosphorylated dimerized, and tyrosine-serine-phosphorylated dimerized STAT3, respectively. The law of mass conservation again claims $k_2 \geq k_3$ and $k_4 \geq k_5$. A schematic representation of the model can be seen in Figure 7.1(a). We are given the following observations:

$$
\begin{aligned}
y_1^\varepsilon(t_i) &= k_6(x_2(t_i) + 2x_3(t_i) + \varepsilon_1(t_i)), \text{ the amount of tyrosine-phosphorylated STAT3} \\
y_2^\varepsilon(t_i) &= k_7(2x_4(t_i) + \varepsilon_2(t_i)), \text{ the amount of tyrosine-serine-phosphorylated STAT3} \\
y_3^\varepsilon(t_i) &= k_8(x_2(t_i) + 2x_3(t_i) + 2x_4(t_i) + \varepsilon_3(t_i)), \text{ the total cytoplasmic STAT3,}
\end{aligned}
$$

at 30 time points $t_1, \ldots, t_{30}$ (in minutes) in the interval $[0, 90]$. Here, k_6, k_7 and k_8 are introduced due to the relativity of the measurements. It has to be noted that k_1 cannot be inferred without any further knowledge about the gp130 measurements as it includes a relativity term of the normalized gp130 measurements. The quantities $\varepsilon_j(t_i)$ denote normally distributed measurement errors included in the data. We assume that all tyrosine-phosphorylated STAT3 molecules are converted into tyrosine-serine-phosphorylated STAT3 (Bohl [2009]). Therefore, k_7 can be approximated via k_6 times the estimated mean quotient r of $y_1^\varepsilon(t_i)$ and $y_2^\varepsilon(t_i)$ which eliminates one parameter. The remaining eight parameters $\boldsymbol{\xi} = (k_1, k_2, k_3, k_4, k_5, \tau, k_6, k_8)^\top$ are time-independent. Again, for readability we omitted the dependence of the solutions $x_i(t)$ to (7.1) on $\boldsymbol{\xi}$. Since the JAK2-STAT5 pathway is directly comparable to the JAK1-STAT3 pathway (Aaronson & Horvath [2002]), we used the rates given in Timmer *et al.* [2004] as prior information, i.e. we chose $k_1 \sim \mathcal{N}_0(0.021, (\frac{0.021}{2})^2)$, $k_2 \sim \mathcal{N}_{k_3}(2.46, (\frac{2.46}{2})^2)$, $k_4 \sim \mathcal{N}_{k_5}(0.107, (\frac{0.107}{2})^2)$, $k_5 \sim \mathcal{N}_0(0.107, (\frac{0.107}{2})^2)$ and $\tau \sim \mathcal{N}_0(6.4, (\frac{6.4}{2})^2)$, where $\mathcal{N}_a(\cdot, \cdot)$ denotes the a left-truncated univariate normal distribution. The rates k_3, k_6, and k_8 were chosen to be uniform on the interval $[0, 1000]$. Here, the lower limit 0 was canonically introduced by the non-negativity constraint for reaction rates. For $\boldsymbol{y} := \{y_1^\varepsilon(t_1), \ldots, y_1^\varepsilon(t_{30}), y_2^\varepsilon(t_1), \ldots, y_2^\varepsilon(t_{30}), y_3^\varepsilon(t_1), \ldots, y_3^\varepsilon(t_{30})\}$ and the prior $\pi(\boldsymbol{\xi})$ this leads to the posterior distribution

$$
\pi(\boldsymbol{\xi}|\boldsymbol{y}) \propto \prod_{i=1}^{30} \Phi(y_1^\varepsilon(t_i)|y_1(t_i), \sigma_1^2) \cdot \Phi(y_2^\varepsilon(t_i)|y_2(t_i), \sigma_2^2) \cdot \Phi(y_3^\varepsilon(t_i)|y_3(t_i), \sigma_3^2) \cdot \pi(\boldsymbol{\xi}),
$$



(a) (b)

Figure 7.1: (a) Schematic representation of the first JAK1-STAT3 pathway model (7.1): Interleukin 6 (IL-6) binds to the gp130 transmembrane receptor. Monomeric STAT3 (x_1) is tyrosine phosphorylated (x_2) by the activated JAK1/gp130 complex in the cytoplasm. After dimerizing the tyrosine-phosphorylated STAT3 dimer (x_3) serine phosphorylates (x_4), gets transported to the nucleus and binds to the promoter target gene region. Subsequently it is dephosphorylated and released back into the cytoplasm. (b) Graphical representation of the alternative JAK1-STAT3 pathway model (7.2): The model contains an additional transfer of tyrosine-phosphorylated STAT3 dimers (x_3) into the nucleus.

where $y_1(t) = k_6(x_2(t) + 2x_3(t))$, $y_2(t) = k_6 r(2x_4(t))$ and $y_3(t) = k_8(x_1(t) + x_2(t) + 2x_3(t) + 2x_4(t))$ for the solutions $x_1(t)$, $x_2(t)$, $x_3(t)$, and $x_4(t)$ of (7.1). The standard deviations σ_1, σ_2, and σ_3 of the measurement errors were inferred by simulated annealing (Chapter 4.6) before starting posterior inference. Like in the JAK2-STAT5 pathway we assume normally distributed measurement/model errors for the likelihood. Again, we performed a Kolmogorov-Smirnov test on normality to ensure that the combined measurement/model error can not possibly be non-Gaussian. The null-hypothesis of non-Gaussian noise could not be rejected on an $\alpha = 5\%$ significance level for tyrosine-phosphorylated STAT3, tyrosine-serine-phosphorylated STAT3, or the total cytoplasmic STAT3. The test was based on the corresponding time course of the MLE of model (7.1). The MLE was obtained by simulated annealing.

As mentioned above it is biologically unclear whether the tyrosine-phosphorylated STAT3 dimer has a strong potential to directly work as transcription factor. We therefore compared our first model (7.1) to a model including an additional direct $x_3(t)$

nucleus import model: once again applying the linear chain trick we have

$$\begin{aligned}
\frac{\mathrm{d}x_1(t)}{\mathrm{d}t} &= -k_1 x_1(t) pgp130(t) + 2k_5 x_8(t) + 2k_5 x_{10}(t) \\
\frac{\mathrm{d}x_2(t)}{\mathrm{d}t} &= -k_2 x_2^2(t) + k_1 x_1(t) pgp130(t) \\
\frac{\mathrm{d}x_3(t)}{\mathrm{d}t} &= -k_3 x_3(t) - k_9 x_3(t) + \frac{1}{2} k_2 x_2^2(t) \\
\frac{\mathrm{d}x_4(t)}{\mathrm{d}t} &= -k_4 x_4(t) + k_3 x_3(t) \\
\frac{\mathrm{d}x_5(t)}{\mathrm{d}t} &= -k_5 x_8(t) + k_4 x_4(t) \\
\frac{\mathrm{d}x_6(t)}{\mathrm{d}t} &= -k_5 x_{10}(t) + k_9 x_3(t) \\
\frac{\mathrm{d}x_7(t)}{\mathrm{d}t} &= \frac{2}{\tau}(x_5(t) - x_7(t)) \\
\frac{\mathrm{d}x_8(t)}{\mathrm{d}t} &= \frac{2}{\tau}(x_7(t) - x_8(t)) \\
\frac{\mathrm{d}x_9(t)}{\mathrm{d}t} &= \frac{2}{\tau'}(x_6(t) - x_9(t)) \\
\frac{\mathrm{d}x_{10}(t)}{\mathrm{d}t} &= \frac{2}{\tau'}(x_9(t) - x_{10}(t)),
\end{aligned} \tag{7.2}$$

with $x_1(0) = 1$ and $x_2(0) \ldots = x_{10}(0) = 0$ in arbitrary units and time delays τ, τ' for the nuclear time spent by the tyrosine-serine-phosphorylated STAT3 dimer and the tyrosine-phosphorylated STAT3 dimer, respectively. Here, $x_5(t), \ldots, x_{10}(t)$ are auxiliary. We claim $k_2 \geq k_3$, and $k_4 + k_9 \geq k_5$. The posterior distribution remains almost unchanged, except that the defining parameter vector of the solutions $x_i(t)$ of (7.2) is given by $\boldsymbol{\xi} = (k_1, k_2, k_3, k_4, k_5, \tau, \tau', k_6, k_8, k_9)^\top$ and the prior distribution of model one is extended by $\tau' \sim \mathcal{N}_0(6.4, (\frac{6.4}{2})^2)$ and $k_9 \sim \mathcal{N}_{k_5 - k_4}(0.107, (\frac{0.107}{2})^2)$ for the additional parameter variables. Due to a different STAT3 dimer transcription activity (Wen *et al.* [1995]), we allow $\tau \neq \tau'$. All rates shared by both models are governed by the very same prior distributions. This means the choice of priors does not influence the model inference process. A schematic representation of (7.2) is depicted in Figure 7.1(b).

7.3 Inference of the JAK1-STAT3 model

We now want to analyze the effect of direct tyrosine-phosphorylated STAT3 dimer import into the nucleus, i.e. we compare the models (7.1) and (7.2) using thermodynamic

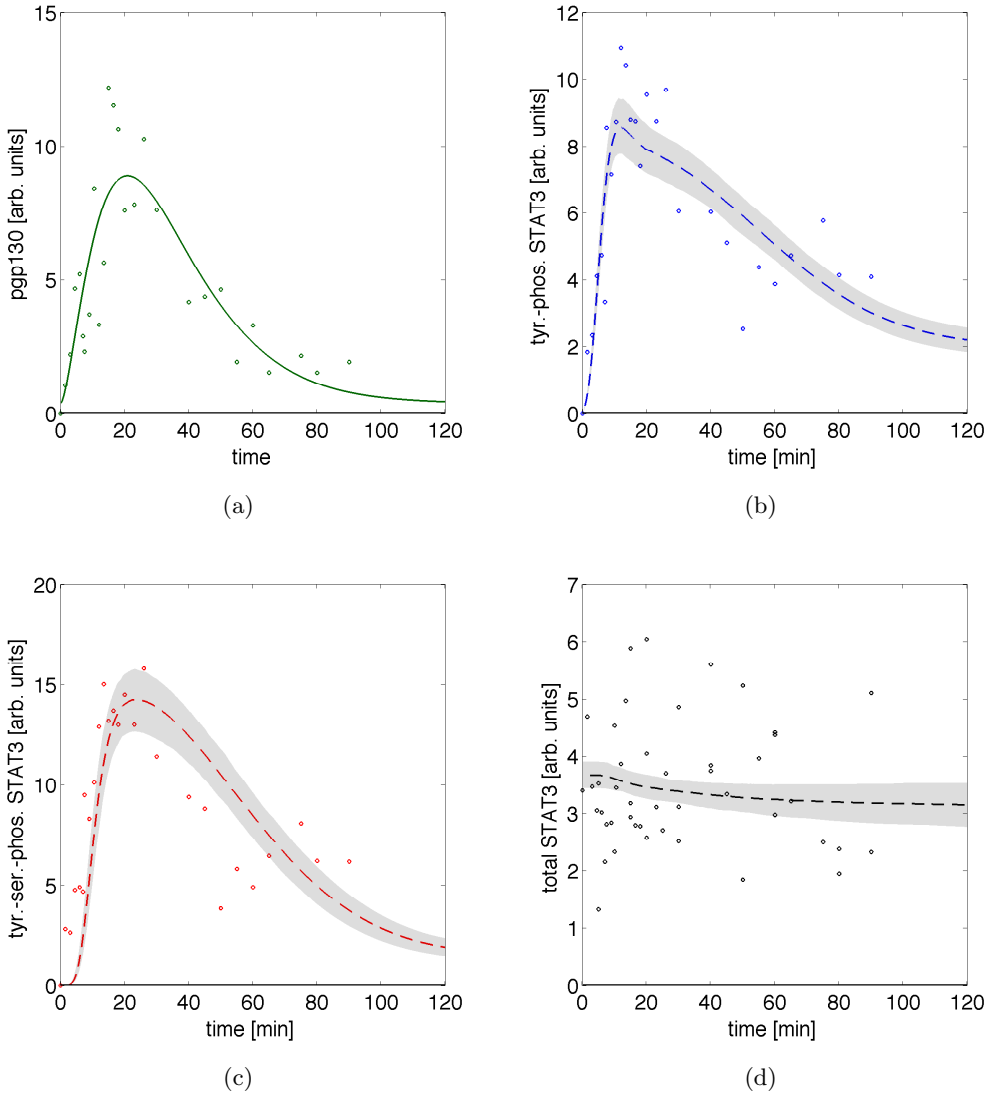


Figure 7.2: JAK1-STAT3 data. (a) *pgp*130 data (green dots) and according fitted *pgp*130 stimulating function (solid green line). (b) Time courses for the numerical solution of tyrosine-phosphorylated STAT3 in the cytoplasm ($y_1(t)$). Depicted are the posterior median solution (dashed blue line) as well as the according 95% credible interval (shaded area) of the thinned $T_{30} = 1$ ACIMH run for model (7.1). Blues dots represent given measurements $y_1^\varepsilon(t_i)$. (c) Similarly to (b), the measurements (red dots), the posterior median solution (dashed red line), and the 95% credible interval (shaded area) for the numerical solution of $y_2(t)$, i.e. tyrosine-serine-phosphorylated STAT3 in the cytoplasm. (d) Similarly to (b), the measurements (black dots), the posterior median solution (dashed black line), and the 95% credible interval (shaded area) for the numerical solution of $y_3(t)$, i.e. total STAT3 in the cytoplasm.

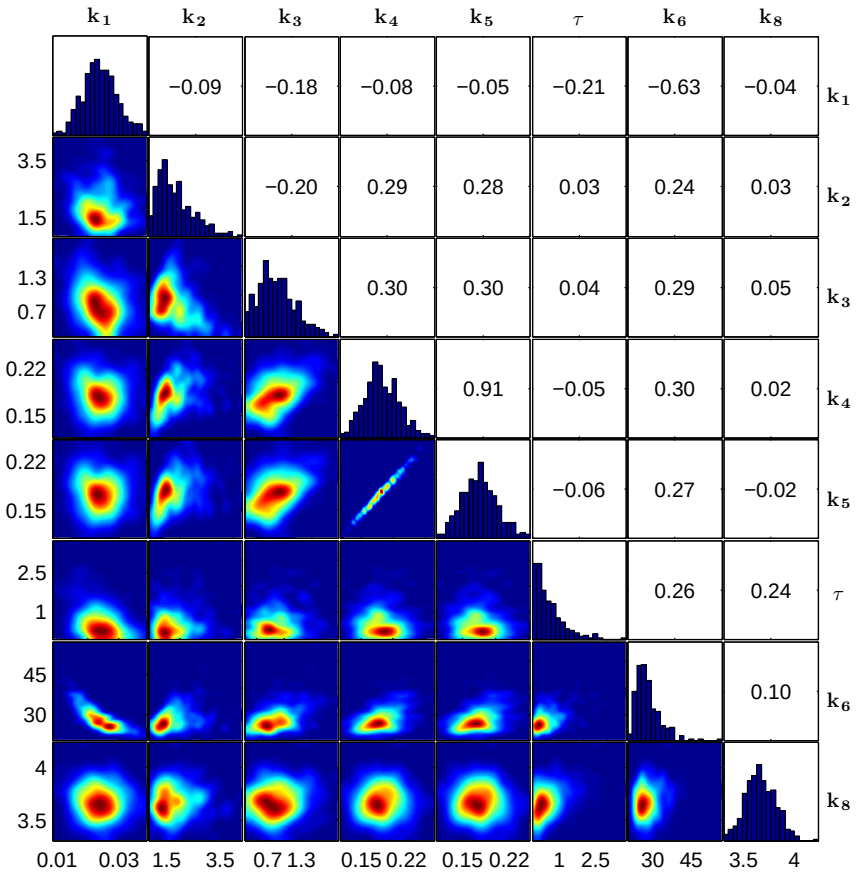


Figure 7.3: JAK1-STAT3 pathway model. Pairwise density plots for all parameter-pairs of the posterior. Red areas depict higher, blue areas lower density values. The diagonal displays the histograms of the sample marginals and the numbers in the upper right triangle the estimated Kendall's τ's.

integration (Chapter 3.4.3). For inference we chose the independence proposal density q_3 of ACIMH to be uniform on $[0, 1000]^8$ for model (7.1) and uniform on $[0, 1000]^{10}$ for model (7.2). Furthermore, $r_1 = 0.7$ and $r_2 = 0.29$, i.e. the copula sampling scheme q_1 was used for ca. 70%, the CovRWMH sampling scheme q_2 for ca. 29%, and q_3 for ca. 1% of the proposals. Following Calderhead & Girolami [2009] the thermodynamic integration schedule $T_i = (i/30)^5$, $i = 0, ..., 30$, was applied (see Equation (3.22)). Based on 1,000,000 prerun samples for each T_i we computed the Bayes factor of model (7.2)

versus model (7.1) using 1,000,000 ACIMH proposals with copula update parameters of R =100,000 and S =4 (see Equation (6.6)) in each T_i. The model prior probability of each model was set to $\frac{1}{2}$. All copulas were fitted on a total of 3,000 samples. Throughout the permutation function ι was chosen to be the identity function. For uniformization we applied fitted normal distributions. Each ACIMH run was started at an arbitrarily drawn sample from the according prior distribution. We applied the Geweke test (Chapter 4.4) to correct for a burn-in phase in every chain. The resulting Bayes factor computed to

$$B_{1,2} = 1.5 \cdot 10^3 \tag{7.3}$$

in favor of model (7.1). Including one standard error, this result is based on 453 ± 59 and 922 ± 149 effective samples for each T_i at average acceptance rates $9.11 \pm 0.87\%$ and $8.10 \pm 0.53\%$ for model (7.1) and model (7.2), respectively. The solutions to (7.1) and (7.2) were derived numerically applying the transformation of Appendix D as done for the JAK2-STAT5 system. The time course for *pgp*130 was fitted prior to inference using a rescaled lognormal density function (Figure 7.2(a)).

According to (7.3) the first model is favored decisively on Jeffreys' scale of evidence. The effect of direct tyrosine-phosphorylated STAT3 dimer import into the nucleus is thus negligible. Henceforth we focus on model (7.1) for further inference. Model (7.1) is identifiable as it differs by a second linear transformation step compared to the JAK2-STAT5 model. The (posterior) time courses are depicted in Figure 7.2. Despite the error prone total STAT data (Figure 7.2(d)) the model can cover the phosphorylated STAT measurements well (Figure 7.2(b) and 7.2(c)). Due to the restriction of $k_4 \geq k_5$ there is a strong correlation between theses two parameters (Figure 7.3). Based on $k_2 \geq k_3$ higher order dependencies arise as can be seen in the pairwise density plot. Especially k_8 seems to inherit many nonlinear dependencies (last row of Figure 7.3). In contrast to the JAK2-STAT5 pathway the nuclear abidance time decreases drastically (Table 7.1). The MAP estimate for τ computed to 0.252 minutes = 15.12 seconds, while the upper bound of the 95% credible interval is 2 minutes. This is unnaturally short. Here, the error prone total STAT data might effect the result. The estimated marginal posterior means, modes (MAP estimates), as well as the 90% posterior quantile based credible intervals are given in Table 7.1. An identifiability analysis as introduced in Chapter 3.3

showed that all parameters are identifiable. Note that k_1 contains a relativity constant as mentioned above and is therefore not meaningful as such.

$\mathbb{E}[k_1\|\boldsymbol{y}]$	0.024	$\mathbb{E}[k_5\|\boldsymbol{y}]$	0.176
$M[k_1\|\boldsymbol{y}]$	0.023	$M[k_5\|\boldsymbol{y}]$	0.176
$CI[k_1\|\boldsymbol{y}]$	(0.014; 0.035)	$CI[k_5\|\boldsymbol{y}]$	(0.131; 0.226)
$\mathbb{E}[k_2\|\boldsymbol{y}]$	1.869	$\mathbb{E}[\tau\|\boldsymbol{y}]$	0.625
$M[k_2\|\boldsymbol{y}]$	1.436	$M[\tau\|\boldsymbol{y}]$	0.252
$CI[k_2\|\boldsymbol{y}]$	(1.014; 3.432)	$CI[\tau\|\boldsymbol{y}]$	(0.017; 2.082)
$\mathbb{E}[k_3\|\boldsymbol{y}]$	0.871	$\mathbb{E}[k_6\|\boldsymbol{y}]$	28.656
$M[k_3\|\boldsymbol{y}]$	0.871	$M[k_6\|\boldsymbol{y}]$	26.305
$CI[k_3\|\boldsymbol{y}]$	(0.356; 1.596)	$CI[k_6\|\boldsymbol{y}]$	(22.900; 38.625)
$\mathbb{E}[k_4\|\boldsymbol{y}]$	0.182	$\mathbb{E}[k_8\|\boldsymbol{y}]$	3.674
$M[k_4\|\boldsymbol{y}]$	0.178	$M[k_8\|\boldsymbol{y}]$	3.657
$CI[k_4\|\boldsymbol{y}]$	(0.136;0.235)	$CI[k_8\|\boldsymbol{y}]$	(3.394; 3.964)

Table 7.1: JAK1-STAT3 pathway model. Estimated marginal posterior means $\mathbb{E}[\cdot|\boldsymbol{y}]$, modes $M[\cdot|\boldsymbol{y}]$ (MAP estimates), and 90% posterior quantile based credible intervals $CI[\cdot|\boldsymbol{y}]$ for the parameters $k_1, k_2, k_3, k_4, k_5, \tau, k_6$ and k_8. The according units are $[\text{min}^{-1}]$ for all k_i's and [min] for τ.

8

Inference of biokinetic models for zirconium processing in humans

Radioactive zirconium (Zr) isotopes are produced in large quantities in nuclear fission reactors; one of the most common fragments in uranium fission is Zirconium-95 (^{95}Zr). For example, the estimated released ^{95}Zr activity of the Fukushima and Chernobyl accidents is considered to have detrimental health effects not only via irradiation, but also via the intake of edibles (Eidgenössisches Nuklearsicherheitsinspektorat Informationsdienst [2011]; UNSCEAR [2008]). The estimation of radiation doses is indispensable for risk analysis for humans exposed to radioactive substances (ICRP [1979, 1988]). They provide limiting values of detrimental effects and build the basis for applications in internal dosimetry (ICRP [2007]), the prediction for radioactive zirconium retention in various organs (ICRP [1998]) as well as retrospective dosimetry, i.e. the estimation of ingested amounts of zirconium for *ex post* measurements. This is crucial for occupational exposure (ICRP [1979]), and for patients undergoing diagnostic and therapeutic nuclear medicine (ICRP [1988]).

In order to calculate the radiation dose and quantify the deposition of radioactivity from the incorporated radionuclide inside the human body, the International Commission on Radiological Protection (ICRP) in ICRP [1989] recommends a compartment model. It incorporates basic processes in the human physiological system (Guyton & Hall [2006]; ICRP [1975, 1979, 1989]). All major organs and tissues are simplified in the model structure as separate compartments that represent kinetically homogeneous

amounts of radionuclides; the connections between organs and tissues are described via *transfer rates*, i.e. model parameters that represent the exchange rates between the individual mutually exclusive compartments. These multi-compartmental systems along with their transfer parameters describing the kinetic behavior of radionuclides in the human body are called *biokinetic models* (ICRP [1989]). Throughout we use the terms biokinetic model and compartment model interchangeably. The transfer of substances into and out of compartments is governed by the law of mass balance. Transfer parameters are frequently evaluated on the basis of experimental data obtained from laboratory animals and, to a lesser extent, human beings (ICRP [1975]). Although animal data is not directly comparable to human data, the former can nevertheless be very helpful as prior information. In order to obtain more reliable dose estimates for humans, Greiter and coworkers developed a novel biokinetic model (Greiter *et al.* [2011]). It is based on the processing of non-radioactive Zr isotopes in 16 investigations with 12 healthy human subjects. In our case *in vivo* measurements were taken in urine and plasma (see Chapter 8.1). Although a global statistical analysis of the HMGU model was provided in Li *et al.* [2011a,b], a thorough comparison of the ICRP and HMGU model by a model selection approach was yet missing.

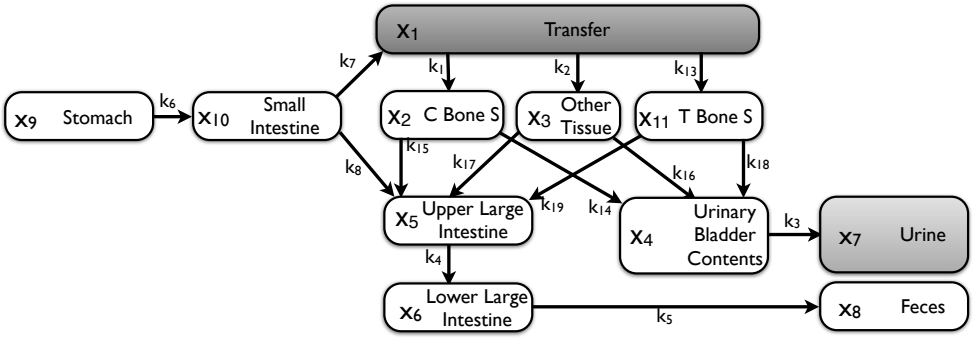
Applying thermodynamic integration in combination with CIMH we compared the HMGU and ICRP models based on *in vivo* plasma and urine data of the 16 investigations. More precisely, the models were evaluated for each investigation individually and for the concatenated data of all investigations. The latter allowed to infer transfer rates (including credible intervals) for an average subject. We also provide an analysis based on the (i) plasma data and (ii) urine data individually. Furthermore, the difference in accretion of zirconium in bones is inferred. The Bayesian framework also yields credible bounds for retrospective dose assessment of an average subject, this is, based on the concatenated data of all 16 investigations. We provide a simple to use estimation table for the prediction of initially ingested zirconium mass for *ex post* measurements. This impacts the estimation of intake and radiation dose, especially the bone dose, when aiming for personalized occupational monitoring data.

8.1 Experimental zirconium data

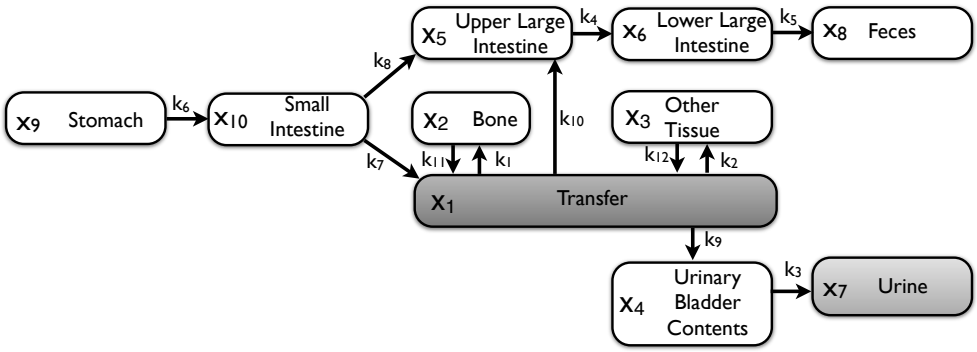
The human biokinetic data was measured in a stable tracer study executed at the Helmholtz Zentrum München (HMGU) (Greiter *et al.* [2011]). It includes 16 investigations with ingestion of a investigation-specific amount of isotopically enriched stable zirconium. The administered amount was based on the individuals weight, aiming at a dose of 0.09mg stable tracer per kg body weight. Tracer concentrations were determined in blood plasma and urine. For the plasma data, samples were taken several times during the first day in increasing intervals, and more scarcely later on. Urine was collected completely in 12-24h periods on several days. The last samples were taken at 100d after tracer administration. Tracer concentrations were normalized to the respective tracer amount ingested in each investigation, such that the total ingested amount corresponds 100% at $t = 0$ in the stomach. Concentrations in blood plasma were expressed as % per kg plasma. The plasma concentrations were scaled by the total amount of plasma in the body to get absolute concentrations (Alberts *et al.* [2002]). Urine data was expressed as excretion rate in % per day.

8.2 Mathematical models for zirconium processing

The currently used compartmental model was recommended by the ICRP in ICRP [1975, 1989, 1993] (Figure 8.1(a)). It consists of eleven compartments and 15 transfer rates. Zirconium enters the body via the stomach compartment x_9 and is processed until it reaches any of the two final compartments urine, x_7, or feces, x_8. The transfer compartment was taken to be identical with blood plasma. Some of the transfer rates and compartments of the ICRP model are however physiologically questionable: The direct mass transport from the two bone compartments to the urinary bladder contents and upper large intestine compartments or the distinction between trabecular bone surface and cortical bone surface as such. In order to address these shortcomings Greiter *et al.* [2011] recently proposed an alternative HMGU model combining the two bone compartments into one single compartment and replacing several mass flows by physiologically more plausible transfer rates (Figure 8.1(b)). Altogether both models



(a)



(b)

Figure 8.1: (a) Schematic representation of the ICRP model. The model consists of eleven compartments x_1,...,x_{11} and 15 time independent transfer rates k_1,...,k_8,k_{13},...,k_{19}. (b) Schematic representation of the HMGU model. The model consists of ten compartments x_1,...,x_{10} and twelve transfer rates k_1,...,k_{12}. In both models zirconium enters the body in the stomach compartment x_9 and diffuses through the system until it reaches either one of the two final compartments urine, x_7, or feces, x_8. The gray compartments x_1 and x_7 are directly related to the datasets measured.

share eight transfer rates, which we denote by $k_1, \ldots, k_8$. Transfers present in just one of the models have a unique index to facilitate distinction.

The dynamics of both models are described by a system of coupled linear first-order ordinary differential equations (ODEs). The ICRP model reads

$$
\begin{aligned}
\frac{\mathrm{d}x_1(t)}{\mathrm{d}t} &= (-k_1 - k_2 - k_{13})\, x_1(t) + k_7 x_{10}(t) \\
\frac{\mathrm{d}x_2(t)}{\mathrm{d}t} &= k_1 x_1(t) + (-k_{14} - k_{15})\, x_2(t) \\
\frac{\mathrm{d}x_3(t)}{\mathrm{d}t} &= k_2 x_1(t) + (-k_{16} - k_{17})\, x_3(t) \\
\frac{\mathrm{d}x_4(t)}{\mathrm{d}t} &= k_{14} x_2(t) + k_{16} x_3(t) - k_3 x_4(t) + k_{18} x_{11}(t) \\
\frac{\mathrm{d}x_5(t)}{\mathrm{d}t} &= k_{15} x_2(t) + k_{17} x_3(t) - k_4 x_5(t) + k_8 x_{10}(t) + k_{19} x_{11}(t) \\
\frac{\mathrm{d}x_6(t)}{\mathrm{d}t} &= k_4 x_5(t) - k_5 x_6(t) \\
\frac{\mathrm{d}x_7(t)}{\mathrm{d}t} &= k_3 x_4(t) \\
\frac{\mathrm{d}x_8(t)}{\mathrm{d}t} &= k_5 x_6(t) \\
\frac{\mathrm{d}x_9(t)}{\mathrm{d}t} &= -k_6 x_9(t) \\
\frac{\mathrm{d}x_{10}(t)}{\mathrm{d}t} &= k_6 x_9(t) + (-k_7 - k_8)\, x_{10}(t) \\
\frac{\mathrm{d}x_{11}(t)}{\mathrm{d}t} &= k_{13} x_1(t) + (-k_{18} - k_{19})\, x_{11}(t).
\end{aligned}
\tag{8.1}
$$

The HMGU model on the other hand is defined by

$$
\begin{aligned}
\frac{\mathrm{d}x_1(t)}{\mathrm{d}t} &= (-k_1 - k_2 - k_9 - k_{10})\, x_1(t) + k_{11} x_2(t) + k_{12} x_3(t) + k_7 x_{10}(t) \\
\frac{\mathrm{d}x_2(t)}{\mathrm{d}t} &= k_1 x_1(t) - k_{11} x_2(t) \\
\frac{\mathrm{d}x_3(t)}{\mathrm{d}t} &= k_2 x_1(t) - k_{12} x_3(t) \\
\frac{\mathrm{d}x_4(t)}{\mathrm{d}t} &= k_9 x_1(t) - k_3 x_4(t) \\
\frac{\mathrm{d}x_5(t)}{\mathrm{d}t} &= k_{10} x_1(t) - k_4 x_5(t) + k_8 x_{10}(t) \\
\frac{\mathrm{d}x_6(t)}{\mathrm{d}t} &= k_4 x_5(t) - k_5 x_6(t) \\
\frac{\mathrm{d}x_7(t)}{\mathrm{d}t} &= k_3 x_4(t) \\
\frac{\mathrm{d}x_8(t)}{\mathrm{d}t} &= k_5 x_6(t) \\
\frac{\mathrm{d}x_9(t)}{\mathrm{d}t} &= -k_6 x_9(t) \\
\frac{\mathrm{d}x_{10}(t)}{\mathrm{d}t} &= k_6 x_9(t) + (-k_7 - k_8)\, x_{10}(t).
\end{aligned}
\tag{8.2}
$$

In both models $x_9(0) = 100\%$ and therefore $x_{j\neq 9}(0) = 0\%$ at time point $t = 0$, this is,

the complete amount of zirconium is initially contained in the stomach compartment. For each investigation i we assume that the data

$$\boldsymbol{y}_i = \{x_1^{i,1}, x_1^{i,2}, \ldots, x_1^{i,n_i^p}, \dot{x}_7^{i,1}, \dot{x}_7^{i,2}, \ldots, \dot{x}_7^{i,n_i^u}\}$$

follows the solution $\boldsymbol{x}_{\boldsymbol{\xi}^m}(t)$ of the differential equation given in (8.1) and (8.2) for any of the two models $\mathcal{M}^m$ and some corresponding parameter vector $\boldsymbol{\xi}^m$. The model index $m \in \{H, I\}$, where $\mathcal{M}^I$ is the ICRP model and $\mathcal{M}^H$ the HMGU model. Corresponding to the notation in Figure 8.1(a) and 8.1(b), $\boldsymbol{\xi}^I = (k_1, \ldots, k_8, k_{13}, \ldots, k_{19})$ and $\boldsymbol{\xi}^H = (k_1, \ldots, k_{12})$. While for investigation i, $x_1^{i,\cdot}$ indicate measurements in plasma, i.e. in the transfer compartment x_1, $\dot{x}_7^{i,\cdot}$ designate measurements of the excretion rate in the urine compartment x_7. The expressions n_i^p denote the number of measurements in plasma and n_i^u the number of measurements in urine for investigation i. Assuming furthermore that all data points contain normally distributed measurement errors, the investigation i and model $\mathcal{M}^m$ specific likelihood function has the form

$$\mathcal{L}_i(\boldsymbol{\xi}^m|\boldsymbol{y}_i, m) = \underbrace{\prod_{\alpha=1}^{n_i^p} \Phi\left(x_1^{i,\alpha}|x_{\boldsymbol{\xi}^m}^p(t_\alpha), \sigma_i^p\right)}_{\mathcal{L}_i^p(\boldsymbol{\xi}^m|\boldsymbol{y}_i,m)} \underbrace{\prod_{\beta=1}^{n_i^u} \Phi\left(\dot{x}_7^{i,\beta}|\frac{\mathrm{d}}{\mathrm{d}t}x_{\boldsymbol{\xi}^m}^u(t_\beta), \sigma_i^u\right)}_{\mathcal{L}_i^u(\boldsymbol{\xi}^m|\boldsymbol{y}_i,m)},$$

where $x_{\boldsymbol{\xi}^m}^p(t_\alpha)$ denotes the solution for the transfer compartment x_1 at time point t_α of the according ODE system, corresponding to the measurement at $x_1^{i,\alpha}$, for the parameter vector $\boldsymbol{\xi}^m$. Furthermore, $\frac{\mathrm{d}}{\mathrm{d}t}x_{\boldsymbol{\xi}^m}^u(t_\beta)$ is the derivative of the solution for the urine compartment x_7 at time point t_β, corresponding to the measurement $\dot{x}_7^{i,\beta}$, while $\Phi(\cdot|\mu, \sigma)$ is the univariate probability density function of the normal distribution with mean μ and standard deviation σ. In order to take into account the biological variability, the combined model/measurement errors for plasma, σ_i^p, and for urine, σ_i^u, are fitted investigation specifically by simulated annealing (Chapter 4.6) before starting the MCMC sampling process. We tested all 16 investigations for non-Gaussian measurement/model error using individual MLE based time courses and applying the Kolmogorov-Smirnov test on normality. The null-hypothesis of non-Gaussian noise could not be rejected on an $\alpha = 5\%$ significance level for any investigation. The complete data (i.e. concatenated data $\boldsymbol{y} = \{by_1, \ldots, \boldsymbol{y}_{16}\}$) likelihood is given by $\mathcal{L}_{ALL}(\boldsymbol{\xi}^m|\boldsymbol{y}_i, m) = \prod_{i=1}^{16} \mathcal{L}_i(\boldsymbol{\xi}^m|\boldsymbol{y}_i, m)$ where in all $\mathcal{L}_i(\boldsymbol{\xi}^m|\boldsymbol{y}_i, m)$ the same fitted investigation independent $\sigma_i^p = \sigma^p$ and

$\sigma_i^u = \sigma^u$ are used. For the calculation of the likelihood $\mathcal{L}_\cdot(\boldsymbol{\xi}^m|\boldsymbol{y}_\cdot, m)$ we solved the according ODE system semi-analytically: Writing the ODE system as

$$\frac{\mathrm{d}\boldsymbol{x}_{\boldsymbol{\xi}^m}(t)}{\mathrm{d}t} = A(\boldsymbol{\xi}^m) \cdot \boldsymbol{x}_{\boldsymbol{\xi}^m}(t),$$

where $\boldsymbol{x}_{\boldsymbol{\xi}^m}(t)$ is the vector of all the compartments of model $\mathcal{M}^m$ and the time independent matrix $A(\boldsymbol{\xi}^m)$ represents all the interactions between these compartments, depending on the transfer rate values $\boldsymbol{\xi}^m$, the corresponding analytical solution is given by

$$\boldsymbol{x}_{\boldsymbol{\xi}^m}(t) = e^{A(\boldsymbol{\xi}^m)t} \cdot \boldsymbol{x}_{\boldsymbol{\xi}^m}(t=0).$$

We computed the matrix exponential $e^{A(\boldsymbol{\xi}^m)t}$ by eigenvalue decomposition, i.e.

$$e^{A(\boldsymbol{\xi}^m)t} = \boldsymbol{U}(\boldsymbol{\xi}^m) \begin{pmatrix} e^{d_1(\boldsymbol{\xi}^m)t} & 0 & \cdots & 0 \\ 0 & e^{d_2(\boldsymbol{\xi}^m)t} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & e^{d_V(\boldsymbol{\xi}^m)t} \end{pmatrix} \boldsymbol{U}(\boldsymbol{\xi}^m)^{-1}$$

for the eigenvalues $d_1(\boldsymbol{\xi}^m), d_2(\boldsymbol{\xi}^m), \ldots, d_V(\boldsymbol{\xi}^m)$ of $A(\boldsymbol{\xi}^m)$ and some orthonormal matrix $\boldsymbol{U}(\boldsymbol{\xi}^m)$. The eigenvalues were estimated using MATLAB's `eig` function. As we have no initial preference for any of the models we chose a uniform model prior. The model specific, investigation independent prior distributions $\pi(\boldsymbol{\xi}^m|m)$ are based on combined human/animal data as specified in the following section.

8.3 Prior information for zirconium processing and algorithmic set up

Since the problem of radiation protection is of great relevance, quite a large number of animal studies have been conducted. These yield excessive prior knowledge for a Bayesian modeling approach. As the ICRP model is the recommended model used for dose estimation, there exists information on the distribution types of the parameters involved in the model along with confidence intervals (Li *et al.* [2011a], and Table 8.1). The prior informations are based on a large number of studies and well-established over the years. Even for the HMGU model, detailed prior information is available from previous studies (Li *et al.* [2011a,b], and Table 8.1). Here, the prior informations are in part directly derived from ICRP recommendations, plus information gained from

additional experiments based on injected zirconium doses (Li *et al.* [2011a]). As we do not use the injection data for our analysis, the HMGU prior informations are not biasing the sampling outcome.

It is noteworthy that of the eight parameters shared in both models, k_8 is the only one having different distributions in the ICRP and HMGU model. Due to a lack of knowledge of the dependency structure between the parameters, the multivariate prior distribution $\pi(\boldsymbol{\xi}^m|m)$ of model $\mathcal{M}^m$ was taken to be the product of the univariate prior distributions $\pi(\boldsymbol{\xi}^m|m)$ for each parameter k_r^m, i.e. $\pi(\boldsymbol{\xi}^m|m) = \prod_r \pi(k_r^m|m)$. Each univariate prior distribution was truncated at zero. While Bayes factors were computed for each investigation separately (see Chapter 8.4.3), the same prior information was applied throughout all investigations. This is reasonable, as the priors contain information from a huge variety of different preceding experiments.

We again used thermodynamic integration (see Chapter 3.4.3) in combination with the CIMH algorithm in order to compare the ICRP model with the HMGU model based on Bayes factors. For all our applications the thermodynamic integration schedule $T_i = (i/29)^5$, $i = 0, ..., 29$, was applied (see Equation (3.22)). The independence proposal function q_3 of CIMH was chosen to be the product of the according prior distributions. Furthermore, $r_1 = 0.89$ and $r_2 = 0.1$, i.e. the copula sampling scheme q_1 was used for ca. 89%, the CovRWMH sampling scheme q_2 for ca. 10%, and q_3 for ca. 1% of the proposals. Throughout the permutation function ι was chosen to be the identity function. Fitting copula distributions was done in preruns containing 1,000,000 unthinned samples each. They were generated for each investigation and model separately. For uniformization of the prerun samples, we naturally applied the according prior distributions of the models at hand. Before starting the MCMC sampling procedure, the maximum a posteriori parameter estimates were computed by simulated annealing and used as initial MCMC sampling values. This makes a burn-in period dispensable. Finally, all Bayes factors were computed based on 30,000 proposals of the CIMH algorithm at each T_i throughout all applications.

ICRP model

Par.	Compartments	Med. (d^{-1})	99.7% CI	Distribution	μ/a	σ/c	b
k_1	TC → CBS	0.69	[0.086, 5.52]	$\mathcal{LN}(\mu,\sigma)$	-0.3711	0.6931	
k_2	TC → Other	1.39	[0.174, 11.12]	$\mathcal{LN}(\mu,\sigma)$	0.3293	0.6931	
k_3	UBC → Urine	12		$\mathcal{T}(a,b,c)$	6	8	24
k_4	UpLI → LoLI	1.8		$\mathcal{T}(a,b,c)$	0.9	1.2	3.6
k_5	LoLI → Feces	1		$\mathcal{T}(a,b,c)$	0.3	1	1.7
k_6	Stomach → SI	24		$\mathcal{T}(a,b,c)$	12	16	48
k_7	SI → TC	0.06	[0.0075, 0.48]	$\mathcal{LN}(\mu,\sigma)$	-2.8134	0.6931	
k_8	SI → UpLI	6		$\mathcal{T}(a,b,c)$	3	4	12
k_{13}	TC → TBS	0.69	[0.086, 5.52]	$\mathcal{LN}(\mu,\sigma)$	-0.3711	0.6931	
k_{14}	CBS → UBC	$5.8 \cdot 10^{-5}$	$[5.8 \cdot 10^{-6}, 1.1 \cdot 10^{-4}]$	$\mathcal{N}(\mu,\sigma)$	$5.8 \cdot 10^{-5}$	$1.7 \cdot 10^{-5}$	
k_{15}	CBS → UpLI	$1.2 \cdot 10^{-5}$	$[1.2 \cdot 10^{-6}, 2.2 \cdot 10^{-5}]$	$\mathcal{N}(\mu,\sigma)$	$1.2 \cdot 10^{-5}$	$3.5 \cdot 10^{-6}$	
k_{16}	Other → UBC	0.083	[0.0083, 0.158]	$\mathcal{N}(\mu,\sigma)$	0.083	0.025	
k_{17}	Other → UpLI	0.0165	[0.00165, 0.0314]	$\mathcal{N}(\mu,\sigma)$	0.0165	0.00495	
k_{18}	TBS → UBC	$5.8 \cdot 10^{-5}$	$[5.8 \cdot 10^{-6}, 1.1 \cdot 10^{-4}]$	$\mathcal{N}(\mu,\sigma)$	$5.8 \cdot 10^{-5}$	$1.7 \cdot 10^{-5}$	
k_{19}	TBS → UpLI	$1.2 \cdot 10^{-5}$	$[1.2 \cdot 10^{-6}, 2.2 \cdot 10^{-5}]$	$\mathcal{N}(\mu,\sigma)$	$1.2 \cdot 10^{-5}$	$3.5 \cdot 10^{-6}$	

HMGU model

Par.	Compartments	Med. (d^{-1})	99.7% CI	Distribution	μ/a	σ/c	b
k_1	TC → Bone	0.10	[0.013, 0.8]	$\mathcal{LN}(\mu,\sigma)$	-2.3026	0.6931	
k_2	TC → Other	1.35	[0.17, 10.8]	$\mathcal{LN}(\mu,\sigma)$	0.3001	0.6931	
k_3	UBC → Urine	12.0		$\mathcal{T}(a,b,c)$	6.0	8.0	24.0
k_4	UpLI → LoLI	1.8		$\mathcal{T}(a,b,c)$	0.9	1.2	3.6
k_5	LoLI → Feces	1.0		$\mathcal{T}(a,b,c)$	0.3	1.0	1.7
k_6	Stomach → SI	24.0		$\mathcal{T}(a,b,c)$	12.0	16.0	48.0
k_7	SI → TC	0.03	$[1.1 \cdot 10^{-3}, 0.81]$	$\mathcal{LN}(\mu,\sigma)$	-3.5066	1.0986	
k_8	SI → UpLI	17.21	[0.64, 464.67]	$\mathcal{LN}(\mu,\sigma)$	2.8455	1.0986	
k_9	TC → UBC	0.031	[0.0011, 0.8370]	$\mathcal{LN}(\mu,\sigma)$	-3.4738	1.0986	
k_{10}	TC → UpLI	0.0062	[0.0002, 0.1674]	$\mathcal{LN}(\mu,\sigma)$	-5.0832	1.0986	
k_{11}	Bone → TC	$6.9 \cdot 10^{-5}$	$[8.7 \cdot 10^{-6}, 5.6 \cdot 10^{-4}]$	$\mathcal{LN}(\mu,\sigma)$	-9.5769	0.6931	
k_{12}	Other → TC	0.53	[0.066, 4.24]	$\mathcal{LN}(\mu,\sigma)$	-0.6349	0.6931	

Table 8.1: Overview of priors for the zirconium models. The tables are based on Li *et al.* [2011a], where the confidence intervals (CI), the medians (Med.) as well as the parameters of the normal and lognormal distributions were calculated according to Appendix F. Abbreviations are: $\mathcal{LN}(\mu,\sigma)$ for a lognormal distribution with location parameter μ and scale parameter σ, $\mathcal{T}(a,b,c)$ for a triangular distribution with lower limit a, upper limit b, and mode c, as well as $\mathcal{N}(\mu,\sigma)$ for a normal distribution with mean μ and standard deviation σ. Furthermore TC= Transfer compartment; CBS = Cortical Bone Surface; Other = Other Tissues; UBC = Urinary Bladder Contents; UpLi = Upper Large Intestine; LoLI = Lower Large Intestine; SI = Small Intestine; TBS = Trabecular Bone Surface.

8.4 Inference of the zirconium models

We now present the results of our analysis, which primarily addresses the question of model selection between the HMGU and ICRP models. First several general results, such as investigation dependency of the Bayes factor and effects of parameter correlations are shown, before turning to the results of the model selection, and their consequences for the HMGU and ICRP models.

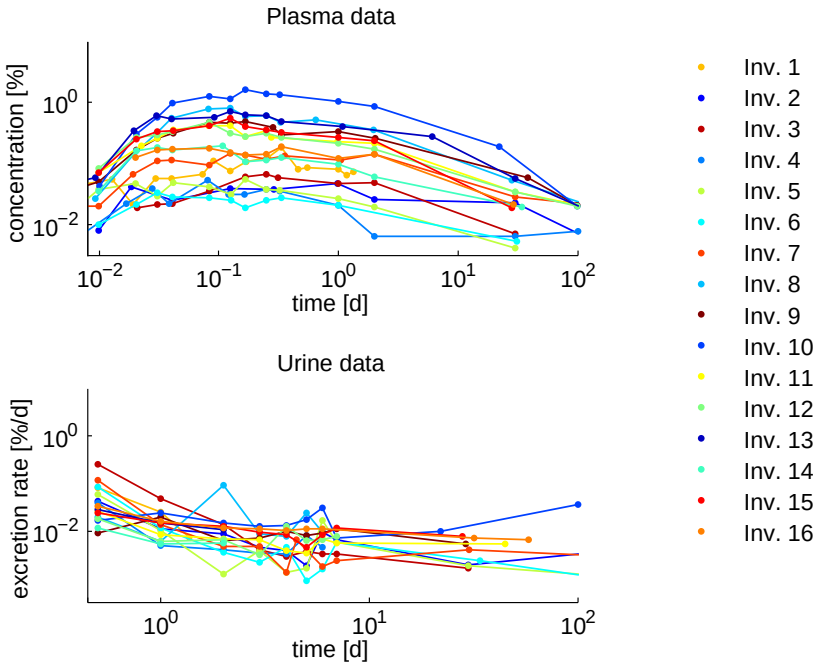


Figure 8.2: Plasma and urine data for investigations 1-16 on log-log-scale.

8.4.1 Investigation specificity of transfer rates

In radiation protection the transfer rates for the biokinetics of radionuclides in the human body are derived from data collected in various independent experiments (ICRP, 2008). We here used plasma and urine measurements of 16 different investigations. This poses the question whether the models should be compared based on the complete dataset $\boldsymbol{y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{16}\}$, or whether statistical evaluation should be done for each investigation individually on $\boldsymbol{y}_i$ for $i = 1, \ldots, 16$. While the former approach

results in one overall Bayes factor, the latter yields 16 investigation specific, not directly comparable Bayes factors. Figure 8.2 shows that all investigations follow the same pulse-like time courses in the transfer compartment x_1 while the excretion rates in the urine compartment x_7 exhibit an exponential decay behavior. However, zirconium tracer concentrations showed up to a 50-fold difference between maximal plasma concentrations, i.e. for investigation 10 (1.616%,) and 6 (0.033%).

To test the hypothesis whether the diversity in concentration also effects transfer rates and therefore the estimated Bayes factors, we pairwise compared the posterior sample marginals of the MCMC run (corresponding to the samples of $T_{29} = 1$) for the parameter k_7 of the ICRP model between all investigations by means of a Kolmogoroff-Smirnov test. Here k_7 was chosen as it directly affects the observed plasma levels (Li *et al.* [2011b]). Except for one pair, all p-values were $< 6 \cdot 10^{-8}$, meaning that the chance of falsely rejecting the hypothesis of comparable marginals is negligible. Therefore, as the posterior marginal distributions are quite different, it can be deduced that the basis for the Bayes factor, the joint posterior distribution, can differ quite strongly with respect to the individuals. This indicated that each investigation should be treated separately. Nevertheless, in order to infer the transfer rates of an average subject (Table 8.5) the concatenated data has to be used. We therefore compared the HMGU and ICRP model based on both the concatenated data $\boldsymbol{y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{16}\}$ and, in order to account for the biological diversity, the individual patient specific datasets $\boldsymbol{y}_i$ $(i = 1, \ldots, 16)$. This could also be the basis for further analysis of influence factors, such as weight or gender.

8.4.2 Parameter correlations

The posterior probabilities of both the HMGU and ICRP model show strong correlation between the parameters k_7 and k_8 throughout all investigations. The estimated Kendall's τ's based on the preruns were $\hat{\tau}_{HMGU} = 0.8027 \pm 0.01$ and $\hat{\tau}_{ICRP} = 0.3452 \pm 0.02$. This can be explained as follows: At time point $t = 0$ the stomach compartment x_9 is the only compartment with non-zero Zr concentration. It is exclusively connected to the small intestines x_{10} in both models. Therefore, all Zr compounds have to pass through x_{10}, which further on distributes them to the observed transfer compartment x_1 via k_7 or to the upper large intestines x_5 via k_8. Aberrations in one of

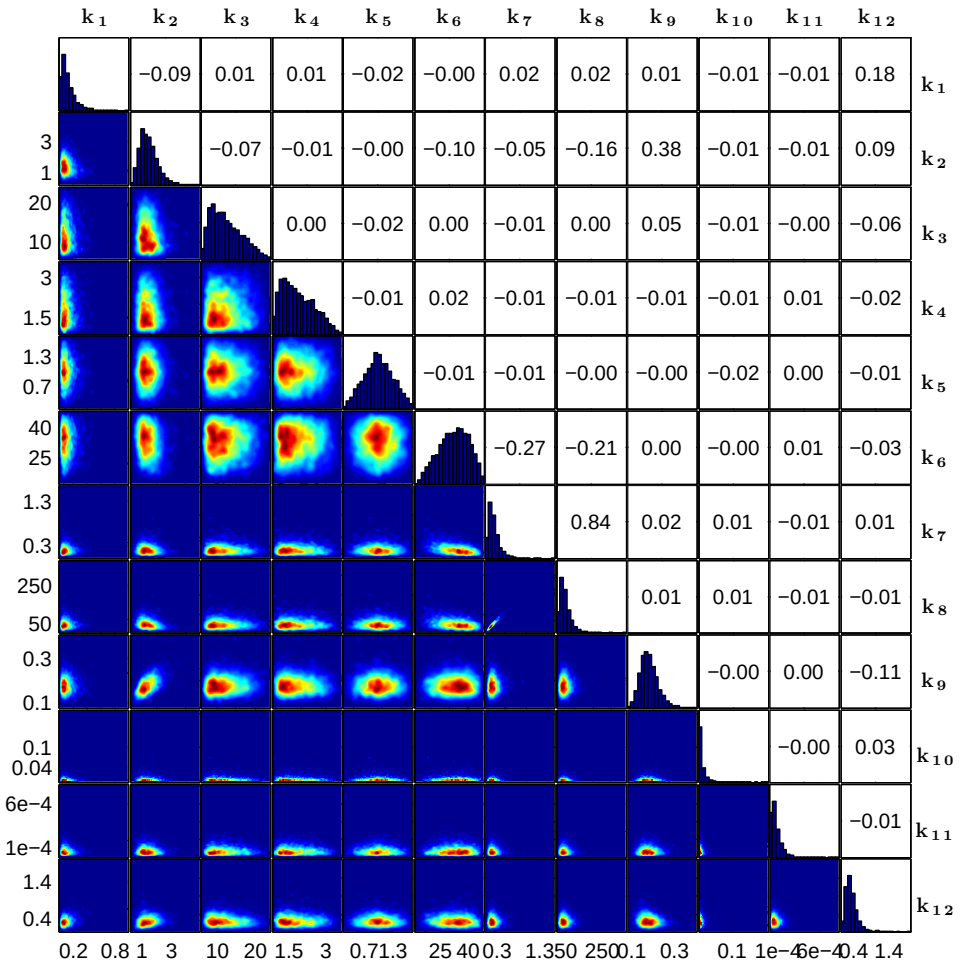


Figure 8.3: Pairwise density plots for all parameter-pairs of the HMGU posterior. Red areas depict higher, blue areas lower density values. The diagonal displays the histograms of the sample marginals and the numbers in the upper right triangle the estimated Kendall's τ's.

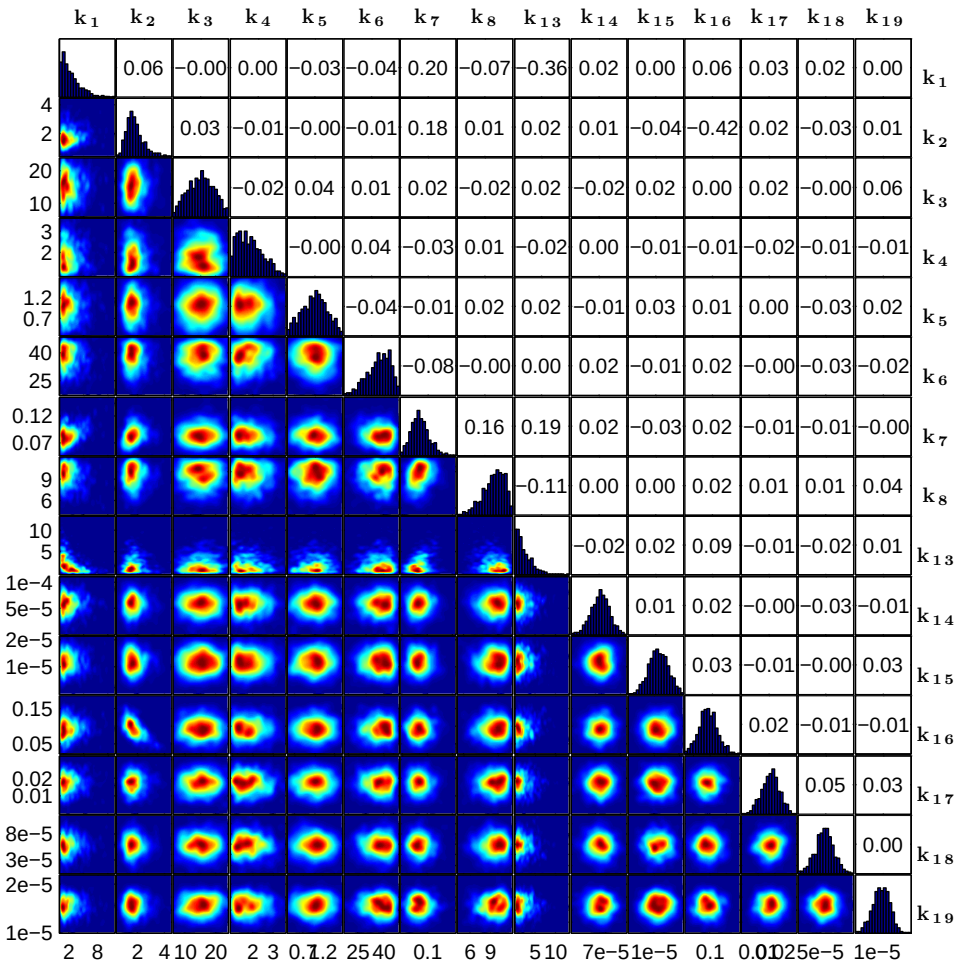


Figure 8.4: Pairwise density plot for all parameter-pairs of the ICRP posterior. Red areas depict higher, blue areas lower density values. The diagonal displays the histograms of the sample marginals and the numbers in the upper right triangle the estimated Kendall's τ's.

the parameters k_7 or k_8 thus have a direct effect on the amount of zirconium predicted for x_1. This affects the according posterior distributions. The same effect is found for the complete data $\boldsymbol{y}$ (see pairwise scatterplots in Figure 8.4 and 8.3). Despite the parameter dependencies, the posterior distributions of the ICRP and HMGU model are identifiable for all 16 investigations, this is, the investigation specific maximum a posteriori estimates are well defined and inferable (Figure 8.4 and 8.3).

8.4.3 Bayesian model comparison of the HMGU and ICRP models

Applying thermodynamic integration in combination with the CIMH algorithm we compared the HMGU and the ICRP model based on (i) the concatenated data $\boldsymbol{y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{16}\}$ and (ii) the individual investigation specific datasets $\boldsymbol{y}_i$ $(i = 1, \ldots, 16)$. This resulted in a total of 17 Bayes factors. We found that all Bayes factors favored the HMGU model; in 14 out of the 17 cases even decisively (Table 8.2). Throughout, the analysis was based on 30,000 proposals for each of the 30 T_i-levels in all 17 cases. Based on 30,000 proposals the average ESS of the HMGU model including one standard error, i.e. taken over all T_i levels and investigations, was 5832 ± 405. In case of the ICRP model we obtained in average 5808 ± 252 (approximately independent) samples from the Markov chains. This implied high acceptance rates for both models. The sampling procedure thus captured the power posteriors very well.

In order to take a closer look at the contribution of the plasma and urine data to the above results, we computed additional Bayes factors based on the likelihoods $\mathcal{L}_i^p(\boldsymbol{\xi}^m|\boldsymbol{y}_i, m)$ and $\mathcal{L}_i^u(\boldsymbol{\xi}^m|\boldsymbol{y}_i, m)$ individually. Here, $i = 1, \ldots, 16, ALL$ and $m \in \{I, H\}$, where I represents the ICRP and H the HMGU model. The time courses already suggested better coverage of plasma data by the HMGU model (Figure 8.5, and individual time courses in Appendix G); for urine the situation is not that clear. This was confirmed by the Bayes factors: all 17 Bayes factors based on plasma data favored the HMGU model; in ten cases even decisively (Table 8.3). For the urine data, three investigations slightly favored the ICRP model (Table 8.4). In summary, all decisive Bayes factors are in favor of the HMGU model. This means the HMGU model was never decisively rejected. On the other hand the ICRP model was decisively rejected in the majority of cases. Hence, the HMGU model is superior over the ICRP model with

Inv.	$B_{H,I}$	AR HMGU [min, med, max]	AR ICRP [min, med, max]	ESS HMGU [min, med, max]	ESS ICRP [min, med, max]
ALL	$1.2010 \cdot 10^{11}$	[0.54, 0.57, 0.64]	[0.27, 0.57, 0.60]	[1500, 4643, 10000]	[639, 4643, 10000]
1	71.7283	[0.56, 0.65, 0.70]	[0.55, 0.63, 0.67]	[1000, 6000, 15000]	[2308, 6000, 10000]
2	114.6109	[0.59, 0.65, 0.69]	[0.54, 0.56, 0.58]	[2000, 7500, 15000]	[1035, 5000, 10000]
3	59532.2127	[0.42, 0.61, 0.66]	[0.47, 0.60, 0.66]	[181, 5500, 10000]	[1667, 6000, 15000]
4	1065.3125	[0.63, 0.66, 0.68]	[0.55, 0.62, 0.64]	[2500, 7500, 15000]	[811, 6000, 15000]
5	219.0939	[0.60, 0.66, 0.69]	[0.56, 0.62, 0.65]	[3750, 7500, 15000]	[1579, 6000, 15000]
6	4642.8755	[0.61, 0.63, 0.68]	[0.57, 0.62, 0.66]	[1072, 6000, 15000]	[1765, 6000, 10000]
7	218.0765	[0.62, 0.66, 0.71]	[0.60, 0.64, 0.67]	[1667, 7500, 15000]	[4286, 7500, 15000]
8	37.5182	[0.47, 0.61, 0.71]	[0.59, 0.64, 0.69]	[2308, 7500, 15000]	[3334, 7500, 15000]
9	462.3241	[0.48, 0.57, 0.71]	[0.41, 0.63, 0.67]	[770, 6000, 15000]	[698, 5500, 15000]
10	861.7574	[0.43, 0.60, 0.71]	[0.44, 0.64, 0.68]	[2000, 7500, 15000]	[126, 5000, 10000]
11	117250.4521	[0.38, 0.49, 0.63]	[0.49, 0.57, 0.59]	[1072, 4286, 15000]	[698, 4286, 7500]
12	177.9964	[0.26, 0.61, 0.72]	[0.46, 0.62, 0.68]	[313, 5000, 10000]	[2308, 6000, 15000]
13	718.7546	[0.10, 0.44, 0.70]	[0.53, 0.58, 0.60]	[169, 4018, 15000]	[2308, 4643, 10000]
14	35.8079	[0.09, 0.41, 0.69]	[0.56, 0.64, 0.69]	[345, 3000, 15000]	[1500, 7500, 15000]
15	6287.6538	[0.22, 0.53, 0.70]	[0.46, 0.64, 0.68]	[121, 5500, 15000]	[1000, 5000, 15000]
16	622.4126	[0.23, 0.56, 0.64]	[0.51, 0.58, 0.59]	[417, 3000, 10000]	[1765, 5000, 10000]

Table 8.2: Bayes factors for the HMGU versus the ICRP model ($B_{H,I}$) for investigation $1, \ldots, 16$ and the complete data model (ALL). Green color indicates a Bayes factor in favor of the HMGU model. Shown are also the minimal, median, and maximal acceptance rates (AR) and effective sampling sizes (ESS) for both models.

Inv.	$B^p_{H,I}$	AR HMGU [min, med, max]	AR ICRP [min, med, max]	ESS HMGU [min, med, max]	ESS ICRP [min, med, max]
ALL	34283.1711	[0.56, 0.61, 0.64]	[0.41, 0.57, 0.60]	[1000, 6750, 10000]	[1875, 5000, 10000]
1	71.1549	[0.53, 0.65, 0.69]	[0.56, 0.62, 0.66]	[834, 6000, 30000]	[455, 6000, 15000]
2	293.4270	[0.58, 0.64, 0.67]	[0.59, 0.62, 0.66]	[338, 7500, 15000]	[546, 7500, 10000]
3	52297.4330	[0.45, 0.62, 0.66]	[0.55, 0.61, 0.65]	[1200, 6000, 15000]	[1765, 6000, 15000]
4	2639.9965	[0.56, 0.60, 0.63]	[0.50, 0.56, 0.57]	[2308, 6000, 15000]	[968, 3334, 10000]
5	473.1182	[0.59, 0.65, 0.69]	[0.59, 0.63, 0.68]	[3334, 8750, 15000]	[1429, 7500, 15000]
6	3926.9639	[0.62, 0.65, 0.70]	[0.48, 0.62, 0.66]	[577, 7500, 10000]	[698, 6000, 10000]
7	229.9698	[0.55, 0.64, 0.72]	[0.59, 0.64, 0.68]	[968, 6000, 15000]	[2143, 6750, 15000]
8	127.7723	[0.38, 0.57, 0.72]	[0.56, 0.64, 0.69]	[667, 6750, 15000]	[2308, 7500, 15000]
9	231.8086	[0.50, 0.57, 0.65]	[0.58, 0.65, 0.69]	[653, 4286, 15000]	[3334, 7500, 15000]
10	115.6091	[0.53, 0.61, 0.65]	[0.52, 0.58, 0.60]	[215, 4643, 15000]	[1667, 6000, 10000]
11	18.0543	[0.56, 0.65, 0.71]	[0.59, 0.65, 0.69]	[1000, 5500, 15000]	[3750, 6750, 15000]
12	5.4764	[0.55, 0.61, 0.64]	[0.56, 0.58, 0.60]	[750, 6000, 15000]	[2728, 5000, 15000]
13	14.1274	[0.50, 0.67, 0.71]	[0.60, 0.65, 0.67]	[1154, 7500, 15000]	[4286, 7500, 15000]
14	7.4250	[0.59, 0.67, 0.72]	[0.62, 0.65, 0.69]	[2500, 7500, 15000]	[2728, 8750, 15000]
15	21.6865	[0.56, 0.61, 0.65]	[0.55, 0.58, 0.59]	[750, 7500, 15000]	[2000, 5000, 10000]
16	13.4114	[0.56, 0.66, 0.70]	[0.59, 0.66, 0.68]	[625, 7500, 30000]	[3334, 7500, 15000]

Table 8.3: Bayes factors for the HMGU versus the ICRP model ($B^p_{H,I}$) for investigation $1, \ldots, 16$ and the complete data model (ALL) based on plasma data only. Green color indicates a Bayes factor in favor of the HMGU model. Shown are also the minimal, median, and maximal acceptance rates (AR) and effective sampling sizes (ESS) for both models.

Inv.	$B^u_{H,I}$	AR HMGU [min, med, max]	AR ICRP [min, med, max]	ESS HMGU [min, med, max]	ESS ICRP [min, med, max]
ALL	47303749.2905	[0.42, 0.58, 0.65]	[0.37, 0.57, 0.60]	[35, 4286, 15000]	[600, 4286, 10000]
1	1.0460	[0.66, 0.70, 0.73]	[0.64, 0.66, 0.68]	[5000, 10000, 15000]	[3750, 7500, 15000]
2	3940.3951	[0.35, 0.54, 0.64]	[0.54, 0.57, 0.60]	[770, 4643, 10000]	[2500, 5000, 7500]
3	1.3352	[0.59, 0.70, 0.73]	[0.60, 0.67, 0.70]	[2143, 8750, 15000]	[4286, 7500, 15000]
4	34.7362	[0.46, 0.65, 0.72]	[0.59, 0.65, 0.69]	[380, 6000, 15000]	[2143, 7500, 15000]
5	133.8984	[0.43, 0.59, 0.64]	[0.55, 0.58, 0.60]	[244, 4018, 15000]	[2308, 5000, 10000]
6	2384.2435	[0.13, 0.48, 0.63]	[0.58, 0.61, 0.62]	[257, 3000, 15000]	[1667, 4286, 7500]
7	1335.8332	[0.13, 0.50, 0.63]	[0.58, 0.61, 0.62]	[136, 3167, 10000]	[2500, 4286, 10000]
8	0.2221	[0.57, 0.69, 0.72]	[0.58, 0.66, 0.68]	[3750, 10000, 15000]	[3334, 7500, 15000]
9	0.1753	[0.33, 0.62, 0.70]	[0.46, 0.63, 0.68]	[235, 4286, 10000]	[770, 7500, 15000]
10	0.1992	[0.48, 0.68, 0.71]	[0.58, 0.64, 0.69]	[1154, 10000, 30000]	[2000, 6750, 15000]
11	2936.7417	[0.33, 0.48, 0.63]	[0.52, 0.57, 0.60]	[273, 3542, 15000]	[2143, 5000, 7500]
12	11.4359	[0.51, 0.59, 0.64]	[0.50, 0.57, 0.60]	[546, 5500, 10000]	[1072, 5000, 7500]
13	4.4105	[0.48, 0.64, 0.71]	[0.59, 0.65, 0.69]	[1000, 6750, 15000]	[2728, 7500, 15000]
14	9.7741	[0.43, 0.54, 0.63]	[0.53, 0.57, 0.60]	[968, 5000, 15000]	[1875, 5000, 10000]
15	160.0045	[0.40, 0.61, 0.71]	[0.52, 0.63, 0.68]	[320, 5000, 15000]	[1875, 6000, 15000]
16	12003.8714	[0.30, 0.50, 0.63]	[0.54, 0.61, 0.62]	[366, 3334, 7500]	[1500, 3750, 10000]

Table 8.4: Bayes factors for the HMGU versus the ICRP model ($B^u_{H,I}$) for investigation $1, \ldots, 16$ and the complete data model (ALL) based on urine data only. Green color indicates a Bayes factor in favor of the HMGU model and red color a Bayes factor in favor of the ICRP model. Shown are also the minimal, median, and maximal acceptance rates (AR) and effective sampling sizes (ESS) for both models.

respect to zirconium processing in the human body. This not only holds investigation specifically, but also based on the complete data.

The posterior median (MAP) as well as the according 95% credible intervals for the HMGU parameter values based on the complete data $\boldsymbol{y}$ are given in Table 8.5. An identifiability analysis as introduced in Chapter 3.3 showed that all parameter rates are in fact identifiable. From a comparison with table 8.1, one can see that some parameters are slightly shifted. Since these parameter values are derived from the concatenated data, they are valid for all subjects and thus represent the parameters of choice for an average subject.

8.4.4 Differences in radioactive ^{95}Zr retention in bone predicted by the HMGU and ICRP models

In internal exposure monitoring, biokinetic models are used to predict the organ retention or daily excretion of incorporated radionuclides (ICRP, 1998). With an interval of

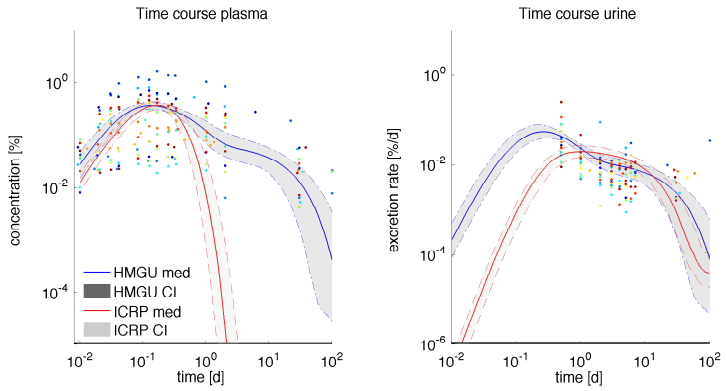


Figure 8.5: Posterior median solution (black line) and according 95% credible interval (shaded area) for the plasma and urinary excretion rate time courses based on the posterior HMGU and ICRP samples for the data $\boldsymbol{y}$. Colored markers are the data points. For readability we truncated the plasma plot at $1 \cdot 10^{-5}$[%] and the urine plot at $1 \cdot 10^{-6}$[%].

Param.	k_1	k_2	k_3	k_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.	k_5	k_6	k_7	k_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.	k_9	k_{10}	k_{11}	k_{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

Table 8.5: Posterior median (MAP) and according 95% credible intervals (CI) for the HMGU parameters based on the complete data $\boldsymbol{y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{16}\}$.

120 days the radioactivity of ^{95}Zr possibly incorporated by occupational workers is routinely monitored by whole body counters. Depending on the intake route, the radiation dose of bone surfaces or colon is taken as regulatory limit for a decision if an individual is requested for person-specific monitoring (BMU, 2007). In this monitoring procedure, the biokinetic model structure and parameters are used implicitly in the background. The organ retention function is the solution of the model in each compartment; the organ doses are directly related to the integral of radioactivity of ^{95}Zr in source organs over 50 years.

In order to compare the retention of ^{95}Zr as predicted by the ICRP and HMGU models, the 90% credible intervals for the time courses in the bone compartments were calculated based on the posterior samples. It is found that there is a significant difference

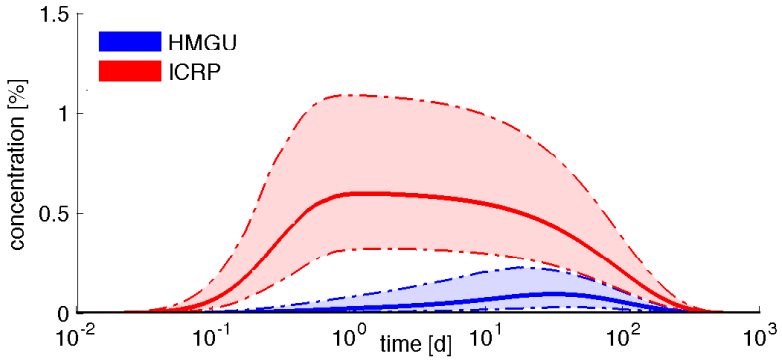


Figure 8.6: Posterior median (solid lines) as well as 90% credible intervals (shaded areas) for the retention of ^{95}Zr in the bone compartments as predicted by the HMGU (blue) and ICRP (red) models with radioactive decay taken into account.

between both models (Figure 8.6), where for the ICRP model we added the concentrations in the two bone compartments. The time courses were derived for stable isotopes of Zr and thus the radioactive decay of ^{95}Zr with half-life of 64.032d (ICRP, 2008) had to be taken into account. The decrease of retention in bone using the HMGU model consequently reduces the radiation dose in bone in comparison to the ICRP bone dose which is currently used in monitoring.

8.4.5 Retrospective dose assessment

Internal doses due to incorporated radionuclides have to be estimated with the help of biokinetic models based on indirect measurements, using for example bioassays for blood or urinary excretion. Normally, bioassay or in vivo data (e.g. radioactivity accumulated in skull or knee detected by a partial body counter) are measured after an accidental intake of radionuclides. Uncertainties of estimated doses are significant and have a large impact on remediation and thus action costs. In contrast to conventional uncertainty analysis (Li *et al.*, 2011a), our Bayesian *Ansatz* naturally integrates the uncertainties of measured data and parameters simultaneously. This trait of the Bayesian approach is helpful as it provides an estimate for the intake and its credible intervals.

For example, if the urinary excretion after accidental exposure is measured, we are able to compute credible intervals for the initial intake of radionuclide ^{95}Zr by exploiting

the posterior distribution together with the linearity of the HMGU model. In order to be as general as possible we use the posterior samples based on the complete data $\boldsymbol{y}$. Given a posterior sample $\boldsymbol{\xi}^H$, a measurement $\dot{x}_7^t$ in [μg/d] for the urinary excretion rate of zirconium at time point t corresponds to a unique solution $\boldsymbol{x}_{\boldsymbol{\xi}^H}(t)$ of the HMGU ODE system. Due to the linearity of the ODE's, the initial concentration $\boldsymbol{x}_{\boldsymbol{\xi}^H}(0)$ is by definition zero for all except the stomach compartment x_9. The latter reads $x_9(0) = \dot{x}_7^t \cdot 100\%/x^9_{\boldsymbol{\xi}^H}(t)$ where $x^9_{\boldsymbol{\xi}^H}(t)$ denotes the value of $\boldsymbol{x}_{\boldsymbol{\xi}^H}(t)$ in the stomach compartment at time point t. Now, assuming that for arbitrary posterior samples $\boldsymbol{\xi}^H$ the measurement $\dot{x}_7^t$ is contained in the 90% credible interval of the solution $\boldsymbol{x}_{\boldsymbol{\xi}^H}(t)$ with initial condition $x_9(0)$ as given above, lower and upper bounds for credible regions of the initial amount of zirconium taken in at $t_0 = 0$h emerge. These are based on the posterior samples. The estimated extrapolation factors for multiplication with a urine measurement (in [μg/d]) after time t (in [h]) are contained in Table 8.6 and yield the initially amount of zirconium contained in the stomach at $t_0 = 0h$. For example, if an amount of $\dot{x}_7^{2\text{d}} = 50\mu$g/d was measured after two days, we find from Table 8.6 that the 90% credible interval for the ingested amount lies between 0.029g and 0.059g. Since the above calculations are based on non-radioactive Zr isotopes, the results have to take into account the radioactive decay of the radionuclide in question, i.e. in many cases ^{95}Zr, for dose assessment.

Time t	6h	12h	18h	24h	30h
lbf for IC	1233.91	1820.44	2614.48	3369.70	4100.16
mf for IC	1763.73	2225.90	3153.70	4228.19	5340.23
ubf for IC	2512.54	2832.49	3978.27	5650.86	7516.00
Time t	36h	42h	48h	54h	60h
lbf for IC	4778.27	5352.64	5800.77	6153.80	6450.74
mf for IC	6364.76	7250.67	7977.31	8557.87	9006.97
ubf for IC	9122.11	10655.01	11878.81	12960.61	13903.07

Table 8.6: Retrospective urine predictions for the HMGU model. Shown are the lower bound factor (lbf), median factor (mf), and upper bound factor (ubf) for multiplication with a urine measurement (in [μg /d]) after time t (in [h]) on a 60h grid yielding the initial intake concentration (IC) at $t_0 = 0$h.

Concluding we have seen that transfer rates can differ quite heavily for the various investigations. However, the HMGU model is able to outperform the current ICRP

model based on the complete data (corresponding to an average individual) and investigation specifically. It can hence improve predictions in internal dosimetry compared to the current ICRP model.

9

Conclusions and outlook

In this thesis we have introduced two novel MCMC sampling schemes: The hybrid vine copula based independence/random walk Metropolis-Hastings algorithm (CIMH) and the adaptive vine copula based independence/random walk Metropolis-Hastings algorithm (ACIMH). A vine copula decomposition of the posterior distribution here exploits higher order parameter dependencies in order to generate efficient problem specific MCMC proposals. The algorithms were applied for parameter inference and model selection in various dynamical systems.

We tested the performance of CIMH and ACIMH on four examples: First of all, we inferred the (i) mean and covariance matrix of a strongly correlated two dimensional normal distribution. The system was analytically tractable and provided a simple proof-of-concept example. Subsequently, an (ii) ordinary differential equations driven steady state as well as an (iii) ordinary differential equations driven compartment model were considered. Finally an existing (iv) delay differential equations model of the JAK2-STAT5 signaling pathway (Swameye *et al.* [2003]) has been inferred.

Throughout, both algorithms were evaluated on the basis of the quotient of acceptance rate versus inefficiency factor ($\mathcal{I}_1$) and the number of independent samples generated per second ($\mathcal{I}_2$). Here, ($\mathcal{I}_1$) was motivated by the antagonistic behavior of high acceptance rates versus high INEFF's, while ($\mathcal{I}_2$) provided an easily interpretable performance statistic. As competing samplers a simple random walk Metropolis-Hastings, a covariance based random walk Metropolis-Hastings, and an independence chain sampler

were chosen for the first three examples. The JAK2-STAT5 pathway was additionally evaluated by SMALA and M-GaA. Our copula based approach generally covered the dependency structure of the posterior very well and outperformed all other sampling schemes in every example. It turned out that the basic CIMH algorithm is doing best on simple systems as it does not lose time on extra copula updates. However, in very complex situations, such as the inference of the JAK-STAT5 pathway, copula updates were needed to fine-tune the proposal distribution and thereby improve the performance.

We applied ACIMH to infer a model of the JAK1-STAT3 signaling pathway. Thermodynamic integration provided a Bayes factor that rejected a model covering direct tyrosine-phosphorylated STAT3 dimer import into the nucleus as compared to a model considering tyrosine-serine-phosphorylated STAT3 dimer import only. The estimated maximum a posteriori estimate for nuclear abidance time of the tyrosine-serine-phosphorylated STAT3 dimer turned out to be unnaturally short (0.252 minutes). The error prone total STAT-data might here affect the result. Additional measurements of total cytoplasmic STAT3 concentrations could supposedly remedy this issue.

Moreover, we evaluated two competing biokinetic models for zirconium processing in the human body after ingestion for *in vivo* plasma and urine measurements. In order to obtain reliable Monte Carlo sampling results, we again combined the numerically stable thermodynamic integration, this time with CIMH. Based on individual Bayes factors for 16 investigations as well as a Bayes factor based on the concatenated dataset the HMGU model was unequivocally superior when compared to the current ICRP model. Also, when restricting the data on plasma and urine measurements only, we found that the HMGU model was clearly favored.

In contrast to the ICRP model, the HMGU model predicted a delayed accumulation of zirconium in bones. Furthermore, we showed that the HMGU model can be applied for retrospective dose assessment, where the initially ingested amount of zirconium can be reconstructed (including credible intervals) from *ex post* urine measurements. This provided estimates that facilitate the decision if measures have to be taken in case of accidental exposure. In future applications the HMGU model together with its posterior samples can readily be used as basis for dose assessment in internal dosimetry.

In this thesis we primarily focused on the issues of model selection and parameter inference in dynamic systems governed by ordinary or delay differential equations. Typically, a closed form solution to the differential equation system is unavailable in real world applications. The computationally expensive numerical solution for every likelihood evaluation calls for a sophisticated MCMC proposal generation scheme. However, the fields of application of CIMH and ACIMH is not limited to this scenario. Both vine copula based algorithms can be applied to any MCMC inference problem, such as Bayesian inference of ARMA or GARCH models used in economics and finance. They are expected to work well on highly dependent posterior distributions, but also very efficiently in simple systems.

Nevertheless, further research is needed to improve the algorithms for sampling from highly complex posterior distributions. A simple first step in this direction could be to apply automated cdf type detection for sample uniformization in each copula adaption step. Monitoring Markov chain convergence by means of convergence statistics could moreover lead to variable adaption of the proposal function rates r_1, r_2, and r_3. The finite copula update scheme of ACIMH might even be generalized to an infinite update scheme. Clearly, this requires a thorough proof of convergence.

CIMH and ACIMH can readily be extended to population MCMC sampling schemes. Much in the sense of thermodynamic integration and path sampling (Gelman & Meng [1998]), tempered MCMC approaches (Liu [2008]) might help to explore the sampling spaces more quickly. This would possibly produce apt copula decomposition of the posterior distribution in a much faster way.

The JAK2-STAT5 pathway analysis indicates that non-standard copula and marginal distributions might be needed to guarantee efficient sampling performances. Fitting non-parametric cdf's for sample uniformization as well as non-parametric pair copula distributions could be a further step to improve the efficiency of CIMH and ACIMH. However, as proposal generation might become computationally more expensive it has to be checked whether speed advantages with respect to ($\mathcal{I}_2$) would still be retained. A similar issue constitutes the choice of the copula decomposition: Although the order of the variables was rather canonical for our examples, introducing more general vine structures, such as R-vines, could further increase the sampling efficiency. Additional vine structure selection methods would however be needed in this case.

Appendix A

Important univariate density functions

The following univariate density functions are used for the copula based independence chain Metropolis-Hastings algorithm:

Normal density function

A random variable X is called *normally distributed*, if its density function is for $\mu \in \mathbb{R}$ and $\sigma > 0$ given by

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

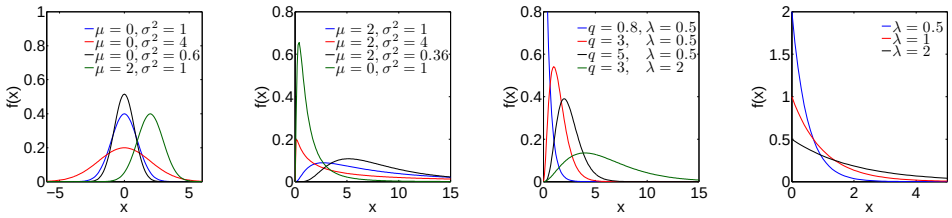
We write $X \sim \mathcal{N}(\mu, \sigma)$. A plot of the normal density function is shown in Figure A.1(a).

Lognormal density function

A random variable X is called *lognormally distributed*, if its density function is for $\mu \in \mathbb{R}$ and $\sigma > 0$ given by

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

where $\mathbb{1}_{(0,\infty)}(x)$ denotes the indicator function on $(0,\infty)$. We write $X \sim \mathcal{LN}(\mu, \sigma)$. A plot of the lognormal density function is shown in Figure A.1(b).



(a) (b) (c) (d)

Figure A.1: Various univariate (a) normal, (b) lognormal, (c) gamma, and (d) exponential density functions.

Gamma density function

A random variable X is called *gamma distributed*, if its density function is for $q, \lambda > 0$ given by

$$f(x) = \frac{\lambda^q}{\Gamma(q)} x^{q-1} \exp(-\lambda x) \mathbb{1}_{(0,\infty)}(x),$$

where $\mathbb{1}_{(0,\infty)}(x)$ denotes the indicator function on $(0, \infty)$ and

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

the Gamma function. We write $X \sim \Gamma(q, \lambda)$. A plot of the Gamma density function is shown in Figure A.1(c).

Exponential density function

A random variable X is called *exponentially distributed*, if its density function is for $\lambda > 0$ given by

$$f(x) = \lambda \exp(-\lambda x) \mathbb{1}_{(0,\infty)}(x),$$

where $\mathbb{1}_{(0,\infty)}(x)$ denotes the indicator function on $(0, \infty)$. We write $X \sim Exp(\lambda)$. A plot of the exponential density function is shown in Figure A.1(d).

Appendix B

Important bivariate copulas

The following bivariate copulas are used for the copula based independence chain Metropolis-Hastings algorithm:

Bivariate elliptical copula density functions

The copula density function of the bivariate *Gaussian copula* (c.f. Aas *et al.* [2009]) is given by

$$c(u_1, u_2|\rho) = \frac{1}{\sqrt{1-\rho^2}} \exp\left(\frac{\rho^2(x_1^2 + x_2^2) - 2\rho x_1 x_2}{2(1-\rho^2)}\right).$$

Here, $\rho \in (-1, 1)$ denotes the (correlation) parameter and $x_i = \Phi^{-1}(u_i)$ for the inverse $\Phi^{-1}(\cdot)$ of the standard normal distribution function. Figure 2.1(c) in Chapter 2.2 shows a plot of the Gaussian copula density function for $\rho = 0.5$.

The copula density function of the bivariate *Student's t copula* (c.f. Aas *et al.* [2009]) is for the copula parameters $\nu > 2$ and $\rho \in (-1, 1)$ given by

$$c(u_1, u_2|\rho, \nu) = \frac{\Gamma(\nu/2+1)/\Gamma(\nu/2)}{\nu\pi t_\nu^d(x_1) t_\nu^d(x_2)\sqrt{1-\rho^2}} \left(1 + \frac{x_1^2 + x_2^2 - 2\rho x_1 x_2}{\nu(1-\rho^2)}\right)^{-\nu/2-0.5}.$$

Here, $\Gamma(\cdot)$ denotes the Gamma function

$$\Gamma(x) = \int_0^\infty y^{x-1} e^{-y}\, \mathrm{d}y$$

and $x_i = t_\nu^{-1}(u_i)$ for the inverse t_ν^{-1} of the univariate standard student's t distribution function with ν degrees of freedom defined via the corresponding density function

$$t_\nu^d(x) = \frac{\Gamma(n/2+0.5)/\Gamma(n/2)}{\sqrt{\nu\pi}} \left(1 + \frac{x^2}{\nu}\right)^{-n/2-0.5}.$$

Figure B.1(a) shows a plot of the student's t copula density function for $\rho = 0.7$ and $\nu = 1$.

Bivariate Archimedean copulas

Archimedean copulas are defined via their generators (see Theorem 2.3). The following table holds a selection of the most prominent Archimedean copulas (c.f. Brechmann & Schepsmeier [2011]):

Name	Generator	Parameter(s)
Independence (I)	$-\log(t)$	–
Clayton (C)	$(t^{-\theta}-1)/\theta$	$\theta > 0$
Gumbel (G)	$(-\log(t))^\theta$	$\theta \geq 1$
Frank (F)	$-\log\left((\exp(-\theta t)-1)/(\exp(-\theta)-1)\right)$	$\theta \in \mathbb{R} \setminus \{0\}$
Joe (J)	$-\log\left(1-(1-t)^\theta\right)$	$\theta > 1$
BB1	$(t^{-\theta}-1)^\delta$	$\theta > 0\,, \delta \geq 1$
BB6	$(-\log(1-(1-t)^\theta))^\delta$	$\theta \geq 1,\ \delta \geq 1$
BB7	$(1-(1-t)^\theta)^{-\delta}-1$	$\theta \geq 1,\ \delta > 0$
BB8	$-\log\left((1-(1-\delta t)^\theta)((1-(1-\delta)^\theta)\right)$	$\theta \geq 1,\ \delta \in (0,1]$

Table B.1: A selection of Archimedean copulas.

Figure 2.1(a) shows a plot of the independence copula density function. All other copula density types are depicted in Figure B.1.

We furthermore get for each elliptical or Archimedean copula C the 90° , 180° and 270° rotated copulas C_{90}, C_{180}, and C_{270} by setting

$$\begin{aligned} C_{90}(u_1,u_2) &= u_2 - C(1-u_1,u_2),\\ C_{180}(u_1,u_2) &= u_1 + u_2 - 1 + C(1-u_1,1-u_2),\\ C_{270}(u_1,u_2) &= u_1 - C(u_1,1-u_2). \end{aligned}$$

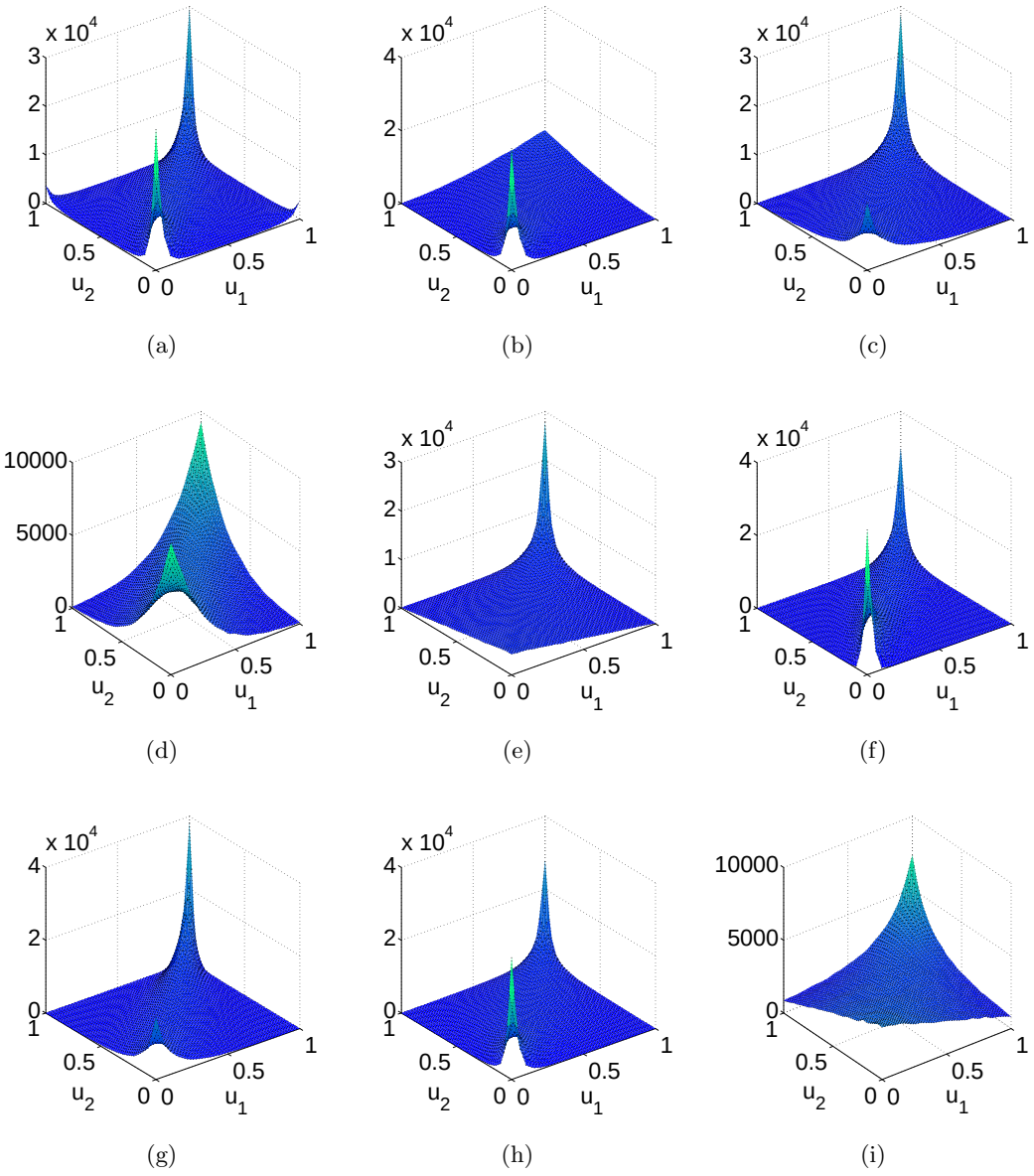
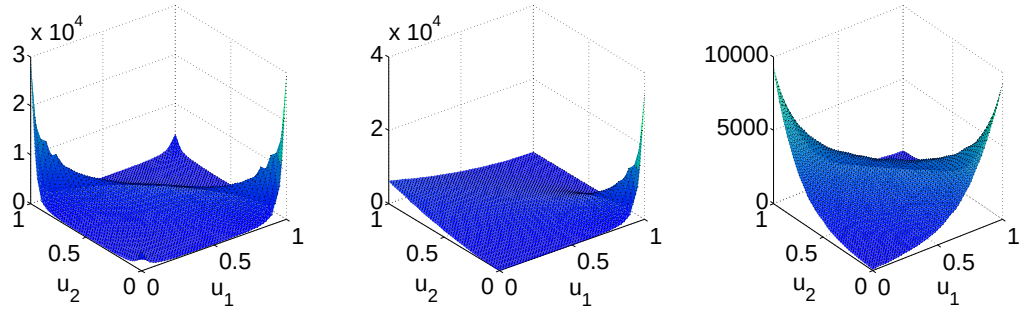


Figure B.1: Copula density functions for the (a) student's t (0.7,1), (b) Clayton (2), (c) Gumbel (2), (d) Frank (5), (e) Joe (2), (f) BB1 (2,2), (g) BB6 (2,2), (h) BB7 (2,2), and (i) BB8 (2,0.8) copula. The bracketed values denote the according parameter values in the order given in Table B.1.



(a) (b) (c)

Figure B.2: Copula density functions for the 90° rotated (a) student's t (0.7,1), (b) Clayton (2), and (c) Frank (5) copula. The bracketed values denote the according parameter values in the order given in Table B.1.

Note that some copula types coincide with some of their rotated versions due to reasons of symmetry. Figures B.2(a), B.2(b), and B.2(c) show a plot of the 90° rotated t, Clayton, and Frank copula density functions.

Appendix C

Calculations for the Bayes factor of the Gaussian mixture model

For our Gaussian mixture example from Chapter 3.4.4 we explicitly compute the power posterior at $t \in [0,1]$ as well as the marginal likelihoods. Subsequently, we exemplary infer the expected value of the log-likelihood with respect to the power posterior for the one component case. We use the notations introduced in Chapter 3.4.4.

C.1 Power posterior and marginal likelihood for the one-component Gaussian (mixture) model

Setting $\bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i$, the t-powered product of the likelihood times prior for $\mathcal{M}_1$ given the observations $\boldsymbol{y} = \{y_1, \ldots, y_m\}$ and the prior distribution $\mu \sim \mathcal{N}(0, \sigma^2)$ computes to

$$\begin{aligned}
\mathcal{L}(\mu|\boldsymbol{y})^t \pi(\mu|\mathcal{M}_1) &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{mt+1} \exp\left(-\frac{1}{2\sigma^2}\left(t\sum_{i=1}^{m} y_i^2 - 2\mu t\sum_{i=1}^{m} y_i + (mt+1)\mu^2\right)\right) \\
&= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{mt+1} \exp\left(\frac{1}{2\sigma^2}\left(\frac{(mt\bar{y})^2}{mt+1} - t\sum_{i=1}^{m} y_i^2\right)\right) \\
&\quad \cdot \exp\left(-\frac{mt+1}{2\sigma^2}\left(\frac{mt\bar{y}}{mt+1} - \mu\right)^2\right) \\
&\propto \exp\left(-\frac{mt+1}{2\sigma^2}\left(\mu - \frac{t\sum_{i=1}^{m} y_i}{mt+1}\right)^2\right),
\end{aligned}$$

where the proportionality is to be understood with respect to μ. The power posterior for $t \in [0,1]$ and $\mu \in \mathbb{R}$ is therefore given by $\mathcal{N}\left(\frac{t\sum_{i=1}^m y_i}{mt+1}, \frac{\sigma^2}{mt+1}\right)$. Furthermore, setting $t = 1$ and integrating over the real line yields the marginal likelihood

$$\begin{aligned}
\pi(\boldsymbol{y}|\mathcal{M}_1) &= \int_{\mathbb{R}} \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{m+1} \exp\left(\frac{1}{2\sigma^2}\left(\frac{m^2\bar{y}^2}{m+1} - \sum_{i=1}^m y_i^2\right)\right) \\
&\qquad \cdot \exp\left(-\frac{m+1}{2\sigma^2}\left(\frac{m\bar{y}}{m+1} - \mu\right)^2\right) \mathrm{d}\mu \\
&= \frac{1}{\sqrt{m+1}}\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^m \exp\left(\frac{1}{2\sigma^2}\left(\frac{m^2\bar{y}^2}{m+1} - \sum_{i=1}^m y_i^2\right)\right) \\
&\qquad \cdot \int_{\mathbb{R}} \frac{\sqrt{m+1}}{\sqrt{2\pi}\sigma} \exp\left(-\frac{m+1}{2\sigma^2}\left(\frac{m\bar{y}}{m+1} - \mu\right)^2\right) \mathrm{d}\mu \\
&= \frac{1}{\sqrt{m+1}}\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^m \exp\left(\frac{1}{2\sigma^2}\left(\frac{m^2\bar{y}^2}{m+1} - \sum_{i=1}^m y_i^2\right)\right).
\end{aligned}$$

C.2 Power posterior for the two-component Gaussian (mixture) model

Similarly to the one-component case, setting $\bar{y}_1 = \frac{1}{m_1}\sum_{i=1}^{m_1} y_i$ and $\bar{y}_2 = \frac{1}{m_2}\sum_{j=m_1+1}^{m} y_j$ for $m_2 := m - m_1$, the t-powered product of the likelihood times prior for the two-component model $\mathcal{M}_2$ given the observations $\boldsymbol{y} = \{y_1, \ldots, y_m, y_{m_1+1}, \ldots, y_m\}$ and the independent prior distributions $\mu_1 \sim \mathcal{N}(2, \sigma^2)$ and $\mu_2 \sim \mathcal{N}(-2, \sigma^2)$ computes to

$$\begin{aligned}
\mathcal{L}(\mu|\boldsymbol{y})^t \pi(\mu_1|\mathcal{M}_2)\pi(\mu_2|\mathcal{M}_2) &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{mt+2} \exp\Bigg(-\frac{1}{2\sigma^2}\Bigg(t\sum_{i=1}^m y_i^2 + 8 - 2(m_1 t\bar{y}_1 + 2)\mu_1 \\
&\qquad - 2(m_2 t\bar{y}_2 - 2)\mu_2 + (m_1 t + 1)\mu_1^2 + (m_2 t + 1)\mu_2^2\Bigg)\Bigg) \\
&\propto \exp\left(-\frac{m_1 t + 1}{2\sigma^2}\left(\mu_1 - \frac{2 + t\sum_{i=1}^{m_1} y_i}{m_1 t + 1}\right)^2\right) \\
&\qquad \cdot \exp\left(-\frac{m_2 t + 1}{2\sigma^2}\left(\mu_2 - \frac{-2 + t\sum_{j=m_1+1}^{m} y_j}{m_2 t + 1}\right)^2\right),
\end{aligned}$$

where the proportionality is to be understood with respect to μ_1 and μ_2. The power posterior for $t \in [0,1]$ and $(\mu_1, \mu_2)^\top \in \mathbb{R}^2$ is hence given by

$$\mathcal{N}_2\left(\begin{pmatrix} \frac{2+t\sum_{i=1}^{m_1} y_i}{m_1 t+1} \\ \frac{-2+t\sum_{j=m_1+1}^{m} y_j}{mt-m_1 t+1} \end{pmatrix}, \begin{pmatrix} \frac{\sigma^2}{m_1 t+1} & 0 \\ 0 & \frac{\sigma^2}{mt-m_1 t+1} \end{pmatrix}\right).$$

For $t = 1$ integration over $\mathbb{R}^2$ yields the marginal likelihood

$$\begin{aligned}
\pi(\boldsymbol{y}|\mathcal{M}_2) &= \int_{\mathbb{R}}\int_{\mathbb{R}} \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{m+2} \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{m} y_i^2 + 8 - 2(m_1\bar{y}_1 + 2)\mu_1\right.\right. \\
&\qquad \left.\left. - 2(m_2\bar{y}_2 - 2)\mu_2 + (m_1+1)\mu_1^2 + (m_2+1)\mu_2^2\right)\right) \mathrm{d}\mu_1 \mathrm{d}\mu_2 \\
&= \frac{1}{\sqrt{(m_1+1)(m_2+1)}} \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{m} \\
&\qquad \cdot \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{m} y_i^2 + 8 - \frac{(m_1\bar{y}_1+2)^2}{m_1+1} - \frac{(m_2\bar{y}_2-2)^2}{m_2+1}\right)\right).
\end{aligned}$$

C.3 Expected value of the log likelihood w.r.t. the power posterior for the one-component Gaussian (mixture) model

Exemplary we compute for the power posterior

$$\pi_t(\mu|\boldsymbol{y}, \mathcal{M}_1) = \frac{\mathcal{L}(\mu|\boldsymbol{y})^t \cdot \pi(\mu|\mathcal{M}_1)}{\pi_t(\boldsymbol{y}|\mathcal{M}_1)}$$

the normalizing constant

$$\begin{aligned}
\pi_t(\boldsymbol{y}|\mathcal{M}_1) &= \int_{\mathbb{R}} \mathcal{L}(\mu|\boldsymbol{y})^t \cdot \pi(\mu|\mathcal{M}_1)\,\mathrm{d}\mu \\
&= \int_{\mathbb{R}} \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{mt+1} \exp\left(-\frac{1}{2\sigma^2}\left(t\sum_{i=1}^{m} y_i^2 - 2m\bar{y}t\mu + (mt+1)\mu^2\right)\right) \mathrm{d}\mu \\
&= \frac{1}{\sqrt{mt+1}} \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{mt} \exp\left(-\frac{1}{2\sigma^2}\left(t\sum_{i=1}^{m} y_i^2 - \frac{(mt\bar{y})^2}{mt+1}\right)\right).
\end{aligned}$$

In order to compute the expectation of the data's log-likelihood, $\log \mathcal{L}(\mu|\boldsymbol{y}, \mathcal{M}_1)$, with respect to the power posterior $\pi_t(\mu|\boldsymbol{y}, \mathcal{M}_1)$, we first have to do the calculations of

C. CALCULATIONS FOR THE BAYES FACTOR OF THE GAUSSIAN MIXTURE MODEL

- $\int_{\mathbb{R}} A(\mu)\,\mathrm{d}\mu := \left(\sum_{i=1}^{m}(y_i^2 + 2\sigma^2\log(\sqrt{2\pi}\sigma))\right)\int_{\mathbb{R}}\exp\left(-\frac{mt+1}{2\sigma^2}\left(\frac{mt\bar{y}}{mt+1}-\mu\right)^2\right)\,\mathrm{d}\mu$
- $\int_{\mathbb{R}} B(\mu)\,\mathrm{d}\mu := -2m\bar{y}\int_{\mathbb{R}}\mu\exp\left(-\frac{mt+1}{2\sigma^2}\left(\frac{mt\bar{y}}{mt+1}-\mu\right)^2\right)d\mu$
- $\int_{\mathbb{R}} C(\mu)\,\mathrm{d}\mu := m\int_{\mathbb{R}}\mu^2\exp\left(-\frac{mt+1}{2\sigma^2}\left(\frac{mt\bar{y}}{mt+1}-\mu\right)^2\right)\,\mathrm{d}\mu.$

For $a := \frac{mt+1}{2\sigma^2}$ and $z := \frac{mt}{mt+1}\bar{y}$ the three integrals are

$$\begin{aligned}\int_{\mathbb{R}} A(\mu)\,\mathrm{d}\mu &= \left(\sum_{i=1}^{m}(y_i^2 + 2\sigma^2\log(\sqrt{2\pi}\sigma))\right)\int_{\mathbb{R}}\exp\left(-\frac{mt+1}{2\sigma^2}\left(\frac{mt\bar{y}}{mt+1}-\mu\right)^2\right)\,\mathrm{d}\mu\\ &= \left(\sum_{i=1}^{m}\left(y_i^2 + 2\sigma^2\log(\sqrt{2\pi}\sigma)\right)\right)\frac{\sqrt{2\pi}\sigma}{\sqrt{mt+1}},\end{aligned}$$

$$\begin{aligned}\int_{\mathbb{R}} B(\mu)\,\mathrm{d}\mu &= -2m\bar{y}\int_{\mathbb{R}}\mu\exp\left(-\frac{mt+1}{2\sigma^2}\left(\frac{mt\bar{y}}{mt+1}-\mu\right)^2\right)\,\mathrm{d}\mu\\ &= -2m\bar{y}\int_{\mathbb{R}}\mu\exp\left(-a\left(z-\mu\right)^2\right)\,\mathrm{d}\mu\\ &= -2m\bar{y}\int_{\mathbb{R}}(\mu-z)\exp\left(-a(z-\mu)^2\right) + z\exp\left(-a(z-\mu)^2\right)\,\mathrm{d}\mu\\ &= -2m\bar{y}\left(\int_{-\infty}^{\infty} y\exp(-ay^2)\,\mathrm{d}y + z\frac{\sqrt{2\pi}\sigma}{\sqrt{mt+1}}\right)\\ &= -2m\bar{y}z\frac{\sqrt{2\pi}\sigma}{\sqrt{mt+1}} = -2\bar{y}^2\frac{\sqrt{2\pi}\sigma m^2 t}{(mt+1)^{\frac{3}{2}}},\end{aligned}$$

and

$$
\begin{aligned}
\int_{\mathbb{R}} C(\mu)\,\mathrm{d}\mu &= m\int_{\mathbb{R}} \mu^2 \exp\left(-\frac{mt+1}{2\sigma^2}\left(\frac{mt\bar{y}}{mt+1}-\mu\right)^2\right)\,\mathrm{d}\mu \\
&= m\int_{\mathbb{R}} \mu^2 \exp\left(-a\,(z-\mu)^2\right)\,\mathrm{d}\mu \\
&= m\int_{\mathbb{R}} (\mu-z)^2 \exp\left(-a\,(z-\mu)^2\right) - z^2\exp\left(-a\,(z-\mu)^2\right) \\
&\qquad + 2z\mu\exp\left(-a\,(z-\mu)^2\right)\,\mathrm{d}\mu \\
&= m\int_{\mathbb{R}} y^2\exp\left(-ay^2\right)\,\mathrm{d}y - mz^2\int_{\mathbb{R}}\exp\left(-a\,(z-\mu)^2\right)\,\mathrm{d}\mu \\
&\qquad + 2mz\int_{\mathbb{R}}\mu\exp\left(-a\,(z-\mu)^2\right)\,\mathrm{d}\mu \\
&= 2m\int_0^{\infty} y^2\exp\left(-ay^2\right)\,\mathrm{d}y - m\left(\frac{mt}{mt+1}\bar{y}\right)^2\frac{\sqrt{2\pi}\sigma}{\sqrt{mt+1}} \\
&\qquad + 2mz\int_{\mathbb{R}}(\mu-z)\exp\left(-a\,(z-\mu)^2\right)\,\mathrm{d}\mu \\
&\qquad + \int_{\mathbb{R}} z\exp\left(-a\,(z-\mu)^2\right)\,\mathrm{d}\mu \\
&= 2m\frac{\Gamma(\frac{3}{2})}{2a^{\frac{3}{2}}} - \frac{m^3t^2\bar{y}^2\sqrt{2\pi}\sigma}{(mt+1)^{\frac{5}{2}}} + 2z^2m\frac{\sqrt{2\pi}\sigma}{\sqrt{mt+1}} \\
&= \frac{m\sqrt{2\pi}\sigma^3}{(mt+1)^{\frac{3}{2}}} + \frac{m^3t^2\bar{y}^2\sqrt{2\pi}\sigma}{(mt+1)^{\frac{5}{2}}}.
\end{aligned}
$$

Defining furthermore

$$
\begin{aligned}
\alpha(\boldsymbol{y}|t) &:= -\frac{1}{Z(\boldsymbol{y}|t)}\frac{1}{2\sigma^2}\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{mt+1}, \\
\beta(\boldsymbol{y}|t) &:= \alpha(\boldsymbol{y}|t)\exp\left(-\frac{1}{2\sigma^2}\left(t\sum_{i=1}^{m} y_i^2 - \frac{(mt\bar{y})^2}{mt+1}\right)\right) = -\frac{1}{2\sigma^2},
\end{aligned}
$$

we finally get:

C. CALCULATIONS FOR THE BAYES FACTOR OF THE GAUSSIAN MIXTURE MODEL

$$
\begin{aligned}
&\mathbb{E}_{\pi_t(\mu|\boldsymbol{y},\mathcal{M}_1)}\left(\log \mathcal{L}(\mu|\boldsymbol{y})\right) \\
&= \int_{\mathbb{R}} \log \mathcal{L}(\mu|\boldsymbol{y}) \frac{\mathcal{L}(\mu|\boldsymbol{y})^t \pi(\mu)}{Z(\boldsymbol{y}|t,\mathcal{M}_1)}\,\mathrm{d}\mu \\
&= \frac{1}{Z(\boldsymbol{y}|t,\mathcal{M}_1)} \int_{\mathbb{R}} \left(\sum_{i=1}^{m} \left(-\frac{1}{2\sigma^2}(y_i^2 - 2\mu y_i + \mu^2) \right) - m\log(\sqrt{2\pi}\sigma) \right) \\
&\qquad \cdot \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^{mt+1} \exp\left(-\frac{1}{2\sigma^2} \left(t\sum_{i=1}^{m} y_i^2 - 2m\bar{y}t\mu + (mt+1)\mu^2 \right) \right) \mathrm{d}\mu \\
&= \alpha(\boldsymbol{y}|t) \cdot \int_{\mathbb{R}} \left(\sum_{i=1}^{m} (y_i^2 - 2\mu y_i + \mu^2) + 2m\sigma^2 \log(\sqrt{2\pi}\sigma) \right) \\
&\qquad \cdot \exp\left(-\frac{1}{2\sigma^2} \left(t\sum_{i=1}^{m} y_i^2 - \frac{(mt\bar{y})^2}{mt+1} \right) \right) \cdot \exp\left(-\frac{mt+1}{2\sigma^2} \left(\frac{mt\bar{y}}{mt+1} - \mu \right)^2 \right) \mathrm{d}\mu \\
&= \beta(\boldsymbol{y}|t) \cdot \int_{\mathbb{R}} \left(\sum_{i=1}^{m} \left(y_i^2 - 2\mu y_i + \mu^2 + 2\sigma^2 \log(\sqrt{2\pi}\sigma) \right) \right) \\
&\qquad \cdot \exp\left(-\frac{mt+1}{2\sigma^2} \left(\frac{mt\bar{y}}{mt+1} - \mu \right)^2 \right) \mathrm{d}\mu \\
&= \beta(\boldsymbol{y}|t) \cdot \left(\int_{\mathbb{R}} A(\mu)\,\mathrm{d}\mu + \int_{\mathbb{R}} B(\mu)\,\mathrm{d}\mu + \int_{\mathbb{R}} C(\mu)\,\mathrm{d}\mu \right) \\
&\qquad \beta(\boldsymbol{y}|t) \Bigg\{ \left(\sum_{i=1}^{m} \left(y_i^2 + 2\sigma^2 \log(\sqrt{2\pi}\sigma) \right) \right) \cdot \frac{\sqrt{2\pi}\sigma}{\sqrt{mt+1}} \\
&\qquad - 2\bar{y}^2 \frac{\sqrt{2\pi}\sigma m^2 t}{(mt+1)^{\frac{3}{2}}} + \frac{m\sqrt{2\pi}\sigma^3}{(mt+1)^{\frac{3}{2}}} + \frac{m^3 t^2 \bar{y}^2 \sqrt{2\pi}\sigma}{(mt+1)^{\frac{5}{2}}} \Bigg\} \\
&= \beta(\boldsymbol{y}|t) \frac{\sqrt{2\pi}\sigma}{\sqrt{mt+1}} \Bigg\{ \left(\sum_{i=1}^{m} \left(y_i^2 + 2\sigma^2 \log(\sqrt{2\pi}\sigma) \right) \right) \\
&\qquad + \frac{m\sigma^2 - 2tm^2\bar{y}^2}{mt+1} + \frac{m^3 t^2 \bar{y}^2}{(mt+1)^2} \Bigg\} = \\
&= -\frac{1}{2\sigma^2} \Bigg\{ \left(\sum_{i=1}^{m} \left(y_i^2 + 2\sigma^2 \log(\sqrt{2\pi}\sigma) \right) \right) + \frac{m\sigma^2 - 2\bar{y}^2 m^2 t}{mt+1} + \frac{m^3 \bar{y}^2 t^2}{(mt+1)^2} \Bigg\}
\end{aligned}
$$

Appendix D

Transformation of the JAK2-STAT5 DDE system

In order to resolve structural identifiability issues of the non-linear JAK2-STAT5 DDE system

$$\begin{aligned}
\frac{\mathrm{d}x_1(t)}{\mathrm{d}t} &= -k_1 x_1(t) Epo(t) + 2k_4 x_3(t+\tau) \\
\frac{\mathrm{d}x_2(t)}{\mathrm{d}t} &= -k_2 x_2^2(t) + k_1 x_1(t) Epo(t) \\
\frac{\mathrm{d}x_3(t)}{\mathrm{d}t} &= -k_3 x_3(t) + \frac{1}{2} k_2 x_2^2(t) \\
\frac{\mathrm{d}x_4(t)}{\mathrm{d}t} &= -k_4 x_3(t+\tau) + k_3 x_3(t),
\end{aligned} \tag{D.1}$$

with observables

$$y_1(t) = k_5(x_2(t) + 2x_3(t)) \qquad \text{and} \qquad y_2(t) = k_6(x_1(t) + x_2(t) + 2x_3(t))$$

of Chapter 6.3.4 we follow Timmer *et al.* [2004] and define

$$\begin{aligned}
z_i(t) &= k_2 x_i(t) \quad \text{for } i = 1, \ldots, 4, \\
k_i' &= \frac{k_i}{k_2} \qquad\qquad \text{for } i = 5, 6.
\end{aligned} \tag{D.2}$$

The transformed DDE system then reads

$$\begin{aligned}
\frac{\mathrm{d}z_1(t)}{\mathrm{d}t} &= -k_1 z_1(t) Epo(t) + 2k_4 z_3(t+\tau) \\
\frac{\mathrm{d}z_2(t)}{\mathrm{d}t} &= -z_2^2(t) + k_1 z_1(t) Epo(t) \\
\frac{\mathrm{d}z_3(t)}{\mathrm{d}t} &= -k_3 z_3(t) + \frac{1}{2} z_2^2(t) \\
\frac{\mathrm{d}z_4(t)}{\mathrm{d}t} &= -k_4 z_3(t+\tau) + k_3 z_3(t),
\end{aligned} \tag{D.3}$$

with observables

$$y_1(t) = k_5'(z_2(t) + 2z_3(t)) \qquad \text{and} \qquad y_2(t) = k_6'(z_1(t) + z_2(t) + 2z_3(t)).$$

Note that this system is structurally identifiable (Timmer *et al.* [2004]) and has the same observables $y_1(t)$ and $y_2(t)$ as the original system (D.1). The posterior distribution with respect to (D.1) can hence be directly transformed into the posterior distribution with respect to (D.3). Since $x_1(0) = 1$ and $x_2(0) = x_2(0) = x_4(0) = 0$, the initial conditions for (D.3) are given by $z_1(0) = k_2$ and $z_2(0) = z_2(0) = z_4(0) = 0$, i.e. k_2 corresponds directly to the initial condition for $z_1(t)$. The transformed system is therefore parametrized by

$$\boldsymbol{\xi} = (k_1, k_2, k_3, k_4, \tau, k_5', k_6')^\top.$$

Inference of the JAK2-STAT5 parameters was done using (D.3). All estimated parameter marginal posterior means, modes, and 90% posterior quantile based credible intervals of Table 6.4 in Chapter 6.3.4 are given with respect to the original parameters $k_1, \ldots, k_4, \tau, k_5, k_6$ by application of the inverse transformation to (D.2).

Appendix E

Geometric tensor for the JAK2-STAT5 DDE system

For the observations $\boldsymbol{y} := \{y_1^\varepsilon(t_1), \ldots, y_1^\varepsilon(t_{16}), y_2^\varepsilon(t_1), \ldots, y_2^\varepsilon(t_{16})\}$ on the time grid $t_1, \ldots, t_{16}$ and the parameter vector $\boldsymbol{\xi} = (k_1, k_2, k_3, k_4, \tau, k_5', k_6')^\top$ we want to infer the posterior distribution

$$\pi(\boldsymbol{\xi}|\boldsymbol{y}) \propto \prod_{i=1}^{16} \Phi(y_1^\varepsilon(t_i)|y_1(t_i), \sigma_{i,1}^2) \cdot \Phi(y_2^\varepsilon(t_i)|y_2(t_i), \sigma_{i,2}^2) \cdot \pi(\boldsymbol{\xi}), \tag{E.1}$$

where Φ denotes the pdf of the univariate normal distribution for the known measurement errors $\sigma_{i,j}^2$, $i = 1, \ldots, 16$, $j = 1, 2$, and $\pi(\boldsymbol{\xi}) = \prod_{j \neq 3} \mathbf{1}_{[0,50]}(\xi_j) \cdot \pi(\xi_3|\xi_4)$ – recall $k_3 \geq k_4$. Furthermore, $y_1(t) = k_5'(z_2(t) + 2z_3(t))$ and $y_2(t) = k_6'(z_1(t) + z_2(t) + 2z_3(t))$ for the solutions $z_1(t)$, $z_2(t)$, and $z_3(t)$ of the DDE system (D.3) of Appendix D. The geometric tensor, as introduced in Chapter 5.1, is given by the expected Fisher information matrix of the log-likelihood minus the Hessian of the log-prior. It involves the partial derivatives of $y_1(t)$ and $y_2(t)$ with respect to the parameters $k_1, k_2, k_3, k_4, \tau, k_5'$ and k_6'. These, however, are not analytically tractable. We therefore (i) approximated the DDE system (D.3) by means of an ODE system and (ii) subsequently apply the so-called *sensitivity equations* to circumvent this issue.

(i) Approximation of the DDE system using the linear chain trick: Simply spoken the *linear chain trick* approximates the time delay of a DDE system by a sequence of linear

segments. More precisely, suppose we are given the differential equation with time delay τ

$$\frac{\mathrm{d}\boldsymbol{x}(t)}{\mathrm{d}t} = \boldsymbol{g}_{\boldsymbol{\xi}}(\boldsymbol{x}(t), \boldsymbol{x}(t-\tau), \boldsymbol{u}(t), t) \tag{E.2}$$

where $t \in \mathbb{R}$, $\boldsymbol{x} \in \mathbb{R}^r$ and – with respect to $\boldsymbol{x}(t)$ and $\boldsymbol{x}(t-\tau)$ – Lipschitz-continuous $\boldsymbol{\xi}$-parametrized function $\boldsymbol{g}_{\boldsymbol{\xi}}(\boldsymbol{x}(t), \boldsymbol{x}(t-\tau), \boldsymbol{u}(t), t)$ with $\boldsymbol{g}_{\boldsymbol{\xi}}(\mathbf{0}, \mathbf{0}, \mathbf{0}, t) = \mathbf{0}$ and external stimulus $\boldsymbol{u}(t)$. Then we can transform (E.2) into a corresponding ODE system using $m \in \mathbb{N}$ segments $\boldsymbol{x}_i(t)$:

$$\begin{aligned} \frac{\mathrm{d}\boldsymbol{x}(t)}{\mathrm{d}t} &= \boldsymbol{g}_{\boldsymbol{\xi}}(\boldsymbol{x}(t), \boldsymbol{x}_m(t), \boldsymbol{u}(t), t), \\ \frac{\mathrm{d}\boldsymbol{x}_i(t)}{\mathrm{d}t} &= \frac{m}{\tau}(\boldsymbol{x}_{i-1}(t) - \boldsymbol{x}_i(t)), \qquad i = 1, \ldots, m \end{aligned}$$

by exchanging the delayed elements $\boldsymbol{x}(t-\tau)$ with $\boldsymbol{x}_m(t)$ (Fall [2002]). In this sense we introduce two auxiliary functions $z_5(t)$ and $z_6(t)$ for the system (D.3) of Appendix D and define the ODE

$$\begin{aligned} \frac{\mathrm{d}z_1(t)}{\mathrm{d}t} &= -k_1 z_1(t) Epo(t) + 2k_4 z_6(t) \\ \frac{\mathrm{d}z_2(t)}{\mathrm{d}t} &= -z_2^2(t) + k_1 z_1(t) Epo(t) \\ \frac{\mathrm{d}z_3(t)}{\mathrm{d}t} &= -k_3 z_3(t) + \frac{1}{2} z_2^2(t) \\ \frac{\mathrm{d}z_4(t)}{\mathrm{d}t} &= -k_4 z_6(t) + k_3 z_3(t) \\ \frac{\mathrm{d}z_5(t)}{\mathrm{d}t} &= \frac{2}{\tau}(z_3(t) - z_5(t)) \\ \frac{\mathrm{d}z_6(t)}{\mathrm{d}t} &= \frac{2}{\tau}(z_5(t) - z_6(t)), \end{aligned} \tag{E.3}$$

with $z_5(0) = z_6(0) = 0$. Nikolov *et al.* [2007] discuss the asymptotic stability of (E.3) and show that the solutions of the original DDE and the approximating ODE model are very close. We can thus use the much simpler differential equation (E.3) to compute the geometric tensor for the JAK2-STAT5 system. Note that the approximation error does not effect the validity but only the efficiency of the posterior inference process in the SMALA algorithm.

(ii) Sensitivity equations: Suppose we are given an arbitrary $\boldsymbol{\xi}$-parametrized ODE system

$$\frac{\mathrm{d}\boldsymbol{x}(t)}{\mathrm{d}t} = \boldsymbol{g}_{\boldsymbol{\xi}}(\boldsymbol{x}(t), \boldsymbol{u}(t), t) \tag{E.4}$$

with initial conditions $\boldsymbol{x}(0) = \boldsymbol{x}_0$. Using the chain rule the derivative with respect to $\boldsymbol{\xi}$ and $\boldsymbol{x}(0)$ yields the following *sensitivity equations* for $\boldsymbol{S}(t) = \frac{\partial \boldsymbol{x}(t)}{\partial \boldsymbol{\xi}}$ and $\boldsymbol{S}_0(t) = \frac{\partial \boldsymbol{x}(t)}{\partial \boldsymbol{x}_0}$:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{S}(t) &= \frac{\partial \boldsymbol{g}_{\boldsymbol{\xi}}(\boldsymbol{x}(t), \boldsymbol{u}(t), t)}{\partial \boldsymbol{x}(t)}\boldsymbol{S}(t) + \frac{\partial \boldsymbol{g}_{\boldsymbol{\xi}}(\boldsymbol{x}(t), \boldsymbol{u}(t), t)}{\partial \boldsymbol{\xi}} \qquad \text{with } \boldsymbol{S}(0) = \boldsymbol{0} \\
\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{S}_0(t) &= \frac{\partial \boldsymbol{g}_{\boldsymbol{\xi}}(\boldsymbol{x}(t), \boldsymbol{u}(t), t)}{\partial \boldsymbol{x}(t)}\boldsymbol{S}_0(t) \qquad \text{with } \boldsymbol{S}_0(0) = \boldsymbol{I}_m
\end{aligned}$$

where $\boldsymbol{I}_m$ is the m-dimensional identity matrix (Wu *et al.* [2008]). This means we can numerically solve the extended differential equations system

$$\begin{aligned}
\frac{\mathrm{d}\boldsymbol{x}(t)}{\mathrm{d}t} &= \boldsymbol{g}_{\boldsymbol{\xi}}(\boldsymbol{x}(t), \boldsymbol{u}(t), t) \\
\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{S}(t) &= \frac{\partial \boldsymbol{g}_{\boldsymbol{\xi}}(\boldsymbol{x}(t), \boldsymbol{u}(t), t)}{\partial \boldsymbol{x}(t)}\boldsymbol{S}(t) + \frac{\partial \boldsymbol{g}_{\boldsymbol{\xi}}(\boldsymbol{x}(t), \boldsymbol{u}(t), t)}{\partial \boldsymbol{\xi}} \\
\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{S}_0(t) &= \frac{\partial \boldsymbol{g}_{\boldsymbol{\xi}}(\boldsymbol{x}(t), \boldsymbol{u}(t), t)}{\partial \boldsymbol{x}(t)}\boldsymbol{S}_0(t)
\end{aligned} \tag{E.5}$$

in order to obtain the solutions for $\boldsymbol{S}(t)$ and $\boldsymbol{S}_0(t)$.

This eventually allows us to compute the geometric tensor for the JAK2-STAT5 system: For $\boldsymbol{z}(t) = (z_1(t), \ldots, z_6(t))^\top$ and $\boldsymbol{\xi} = (k_1, k_2, k_3, k_4, \tau, k_5', k_6')^\top$ let

$$\frac{\mathrm{d}\boldsymbol{z}(t)}{\mathrm{d}t} = \boldsymbol{g}_{\boldsymbol{\xi}}(\boldsymbol{z}(t), Epo(t))$$

be the ODE system defined by (E.3) with $y_1(t) = k_5'(z_2(t) + 2z_3(t))$ and $y_2(t) = k_6'(z_1(t) + z_2(t) + 2z_3(t))$ intrinsically dependent on the parameter vector $\boldsymbol{\xi}$. Using the covariance matrices $\boldsymbol{\Sigma}_1 = \mathrm{diag}(\sigma_{1,1}^2, \ldots, \sigma_{16,1}^2)$, $\boldsymbol{\Sigma}_2 = \mathrm{diag}(\sigma_{1,2}^2, \ldots, \sigma_{16,2}^2)$ along with the vectors $\boldsymbol{v}_1 = (y_1^\varepsilon(t_1) - y_1(t_1), \ldots, y_1^\varepsilon(t_{16}) - y_1(t_{16}))^\top$, $\boldsymbol{v}_2 = (y_2^\varepsilon(t_1) - y_2(t_1), \ldots, y_2^\varepsilon(t_{16}) - y_2(t_{16}))^\top$ we can rewrite the posterior (E.1) as

$$\pi(\boldsymbol{\xi}|\boldsymbol{y}) \propto \prod_{i=1}^{2} \Phi_{16}(\boldsymbol{v}_i|\boldsymbol{\Sigma}_i) \cdot \pi(\boldsymbol{\xi}),$$

for the pdf's $\Phi_{16}(\cdot|\boldsymbol{\Sigma}_i)$ of the 16 dimensional normal distribution $\mathcal{N}_{16}(\boldsymbol{0}, \boldsymbol{\Sigma}_i)$. Defining $\boldsymbol{v}_j^i := \frac{\partial \boldsymbol{v}_j}{\partial \xi_i}$ the partial derivative of the log-likelihood $\log(\mathcal{L}(\boldsymbol{\xi}|\boldsymbol{y})) = \log\left(\prod_{i=1}^{2} \Phi_{16}(\boldsymbol{v}_i|\boldsymbol{\Sigma}_i)\right)$

with respect to the i^{th} parameter ξ_i computes to

$$\begin{aligned}\frac{\partial}{\partial\xi_i}\log(\mathcal{L}(\boldsymbol{\xi}|\boldsymbol{y})) &= \frac{\partial}{\partial\xi_i}\sum_{j=1}^{2}\left(-\frac{1}{2}\boldsymbol{v}_j^\top\boldsymbol{\Sigma}_j^{-1}\boldsymbol{v}_j\right)\\ &= \sum_{j=1}^{2}\left(-\frac{1}{2}(\boldsymbol{v}_j^i)^\top\boldsymbol{\Sigma}_j^{-1}\boldsymbol{v}_j - \frac{1}{2}\boldsymbol{v}_j^\top\boldsymbol{\Sigma}_j^{-1}\boldsymbol{v}_j^i\right)\\ &= -\sum_{j=1}^{n}(\boldsymbol{v}_j^i)^\top\boldsymbol{\Sigma}_j^{-1}\boldsymbol{v}_j\end{aligned}$$

It is easy to see that $(\boldsymbol{v}_j^i)^\top\boldsymbol{\Sigma}_j^{-1}\boldsymbol{v}_j = \boldsymbol{v}_j^\top\boldsymbol{\Sigma}_j^{-1}\boldsymbol{v}_j^i$. According to Chapter 5.1 the $(i,j)^{th}$ element $(i,j=1,\ldots,7)$ of the geometric tensor is given as

$$\begin{aligned}\boldsymbol{G}_{i,j}(\boldsymbol{\xi}) &= \operatorname{cov}\left[\frac{\partial}{\partial\xi_i}\log(\mathcal{L}(\boldsymbol{\xi}|\boldsymbol{y}))^\top, \frac{\partial}{\partial\xi_j}\log(\mathcal{L}(\boldsymbol{\xi}|\boldsymbol{y}))^\top\right] - \frac{\partial^2}{\partial\xi_i\partial\xi_j}\log(\pi(\boldsymbol{\xi}))\\ &= \operatorname{cov}\left[\sum_{k=1}^{2}(\boldsymbol{v}_k^i)^\top\boldsymbol{\Sigma}_k^{-1}\boldsymbol{v}_k, \sum_{l=1}^{2}\boldsymbol{v}_l^\top\boldsymbol{\Sigma}_l^{-1}\boldsymbol{v}_l^j\right] - \frac{\partial^2}{\partial\xi_i\partial\xi_j}\log(\pi(\boldsymbol{\xi}))\\ &= \sum_{k=1}^{2}(\boldsymbol{v}_k^i)^\top\boldsymbol{\Sigma}_k^{-1}\boldsymbol{v}_k^j - \frac{\partial^2}{\partial\xi_i\partial\xi_j}\log(\pi(\boldsymbol{\xi}))\end{aligned} \tag{E.6}$$

(see Girolami & Calderhead [2011]).

The k^{th} element of $\boldsymbol{v}_j^i$ is $\frac{\partial y_j(t_k)}{\partial\xi_i}$, where $i=1,\ldots,7$ is the parameter index, $j=1,2$ the species index, $k=1,\ldots,16$ the time index, and $y_j(t)$ the solution to the ODE system (E.3). Setting $v_j^i(t) := \frac{\partial y_j(t)}{\partial\xi_i}$ we now have for $i=1,\ldots,7$

$$\begin{aligned}\frac{\mathrm{d}v_1^i(t)}{\mathrm{d}t} &= \sum_{l=1}^{6}\left(\left(\frac{\partial}{\partial z_l(t)}\frac{\mathrm{d}y_1(t)}{\mathrm{d}t}\right)\frac{\partial z_l(t)}{\partial\xi_i}\right) + \frac{\partial}{\partial\xi_i}\frac{\mathrm{d}y_1(t)}{\mathrm{d}t}\\ &= \sum_{l=1}^{6}\left(\frac{\partial k_5'(k_1z_1(t)Epo(t) - 2k_3z_3(t))}{\partial z_l(t)}\frac{\partial z_l(t)}{\partial\xi_i}\right)\\ &\qquad + \frac{\partial k_5'(k_1z_1(t)Epo(t) - 2k_3z_3(t))}{\partial\xi_i}\\ \frac{\mathrm{d}v_2^i(t)}{\mathrm{d}t} &= \sum_{l=1}^{6}\left(\left(\frac{\partial}{\partial z_l(t)}\frac{\mathrm{d}y_2(t)}{\mathrm{d}t}\right)\frac{\partial z_l(t)}{\partial\xi_i}\right) + \frac{\partial}{\partial\xi_i}\frac{\mathrm{d}y_2(t)}{\mathrm{d}t}\\ &= \sum_{l=1}^{6}\left(\frac{\partial k_6'(2k_4z_6(t) - 2k_3z_3(t))}{\partial z_l(t)}\frac{\partial z_l(t)}{\partial\xi_i}\right) + \frac{\partial k_6'(2k_4z_6(t) - 2k_3z_3(t))}{\partial\xi_i}.\end{aligned} \tag{E.7}$$

with $v_1^i(0)=0$ for all i, $v_2^i(0)=0$ for $i=1,3,4,5,6$, $v_2^2(0)=k_6'$, and $v_2^7(0)=k_2$ which can be seen by straightforward application of the definition of $y_1(t)$ and $y_2(t)$. The

expressions $\frac{\partial z_l(t)}{\partial \xi_i}$ can be computed via the sensitivity equations

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial z_l(t)}{\partial \xi_i} &= \sum_{m=1}^{6}\left(\left(\frac{\partial \boldsymbol{g}_{\boldsymbol{\xi}}(\boldsymbol{z}(t), Epo(t))}{\partial z_m(t)}\right)\frac{\partial z_m(t)}{\partial \xi_i}\right) \\
&\qquad + \frac{\partial \boldsymbol{g}_{\boldsymbol{\xi}}(\boldsymbol{z}(t), Epo(t))}{\partial \xi_i(t)} \qquad \text{for } i = 1, 3, \ldots, 7,\ l = 1, \ldots, 6 \\
\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial z_l(t)}{\partial \xi_2} &= \sum_{m=1}^{6}\left(\left(\frac{\partial \boldsymbol{g}_{\boldsymbol{\xi}}(\boldsymbol{z}(t), Epo(t))}{\partial z_m(t)}\right)\frac{\partial z_m(t)}{\partial \xi_2}\right) \qquad \text{for } l = 1, \ldots, 6 \\
\frac{\partial z_2(0)}{\partial \xi_2} &= \frac{\partial z_l(0)}{\partial k_2} = 1 \\
\frac{\partial z_l(0)}{\partial \xi_i} &= 0 \qquad \text{for } i = 1, \ldots, 7,\ l = 1, \ldots, 6,\ (i,j) \neq (2,2).
\end{aligned} \tag{E.8}$$

Hence, $\boldsymbol{v}_j^i$ can be numerically computed by means of the extended ODE system (E.3), (E.7), and (E.8). We used Matlab's `ode15s` solver for the solution of the extended ODE system, which is faster than the `dde23` solver applied for the RWMH, IMH, CovRWMH, M-GaA, CIMH, and ACIMH algorithms. SMALA might thus have a slight advantage in speed when solving the differential equation systems (D.1).

For the second term of the right hand side of Equation (E.6) we have due to

$$k_1, k_2, k_4, \tau, k_5, k_6 \overset{i.i.d.}{\sim} \mathcal{U}[0, 50] \text{ and } k_3 \sim \mathcal{U}[k_4, 50] \text{ that}$$

$$\frac{\partial^2}{\partial \xi_i \partial \xi_j}\log(\pi(\boldsymbol{\xi})) = 0 \qquad \text{for all } i, j \text{ except } i = j = 4.$$

On the other hand

$$\begin{aligned}
\frac{\partial^2}{\partial k_4 \partial k_4}\log(\pi(\boldsymbol{\xi})) &= -6\log(50)\frac{\partial^2}{\partial k_4 \partial k_4}\log(\pi(k_3|k_4)) \\
&= -6\log(50)\frac{\partial^2}{\partial k_4 \partial k_4}\log\left(\frac{1}{50 - k_4}\right) \\
&= -6\log(50)\frac{\partial}{\partial k_4}\frac{1}{50 - k_4} \\
&= 6\log(50)\frac{\partial}{\partial k_4}\frac{1}{(50 - k_4)^2}
\end{aligned}$$

E. GEOMETRIC TENSOR FOR THE JAK2-STAT5 DDE SYSTEM

Appendix F

Parameters for prior distributions of the zirconium models

The prior distributions of the HMGU and ICRP model were computed in Li *et al.* [2011a]. Since estimated confidence intervals as well as estimated medians were provided only, we need to infer the location (μ) and scale (σ) parameters for all lognormal distributions $\mathcal{LN}(\mu, \sigma)$. Furthermore, the means μ and standard deviations σ for all normal distributions $\mathcal{N}(\mu, \sigma)$ need to be computed. The formulas are derived in the following.

Location and scale parameters for the lognormal distribution given the estimated median and geometric standard deviation

Suppose we are given the estimation $\hat{m}$ of the median m of a univariate lognormal distribution $\mathcal{LN}(\mu, \sigma)$. According to Johnson *et al.* [1994], $m = \exp(\mu)$ and therefore

$$\mu = \log(m) \approx \log(\hat{m}).$$

Furthermore, the geometric standard deviation (GSD) is provided for all lognormally distributed parameters and is either $GSD = 2$ or $GSD = 3$. Since we have $GSD =$

$\exp(\sigma)$ in case of a lognormal distribution, this naturally yields for the scale parameter:

$$\sigma = \ln(2) \text{ or } \sigma = \ln(3)$$

Mean and standard deviation for the normal distribution given the estimated median and the coefficient of variation

Since for the normal distribution $\mathcal{N}(\mu, \sigma)$ the mean and median coincide, we do not further distinguish between the two of them and simply denote their estimates by $\hat{\mu}$. Clearly, we have

$$\mu \approx \hat{\mu}.$$

Also, for the normally distributed parameters, we are given a coefficient of variation of $c_V = 0.3$. Since $c_V = \frac{\sigma}{\mu}$, we obtain

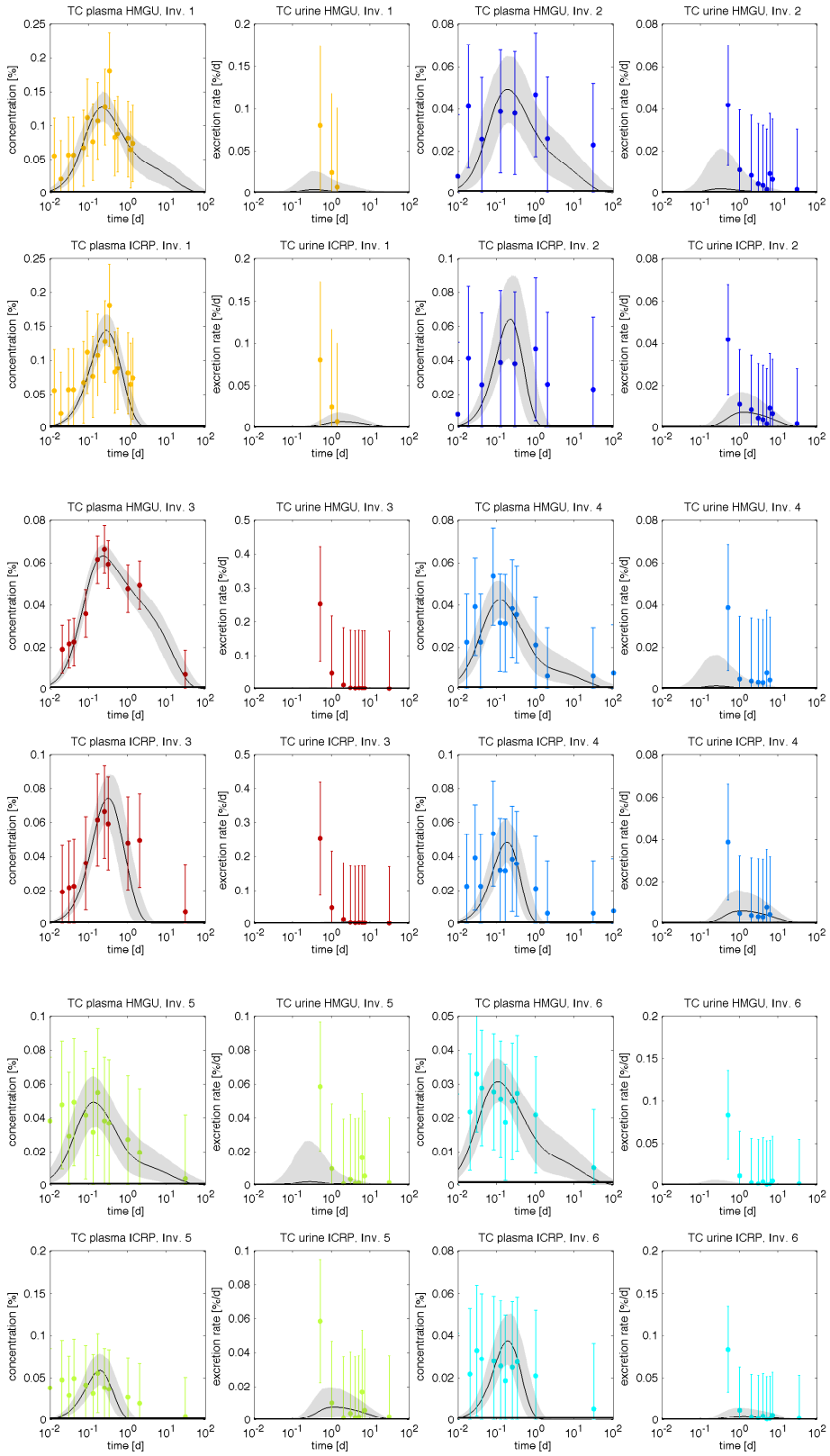
$$\sigma = 0.3\mu.$$

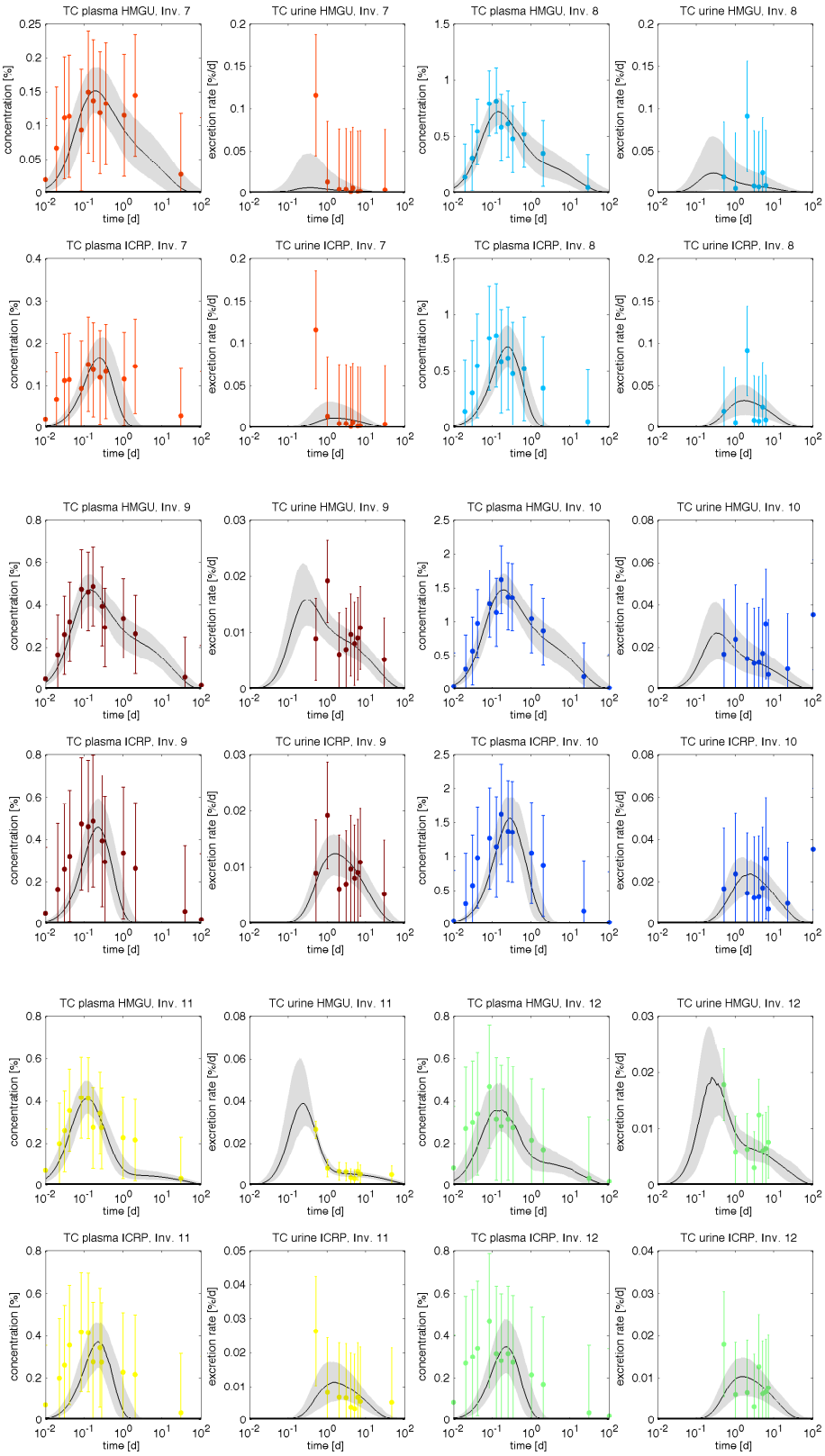
Appendix G

Investigation specific time courses for the ICRP and HMGU models

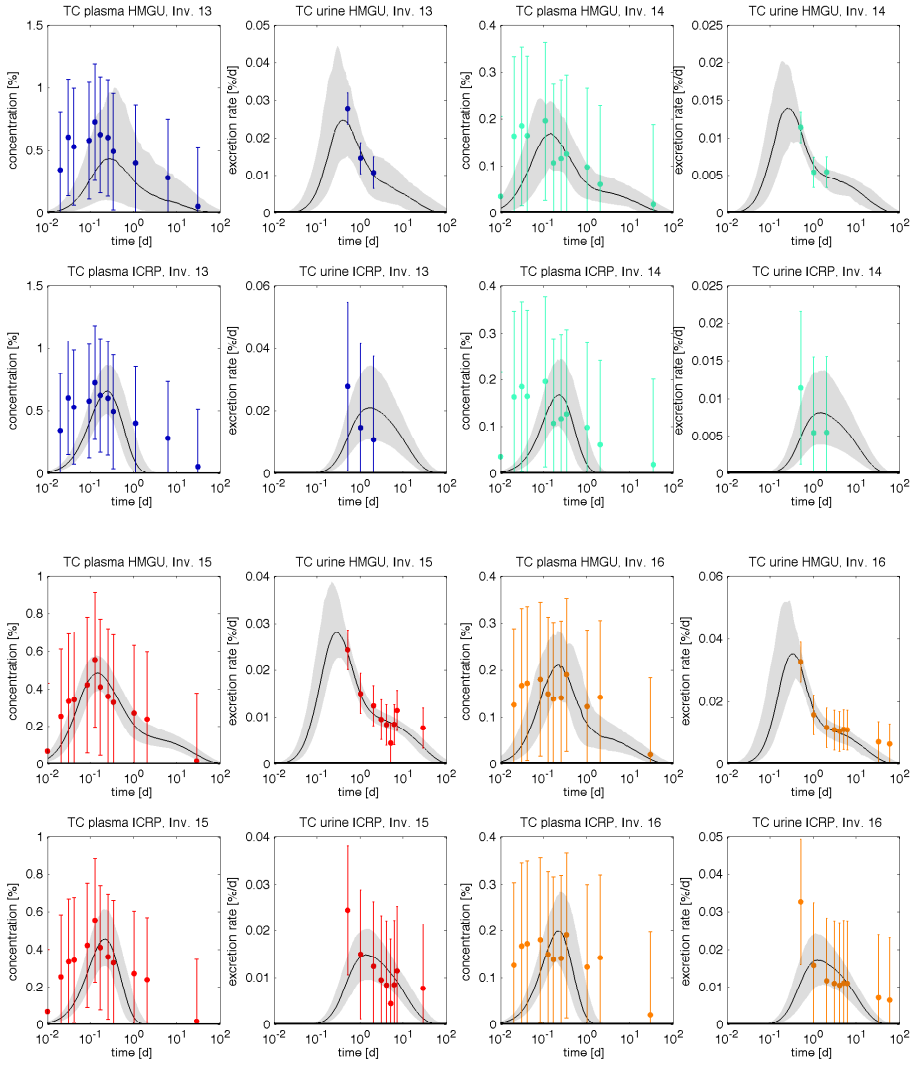
We here present the investigation specific time courses (TC) based on the posterior samples. Depicted are the time courses for the transfer compartment and for the excretion rate in the urine compartment together with the corresponding data. The black lines are the posterior median solutions of the time courses, while the shaded areas denote the 90 % posterior credible intervals. In addition, the investigation specific zero-truncated measurement errors, fitted by simulated annealing, are depicted. Note that neither the upper nor the lower confidence bound, nor the median needs to represent a solution to the according ordinary differential equation. While plasma time courses are generally covered well by both models, especially the ICRP model struggles from time to time with the urinary data. The coloring corresponds to the coloring of the individuals in Figure 8.2 and Figure 8.5.

G. INVESTIGATION SPECIFIC TIME COURSES FOR THE ICRP AND HMGU MODELS




TC plasma HMGU, Inv. 7
TC urine HMGU, Inv. 7
TC plasma HMGU, Inv. 8
TC urine HMGU, Inv. 8
TC plasma ICRP, Inv. 7
TC urine ICRP, Inv. 7
TC plasma ICRP, Inv. 8
TC urine ICRP, Inv. 8
TC plasma HMGU, Inv. 9
TC urine HMGU, Inv. 9
TC plasma HMGU, Inv. 10
TC urine HMGU, Inv. 10
TC plasma ICRP, Inv. 9
TC urine ICRP, Inv. 9
TC plasma ICRP, Inv. 10
TC urine ICRP, Inv. 10
TC plasma HMGU, Inv. 11
TC urine HMGU, Inv. 11
TC plasma HMGU, Inv. 12
TC urine HMGU, Inv. 12
TC plasma ICRP, Inv. 11
TC urine ICRP, Inv. 11
TC plasma ICRP, Inv. 12
TC urine ICRP, Inv. 12
concentration [%]
excretion rate [%/d]
time [d]

G. INVESTIGATION SPECIFIC TIME COURSES FOR THE ICRP AND HMGU MODELS



References

Aaronson, D. & Horvath, C. (2002). A road map for those who don't know JAK-STAT. *Science*, **296**, 1653 – 1655. 38, 119, 129

Aas, K., Czado, C., Frigessi, A. & Bakken, H. (2009). Pair-copula constructions of multiple dependence. *Insurance, Mathematics and Economics*, **44**, 182–198. 5, 19, 21, 22, 24, 95, 99, 106, 163

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. & Walter, P. (2002). *Molecular Biology of the Cell*, vol. 4. Garland Science. 36, 139

Amari, S. & Nagaoka, H. (2007). *Methods of information geometry*, vol. 191. American Mathematical Society. 87

Anderson, R. & May, R. (1992). *Infectious diseases of humans: dynamics and control*, vol. 26. Wiley Online Library. 45

Arbenz, P. (2011). Bayesian copulae distributions, with application to operational risk management – some comments. *Methodology and Computing in Applied Probability*, 1–4. 19

Bartlett, M. (1966). *An introduction to stochastic processes*. Cambridge University Press. 76

Bartolucci, F., Scaccia, L. & Mira, A. (2006). Efficient Bayes factor estimation from the reversible jump output. *Biometrika*, **93**, 41–52. 82

Bedford, T. & Cooke, R. (2001). Probability density decomposition for conditionally dependent random variables modeled by vines. *Annals of Mathematics and Artificial Intelligence*, **32**, 245–268. 22

Bedford, T. & Cooke, R. (2002). Vines – a new graphical model for dependent random variables. *Annals of Statistics*, **30**, 1031–1068. 22

Beichl, I. & Sullivan, F. (2000). The Metropolis algorithm. *Computing in Science & Engineering*, **2**, 65–69. 2, 69

Berger, J. (1985). *Statistical decision theory and Bayesian analysis*. Springer. 52

REFERENCES

Bernardo, J., Smith, A. & Berliner, M. (1994). *Bayesian theory*, vol. 62. Wiley. 52, 55

BMU (2007). *Richtlinie für die physikalische Strahlenschutzkontrolle zur Ermittlung der Körperdosis. Teil 2: Ermittlung der Körperdosis bei innerer Strahlenexposition (Inkorporationsüberwachung) (§§40, 41 und 42 StrlSchV)*. Bonn: Bundesministerium für Umwelt, Naturschutz und Reaktorsicherheit. 153

Bohl, S. (2009). *Dynamic modeling of signal processing for IL-6-induced STAT3 signal transduction in primary mouse hepatocytes*. Ph.D. thesis, Ruperto-Carola University of Heidelberg, Germany. 128, 129

Bonizzi, G. & Karin, M. (2004). The two nf-κb activation pathways and their role in innate and adaptive immunity. *Trends in immunology*, **25**, 280–288. 127

Bornholdt, S. (2008). Boolean network models of cellular regulation: prospects and limitations. *Journal of the Royal Society Interface*, **5**, S85–S94. 43

Brechmann, E. (2010). Truncated and simplified regular vines and their applications, diploma thesis, Technische Universität München, Germany. 99

Brechmann, E. & Schepsmeier, U. (2011). Dependence modeling with C- and D-vine copulas: The R-package CDVine, *Preprint* (*http://www-m4.ma.tum.de/Papers/index.html*). 99, 106, 164

Brooks, S. (1998). Markov chain Monte Carlo method and its application. *Journal of the Royal Statistical Society: Series D (The Statistician)*, **47**, 69–100. 2

Brooks, S. & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 434–455. 79

Brown, K. & Sethna, J. (2003). Statistical mechanical approaches to models with many poorly known parameters. *Physical Review E*, **68**, 021904–1–021904–9. 1

Calderhead, B. & Girolami, M. (2009). Estimating Bayes factors via thermodynamic integration and population MCMC. *Computational Statistics & Data Analysis*, **53**, 4028–4045. 63, 65, 66, 133

Cao, Y., Li, H. & Petzold, L. (2004). Efficient formulation of the stochastic simulation algorithm for chemically reacting systems. *The journal of chemical physics*, **121**, 4059. 41

Casella, G. & Berger, R. (2001). *Statistical inference*. Duxbury Press. 56

Černỳ, V. (1985). Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *Journal of Optimization Theory and Applications*, **45**, 41–51. 47

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. 43

Chung, K. (1982). *Lectures from Markov processes to Brownian motion*. Springer-Verlag. 87

Czado, C. (2010). Pair-copula constructions of multivariate copulas. In P. Jaworki, F. Durante, W. Härdle & T. Rychlik, eds., *Copula Theory and its Applications*, 93–110, Springer-Verlag. 99

D., K.Z.Y.P., Shaw, B., Kou, B., McAuley, K. & Bacon, D. (2003). Modeling ethylene/butene copolymerization with multi-site catalysts: parameter estimability and experimental design. *Polymer Reaction Engineering*, **11**, 563–588. 50

Dargatz, C. (2010). *Bayesian Inference for Diffusion Processes with Applications in Life Sciences*. Ph.D. thesis, Ludwig-Maximilians-Universität München, Germany. 28, 29, 42

Davison, A. (2003). *Statistical models*, vol. 11. Cambridge University Press. 121

De Jong, H. (2002). Modeling and simulation of genetic regulatory systems: a literature review. *Journal of Computational Biology*, **9**, 67–103. 38, 42

Diestel, R. (2000). Graph theory (graduate texts in mathematics vol 173). 23

Dissmann, J., Brechmann, E., Czado, C. & Kurowicka, D. (2011). Selecting and estimating regular vine copulae and application to financial returns. *Preprint*. 106

Duane, S., Kennedy, A., Pendleton, B. & Roweth, D. (1987). Hybrid Monte Carlo. *Physics letters B*, **195**, 216–222. 3

Eidgenössisches Nuklearsicherheitsinspektorat Informationsdienst (2011). *Radiologische Auswirkungen aus den kerntechnischen Unfällen in Fukushima vom 11.3.2011*. Eidgenössisches Nuklearsicherheitsinspektorat Informationsdienst. 137

Evett, I. (1991). Implementing bayesian methods in forensic science. In *Fourth Valencia International Meeting on Bayesian Statistics*. 59

Fall, C. (2002). *Computational cell biology*, vol. 20. Springer-Verlag. 176

Fearnhead, P. (2008). Editorial: Special issue on adaptive Monte Carlo methods. *Statistics and Computing*, **18**, 341–342. 101

Fletcher, R. (1987). *Practical methods of optimization, volume 1*. Wiley. 47

Fraser, A. & Burnell, D. (1970). *Computer models in genetics*. McGraw-Hill Book Company. 47

Friel, N. & Pettitt, A. (2008). Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**, 589–607. 60, 63, 65, 66

Gamerman, D. & Lopes, H. (2006). *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*, vol. 68. Chapman & Hall/CRC. 2, 60

REFERENCES

Gelman, A. & Meng, X. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, 163–185. 63, 159

Gelman, A. & Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457–472. 79

Gelman, A., Roberts, G. & Gilks, W. (1996). Efficient Metropolis jumping rules. *Bayesian statistics*, **5**, 599–608. 78

Geweke, J. (1992). *Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments*. Oxford University Press. 79

Geyer, C. (1992). Practical Markov chain Monte Carlo. *Statistical Science*, **7**, 473–483. 80

Geyer, C. & Møller, J. (1994). Simulation procedures and likelihood inference for spatial point processes. *Scandinavian Journal of Statistics*, 359–373. 80

Gibson, M. & Bruck, J. (2000). Efficient exact stochastic simulation of chemical systems with many species and many channels. *The journal of physical chemistry A*, **104**, 1876–1889. 41

Gilks, W., Roberts, G. & Sahu, S. (1998). Adaptive Markov chain Monte Carlo through regeneration. *Journal of the American Statistical Association*, 1045–1054. 3, 89

Gillespie, D. (1992). A rigorous derivation of the chemical master equation. *Physica A: Statistical Mechanics and its Applications*, **188**, 404–425. 40, 41, 42

Gillespie, D. (2007). Stochastic simulation of chemical kinetics. *Annual Review of Physical Chemistry*, **58**, 35–55. 39

Girolami, M. & Calderhead, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**, 123–214. 3, 5, 86, 88, 103, 178

Goodwin, B. (1963). *Temporal organization in cells*. Academic Press New York. 43

Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732. 4, 80, 81

Greiter, M., Giussani, A., Höllriegl, V., Li, W. & Oeh, U. (2011). 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. 5, 138, 139

Grenander, U. & Miller, M. (1994). Representations of knowledge in complex systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 549–603. 85

Guyton, A. & Hall, J. (2006). *Textbook of Medical Physiology*, vol. 11. Elsevier Saunders. 137

Haario, H., Saksman, E. & Tamminen, J. (1999). Adaptive proposal distribution for random walk Metropolis algorithm. *Computational Statistics*, **14**, 375–396. 3, 90

Haario, H., Saksman, E. & Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli*, 223–242. 3, 90, 92

Hastie, T., Tibshirani, R. & Friedman, J.J.H. (2009). *The elements of statistical learning*. Springer-Verlag. 80, 93

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109. 2, 69, 70

Hengl, S., Kreutz, C., Timmer, J. & Maiwald, T. (2007). Data-based identifiability analysis of non-linear dynamical models. *Bioinformatics*, **23**, 2612. 50

Hobæk Haff, I. (2010). Parameter estimation for pair-copula constructions. *Preprint*. 99

Hobæk Haff, I., Aas, K. & Frigessi, A. (2010). On the simplified pair-copula construction–simply useful or too simplistic? *Journal of Multivariate Analysis*, **101**, 1296–1310. 21

Holden, L. (2000). Convergence of Markov chains in the relative supremum norm. *Journal of Applied Probability*, **37**, 1074–1083. 97

Holden, L., Hauge, R. & Holden, M. (2009). Adaptive independent Metropolis-Hastings. *The Annals of Applied Probability*, **19**, 395–413. 3, 89, 98

Horbelt, W., Timmer, J. & Voss, H. (2002). Parameter estimation in nonlinear delayed feedback systems from noisy data. *Physics Letters A*, **299**, 513–521. 1, 47

Hou, S., Zheng, Z., Chen, X. & Perrimon, N. (2002). The JAK/STAT pathway in model organisms: Emerging roles in cell movement. *Developmental Cell*, **3**, 765–778. 119

Hu, L. (2006). Dependence patterns across financial markets: a mixed copula approach. *Applied Financial Economics*, **16**, 717–729. 126

ICRP (1975). ICRP publication 23. Report on the task group on reference man. *Annals of the ICRP*. 137, 138, 139

ICRP (1979). ICRP publication 30. Limits for intakes of radionuclides by workers (part 1). *Annals of the ICRP*, **8**. 137

ICRP (1988). ICRP publication 53. Radiation dose to patients from radiopharmaceuticals. *Annals of the ICRP*, **18**. 137

ICRP (1989). ICRP publication 56. Age-dependent doses to members of the public from intake of radionuclides (part 1: Ingestion dose coefficients). *Annals of the ICRP*, **20**. 137, 138, 139

ICRP (1993). ICRP publication 67. Age-dependent doses to members of the public from intake of radionuclides (part 2: Ingestion dose coefficients). *Annals of the ICRP*, **23**. 139

REFERENCES

ICRP (1998). ICRP publication 78. Individual monitoring for internal exposure of workers. *Annals of the ICRP*, **27**. 137, 152

ICRP (2007). ICRP publication 103. The 2007 Recommendations of the International Commission on Radiological Protection. *Annals of the ICRP*, **37**, 2. 137

ICRP (2008). ICRP publication 107. Nuclear decay data for dosimetric calculations. *Annals of the ICRP*, **38**. 146, 154

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. 37

Jacquez, J. (1985). *Compartmental analysis in biology and medicine*. University of Michigan Press Ann Arbor, MI. 45

Jefferys, W. & Berger, J. (1992). Ockham's razor and Bayesian analysis. *American Scientist*, **80**, 64–72. 60

Jeffreys, H. (1961). *Theory of probability*. Clarendon Press. 59

Joe, H. (1996). Families of m-variate distributions with given margins and m(m-1)/2 bivariate dependence parameters. In L. Rüschendorf and B. Schweizer and M. D. Taylor, ed., *Distributions with Fixed Marginals and Related Topics*, vol. 28. 19, 21

Joe, H. (1997). *Multivariate Models and Dependence Concepts*. Chapman & Hall. 16

Joe, H., Li, H. & Nikoloulopoulos, A. (2010). Tail dependence functions and vine copulas. *Journal of Multivariate Analysis*, **101**, 252–270. 22

Johnson, N., Kotz, S. & Balakrishnan, N. (1994). Continuous univariate distributions, vol. 1. 181

Kaplan, S., Bren, A., Dekel, E. & Alon, U. (2008). The incoherent feed-forward loop can generate non-monotonic input functions for genes. *Molecular Systems Biology*, **4**, 1–9. 46

Kass, R. (1993). Probabilistic inference using Markov chain Monte Carlo methods. *Technical Report*. 76

Kass, R. & Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, 773–795. 57, 58, 59, 60

Kass, R., Carlin, B., Gelman, A. & Neal, R. (1998). Markov chain Monte Carlo in practice: A roundtable discussion. *American Statistician*, **52**, 93–100. 76

Kauffman, S. (1969). Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of Theoretical Biology*, **22**, 437–467. 43

Kim, G., Silvapulle, M. & Silvapulle, P. (2007). Comparison of semiparametric and parametric methods for estimating copulas. *Computational Statistics & Data Analysis*, **51**, 2836–2850. 126

Kirkpatrick, S., Gelatt, C.D. & Vecchi, M.P. (1983). Optimization by simulated annealing. *Science*, **220**, 671–680. 4, 47, 84

Klenke, A. (2008). *Wahrscheinlichkeitstheorie*. Springer-Verlag. 9

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. 37

Kullback, S. & Leibler, R. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, **22**, 79–86. 65

Kurowicka, D. & Cooke, R. (2006a). Completion problem with partial correlation vines. *Linear Algebra and its Applications*, **418**, 188–200. 5

Kurowicka, D. & Cooke, R. (2006b). *Uncertainty Analysis with High Dimensional Dependence Modelling*. Wiley. 24

Kurowicka, D. & Joe, H. (2011). *Dependence Modeling - Handbook on Vine Copulae*. World Scientific Publishing Co. 5, 99

Lartillot, N. & Philippe, H. (2006). Computing Bayes factors using thermodynamic integration. *Systematic biology*, **55**, 195–207. 60, 63

Lawrence, N., Girolami, M., Rattray, M. & Sanguinetti, G. (2010). *Learning and Inference in Computational Systems Biology*. The MIT Press. 1

Lecourtier, Y., Lamnabhi-Lagarrigue, F. & Walter, E. (1987). Volterra and generating power series approaches to identifiability testing. *Identifiability of parametric models*, 50–66. 50

Lewis, S. (1994). *Multilevel modeling of discrete event history data using Markov chain Monte Carlo methods*. Ph.D. thesis, University of Washington, USA. 60

Li, S., Brazhnik, P., Sobral, B. & Tyson, J. (2008). A quantitative study of the division cycle of Caulobacter crescentus stalked cells. *PLoS Comput Biol*, **4**, e9. 43

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. 115, 138, 143, 144, 145, 154, 181

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**, 677. 115, 138, 143, 147

Liu, J. (2008). *Monte Carlo Strategies in Scientific Computing*. Springer-Verlag. 85, 103, 159

Ljung, L. & Glad, T. (1994). On global identifiability for arbitrary model parametrizations. *Automatica*, **30**, 265–276. 50

Lodewyckx, T., Kim, W., Lee, M., Tuerlinckx, F., Kuppens, P. & Wagenmakers, E. (2011). A tutorial on Bayes factor estimation with the product space method. *Journal of Mathematical Psychology*, **55**. 60

Maiwald, T. & Timmer, J. (2008). Dynamical modeling and multi-experiment fitting with potterswheel. *Bioinformatics*, **24**, 2037–2043. 47, 49

Marin, J. & Robert, C. (2007). *Bayesian core: a practical approach to computational Bayesian statistics*. Springer-Verlag. 54, 55

Meeker, W. & Escobar, L. (1995). Teaching about approximate confidence regions based on maximum likelihood estimation. *American Statistician*, 48–53. 48, 56

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., Teller, E. *et al.* (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, **21**, 1087–1092. 2, 69, 70

Meyn, S., Tweedie, R., Glynn, P. & Corporation, E. (1996). *Markov chains and stochastic stability*. Springer-Verlag. 26, 29

Min, A. & Czado, C. (2010). Bayesian inference for multivariate copulas using pair-copula constructions. *Journal of Financial Econometrics*, **8(4)**, 511–546. 99

Min, A. & Czado, C. (2011). Bayesian model selection for multivariate copulas using pair-copula constructions. *Canadian Journal of Statistics*, **39**, 239–258. 99

Mitra, D., Romeo, F. & Sangiovanni-Vincentelli, A. (1986). Convergence and finite-time behavior of simulated annealing. *Advances in Applied Probability*, 747–771. 84

Morales-Nápoles, O., Cooke, R. & Kurowicka, D. (2010). About the number of vines and regular vines on n nodes. *Submitted for publication*. 24

Müller, C. & Sbalzarini, I. (2010). Gaussian Adaptation as a unifying framework for continuous black-box optimization and adaptive Monte Carlo sampling. In *Evolutionary Computation (CEC), 2010 IEEE Congress on*, 1–8, IEEE. 3, 5, 90, 91

Murphy, S. & Van der Vaart, A. (2000). On profile likelihood. *Journal of the American Statistical Association*, 449–465. 48

Myung, I. & Pitt, M. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, **4**, 79–95. 60

Neal, R. (1993). Probabilistic inference using Markov chain Monte Carlo methods. Tech. Rep. CRG-TR-93-1, University of Toronto. Department of Computer Science. 76

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`. 61

Nelsen, R. (2006). *An Introduction to Copulas*. Springer. 16, 17, 18

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. 60, 61, 62

Nikolov, S., Kotev, V. & Petrov, V. (2007). Stability analysis of a time delay model for the JAK-STAT signaling pathway. *Series on Biomechanics*, **23**, 52–65. 176

Nummelin, E. (2004). *General irreducible Markov chains and non-negative operators*, vol. 83. Cambridge University Press. 33, 71, 90

Øksendal, B. (2003). *Stochastic differential equations: an introduction with applications*. Springer-Verlag. 28, 42

Palsson, B. (2006). *Systems Biology: Properties of Reconstructed Networks*. Cambridge University Press. 38

Papoulis, A., Pillai, S. & Unnikrishna, S. (1965). *Probability, random variables, and stochastic processes*, vol. 196. McGraw-Hill. 74

Pitt, M., Myung, I. & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, **109**, 472. 60

Press, W., Flannery, B., Teukolsky, S., Vetterling, W. *et al.* (1986). *Numerical recipes*, vol. 547. Cambridge Univ Press. 48

Raia, V., Schilling, M., Böhm, M., Hahn, B., Kowarsch, A., Raue, A., Sticht, C., Bohl, S., Saile, M., Möller, P. *et al.* (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. 43

Ramsay, J., Hooker, G., Campbell, D. & Cao, J. (2007). Parameter estimation for differential equations: a generalized smoothing approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69**, 741–796. 3

Rao, C. (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bulletin Calcutta Mathematical Society*, **37**, 81–91. 86

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. 48, 121

REFERENCES

Revuz, D. (1984). *Markov chains*, vol. 11. North Holland. 34

Ripley, B. (1977). Modelling spatial patterns. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 172–212. 80

Robert, C. & Casella, G. (2004). *Monte Carlo statistical methods*. Springer-Verlag. 26, 30, 51, 55, 71, 82, 84

Roberts, G. & Rosenthal, J. (2007). Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. *Journal of Applied Probability*, **44**, 458–475. 3, 88, 90, 99

Roberts, G. & Stramer, O. (2002). Langevin diffusions and Metropolis-Hastings algorithms. *Methodology and computing in applied probability*, **4**, 337–357. 3, 86

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. 103, 104

Rosenthal, J. (2011). *Optimal proposal distributions and adaptive MCMC*. Chapman & Hall/CRC Press. 3, 88

Schervish, M.J. (1995). *Theory of statistics*. Springer-Verlag. 87

Schilling, M., Maiwald, T., Bohl, S., Kollmann, M., Kreutz, C., Timmer, J. & Klingmüller, U. (2005). Computational processing and error reduction strategies for standardized quantitative data in biological networks. *FEBS Journal*, **272**, 6400–6411. 128

Schmidl, D., Czado, C. & Theis, F. (2012a). A vine copula based adaptive MCMC sampler for efficient inference of dynamical systems. *Bayesian Analysis*, **under revision**. 5

Schmidl, D., Hug, S., Li, W., Greiter, M. & Theis, F. (2012b). Bayesian model selection validates a biokinetic model for zirconium processing in humans. *BMC Systems Biology*, **6**, 95. 6

Serban, R. & Hindmarsh, A. (2005). Cvodes: the sensitivity-enabled ode solver in Sundials. In *Proceedings of IDETC/CIE*, vol. 24. 45

Sethuraman, J., Athreya, K. & Ddoss, H. (1992). A Proof of convergence of the Markov Chain simulation Method. *Technical Report 868*. 34

Shampine, L. & Reichelt, M. (1997). The Matlab ode suite. *SIAM journal on scientific computing*, **18**, 1–22. 44

Shampine, L. & Thompson, S. (2001). Solving DDEs in Matlab. *Applied Numerical Mathematics*, **37**, 441–458. 45

Sklar, A. (1959). Fonctions dé repartition á n dimensions et leurs marges. *Publ. Inst. Stat. Univ. Paris*, **8**, 229–231. 16

Smith, M., Min, A., Almeida, C. & Czado, C. (2010). Modeling longitudinal data using a pair-copula construction decomposition of serial dependence. *Journal of the American Statistical Association*, **105**, 1467–1479. 99

Spiegelhalter, D. & Smith, A. (1982). Bayes factors for linear and log-linear models with vague prior information. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 377–387. 60

Subramaniam, P., Torres, B. & Johnson, H. (2001). So many ligands, so few transcription factors: a new paradigm for signaling through the STAT transcription factors. *Cytokine*, **15**, 175–187. 37

Suykens, J. & Vandewalle, J. (2002). Coupled local minimizers: alternative formulations and extensions. In *Neural Networks, 2002. IJCNN'02. Proceedings of the 2002 International Joint Conference on*, vol. 3, 2039–2043, IEEE. 47

Suykens, J., Vandewalle, J. & De Moor, B. (2002). Intelligence and cooperative search by coupled local minimizers. *Arxiv Preprint cs/0210030*. 47

Swameye, I., Müller, T., Timmer, J., 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 of the United States of America*, **100**, 1028–1033. 102, 111, 120, 121, 125, 157

Ter Braak, C. & Vrugt, J. (2008). Differential evolution Markov chain with snooker updater and fewer chains. *Statistics and Computing*, **18**, 435–446. 3

Theis, F. (2002). *Mathematics in Independent Component Analysis*. Ph.D. thesis, University of Regensburg, Germany. 9

Thomas, R. (1991). Regulatory networks seen as asynchronous automata: a logical description. *Journal of Theoretical Biology*, **153**, 1–23. 43

Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics*, 1701–1728. 26, 30, 33, 34

Timmer, J., Müller, T., Swameye, I., Sandra, O. & Klingmüller, U. (2004). Modeling the nonlinear dynamics of cellular signal transduction. *International Journal of Bifurcation and Chaos*, **14**, 2069–2079. 121, 129, 173, 174

UNSCEAR (2008). *Sources and Effects of Ionizing Radiation*, vol. 2. United Nations Publications. 137

Venzon, D. & Moolgavkar, S. (1988). A method for computing profile-likelihood-based confidence intervals. *Applied Statistics*, 87–94. 48

Vuong, Q. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, 307–333. 57

REFERENCES

Wen, Z., Zhong, Z. & Darnell, J. (1995). Maximal activation of transcription by Stat1 and Stat3 requires both tyrosine and serine phosphorylation. *Cell*, **82**, 241–250. 38, 127, 131

Wilkinson, D. (2006). *Stochastic Modelling for Systems Biology*. Chapman & Hall/CRC. 1, 26, 39, 69

Wilkinson, D. (2007). Bayesian methods in Bioinformatics and computational Systems Biology. *Briefings in Bioinformatics*, **8**, 109–116. 2

Wu, W., Wang, F. & Chang, M. (2008). Dynamic sensitivity analysis of biological systems. *BMC Bioinformatics*, **9**, S17. 177

Index

INDEX