

TECHNISCHE UNIVERSITÄT MÜNCHEN

Development and Evaluation of Sampling-based
Parameter Estimation Methods for Dynamic
Biological Processes

Benjamin Ballnus

April 2019

TECHNISCHE UNIVERSITÄT MÜNCHEN

Fakultät für Mathematik — Lehrstuhl M12 (Mathematische Modellierung
biologischer Systeme)

**Development and Evaluation of
Sampling-based Parameter Estimation
Methods for Dynamic Biological
Processes**

Benjamin Ballnus

Vollständiger Abdruck der von der Fakultät für Mathematik — Lehrstuhl M12
(Mathematische Modellierung biologischer Systeme) der Technischen Univer-
sität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzende:

Prof. Claudia Czado, Ph.D.

Prüfer(-innen) der Dissertation:

1. Prof. Dr. Dr. Fabian J. Theis
2. Prof. Dr. Christina Kuttler
3. Prof. Dr. Jörg Stelling

Die Dissertation wurde am 18.04.2019 bei der Technischen Universität München
eingereicht und durch die Fakultät für Mathematik am 30.07.2019 angenommen.

Acknowledgment

I would like to thank

Linus and Fabian, who created the framework for this work and made the exceptional collaboration between Bayer and the Helmholtz Center Munich possible for me.

Jan for his high durability, patience and dedication without which this work could not have come about.

Fabian, Thomas, Steffen and Kathrin for the fun and educational time at Bayer.

Sabine for the many things I have learned from her.

Caro, Maren, Phil, Lisa, Valle, Susi, Sabrina, Norbert, Atefeh, Julia, Fabi, David, Hans, Adrinana, Lisa, Hannah, Christian, Lena, Vanessa, Pavel, Paul, Leonard, Elba, Aaron, Michi, Anna, Turid, Jessi, Dennis, Sascha, Katrin and all the other great fellows who have accompanied me on my way. Our fun was my fuel.

My wonderful family, especially my parents, who have prepared me for this journey and have unceasingly supported it.

Romina, my north star, which has allowed me to walk the path to the very end.

Abstract

In systems biology, scientific hypotheses can rarely be tested by directly measuring the quantities of interest, e.g. certain protein concentrations in living cells. Instead, quantities which can be measured directly, e.g. certain biomarkers, are exploited to calibrate the parameters of complex mathematical models. Mathematical models are formalized simplifications of reality and comparing their output with experimental data can support falsifications of scientific hypotheses. For example, this procedure often allows to derive latent variables of the biological system and facilitates predictions of untested experimental conditions.

However, the process of calibrating the parameters of a mathematical model given experimental data, called parameter estimation, is in general a challenging task. The performance of algorithms suited for this task is known to be highly problem dependent and the sparsity of biological data causes challenges as well.

Sampling algorithms such as Markov chain Monte Carlo methods are commonly employed for parameter estimation. In comparison to optimization algorithms, Markov chain Monte Carlo methods allow to train models and to assess uncertainty of the result at the same time. Uncertainty analysis is often required in order to estimate how reliable conclusions drawn from model-based data analysis are.

The optimal choice of a particular Markov chain Monte Carlo algorithm and its tuning parameters is unknown for a given problem while these choices vastly impact the computation time the parameter estimation requires. This is problematic as dynamical models in computational biology are often expensive in computation time per evaluation while Markov chain Monte Carlo algorithms require a large amount of model evaluations. Unfortunately, Markov chain Monte Carlo performance has never been quantified thoroughly between multiple algorithms and within multiple realistic use-cases for parameter estimation arising in biological sciences.

In this thesis, a comprehensive benchmark study about Markov chain Monte Carlo is implemented and presented based on a novel, fair analysis framework. The lessons learned during this evaluation are then used to construct a novel Markov chain Monte Carlo method. The novel algorithm combines the strengths of multiple existing algorithms including parallel tempering and adaptive Metropolis while dividing the parameter space into regions with simple probability density structure. The novel method is shown to be capable of outperforming state-of-the-art methods while maintaining strong self-tuning capabilities. This effort enables an straightforward, efficient and robust application of state-of-the-art Markov chain Monte Carlo across many biological applications.

Overall, these contributions aid the future usage and development of Markov chain Monte Carlo in systems biology, a scientific discipline which often requires strict assessment of uncertainty. The benchmark study closes an important gap in the literature as Markov chain Monte Carlo methods are typically selected uninformed for a given application. Multi-purpose self-tuning algorithms as presented in this thesis can help to increase overall Markov chain Monte Carlo performances across the computational biology community, creating a higher perception of the overall Markov chain Monte Carlo usability in practice.

List of contributed articles

- (i) B. Ballnus, S.Hug, K. Hatz, L. Görlitz, J. Hasenauer, F. J. Theis. **Comprehensive benchmarking of Markov chain Monte Carlo methods for dynamical systems.** BMC Systems Biology 11(1):63 (2017).
- (ii) B. Ballnus, S. Schaper, F. J. Theis, J. Hasenauer. **Bayesian parameter estimation for biochemical reaction networks using region-based adaptive parallel tempering.** Bioinformatics 34(13):i494–i501 (2018).

I am the sole first author and author in charge of all articles listed above. On top of these, I contributed to the following articles:

- (i) P. Stapor, D. Weindl, B. Ballnus, S. Hug, C. Loos, A. Fiedler, S. Krause, S. Hross, F. Fröhlich and J. Hasenauer **PESTO: Parameter ESTimation TOolbox.** Bioinformatics 34(4):705–707 (2017).
- (ii) C. Moore, D. Liu, B. Ballnus, M. Karbach and G. Müller **Disks in a narrow channel jammed by gravity and centrifuge: profiles of pressure, mass density and entropy density.** Journal of Statistical Mechanics: Theory and Experiment 2014(4):P04008 (2014).

Contents

1	Introduction	15
1.1	Motivation & research overview	15
1.2	Contribution of the thesis	19
1.3	Outline of the thesis	21
2	Background	23
2.1	Bayesian parameter estimation methods	23
2.1.1	Bayesian statistics	23
2.1.2	Assessment of high dimensional probability distributions	24
2.1.3	Monte Carlo principle	25
2.1.4	Classic Monte Carlo methods	26
2.1.5	The need for Markov chain Monte Carlo	27
2.1.6	Other Monte Carlo methods	28
2.2	Parameter estimation in ordinary differential equations	29
2.3	Biochemical reaction networks	31
2.4	Markov chain Monte Carlo	32
2.4.1	Theory of Markov chains in MCMC	33
2.4.2	Finite MCMC samples	40
2.4.3	MCMC methods	41
2.5	Uncertainty quantification	47
3	Quantification of sampling quality	51
3.1	Introduction and problem statement	52
3.2	Symptoms of MCMC failure	52

3.2.1	Burn-in	53
3.2.2	High auto-correlation	54
3.2.3	Insufficient exploration	56
3.3	Identification of MCMC failure	56
3.3.1	Calculation of burn-in	57
3.3.2	Assessment of exploration quality	60
3.3.3	Conditional effective sample size	62
3.3.4	Computation time	63
3.4	Example analysis: Multiple posterior modes	63
3.5	Summary and discussion	66
4	Benchmarking of MCMC in ODE-constraint models	69
4.1	Introduction and problem statement	70
4.2	Initialization	71
4.3	Benchmark collection	75
4.3.1	(M1) mRNA transfection	75
4.3.2	(M2) bistable switch	75
4.3.3	(M3) saturated growth	76
4.3.4	(M4) biochemical system with Hopf bifurcation	76
4.3.5	(M5) driven Van Der Pol Oscillator	76
4.3.6	(M6) Lorenz Attractor	76
4.3.7	Data and priors	77
4.4	Implementation	77
4.5	Performance of MCMC algorithms	77

<i>CONTENTS</i>	13
4.5.1 Revealing challenges in small sized application problem	79
4.5.2 Boosting sampling efficiency by resolving non-identifiabilities	79
4.5.3 Influence of posterior properties on sampling performance	81
4.5.4 Comparison of single- and multi-chain methods	82
4.5.5 Comparison of initialization strategies	84
4.5.6 Selection of tuning parameters and algorithmic settings	85
4.6 Summary and discussion	86
5 Region-based adaptive parallel tempering	91
5.1 Introduction and problem statement	92
5.2 Region-based adaptive parallel tempering	93
5.2.1 Motivation	93
5.2.2 Method	94
5.2.3 Implementation	98
5.2.4 Ergodicity	102
5.3 Performance benchmark	106
5.3.1 Artificial problems	107
5.3.2 Application problems	111
5.4 Aimed temperature selection	117
5.4.1 Motivation	117
5.4.2 Method	118
5.4.3 Evaluation	120
5.5 Summary and discussion	121
6 Conclusion	123

6.1	Summary and conclusions	123
6.2	Outlook	125
6.2.1	Further improvements of sample analysis	125
6.2.2	Parameter estimation and mathematical modeling	127
6.2.3	Further hybrid MCMC methods	129
6.2.4	Model based data integration	129

Chapter 1

Introduction

1.1 Motivation & research overview

In the field of computational systems biology, mechanistic models are developed to explain experimental data, to gain a quantitative understanding of processes and to predict the process dynamics under new experimental conditions (Gábor and Banga, 2015; Klipp et al., 2005b; Kitano, 2002). The parameters of these mechanistic models are typically unknown and need to be estimated from available experimental data. The parameter estimation may provide novel insights of the biological processes and facilitates flexible data integration.

Introducing example: *The utility of parameter estimation can be demonstrated using a simple example of estimating the gravity acceleration on earth, g , from the observation of a falling apple (see Figure 1.1a). In basic Newtonian mechanics, the trajectory of a non-relativistic apple with mass m and position $x(t)$ at time t is modeled by the Newtonian law,*

$$\ddot{x}(t) = g, \tag{1.1}$$

where gravity acceleration g is pretended to be unknown for the purpose of this example and must be estimated from experimental data. The trajectory of the apple is given by the analytical solution of 1.1 which reads

$$x(t) = gt^2 + v_0t + x_0, \tag{1.2}$$

with the initial velocity $v_0 = 0$ and the initial position $x_0 = 40$ for an apple just starting to fall at $t = 0$ from a tree of 40 meters height. In this example, one can not measure g directly. Instead, one obtains the apple position with time stamps and exploits Equation 1.2 as a mathematical model in order to estimate g . Let the (noisy) data be given by $x_1 = x(t_1 = 0.1) = 40$, $x_2 = x(t_2 = 1/\sqrt{2}) = 35$, $x_3 = x(t_3 = \sqrt{2}) = 20$ and $x_4 = x(t_4 = 3.9) = 0$. From Equation 1.2, one derives the distance, the apple has traveled,

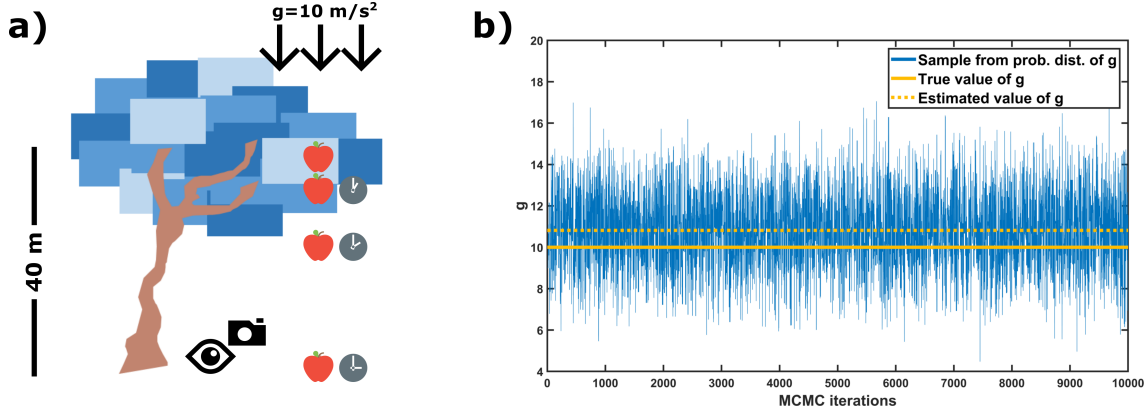


Figure 1.1: **Caption: A simple example of parameter estimation.** a) Observations of a falling apple. The time stamps possess measurement errors. b) Parameter estimation of the gravity acceleration g using Markov chain Monte Carlo.

$\Delta x_i(g) = x_i - x_0 = gt_i^2$. *Formulating an optimization problem*

$$\operatorname{argmin}_g \pi(g|x_1, x_2, x_3, x_4) \quad (1.3)$$

with an objective function

$$\pi(g|x_1, x_2, x_3, x_4) = (\Delta x_1(g))^2 + (\Delta x_2(g) - 5)^2 + (\Delta x_3(g) - 20)^2 + (\Delta x_4(g) - 40)^2 \quad (1.4)$$

yields an optimal solution $g = 10.8$ which is an estimate for the true but unknown $g_{\text{true}} \approx 10$. If one is interested in an estimate for the confidence of the previously obtained point estimate of g , one could sample from $\pi(g|x_1, x_2, x_3, x_4)$ as shown in Figure 1.1b. Assuming a measurement error of ± 10 meters, the obtained sample yields an estimated standard deviation of $\sigma = 1.7$, thus, one obtains $g = 10.8 \pm 1.7$. While this example is fairly simple, it demonstrates the basic concepts of parameter estimation and its utility of revealing latent variables, as g , which can not be measured directly. In addition to revealing latent variables, a trained model (Equation 1.2 with the estimated g) could be used to predict untested scenarios, e.g. how long the apple would fall if the tree was 60 meters tall instead.

The usage of mathematical modeling in cell biology goes back to 1950 (Monod, 1950; Turing, 1952; Hodgkin and Huxley, 1952). Nowadays, a variety of mechanistic model approaches are used, e.g. ordinary differential equations (ODEs), partial differential equation, and stochastic differential equations. However, there exist great interest in phenomenological and statistical models, e.g. based on neural networks, which focus more on the prediction of untested scenarios than on explaining the experimental data mechanistically. For the inverse problem of parameter estimation, the type of model is important,

because for some types of model, e.g. stochastic differential equations, the evaluation of the objective function may not be feasible or computationally demanding.

The parameters of biological processes are usually estimated using frequentist or Bayesian approaches (Raue et al., 2013a; Hug et al., 2013). Frequentist approaches tend to exploit optimization methods to determine the maximum likelihood estimate and its uncertainty, e.g., using bootstrapping or profile likelihoods (Joshi et al., 2006; Fröhlich et al., 2014b; Raue et al., 2011). Bayesian approaches often rely on the sampling of the parameter posterior distribution using Monte Carlo algorithms (Wilkinson, 2007; Xu et al., 2010; Krauss et al., 2013). Both, optimization and sampling, are challenging for a wide range of applications encountered in computational systems biology (Hug et al., 2013; Raue et al., 2013b).

Likelihood functions and posterior densities are frequently multi-modal and possess pronounced tails (see, e.g., (Raue et al., 2013a; Hug et al., 2013)), and many application problems possess structural and practical non-identifiabilities (see, e.g., (Raue et al., 2009; Balsa-Canto et al., 2010; Chis et al., 2011; Weber et al., 2011) and references therein). This is, amongst others, due to scarce and noise-corrupted experimental data. Furthermore, features of the underlying dynamical systems, such as bistability (Gardner et al., 2000; Ozbudak et al., 2004), oscillation (Tyson, 1991; Kholodenko, 2000; Calderhead, 2007) and chaos (Kosuta et al., 2008; Ngonghala et al., 2016; Braxenthaler et al., 1997) are expected to impact the structure of the likelihoods and posteriors. It is for instance well-known that the posterior distribution for systems exhibiting oscillations often possesses a large number of modes (Casey, 2004). To pinpoint limitations of optimization algorithms in the presence of such difficulties, large collections of parameter estimation problems which facilitate diverse model properties (see e.g. (Villaverde et al., 2015)) were established. Unfortunately, for sampling methods, (i) comprehensive benchmark collections and (ii) a rigorous comparison of their performance is missing. As a consequence, it is not clear how well state-of-the-art methods are suited for a given problem and the development of novel methods can not be supported by comprehensive benchmarks against existing methods. In practice, this shortage does often render the selection of methods in the presence of a given problem complicated.

Markov chain Monte Carlo (MCMC) methods provide samples from the posterior distribution of the parameters given the experimental data and prior knowledge (Andrieu et al., 2003). In contrast to point estimates provided by optimization algorithms, posterior samples facilitate the assessment of parameter- and prediction uncertainties (Vanlier et al., 2012). While uncertainty analysis is beneficial across research fields, in biology

these methods are especially important because experimental data is often sparse or noise corrupted and parameters are therefore often non-identifiable (Chis et al., 2011; Fröhlich et al., 2014b; Raue et al., 2013a; Eisenberg and Hayashi, 2013). Unraveling parameter uncertainties is essential to avoid incorrect conclusions and to facilitate reliable predictions. For that reason sampling methods as MCMC are widely used in systems and computational biology to parameterize computational models (Wilkinson, 2007). However, in order to exploit the benefits of sampling methods, in particular MCMC, one must ensure that the sample is representative. While for most MCMC algorithms convergence has been proven, for finite samples one has to apply statistical tests and heuristics to assess representativity. However, these tests tend to overestimate sample quality and often include subjective decision making. In the worst case, non-representative samples generate a false picture of the posterior and lead to an underestimation of uncertainties (Woodard, 2007; Brooks and Gelman, 1998; Raue et al., 2013a; Kreutz et al., 2013; Bayarri and Berger, 2004).

Bayesian methods have been successfully applied in problems focused on linear-, mixed-, hierarchical- (Albert, 1988; Carlin et al., 1992; Bennett et al., 1996), graphical- (Geman and Geman, 1984), and dynamical models (Wilkinson, 2006) as well as neural networks (Andrieu et al., 2003; Freitas et al., 2000; Marwala, 2007), diffusion and stochastic models (Roberts and Stramer, 2001; Roberts et al., 2004) and hidden trees (Larget and Simon, 1999; Huelsenbeck and Ronquist, 2001). Each of the mentioned model types is applied in a variety of biological applications, e.g. dynamical models were used for modeling biochemical reaction networks (Xu et al., 2010; Hug et al., 2013), spatio-temporal processes (Jagiella et al., 2017), single-cell data (Zechner et al., 2014) and more. Thus, Bayesian methods are widely applied and still under development. MCMC is no exception for this rule as it is one of the most generally applicable Bayesian methods and methodological and theoretical aspects are still an active field of research (Green et al., 2015). Lately, there has been a trend towards GPU usage and parallelization (Lee et al., 2010; Jacob et al., 2011; Strid, 2010; Suchard et al., 2010; Scott et al., 2016; Calderhead, 2014; Green et al., 2015; Jacob et al., 2017) and approximate approaches to cope with the increasing complexity of data and models (Green, 1995). However, the improvement of non-approximative MCMC methods continues as well (see e.g., (Lacki and Miasojedow, 2015; Hoffman and Gelman, 2014)). Overall, there is a distinctive desire for more efficient and reliable MCMC methods in order to cope with large-scale models (Fröhlich, 2018) and data across a variety of different parameter estimation problems, e.g. for the frequently occurring problem of parameter inference for ODE based models. Furthermore, as parameter estimation problems in practice are often treated as black-boxes, auto-tuning capabilities and improved convergence rates of the methods, e.g. by employing quasi-Monte Carlo (Owen and Tribble,

2005; Beentjes and Baker, 2019; Buchholz and Chopin, 2019), are of particular interest.

1.2 Contribution of the thesis

Parameter estimation is a difficult task in general. Employing MCMC as a tool for parameter estimation provides great benefits, e.g. the assessment of uncertainty of the estimated parameters and model predictions. However, the application and quality assessment of MCMC is not straightforward and practical applicability often remains unclear. The following key issues were outlined in the background in Section 1.1 and motivated this thesis:

- (i) MCMC is a powerful tool, which requires very few assumptions regarding the numeric properties of the problem of interest. As long as one can evaluate an objective function or posterior density point-wise, MCMC can be employed to generate a sample from such a target. This generality and the existence of a great number of different MCMC algorithms in the literature renders a quantitative comparison of MCMC algorithms valuable and challenging. However, it is not clear how to make MCMC results comparable because established measures such as effective sample size tend to overestimate sample quality especially for poorly performing methods (Chapter 3). In particular, it is not clear how one can evaluate MCMC in the presence of challenging posteriors (e.g. multiple modes) automatically and without visual inspections of the chains while being fair and robust across different problems and MCMC algorithms.
- (ii) Due to the general bias towards positive results in peer reviewed literature (Mahoney, 1977; Lee et al., 2013; Smith, 2006), novel MCMC algorithms are likely to get presented in favorable or well suited problems. Unfortunately, there are few independent, comprehensive and quantitative comparison studies of MCMC methods across different benchmark problems. This leaves MCMC algorithms in an odd position regarding its application to actual problems.
- (iii) In computational biology, mechanistic models are often defined by ODEs. The parameters of these models may provide valuable information about the underlying biological mechanism. For mechanistic models, the application of MCMC is beneficial because it allows to estimate uncertainty for these parameters supporting a correct biological interpretation of the parameter point-estimates. However, ODE models are very diverse in their behavior and properties of the model are expected to impact the posterior distribution which impacts the performance of parameter esti-

mation via MCMC. This connection was rarely evaluated quantitatively. Examples, e.g. covering the connection between oscillations of an ODE solution and MCMC performance (Calderhead and Girolami, 2011), are few.

- (iv) Due to the lack of standardized evaluations of the performance of MCMC algorithms (i), the lack of comprehensive benchmarks (ii), the unclear connections of model properties and algorithmic performances (iii), and the tuning intensive nature of MCMC methods, the beneficial employment of MCMC can be rather challenging in practice. Employing a poorly tuned MCMC algorithm will generate results which are likely to be non-representative of the posterior and misleading in their interpretation without any sign of warning (due to the lack of ground truth regarding the posterior shape). If one could close the gaps in (i-iii), the gathered insights would be beneficial for the construction of a novel MCMC method which improves state-of-the-art MCMC performance while performing well across diverse problems in computational biology.

The overall goal of the thesis is to address these open problems (i-iii) and possibility (iv) in order to assess and improve MCMC sampling based parameter estimation for problems frequently arising in computational biology applications. While the presented approaches and methods feature a fairly general usage across a variety of other disciplines, this thesis focuses on parameter estimation for biologically motivated ODE models. These models represent a particularly important sub-class of problems in computational biology. The overall novelty provided in the scope of this thesis can be summarized as:

- As a basis, in this thesis **a robust sampling result analysis pipeline for the assessment of sample quality is introduced**. In comparison to established assessments, the presented analysis pipeline aims towards increasing objectivity of the analysis and reducing the probability to overestimate sample quality in the presence of challenging posterior landscapes. The analysis pipeline is suited for quantitative benchmarks and the classification of novel MCMC methods as it does not rely on additional information about the given problem.
- Given a problem independent evaluation framework as the introduced analysis pipeline, it is possible to test and compare MCMC algorithms for a large number of problems. This is essential as posterior distributions arising in ODE constraint parameter estimation problems are diverse. In order to provide a broad evaluation of MCMC algorithms, **a benchmark collection of ODE constrained parameter estimation problems is introduced**. The collection of ODE constraint problems

includes a variety of ODE properties such as bifurcations, multi-stability and oscillations to study the connection between model properties and sampling performance quantitatively. In this thesis, for the first time **a comprehensive benchmarking of state-of-the-art sampling algorithms is performed**. In total, over 300.000 hours of CPU were executed and the results were analyzed using the analysis pipeline. This closes an important gap between optimization and sampling literature supporting future development of MCMC methods and informs a more selective choice of algorithms or tunings to apply efficiently for a given parameter estimation problem in computational biology and beyond.

- The comprehensive benchmark of state-of-the-art MCMC methods revealed major differences in sampling performance. Beside that, some valuable insights about MCMC behavior in ODE-constraint parameter estimation problems were gathered. These insights got employed for **developing a novel, well performing MCMC method, RAmPART**, which combines the benefits of multiple existing methods. A benchmark reveals, that RAmPART has the potential to outperform its ancestor methods significantly by up to a factor of 10 – 100. It does not require derivatives and features strong auto-tuning capabilities enabling its application to be beneficial in a variety of problems.

Overall, the thesis provides insights and algorithms for practical parameter estimation problems arising in computational biology and beyond. In particular, the thesis provides a framework for quantitative comparison of sampling results, a comprehensive benchmark of state-of-the-art MCMC methods and tunings, and a novel, well performing MCMC method.

1.3 Outline of the thesis

As a basis for comparisons, development and application of MCMC methods in ODE constraint parameter estimation problems, Chapter 2 starts with an introduction to Bayesian parameter estimation. The general concept of MCMC algorithms based on the corresponding Markov chain theory is motivated. The introduction of Bayesian parameter estimation is complemented by an introduction of general concepts of mathematical modeling, data integration, predictions and uncertainty analysis in particular regarding ODE constraint problems.

In Chapter 3, the basis for quantitative comparisons of sampling performance is set. A

pipeline for quantitative and unbiased benchmarking of MCMC methods and the assessment of sampling convergence for finite chains is discussed. Furthermore, potential pitfalls of general sampling quality assessment which the pipeline resolves are disentangled. The developed framework is provided with an example.

In Chapter 4, the pipeline developed in Chapter 3 is employed to quantitatively analyze MCMC samples obtained using multiple state-of-the-art algorithms and tunings in several benchmark problems. The results are used to draw conclusions about the performance of MCMC algorithms, e.g. about the selection of methods, optimal tunings or chain initialization using optimization in practice.

In Chapter 5, the insights gathered in Chapter 4 are used to develop a novel MCMC method fixing some of the identified problems while facilitating strong self-tuning capabilities. The novel method is quantitatively compared to similar methods using the pipeline from Chapter 3 once more. Furthermore, theoretical aspects of the methods are presented.

Chapter 2

Background

In this chapter, the mathematical and biological background required to follow the thesis is introduced. In particular, this Chapter gives an overview about quantitative mathematical modeling by employing parameter estimation methods. In order to formally introduce all required aspects of Markov chain Monte Carlo algorithms – a class of algorithms suited for comprehensive parameter estimations – a short introduction to the general field of Monte Carlo sampling methods is provided. The discussions about the theoretical aspects of MCMC are complemented by the introduction of a common, ODE-constraint parameter estimation problem, several MCMC algorithms and an important application of MCMC, uncertainty analysis.

2.1 Bayesian parameter estimation methods

The following sections introduce Bayesian parameter estimation (Section 2.1.1) and motivate sampling as a powerful tool to solve these problems stochastically (Section 2.1.2 and 2.1.3). First, sampling techniques which have evolved during the last century get introduced (Section 2.1.4) followed by more recently developed sampling methods such as MCMC (Section 2.1.5) and others (Section 2.1.6).

2.1.1 Bayesian statistics

“There are two major classes of numerical problem that arise in statistical inference, optimization [...] and integration problems” (Robert and Casella, 2013, Chapter 3). Optimization problems are often related to frequentist approaches while integration problems are typically associated with Bayesian methods (Robert and Casella, 2013). However, these connections are not strictly true (Robert and Casella, 2013) and Bayesian parameter estimation may be employed to solve both problems. A central idea of Bayesian parameter estimation is to merge currently available knowledge about the probability distribution of parameter $\theta \in \mathbb{R}^{n_\theta}$ with novel information encoded in the data \mathcal{D} (Kramer and Radde, 2010; Krauss et al., 2012). This merging of prior knowledge and evidence is

formalized as a posterior probability density $\pi(\theta|\mathcal{D})$ within the Theorem of Bayes,

$$\pi(\theta|\mathcal{D}) = \frac{\pi(\mathcal{D}|\theta)\pi(\theta)}{\pi(\mathcal{D})}, \quad (2.1)$$

in which $\pi(\theta)$ denotes the prior, $\pi(\mathcal{D}|\theta)$ denotes the likelihood and $p(\mathcal{D})$ denotes the marginal probability (being a normalization constant). In practice, the likelihood function $\pi(\mathcal{D}|\theta)$ depends on the model, e.g. of the considered dynamical system, and the model of the measurement process. Please note, priors may also be used for regularization purpose, e.g. Gaussian priors for an l_2 - and Laplace priors for an l_1 -regularization (Chaari et al., 2014).

2.1.2 Assessment of high dimensional probability distributions

“Only few problems allow for explicit computation of the likelihood, and even fewer for an explicit formula of [...] $\pi(\theta|\mathcal{D})$ ” (Hasenauer, 2013, Section 2.3.1) – e.g., due to the typically high dimensional integral which is the marginal,

$$\pi(\mathcal{D}) = \int_{\mathbb{R}^m} \pi(\mathcal{D}|\theta)\pi(\theta)d\theta. \quad (2.2)$$

Fortunately, it is possible to learn about the statistical quantities of $\pi(\theta|\mathcal{D})$ numerically while only using $\pi(\theta|\mathcal{D})$ up to a constant. Depending on the level of desired information about $\pi(\theta|\mathcal{D})$ one can apply different approaches, e.g. optimization to obtain maximum-a-posteriori estimates,

$$\theta^* = \arg \max_{\theta \in \mathbb{R}^{n_\theta}} (\pi(\theta|\mathcal{D})), \quad (2.3)$$

or profile calculation to obtain profiles of $\pi(\theta|\mathcal{D})$ ¹ for each parameter dimension $\theta^{(i)}$, $i = 1, \dots, n_\theta$,

$$PL(\theta^{(i)}) = \max_{\theta^{(j \neq i)} \in \mathbb{R}^{n_\theta - 1}} \pi(\theta^{(1)}, \dots, \theta^{(i)}, \dots, \theta^{(n_\theta)}|\mathcal{D}) \quad (2.4)$$

(Raue et al., 2010; Kreutz et al., 2013; Boiger et al., 2016; Fröhlich, 2018). If one desires more information about parameters θ and their distribution $\pi(\theta|\mathcal{D})$ one may use Monte Carlo methods.

¹See Section 2.5 for more information about uncertainty analysis.

2.1.3 Monte Carlo principle

Monte Carlo methods ‘simulate’ or ‘sample’ from $\pi(\theta|\mathcal{D})$. Those samples may be used to assess the underlying probability distribution. The assessment gets generally more precise with an increasing number of simulations.

One particular important case and common problem where a sample of $\pi(\theta|\mathcal{D})$ is needed is the evaluation of expectation values of a function $g : \mathbb{R}^{n_\theta} \rightarrow \mathbb{R}$ as it typically requires the evaluation of an integral,

$$E_\pi(g) = \int_{\mathbb{R}^{n_\theta}} g(\theta) d\Pi(\theta|\mathcal{D}) = \int_{\mathbb{R}^{n_\theta}} g(\theta) \pi(\theta|\mathcal{D}) d\theta, \quad (2.5)$$

where π is a probability density function and Π is the corresponding cumulative probability distribution. This problem can be solved using Monte Carlo integration. Therefore, random samples $\theta_1, \dots, \theta_n \sim \pi$ are generated to approximate Equation 2.5 with the empirical average (Robert and Casella, 2004),

$$\bar{g}_n = \frac{1}{n} \sum_{i=1}^n g(\theta_i). \quad (2.6)$$

Hereby, \bar{g}_n will almost certainly converge towards $E_\pi(g)$ by the strong law of numbers (Robert and Casella, 2004, Section 3.2). Furthermore, for functions g , whose expectation of g^2 is finite, the variance

$$\text{Var}_\pi(g) = \int_{\mathbb{R}^{n_\theta}} (g(\theta) - E_\pi(g))^2 \pi(\theta|\mathcal{D}) d\theta \quad (2.7)$$

exists and can be estimated using the sample variance

$$\bar{v}_n = \frac{1}{n} \sum_{i=1}^n (g(\theta_i) - \bar{g}_n)^2. \quad (2.8)$$

Thus, for $n \rightarrow \infty$,

$$\frac{\bar{g}_n - E(g(X))}{\bar{v}_n} \rightarrow \mathcal{N}(0, 1) \quad (2.9)$$

which provides a distribution for the error made for those methods regardless of the distribution of π and the non-linearity of g . Equation 2.9 is often part of central limit theorems (Andrieu et al., 2003). One direct application of integrals of the form in Equation 2.5 are Ito-integrals found in the evaluation of stochastic differential equations (Robert and Casella, 2004, Section 3.7.2) or – in the special case $g(\theta) = y_t(\theta)$ with $y_t(\theta)$ being the

output of a biological model as introduced in Section 2.2 – one can estimate the expectation curve of ODE solutions using a propagated sample of the posterior. This concept can be generalized to prediction uncertainty (see Section 2.5). Hereby, if the sample $\theta_1, \theta_2, \dots$ is propagated through a model $g, g(\theta_1), g(\theta_2), \dots$, one can estimate other statistical moments beside the expectation value as well, e.g. the quantiles. Additionally, the distribution of θ and correlation structures within θ hold valuable information, in particular for parameters of mechanistic models.

2.1.4 Classic Monte Carlo methods

Monte Carlo methods yield samples $\theta_1, \dots, \theta_N$ of a target density, e.g. the posterior $\pi(\theta|\mathcal{D})$. One basic Monte Carlo method is rejection sampling (Andrieu et al., 2003). It uses a proposal density $q(\theta)$ such that there exists an $M \in \mathbb{R}, M < \infty$ such that $\pi(\theta|\mathcal{D}) \leq Mq(\theta)$ to propose points $\theta_* \sim q$. The points θ_* are either accepted or rejected by chance based on how likely $\pi(\theta|\mathcal{D})$ is compared to $Mq(\theta)$ (see Figure 2.1 for the corresponding pseudo-code). The generated sample $\theta_1, \dots, \theta_N$ follows the probability $\pi(\theta|\mathcal{D})$ (Andrieu et al., 2003). The main idea behind rejection sampling is to choose q such that it can be sampled easily (by standard pseudo-number generators), e.g. being normal or uniform. The performance of rejection sampling scheme gets better, if the proposal density q has a similar location and shape as the target density π which will be often true for low dimensional problems. However, with increasing dimension one will almost certainly propose points along the tails of $\pi(\theta|\mathcal{D})$, which are unlikely to get accepted (Andrieu et al., 2003). This is especially true for uniform proposal densities or scenarios where large M are required to bound π/q across the whole parameter space (Andrieu et al., 2003).

Another well known Monte Carlo method is importance sampling. Importance sampling is directly connected to the concept of Monte Carlo integration, Equation 2.6. Here, Equation 2.5 is rewritten as

$$E_\pi(g) = \int_{\mathbb{R}^{n_\theta}} g(\theta)w(\theta)q(\theta)d\theta \quad (2.10)$$

with $w(\theta) = \pi(\theta|\mathcal{D})/q(\theta)$ being an importance weight and $q(\theta)$ being a proposal density similar to the one for rejection sampling, which should be easy to sample from (Andrieu et al., 2003). Instead of using Equation 2.6, one uses

$$\hat{g}_n = \frac{1}{n} \sum_{i=1}^n g(\theta_i)w(\theta_i) \quad (2.11)$$

which will almost certainly converge towards $E_\pi(g)$. One samples from $q(\theta)$ and calculates

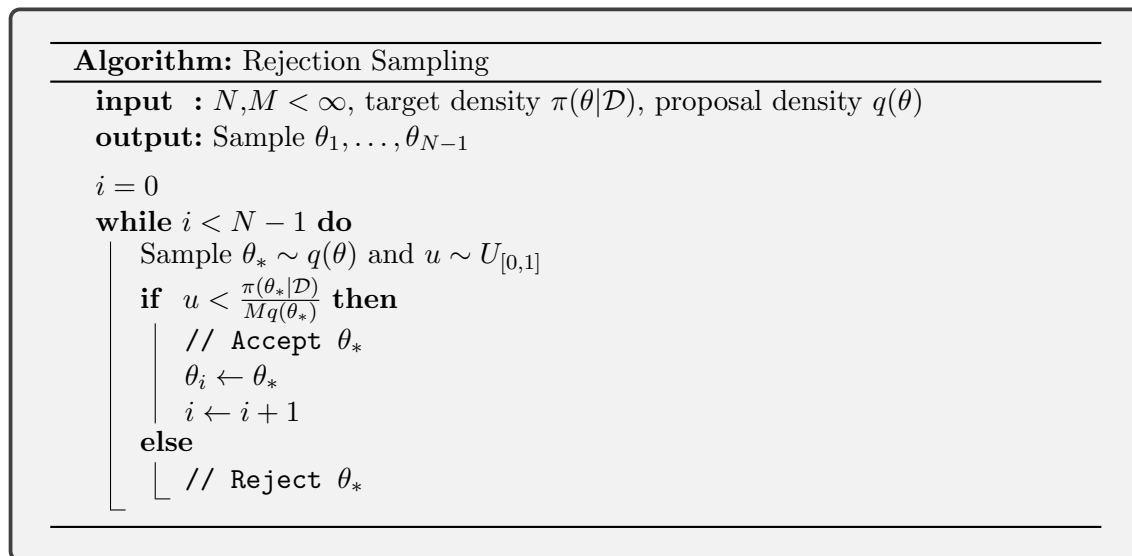


Figure 2.1: **The pseudocode of rejection sampling** (Andrieu et al., 2003).

the weights afterwards. If q is chosen similar to π so that it encourages points with large weights, importance sampling gets most efficient, as the variance between estimate \hat{g}_n and the true expectation value $E_\pi(g)$ decreases (Andrieu et al., 2003).

One main difference between rejection and importance sampling is the ensemble of what is kept guaranteed: For rejection sampling, one gets a guaranteed number of independent samples, but the time spent finding those is variable. For importance sampling, one controls the number of necessary evaluations of $\pi(\theta|\mathcal{D})$ but the precision of the result may be low in cases where most weights w were small. Both methods are classic Monte Carlo methods and need an appropriate q to function properly. This inherent weakness led to the invention of more advanced adaptive variants and Markov chain Monte Carlo methods.

2.1.5 The need for Markov chain Monte Carlo

Classic Monte Carlo methods generate an independent sample from a probability distribution, e.g. $\pi(\theta|\mathcal{D})$. Once the proposal density q is chosen similar to the target density $\pi(\theta|\mathcal{D})$, classic Monte Carlo methods work efficiently. Unfortunately, with increasing dimension and complexity of $\pi(\theta|\mathcal{D})$, it is almost impossible to choose q satisfactory without knowing additional properties of $\pi(\theta|\mathcal{D})$ before sampling. MCMC methods are a special class of Monte Carlo methods, which aims to address this problem. MCMC methods generate an irreducible Markov chain, whose equilibrium distribution is the probability distribution of interest. While the classic approaches are often not efficient in multivariate problems as required in most applications (Geyer, 1992), MCMC facilitates the generation

of samples in high dimensions but introduces autocorrelation to the sample as a trade-off (Geyer, 1992).

As MCMC algorithms are of particular interest for the thesis, they will be discussed in detail during Section 2.4. All considered MCMC algorithms (see Section 2.4.3 for examples) target $\pi(\theta|\mathcal{D})$ as stationary probability density (see Section 2.4.1 for theoretical aspects of MCMC methods).

2.1.6 Other Monte Carlo methods

MCMC methods are well suited in cases of deterministic dependency between model dynamics and observables (formally introduced in Section 2.2) or – more general – when evaluating $\pi(\theta|\mathcal{D})$ up to a constant is straightforward. This setting is given, i.e. in the case of parameter estimation problems of ODE-constraint models (see Section 2.2). However, there exist cases where alternative Monte Carlo approaches have to be used.

In cases, when there is an additional layer of stochasticity between model states and observables the corresponding parameter estimation problem is often called a filtering problem (Doucet et al., 2000). Examples are parameter estimation in hidden Markov models or GPS tracking. For filtering problems sequential Monte Carlo (SMC)² methods are more suitable than MCMC as they work recursively (Doucet et al., 2001). The general idea of SMC is to approximate sequential probabilities $p(x_k|y_1, \dots, y_{k-1})$ where x_k is the k^{th} state of the model given all preceding observations. In contrast, other Monte Carlo methods as MCMC or importance sampling would employ the full probability $p(x_1, \dots, x_k|y_1, \dots, y_k)$ instead.

Another interesting case are problems where the evaluation of $\pi(\theta|\mathcal{D})$ is not straightforward or extraordinarily expensive in computation, e.g. for parameter estimation of stochastic differential equations. In those cases, Approximate Bayesian Computing (ABC) is often employed to approximate $\pi(\theta|\mathcal{D})$ instead of evaluating it directly (Lillacci and Khammash, 2013; Jagiella et al., 2016; Stram et al., 2015; Toni et al., 2009). ABC methods generally construct a posterior approximation by accepting parameter points θ for which a certain distance measure $\rho(\mathcal{D}, \hat{\mathcal{D}}(\theta))$ between data \mathcal{D} and simulated data $\hat{\mathcal{D}}(\theta)$ is smaller than $\varepsilon > 0$. From this, one can derive the probability $p(\theta|\rho(\mathcal{D}, \hat{\mathcal{D}}(\theta)) \leq \varepsilon)$ as an approximation for the true posterior $\pi(\theta|\mathcal{D})$. The concept of ABC can be combined with several other Monte Carlo methods, e.g. rejection sampling, MCMC (Wegmann et al., 2009) or SMC (Toni et al., 2009).

² SMC methods are also known as particle filters, sampling importance resampling or sequential importance sampling.

2.2 Parameter estimation in ordinary differential equations

Before continuing the discussion about MCMC and the theory behind it in Section 2.3, an important application for Bayesian inference in systems biology, in particular for MCMC sampling, is presented. Specifically, the sampling of posterior distributions $\pi(\theta|\mathcal{D})$ arising for parameters θ of an ODEs-constraint model given data \mathcal{D} is introduced. While the general concepts of Bayesian inference were introduced in Section 2.1, here a concrete estimation problem and the construction of the likelihood, $\pi(\mathcal{D}|\theta)$, will be discussed.

ODE models are used for the mechanistic description of biological processes, e.g. gene regulation (Polynikis et al., 2009), signal transduction and metabolism (Covert et al., 2008), and pharmacokinetics (Klipp et al., 2005a; Krauss et al., 2013). Mathematically, ODE models can be defined as

$$\dot{x} = f(x, t, \eta), \quad x(t_0) = x_0(\eta), \quad (2.12)$$

with time $t \in [t_0, t_{\max}]$, state vector $x(t) \in \mathbb{R}^{n_x}$ and a parameter vector $\eta \in \mathbb{R}^{n_\eta}$. The vector field $f(x, t, \eta)$, $f : \mathbb{R}^{n_x} \times \mathbb{R} \times \mathbb{R}^{n_\eta} \rightarrow \mathbb{R}^{n_x}$, and the initial conditions $x_0(\eta)$, $x_0 : \mathbb{R}^{n_\eta} \rightarrow \mathbb{R}^{n_x}$, define the temporal evolution of the state variables as functions of η . Existence and uniqueness of the solution (Coddington and Levinson, 1955) is usually ensured by the structure of the right hand side, f . Hereby, continuity is sufficient for the existence of a local solutions using the Theorem of Peano (Peano, 1890) while Lipschitz continuity of f is sufficient for an *unique* local solution of Equation 2.12 using the Theorem of Picard-Lindelöf (Lindelöf, 1894). Fortunately, for biological applications often Lipschitz continuity can be assumed for f as will be done in the following.

For biological processes, experimental limitations usually prevent the direct measurement of the state vector $x(t)$. Instead, measurements provide information about the observable vector y . The observables y depend on the state of the process, $y = h(x, t, \eta)$, in which h denotes the output map $h : \mathbb{R}^{n_x} \times \mathbb{R} \times \mathbb{R}^{n_\eta} \rightarrow \mathbb{R}^{n_y}$. Examples for $f(x, t, \eta)$ and $h(x, t, \eta)$ can be found in Section 4.3. Please note, in cases where the observables h are stochastic, one would face a filtering problem (Doucet et al., 2000). However, for the problems treated during this thesis, h will always be deterministic.

The measurement of the observables y yields noise corrupted experimental data $\mathcal{D} = \{(t_k, \tilde{y}_k)\}_{k=1}^{n_t}$, where n_t is the number of time resolved data points. Please note, one could define observables independent of the time as well, e.g. an area under curve is common

when estimating parameters of physiologically based pharmacokinetic models. In the following, independent, additive normally distributed measurement noise

$$\tilde{y}_{ik} = y_i(t_k) + \epsilon_{ik}, \quad \epsilon_{ik} \sim \mathcal{N}(0, \sigma_i^2) \quad (2.13)$$

is assumed in which σ denotes the standard deviation of the measurement noise and $i = 1, \dots, n_y$ accounts for different observables. While this error model is most common, other error models may be used instead, e.g. one assuming Laplace distributed noise (Maier et al., 2017).

The standard deviations σ_i are usually unknown and part of the parameter vector, i.e., $\theta = (\eta, \sigma)$. The likelihood of observing the data \mathcal{D} given the parameters θ is

$$\pi(\mathcal{D}|\theta) = \prod_{i=1}^{n_y} \prod_{k=1}^{n_t} \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{(\tilde{y}_{ik} - y_i(t_k))^2}{2\sigma_i^2}\right), \quad (2.14)$$

in which $y(t_k)$ depends implicitly on η . Throughout this thesis, if not stated otherwise, this particular form of likelihood will be used (Section 2.1).

The solution $x = x(t, \theta)$ of the ODE directly impacts the structure of the likelihood $\pi(\mathcal{D}|\theta)$. Chapter 4 covers a range of different ODE-constraint problems which possess mathematical properties as multiple attractor sets known as multiple-stability, periodical orbits and bifurcations of the flow topology. These give rise to a diverse set of posteriors featuring rims, multiple modes and pronounced correlation structures in parameter space.

As for most ODEs the analytical solution is unknown, one typically applies numerical methods to obtain $x = x(t, \theta)$. When choosing a solver, a variety of different things must be considered and inventing scalable and robust solvers is still topic of ongoing research (Fröhlich, 2018). For ODEs in computational biology, implicit solvers (Butcher, 1964; Alexander, 1977; Rosenbrock, 1963) are suited because they can handle stiffness (Resat et al., 2009; Jia et al., 2011) fairly well (Gonnet et al., 2012). Sometimes, biologically motivated ODEs possess very different time scales. In these cases, one may apply techniques relying on discrete events (Le Novère, 2015; McAdams and Shapiro, 1995). Depending on whether one requires derivatives $\partial x / \partial \theta_i, i = 1, \dots, n_\theta$, techniques as forward sensitivity analysis or adjoint sensitivity analysis are applied (Fröhlich, 2018). In literature, there exist several solvers, e.g. the CVODE solver (Serban and Hindmarsh, 2005) and enhancements of it (Fröhlich et al., 2017c).

2.3 Biochemical reaction networks

During the last Section, 2.2, parameter estimation in ordinary differential equations was formally introduced. Before continuing with methods as MCMC suited for executing parameter estimation in practice, in this Section a particular class of ODE models derived from biochemical reaction networks (BRNs) will get introduced. Models of BRNs play an important role in computational biology as they allow the formulation of complex hypotheses on molecular level.

Generally, the state of a chemical reaction network can be represented by a vector of molecule numbers n (Van Kampen, 2007). The probability $P : \mathbb{R} \times \mathbb{R}^{N_s} \times \mathbb{R}^{N_r}$ of a certain state n_s for species s at time t is then determined by the Chemical Master Equation (CME) (Gillespie, 1992),

$$\dot{P}(t, n_s, \theta) = \Omega \sum_{r=1}^{N_r} (\hat{v}_r(n_s - \nu_r, \theta)P(t, n_s - \nu_r, \theta) - \hat{v}_r(n_s, \theta)P(t, n_s, \theta)), \quad (2.15)$$

where $\hat{v}_r(n_s, \theta)$ denotes the microscopic propensity function, i.e., the probability per unit time for reaction r to occur in an infinitesimal volume, and a macroscopic volume Ω (Fröhlich, 2018). The analytical solution of the CME and its simulation via the Stochastic Simulation Algorithm (Gillespie, 1977) is typically intractable due to the large number of attainable system states (Fröhlich, 2018). However, there exist approximative approaches (Munsky and Khammash, 2006; Gillespie, 2000; Risken, 1996), and often only the statistical moments of the concentration,

$$\mu_s = \mathbb{E}(n_s/\Omega), \quad (2.16)$$

are estimated. The most commonly applied method to obtain estimates, c_s , for μ_s are Reaction Rate Equations (RREs). Under the assumption of mass action kinetics, the RRE for the r th non-reversible reaction can be defined by the reaction flux $v_{r,s} : \mathbb{R}^{N_s} \times \mathbb{R}^{N_\theta} \rightarrow \mathbb{R}$,

$$v_{r,s}(c, \theta) = k_r(\theta) \prod_{s=1}^{N_s} c_s^{\nu_{sr}}, \quad (2.17)$$

where c_s is a mean concentration and k_r is a kinetic rate constant. The stoichiometric matrix S is given by

$$S = (\nu_{sr})_{s=1, \dots, N_s; r=1, \dots, N_r}. \quad (2.18)$$

The reaction flux v and the stoichiometric matrix S define the RRE,

$$\dot{c} = S \cdot v(c, \theta), \quad (2.19)$$

whose solution is an approximation for the concentration c at time t . The approximation tends to become more precise with larger volumes Ω and molecule numbers and will eventually converge towards the exact average solution of 2.15 for $\lim_{\Omega \rightarrow \infty}$ (Engblom, 2006; Grima, 2012; Van Kampen, 2007). Due to this asymptotic property, the RRE is often seen as macroscopic description of the system. Please note, there exist higher order moment approximations of the CME, e.g., System Size Expansion and Moment-Closure Approximation and more (Fröhlich, 2018; Hespanha, 2008; Singh and Hespanha, 2007; Gandhi et al., 2000), which all make certain assumptions about P , which are hard to verify a priori (Fröhlich, 2018; Grima, 2012).

From a mathematical point of view, RREs are a particular special case of ODEs which are known to facilitate complex behavior as multi-stability, oscillations, chaos and more. In fact, BRNs described by ODEs make no exception from this. E.g., "Feedback [loops in the network] can result in bistable behavior with discrete steady-state activities, well-defined input thresholds for transition between states and prolonged signal output, and signal modulation in response to transient stimuli. These properties of signaling networks raise the possibility that information for 'learned behavior' of biological systems may be stored within intracellular biochemical reactions that comprise signaling pathways." (Bhalla and Iyengar, 1999). In fact, one can define precise conditions under which a BRN can or will behave in a certain topological way (Feinberg and Horn, 1977; Klamt and Gilles, 2004; Clarke, 1980). Nowadays, there exist several tools to treat, exchange and analyze BRNs, e.g. the systems biology markup language (SBML) (Hucka et al., 2003), the CellDesigner toolbox (Funahashi et al., 2008), or the COPASI toolbox (Hoops et al., 2006) to name a few.

2.4 Markov chain Monte Carlo

MCMC methods are a class of Monte Carlo methods suited for sampling from complicated, unknown and high-dimensional probability distributions, e.g. the posterior found in ODE based parameter estimation problems described in Section 2.2 and Section 2.3. In this context, sampling proceeds by running a Markov chain, whose stationary distribution is $\pi(\theta|\mathcal{D})$ while the sample gets more representative as its size grows.

This thesis focuses on application, comparison and improvement of MCMC methods. Therefore, the mathematical background and notation of MCMC Methods is introduced in detail (Section 2.4.1) followed by examples of state-of-the-art algorithms (Section 2.4.3).

2.4.1 Theory of Markov chains in MCMC

MCMC algorithms are designed to sample from a target distribution π by employing Markov chains which possess the target distribution π as an (unique) stationary distribution. This subsection will formally introduce several important definitions and theorems about Markov chains in the context of MCMC in order to summarize what is necessary for such a stationary distribution to exist and the Markov chain to converge towards it. If not stated otherwise, definitions and statements provided in this chapter are based on (Meyn and Tweedie, 2012; Robert and Casella, 2013; Geyer, 1992) and (Olsen, 2015), which provide excellent introductions to Markov chain theory as well.

In order to define a Markov chain, the definition of a transition kernel is required. Let (Ω, \mathcal{A}) be a probability space with a state space, $\Omega \subseteq \mathbb{R}^k$, a σ -Algebra \mathcal{A} defined by the power set of Ω with measurable elements $A \in \mathcal{A}$.

Definition 2.1 (Transition kernel). *A (transition) kernel $P(\cdot, \cdot)$ is a function defined on (Ω, \mathcal{A}) such that*

- (i) $P(x, \cdot)$ is a probability measure on \mathcal{A} for all $x \in \Omega$ and
- (ii) $P(\cdot, A)$ is a non-negative measurable function on Ω for all $A \in \mathcal{A}$.

Furthermore, an n -step (transition) kernel is defined as

$$P^n(x, A) = \int_{\Omega} P^{n-1}(y, A)P(x, dy). \quad (2.20)$$

A Markov chain is defined in terms of its transition kernel.

Definition 2.2 (Markov chain). *A time discrete, time invariant stochastic process $\{X_i\}_{i=1 \dots n}$ on (Ω, \mathcal{A}) is called Markov chain, if it is defined by an initial distribution, X_1 , and transition probabilities*

$$\begin{aligned} \mathcal{P}((x_i \sim X_i) \in A | X_{i-1} = x_{i-1}, \dots, X_1 = x_1) &= \mathcal{P}((x_i \sim X_i) \in A | X_{i-1} = x_{i-1}) \\ &= \int_A P(x_{i-1}, dy) \end{aligned} \quad (2.21)$$

for $i = 2, \dots, n$, realizations $x_i \sim X_i$ and a measurable set $A \in \mathcal{A}$.

Equation 2.21 generates a stochastic process $\{X_i\}_{i=1\dots n}$ whose next realization only depends on the current state of the chain. Note, that a transition kernel as defined in 2.1 will often be expressed in terms of its corresponding transition density $p(x, y)$ so that

$$P(x, A) = \int_A p(x, y)\mu(dy) \quad (2.22)$$

where μ denotes a positive σ -finite measure on (Ω, \mathcal{A}) . The existence of p is ensured by the Radon-Nikodym Theorem (Resnick, 2013) and will be often used throughout the notation of this thesis. Markov chains on continuous times, are called Markovian processes (Robert and Casella, 2013) and are not treated during this thesis.

Example 2.3 (Metropolis-Hastings Transition Kernel). *The most common type of transition kernel in MCMC literature is the kernel introduced by Metropolis et al. (1953) and Hastings (1970). It founds the basis for a variety of modern algorithms (see Section 2.4.3). Let π be a target measure on (Ω, \mathcal{A}) , e.g. as defined for a parameter estimation problem as in Section 2.2. Here, the transition density p is defined in terms of an algorithm-specific proposal density q and the Metropolis-Hastings acceptance criteria p_{acc} ,*

$$p(x, y) = p_{acc}(x, y)q(y|x)(1 - \delta_x(y)) + \left(1 - \int_{\Omega} q(z|x)p_{acc}(x, z)dz\right) \delta_x(y). \quad (2.23)$$

The Metropolis-Hastings acceptance criteria is defined as

$$p_{acc}(x, y) = \min\left(1, \frac{\pi(y)q(x|y)}{\pi(x)q(y|x)}\right). \quad (2.24)$$

Note, that Equation 2.23 describes the transition of the chain starting from point x either towards a novel point y or the old position x once more. Hereby, the first term in Equation 2.23 covers novel accepted points y and the second term covers rejections (see Sections 2.4.3 or 5.2.2 for additional details of the Metropolis-Hastings algorithm).

Markov chains defined by MCMC algorithms share certain properties which guarantee proper behavior. In the following, these properties are introduced and finally connected to the important concepts of ergodicity and central limit theorems which guarantee chain convergence in distribution within controlled terms.

The first definition is φ -irreducibility, which basically states, that a Markov chain is able to move from any x to any A , as long as A has a non-vanishing measure regarding φ .

Definition 2.4 (φ -irreducibility). *A Markov chain $\{X_i\}_{i=1,\dots,n}$ with kernel $P(\cdot, \cdot)$ is called φ -irreducible for a measure φ on \mathcal{A} if for every $A \in \mathcal{A}$ with $\varphi(A) > 0$ there exists an $b \in \mathbb{N}$ so that $P^b(x, A) > 0$ for all $x \in \Omega$.*

Building on φ -irreducibility, (Meyn and Tweedie, 2012) introduced ψ -irreducibility, which follows Definition 2.4 while guarantying that negligible sets $B \in \mathcal{A}$ with $\varphi(B) = 0$ are avoided with probability one from most starting points. While ψ -irreducibility is interesting for theoretical purpose, the existence of measure ψ conveniently follows from φ -irreducibility and there is a way of constructing it (Meyn and Tweedie, 2012, Proposition 4.2.2).

While irreducibility is important for the chain to be able to visit all places of the state space Ω the next important property, aperiodicity, is important to ensure that there is no repeating structure in the way it moves.

Definition 2.5 (Aperiodicity). *A ψ -irreducible Markov chain $\{X_i\}_{i=1,\dots,n}$ with kernel $P(\cdot, \cdot)$ possesses d disjoint sets $D_1, \dots, D_d \in \mathcal{A}$ such that*

$$P(x, D_{i+1}) = 1 \text{ for } x \in D_i, \quad i = 1, \dots, d-1 \quad (2.25)$$

and

$$P(x, D_1) = 1 \text{ for } x \in D_d \quad (2.26)$$

while $\psi((\cup_{i=1}^d D_i)^c) = 0$. If $d = 1$, the chain $\{X_i\}_{i=1,\dots,n}$ is aperiodic (Meyn and Tweedie, 2012, Theorem 5.4.4).

Aperiodicity (and therefore irreducibility) is one of the necessary ingredients for Harris ergodicity introduced below. Another ingredient for Harris ergodicity is recurrence. Recurrence of a Markov chain makes a statement about how often a chain would visit a certain set $A \in \mathcal{A}$ if the chain was run infinitely long.

In order to define recurrence, one needs the indicator function

$$1_A(x) = \begin{cases} 1, & x \in A \\ 0, & x \notin A \end{cases} \quad (2.27)$$

and the number of passages of x_i through $A \in \mathcal{A}$,

$$\eta_A = \lim_{n \rightarrow \infty} \sum_{i=1}^n 1_A(x_i). \quad (2.28)$$

Furthermore, the notations for the conditioned probability $P_x(\cdot) = P(\cdot | X_1 = x)$ and expectation $E_x(\cdot) = E(\cdot | X_1 = x)$ given a certain initial state of the chain will be used in the following.

Definition 2.6 (Recurrence). *A ψ -irreducible Markov chain is recurrent if for every $A \in \mathcal{A}$ with non-zero measure, $\psi(A) > 0$, the expected number of passages through A , $E_x(\eta_A) = \infty$, is infinite.*

Recurrence implies that a chain is expected to visit non-zero measure sets A infinitely often. There is an even stronger property called Harris recurrence, from which recurrence follows.

Definition 2.7 (Harris recurrence). *A ψ -irreducible Markov chain is Harris recurrent if for every $A \in \mathcal{A}$ with $\psi(A) > 0$ is $P_x(\eta_A = \infty) = 1$ for all $x \in A$.*

While recurrence indicates an infinite expectation of the number of visits, Harris recurrence indicates that each chain almost surely visits a non-zero measure set A infinitely often (Meyn and Tweedie, 2012) – which is a stronger condition.

Before irreducibility, aperiodicity and recurrence will be used to introduce chain convergence, first the concept behind convergence is formalized. MCMC methods have the purpose to sample from a probability distribution π . The first step is to define π as an invariant measure of the Kernel P .

Definition 2.8 (Invariant measure). *Let π be a probability measure for which*

$$\pi(A) = \int_{\Omega} P(x, A)\pi(dx), \quad \forall A \in \mathcal{A} \quad (2.29)$$

holds. Then π is invariant for $P(\cdot, \cdot)$ and the corresponding Markov chain. The corresponding Markov chain is called positive if such an invariant measure exists. Otherwise it is called null. If the marginal of the Markov chain, $X_1 = \pi$, is invariant, the chain is stationary.

Note, that π being invariant for P does not guarantee the convergence of the Markov chain towards π . A Markov chain possessing an invariant transition kernel is not the same as being stationary. The former is determined just by the kernel of the chain, while the latter is determined by the kernel and the initial distribution X_1 (Geyer, 1992). In practice, this difference may cause burn-in which is discussed in Section 3.2.1. A useful connection is the following theorem as it shows that the existence of an invariant measure is quite powerful.

Theorem 2.9 (Uniqueness of invariant measure). *ψ -irreducible positive chains are Harris recurrent (Meyn and Tweedie, 2012, Proposition 10.1.1) and the corresponding invariant measure π is unique up to a constant (Meyn and Tweedie, 2012, Theorem 10.4.4).*

Thus, for irreducible chains, if an invariant measure exists, it is unique as well. Because positiveness implies Harris recurrence, these terms are often used together. However, the existence of an invariant measure π is questionable in general. One common approach to prove existence of an invariant measure π is to apply the detailed balance criteria. While it is easy to apply, it is a rather restricting sufficient condition for the existence of an invariant π .

Definition 2.10. *A Markov chain $\{X_i\}_{i=1,\dots,n}$ and a kernel density p satisfy detailed balance if there exists an function f such that*

$$p(y, x)f(y) = p(x, y)f(x) \tag{2.30}$$

for all $x, y \in \Omega$.

Another implication of detailed balance is reversibility, a property which is commonly discussed in MCMC literature.

Definition 2.11. *A Markov chain $\{X_i\}_{i=1,\dots,n}$ is reversible if the conditional distribution $P(X_i = x|X_{i+1}) = P(X_i = x|X_{i-1})$ for all $i \geq 2$.*

The following theorem connects detailed balance with the existence of an invariant set and reversibility.

Theorem 2.12. *A Markov chain $\{X_i\}_{i=1,\dots,n}$ with a kernel P fulfilling detailed balance regarding a probability measure, π ,*

- (i) has π as an invariant density and*
- (ii) is reversible.*

Thus, detailed balance is one *particular* way of showing the existence of an invariant measure π for the transition kernel. The existence of an invariant measure is, however, a prerequisite for chain converge which will be introduced next by the definitions of total variance distance and Harris ergodicity³. Please note, reversibility is not crucial to prove Harris ergodicity and thus the overall convergence of a chain. However, reversibility may be used beneficially in order to treat convergence rates as discussed below ([Kontoyiannis and Meyn, 2009](#); [Jones et al., 2004](#)).

Harris ergodicity guarantees convergence of the kernel towards an equilibrium density π .

³ Sometimes just called ergodicity.

Definition 2.13 (Harris ergodicity). *An aperiodic and positive Harris recurrent (and thus ψ -irreducible by the prerequisites of thee) Markov chain $\{X_i\}_{i=1,\dots,n}$ is called Harris ergodic.*

In order to talk about the two major benefits of Harris ergodic chains, first the total variance distance has to be introduced.

Definition 2.14 (Total variance distance). *Denote μ_1 and μ_2 as two measures on (Ω, \mathcal{A}) . The total variance distance is defined as*

$$\|\mu_1(\cdot) - \mu_2(\cdot)\|_{TV} = \sup_{A \in \mathcal{A}} |\mu_1(A) - \mu_2(A)|. \quad (2.31)$$

The definition of total variance distance gives rise to the following theorem.

Theorem 2.15 (Theorem 13.3.3, (Meyn and Tweedie, 2012)). *For a Harris ergodic chain with kernel P and a stationary distribution $\pi(\cdot)$,*

$$\lim_{n \rightarrow \infty} \|P^n(x, \cdot) - \pi(\cdot)\|_{TV} = 0 \quad (2.32)$$

holds for every $x \sim X_1$ and any initial distribution X_1 .

Therefore, a Harris ergodic chain converges in total variance distance towards its stationary distribution regardless of the initial distribution which is a strong result. Another equally strong result is the following theorem.

Theorem 2.16 (Law of large numbers, Theorem 17.0.1, (Meyn and Tweedie, 2012)). *For a Harris ergodic chain $\{x_i\}_{i=1,\dots,n}$ with target density π and a measurable function g , such that $E_\pi(|g|) = \lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n |g(x_i)| < \infty$, the law of large numbers,*

$$\lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n g(x_i) = E_\pi(g), \quad (2.33)$$

holds almost surely for any initial value $x_1 \sim X_1$ and every initial distribution X_1 .

Thus, for a Harris ergodic chain the average function values do almost surely converge towards the desired expectation value. This is important in practice, as it allows for precise propagation of parameter samples through a non-linear model in order to obtain model prediction uncertainty (see Section 2.2 and 2.5).

There are multiple stronger versions of ergodicity. The most common ones are geometric and uniform ergodicity (Meyn and Tweedie, 2012). In addition to the convergence in variation and law of large number guaranteed by Harris ergodicity, they further allow to appraise upper bounds of convergence rates.

Definition 2.17 (Geometric and uniform ergodicity). *A Harris ergodic chain is geometrical ergodic if for some function $M : \Omega \rightarrow \mathbb{R}^+$ and any $x \in \Omega$ there exists a constant $r \in]0, 1[$, such that*

$$\|P^n(x, \cdot) - \pi(\cdot)\|_{TV} \leq M(x)r^n, \quad (2.34)$$

for any positive integer n . A special case and even stronger version of geometric ergodicity is if $M(x)$ is bounded. This case is called uniform ergodicity.

The upper definition states, that the difference between an n -step kernel P^n and the target density π does decrease in each step by at least a factor r which is thus an upper bound for convergence of the chain. While Harris ergodicity is common for MCMC algorithms, geometric ergodicity is less frequent and more difficult to show (Olsen, 2015). Geometric ergodicity can be proven using so called minorization and drift conditions (Meyn and Tweedie, 2012) or small-sets (Meyn and Tweedie, 2012).

Note, reversibility impacts calculus in spectral gap theory (Kontoyiannis and Meyn, 2009), which can be used for the determination of convergence rates r of geometrically ergodic chains (Rosenthal, 1995; Meyn et al., 1994). In particular, it determines whether Hilbert spaces, L_2 , or L_∞ spaces have to be used in order to obtain spectral theory results (Kontoyiannis and Meyn, 2009). In the latter, the kernel P does not possess the properties of a linear, self-adjoint operator, making the calculus arguably harder (Kontoyiannis and Meyn, 2009).

In addition to the law of large numbers, central limit theorems (CLTs) are of practical interest, as they quantify the asymptotic variance of the difference between estimate and true functional of π (Jones et al., 2004). They basically determine how precise a functional estimate $g_n = n^{-1} \sum_{i=1}^n g(X_i)$ can get, by guaranteeing

$$\sqrt{n}(g_n - E_\pi(g)) \rightarrow \mathcal{N}(0, \sigma_g^2), \quad (2.35)$$

almost certainly for $n \rightarrow \infty$ with a variance σ_g^2 which often has to be estimated (Olsen, 2015). There exist several criteria for CLTs to hold, e.g. (Jones et al., 2004, Theorem 9). Interestingly, reversibility does impact the calculus of central limit theorems as it is shown, that if the detailed balance (and thus reversibility) holds, a central limit theorem

can be proven for geometric ergodic chains for which $E_\pi(g^2) < \infty$ holds. In contrast, without the assumption of reversibility, higher moments must be assumed to be finite, e.g. $E_\pi(|g|^{\delta+2}) < \infty$ for some $\delta > 0$ in order to prove a CLT. Furthermore, the type of ergodicity, e.g. geometric or polynomial (Jones et al., 2004), do determine the behavior of σ_g^2 (Jones et al., 2004). As Markov chains are auto-correlated, $\sigma_g^2 \neq \text{Var}_\pi(g)$ would be a poor estimate. However, low auto-correlation of the chain implies smaller σ_g^2 and thus, by Equation 2.35, higher precision of the MCMC approximation towards $E_\pi(g)$.

In this section, the theoretical results for time homogeneous Markov chains in the context of MCMC were summarized. However, there exists several adaptive algorithms (see Section 2.4.3) which are not time homogeneous anymore. Fortunately, the theory can often be extended towards those cases (Roberts and Rosenthal, 2007; Rosenthal and Yang, 2017; Craiu et al., 2015). Some of these results will be used during Chapter 5 in order to prove ergodicity for a novel adaptive method.

2.4.2 Finite MCMC samples

For MCMC algorithms, Harris ergodicity (as discussed in Section 2.4.1) provides a theoretical statement of MCMC convergence for infinite long chains. However, in practice the generated Markov chains are finite and Harris ergodicity becomes a necessary but not a sufficient condition for the generation of a representative sample of the target distribution π . Hereby, the term representative belongs to chains for which the number of iterations is sufficiently large in order for the CLT, Equation 2.35, to hold. This implies that MCMC estimates are typically biased if n is too small (Jacob et al., 2017). Unfortunately, for a given problem it is typically not clear how many samples are required for the theorems in Section 2.4.1 to apply. This unknown number of required samples, n , does depend on the specific problem, the performing algorithm, its tuning parameters, the position where the chain is initialized and even the random seed used in a certain run as will be discussed and exemplified in Chapter 4. In literature, there exists a variety of statistical tests and convergence diagnostics, e.g. (Cowles and Carlin, 1996b). However, there are several reasons which can cause a finite MCMC sample to become non-representative (see Chapter 3) which are difficult to address with just one of these approaches. Generally, existing convergence diagnostics tend to systematically overestimate sampling quality. The reasons for non-representativity will be discussed in detail within the Problem Statement of Chapter 3 to motivate the development of an analysis pipeline during Chapter 3. The pipeline is meant to identify non-representative MCMC samples and suited for the analysis of diverse MCMC results for an arbitrary posterior density. It combines several convergence diagnostic approaches to decrease the the likelihood of overestimating sampling quality.

Table 2.1: An incomplete list of MCMC methods.

Algorithm	Signature Mechanism	Source
MH	Classic MCMC with fixed proposal	1953; 1970
AM	Adapts proposal density over time	2008
DRAM	Repeated proposals after rejection	2006
MALA	Uses derivative information about the objective to propose	2011
HMC	Hamiltonian movement through objective landscape	2011
PT	Samples and exchanges multiple tempered objectives	2013; 2016; 2013
PHS	Employs and exchanges multiple chains	2012
Gibbs	Samples from full conditional distribution	1992; 1984; 2003
MiG	Metropolis techniques within Gibbs	2017
RJ MCMC	Reversible-jump MCMC extends to spaces of varying dimensions	1995; 2011
SS	Slice Sampling of the objective landscape	2003
Copula MCMC	Disentangling of marginals and dependency of parameters	2013
Wormhole MCMC	Special HMC adding shortcuts to the parameter space	2014
No-U-turn	Special HMC with forced directions	2014

While the pipeline employs multiple MCMC runs in order to decrease the likelihood of underestimating the convergence time, other novel approaches employ multiple chains as well in parallel in order to directly obtain unbiased Monte Carlo estimates (Jacob et al., 2017).

2.4.3 MCMC methods

The following section is based on (Ballnus et al., 2017).

A well-known MCMC algorithm is the Metropolis-Hastings (MH) algorithm (Metropolis et al., 1953; Hastings, 1970). The MH algorithm samples from the posterior via a weighted random walk. Parameter candidates are drawn from a proposal distribution and accepted or rejected based on the ratio of the posterior at the parameter candidate and the current parameter. The choice of the proposal distribution is a design parameter. In practice the distribution is frequently chosen to be symmetric, e.g., a normal distribution, and centered at the current point.

While being the theoretical foundation of modern MCMC algorithms, in practice the MH algorithm has several shortcomings, including the need for manual tuning of the proposal covariance and high autocorrelation (Andrieu et al., 2003). Accordingly, a large number of extensions has been developed. In the following two subsections, multiple extensions and alternatives for MH MCMC sampling (see Table 2.1) are shortly presented.

In Chapter 4 of this thesis some of these extensions are benchmarked. Figure 2.2 highlights the differences between the sampling methods employed in this thesis using a pseudo-code

representation. In the following, the three single-chain and the two multi-chain methods employed in Chapter 4 are introduced.

Adaptive Metropolis (AM): The AM algorithm is an extension of the standard MH algorithm. Instead of using a fixed proposal distribution which is tuned manually, the distribution is updated based on the already available samples. In particular, for posteriors with high correlation, this improves sampling efficiency by aligning the proposal with the posterior distribution (Andrieu and Thoms, 2008). In addition to the correlation structure, the scale of the proposal is also adapted. A commonly applied scaling scheme is based on the dimension of the problem (Haario et al., 2001, 2006) while other possible schemes are based on the chain acceptance rate (Miasojedow et al., 2013). These scaling schemes are in the following (in particular Chapter 4) indicated by ‘dim’ and ‘acc’, respectively.

Delayed Rejection Adaptive Metropolis (DRAM): To further decrease the in-chain auto-correlation, the AM algorithm has been combined with a delayed rejection method, yielding the DRAM algorithm (Haario et al., 2006). When a candidate parameter is rejected, the algorithm tries to find a new point using the information about the rejected point. This is repeated multiple times until a certain number of tries is reached or a point is accepted. During Chapter 4, the implementation provided in (Haario et al., 2006) is employed and is exclusively based on the previously mentioned ‘dim’ adaption scheme.

Metropolis-adjusted Langevin Algorithm (MALA): Both, AM and DRAM, work best if the proposal density suits the local shape of the posterior. Otherwise, the performance of the algorithm suffers, i.e. the in-chain auto-correlation increases. To circumvent this problem, the MALA makes use of the gradient, $\nabla_{\theta} p(\theta|\mathcal{D})$, and Fisher Information Matrix (Girolami and Calderhead, 2011) of the estimation problem at the current point in parameter space. This information is used to construct a proposal which is adapted to the local posterior shape (Calderhead, 2011; Girolami and Calderhead, 2011). Gradient and Fisher Information Matrix can be computed using forward sensitivity equations (Raue et al., 2009).

Parallel Tempering (PT): All of the algorithms, AM, DRAM and MALA, discussed so far are single-chain algorithms which exploit local posterior properties to tune their global movement. This can make transitions between different posterior modes unlikely if they are separated by areas of low probability density. To address the issue, PT algorithms

Algorithm 1: MCMC algorithms used in this study

```

input : Initial point  $\theta^0$ , lower bounds  $\theta_{min}$ , upper bounds  $\theta_{max}$  and number of samples  $N_{sample}$ 
input : Initial covariance  $\Sigma^0$  // This is required for all algorithms but MALA
input : Algorithm-specific options for AM, DRAM, PT, PHS and MALA
output:  $\theta^1, \dots, \theta^{N_{sample}}$ 

Initialize
for  $i \leftarrow 1$  to  $N_{sample}$  do
  // AM, DRAM and MALA use a single chain  $L = 1$  while PT and PHS use multiple
  // chains  $L > 1$ . The chain index is denoted by  $l = 1, \dots, L$  and the chain
  // position by  $\theta^{l,i}$ 
  for  $l \leftarrow 1$  to  $L$  do
    // AM, MALA, PHS and PT propose a single new candidate  $\theta_{cand}^{l,k}$  in each
    // iteration  $i$  per chain  $l$  ( $N_{tries} = 1$ ). DRAM exploits multiple tries
    //  $N_{tries} > 1$  to decrease the auto-correlation.
    for  $k \leftarrow 1$  to  $N_{tries}$  do
      Propose a candidate  $\theta_{cand}^{l,k} \sim \mathcal{N}(\theta^{l,i-1}, \Sigma^{l,i-1})$ . // All algorithms in this study
      // use a normal distribution for proposing new candidates.
      if  $\theta_{min} < \theta_{cand}^{l,k} < \theta_{max}$  then
        Evaluate the acceptance probability  $p_{acc}^{l,k}$  as a function of posterior values and
        // transition probabilities // For DRAM the acceptance probability accounts
        // for the multiple tries. PT compares the tempered posterior
        // values.
      else
        |  $p_{acc} \leftarrow -\infty$ 
      end if
      if  $u \sim U(0,1) \leq p_{acc}^{l,k}$  then
        | Accept candidate  $\theta^{l,i} \leftarrow \theta_{cand}^{l,k}$ 
        | break for // Necessary in case of DRAM.
      else if  $k = N_{tries}$  then
        | Reject candidate  $\theta^{l,i} \leftarrow \theta^{l,i-1}$ 
      end if
    end for
    Calculate new proposal covariance matrix  $\Sigma^{l,i}$ . // For MH  $\Sigma$  is fixed. For AM, DRAM,
    // PT and PHS,  $\Sigma^{l,i}$  is calculated from  $\Sigma^{l,i-1}$  and  $\theta^{l,i}$ . For MALA  $\Sigma$  is
    // approximated using local gradient and Hessian information at the current
    // point  $\theta^{l,i}$ .
    Adapt scaling factor  $\eta$ . // For AM, DRAM, PT and PHS,  $\Sigma$  is usually multiplied
    // with a scalar factor  $\eta$  to ensure 23.4% acceptance of the chain.
  end for
  Swap chains // Only for PT and PHS. Swaps of PT chains are executed by chance,
  // applying a swapping strategy as for example PTEE. PHS swaps its main chain
  // ( $l = 1$ ) with one of the auxiliary chains ( $l \in \{2, \dots, L\}$ ) in each iteration. The
  // auxiliary chain is chosen uniformly random.
  Adapt inverse temperatures  $\beta^{1, \dots, L}$  // This is performed for some PT versions based
  // on the acceptance rates of swaps between chains.
  Adapt the number of chains  $L$  // This is performed for some PT versions.
end for

```

Figure 2.2: **Pseudo-code for the AM, DRAM, PT, PHS and MALA routines as employed during Chapter 4.** The pseudo-code highlights differences between MCMC methods using comments indicated by “//” and the color-coded name of the relevant algorithm either AM, DRAM, PT, PHS or MALA.

have been introduced. These algorithm sample from multiple tempered versions of the posterior $p(\mathcal{D}|\theta)^{\frac{1}{\beta_l}}p(\theta)$, $\beta_l \geq 1$, $l = 1, \dots, L$, at the same time (Miasojedow et al., 2013; Sambridge, 2013; Vousden et al., 2016). The tempered posteriors are flattened out in comparison to the posterior, rendering transitions between posterior modes more likely. Allowing the tempered chains to exchange their position by chance enables the untempered chain, which samples from the posterior, to ‘jump’. During Chapter 4, the PT algorithm is implemented as formulated by Lacki and Miasojedow (2015) using AM with ‘acc’ adaption scheme or MALA for each tempered chain.

Different initial numbers L_0 of tempered chains, adaptive $L \leq L_0$ or fixed numbers $L = L_0$ and two different swapping strategies (Lacki and Miasojedow, 2015) were considered:

- Swaps between all adjacent chains (aa)
- Swaps of chains with equal energy (ee)

are employed. A more detailed discussion is provided along the motivation of the novel method presented in Chapter 5, as PT provides one of its core mechanics.

Parallel Hierarchical Sampling (PHS): An alternative to PT is PHS, which employs several chains sampling from the posterior (Rigat and Mira, 2012). Similar to PT, the idea is to start multiple auxiliary chains at different points in parameter space and to swap the main chain with a randomly picked one in each iteration. The main differences between PT and PHS are that all chains of PHS sample from the same distribution and that a swap between main and auxiliary chains is always accepted in PHS. The use of multiple chains can improve the mixing as different chains can employ different proposal distributions (Hug et al., 2013). During Chapter 4, for each of the auxiliary chains AM(‘acc’) is applied.

Further methods The aforementioned methods were selected with the rationale to cover the most widely employed archetypes of MCMC. For example the MALA was chosen to cover gradient based MCMC algorithms. However, beside the aforementioned algorithms there exists large set of other MCMC methods which are not quantitatively assessed in this thesis. Some of thee are described in the following.

Hamilton or hybrid Monte Carlo (HMC) methods (Neal, 2011) are MCMC algorithms which exploit gradient information to propose parameter candidates without the arguably slower mixing of the diffusion like proposal strategies other methods as the AM employ (Neal, 2011). HMC is motivated by Hamiltonian mechanics calculus from physics, where

a body possesses a certain position q , momentum p , kinetic energy K and potential energy U and where its movements are treated similar to a marble toss through a landscape via the Hamilton equations

$$\begin{aligned}\frac{dq}{dt} &= \frac{\partial H}{\partial p} \\ \frac{dp}{dt} &= -\frac{\partial H}{\partial q}\end{aligned}\tag{2.36}$$

where

$$H(q, p) = U(q) + K(p)\tag{2.37}$$

is the Hamiltonian. The potential $U(q)$ is defined by the target density, i.e. it is the negative logarithm of the posterior density,

$$U(q) = -\log(\pi(q|\mathcal{D})).\tag{2.38}$$

The kinetic energy $K(p)$ is a function of the momentum p , typically

$$K(p) = p^T M^{-1} p / 2,\tag{2.39}$$

where M is some design matrix often chosen diagonally. It was suggested to choose M so that the acceptance of the chain is $\in [0.6, 0.9]$ (Beskos et al., 2013). For the purpose of HMC, the initial position, $q_0 = \theta_{i-1}$, is defined by the last point of the chain, θ_{i-1} . The initial momentum, p_0 , is treated as a random variable, e.g. $p_0 \sim \mathcal{N}$ where \mathcal{N} is the multivariate standard normal distribution in n_θ dimensions. Given the initial conditions, the solution of the Equations 2.36 can be approximated, e.g. by Euler, modified Euler or leapfrog solvers (Neal, 2011) by evaluating the posterior multiple times depending on the discretization and step sizes used. For the usual default choice, leapfrog, the optimal step sizes were investigated (Beskos et al., 2013; Betancourt et al., 2014). In order to accept a proposed point, in HMC one must evaluate the acceptance probability,

$$p_{acc} = \min\{1, \exp(H(q_{i-1}, p_{i-1}) - H(q^*, p^*))\},\tag{2.40}$$

given the old positions and momentum q_{i-1}, p_{i-1} and the candidates q^*, p^* . The advantage of HMC is that it may significantly speed up the performance of MCMC in problems where there is a strong dependency structure in the parameters, e.g. a banana shaped mode. However, the required choice of the distribution of the momentum and the design matrix M , and the need for multiple evaluations of the posterior density $\pi(\theta|\mathcal{D})$ and its required

derivative $\partial\pi(\theta|\mathcal{D})/\partial\theta$ are drawbacks of the method. HMC algorithms are getting actively developed and novel methods, e.g. the no-U-turn sampler (Hoffman and Gelman, 2014) or wormhole HMC (Lan et al., 2014), improve the overall performance or by adding self-tuning components and novel mechanics as the prevention of u-turns or shortcuts between posterior modes to the method.

Gibbs sampling (Casella and George, 1992; Geman and Geman, 1984; Andrieu et al., 2003) is a sampling algorithm designed for obtaining samples from multivariate probability distributions by sampling from the full conditional marginals in order to improve efficiency. Gibbs sampling employs a MH algorithm with a particular structured proposal density,

$$q(\theta_j^*|\theta^{i-1}) = \pi(\theta_j^*|\mathcal{D}; \theta_1^{i-1}, \dots, \theta_{j-1}^{i-1}, \theta_{j+1}^{i-1}, \dots, \theta_{n_\theta}^{i-1}) \quad \forall j = 1, \dots, n_\theta. \quad (2.41)$$

Thus, in order to employ Gibbs sampling the full conditional target density has to be evaluable. The corresponding acceptance probability, p_{acc} , can be shown to be $p_{acc} = 1$ (Andrieu et al., 2003) which makes Gibbs sampling efficient if applicable. In cases where the full conditional probability, $\pi(\theta_j^*|\mathcal{D}; \theta_1^{i-1}, \dots, \theta_{j-1}^{i-1}, \theta_{j+1}^{i-1}, \dots, \theta_{n_\theta}^{i-1})$, is unknown, *Algorithm-within-Gibbs* methods can be applied instead. These methods employ sequential parameter updates as Gibbs sampling but use other proposal densities than the full conditional probability of a parameter, e.g. in case of Metropolis-within-Gibbs,

$$q(\theta_j^*|\theta^{i-1}) = \mathcal{N}(\theta_j^*|\theta_j^{i-1}, \sigma^2) \quad \forall j = 1, \dots, n_\theta, \quad (2.42)$$

with σ being the fixed standard deviation of the proposal density. There exists a variety of more sophisticated hybrids, e.g. HMC-within-Gibbs (Dang et al., 2017). Note, for algorithm-within-Gibbs methods, the acceptance probability, p_{acc} , does no longer necessarily equal one as for the Gibbs sampler. In general, Gibbs sampling based algorithms have problems of sampling strongly correlated or anti-correlated parameters and can get inefficient if there is a single area of high probability in parameter space. There exist several extensions of Gibbs sampling, e.g. block-wise Gibbs sampling (Djuric and Chun, 2002) or particle Gibbs (Chopin et al., 2015) to overcome problems the original algorithm possesses.

Reversible-jump MCMC methods extends MCMC so that spaces of varying dimension can be sampled (Green, 1995; Brooks et al., 2011). This is particularly useful in problems where the true number of parameters is unknown, e.g. in model selection problems (Green, 1995).

Slice sampling (Neal, 2003) is a class of MCMC algorithms similar to the class of Gibbs-

and Metropolis samplers. The proposal density of MH and Gibbs are replaced by slicing the target density $\pi(\theta|\mathcal{D})$ in order to obtain a Markov chain which samples from the posterior. Given a θ , the idea of slice sampling is to

- (i) sample y uniformly between $[0, \pi(\theta|\mathcal{D})]$
- (ii) use y to obtain all θ so that $\pi(\theta|\mathcal{D}) \geq y$, namely the set $\pi^{-1}([y, \infty))$
- (iii) to uniformly draw another θ from the set $\pi^{-1}([y, \infty))$.

Hereby, the problem of finding a good proposal density for drawing candidate points, as in MH, is replaced by obtaining slice intervals. Thus, the method naturally adapts to the shape of $\pi(\theta|\mathcal{D})$. However, in high dimensions, finding slices is not straightforward. A collection of approaches, e.g. using reflective slice sampling were discussed (Neal, 2003).

Sequential Monte Carlo methods can be used to construct proposal densities in high dimensions (Andrieu et al., 2010). The corresponding MCMC method is called *particle MCMC* (Andrieu et al., 2010). There exist tempered versions of sequential Monte Carlo (Latz et al., 2018). Another approach to construct problem specific proposal densities is *copula MCMC* (Schmidl et al., 2013). This approach uses a D-vine copula decomposition (Hofmann and Czado, 2010) of a pre-run sample to construct a joint density function of the form

$$q_1(\theta|\gamma, \eta) = c(G_1(\theta_1|\gamma), \dots, G_{n_\theta}(\theta_{n_\theta}|\gamma)|\eta) \prod_{i=1}^{n_\theta} g_i(\theta_i|\gamma), \quad (2.43)$$

where c denotes the D-vine copula density, G and g are the marginal cumulative distribution functions and densities and γ, η denote the parametrization of the copula representation found by the pre-sample. q_1 is combined with fixed MH proposals, q_2 and q_3 , for improved robustness of the method (Schmidl et al., 2013).

2.5 Uncertainty quantification

Maximum-a-posteriori estimates (as in Equation 2.3) are point estimates θ^* for the true parameter value, namely, the most likely realization of the parameter, θ_{true} , given data and prior knowledge. Unfortunately, in general θ_{true} may be non-unique in which case the parameter is called non-identifiable (Raue et al., 2010). In these cases, the estimated value θ^* is unreliable and conclusions based on the estimate have little meaning (Hasenauer, 2013). In particular, the value of θ^* may differ from estimate to estimate and each different

value leads to a different prediction which establishes wrong expectations (Hasenauer, 2013).

Non-identifiabilities can be driven by the structure of the underlying model (Chis et al., 2011; Raue et al., 2009). For example, for ODE-constrained parameter estimation (see Section 2.2 for an introduction), if the fraction of parameters θ_1/θ_2 appears exclusively throughout the ODE while θ_1 and θ_2 do not appear in other constellations, then θ_1 and θ_2 are structurally non-identifiable as a larger value of θ_1 could be compensated by a smaller value of θ_2 and vice versa in order to obtain the same output of the model. Structural identifiability analysis provides information about the topology of the model regarding its output and is independent of specific data sets (Fröhlich, 2018). Complementary, practical non-identifiability (Raue et al., 2009) arises given data. This type of non-identifiability is driven by structural properties of the model and redundancies or noise in the data.

Generally, in order to account for non-identifiabilities, an uncertainty analysis needs to be carried out. Uncertainty analysis incorporates global information about the posterior $\pi(\theta|\mathcal{D})$ beyond what point estimates, θ^* , can provide. One approach are profile calculations (Raue et al., 2010, 2013a; Stapor et al., 2017) as in Equation 2.4. Flat profiles are necessary but not sufficient for a non-identifiability to exist (Hug, 2015). Alternatively, a more comprehensive and expensive approach is sampling as it provides parameter dependencies as well (see below) and guarantees a sufficient detection of non-identifiability. In general, one is interested in either obtaining confidence regions in case of a frequentist view or credible regions in case of Bayesian statistics. Both are regions $R \in \mathbb{R}^{n_\theta}$ in the posterior landscape which provide a quantified measure of parameter uncertainty given a confidence level $1 - \alpha$ so that, e.g., in the case of a posterior credible region,

$$\int_R \pi(\theta|\mathcal{D})d\theta = 1 - \alpha. \quad (2.44)$$

Both views on the region R differ in their philosophy. Confidence regions assume the (unknown) parameter value θ_{true} to be fixed and the bounds of the confidence region to be random variables. In contrast, credible regions treat the parameter, θ , as random variable and the parametrization of R as fixed. As can be seen from 2.44, confidence and credible intervals are not unique.

There exist a variety of different methods to calculate confidence or credible regions in practice (Lu et al., 2012; Raue et al., 2010; Chen and Shao, 1999; DiCiccio and Efron, 1996; Joshi et al., 2006). Some of the methods are more precise or make fewer assumptions, e.g. by not assuming symmetric marginals (Chen and Shao, 1999). An MCMC sample can naturally be employed for obtaining credible intervals by sorting the samples

$\theta^1, \dots, \theta^n$ regarding their posterior values $\pi(\theta^1|\mathcal{D}), \dots, \pi(\theta^n|\mathcal{D})$ and taking only the first $1 - \alpha$ percentile into account for a region estimate, e.g. via kernel smoothing density estimate.

On top of parameter uncertainties, one is often interested in prediction uncertainties in order to quantify the predictive power of the model. This can be accomplished by propagating the sample through the model, y , e.g. as defined in Section 2.2, and calculating credible intervals for each time point (Hasenauer, 2013) separately.

Parameter uncertainties often include dependencies between different parameters. In this case, some directions can be determined precisely while others carry large uncertainties. This is known as sloppiness. The simplest example is a linear dependency between two parameters as strong positive or negative correlations often imply that individual parameters get very uncertain. This often happens with sums or products of parameters, e.g. as in the example θ_1/θ_2 . However, there often exist non-linear dependencies or dependencies between more than two parameters as well, which can also be seen as non-identifiabilities. For example, in Chapter 5 a 20-dimensional Gaussian ring example will be introduced, where two parameters possess a non-linear dependency and non-identifiability. There exists a variety of different approaches to reveal parameter dependencies, e.g. via principal component analysis (Jolliffe, 2002), Correlation analysis (Rodgers and Nicewander, 1988), maximum information coefficients (Reshef et al., 2011). In some cases, the knowledge about parameter dependencies can be used to identify uncertain directions and sloppiness in parameter space (Apgar et al., 2010) or for model reduction (Balsa-Canto et al., 2010).

Throughout this background chapter, the required introduction and notation for the following main chapters was introduced. In particular, Bayesian parameter estimation with application to ODE-constraint problems was formalized and Monte Carlo methods, in particular MCMC were discussed in detail. Chapter 3 will focus on the analysis and quality assessment of samples generated by Monte Carlo methods. Chapter 4 will critically evaluate a variety of standard MCMC methods in diverse ODE-constraint parameter estimation problems and Chapter 5 will introduce a novel MCMC method to improve practical inference for a broad range of problems.

Chapter 3

Quantification of sampling quality

This chapter is based on (Ballnus et al., 2017).

In Chapter 2.4.2, it was first mentioned that finite sample sizes may result in non-representative MCMC samples. While this problem is common across different applications, there is no standard procedure for the assessment of sample quality. Additionally, it is still common practice to judge sample quality by a visual (and thus subjective) inspection of the distribution marginals only. The development of reliable quality measures is naturally harder for sampling results than it is for optimization results as optimization yields point estimators of the posterior while a sample has to cover many other statistical moments (e.g. variance) of the posterior as well.

In this Chapter, the possible finite-time sampling behaviors are outlined and used to develop a semi-automated analysis pipeline suited for quality assessment of MCMC samples. The pipeline increases robustness and lowers subjectivity compared to established approaches. In particular, it allows for quantitative assessment of sampling performance across different methods with different properties. The pipeline will be used as basis for the results presented in Chapter 4 and Chapter 5.

3.1 Introduction and problem statement

MCMC algorithms are designed to be ergodic in order to converge independently of initialization towards a certain stationary probability distribution (see (Meyn and Tweedie, 2012) and Chapter 2). As ergodicity is an asymptotic property, for finite sample sizes there is no guarantee that the chain has already become representative for the posterior density. This, however, is crucial for real world applications. While it is true, that an MCMC chain will be representative at some point, it is questionable whether a finite chain has *already* become representative towards the underlying probability distribution. There are multiple reasons which may cause an MCMC chain to be not representative, yet. In practice, only few or none of the reasons are analyzed quantitatively often making quality judgment of results prone to errors and method selection for a given problem a matter of empirical experience.

For fully assessable posterior distributions, one could directly check, if a chain is representative, e.g. by comparing the empirical density estimate of the sample with the posterior density by using statistical tests as Kolmogorov–Smirnov (or multivariate extensions). Unfortunately, in most applications, the full posterior distribution is not known. Thus, one often has to purely rely on the information gained during the sampling process and judge the quality of an MCMC sample without knowing the ground truth. In literature there exist a variety of so-called convergence tests (Brooks and Roberts, 1998; Cowles and Carlin, 1996a; Mengersen et al., 1999). These rejection tests do typically assess one of the empirical properties of the chain to judge whether a chain is significantly non-representative. All of them have in common, that they can only suggest a chain is representative by not finding a significant result for the opposite. Obviously, even if such a convergence test is passed, it is not ensured that a chain is representative. Indeed, there is a high chance of systematic false negatives driven by a combination of common problems observed in MCMC samples (as discussed below) which are hard to capture by a single test at a time.

In this chapter, the lack of methods for a thorough comparison and robust, quantified assessment of MCMC results is addressed.

3.2 Symptoms of MCMC failure

In order to develop comprehensive analysis tools to assess MCMC sampling quality quantitatively, first the potential problems which may arise in applications have to be identified.

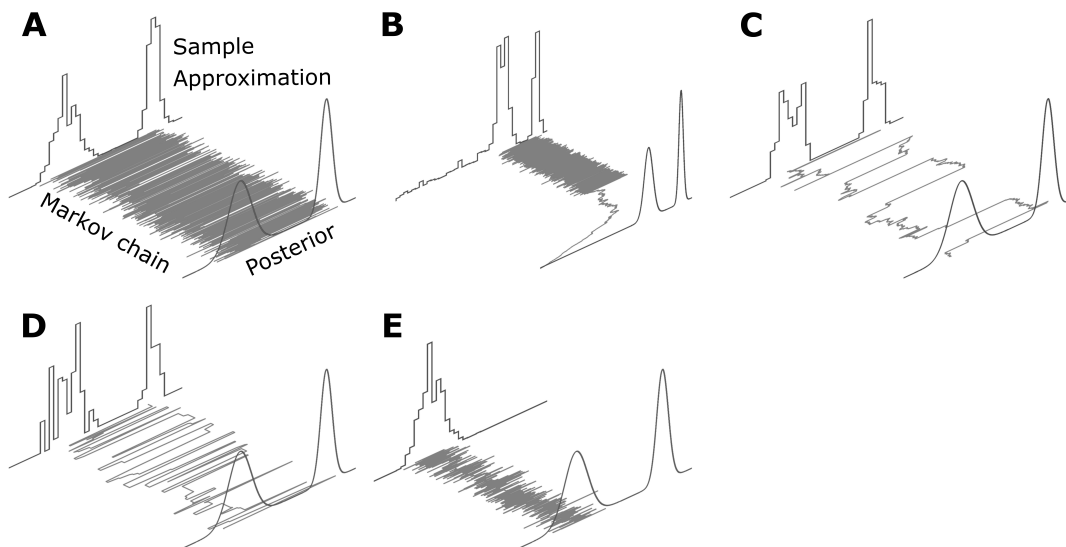


Figure 3.1: **Representative and non-representative MCMC samples of a bimodal one-dimensional posterior density.** The figure is inspired by (Andrieu et al., 2003). (A) A representative sample of the posterior. (B) Non-representative sample due to burn-in bias. (C) Non-representative sample due to high auto-correlation caused by small movements. (D) Non-representative sample due to high auto-correlation caused by a low acceptance rate. (E) Non-representative sample due to insufficient exploration.

While some of the discussed problems are covered in MCMC literature, they are typically not addressed as a whole. This is problematic because they interact, e.g. a chain which suffers from burn-in bias, will very likely possess a very low effective sample size. In the following, these terms will be introduced in detail and discussed.

3.2.1 Burn-in

A common reason for finite chains to become non-representative of the posterior distribution is a distortion in the sample due to initialization called burn-in. Hereby, the first part of a Markov chain is influenced by the starting point and, for adaptive methods, by the initial choice of the adaptation parameters (Calderhead, 2011). Please note, for finite chains burn-in may exist regardless of ergodicity and might substantially impact the results.

Informally spoken, if the chain is initialized far from the modes of the posterior density, the chain usually requires some iterations before it reaches the more likely areas of the parameter space. Depending on the given problem, algorithm and initialization, burn-in may take substantially many iterations causing a distortion of the empirical sample

regarding the true posterior as unlikely areas of the posterior are overrepresented (Figure 3.1B). In the worst case, for the given amount of MCMC iterations a chain may not be able to reach the relevant areas in the parameter space at all.

Example: Let a posterior density π be given by a 1-dimensional standard normal density

$$\pi(\theta) = N(\theta|0, 1).$$

For a point to be realized between $\theta = -0.5$ and $\theta = 0.5$, the probability is $P(\theta \in [-0.5, 0.5]) = 0.38$ while points far away from the origin have the chance $P(\theta \leq -10) = 7.6 \cdot 10^{-24}$. If the chain is initialized at -10 , it would almost certainly move quickly towards a region around 0. However, -10 would still be part of the chain. In contrast, if the chain has been started in 0, it would probably need about 10^{24} iterations to draw one single point beyond -10 . This said, from a theoretical perspective -10 is a legitimate point of the chain. Yet, for finite samples of length noticeably smaller than 10^{24} such a point adds a statistically distortion to the result.

In addition to the positional initialization of the chain, initialization of adaptive parameters may cause burn-in effects, e.g., the Adaptive Metropolis algorithm learns and scales the parameters of its proposal density on the fly (Haario et al., 2006) causing qualitative changes in the chain behavior over iterations.

To identify whether a chain possesses burn-in bias, either visual inspections or a statistical tests, as the Geweke test (Geweke, 1992) and others (Brooks and Roberts, 1998; Cowles and Carlin, 1996a; Mengersen et al., 1999) are applied. Also, burn-in related samples are often discarded (Brooks and Roberts, 1998).

3.2.2 High auto-correlation

In addition to the tests mentioned at the end of the previous section, the estimators for the Effective Sample Size (ESS) are frequently employed to assess sampling quality quantitatively. An ESS estimate is defined as the number of MCMC iterations remaining after removing iterations which are subject to the estimated auto-correlation of the chain. As each MCMC chain is generated by a stochastic process with the Markovian property, it naturally possesses a certain degree of auto-correlation. A lower auto-correlation and thus a larger ESS is desirable because it allows for shorter lengths for a chain to be representative. Please note, that thinning a chain to its effective sample is often not desirable, e.g. because it would increase the variance of the maximum a posteriori estimate (Geyer, 1992). Thus, in practice ESS should get exclusively employed for quality assessments.

For an MCMC chain there are two main sources of auto-correlation. Given, that the chain has overcome its burn-in, the first potential source lays in small step sizes (Figure 3.1C) and the second one in small acceptance rates of the chain (Figure 3.1D). Small step sizes in the Markov process makes it likely to generate chain points close to each other thus increasing auto-correlation. In contrast, due to how MCMC algorithms are designed, high rejection rates result in the same parameter points getting drawn multiple times in a row causing high auto-correlation as well. Interestingly, the concept of large step sizes and high acceptance rates often rival against each other (Link and Eaton, 2012). If the steps are large, it is typically more likely to propose unlikely points resulting in higher rejection rates.

Formally, for discrete, M dimensional stationary processes θ , for each dimension (without writing the dimension index) the autocorrelation can be defined as expectation value

$$R_\tau = \frac{1}{N} \sum_{i=1}^{N-\tau} \frac{(\theta_i - \mu)(\theta_{i+\tau} - \mu)}{\sigma^2} \quad (3.1)$$

with the posterior distribution average estimate, $\mu = 1/N \sum_{i=1}^N \theta_i$, variance estimate $\sigma^2 = R_0$, and lag $\tau \in \{0, \dots, N/4\}$ (Box et al., 2015). The expression in Equation 3.1 could be evaluated for a given MCMC result resulting in a computational complexity of $\mathcal{O}(N^2)$.

In order to reduce the computational effort one notes that – for continuous times $N \rightarrow \infty$ – the right-hand-side of Equation 3.1 becomes a convolution integral. Therefore, it is possible to apply the Wiener-Khinchin theorem (Chatfield, 2016, Chapter 6) and the right-hand-side of Equation 3.1 can be expressed using the spectral density S of θ (Box et al., 2015, Chapter 2). This yields

$$R_\tau = \mathcal{F}^{-1}[S_f]_\tau \quad (3.2)$$

where \mathcal{F}^{-1} denotes the inverse Fourier transformation. S_f has to be estimated from the chain θ . There exists a variety of estimators, either parametric, non-parametric or subspace methods. Examples are the periodogram- (Brockwell and Davis, 2016), Sacchi- (Sacchi et al., 1998b), Sokal's truncated periodogram- (Sokal, 1997), Welch- (Welch, 1967), multitaper- (Percival and Walden, 1993; Thomson, 1982) and Yule-Walker- (Friedlander and Sharman, 1985) estimator differing in computational effort and statistical precision. For example, the periodogram estimate for the spectral density S is given by

$$S_f = |\mathcal{F}[\theta_\tau]_f|^2 \quad (3.3)$$

where $\mathcal{F}[\theta_\tau]_f$ denotes the discrete Fourier transformation of the time dependent chain samples θ_τ (Brockwell and Davis, 2016). Using overall two Fast Fourier Transformations (FFTs) to obtain the quantities in Equation 3.2 and 3.3 reduces the computational complexity to $\mathcal{O}(N \log(N))$ in comparison to the complexity for directly obtaining the right-hand-side of Equation 3.1.

3.2.3 Insufficient exploration

A chain which possesses high ESS estimate and/or passes every applied burn-in test may still be not representative due to insufficient exploration of the posterior (see Figure 3.1E). The reason typically is an underrepresented part of the posterior (Andrieu et al., 2003). In literature this is often seen as special case of burn-in (Calderhead, 2007) because an underrepresented part of the parameter space does naturally cause an overrepresented part as well. Other sources call this effect pseudo-converged (Brooks et al., 2011, Chapter 1). However, in contrast to burn-in, here the driving phenomenon is that the chain misses a certain part of the posterior entirely (Figure 3.1E). This makes the detection of insufficient exploration particularly challenging – especially if only one sample is analyzed – because there often is no sign of an underrepresented part. In most cases, this renders conventional convergence tests and ESS estimates unreliable if applied to only one MCMC sample. The problem of insufficient exploration does often occur in cases where the posterior density possesses multiple modes, pronounced tails or complicated dependency structures between the parameters. Insufficient exploration is particularly dangerous for quantified comparisons of MCMC algorithms because the ESS estimates of MCMC results are often particularly good for insufficiently exploring chains (which may have passed certain convergence tests as well) because the captured structure is simpler than the full structure captured in a representative sample. For example, the chain in Figure 3.1A may have a lower ESS than the chain in Figure 3.1E.

3.3 Identification of MCMC failure

As the main result of this chapter, a semi-automated sampling analysis pipeline is introduced, which is employed as the basis for the analysis in Chapter 4 and 5. The pipeline deploys a combination of burn-in time calculation, exploration assessment, effective sample size estimation and computation time evaluation. Its main advantages are a high robustness against overestimation of sampling quality and the reduction of subjective judgment by the testing scientist. For that, the pipeline uses more than one MCMC sample from the same posterior distribution in order to deal with the particularly hard to detect insufficient

exploration.

The analysis pipeline is illustrated in Figure 3.2. The individual parts of the pipeline are based on a combination of heuristics and statistical tests which are applied sequentially in a filter-like structure. The overall idea is to only let proper samples contribute to the final ESS statistics. Thus, the final statistic could be called “conditional effective sample size”. Details of pipeline filtering process are covered in the following.

3.3.1 Calculation of burn-in

The first pipeline step addresses the elimination of the burn-in. In particular, one is interested in the last chain iteration identified as being part of the transient phase before leaving the burn-in phase denoted as N_{BI} . Once this number was identified, only the shortened chains with iteration numbers $N_{BI} + 1$ to total iterations N are considered for further analysis steps (Figure 3.2c).

In order to identify N_{BI} , the Geweke test (Geweke, 1992), which is described below and illustrated in Figure 3.3a, is employed multiple times on different parts of the same chain. The presented approach for burn-in calculation is automated using a sequence of Geweke tests taking Bonferroni-Holm adaptation (Holm, 1979) into account.

The Geweke test was originally introduced to test for convergence of the Gibbs Sampler (Geweke, 1992). However, it is suited to be applied for samples generated with other algorithms as well. The test compares the sub-sample means $\bar{\mu}_{0\%-10\%}$ and $\bar{\mu}_{50\%-100\%}$ of the first 10% and the last 50% of a chain (Geweke, 1992) while accounting for the respective spectral variance approximations $\hat{\sigma}_{0\%-10\%}^2$ and $\hat{\sigma}_{50\%-100\%}^2$ of the interval sample means. The spectral variance approximations are based on a periodogram estimate (see Section 3.2.2) using a Daniell window (Geweke, 1992). The spectral approach is employed, as it corrects the empirical variance estimate for auto-correlation between samples (Geweke, 1992). The very essence of the Geweke test is the z-score:

$$z = \frac{\bar{\mu}_{0\%-10\%} - \bar{\mu}_{50\%-100\%}}{\sqrt{\hat{\sigma}_{0\%-10\%}^2 + \hat{\sigma}_{50\%-100\%}^2}}. \quad (3.4)$$

Asymptotically, the score goes $z \rightarrow N(0, 1)$ for $N \rightarrow \infty$. In practice, this condition is reached faster if the sample has no burn-in. Thus, for sufficiently large iteration numbers N , one can derive interval probabilities, e.g. $P(z \in [-2, 2]) = 95.4\%$. For the null hypothesis that the sample means $\bar{\mu}_{0\%-10\%}$ and $\bar{\mu}_{50\%-100\%}$ are equal, those probabilities can be used to define test-thresholds for a given significance level. Indeed, the threshold $z_0 = 2$ with $z \in [-z_0, z_0]$ for a significance level of $\alpha = 4.6\%$ is a common choice in

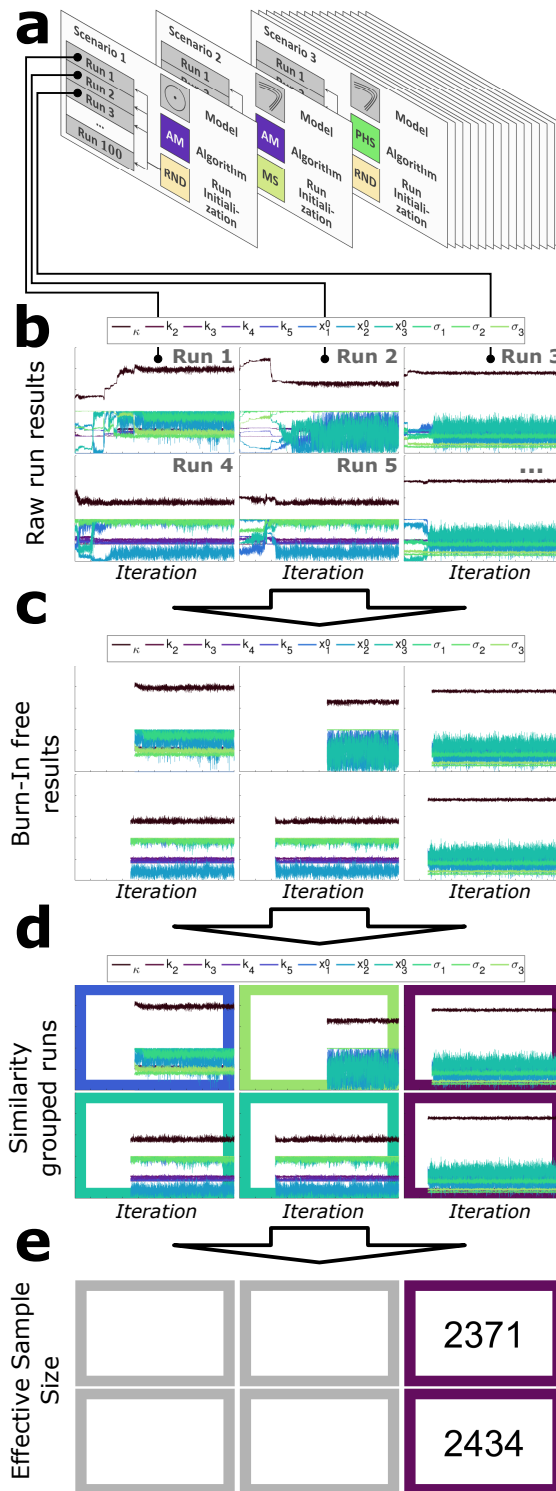


Figure 3.2: Analysis pipeline for the quantitative comparison of sampling methods featuring one of the problems from Chapter 4 as an example. (a) Multiple MCMC runs per unique problem are selected. (b) Diversity of raw results as trace plots of all parameters. (c) Automated removal of burn-in. (d) Similarity grouping of runs. Each frame color belongs to a group of similar chains. (e) Identification of groups with good exploration by group-wise comparison.

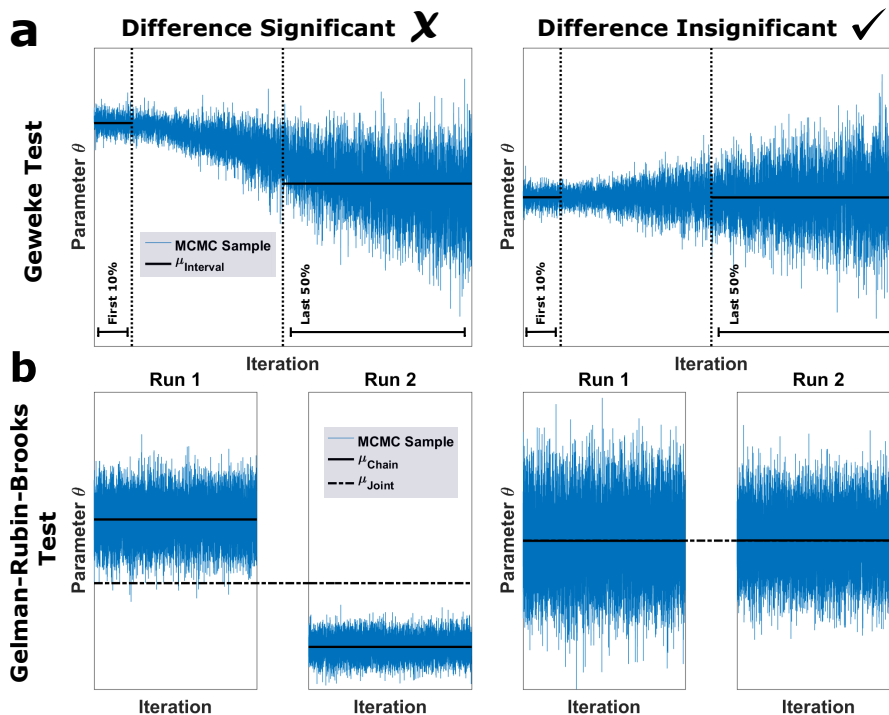


Figure 3.3: **Visual representation of statistical chain diagnosis.** (a) Geweke test. (b) Gelman-Rubin-Brooks test.

literature. Thus, empirically found values for z beyond the interval $[-2, 2]$ are statistically significant and the null hypothesis gets rejected meaning that the MCMC chain probably is subject to burn-in.

For automatic detection of the burn-in length, the raw chain is split into 40 equally sized intervals. Then, a Geweke test is performed on subsets of these intervals. Initially, the Geweke test is performed on the chain composed of all segments $1, \dots, 40$. If the resulting z -score is significant, the chain composed of the second to the last segment $2, \dots, 40$ is tested. If the z -score is still significant, the third to the last segment $3, \dots, 40$ is tested and so on until all sub-chains were tested or the z -score falls below the threshold.

To account for multiple testing, the significance threshold z_0 is adjusted for each subsequent test with the Bonferroni-Holm correction (Holm, 1979). For that, the obtained z -scores are sorted in descending order (as they are anti-proportionally related to the p-values, which should be sorted ascendingly) and the corresponding z_0 of the test are normalized as $\tilde{z}_0 = z_0/k$ where k is the sorted index $k \in \{1, \dots, 40\}$ of the test.

Once z is insignificant regarding the level \tilde{z}_0 , all iterations with $i \leq N_{BI}$ are chosen to

be discarded. As MCMC chains are typically multidimensional with $N_\theta > 1$ while the Geweke test is a univariate test, all parameter dimensions are tested individually and then the worst case is used, i.e. the highest burn-in iteration found for all dimensions is taken.

3.3.2 Assessment of exploration quality

Following the discussion in Section 3.2.3, the second pipeline step for the analysis of sampling results is the assessment of chain exploration. Due to a typically unknown posterior, the exploration of a chain has to be verified indirectly by using chain properties only while having no access to the ground-truth. Unfortunately, insufficient exploration can be rarely detected by tools which are solely based on the chain properties, e.g. by ESS calculations or burn-in tests. However, there is a possible solution for the assessment of chain exploration inspired by a multi-start local optimization (Raue, 2013). A Multi-start local optimization yields point estimates of the posterior which are easy to compare. A comparison of multiple differently initialized MCMC runs may reveal insufficient exploration of individual chains. However, comparing samples is more difficult than comparing point estimates and will be discussed in detail below. The presented approach shifts the analysis from the level of an individual run to the level of a run statistic in the overall problem. Thus, this pipeline step demands multiple runs which may be a problem if computational resources are limited. The fraction of runs which pass this pipeline step is denoted as Exploration Quality (EQ).

In order to obtain the EQ, individual MCMC run results are classified into groups based on certain similarity measures discussed below (see Figure 3.2d). Groups which are smaller than 5% of all runs are neglected from further analysis in order to simplify the next steps. For each of the remaining groups, it is assessed whether the posterior is explored by the group members by comparing the groups with each other. In particular, for each group it is evaluated if (i) all regions of high posterior probability and (ii) tails, found in the other groups, have been covered. In this way, one can tell if a group has missed relevant parameter regimes found by others. This facilitates the selection of the group(s) with the best exploration properties and the inspection of groups replaces the inspection of individual chains, resulting in improved efficiency and decrease of subjective judgment regarding chain convergence.

The grouping is based on a pairwise distance measure between chains using a combination of multivariate Gelman-Rubin-Brooks and Geweke diagnostics (Brooks and Gelman, 1998; Geweke, 1992). If both tests are passed, the corresponding runs are assumed to be similar. Each time two runs are identified as similar, they form a group. If one of the members

of a group is classified as similar to a run not yet included in the group the latter run is assigned to the entire group as well. As this is a rather greedy approach, this procedure is not independent of the amount of runs getting analyzed. Here, careful adjustment of the distance measure thresholds may help. In general, one should aim to tune the threshold so that there is a tendency towards more and smaller groups.

While the Geweke test considers differences in the means of two signals (see Section 3.2.1 for a formal introduction), the Gelman-Rubin-Brooks diagnostic focuses on within-chain and between-chain variance comparison (see Figure 3.3b for a visual representation). The convergence diagnostics consider selected summary statistics, mostly the sample means, and might miss differences in the samples which are easy to spot visually (see, e.g. the accepted cases in Figure 3.3 (right panel)). Therefore, here a combination of such methods is applied to increase the conservativeness of the pipeline step. In order to further enhance the robustness of the similarity grouping, additional tests could be employed (with adequate accounting of multiple testing cumulation). Other than in Section 3.2.1, here the Geweke test is only applied once between two *different* chains. Furthermore, in order to compare unbiased the chains are getting preprocessed by shortening them regarding their burn-in and thinning them to the size of the shorter one if necessary (see Figure 3.2c-d).

The Gelman-Rubin-Brooks (GRB) test compares the variance-covariance matrix W of θ within a set of chains with the variance-covariance B of θ between the set of chains (Brooks and Gelman, 1998). Here $\theta_i^{(jt)}$ is the i th element of the parameter vector in chain j at iteration t . The variance-covariance matrix within the chains is defined as

$$W = \frac{1}{(n-1)m} \sum_{j=1}^m \sum_{t=1}^n (\theta^{(jt)} - \bar{\theta}^{(j)}) (\theta^{(jt)} - \bar{\theta}^{(j)})^T \quad (3.5)$$

and the variance-covariance matrix between the chains as

$$B/n = \frac{1}{m-1} \sum_{j=1}^m (\bar{\theta}_j - \bar{\theta}) (\bar{\theta}_j - \bar{\theta})^T. \quad (3.6)$$

with means

$$\bar{\theta}^{(j)} = \frac{1}{n} \sum_{t=1}^n \theta^{(jt)} \quad \text{and} \quad \bar{\theta} = \frac{1}{nm} \sum_{j=1}^m \sum_{t=1}^n \theta^{(jt)}. \quad (3.7)$$

In this formula, n is the number of iterations and m the number of chains being compared (here always $n = 2$). Any rotationally invariant distance measure between W and B/n can be used to determine if the chains are sufficiently similar or not (Brooks and Gelman,

1998). Brooks and Gelman proposed

$$R = \frac{n-1}{n} + \frac{m+1}{m}\lambda, \quad (3.8)$$

as distance measure where λ is the largest eigenvalue of $W^{-1}B/n$. As W and B are assumed to converge towards the same matrix for large n , the expression in 3.8 goes to 1 asymptotically. This multivariate measure is an upper bound for its univariate counterpart (see (Brooks and Gelman, 1998)).

The conservativeness of both diagnostics can be controlled by modifying the testing thresholds R_0 and z_0 . The tests are passed, if $R < R_0$ and $|z| < z_0$. Empirically, $R_0 = 1.05$ and $z_0 = 0.05$ were found to be fairly conservative thus making it more likely to overestimate the number of groups than to underestimate it.

While computational resources are often limited, a large statistic of runs is beneficial for this analysis pipeline step, as it increases the certainty regarding the quality of passing groups of chains. For example, in Chapter 4 100 runs per combination of algorithm and benchmark problem were evaluated, 2300 runs per benchmark problem in total. This size was found to be sufficiently large for robust assessments.

3.3.3 Conditional effective sample size

For the groups of well exploring members the ESS is computed (Calderhead, 2011; Girolami and Calderhead, 2011; Schmidl et al., 2013). The ESS accounts for the in-chain autocorrelation, which was introduced in Section 3.2.2, and is an important measure for the quality of the posterior approximation of individual chains. The ESS is often overestimated if chains miss individual modes of the posterior density. Thus, here only chains assigned to groups which explore the posterior well are considered (see Figure 3.2e). For these chains, autocorrelation for individual parameters θ_i is determined using Sokal's adaptive truncated periodogram estimator (Haario et al., 2006; Sacchi et al., 1998a) which is implemented in the DRAM toolbox (Haario et al., 2006). As this is a univariate measure, the maximum of the autocorrelation across all θ_i is taken into account in order to determine the ESS via

$$ESS = \max_{i=1, \dots, N_\theta} \left(\frac{N}{\tau_i} \right) \quad (3.9)$$

while the auto-correlation lag τ is chosen for each of the parameter dimensions $i = 1, \dots, N_\theta$ via

$$\min_{\tau} \left\{ \tau \in \mathbb{N} : -\frac{1}{3} + \sum_{t=1}^{\tau} \left(\frac{R_t}{R_0} - \frac{1}{6} \right) < 0 \right\}. \quad (3.10)$$

Hereby the auto-correlation R_t is estimated as described in Section 3.2.2.

3.3.4 Computation time

Different sampling methods demand different computational efforts. For example, MALA requires gradient information while multi-chain methods require multiple evaluations of the (tempered) posterior probability in each MCMC iteration. To account for these differences, the conditional ESS is evaluated per central processing unit (CPU) second, which provides a comparable measure of computational efficiency. Furthermore, the efficiency reduction caused by runs which lack proper exploration is taken into account as well. Therefore, the ESS/s value of each run is multiplied with the EQ fraction of all run. This normalization is chosen because bad runs are sometimes much faster in execution than well behaving runs, e.g. a run only proposing parameter values outside the parameter bounds is extremely swift since neither cost function nor gradients are calculated.

3.4 Example analysis: Multiple posterior modes

In general, the analysis of MCMC samples is not straight-forward. In the preceding sections, an MCMC analysis pipeline was motivated and introduced with the aim to facilitate an automated and robust analysis of samples. To showcase the differences between state-of-the-art approaches and the pipeline, 15 MCMC chains with 10^6 iterations were generated in a certain posterior. As the qualitative properties of the posterior are sufficient for this example, it gets formally introduced later in Section 5.3.1. The posterior possesses two mass equal modes in the first two of twenty parameters θ_1 and θ_2 . In Chapter 4, it will be shown that such a structure is challenging for state-of-the-art MCMC algorithms. Thus, it is expected, that some of the results are non-representative and it is interesting to see, how standard approaches and the pipeline do handle this. At this point it does not (yet) matter which algorithms were used to generate the chains, however, the results were selected so that their representativity differs in a visual inspection (Figure 3.4, first two rows).

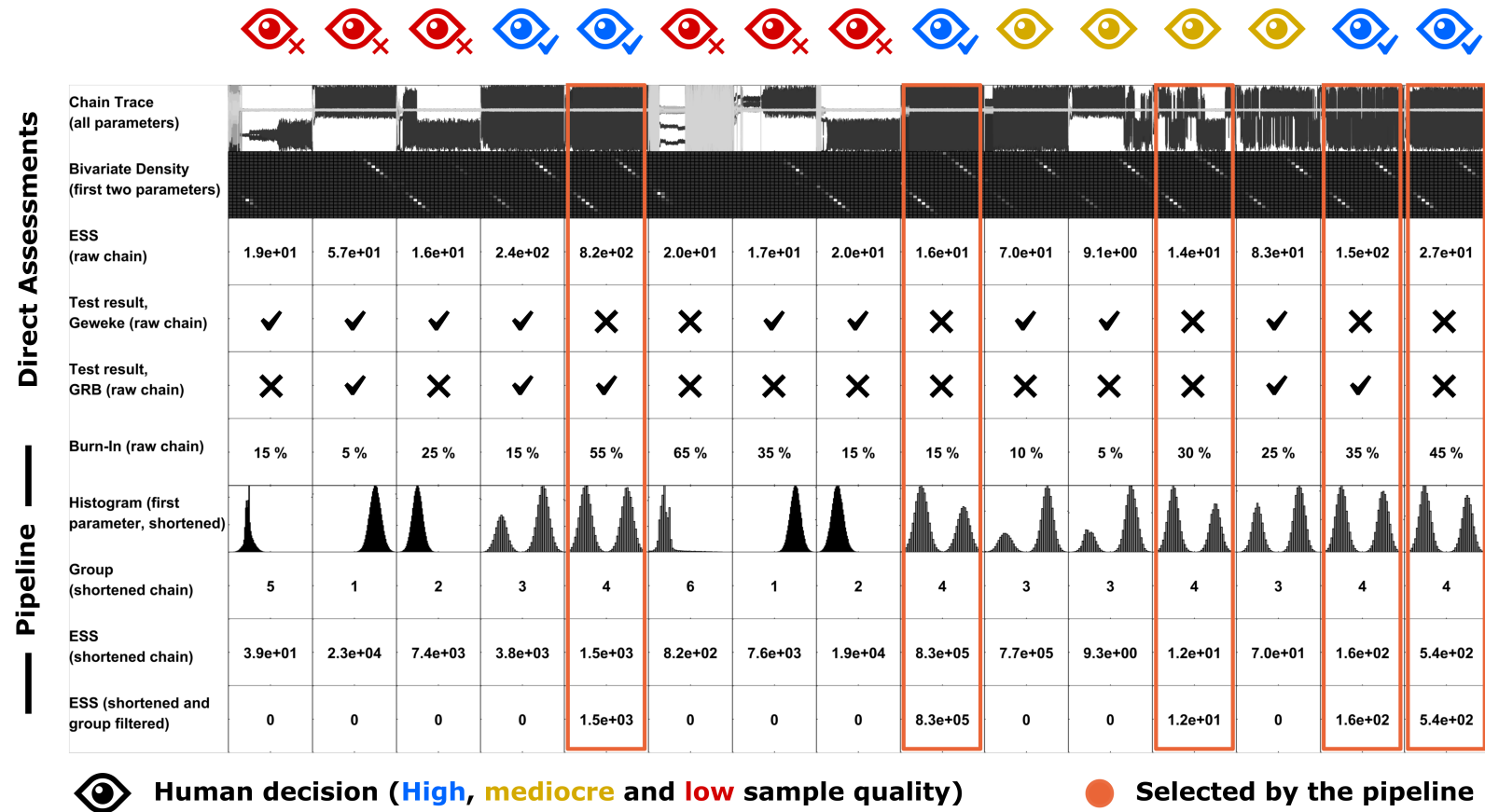


Figure 3.4: Comparison of standard MCMC quality assessments and pipeline framework using sampling results from a bimodal posterior. The samples are presented as one-dimensional parameter traces and bivariate density plots. Only some of them cover both posterior modes, indicated by jumps in the trace and two bright spots in the bivariate density plot.

In order to account for sampling quality assessments found in the literature, an ESS estimation, Geweke test and Gelman-Rubin-Brooks test were applied to the raw chain (Figure 3.4, rows 3–5). Due to the existence of burn-in in the samples, the ESS values estimated on raw results are all fairly low and do not correspond to a visual inspection. The reason for that observation is that non-representative samples in the beginning of the chain (burn-in) may impact the whole auto-correlation estimate significantly. At the same time – as can be seen visually – some of the chains do only sample from one of the two modes while the overall ESS is as high as for the chains sampling from both modes. Thus, in those cases the ESS gets overestimated.

The two convergence tests applied to the raw chains, do not necessarily agree with each other. The GRB test appears to be slightly more conservative. Here, the tests were applied with $z_0 = 2$ and $R_0 = 1.05$. Both tests struggle to identify cases where only one mode is sampled reliably. This was expected, because a run only sampling one of the two modes can not be identified by using within-chain properties only (see Section 3.2.3). On the other hand, runs, which yield an almost perfect weight balance between the two modes would be rejected in all cases. The most likely reason for those test results is the distortion in sample moments caused by the burn-in (see Section 3.2.1).

In order to apply the analysis pipeline, the first step is to estimate the burn-in phase. Across the presented runs, the pipeline method selects ranges between 5% and 55% of the chains as being subject to burn-in (Figure 3.4, row 6). For some of the runs, this selection is fairly intuitive compared to visual assessments, for others – in particular cases, where both modes were captured – the selection is less intuitive and often rather conservative. Employing the according shortened samples, one can inspect the adjusted weight relation between both modes captured in the remaining sample by plotting a histogram of thee (Figure 3.4, row 7). Knowing that both modes should have equal mass by construction, only some of the histograms show reasonably representative samples.

The shortened chains are also employed to perform similarity clustering. In total, the pipeline identified 6 different groups of similar chains. A visual inspection reveals, that group 4 has members with most equal mode weights. It has 5 members. The second best group is group 3 whose members do sample from both modes, but miss the correct weighting. The other groups are smaller and consist of chains sampling from only one of the two modes. In an application (and thus without knowing the true weighting of the modes is perfectly even) one would choose either group 3 and 4 or just group 4 as all other groups do miss at least one of the two modes. However, taking into account additional information as posterior values and local covariance structures of the posterior

Table 3.1: Confusion matrix of Geweke, GRB and pipeline decisions regarding the ground truth based on human decisions in the bi-modal posterior example. Please note, mediocre results were counted as positives.

Geweke	Positive	Negative	GRB	Positive	Negative
True	4	1	True	4	5
False	5	5	False	1	5
	Pipeline	Positive	Negative		
	True	5	6		
	False	0	4		

modes would probably be sufficient to justify the exclusive selection of group 4 in this problem.

Depending on the previous steps, the ESS estimate is calculated. Now that the burn-in and obviously non-representative groups were removed, the ESS gets estimated more accurately compared to the naive ESS calculation. For example, the three most-right selected runs have a significantly lower ESS estimate than the first two selected runs due to differences in the jumping frequency between both modes. However, the results are still not perfect. For example, the second selected run yields the highest ESS while its mode weights are a bit off.

In conclusion, the pipeline related analysis draws an overall correct picture and yields more accurate estimates than the direct assessments (Table 3.1) as these basically fail to capture any correct quality conclusions. However, as the sample sizes are rather small and just one posterior example was presented general conclusions can not be drawn. Additional testing has been performed with the applications outlined in Chapter 4. In particular, the analysis pipeline was successfully applied while observing that its decisions are congruent to human expectations for 16100 different MCMC runs.

3.5 Summary and discussion

The result of this chapter is an analysis pipeline facilitating robust assessment of sampling quality in the presence of burn-in, insufficient exploration quality and diverse ESS ranges. The pipeline founds the basis for the analysis carried out in Chapter 4 and 5.

In the beginning of this Chapter, the common problems of MCMC samples, burn-in, high auto-correlation and insufficient exploration were introduced. These discussions were used as a basis to motivate a sampling pipeline suited to assess each of the potential problems

in one framework. The benefits of the pipeline were demonstrated in an example at the end of the Chapter by comparing its analysis results to standard approaches in a simple structured bimodal posterior. Indeed, the proof of concept highlights the need for a combination of multiple sampling quality assessments and the benefits of taking into account more than one sample in order to obtain reliable quality estimates.

The pipeline can be technically employed regardless of the employed algorithm or the application problem, in particular, when the full posterior distribution is unknown (as typically being the case in applications). The pipeline does also limit the subjective influence of the performing scientist, as the inspection of individual chains is replaced with the inspection of topologies of groups of chains. Overall, this makes the pipeline a good tool for benchmarking.

There is still room for improvement regarding automation, e.g. by fully eliminating the need for a visual inspection of the groups. As the pipeline consists of multiple sequential procedures, it is currently not possible to address certainty about the overall analysis result obtained by the pipeline, e.g. by providing a p-value. However, the certainty does increase with the number of chains taken into account making it possible to exploit parallelized MCMC runs for increasingly reliable conclusions. Each of the pipeline steps can be further improved by taking into account additional or more fine tuned statistical tests and automatically tuned significance levels, e.g. by making the thresholds for the similarity grouping depend on the overall number of analyzed chains.

Chapter 4

Benchmarking of MCMC in ODE-constraint models

This chapter is based on (Ballnus et al., 2017).

In quantitative biology, mathematical models are employed to describe and analyze biological processes. The parameters of these models are usually unknown and need to be estimated from experimental data using statistical methods (see Section 2.2). For this, MCMC methods have become increasingly popular. A broad spectrum of MCMC algorithms have been proposed (summary in Section 2.4). However, selecting and tuning sampling algorithms suited for a given problem remains challenging and a comprehensive comparison of different methods is so far not available.

In this chapter, the results of a thorough benchmarking of state-of-the-art single- and multi-chain sampling methods are presented, including Adaptive Metropolis, Delayed Rejection Adaptive Metropolis, Metropolis adjusted Langevin algorithm, Parallel Tempering and Parallel Hierarchical Sampling. Different initialization and adaptation schemes are considered. To ensure a comprehensive and fair comparison, ODE-constraint problems with a range of features such as bifurcations, periodical orbits, multistability of steady-state solutions and chaotic regimes were considered. These problem properties give rise to various posterior distributions including uni- and multi-modal densities and non-normally distributed mode tails. For an objective comparison, the analysis pipeline introduced in Chapter 3 is employed. Some of the insights gained in this chapter are accumulated in Chapter 5 to develop a novel MCMC method, RAmPART.

4.1 Introduction and problem statement

For the evaluation of optimization methods, large collections of benchmarking problems were established to facilitate a fair comparison of methods (see, e.g. (Villaverde et al., 2015)). Furthermore, optimization toolboxes are available and provide access to a large number of different optimization schemes (Kronfeld et al., 2010; Egea et al., 2014). The availability of both, benchmark problems and toolboxes, is more problematic for sampling methods. Apparently, there is no similar effort for establishing a collection of benchmarking problems for sampling methods in particular featuring dynamical systems.

In this chapter, state-of-the-art MCMC algorithms are evaluated by comprehensive means in multiple ODE-constraint problems in order to inform future MCMC quality assessments, method development and selection of methods for a given problem. The aforementioned needs are addressed by (i) providing generic implementations for several MCMC algorithms and (ii) compiling a collection of benchmark problems. For a discussion of the MCMC methods, in particular AM, DRAM, PT, PHS and MALA, and a the pseudo-code employed in this Chapter, please refer to Section 2.4.

It is expected, that chain initialization is crucial for the overall performance of a MCMC method (see Chapter 3.2.1). Thus, it is evaluated how the additional effort of a preceding *multi-start local optimization* (Raue et al., 2013b) based initialization does impact the overall performance of the methods in comparison to a prior-based initialization. A detailed discussion is covered in Section 4.2.

The sampling methods are evaluated in a collection of ODE-constraint benchmark problems featuring dynamical systems with different properties such as periodic attractors, bistability, saddle-node, Hopf and period-doubling bifurcations as well as chaotic parameter regimes and non-identifiabilities. This implies posterior densities with uni- and multi-modal, pronounced tails and non-linear dependency structures of parameters. This collection of features which are commonly encountered in systems biology facilitates the evaluation of the sampling methods under realistic, challenging conditions. To ensure realism of the evaluations, knowledge about the posterior distribution, which is not available in practice, is not employed for selection, adaptation or tuning of methods. The problems will be introduced in Section 4.3.

To ensure a rigorous and efficient evaluation of sampling methods in multiple benchmark problems, the semi-automatic analysis pipeline developed in Chapter 3 is employed. This enabled the evaluation of $> 16,000$ MCMC runs covering a wide spectrum of sam-

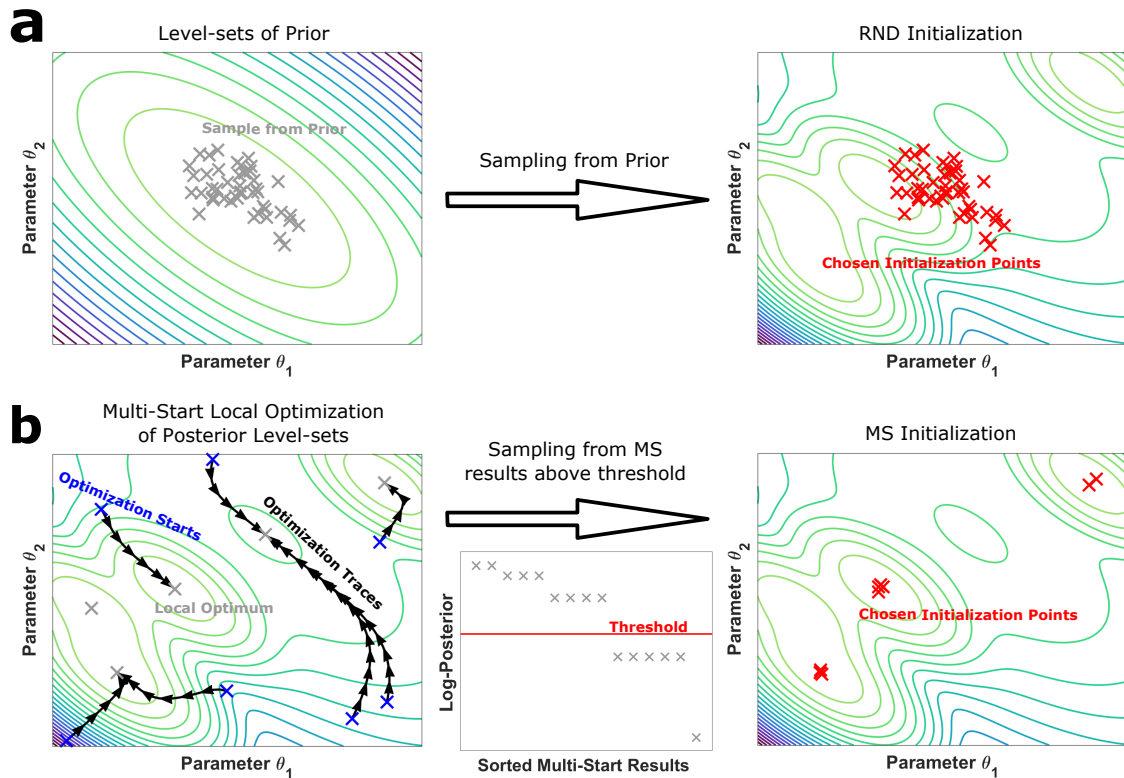


Figure 4.1: **Graphical representation of initialization schemes.** (a) Drawn from the prior distribution. (b) Drawn from the best results of a multi-start local optimization.

pling methods and benchmarks. Overall, this comprehensive assessment required roughly 300,000 CPU hours. The benchmark is organized into several *scenarios*. Each scenario consists of a unique combination of one algorithm with a particular tuning performing in a specific benchmark problem possessing 100 MCMC runs in total.

4.2 Initialization

The performance of sampling methods may be sensitive to their initialization (Andrieu et al., 2003). Here, two alternative initialization schemes are compared: Initialization using samples from the prior distribution; and initialization using multi-start local optimization results. The two methods are illustrated in Figure 4.1.

As the initialization schemes slightly differ between single- and multi-chain algorithms, an illustration of the different initialization schemes (Figure 4.2) visualizes the differences.

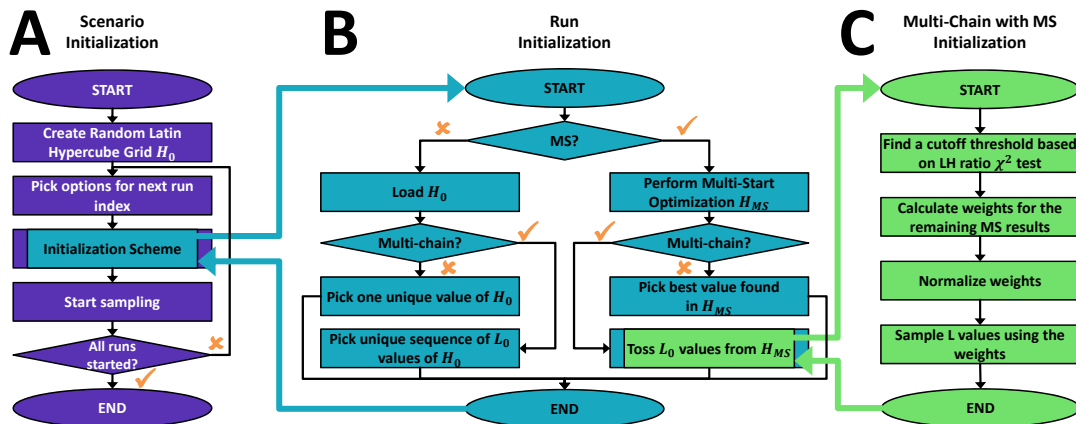


Figure 4.2: **Sampling and initialization workflowchart.** (A) Flowchart for sampling. (B) Flowchart for initialization of sampler, distinguishing single- and multi-chain methods as well as random and optimization-based initialization. (C) Flowchart for the selection of starting points for optimization-based initialization.

The schemes are discussed in detail in the following.

Sampling From Prior Distribution (RND): In many applications, sampling is initialized with parameters drawn from the prior distribution (Figure 4.2B, left column). As the prior distributions are often available in closed-form, this is usually straightforward and computationally inexpensive. Here, 100 parameter points, H_0 , are drawn from the uniform priors. For single-chain algorithms, these points are distributed to the 100 runs per scenario. For multi-chain algorithms, different (but overlapping) sequences of L points from the 100 available points were taken into account and distributed onto the L chains of each run. Each scenario has a different random seed.

Multi-start Local Optimization (MS): Sampling from the prior distribution $\pi(\theta)$ frequently yields starting points with low posterior probability. Sampling methods started at these points may require a large number of iterations to reach a parameter regime with high posterior probabilities (effectively causing a long burn-in as discussed in Chapter 3). To address this problem, initialization using multi-start local optimization has been proposed (Hug et al., 2013). The results of multi-start local optimization H_{MS} provide a map of the local optima of the posterior density where the frequency of occurrences of a local optimum is connected to the size of its basin of attraction.

To initialize the multi-start local optimization, for each run 1000 parameters vectors are drawn from the prior distribution (as presented in (Raue et al., 2013a)). For each starting

point, the local optimization is performed using the MATLAB build-in routine *fmincon* with ‘interior-point’. In the following, the results H_{MS} of such a multi-start local optimization are used to inform the initialization of an MCMC run (Figure 4.2B, right column).

Single-chain methods are initialized at the local optimum with the highest posterior probability corresponding to the maximum a posteriori (MAP) estimate. For multi-chain methods, the scheme is more complicated than for single-chain methods (Figure 4.2C), as by means of experience each sub-chain should be initialized differently to decrease the redundancy across sub-chains. To prevent initialization at points with low posterior probability, the optimization results are getting filtered based on the difference to the best optimization result. Then, initial points for each of the individual sub-chains are sampled from the remaining results. In particular, only the best $u \leq 1000$ optimization results $\{\theta^{(j)}, j = 1, \dots, u\}$ are taken into account for the MCMC initialization. To determine a reasonable number u , the sorted results are used to obtain the fraction

$$R_i = 2 \log \left(\frac{\pi(\theta^{(1)}|\mathcal{D})}{\pi(\theta^{(i)}|\mathcal{D})} \right), \quad (4.1)$$

$i = 1, \dots, 1000$, such that

$$\pi(\theta^{(1)}|\mathcal{D}) \geq \pi(\theta^{(2)}|\mathcal{D}) \geq \dots \geq \pi(\theta^{(1000)}|\mathcal{D}) \quad (4.2)$$

holds. From this sequence u is obtained such that for all $i = 1, \dots, u$, R_i is larger than the inverse χ^2 distribution with one-degree of freedom with an α -level of 0.001. This is motivated by the likelihood ratio test (Hross, 2016). After u was obtained, weights

$$w_j = 1 - \frac{\pi(\theta^{(1)}|\mathcal{D}) - \pi(\theta^{(j)}|\mathcal{D})}{\pi(\theta^{(1)}|\mathcal{D}) - \pi(\theta^{(u)}|\mathcal{D})} \in [0, 1] \quad (4.3)$$

canonically define a discrete distribution $w_j / \sum_{i=1}^u w_i, j = 1, \dots, u$. From this distribution, the initial values for each of the sub-chain of the multi-chain MCMC run are drawn.

The overall heuristic described above takes into account both, the height and area of the modes of the posterior density. The area is encoded in the frequency which a certain local optima is recovered (Figure 4.1b).

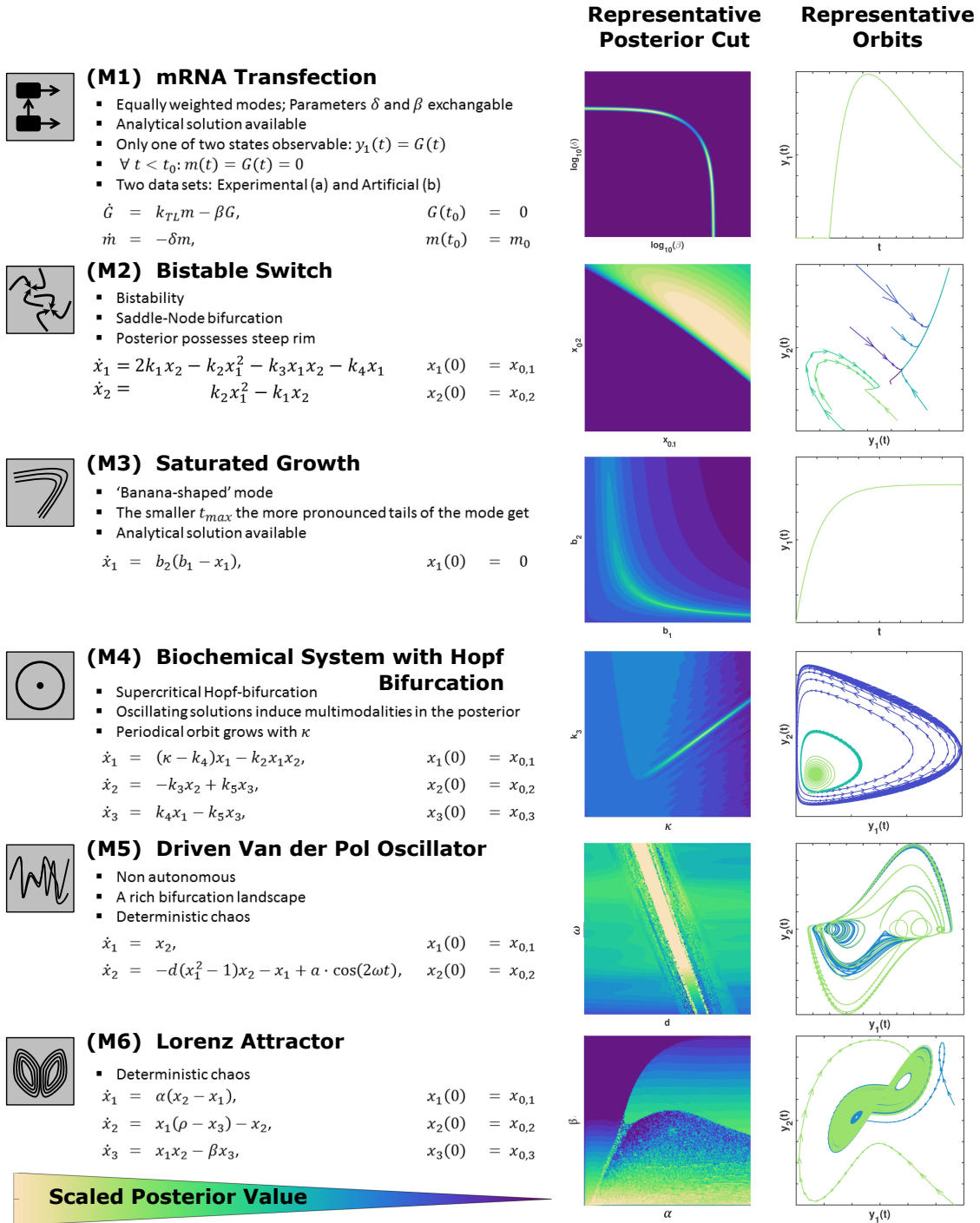


Figure 4.3: **Visual summary of benchmark problems.** (Left) ODE model and its properties, e.g. bifurcations. (Right) Illustration of system dynamics using posterior cuts and orbits.

4.3 Benchmark collection

For the evaluation of the sampling algorithms, six benchmark problems for ODE constrained parameter estimation were established. Each benchmark problem is related to behaviors found in biologically motivated ODE models. The considered problems are low dimensional. This ensures a low computational cost per evaluation and facilitates a short evaluation time to allow the large number of long MCMC runs necessary for a comprehensive comparison. Yet, the ODE models possess properties such as structural non-identifiabilities, bifurcations, limit-cycle oscillations and chaotic behavior. This yields posterior densities with pronounced tails, multi-modalities and rims which makes them difficult to sample. These are common scenarios for many application problems in systems biology (Raue et al., 2013a; Hug et al., 2013; Raue et al., 2009; Balsa-Canto et al., 2010; Chis et al., 2011; Weber et al., 2011; Gardner et al., 2000; Ozbudak et al., 2004; Tyson, 1991; Kholodenko, 2000; Calderhead, 2007; Kosuta et al., 2008; Ngonghala et al., 2016; Braxenthaler et al., 1997) which are difficult to identify prior to the parameter estimation. In the following, the properties of the benchmark problems are described. A visual summary is provided in Figure 4.3.

4.3.1 (M1) mRNA transfection

The first model describes the transfection of cells with GFP mRNA, its translation and degradation (Leonhardt et al., 2014). The observable is the protein concentration. The posterior of the estimation problem is bimodal as the exchange of the degradation rates of mRNA and protein results in the same dynamics for $G(t)$. This ODE model is studied for experimental data (M1a) and for artificial data (M1b).

4.3.2 (M2) bistable switch

This model describes a bistable switch (Wilhelm, 2009), a frequent motif in gene regulation (Chaves et al., 2008), neuronal science (Guevara, 2003) and population dynamics (Sgro et al., 2015). The parameter vector includes parameters influencing the vector field of the ODE (dynamical parameters), as well as the unknown initial conditions. Depending on the dynamical parameters, the system holds one or more stable equilibria and undergoes a saddle-node bifurcation. Given that the topology holds multiple equilibria – depending on the initial condition – the state orbit converges to one of two steady states. This leads to a steep rim in the posterior if altering the initial condition parameters. As (M2) possesses a saddle-node bifurcation, some choices for the dynamic parameters result in the absence of multiple stable equilibria. In those cases, the steep rim vanishes.

4.3.3 (M3) saturated growth

This model describes the growth of a population in an environment with limited resources. It is widely used to model population dynamics, i.e. immigration-death processes (Zimmer et al., 2015), and a variety of extensions are available. Already for the simplest model, the parameters are strongly correlated and the posterior density possesses ‘banana’ shaped tails if the measurement is stopped before the steady state is reached (Solonen et al., 2012). This effect can be enhanced by decreasing the maximum measurement time t_{\max} when creating synthetic data.

4.3.4 (M4) biochemical system with Hopf bifurcation

This model describes a simple biochemical reaction network (Kirk et al., 2008) with a supercritical Hopf bifurcation (Crawford, 1991; Kuznetsov, 2013; Dercole and Rinaldi, 2011) as found in many biological applications (Sgro et al., 2015; Heldt et al., 2002; Feinberg and Horn, 1977). Depending on the parameter values, the orbits of the system approach a stable limit cycle or a stable fixed point. The posterior density for this problem is multi-modal but most of the probability mass is contained in the main mode.

4.3.5 (M5) driven Van Der Pol Oscillator

This model is an extension of the Van der Pol oscillator with an oscillating input (Tsat-sos, 2006; Mettin et al., 1993; Parlitz and Lauterborn, 1987; Leonov et al., 2011). The input causes deterministic chaos by creating a strange attractor. Chaotic behavior can be observed in biological applications e.g. in cardiovascular models with driving pacemaker compartment (Glass et al., 1983; Heldt et al., 2002). The posterior density possesses a large number of modes of different sizes and masses. This effect can be increased by creating synthetic data with larger t_{\max} . For chaotic systems sampling is known to be very challenging (Du and Smith, 2017).

4.3.6 (M6) Lorenz Attractor

The Lorenz attractor provides an idealized description of a hydrodynamic process and can be interpreted as chemical reaction network (Poland, 1993). Similar to (M5), this system is chaotic and thus possesses a multi-modal posterior density. However, its topology strongly differs from the one of (M5) and the chaotic behavior does not arise from a driving term.

4.3.7 Data and priors

In this Chapter, benchmark settings with measured data (M1a) or synthetic data (M1b-M6) are considered. The simulated data is obtained by simulating the models for the parameters θ_{true} (Table 4.1) and adding normally distributed measurement noise. If not stated otherwise, each system state x corresponds to one canonical observable y (see Section 2.2). The prior densities are uniform¹ in the interval $\theta \in [\theta_{\min}, \theta_{\max}]$ equidistantly spaced points in time. Information about observables is provided in Figure 4.3.

4.4 Implementation








The implementation of the benchmark problems is available as supplementary material in (Ballnus et al., 2017). The sampling algorithms are implemented in the Parameter ESTimation TOolbox (PESTO) (Stapor et al., 2017). PESTO comes with a detailed documentation of all functionalities and the respective methods. For numerical simulation and sensitivity calculation the Advanced MATLAB Interface for CVODES and IDAS (AMICI) (Fröhlich et al., 2014b, 2017a) was employed. Both toolboxes are developed and available via GitHub (<https://github.com/ICB-DCM/PESTO> and <https://github.com/ICB-DCM/AMICI>). The entire code basis could be transferred to other programming languages similar to MATLAB, such as Python, Octave or Julia, without major changes. A re-implementation of the tool in R would also be conceptually possible and allow for the comparison with other packages, e.g. (Vihola, 2012).

4.5 Performance of MCMC algorithms

In this section, five state-of-the-art sampling approaches for multiple settings of tuning parameters are benchmarked in the afore-described benchmark problems. In the following, 23 scenarios (combination of algorithms and benchmark problem) are considered. To obtain reliable results and to analyze them employing the pipeline from Chapter 3, there were obtained 100 runs per scenario, thus performing 2300 runs per benchmark problem. Each run comprises 10^6 iterations of a single- or multiple chains depending on the performing algorithm. In the following, different aspects of the analysis are presented.

¹In principle, one could employ other prior densities as well.

Table 4.1: Parameters, parameter bounds and true parameter values for the benchmark problems.

	θ	θ_{\min}	θ_{\max}	θ_{true}
(M1a)  $n_t = 150$ $t \in [2, 27]$	$\log_{10}(t_0)$	-2	1	-
	$\log_{10}(k_{TL}m_0)$	-5	5	-
	$\log_{10}(\beta)$	-5	5	-
	$\log_{10}(\delta)$	-5	5	-
	$\log_{10}(\sigma)$	-2	2	-
(M1b)  $n_t = 51$ $t \in [0, 10]$	$\log_{10}(t_0)$	-2	1	$\log_{10}(2)$
	$\log_{10}(k_{TL}m_0)$	-5	5	$\log_{10}(5)$
	$\log_{10}(\beta)$	-5	5	$\log_{10}(0.8)$
	$\log_{10}(\delta)$	-5	5	$\log_{10}(0.2)$
	$\log_{10}(\sigma)$	-2	2	-1
(M2)  $n_t = 101$ $t \in [0, 200]$	k_1	2	20	8
	k_2	0	5	1
	k_3	0	5	1
	k_4	0	5	1
	$x_{0,1}$	-3	3	2
	$x_{0,2}$	-3	3	0.25
	σ_1^0	10^{-3}	1	0.3
	σ_2^0	10^{-3}	1	0.3
(M3)  $n_t = 101$ $t \in [0, 2.5]$	b_1	0	5	1
	b_2	0	5	0.2
	σ_1	10^{-3}	10^2	0.03
(M4)  $n_t = 101$ $t \in [0, 200]$	κ	1	5	3.8
	k_2	0.8	1.2	1
	k_3	0.8	1.2	1
	k_4	0.8	1.2	1
	k_5	0.8	1.2	1
	$x_{0,1}$	0	2	1
	$x_{0,2}$	0	2	1
	$x_{0,3}$	0	2	1
	σ_1	10^{-2}	2	0.75
	σ_2	10^{-2}	2	0.32
	σ_3	10^{-2}	2	0.46
(M5)  $n_t = 101$ $t \in [0, 200]$	a	2	8	5
	d	2	8	5
	ω	2	8	2.464
	$x_{0,1}$	-1	3	0
	$x_{0,2}$	-1	3	0
	$x_{0,3}$	-1	3	1
	σ_1	10^{-2}	2	0.2
	σ_2	10^{-2}	2	0.8
σ_3	10^{-2}	2	0.2	
(M6)  $n_t = 101$ $t \in [0, 200]$	α	0	20	10
	β	0	10	$\frac{8}{3}$
	ρ	10	30	28
	$x_{0,1}$	0	35	26.61
	$x_{0,2}$	-10	10	-2.74
	$x_{0,3}$	-5	5	0.95
	σ_1	10^{-4}	10^2	1
	σ_2	10^{-4}	10^2	1
σ_3	10^{-4}	10^2	1	

4.5.1 Revealing challenges in small sized application problem

To illustrate the behavior and the properties of the different sampling methods, the process of mRNA transcription ((M1), Figure 4.4a) is considered. This process has been modeled and experimentally assessed by (Leonhardt et al., 2014). The ODE model possesses two state variables and five parameters. Structural analysis using the MATLAB toolbox GenSSI (Chis et al., 2011) indicated one structural non-identifiability but did not reveal its nature. (Leonhardt et al., 2014) derived the analytical solution of the ODE model and showed that the parameters β and δ can be interchanged without altering the output y . This implied that the parameters are locally but not globally structurally identifiable, giving rise to a bimodal posterior density (Figure 4.4b,c). As the analytical solution is in general not available, here, the information about the interchangeability of β and δ for the initial assessment was disregarded as well.

The posterior distribution was sampled using several single- and multi-chain methods as well as settings and initialization schemes. The analysis of the sampling results revealed that many methods fail to sample from both modes of the posterior within 10^6 iterations (see Figure 4.4d,e). Accordingly, the exploration quality (see Chapter 3 for definitions) of many methods is low (Figure 4.4f). The single-chain methods, AM, DRAM and MALA, were expected to always sample close to the starting point, which was indeed the case. Interestingly, it was found that PHS often succeeded in moving its chain between both modes but failed to explore the mode tails properly. Merely PT, either MS or RND initialized, captured both modes in most runs (Figure 4.4f). Thus, in (M1a) the conditional ESS – the ESS for the chains sampling both modes and the tails – was > 0 only for PT.

This application example highlights challenges arising from missing information about parameter identifiability and limitations of available sampling methods. Some of these limitations were not encountered in the manuscripts introducing the methods (e.g. (Lacki and Miasojedow, 2015) or (Rigat and Mira, 2012)) as the study focused on different aspects or only considered problems for which the proposed algorithms were well-suited. The analysis of (M1a) demonstrates that even simple linear ODE models can give rise to posterior landscapes that are difficult to sample. This motivates the analysis of other (small-scale) benchmark problems.

4.5.2 Boosting sampling efficiency by resolving non-identifiabilities

For most ODE constrained parameter estimation problems, information about the identifiability properties of parameters will not be available prior to the sampling. This is unfortunate as the sampling performance of all methods could be improved by exploiting

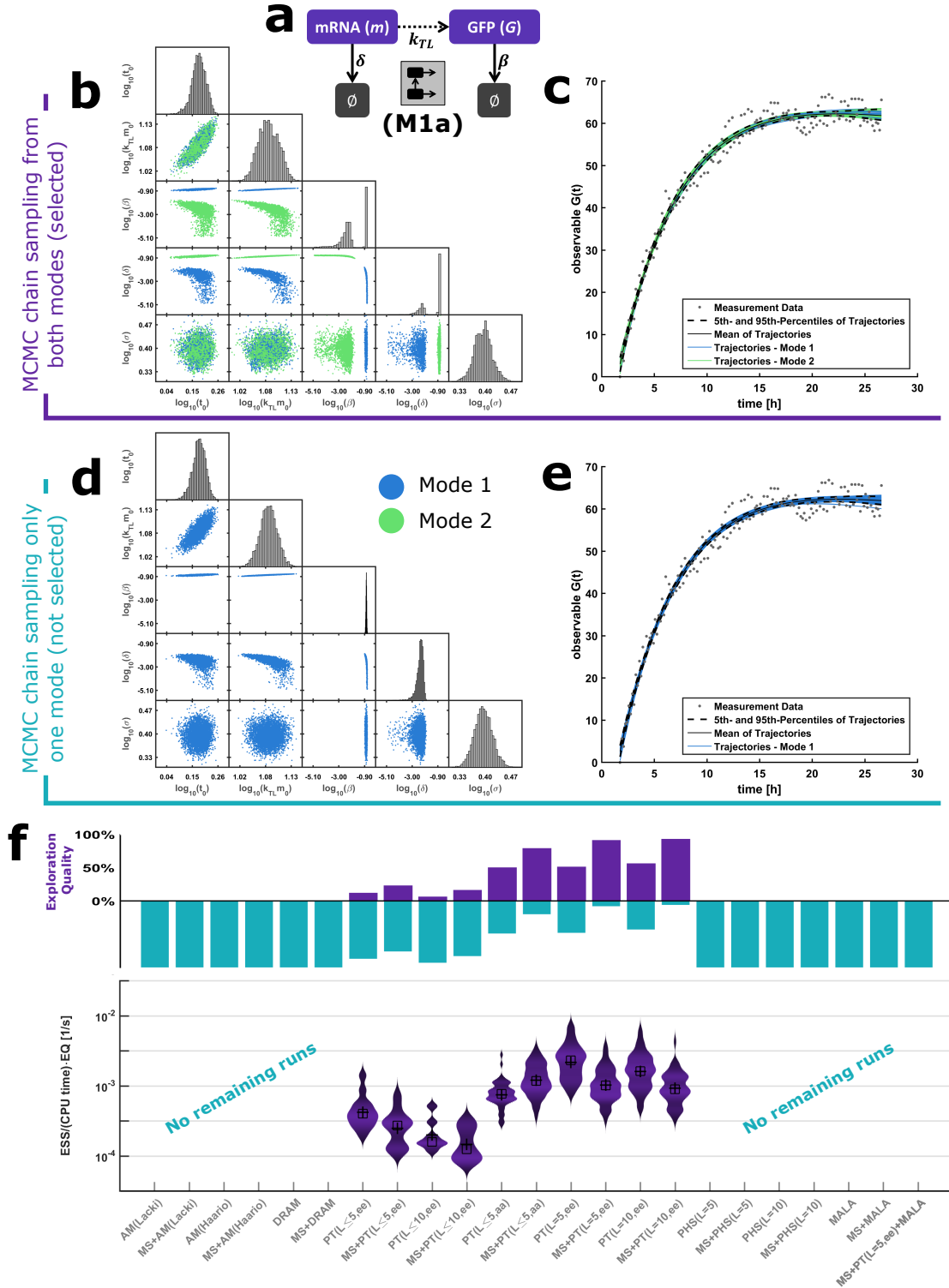


Figure 4.4: **Results from benchmark problem (M1a).** (a) Sketch of the translation process. (b) A bivariate scatter plot of a chain which explored both modes. (c) The corresponding trajectories of the sampled parameter points of both modes. (d) A representative chain which was not able to cover both modes. (e) The corresponding trajectories of the sample of one mode. (f) Effective sample sizes of chains which explored both modes. For several methods, no chain explored both modes, implying an effective sample size of zero.

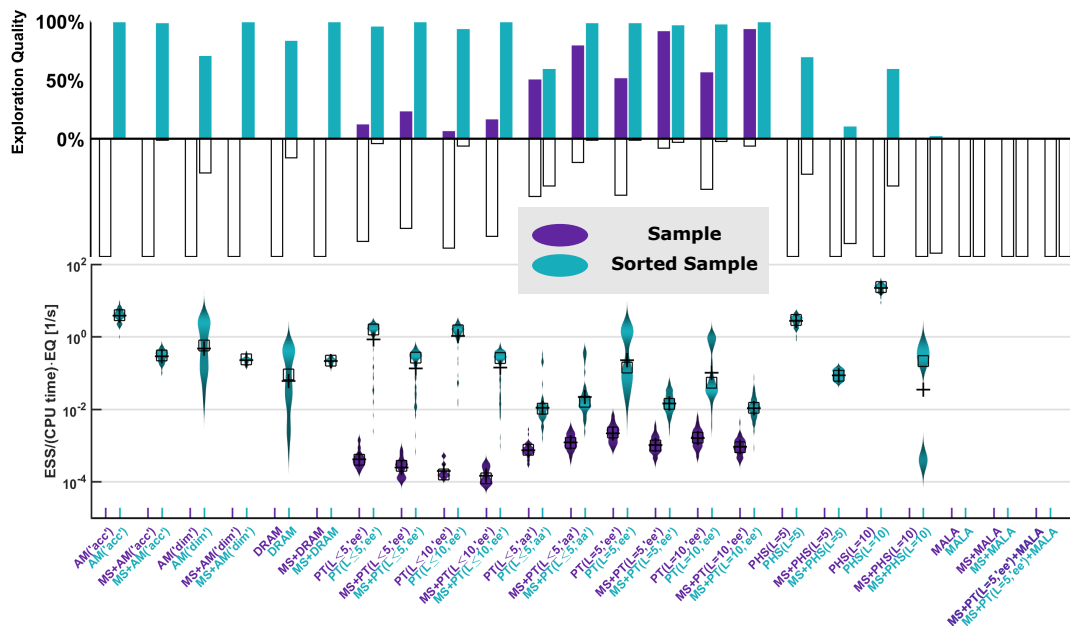


Figure 4.5: **Sampling performance comparison for (M1a).** **Upper Panel:** Exploration Quality. **Lower Panel:** Distribution of Effective Sample Size per second regarding runs which contribute to Exploration Quality. Results for the original sample and the post-processed (= sorted) sample are shown.

such additional information. Models with parameter interchangeabilities such as (M1) are well studied in the context of mixture models. Tailored methods for such problems include post-processing methods or a random permutation sampler (Jasra et al., 2005; Papastamoulis and Iliopoulos, 2013).

As mentioned in Section 4.5.1, the parameters β and δ are interchangeable. Here, the benefits of applying a post-processing strategy was evaluated. Instead of applying the analysis pipeline to the raw chains, the chains were first processed by switching β and δ so that $\beta > \delta$ holds. The resulting increase in EQ and ESS per second are shown in Figure 4.5. The sampling performance improves significantly for almost all algorithms. The single-chain algorithms benefit the most. This highlights the importance of usage of additional information for sampling problems – if available.

4.5.3 Influence of posterior properties on sampling performance

The former analysis of (M1a) revealed, that small sized problems can be challenging for modern MCMC algorithms. This motivates further investigation of other small sized models (M1b)–(M6) which potentially hold interesting ODE properties, e.g. oscillating

orbits, as well. Please note, that the data used for these models is synthetic (see Section 4.3).

It was investigated how EQ depends on the benchmark problem and its properties. The size of the groups of runs identified by the analysis pipeline (Figure 4.6a) and the overall EQ (Figure 4.6b) do strongly vary between the benchmark problems. For problems with uni-modal (M2-3) and weakly² multi-modal (M4) posteriors, the average EQ of the sampling methods was estimated higher than 50%. For the problems with bimodal posteriors (M1a,b), 79% of the runs sampled from one of the modes and failed to explore the posterior, while 21% of the chains sampled from both modes and achieved a good exploration. For posteriors with strong multi-modalities (M5-6), all chains behaved differently and no large groups could be identified (Figure 4a).

In terms of the dynamical properties of the underlying dynamical system, the results for the benchmark problems indicate that state-of-the-art sampling methods work well with multiple steady states and saddle-node bifurcations, as well as Hopf bifurcations and (limit cycle) oscillations resulting in weak multi-modalities of the posterior. However, these methods fail in case of chaotic behavior and local non-identifiability resulting in strong multi-modalities of the posterior.

The analysis on the level of sampling methods revealed that for (M2-4) most algorithms worked appropriately (Figure 4.6b) while for (M5-6) all algorithms failed. For (M1), a benefit if using PT and PHS was observed. As the EQ directly impacts the ESS, these observations hold true for the ESS per CPU second (Figure 4.4f). Indeed, a strong correlation of exploration quality and sampling efficiency was found and identified as the major limiting performance factor for (M1a,b) and (M5-6) because a majority of the runs were filtered by the pipeline (Figure 4.6b).

4.5.4 Comparison of single- and multi-chain methods

Following the analysis of the differences between benchmark problems, the single- and multi-chain methods were directly compared. The average performance characteristics for single- and multi-chain methods were computed by averaging over sampling methods, initialization schemes and tuning parameter choices (Figure 4.7). It was found that for all considered benchmark problems, multi-chain methods achieved better EQs than single-chain methods (Figure 4.7a). Indeed, for several problems, multi-chain methods provided representative samples from the posterior distributions while single-chain methods sampled only individual modes. Interestingly, the improved mixing of multi-chain methods

² Meaning that there is one global mode and several small side modes.

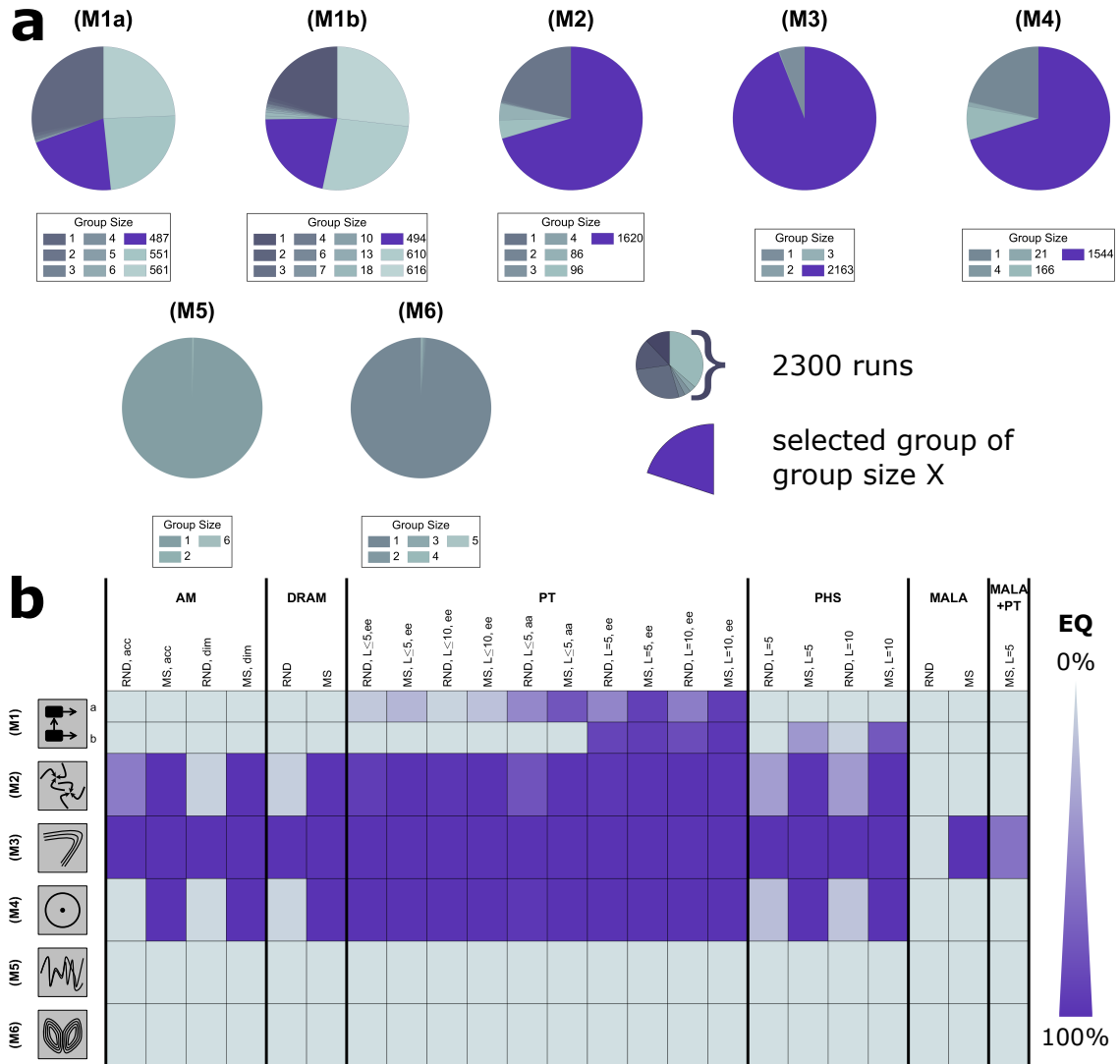


Figure 4.6: **Overview of observed exploration qualities.** (a) Distribution of group sizes regarding chain similarity. All groups with the same groups sizes are colored identically. The coloring scheme is indicated below the individual plots. (b) Exploration quality by benchmark problem (row) and algorithm (column). Each colored square is based on the fraction of 100 runs which were able to explore well.

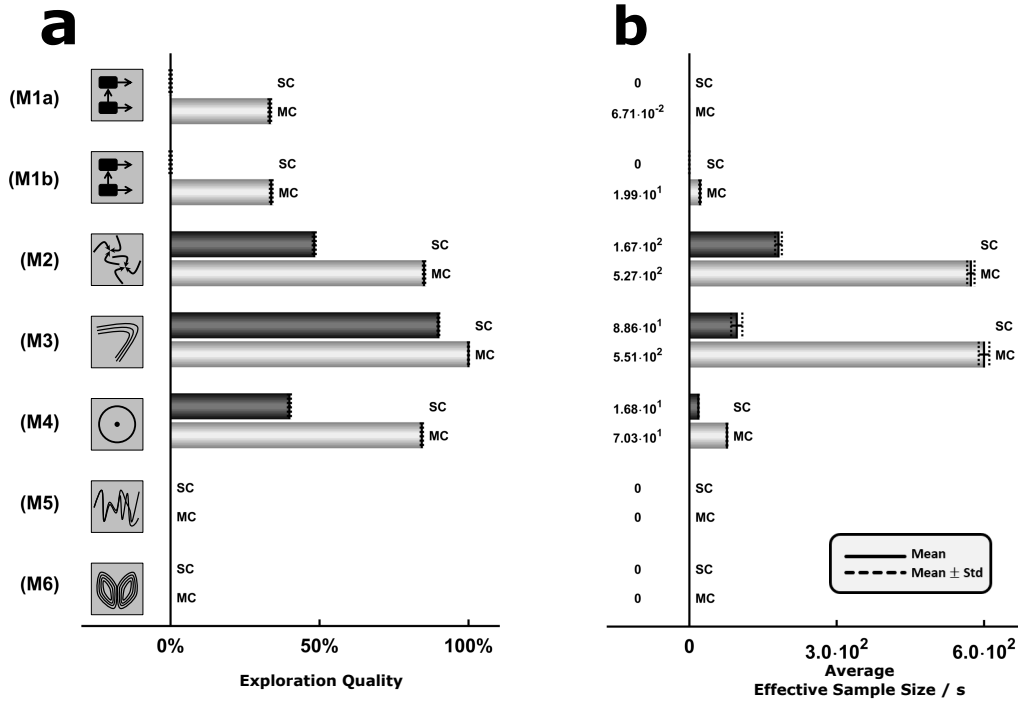


Figure 4.7: **Benchmark problem wise comparison of single- and multi-chain based sampling methods.** (a) EQ and (b) ESS per second computed by averaging across scenarios using single- or multi-chain sampling methods.

outweighed the higher computational complexity even for benchmark problems with one mode. As a result, multi-chain methods produced higher effective samples sizes and were overall computationally more efficient (Figure 4.7b).

4.5.5 Comparison of initialization strategies

In addition to characteristics of methods, the importance of initialization schemes was addressed. The average performance characteristics for RND and MS initialization were computed by averaging over sampling methods and tuning parameter choices (Figure 4.8). This revealed that multi-start local optimization substantially improved the EQ (Figure 4.8a). The difference in the sampling efficiency (conditioned ESS per CPU second) was less pronounced than for the EQ as multi-start local optimization required additional computation time (Figure 4.8b).

A detailed analysis revealed that some methods were more sensitive to the initialization than others. The performance of PT appeared to be almost independent of the initialization scheme (Figure 4.6b), making it a robust choice if no optimization is available. PHS

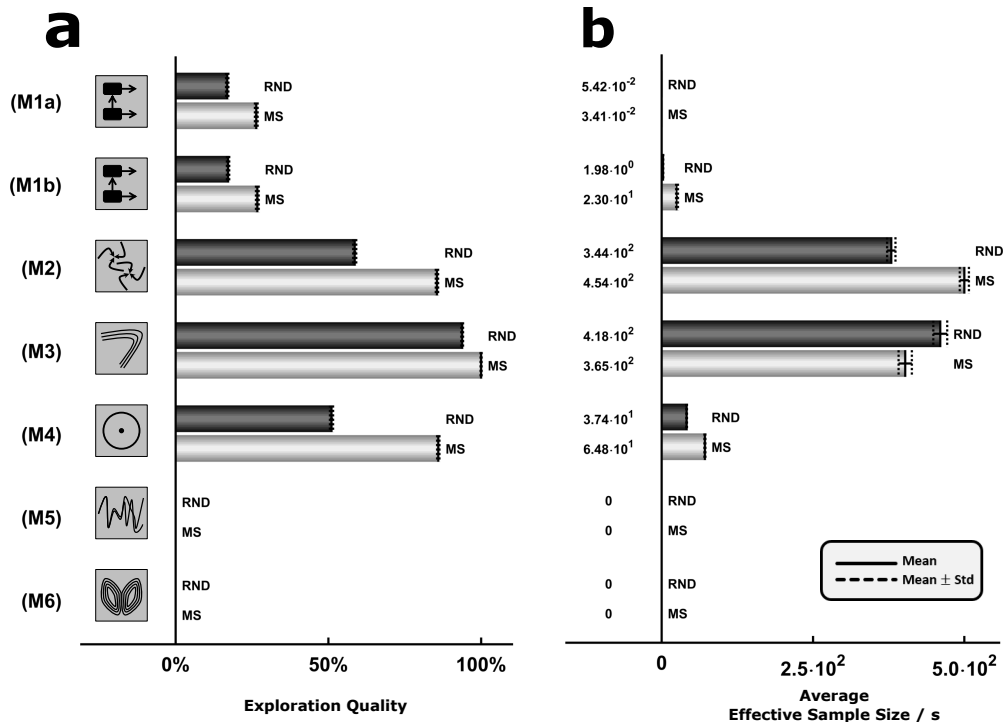


Figure 4.8: **Benchmark problem wise comparison of initialization using samples from the prior (RND) and multi-start local optimization results (MS).** (a) EQ and (b) ESS per second computed by averaging across scenarios using RND or MS initialization.

required initialization using multi-start optimization results to achieve good EQ (Figure 4.6b). Indeed, PHS initialized using samples from the prior performed poorly while PHS initialized using multi-start optimization outperformed the other methods in some cases.

4.5.6 Selection of tuning parameters and algorithmic settings

To provide guidelines regarding tuning parameters and adaptation mechanisms, a fine-grained analysis of sampling methods and subclasses of them was carried out. The assessment of single-chain samplers revealed that the adaptive Metropolis methods with acceptance rate dependent proposal scaling (AM(acc), see Section 2.4.3 for details on 'acc' and 'dim') outperformed methods with dimension-dependent proposal scaling (AM(dim) and DRAM(dim)) as shown in Figure 4.4f. Delayed rejection implemented in DRAM could not compensate for the improved proposal scaling implemented in AM(acc). Furthermore, for the benchmark problems considered here, AM(acc) outperformed MALA. While AM(acc) worked for the benchmark problems with mono-modal posterior densi-

ties, AM(dim), DRAM and MALA mostly failed to explore the posterior densities (see Figure 4.4f and 4.6b).

The PT algorithms employed in this study used temperature and proposal density adaptation. In particular, different strategies for swapping sub-chains and selecting the number of temperatures were analyzed. The best performance characteristics were achieved with a large, fixed number of temperatures (see Figure 4.4f). If too few temperatures or an adaptive reduction of the number of temperatures are used, the methods are more likely to fail to explore, e.g. by sampling only from one mode. This indicates that the available methods for the reduction of the number of temperatures (Lacki and Miasojedow, 2015) — which worked for a series of simple examples — is not sufficiently robust, yet. In contrast, the parallel tempering algorithms appeared to be robust with respect to the swapping strategy, with equi-energy (ee) swaps (see Section 2.4.3 for details) yielding superior performance.

4.6 Summary and discussion

The quantitative and qualitative properties of biological models depend on the values of their parameters. These parameters values are usually inferred using optimization or sampling methods. For optimization schemes comprehensive benchmarking results are available (Moles et al., 2003; Raue et al., 2013b; Villaverde et al., 2015; Hross and Hasebauer, 2016). In this work these results were complemented and a selection of sampling methods was benchmarked.

For this purpose, a collection of small-sized benchmark problems for ODE constrained parameter estimation with oscillating, bifurcating and chaotic solutions as well as multi-stable steady states and non-identifiabilities was studied. These model properties lead to pronounced tails, multiple modes and rims in the posterior densities. It was found, that some of these challenges can be addressed by employing additional information about the model and tools like structural identifiability analysis (see Section 4.5.1). However, in applications, it is typically not possible to avoid non-identifiabilities and tools as GenSSI (Ligon et al., 2017) do not scale well with raising parameter dimensions or reveal the exact nature of the non-identifiability.

As a by-product of the presented benchmarking study the effect of properties of the ODE model, such as Hopf bifurcation and multi-stability, onto the performance of sampling algorithms was considered. As most models of biological systems are nonlinear, high-

dimensional and possess multiple positive and negative feedback loops (Alon, 2006), a single model can usually exhibit different properties in different parameter regimes. As the biologically relevant regimes in parameter spaces are typically unknown prior to the parameter estimation, knowledge about the dynamic properties cannot be employed and the use of robust sampling methods is beneficial. We expected bifurcations to strongly impact the sampling efficiency. This, however, was not found here. Instead, it was observed that chaotic regimes have a strong influence on the sampling efficiency and might even render it intractable. This is consistent with previous findings and expected as “chaotic likelihood functions, while ultimately smooth, have such complicated small scale structure” (Du and Smith, 2017).

To derive guidelines for sampling method selection, a range of single- and multi-chain samplers were assessed. This revealed that most state-of-the-art sampling methods require a large number of iterations to provide a representative sample from multi-modal posterior densities even in low-dimensional parameter spaces. Multi-chain methods clearly outperformed single-chain methods, as reported earlier for individual examples (see, e.g., (Calderhead, 2007; Hug et al., 2013) and references therein). The presented, comprehensive evaluation confirms that the intuition was correct. Surprisingly, this was also the case for unimodal posterior densities. In addition, it was shown for the first time that reliability and performance of all sampling methods except PT was substantially improved when initialized using optimization results instead of samples from the prior. Interestingly, for the benchmarks considered in this chapter, PT performed better without novel adaptation schemes for the number of temperatures (Lacki and Miasojedow, 2015). This is in contrast to results for posterior distributions in the original publication (Lacki and Miasojedow, 2015) – for which the employed implementation achieved the same results –, suggesting that additional research is required. Furthermore, this emphasizes the importance of realistic test problems. The comparison of dimension-dependent proposal scaling (Haario et al., 2006) and acceptance-rate-dependent proposal scaling (Miasojedow et al., 2013), which was to the best of the authors knowledge not published before, revealed the superiority of the latter. From this insight a range of single- and multi-chain methods can benefit. Overall, PHS with optimization-based initialization performed best for uni-modal posterior landscapes while PT performed most robustly regarding all posterior types.

While multi-chain samplers were found to be superior, the overall low ESSs of methods even for methods with good EQ suggest, that there is potential for improvement for these methods as well. In particular, adaptation strategies *acc* or *dim* with a simple normal distributed proposal density may behave poorly if the local posterior landscape strongly differs between different regions in parameter space. This finding will be discussed in

more detail in Chapter 5 and a novel method, RAmPART, is proposed to overcome these shortcomings, further improving the performance of multi-chain samplers.

Both, the single-chain MALA and the PT-MALA almost always failed to generate representative samples. It is expected that this is due to often proposing new chain points outside the box constraints θ_{\min} and θ_{\max} as the current implementation calculates gradients based on the most recent parameter point exclusively and does not know about parameter constraints while doing so. It is expected, that this effect gets worse with larger system dimension due to the curse of dimensionality. It would be interesting to consider a MALA like algorithm with a regularization of the proposal density.

Beyond the evaluation of algorithms, the results demonstrate the importance of performing multiple independent runs of sampling methods starting from different points in parameter space (Hug et al., 2013). Most algorithms merely provide a representative sample in a fraction of the runs. In addition to standard sampling diagnostics (e.g. convergence tests like Gelman-Rubin-Brooks (Geweke, 1992)), the extended analysis pipeline takes into account the EQ while minimizing the need for subjective visual inspection. The results confirm the need to evaluate sampling methods by not only taking into account the ESS of the generated runs but the overall EQ as important measure for algorithmic robustness. Thus, only a combination of ESS and EQ should be used to decide whether a sample is representative of the given posterior.

The benchmark problems considered in this chapter are low-dimensional yet they resemble some essential features of parameter estimation problems in systems biology. The precise quantitative results depend on the selection of the benchmark problems and there is no general theory regarding the qualitative findings, yet. Thus, the presented effort is explicitly meant as a starting point for additional research including a broader range of application problems. Furthermore, while several classes of sampling methods have been considered, the study of additional methods would be beneficial as well. In particular, the assessment of Hamiltonian Monte Carlo (HMC) based algorithms such as NUTS or Wormhole Monte Carlo (Hoffman and Gelman, 2014; Lan et al., 2014), region-based methods (Craiu et al., 2009), Metropolis-in-Gibbs methods (Bédard, 2017), Transitional MCMC (Betz et al., 2016), sequential Monte Carlo methods (Yanagita and Iba, 2009) or additional proposal adaptation strategies (Vihola, 2012) would be valuable. Using a copula decomposition for the proposal densities is certainly an interesting approach in order to obtain well suited proposal densities (Schmidl et al., 2013). For ODE models for which conditional distributions of the parameters can be derived, also Gibbs samplers might be used (Casella and George, 1992). Furthermore, a comparison with non-sampling-based approximation

methods, e.g. variational methods (MacKay, 2005) or approximation methods (Fröhlich et al., 2014a) could be interesting.

In this thesis, box constraint parameter estimation problems are considered exclusively, as they are the most commonly used in computational biology. However, there exist problems with more complex parameter constraints as dependencies and non-linearities. It is known, that sampling gets significantly more challenging in those problem classes (Chen et al., 2017). It would be interesting to investigate the performance of state-of-the-art MCMC in such settings as well and to construct better performing algorithms if necessary.

In summary, the presented, comprehensive evaluation revealed that even state-of-the-art MCMC algorithms have problems to sample efficiently from many posterior distributions arising in systems biology. Problems arose in particular in the presence of non-identifiabilities and chaotic regimes. The examples provided in manuscripts presenting new algorithms are often not representative and a more thorough assessment on benchmark collections should be required (as is common practice in other fields). Chapter 4 provides a basis for future developments of such benchmark collections allowing for a rigorous assessment of novel sampling algorithms. Here, six benchmark problems with common challenges to provide practical guidelines for the selection of sampling algorithms, adaptation and initialization schemes were employed already as a start. Furthermore, the presented results highlight the need to address chain exploration quality by taking into account multiple MCMC runs which can be compared with each other before calculating effective sample sizes. The availability of the code will simplify the extension of the methods and the extension of the benchmark collection.

Chapter 5

Region-based adaptive parallel tempering

This chapter is based on (Ballnus et al., 2018).

Biochemical reaction networks (BRNs, see Section 2.3) have become a standard tool for the investigation of cellular processes and the unraveling of signal processing mechanisms. The parameters of these models are usually derived from the available data using optimization and sampling methods. However, the efficiency of these methods is limited by the properties of the mathematical model (see Chapter 4). Parameter probability densities of BRNs are known to possess, e.g., multi-modal posterior densities with long valleys or pronounced tails which make optimization and sampling challenging. Thus, the development or improvement of optimization and sampling methods in particular suited for the application in challenging model classes as BRNs is subject to ongoing research. However, when evaluating critically (see Section 4) it is revealed, that state-of-the-art sampling algorithms under-perform even for small models with mathematical properties commonly found in BRNs.

In this chapter, a Region-based Adaptive PARallel Tempering algorithm (RAmPART) which adapts to the problem-specific posterior densities, i.e. modes and valleys, is suggested. The algorithm combines several established algorithms to overcome their individual shortcomings and to improve sampling efficiency. Its properties were assessed for established benchmark problems and two ordinary differential equation models of biochemical reaction networks. The proposed algorithm outperformed state-of-the-art methods in terms of calculation efficiency and mixing. Since the algorithm does not rely on a specific problem structure, but adapts to the posterior density, it is suitable for a variety of model classes.

5.1 Introduction and problem statement

One natural way to maximize the efficiency of sampling algorithms is to reduce the auto-correlation of the generated chains (Andrieu et al., 2003) (see Chapter 3). Decreasing the auto-correlation lowers the necessary chain length required for a representative sample from the posterior distribution. The auto-correlation achieved using an MCMC algorithm can be reduced by employing a tailored proposal density. In the literature, three prominent concepts for the tailoring of the proposal density, which are independent of the underlying model, are present (see also Section 2.4.3):

- (i) Adaptive Metropolis (AM) samplers improve the global proposal density based on the already available chain (Haario et al., 2001; Roberts and Rosenthal, 2009; Andrieu and Thoms, 2008).
- (ii) Hamiltonian Monte Carlo, Riemannian Monte Carlo and related sampling approaches exploit the local geometry of the posterior, such as 1st and 2nd order derivatives, to construct an appropriate local proposal density (Girolami and Calderhead, 2011; Hoffman and Gelman, 2014; Lan et al., 2014; Graham and Storkey, 2017).
- (iii) Region-based methods split the parameter domain in different regions and assign appropriate proposal densities for each of the individual regions (Craiu et al., 2009; Bai et al., 2010). Some of these methods have also been combined, e.g., region-based adaptive Metropolis (RB-AM) samplers (Bai et al., 2010).

Complementary, small-world sampling (Yang et al., 2016; Guan and Krone, 2007) and delayed rejection adaptive metropolis (Haario et al., 2006) has been introduced. These methods employ multi-component and multi-try proposals, respectively, and can be combined with the aforementioned approaches. All of these different concepts boosted the sampling performance of MCMC methods, but the auto-correlation for posteriors with multiple modes usually remains high (Chapter 4). To address this, multi-chain algorithms such as Parallel Tempering (PT) (Lacki and Miasojedow, 2015; Sambridge, 2013; Miasojedow et al., 2013; Vousden et al., 2016) and Parallel Hierarchical Sampling (Rigat and Mira, 2012; Hug et al., 2013) have been introduced. These algorithms aim to improve mixing by increasing the frequency of jumps between modes and the exploration of tails. From a visual inspection of the individual results of Chapter 4 one can see that for multi-chain methods the local exploration may become slow and limits the mixing.

Analogue to the definition in Equation 2.1 in Chapter 2, here the problem of sampling from the posterior density $\pi(\theta|\mathcal{D})$ of the model parameter $\theta \in \Omega \subset \mathbb{R}^n$ given the data \mathcal{D} is

considered. Depending on the model and dataset, posterior densities may possess different properties, including multiple modes, pronounced tails and differences between local and global structure (Chapter 3 and 4, Section 4.3 for ODE examples) – properties often found in biological problems with non-identifiabilities (Chis et al., 2011; Fröhlich et al., 2014b; Raue et al., 2013a; Eisenberg and Hayashi, 2013).

In this chapter, a sampling algorithm which addresses differences between local- and global posterior properties to improve the mixing and to decrease the autocorrelation is proposed. The algorithm combines PT with region-based adaptation. This allows efficient transitions between modes and a good local exploration at the same time. To facilitate the application of the algorithm, all steps are automatized, including the construction of the regions and the adaptation of the proposal density. The proposed algorithm is systematically assessed and evaluated using established benchmarks as well as application problems, and compared it to state-of-the-art RB-AM and PT methods. It is shown by applying the analysis pipeline developed in Chapter 3, that the presented method has the potential to improve sampling efficiency and robustness for posteriors densities with multiple modes or pronounced tails substantially. As these properties are usually unknown prior to the sampling, here the problem of developing a robust multi-purpose method is considered.

5.2 Region-based adaptive parallel tempering

Here, a sampling algorithm which combines region-based adaptive Metropolis and parallel tempering is proposed. In this section, Region-based Adaptive PARallel Tempering (RAmPART) is motivated, proposed, its implementation is presented and its convergence properties are discussed.

5.2.1 Motivation

The proposed method is based on two existing algorithms, which are briefly summarized in the following. Additional details about MCMC algorithms can be found in Chapter 2.4 as well.

Region-based adaptive Metropolis (RB-AM) algorithms construct a Markov chain for the parameters $\theta \in \Omega$ with target density $\pi(\mathcal{D}|\theta)$. To account for the potentially complex geometry of the posterior density, the parameter domain Ω is split into regions Ω_r , $r = 1, \dots, R$ (Craiu et al., 2009). Each region Ω_r possesses an individual proposal density $q_r^{[i]}(\theta'|\theta^{[i]})$, in which i is the index of the iteration. The union of all regions corresponds

to Ω . The regions and region-specific proposal densities are constructed adaptively for instance using Gaussian Mixture Models (GMMs). In comparison to classic AM algorithms which use a single, global proposal density, RB-AM adds a high degree of freedom to the way new sample points are proposed across different parts of the posterior.

Parallel tempering (PT) algorithms construct a Markov chain on a product space $\Omega^L := \{\theta = (\theta_1, \dots, \theta_L) | \theta_\ell \in \Omega, \ell = 1, \dots, L\}$ where L is the number of temperature levels (Lacki and Miasojedow, 2015; Vousden et al., 2016). The target density on the product space is defined as the product of tempered posterior densities,

$$\pi(\theta | \mathcal{D}) \propto \prod_{\ell=1}^L \pi^{1/\tau_\ell}(\theta_\ell | \mathcal{D}), \quad (5.1)$$

with temperatures τ_ℓ , $\ell = 1, \dots, L$, such that $1 = \tau_1 < \tau_2 < \dots < \tau_L$. The sequence of points $\theta_\ell^{[i]}$ for a temperature ℓ , is referred to as the ℓ th chain. The Markov chain on the product space performs random walk steps for each chain and random swaps between chains. As chains with high temperatures travel more easily between different areas in the parameter space, e.g., different modes and tails, chains generated using PT often possess a lower auto-correlation than single-chain algorithms (see Chapter 4).

RB-AM and PT are applicable to a wide range of problems, yet, their sampling performance is often unsatisfactory. In Section 5.3 it is illustrated, that RB-AM has difficulties to travel between different parameter regions with high probability mass and PT suffers from differences between local and global correlation structure. To address these shortcomings, the RAMPART algorithm is proposed. RAMPART constructs a Markov chain on a product space Ω^L by interweaving random walk steps and random swaps, as in PT. For each of the tempered sub-chains, a region-based proposal density, constructed and adapted over time as in RB-AM algorithms, is employed while the regions are chosen to be identical for all temperatures. In the following, the mathematical details are discussed.

5.2.2 Method

The proposed method has two phases: warm-up and sampling phase. In the *warm-up phase*, parameter regions Ω_r , $r = 1, \dots, R$, are constructed which are suited for efficient sampling with Gaussian proposal densities. In principle, different approaches could be employed to determine such regions, including the use of information obtained in a preceding multi-start local optimization. Here, it was chosen to employ a PT algorithm (Lacki and Miasojedow, 2015; Vousden et al., 2016; Ballnus et al., 2017), which uses an adaptive Gaussian proposal for the random walk steps (Figure 5.1a), an adjacent proposal for ran-

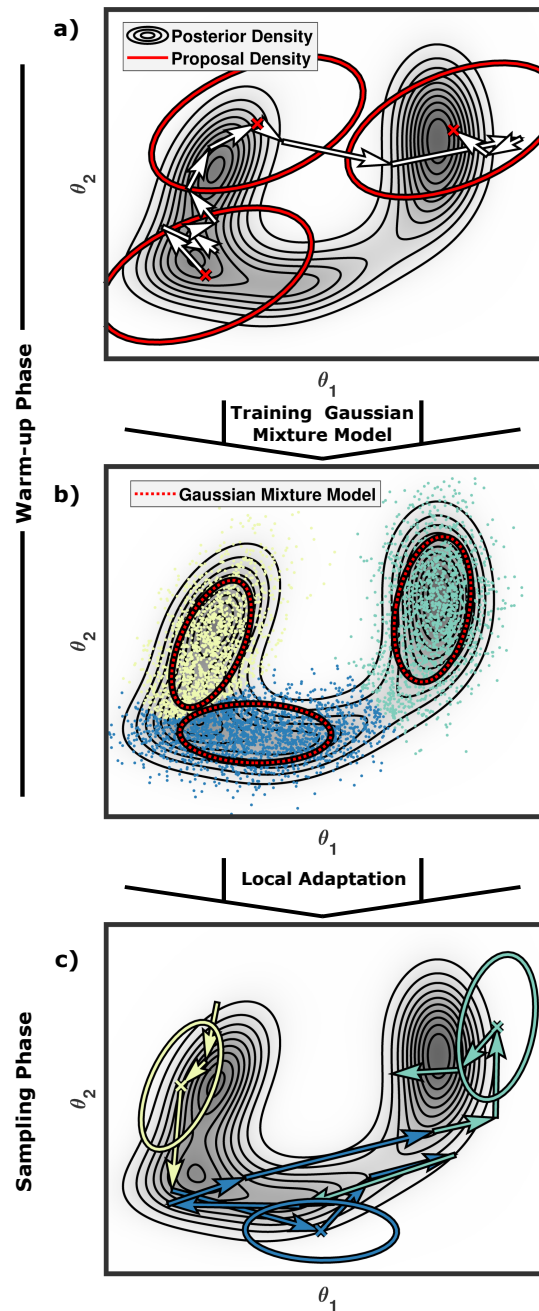


Figure 5.1: **Visualization of the 2-phase sampling process employed by RAM-PART.** a) In the warm-up phase, the posterior distribution is sampled with a parallel tempering algorithm which adapts to the global covariance structure of the corresponding posterior density. b) The posterior samples (which might not be representative) are used to define regions in the posterior which can be approximated using a Gaussian mixture model. The Gaussian mixture model defines three regions and each of the sample points is associated with one them represented by the yellow, green or blue color. c) In the sampling phase, the regions are used to adapt region-specific proposal densities. The adaptation of the covariance matrices in the warm-up and sampling phase is performed for each temperature separately.

dom swaps and temperature adaptation following (Vousden et al., 2016). This choice is motivated by the versatility of PT found as a result in Chapter 4. A short run of this PT algorithm yields a sample which is supposed to capture the high-probability regions of the posterior. This sample is usually not representative for the posterior distribution, but is sufficient for the warm-up phase. Using the sample, a Gaussian Mixture Model (GMM) is trained with an expectation-maximization (EM) algorithm (Murphy, 2012) (Figure 5.1b), yielding

$$\text{GMM}(\theta) = \sum_{r=1}^R w_r \mathcal{N}(\theta | m_r, C_r), \quad (5.2)$$

with weights $w_r > 0$, $\sum_{r=1}^R w_r = 1$, means m_r and covariance matrices C_r , $r = 1, \dots, R$. A reasonable number of mixture components R is determined using 5-fold cross-validation (Kohavi et al., 1995) with the BIC (Claeskens et al., 2008) as selection criterion. As the GMM approximates the posterior density, it is assumed that in the parameter region

$$\Omega_r = \{\theta \in \Omega | \forall r' \in \{1, \dots, R\} \setminus r, w_r \mathcal{N}(\theta | m_r, C_r) \geq w_{r'} \mathcal{N}(\theta | m_{r'}, C_{r'})\}, \quad (5.3)$$

in which the r th mixture component dominates, a Gaussian proposal density (see below for definition) with covariance matrix C_r can achieve a good sampling performance. Accordingly, in the following, the parameter regions Ω_r , $r = 1, \dots, R$, are used in random walk steps (for all temperatures).

Remark: Information about the posterior density, e.g. number of modes, known non-identifiabilities, most relevant parameter dimensions or correlation structures, can be exploited to provide user-defined regions Ω_r or to restrict the GMMs to a subset of the parameters. This does potentially improve robustness and computational efficiency, or allows to skip the warm-up phase entirely.

In the *sampling phase* the afore-derived parameter regions Ω_r , $r = 1, \dots, R$, are used to construct random walk proposals which are tailored to the shape of the tempered posterior densities. The random walks for the different temperatures are performed independently and use a Gaussian mixture as proposal density, also known as small world proposal density (Figure 5.1c). For a parameter vector $\theta_\ell^{[i]}$ in the r th region, $\theta_\ell^{[i]} \in \Omega_r$, the region-based proposal density is

$$q_{\ell,r}^{[i]}(\theta'_\ell | \theta_\ell^{[i]}) = (1 - p_g) \mathcal{N}(\theta'_\ell | \theta_\ell^{[i]}, e^{2\eta_{\ell,r}^{[i]}} C_{\ell,r}^{[i]}) + p_g \mathcal{N}(\theta'_\ell | \theta_\ell^{[i]}, e^{2\eta_\ell^{[i]}} C_\ell^{[i]}). \quad (5.4)$$

The two mixture components capture the correlation structures of the tempered posterior $\pi^{1/\tau_\ell}(\theta_\ell | \mathcal{D})$ in Ω_r and Ω , respectively. $C_{\ell,r}^{[i]}$ and $C_\ell^{[i]}$ denote estimates of regional- and

global covariance matrices while $\eta_{\ell,r}^{[i]}$ and $\eta_{\ell}^{[i]}$ denote the corresponding scaling factors. The fraction of steps using the global proposal density is denoted by p_g . The proposed points $\theta' \sim q_{\ell,r}^{[i]}(\theta'_\ell | \theta_\ell^{[i]})$ are accepted with probability

$$p_{\text{acc}}(\theta_\ell, \theta'_\ell, 1/\tau_\ell) = \min \left\{ 1, \left(\frac{\pi(\theta'_\ell | \mathcal{D})}{\pi(\theta_\ell^{[i]} | \mathcal{D})} \right)^{1/\tau_\ell} \frac{q_{\ell,r'}^{[i]}(\theta_\ell^{[i]} | \theta'_\ell)}{q_{\ell,r}^{[i]}(\theta'_\ell | \theta_\ell^{[i]})} \right\}, \quad (5.5)$$

for region index r' such that $\theta' \in \Omega_{r'}$. If accepted, $\theta_\ell^{[i+1]} = \theta'_\ell$, otherwise $\theta_\ell^{[i+1]} = \theta_\ell^{[i]}$. Estimates for regional and global covariance are adapted during the sampling phase. For $\theta_\ell^{[i]} \in \Omega_r$ updates of local mean and covariance are given by

$$m_{\ell,r}^{[i]} = (1 - \gamma^{[i]})m_{\ell,r}^{[i-1]} + \gamma^{[i]}\theta_{\ell,r}^{[i]}, \quad (5.6)$$

$$C_{\ell,r}^{[i]} = (1 - \gamma^{[i]})C_{\ell,r}^{[i-1]} + \gamma^{[i]}(\theta_\ell^{[i]} - m_{\ell,r}^{[i]})(\theta_\ell^{[i]} - m_{\ell,r}^{[i]})^T, \quad (5.7)$$

with adaptation strength $\gamma^{[i]} = i^{-\alpha}$, with $\alpha \in (0.5, 1)$, and initial value obtained computed in the warm-up phase, $C_{\ell,r}^{[0]} = C_r$ (Lacki and Miasojedow, 2015). The scaling factors for regional and global covariance are adapted to achieve an acceptance rate of 23.4% (Lacki and Miasojedow, 2015; Haario et al., 2001),

$$\eta_{\ell,r}^{[i]} = \eta_{\ell,r}^{[i-1]} \exp \left(\gamma^{[i]}(p - 0.234) \right), \quad (5.8)$$

with $\eta_{\ell,r}^{[0]} = 1$. These adaptation strategies are similar to the *acc* strategy employed in Chapter 4, which was found to perform well in ODE based parameter estimation problems.

The region-based random walk proposal is incorporated in a PT algorithm with an adjacent swap proposal and temperature adaptation. For the temperature adaptation the method by (Vousden et al., 2016) is employed and a maximum temperature τ_{max} is selected. Please note, that this swapping and temperature adaptation strategy is slightly different from the ones used in Chapter 4. This is motivated by empirical studies, while in principle other strategies could be used as well. The temperatures of the sub-chains, $1 = \tau_1 < \tau_2 < \dots < \tau_L = \tau_{\text{max}}$, are adapted such that the swap acceptance for adjacent sub-chains gets balanced:

$$\tau_\ell^{[i]} = \sum_{m=2}^{\ell} \left(\tau_m^{[i-1]} - \tau_{m-1}^{[i-1]} \right) \exp \left\{ -\kappa^{[i]} \left(A_m^{[i]} - A_{m-1}^{[i]} \right) \right\}, \quad (5.9)$$

with $\kappa^{[i]} = \frac{\nu_\tau}{\eta_\tau(i+1 + \nu_\tau)}$

for $\ell = 1, \dots, L$ (Vousden et al., 2016). The adaptation parameters are set $\nu_\tau = 10^3$ and $\eta_\tau = 10$ according to personal experience. $A_\ell^{[i]}$ indicates whether the proposed swap

between chain ℓ and $(\ell + 1)$ has been accepted. Please note that the temperatures $\tau_1 = 1$ and $\tau_L = \tau_{\max}$ are constant. The optimal choice for the maximum temperature τ_{\max} is problem specific yet known to be crucial for the success for any PT like algorithm. In Section 5.4, a automated selection for this important tuning parameter is presented and evaluated.

The resulting algorithm, RAmPART, is flexible as it possesses the RB-AM and the adaptive PT as special cases. A RB-AM is obtained for $L = 1$ and an adaptive PT for $R = 1$, respectively.

5.2.3 Implementation

In this section, the corresponding pseudocode (Figures 5.2 and 5.3), default values (Table 5.1) and tuning aspects for RAmPART are discussed. Warm-up and sampling phase are presented individually. In the warm-up phase, an adaptive parallel tempering algorithm is used to gather a sample to estimate the parameters of a Gaussian Mixture Model (GMM) (Figure 5.2). In the sampling phase, the Gaussian mixture model is used to initialize a region-based parallel tempering algorithm which adapts the regional proposal covariances (Figure 5.3). In the pseudocode, the number sample iterations N , the chain initialization point $\theta^{[0]}$, the initial proposal covariance matrix C , the initial inverse temperature latter β (with $\beta_1 = 1$ and $\beta_L = 1/\tau_{\max}$), the number of parallel tempered chains L , the maximum number of regions R_{\max} , the number of cross validations when training the GMM N_{rep} , the adaptation parameters α , ν_τ and η_τ and the fraction of global proposals compared to local proposals p_g are denoted. Note, a heuristic for initializing β is used, which is based on the discussion in (Vyshemirsky and Girolami, 2008) and personal experience. An overview of all input parameters is shown in Table 5.1. The algorithm returns the posterior parameter chain $\theta_1^{[0]}, \dots, \theta_1^{[N-1]}$ as well as the corresponding posterior values $p^{[0]}, \dots, p_1^{[N-1]}$. An overview of all output parameters is provided in Table 5.2.

As part of RAmPART’s implementation, parameters of a GMM are estimated using an EM algorithm (Murphy, 2012). The convergence and robustness of the EM is directly influenced by the properties of the posterior. As the EM algorithm is a local (sometimes called “greedy”) optimization routine, a single run often only finds a local optimum (e.g. (Zhang et al., 2008)). To account for this problem, the EM algorithm is initialized at multiple parameter points (5 different points for the examples discussed in the Section 5.3) and the best result regarding BIC is taken for running the sampling phase.

In practice, an important aspect for RAmPART is the generation of a sufficiently representative sample during the warm-up phase, which is supposed to capture the rough

Algorithm 1: Warm-up Phase

```

input :  $N, \theta^{[0]}, \mathbf{C}, \beta, L, R_{max}, N_{rep}, \alpha, \nu_\tau, \eta_\tau$ 
output:  $R, \theta^{[N-1]}, \mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \beta, \mathbf{m}, \mathbf{C}$ 

// Initialize
for  $\ell \leftarrow 1$  to  $L$  do
   $\eta_\ell \leftarrow 1, \mathbf{m}_\ell \leftarrow \theta^{[0]}$ 
   $p_\ell = \pi(\theta^{[0]} | \mathcal{D})$ 
for  $i \leftarrow 0$  to  $N - 1$  do
  // Random walk step
  for  $\ell \leftarrow 1$  to  $L$  do
    Propose parameter  $\theta' \sim \mathcal{N}(\theta_\ell^{[i]} | \mathbf{m}_\ell, \eta_\ell^2 \mathbf{C}_\ell)$ 
    Evaluate posterior  $p' \leftarrow \pi(\theta' | \mathcal{D})$ 
    Acceptance chance  $p \leftarrow (p'/p_\ell)^{\beta_\ell}$ 
    Draw uniformly  $v \sim U([0, 1])$ 
    if  $v < p$  then
       $\theta_\ell^{[i+1]} \leftarrow \theta'$ 
       $p_\ell = p'$ 

  // Random walk proposal density adaptation
  for  $\ell \leftarrow 1$  to  $L$  do
     $\gamma^{[i]} = 1/(1+i)^\alpha$ 
     $\mathbf{m}_\ell \leftarrow (1 - \gamma^{[i]})\mathbf{m}_\ell + \gamma^{[i]}\theta_\ell^{[i]}$ 
     $\mathbf{C}_\ell \leftarrow (1 - \gamma^{[i]})\mathbf{C}_\ell + \gamma^{[i]}(\theta_\ell^{[i]} - \mathbf{m}_\ell)(\theta_\ell^{[i]} - \mathbf{m}_\ell)^T$ 
     $\eta_\ell \leftarrow \eta_\ell \exp(\gamma^{[i]}(p - 0.234))$ 

  // Chain swapping
  for  $\ell \leftarrow L$  to  $2$  do
     $\Delta\beta_{\ell-1} \leftarrow \beta_{\ell-1} - \beta_\ell$ 
    Swap probability  $p_{swap, \ell-1} \leftarrow (p_\ell/p_{\ell-1})^{\Delta\beta_{\ell-1}}$ 
    Draw uniformly  $w \sim U([0, 1])$ 
     $A_{\ell-1} \leftarrow (w < p_{swap})$ 
    if  $A_{\ell-1}$  then
       $\theta_\ell \leftrightarrow \theta_{\ell-1}$ 
       $p_\ell \leftrightarrow p_{\ell-1}$ 

  // Temperature adaptation
   $\kappa^{[i]} \leftarrow \nu_\tau / (\eta_\tau(i + 1 + \nu_\tau))$ 
  for  $\ell \leftarrow 1$  to  $L - 2$  do
     $\Delta S_\ell \leftarrow \kappa^{[i]}(A_\ell - A_{\ell+1})$ 
     $\Delta\tau_\ell \leftarrow 1/\beta_{\ell+1} - 1/\beta_\ell$ 
  for  $\ell \leftarrow 2$  to  $L - 1$  do
     $\beta_\ell \leftarrow 1/\sum_{m=2}^\ell \Delta\tau_{m-1} \exp(\Delta S_{m-1})$ 

// Train GMM
for  $k \leftarrow 1$  to  $N_{replicates}$  do
  for  $n \leftarrow 1$  to  $R_{max}$  do
    Fit GMM  $\mathbf{w}_{n,k}, \boldsymbol{\mu}_{n,k}, \boldsymbol{\Sigma}_{n,k} \leftarrow \text{EM-Algorithm}(\theta_1^{[0]}, \dots, \theta_1^{[N]}; n, \text{random seed } k)$ 
     $BIC_{n,k} \leftarrow -2 \log \left( \sum_{m=0}^N \text{GMM}(\theta_\ell^{[m]} | \mathbf{w}_{n,k}, \boldsymbol{\mu}_{n,k}, \boldsymbol{\Sigma}_{n,k}) \right) +$ 
       $n \log(N) ((n-1)/n + 2 \dim_\theta + (\dim_\theta - 1) \dim_\theta / 2)$ 

  Select best GMM  $(R, K) \leftarrow \underset{n,k}{\text{argmin}}(BIC_{n,k})$ 

  Define regions  $\mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\Sigma} \leftarrow \mathbf{w}_{R,K}, \boldsymbol{\mu}_{R,K}, \boldsymbol{\Sigma}_{R,K}$ 

```

Figure 5.2: Pseudocode for warm-up phase of RAMPART.

Algorithm 2: Sampling Phase

```

input :  $N, N_{warmUp}, \theta^{[0]}, \mathbf{C}, \beta, L, R_{max}, N_{rep}, \alpha, \nu_\tau, \eta_\tau, p_g$ 
output: non-tempered parameter chain  $\theta_1^{[0]}, \dots, \theta_1^{[N-1]}$ ,
          non-tempered posterior values  $p^{[0]}, \dots, p_1^{[N-1]}$ 

// Run warm-up phase
 $(R, \theta^{[0]}, \mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \beta, \mathbf{m}, \mathbf{C}) \leftarrow \text{WarmUpPhase}(N_{warmUp}, \theta^{[0]}, \mathbf{C}^{[0]}, \beta^{[0]}, L, R_{max}, N_{rep}, \alpha, \nu_\tau, \eta_\tau)$ 
// Initialize
for  $\ell \leftarrow 1$  to  $L$  do
  New region label  $r_{prop, \ell} \leftarrow \underset{r}{\text{argmax}}(\mathcal{N}(\theta_\ell^{[0]} | \boldsymbol{\mu}_r, w_r \boldsymbol{\Sigma}_r))$ 
   $p_\ell = \pi(\theta^{[0]} | \mathcal{D})$ 
  for  $r \leftarrow 1$  to  $R$  do
    Local proposal parameters  $\eta_{\ell, r} \leftarrow 1, \mathbf{m}_{\ell, r} \leftarrow \mathbf{m}, \mathbf{C}_{\ell, r} \leftarrow \mathbf{C}$ 
    Adaptation times  $j_{\ell, r} \leftarrow 0$ 
for  $i \leftarrow 0$  to  $N - 1$  do
  // Random Walk Step
  for  $\ell \leftarrow 1$  to  $L$  do
    Set old region label  $r_\ell \leftarrow r_{prop, \ell}$ 
    Adaptation fading  $j_{\ell, r_\ell} \leftarrow +$ 
    Draw uniformly  $u \sim U([0, 1])$ 
    if  $u < 0.5$  then
       $\theta' \sim \mathcal{N}(\theta_\ell^{[i]} | \mathbf{m}_{\ell, r_\ell}, \eta_{\ell, r_\ell}^2 \mathbf{C}_{\ell, r_\ell})$ 
    else
       $\theta' \sim \mathcal{N}(\theta_\ell^{[i]} | \mathbf{m}_\ell, \eta_\ell^2 \mathbf{C}_\ell)$ 
    Get new region label  $r_{prop, \ell} \leftarrow \underset{r}{\text{argmax}}(\mathcal{N}(\theta' | \boldsymbol{\mu}_r, w_r \boldsymbol{\Sigma}_r))$ 
    Evaluate posterior  $p' \leftarrow \pi(\theta' | \mathcal{D})$ 
    Forward probability  $T_{for} \leftarrow (1 - p_g) \mathcal{N}(\theta' | \theta_\ell, \mathbf{C}_{\ell, r_\ell}) + p_g \mathcal{N}(\theta' | \theta_\ell, \mathbf{C}_\ell)$ 
    Backward probability  $T_{back} \leftarrow (1 - p_g) \mathcal{N}(\theta_\ell | \theta', \mathbf{C}_{\ell, r_{prop, \ell}}) + p_g \mathcal{N}(\theta_\ell | \theta', \mathbf{C}_\ell)$ 
    Acceptance chance  $p \leftarrow (p' / p_\ell)^{\beta_\ell} (T_{back} / T_{for})$ 
    Draw uniformly  $v \sim U([0, 1])$ 
    if  $v < p$  then
       $\theta_\ell^{[i+1]} \leftarrow \theta'$ 
       $p_\ell = p'$ 
  // Random walk proposal density adaptation
  for  $\ell \leftarrow 1$  to  $L$  do
     $\gamma = 1 / j_{\ell, r_\ell}^\alpha$ 
     $\mathbf{m}_{\ell, r_\ell} \leftarrow (1 - \gamma) \mathbf{m}_{\ell, r_\ell} + \gamma \theta_\ell^{[i]}$ 
     $\mathbf{C}_{\ell, r_\ell} \leftarrow (1 - \gamma) \mathbf{C}_{\ell, r_\ell} + \gamma (\theta_\ell^{[i]} - \mathbf{m}_{\ell, r_\ell})(\theta_\ell^{[i]} - \mathbf{m}_{\ell, r_\ell})^T$ 
     $\eta_{\ell, r_\ell} \leftarrow \eta_{\ell, r_\ell} \exp(\gamma(p - 0.234))$ 
  // Chain swapping
  for  $\ell \leftarrow L$  to  $2$  do
     $\Delta \beta_{\ell-1} \leftarrow \beta_{\ell-1} - \beta_\ell$ 
    Swap probability  $p_{swap, \ell-1} \leftarrow (p_\ell / p_{\ell-1})^{\Delta \beta_{\ell-1}}$ 
    Draw uniformly  $w \sim U([0, 1])$ 
     $A_{\ell-1} \leftarrow (w < p_{swap})$ 
    if  $A_{\ell-1}$  then
       $\theta_\ell \leftrightarrow \theta_{\ell-1}$ 
       $p_\ell \leftrightarrow p_{\ell-1}$ 
       $r_{prop, \ell} \leftrightarrow r_{prop, \ell-1}$ 
  // Temperature adaptation
   $\kappa^{[i]} \leftarrow \nu_\tau / (\eta_\tau (i + 1 + \nu_\tau))$ 
  for  $\ell \leftarrow 1$  to  $L - 2$  do
     $\Delta S_\ell \leftarrow \kappa^{[i]} (A_\ell - A_{\ell+1})$ 
     $\Delta \tau_\ell \leftarrow 1 / \beta_{\ell+1} - 1 / \beta_\ell$ 
  for  $\ell \leftarrow 2$  to  $L - 1$  do
     $\beta_\ell \leftarrow 1 / \sum_{m=2}^\ell \Delta \tau_{m-1} \exp(\Delta S_{m-1})$ 

```

Figure 5.3: Pseudocode for sampling phase of RAmPART.

Table 5.1: Input parameters used in RAmPART and their default values.

Description	Symbol	Default Value
Number of sampling iterations	N	10^6
Number of warm-up iterations	N_{warmUp}	10^5
Initial chain positions	$\theta_\ell^{[0]}$	Initial chain positions
Initial covariance matrix	\mathbf{C}	$10^6 \cdot \mathbf{I}$, \mathbf{I} Identity
Number of tempered chains	L	20
Maximum temperature	τ_{\max}	2000
Initial inverse temperatures	β	$\left(\frac{L-1-i}{L-1} + \frac{i}{L-1} \cdot \tau_{\max}^{(1/1000)}\right)^{-1000}$ for $i = 0, \dots, L-1$
Maximum number of regions allowed	R_{\max}	10
Number of EM runs on the training sample	N_{rep}	5
Covariance adaptation velocity factor	α	0.51
Temperature adaptation velocity factor 1	ν_τ	10^3
Temperature adaptation velocity factor 2	η_τ	10
Global proposal density contribution factor	p_g	0.5

structure of the posterior, including multiple modes or pronounced tails. In the worst case, the sample generated during the warm-up phase is not representative. In this case, the GMM does not capture the structure of the underlying posterior and the regional adaptation during the sampling phase may not be beneficial. Fortunately, in this case RAmPART would simply perform as good/bad as a standard PT with minimal computational overhead. This overhead is negligible for problems in which the evaluation of the objective function requires a substantial amount of computation time.

RAmPART is implemented in the MATLAB toolbox PESTO (Stapor et al., 2017), which

Table 5.2: Output parameters used in RAmPART.

Phase	Description	Symbol
Warm-up	Selected number of GMM modes	R
Warm-up	Last chain position in warm-up phase	$\theta^{[N-1]}$
Warm-up	Selected GMM mode weights	\mathbf{w}
Warm-up	Selected GMM mode centers	μ
Warm-up	Selected GMM mode covariances	Σ
Warm-up	Training sample mean	\mathbf{m}
Warm-up	Training sample covariance	\mathbf{C}
Sampling	Non-tempered parameter chain	$\theta_1^{[0]}, \dots, \theta_1^{[N-1]}$
Sampling	Non-tempered posterior values	$p_1^{[0]}, \dots, p_1^{[N-1]}$

is available on GitHub <https://github.com/ICB-DCM/PESTO/>. For the numerical simulation the SUNDIALS toolbox CVODES (Serban and Hindmarsh, 2005) via the MATLAB interface AMICI (Fröhlich et al., 2017b) has been employed.

5.2.4 Ergodicity

Beside the performance of RAMPART in practice, which will be considered in the following sections, the convergence of RAMPART has to be proven. As the warm-up phase is finite and the samples are disregarded, only the sampling phase has to be studied.

The argument is initialized with a couple of definitions. Let (Ω, \mathcal{A}) be a Borel-space with σ -Algebra \mathcal{A} defined by the power set $\mathcal{A} = \mathcal{P}(\Omega)$ and a compact set $\Omega \subseteq \mathbb{R}^k$, so that the posterior density $\pi(x|\mathcal{D})$ is continuous for $x \in \Omega$, positive on the interior of Ω and zero outside of Ω . Please note, the compactness directly follows from the parameter box constraints in the considered problems (e.g. as used in Chapter 4), while positivity and continuity of $\pi(x|\mathcal{D})$ directly follow from the fact, that the posterior is based on a continuous ODE with positive squared residuals towards the data due to measurement noise. These definitions can be used to construct a product space $\Omega^L = \underbrace{\Omega \times \dots \times \Omega}_{L \text{ times}}$ with the corresponding σ -Algebras $\mathcal{A}^L = \underbrace{\mathcal{A} \times \dots \times \mathcal{A}}_{L \text{ times}}$ defining the space in which the RAMPART sub-chains move. Furthermore, $\{P_{(\gamma, \beta)}(x, A) : x \in \Omega^L, A \subseteq \mathcal{A}^L, \gamma \in \Gamma \subseteq \mathbb{N}\}$ denotes the set of one-step Markov kernels used in RAMPART, where Γ is an index set matching the corresponding adaptation steps from Equation 5.6–5.7 and β be defined by Equation 5.9 with $\tau = 1/\beta$. The Markov kernel is defined as

$$P_{(\gamma, \beta)}(x, A) = \prod_{\ell=1}^L P_{(\gamma, \beta^{(\ell)})}(x^{(\ell)}, A_\ell) \quad (5.10)$$

with the Markov kernel for the sub-chains

$$P_{(\gamma, \beta^{(\ell)})}(x^{(\ell)}, A_\ell) = \int_{A_\ell} p_{acc}(x^{(\ell)}, y, \beta^{(\ell)}) q_\ell^{(\gamma)}(y|x^{(\ell)}) \mu^{Leb}(dy) \quad (5.11)$$

which depends on the acceptance ratio p_{acc} and the proposal density

$$q_\ell^{(\gamma)}(y|x^{(\ell)}) = \sum_{r=1}^R 1_{\Omega_r}(x^{(\ell)}) q_{\ell, r}^{(\gamma)}(y|x^{(\ell)}) \quad (5.12)$$

which is constructed by employing a mixture of regional and global proposal densities

$$q_{\ell,r}^{(\gamma)}(y|x^{(\ell)}) = (1 - p_g)\mathcal{N}(y|x^{(\ell)}, \Sigma_r^{(\gamma)}) + p_g\mathcal{N}(y|x^{(\ell)}, \Sigma^{(\gamma)}) \quad (5.13)$$

and a carrier

$$1_{\Omega_r}(x^{(\ell)}) = \begin{cases} 1, & x^{(\ell)} \in \Omega_r, \\ 0, & \textit{otherwise}, \end{cases} \quad (5.14)$$

where $P_{(\gamma,\beta^{(\ell)})}$ is a Markov kernel for one of the sub-chains targeting the ℓ th tempered posterior density $\pi^{\beta^{(\ell)}}(\cdot|\mathcal{D})$ and Ω_r , p_{acc} and $q_{\ell,r}^{(\gamma)}$ are being defined in Equations 5.3–5.5 respectively while μ^{Leb} denotes a Lebesgue measure. μ^{Leb} does exist, because $\pi(\cdot|\mathcal{D})$ and thus the integrand is positive and continuous. Please note, the Markov kernel $P_{(\gamma,\beta^{(\ell)})}$ has no index of a specific region r , as 5.12–5.14 include the possibility of all region kernels.

The theorems in (Miasojedow et al., 2013) prove ergodicity for a PT with an adaptive normal distributed proposal density in each of the sub-chains. As suggested in Section 2 of (Miasojedow et al., 2013), this proof can be extended towards more sophisticated proposal densities. Consequently, the kernel definitions in the Equations (5-7) of (Miasojedow et al., 2013) are replaced with Equation (5.10–5.14). To ensure that a PT algorithm is still ergodic after altering the sub-chain proposal densities – as is the case for RAmPART –, the assumptions made in (Miasojedow et al., 2013) have to be checked. The first assumption, (A1) in (Miasojedow et al., 2013), is based on properties of $\pi(x|\mathcal{D})$ and thus independent of changes in the algorithm. The second and third assumptions (A2-3) refer to the adaptation velocities κ . However, these were chosen identically for RAmPART as defined in Equation 5.9. In addition to (A1-3), (Miasojedow et al., 2013) explicitly uses an adaptive normal distribution for the sub-chains in its theorems. Fortunately, the proofs in (Miasojedow et al., 2013) are not tied to the usage of a simple normal distribution and may be generalized. Following the statement in (Miasojedow et al., 2013, 5.1), it is only required that individual one-step kernels $P_{(\gamma,\beta^{(\ell)})}(\cdot, A)$ satisfy ergodicity towards $\pi^{(\beta^{(\ell)})}(A|\mathcal{D})$ for all ℓ while the proof components regarding chain swapping hold independently. Thus, in the following, ergodicity of individual kernels $P_{(\gamma,\beta^{(\ell)})}(\cdot, A)$ is proven.

To prove ergodicity of individual kernels, the proof and notation in (Craiu et al., 2009) is adapted. Please note, without loss of generality uniform ergodicity has to be shown for one of the sub-chains with $\ell \in \{1, \dots, L\}$ being arbitrary but fixed. To improve readability, in the following ℓ will be left out of the notation. It is sufficient (Roberts and Rosenthal,

2007) to prove *diminishing adaptation*,

$$\lim_{n \rightarrow \infty} \sup_{x \in \Omega} \sup_{A \subseteq \mathcal{A}} |P_{\gamma_{n+1}}(x, A) - P_{\gamma_n}(x, A)| = 0, \quad (5.15)$$

using the sub-chain kernel defined in Equation 5.11 for fixed β meaning that the adaptation is vanishing over time, and *simultaneous uniform ergodicity* for $\rho < 1$,

$$|P_\gamma^n(x, A) - \pi(A)| \leq \rho^n, \quad n \in \mathbb{N}, \gamma \in \Gamma, x \in \Omega, A \in \mathcal{A}. \quad (5.16)$$

meaning that ergodicity holds for an arbitrary but fixed state of adaptation γ . Please note, for Equation 5.15, the total variance distance as defined in Section 2.4.1, Equation 2.14, was employed. In the following, Equation 5.15 and 5.16 are proven analogue to (Craiu et al., 2009).

To prove that Equation 5.16 holds, one defines

$$\varepsilon = \min(1 - p_g, p_g) \min_{r \in \{1, \dots, R\}} \left(\inf_{x, y \in \Omega} (\mathcal{N}(y|x, \Sigma_r^{(\gamma)})), \inf_{x, y \in \Omega} (\mathcal{N}(y|x, \Sigma^{(\gamma)})) \right). \quad (5.17)$$

Due to compactness (and thus bounding) of Ω and $\Sigma^{(\gamma)}$ as defined in Equation 5.4, it holds that $\varepsilon > 0$. Using Equation 5.17 and the Definition 5.12, it follows that

$$q^{(\gamma)}(y|x) \geq \varepsilon \quad (5.18)$$

for all $x, y \in \Omega$. Note, that Equation 5.18 holds for any $\gamma \in \Gamma$ as the matrices $\Sigma^{(\gamma)}, \Sigma_r^{(\gamma)}$ are positive definite (Craiu et al., 2009, proof of Theorem 2) and (Haario et al., 2001, Theorem 1). Furthermore, there exists

$$d = \sup_{x \in \Omega} (\pi(x|\mathcal{D})) < \infty. \quad (5.19)$$

By splitting any $B \subseteq \mathcal{A}$ into a set $R_\gamma(x, B) = \{y \in B : \left(\frac{\pi(y|\mathcal{D})}{\pi(x|\mathcal{D})}\right)^{1/\tau} \frac{q^{(\gamma)}(y|x)}{q^{(\gamma)}(x|y)} < 1\}$ and $A_\gamma(x, B) = B \setminus R_\gamma(x, B)$ with $x \in \Omega$ one can show for a one-step kernel as defined in Equation 5.11,

$$\begin{aligned} P_{(\gamma)}(x, B) &= \int_{R_\gamma} q^{(\gamma)}(y|x) \min \left\{ 1, \left(\frac{\pi(y|\mathcal{D})}{\pi(x|\mathcal{D})}\right)^{1/\tau} \frac{q^{(\gamma)}(x|y)}{q^{(\gamma)}(y|x)} \right\} \mu^{Leb}(dy) \\ &\quad + \int_{A_\gamma} q^{(\gamma)}(y|x) \min \left\{ 1, \left(\frac{\pi(y|\mathcal{D})}{\pi(x|\mathcal{D})}\right)^{1/\tau} \frac{q^{(\gamma)}(x|y)}{q^{(\gamma)}(y|x)} \right\} \mu^{Leb}(dy) \end{aligned}$$

$$\begin{aligned}
&= \int_{R_\gamma} q^{(\gamma)}(x|y) \left(\frac{\pi(y|\mathcal{D})}{\pi(x|\mathcal{D})} \right)^{1/\tau} \mu^{Leb}(dy) + \int_{A_\gamma} q^{(\gamma)}(y|x) \mu^{Leb}(dy) \\
&\geq \frac{\varepsilon}{d^{(1/\tau)}} \int_{R_\gamma} \pi(y|\mathcal{D})^{(1/\tau)} \mu^{Leb}(dy) + \frac{\varepsilon}{d^{(1/\tau)}} \int_{A_\gamma} \pi(y|\mathcal{D})^{(1/\tau)} \mu^{Leb}(dy) \\
&= \frac{\varepsilon}{d^{(1/\tau)}} \pi(B|\mathcal{D})^{(1/\tau)}. \tag{5.20}
\end{aligned}$$

By (Meyn and Tweedie, 2012, Definition 5.14) and Equation 5.20, $B = \Omega$ is v -small with $P_{(\gamma)}(x, \Omega) \geq \frac{\varepsilon}{d^{(1/\tau)}}$. From this it follows that the chain is uniformly ergodic (Meyn and Tweedie, 2012, Theorem 16.0.2). Thus, Equation 5.16 holds. It remains to be proven, that Equation 5.15 holds as well.

To prove Equation 5.15, let $M = \max_{r \in \{1, \dots, R\}} \{\sup_{x, y \in \Omega} q_r^{(\gamma)}(y|x) : r = 1, \dots, R\}$ where $q_r^{(\gamma)}(y|x)$ is defined by Equation 5.13. M is finite. Without loss of generality, let $x \in \Omega_1$ and $A \subseteq \mathcal{A}$. As $\Omega = \cup_{r=1}^R \Omega_r$ by construction, one can write

$$\begin{aligned}
P_{(\gamma_i)}(x, A) &= \sum_{r=1}^R \int_{A \cap \Omega_r} q_r^{(\gamma_i)}(y|x) \cdot \\
&\quad \min \left\{ 1, \left(\frac{\pi(y|\mathcal{D})}{\pi(x|\mathcal{D})} \right)^{1/\tau} \frac{q_r^{(\gamma_i)}(x|y)}{q_1^{(\gamma_i)}(y|x)} \right\} \mu^{Leb}(dy). \tag{5.21}
\end{aligned}$$

Consistent to earlier definitions, let

$$p_{acc,r}^{(\gamma_i)}(x, y) = \min \left\{ 1, \left(\frac{\pi(y|\mathcal{D})}{\pi(x|\mathcal{D})} \right)^{1/\tau} \frac{q_r^{(\gamma_i)}(x|y)}{q_1^{(\gamma_i)}(y|x)} \right\}. \tag{5.22}$$

Thus, to calculate $|P_{(\gamma_{i+1})}(x, A) - P_{(\gamma_i)}(x, A)|$ one observes terms as

$$I_r = \left| \int_{A \cap \Omega_r} \left(q_r^{(\gamma_{i+1})}(y|x) p_{acc,r}^{(\gamma_{i+1})}(x, y) - q_r^{(\gamma_i)}(y|x) p_{acc,r}^{(\gamma_i)}(x, y) \right) \mu^{Leb}(dy) \right| \tag{5.23}$$

for all $r = 1, \dots, R$. One may appraise,

$$\begin{aligned}
I_r &\leq \int_{A \cap \Omega_r} \left| \left(q_r^{(\gamma_{i+1})}(y|x) p_{acc,r}^{(\gamma_{i+1})}(x, y) - q_r^{(\gamma_i)}(y|x) p_{acc,r}^{(\gamma_i)}(x, y) \right) \right| \mu^{Leb}(dy) \\
&= \int_{A \cap \Omega_r} \left| q_r^{(\gamma_{i+1})}(y|x) p_{acc,r}^{(\gamma_{i+1})}(x, y) + q_r^{(\gamma_{i+1})}(y|x) p_{acc,r}^{(\gamma_i)}(x, y) \right. \\
&\quad \left. - q_r^{(\gamma_{i+1})}(y|x) p_{acc,r}^{(\gamma_i)}(x, y) - q_r^{(\gamma_i)}(y|x) p_{acc,r}^{(\gamma_i)}(x, y) \right| \mu^{Leb}(dy) \\
&\leq \int_{A \cap \Omega_r} q_r^{(\gamma_{i+1})}(y|x) \left| p_{acc,r}^{(\gamma_{i+1})}(x, y) - p_{acc,r}^{(\gamma_i)}(x, y) \right| \mu^{Leb}(dy) \\
&\quad + \int_{A \cap \Omega_r} p_{acc,r}^{(\gamma_i)}(x, y) \left| q_r^{(\gamma_{i+1})}(y|x) - q_r^{(\gamma_i)}(y|x) \right| \mu^{Leb}(dy)
\end{aligned}$$

$$\begin{aligned}
&\leq M \int_{A \cap \Omega_r} \left| p_{acc,r}^{(\gamma_{i+1})}(x, y) - p_{acc,r}^{(\gamma_i)}(x, y) \right| \mu^{Leb}(dy) \\
&\quad + \int_{A \cap \Omega_r} \left| q_r^{(\gamma_{i+1})}(y|x) - q_r^{(\gamma_i)}(y|x) \right| \mu^{Leb}(dy)
\end{aligned} \tag{5.24}$$

and the first term can be further appraised to

$$\begin{aligned}
&M \int_{A \cap \Omega_r} \left| p_{acc,r}^{(\gamma_{i+1})}(x, y) - p_{acc,r}^{(\gamma_i)}(x, y) \right| \mu^{Leb}(dy) \\
&= M \int_{A \cap \Omega_r} \left(\frac{\pi(y|\mathcal{D})}{\pi(x|\mathcal{D})} \right)^{1/\tau} \left| \frac{q_r^{(\gamma_{i+1})}(x|y)}{q_1^{(\gamma_{i+1})}(y|x)} - \frac{q_r^{(\gamma_i)}(x|y)}{q_1^{(\gamma_i)}(y|x)} \right| \mu^{Leb}(dy) \\
&\leq \frac{M d^{1/\tau}}{\pi(x|\mathcal{D})^{1/\tau}} \int_{A \cap \Omega_r} \left| \frac{q_r^{(\gamma_{i+1})}(x|y)}{q_1^{(\gamma_{i+1})}(y|x)} - \frac{q_r^{(\gamma_i)}(x|y)}{q_1^{(\gamma_i)}(y|x)} \right| \mu^{Leb}(dy).
\end{aligned} \tag{5.25}$$

Other than in the setting of (Crainu et al., 2009), RAMPART uses a fixed ratio p_g over all iterations. Thus, overall it remains to be proven, that $\left| q_r^{(\gamma_{i+1})}(y|x) - q_r^{(\gamma_i)}(y|x) \right|$ and $\left| \frac{q_r^{(\gamma_{i+1})}(x|y)}{q_1^{(\gamma_{i+1})}(y|x)} - \frac{q_r^{(\gamma_i)}(x|y)}{q_1^{(\gamma_i)}(y|x)} \right|$ do vanish with increasing i . This, however, follows from Corollary 3 (Haario et al., 2001) for the proof for the AM algorithm by Haario et al., since $\lim_{i \rightarrow \infty} \Sigma_r^{(\gamma_i)} \propto \text{Cov}(\Omega_r)$ and $\lim_{i \rightarrow \infty} \Sigma^{(\gamma_i)} \propto \text{Cov}(\Omega)$. Therefore, $I_r \rightarrow 0$ follows for all r and simultaneous uniform ergodicity 5.16 is proven.

In summary this shows, that samples generated by RAMPART will converge regardless of the chain-initial point and tuning parameters. However, as discussed in Chapter 2 and 3, this finding provides a necessary but not a sufficient criterion for the generation of representative samples in practice. Therefore, in the following the practical performance of RAMPART will be quantified to provide the full picture.

5.3 Performance benchmark

To assess the performance and robustness of RAMPART, two simulation examples and two biological problems are studied. As a reference, own implementations of state-of-the-art RB-AM and adaptive PT algorithms are considered.

To ensure that the results are representative, 100 runs with 10^6 iterations per algorithm and problem were performed. Overall, the assessment is based on all these runs, which in total correspond to 18.000 CPU hours. RAMPART is compared to different state-of-the-art algorithms. To ensure an unbiased comparison, a semi-automated analysis pipeline

described in Chapter 3 is employed.

5.3.1 Artificial problems

To illustrate the properties of RAmPART, two established artificial problems were considered: Gaussian mixture density (Lacki and Miasojedow, 2015) and blurred ring (Kramer, 2016).

- The **20-dimensional Gaussian mixture** density is given by

$$\pi_{\text{gm}}(\theta|\mathcal{D}) \propto \left(\sum_{i=1}^2 \mathcal{N} \left(\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \mid \begin{pmatrix} \mu_{i,1} \\ \mu_{i,2} \end{pmatrix}, \Sigma \right) \right) \prod_{j=3}^{20} \mathcal{N}(\theta_j | 25, \sigma^2), \quad (5.26)$$

with $\mu_1 = (-50, -50)^t$, $\mu_2 = (50, 50)^t$, $\Sigma = s \cdot 250 \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} + \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ and $s, \sigma = 1$. The box constraints are summarized in Table 5.3. This defines a mixture of two separated Gaussian modes in two dimensions, whose largest eigenvectors are orthogonal to the connection line between the mode centers. The 18 other dimensions of this problem possess simple uncorrelated Gaussian modes (Figure 5.4a). Here, s does control the separation of the mixture-modes, while σ denotes the width of the simple Gaussian modes.

- The **20-dimensional blurred ring** density is defined by

$$\pi_{\text{ring}}(\theta|\mathcal{D}) \propto \mathcal{N}(r(\theta) | r_0, \sigma_r^2) \prod_{j=3}^{20} \mathcal{N}(\theta_j | 0, \sigma^2) \quad (5.27)$$

with $r(\theta) = \sqrt{\theta_1^2 + \theta_2^2}$, $\sigma = 1$, $r_0 = 50$ and $\sigma_r = 5$. The box constraints are summarized in Table 5.4. This defines a ring-like correlation structure in the first two parameters, while r_0 controls the radius and σ_r the width of the blurred ring. The other 18 parameters are again distributed by uncorrelated normal densities with standard deviation σ (Figure 5.4b).

Table 5.3: The parameter constraints for the 20-dimensional Gaussian mixture density.

Parameter Name	θ_{\min}	θ_{\max}
$\theta_i, i = 1, \dots, 20$	-100	100

The benchmark problems were sampled with RB-AM, PT and RAmPART. For PT and RAmPART, 40 temperature levels with $\tau_{\max} = 2000$ were employed. RAmPART has

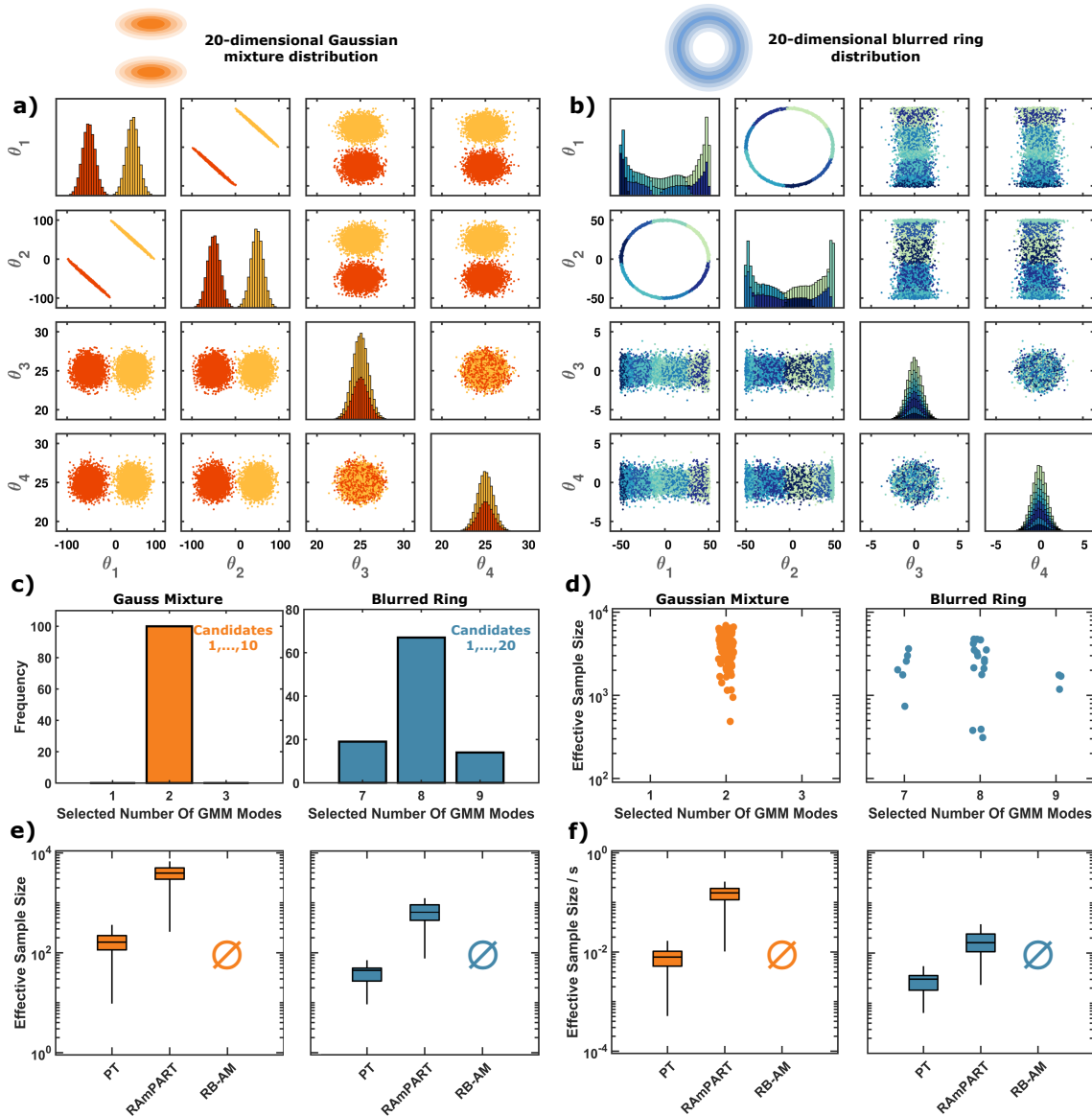


Figure 5.4: **RAMPART** outperforms established methods for simulation examples: **20-dimensional Gaussian mixture** and the **20-dimensional blurred ring**. **a-b)** Bivariate scatter plot matrix and histograms for a representative chain generated using **RAMPART**. Parameters $\theta_1 - \theta_4$ are illustrated. The parameters $\theta_5 - \theta_{20}$ possess the same density as θ_3 and θ_4 . **c)** Variability of the selected GMM complexity between **RAMPART** runs. **d)** The remaining ESS for different selected GMM complexities after application of the analysis pipeline. **e-f)** Quantitative assessment of the effective sample size and the effective sample size per second computation time for different algorithms.

Table 5.4: The parameter constraints for the 20-dimensional blurred ring density.

Parameter Name	θ_{\min}	θ_{\max}
$\theta_i, i = 1, 2$	-200	200
$\theta_i, i = 3, \dots, 20$	-20	20

Table 5.5: Summarized run times (in seconds) per iteration. This includes runs which were not able to converge as well.

	Gaussian Mixture	Blurred Ring
RB-AM	$1.4 \cdot 10^{-1}$	$1.2 \cdot 10^{-1}$
PT	1.4	2.1
RAmPART	3.0	3.7

been initialized with the last 50% of a PT run with 10^5 samples. For this sample size, PT explored a large fraction of the parameter space but samples were not representative for the posterior distribution. However, the inference of GMMs using EM already provided reasonable definitions of regions (Figure 5.4a-d).

The number of regions and their location differed between runs. Thus, it is interesting to observe whether the sampling performance during the sampling phase depends on the number of regions and if the automatically chosen number of regions varies. For the 20-dimensional Gaussian mixture example, it was found that the cross-validations always yielded 2 regions (Figure 5.4c, left panel). For the 20-dimensional blurred ring, on average 8 regions were selected for the considered sample size (Figure 5.4c, right panel). Here, the covariance matrices of regions differed substantially, which allowed for a good approximation of the density. For both examples, the sampling efficiencies were independent of the number of regions and their locations (Figure 5.4d).

The evaluation of the sampling results revealed that RB-AM and PT suffer from convergence problems for the considered sample size (Figure 5.4e,f). The RB-AM did neither provide representative samples for the Gaussian mixture nor for the blurred ring as it failed to explore the posterior. In most runs, PT provided a representative sample from the Gaussian mixture, but not for the blurred ring. RAmPART converged in 95 of 100 runs for the Gaussian mixture and in 25 of 100 runs for the blurred ring. The computation time per iteration was highest for RAmPART (Table 5.5). Yet, due the decreased auto-correlation, the effective sample size was highest for RAmPART, both in absolute terms and relative to the computation time (Figure 5.4e,f).

The comparison of PT and RAmPART revealed that the use of region-based random walk

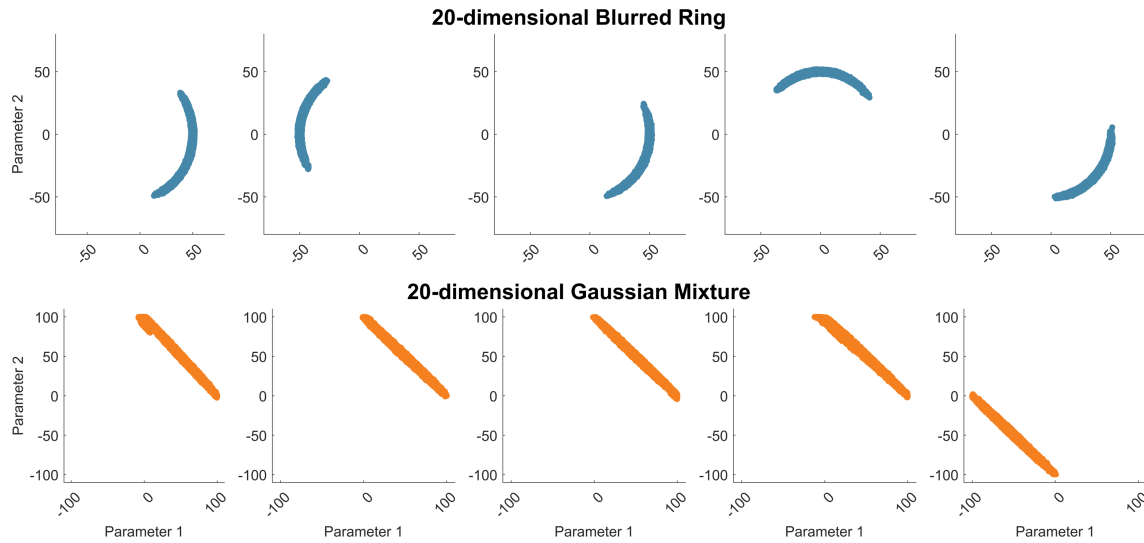


Figure 5.5: **Bivariate scatter plots of AM samples in the 20-dimensional blurred ring and 20-dimensional Gaussian mixture. Upper panel: 20-dimensional blurred ring. Lower panel: 20-dimensional Gaussian mixture.**

proposals ($p_g = 0.5$) improved convergence and computational efficiency. Accordingly, it was not clear whether a pure region-based random walk proposal would perform even better ($p_g = 1$). Interestingly, this was not the case for the considered benchmark problems. This indicates, that the small-world proposal used by RAMPART for $p_g = 0.5$ exploits benefits of local and global proposals and improves the overall robustness in particular across different regions.

Driven by the experience gathered in Chapter 4, it was expected that the standard Metropolis Hastings (MH) method and even its adaptive variates will not perform well for the considered problems. To confirm this hypothesis, 5 runs using an adaptive Metropolis (AM) algorithm for the 20-dimensional blurred ring and 20-dimensional Gaussian mixture examples were performed (Figure 5.5).

In comparison to the results in Figure 5.4, one can see, that the generated samples are not representative for the underlying posterior structure as either parts of the ringed mode were missed or only one of the two Gaussian modes were captured. These runs would have been filtered by the employed analysis pipeline and thus would have not contributed to the overall ESS of the method.

5.3.2 Application problems

To evaluate RAmPART in practice, the processes of mRNA transfection and Epo-induced JAK2/STAT5 signaling were considered:

- **mRNA transfection** is a promising treatment option, among others, in immunotherapy (Kuhn et al., 2011). In mRNA transfection, mRNA is encapsulated in lipoplexes, transported across the cell membrane, released into the cytosol and being translated (Figure 5.6a). Single cell time-lapse data for this process have been collected and modeled by (Leonhardt et al., 2014) (Figure 5.6b). In this study, the model introduced by (Leonhardt et al., 2014) is considered, inferring its five parameters from a representative single-cell trace. The model was already employed as a benchmarking problem in Chapter 4 and found to give rise to challenging posterior landscapes. For readability, some redundant details are covered in the following. The ODE of the model reads

$$[\dot{\text{GFP}}] = k_{TL}[\text{mRNA}] - \beta[\text{GFP}], \quad [\text{GFP}(t \leq 0)] = 0, \quad (5.28)$$

$$[\dot{\text{mRNA}}] = -\delta[\text{mRNA}], \quad [\text{mRNA}(t)] = \begin{cases} 0, & t < 0 \\ m_0, & t = t_0 \end{cases} \quad (5.29)$$

in which $[\text{GFP}]$ denotes the concentration of green fluorescent protein and $[\text{mRNA}]$ denotes the concentration of mRNA of green fluorescent protein. The mRNA is released into the cell at t_0 and the initial concentration is m_0 . The mRNA is translated with rate k_{TL} . Degradation rates for mRNA and protein are δ and β , respectively. The analytical solution for the ODE model is given by

$$[\text{mRNA}(t)] = \begin{cases} 0, & t < t_0, \\ m_0 e^{-\delta(t-t_0)}, & \text{otherwise,} \end{cases} \quad (5.30)$$

and

$$[\text{GFP}(t)] = \begin{cases} 0, & t < t_0 \\ k_{TL} m_0 (e^{-\beta(t-t_0)} - e^{-\delta(t-t_0)}) / (\delta - \beta), & t \geq t_0 \wedge \delta \neq \beta \\ k_{TL} m_0 (t - t_0) e^{-\delta(t-t_0)}, & t \geq t_0 \wedge \delta = \beta \end{cases} \quad (5.31)$$

The experimental data was collected for $[\text{GFP}(t)]$ at 150 time points in the time interval $t \in [2, 27]$ hours (Leonhardt et al., 2014). For such data, the parameters k_{TL} and m_0 are structurally non-identifiable, as $[\text{GFP}(t)]$ merely depends on the product

$\kappa = k_{TL}m_0$. This problem is addressed by estimating merely κ . In addition, β and δ are only locally structural identifiable. The values of the two parameters can be interchanged without altering the observable $[GFP(t)]$. In previous studies, it was assumed that the mRNA half-life is smaller than the protein half-life, implying $\beta < \delta$. As this is not necessarily correct, this constraint is not applied here. Simulation and data are compared using an error model assuming the measurement noise is normally distributed with standard deviation σ . As σ is unknown, it is taken into account as an additional parameter for parameter estimation. The parameter constraints are reported in Table 5.6.

- **Epo-induced JAK2/STAT5 signaling** is essential for survival, proliferation and differentiation in hematopoiesis (Bachmann et al., 2011). Epo binds to the complex of the Epo receptor (EpoR) and JAK2. The complex induces phosphorylation of STAT5, which subsequently dimerises and translocates to the nucleus to regulate gene expression (Figure 5.6c). An initial model of this process has been developed by (Swameye et al., 2003). Here, the implementation of the model by (Maier et al., 2017) with 17 parameters and fit it to quantitative immunoblotting data is considered (Figure 5.6d).

For JAK2/STAT5 signaling the ODE model introduced by (Maier et al., 2017) is considered, which is based on the original publication by (Swameye et al., 2003). The ODE model is defined by

$$\begin{aligned}
\frac{d}{dt}[\text{STAT}] &= (\Omega_{nuc} \cdot p_4 \cdot [\text{nSTAT}_5] - \Omega_{cyt} \cdot [\text{STAT}] \cdot p_1 \cdot u(t)) / \Omega_{cyt} \\
\frac{d}{dt}[\text{pSTAT}] &= [\text{STAT}] \cdot p_1 \cdot u(t) - 2p_2 \cdot [\text{pSTAT}]^2 \\
\frac{d}{dt}[\text{pSTAT2}] &= p_2 \cdot [\text{pSTAT}]^2 - p_3 \cdot [\text{pSTAT2}] \\
\frac{d}{dt}[\text{npSTAT2}] &= -(\Omega_{nuc} \cdot p_4 \cdot [\text{npSTATmpSTAT}] - \Omega_{cyt} \cdot p_3 [\text{pSTAT2}]) / \Omega_{nuc} \\
\frac{d}{dt}[\text{nSTAT}_1] &= -p_4([\text{nSTAT}_1] - 2[\text{npSTAT2}]) \\
\frac{d}{dt}[\text{nSTAT}_2] &= p_4([\text{nSTAT}_1] - [\text{nSTAT}_2]) \\
\frac{d}{dt}[\text{nSTAT}_3] &= p_4([\text{nSTAT}_2] - [\text{nSTAT}_3]) \\
\frac{d}{dt}[\text{nSTAT}_4] &= p_4([\text{nSTAT}_3] - [\text{nSTAT}_4]) \\
\frac{d}{dt}[\text{nSTAT}_5] &= p_4([\text{nSTAT}_4] - [\text{nSTAT}_5])
\end{aligned} \tag{5.32}$$

in which u is the time dependent level of phosphorylated Epo receptor and the initial

conditions are defined by $x(0) = 0$ for all states except for $[\text{STAT}](0) = [\text{STAT}]_{tot}$. The phosphorylated Epo receptor initiates JAK2/STAT5 signalling and it is modeled using a cubic spline function. This function has the values $u(0) = sp_1$, $u(5) = sp_2$, $u(10) = sp_3$, $u(20) = sp_4$, $u(60) = sp_5$. For the process, measurement data for the amount of phosphorylated STAT and total STAT in the cytosol is taken into account,

$$y_{[\text{pSTAT}]} = O_{[\text{pSTAT}]} + s_{[\text{pSTAT}]} / [\text{STAT}]_{tot} ([\text{pSTAT}] + 2[\text{pSTAT}2]) \quad (5.33)$$

$$y_{[\text{tSTAT}]} = O_{[\text{tSTAT}]} + s_{[\text{tSTAT}]} / [\text{STAT}]_{tot} ([\text{STAT}] + [\text{pSTAT}] + 2[\text{pSTAT}2]) \quad (5.34)$$

using offset parameters O and scaling parameters s and the concentration of phosphorylated Epo receptor

$$y_{[\text{pEpoR}]} = u(t). \quad (5.35)$$

The unknown parameters of the models are $\theta = (p_1, p_2, p_3, p_4, [\text{STAT}]_{tot}, sp_1, sp_2, sp_3, sp_4, sp_5, O_{[\text{tSTAT}]}, O_{[\text{pSTAT}]}, s_{[\text{tSTAT}]}, s_{[\text{pSTAT}]}, \sigma_{[\text{pSTAT}]}, \sigma_{[\text{tSTAT}]}, \sigma_{[\text{pEpoR}]})^T$, the volumes of cytosol and nucleus (Ω_{cyt} and Ω_{nuc}) are known. The box constraints are defined in Table 5.7.

For both applications, \log_{10} -transformed parameters were employed and uniform prior densities were used.

The posterior distributions of the models were sampled using RB-AM, PT and RAmPART with the same settings as for the simulation examples (Figure 5.4) but with $L = 60$, $\tau_{\max} = 4000$ for the JAK2/STAT5 problem and $L = 30$, $\tau_{\max} = 2000$ for the mRNA transfection problem. RAmPART provided the most runs with a representative sample of the posterior distribution, while especially RB-AM suffered from convergence problems. The trajectories contained in the representative samples by PT and RAmPART provide a good description of the experimental data (Figure 5.6b,d).

Table 5.6: The parameter constraints for the model of mRNA transfection.

Parameter Name	θ_{\min}	θ_{\max}
$\log_{10}(t_0)$	-2	1
$\log_{10}(k_{TL}m_0)$	-5	5
$\log_{10}(\beta)$	-5	5
$\log_{10}(\delta)$	-5	5
$\log_{10}(\sigma)$	-2	2

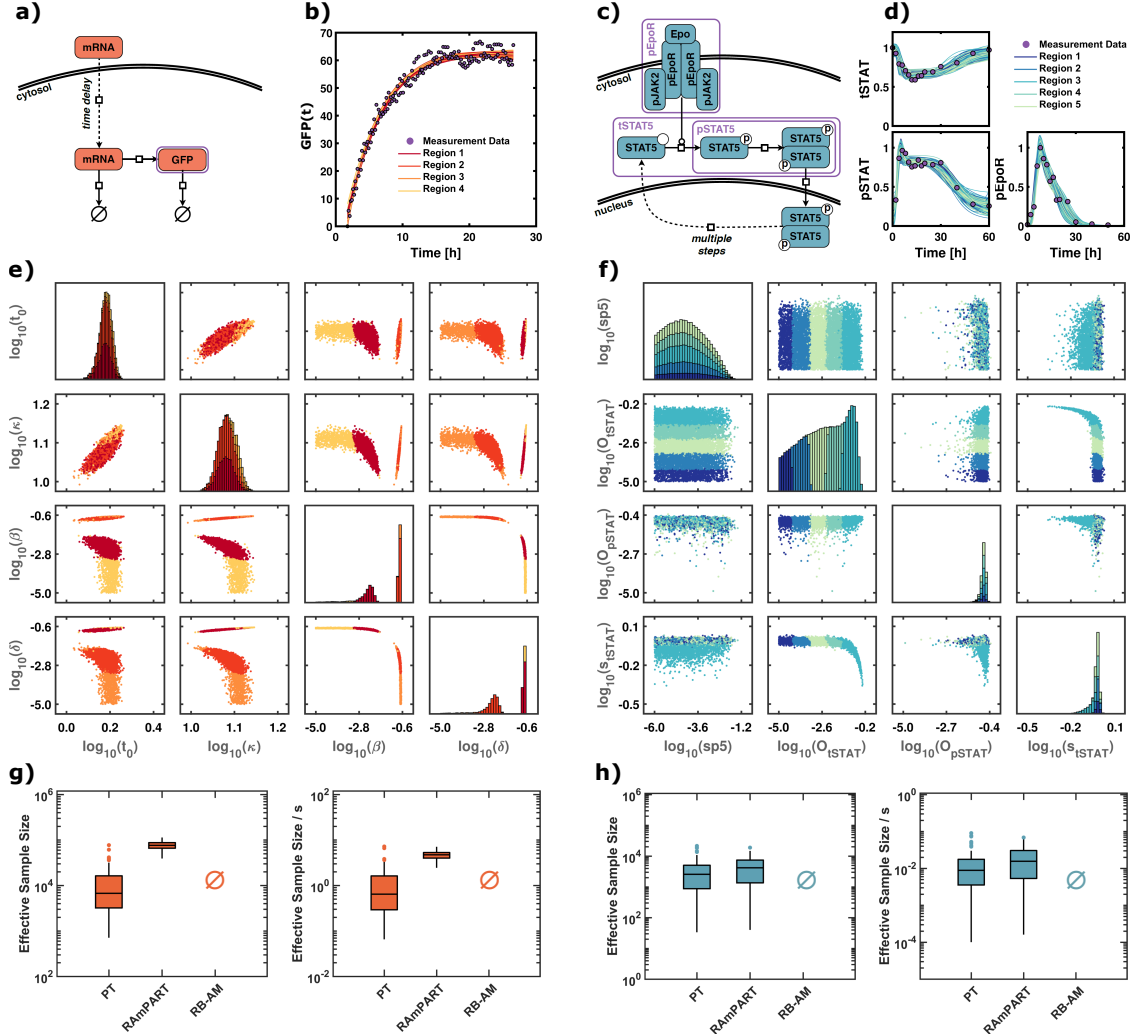


Figure 5.6: **RAMPART** adapts to the posterior landscape and outperforms established methods for models of mRNA transfection and JAK2/STAT5 signaling. **a)** Biochemical reaction network for model of mRNA transfection. **b)** Measurement data and propagated model trajectories of the observable $G(t)$ derived from the parameter sample points. **c)** Biochemical reaction network for model of JAK2/STAT5 signaling. **d)** Measurement data and propagated model trajectories of the observables tSTAT, pSTAT and pEpoR derived from the parameter sample points. **e),f)** Bivariate scatter plot of one RAMPART MCMC chain. The colors indicate the different regions. **g),h)** Comparison of the effective sample size and the effective sample size per second computation time for the sampling algorithms.

Table 5.7: The parameter constraints for the JAK2/STAT5 signaling application.

Parameter Name	θ_{\min}	θ_{\max}
$\log_{10}(p_1)$	-5	5
$\log_{10}(p_2)$	-3	6
$\log_{10}(p_3)$	-5	5
$\log_{10}(p_4)$	-3	6
$\log_{10}([\text{STAT}]_{tot})$	-5	5
$\log_{10}(sp_1)$	-5	5
$\log_{10}(sp_2)$	-5	5
$\log_{10}(sp_3)$	-5	5
$\log_{10}(sp_4)$	-5	5
$\log_{10}(sp_5)$	-6	5
$\log_{10}(O_{[t\text{STAT}]})$	-5	5
$\log_{10}(O_{[p\text{STAT}]})$	-5	5
$\log_{10}(s_{[t\text{STAT}]})$	-5	5
$\log_{10}(s_{[p\text{STAT}]})$	-5	5
$\log_{10}(\sigma_{[p\text{STAT}]})$	-5	5
$\log_{10}(\sigma_{[t\text{STAT}]})$	-5	5
$\log_{10}(\sigma_{[p\text{EpoR}]})$	-5	5

The sampling results for the model of mRNA transfection revealed a bimodal posterior (Figure 5.6e). Accordingly, in all runs RAmPART selected at least 2 regions, on average 4.8. In all runs, the two modes were separated into different regions, allowing RAmPART to account for the different local correlation structure. Furthermore, each mode was often split into several regions to cover the tails (see Figure 5.6e). RAmPART identified the symmetry between the degradation rates, β and γ , on the level of the sample and the level of the regions. This symmetry is associated to structural identifiability problems. However, it can not be identified by established tools for structural identifiability analysis, such as GenSSI 2.0 (Ligon et al., 2017). Thus, the automatic identification of symmetries offered by RAmPART is important.

The sample for the model of JAK2/STAT5 signaling, did not reveal multi-modality as for the mRNA transfection but practical non-identifiabilities (Figure 5.6f). Practical non-identifiabilities manifested as tails in the posterior densities and were visible, among others, for $O_{[t\text{STAT}]}$ and $s_{[t\text{STAT}]}$. If these parameter dimensions are considered for the training of the GMM, RAmPART selects 4 - 5 regions. These regions partition the densities and possess different correlations structure, facilitating the construction of a tailored proposal density.

The overall run times (Table 5.8) and the number of runs which provide a representative

Table 5.8: Summarized run times (in seconds) per iteration. This includes runs which were not able to generate representative samples.

	mRNA Transfection	JAK2/STAT5 Signaling
RB-AM	$7.5 \cdot 10^{-2}$	$2.6 \cdot 10^{-1}$
PT	1.1	$2.7 \cdot 10^1$
RAmPART	1.6	$2.5 \cdot 10^1$

Table 5.9: Summarized the number of converged runs in each of the examples. Any chain which has a non-vanishing ESS counts as converged. In total, 100 runs were started.

	mRNA Transfection	JAK2/STAT5 Signaling
RB-AM	0	0
PT	100	86
RAmPART	100	99

sample (Table 5.9) were assessed. In comparison to PT, RAmPART requires additional computational effort. Please note, in Chapter 3 and 4 the fraction of runs which provide a representative sample was called exploration quality (EQ) of the algorithm. For the mRNA transfection model, PT and RAmPART both showed a full EQ while RB-AM, as expected, did not generate a representative sample at all (Table 5.9). For the JAK2/STAT5 model, the additional computational effort is small compared to the objective function evaluation and leads to slightly higher EQ. This is probably the most realistic scenario for parameter estimation problems in systems biology.

Overall, the evaluation of the sampling performance and robustness revealed that RAmPART is more efficient than the established methods for both application problems (Figure 5.6g,h). For mRNA transfection, RAmPART achieves a 6.6-fold higher ESS/t than PT (Figure 5.6g). The key reason probably was the improved alignment of the regional proposal densities of RAmPART compared to the coverage of the two modes by the global proposal density of PT. For JAK2/STAT5 signaling, it was found that RAmPART doubled the ESS/t in presence of an apparently rather simple posterior structure (Figure 5.6h). Apparently, even though the posterior structure is uni-modal, a Gaussian mixture model provides a substantially better approximation than a single Gaussian due to the pronounced tails of the posterior which allows RAmPART to slightly outperform PT.

5.4 Aimed temperature selection

Applying tailored algorithms for estimating parameter distributions in application problems can increase the performance significantly. This typically demands the selection of problem specific tuning parameters, which often make or break the given method. Unfortunately, in practical parameter estimation problems one typically faces an unknown posterior and it is not clear how to choose proper tuning parameters. By applying multiple adaptation schemes and an automated region learning, RAmPART limits the number of problem specific tuning parameters. However, critical tuning parameters exist for RAmPART as well. The most crucial one of them appears to be the maximum temperature τ_{\max} . Too low values will limit the transition between different modes of the posterior distribution, while too high values render multiple tempered sub-chains redundant and lower the efficiency of the algorithm (Vousden et al., 2016; Lacki and Miasojedow, 2015). In this section, an approach for the selection of τ_{\max} is proposed and evaluated.

5.4.1 Motivation

In each of the L sub-chains, a candidate θ' proposed from θ is accepted with an acceptance ratio defined by Equation 5.5 which is proportional to $(\pi(\theta'|\mathcal{D})/\pi(\theta|\mathcal{D}))^{(1/\tau)}$ where $\tau \geq 1$ is the temperature. Proposed candidates θ' with $\pi(\theta'|\mathcal{D}) \ll \pi(\theta|\mathcal{D})$ are unlikely to get accepted for $\tau = 1$. However, depending on the structure of the posterior, one may want a chain whose tempered sub-chains ($\tau > 1$) are able to move through unlikely sets in the parameter space

$$\omega_{\mathcal{M}} = \{\theta \in \Omega : \pi(\theta|\mathcal{D}) \ll \pi(\theta_{\text{opt}}|\mathcal{D}), \theta_{\text{opt}} \in \mathcal{M}\}, \quad (5.36)$$

where \mathcal{M} denotes the set of local optima of the posterior, aka. modes of the posterior,

$$\mathcal{M} = \{\theta \in \Omega : \exists \delta > 0, \text{ s.t. } \forall \tilde{\theta} \in S_{\delta}(\theta) \subseteq \Omega \text{ is } \pi(\tilde{\theta}|\mathcal{D}) \leq \pi(\theta|\mathcal{D})\}, \quad (5.37)$$

in order to explore the posterior representatively (see Chapter 3 and 4). Here, $S_{\delta}(\theta) \subseteq \Omega$ denotes a k -dimensional sphere with radius δ centered around θ . In particular, for a posterior with multiple modes, it is important, that at least the hottest sub-chain ($\tau = \tau_{\max}$) is able to move between the modes. Thus, finding a sufficiently high τ_{\max} such that $(\pi(\theta'|\mathcal{D})/\pi(\theta|\mathcal{D}))^{(1/\tau_{\max})} \approx 1$ with $\pi(\theta'|\mathcal{D}) \ll \pi(\theta|\mathcal{D})$ holds for all $\theta' \in \omega_{\mathcal{M}}, \theta \in \Omega$ is crucial.

While it should be avoided to choose τ_{\max} too small, too large values are not beneficial either. As the swap acceptance depends on the inverse temperature difference $\Delta\beta$ between adjacent sub-chain, higher τ_{\max} typically demand larger L in order to maintain frequent

swaps between each pair of adjacent sub-chains. This causes additional computational effort. In addition, larger L may cause (indirect) swaps between the sub-chains of $\tau = 1$ and $\tau = \tau_{\max}$ to get less likely because all adjacent intermediate swaps have to be accepted first. Thus, it is desirable to use the smallest τ_{\max} possible.

5.4.2 Method

In the following, a heuristic based on multi-start local optimization for selecting τ_{\max} is proposed. Let \mathcal{M} be the set of local optima of the posterior density as defined in Equation 5.37. If \mathcal{M} is known, it can be used to define the global maximum of the posterior

$$\pi_{\max} = \max_{\theta \in \mathcal{M}} (\{\pi(\theta|\mathcal{D})\}) \quad (5.38)$$

and the minimum of the posterior across the straight connection lines between all local optima

$$\pi_{\min} = \min_{\theta_1 \neq \theta_2 \in \mathcal{M}} \left(\left\{ \min_{b \in [0,1]} \pi(\theta_1 + (\theta_2 - \theta_1)b|\mathcal{D}) \right\} \right). \quad (5.39)$$

This set of equations defines a maximum height difference between the global optimum π_{\max} and the smallest constrained local minimum π_{\min} along the Euclidean connections between all pairs of local optima (see Figure 5.7),

$$\varepsilon = \log \left(\pi_{\max}^{(1/\tau_{\max})} \right) - \log \left(\pi_{\min}^{(1/\tau_{\max})} \right). \quad (5.40)$$

By construction ε is positive and a user defined upper bound of height differences in log-posterior space between local optima and the lowest points connecting them.

As \mathcal{M} is the set of local optima of the posterior, its elements can be found using sufficiently many multi-start local optimizations. However, \mathcal{M} may have infinite many elements, e.g. in the presence of non-identifiabilities, in which case finite many multi-start local optimizations would yield a subset $\mathcal{M}' \subseteq \mathcal{M}$ which is expected to cause no problems if \mathcal{M}' is sufficiently diverse. In cases where \mathcal{M} has ≤ 1 elements, the equations are not well defined. In this case, one could simply set τ_{\max} to a arbitrary constant without using the discussion of this section with no expected loss of performance as the problem is mono-modal.

Equation 5.40 can be used to inform a problem specific maximum temperature

$$\tau_{\max}(\varepsilon) = \frac{1}{\varepsilon} \log \left(\frac{\pi_{\max}}{\pi_{\min}} \right). \quad (5.41)$$

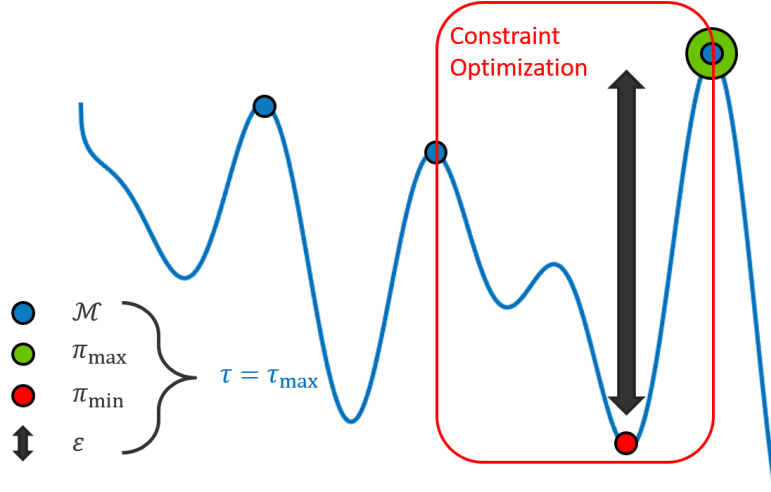


Figure 5.7: **A scheme for obtaining a problem specific τ_{\max} .** The blue line represents the posterior density $\pi(\theta|\mathcal{D})$. Missing blue points, correspond to local optima missed by the multi-start local optimization.

The chance for a Markov chain to move between modes located within $S_\delta(\theta_1)$ and $S_\delta(\theta_2)$ is based on both, the gap between π_{\max} and π_{\min} , and the corresponding distance in parameter space $\|\theta_1 - \theta_2\|_2$, $\theta_1, \theta_2 \in \mathcal{M}$. This motivates a heuristic for ε ,

$$\varepsilon = \max_{\theta_1, \theta_2 \in \mathcal{M}} (\|\theta_1 - \theta_2\|_2), \quad (5.42)$$

making τ_{\max} account for both.

In order to obtain the quantities required in Equation 5.41, one needs multiple local optimizations of the posterior to accumulate a reasonable subset of \mathcal{M} . However, for most MCMC based parameter estimation applications, a preceding multi-start local optimization is advisable anyways (see Chapter 4). Performing additional 1-dimensional optimizations as required for obtaining π_{\min} from Equation 5.39 is typically cheap making the overall approach for obtaining a problem specific τ_{\max} efficient. However, the computational effort increases quadratically with $|\mathcal{M}|$, which could be overcome by only considering the highest n_M local optima as a further simplification if necessary. Note, while a good choice of τ_{\max} is expected to improve sampling performance, ergodicity is independent of the choice of τ_{\max} . Additional research on heuristics similar to the one presented here, is expected to further improve the estimation of τ_{\max} .

Table 5.10: Selection of aimed τ_{\max} for each of the 3 20-dimensional Gaussian Mixture Problems.

	$s = 10^{-1}$	$s = 10^0$	$s = 10^1$
$\log(\pi_{\max})$	$-2.09 \cdot 10^1$	$-2.68 \cdot 10^2$	$-2.50 \cdot 10^4$
$\log(\pi_{\min})$	$-2.50 \cdot 10^4$	$-2.77 \cdot 10^3$	$-2.53 \cdot 10^4$
Aimed τ_{\max}	$1.77 \cdot 10^2$	$1.77 \cdot 10^1$	$1.76 \cdot 10^0$

5.4.3 Evaluation

To quantify the practical behavior of an aimed maximum temperature selection, an empirical study was executed using the 20-dimensional Gaussian mixture problem from the preceding sections. The problem was used in three versions, multiplying the covariances of the Gaussians with $s = 0.1$, $s = 1$ (as in earlier sections) and $s = 10$. For $s = 0.1$ the transition between the modes is rather difficult while for $s = 10$ the disconnection between the two modes is less pronounced. For each of the 3 problems, 10 runs with each combination of $L = 5, 10, 20$ and 40 with $\tau_{\max} = 10^{-6}, 10^{-5}, \dots, 1, 10, 10^2, \dots, 10^6$ were obtained and compared to runs using $\tau_{\max, \text{aimed}}$. Table 5.10 shows an overview of π_{\max} , π_{\min} and $\tau_{\max, \text{aimed}}$ for each of the problems. For all three problems $\varepsilon \approx 141$ was selected based on Equation 5.42. The results are analyzed using the analysis pipeline presented in Chapter 3.

While the results (visualized in Figure 5.8) are obtained in only three (artificial) problems and are thus not comprehensive, they already suggest that the performance of RAmPART does sensitively depend on the selection of L and τ_{\max} . This observation underpins the point of gaining or loosing performance benefits due to proper or improper tuning parameters. As expected, the number of sub-chains, L , directly impacts the ESS of the runs as too low numbers let the runs fail entirely. Raising τ_{\max} from low to high values gives rise to failure, good- ($ESS > 10^4$) and moderate ($ESS > 10^3$) sampling performance if L is sufficiently high. Furthermore, for sufficiently large L there apparently exist optimal maximum temperatures. Interestingly, the optimal maximum temperature τ_{\max} changes fewer than by a factor of 10 when the posterior density mode covariances are scaled with 0.1, 1 or 10 respectively. The heuristic for aimed tempering shows moderate to good results for these three problems typically being a bit higher than the empirically optimal maximum temperature τ_{\max} .

In summary, the aimed temperature selection presented in this section, is an attempt to further automatize RAmPART regardless of the given problem. First results showed, that a careful selection of critical tuning parameters as L or τ_{\max} are indeed crucial and

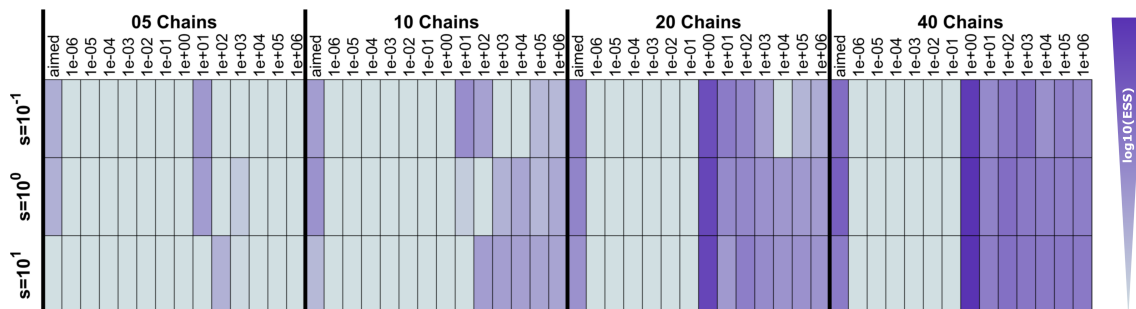


Figure 5.8: **Average ESS in three versions of the 20-dimensional Gaussian mixture problem.** Each square corresponds to the results of 10 runs possessing a certain set of tuning parameters.

the presented heuristic for a problem specific determination of τ_{\max} performed reasonably well. As only three similar problems were included in this study (in total 1680 runs with 10^6 iterations each) and a high performance impact of L and τ_{\max} was observed, further investigation in scope of future research is required.

5.5 Summary and discussion

The most common approach for comprehensive assessment of parameter probability distributions is Markov chain Monte Carlo sampling. However, for computationally demanding problems with posterior densities which possess multiple modes or pronounced tails, standard methods (i.e. the Metropolis Hastings or AM), are known to require massive computational resources in order to provide representative samples (Chapter 4).

Here, the region-based adaptive parallel tempering algorithm, RAmPART, which adapts to the tempered posterior densities and constructs tailored proposal densities on the fly, is proposed. Following the requirements formulated by (Craiu et al., 2009, pages 1464–1465) to ensure ergodicity, this multi-level adaptation is designed to achieve good sampling properties “within each region” and transition between “all regions”. In order to provide a theoretical basis a formal proof of ergodicity of RAmPART generated chains has been provided.

The performance of RAmPART was evaluated for benchmark and application problems. The analysis revealed that RAmPART possesses a higher computation cost per iteration than RB-AM and PT, but it also provides a higher effective sample size. The increased computational cost is compensated by the improved mixing which resulted in a higher

effective sample size per unit computation time. RAMPART outperformed the reference implementations of RB-AM and PT for all considered problems, by providing an improved ESS and a higher reliability of individual runs. Both aspects are highly relevant in practice and will allow for a consideration of higher-dimensional models with more involved posterior distributions. However, the results should be corroborated by analyzing a larger set of application problems in the future. As stronger parameter dependency structures are expected to enable RAMPART to provide greater benefits compared to state-of-the-art methods, a study with focus on the co-dimension of such dependency manifolds would be very interesting.

The robustness and degree of automation of RAMPART was further improved by the development of aimed maximum temperature selection. This is a novel approach that solves an open problem for tempering algorithms (Vousden et al., 2016; Lacki and Miasojedow, 2015) and initial evaluation results were very positive. However, there still is opportunity for further improvement of RAMPART, e.g. by adapting the regions during the sampling, instead of fixing them after the warm-up phase. Ideas by (Craiu et al., 2009) might be employed, as implemented in the single-chain algorithm RAPTOR. Complementary, more robust clustering approaches could be used (Levenstien et al., 2003; Gesteira Costa Filho, 2008) to enhance the robustness of region learning implemented in RAMPART. Alternatively, instead of using region-based proposal densities, Hamiltonian Monte Carlo methods (Hoffman and Gelman, 2014; Graham and Storkey, 2017) might be employed for the different temperatures.

In summary, RAMPART was introduced and a comprehensive evaluation was provided. The proposed algorithm has substantial practical value and is publicly available in the MATLAB toolbox PESTO. This will facilitate its reuse and application to a broad class of problems in particular ODE-constraint parameter estimation arising in computational biology.

Chapter 6

Conclusion

6.1 Summary and conclusions

Computational models are an important tool in systems biology. The available information about model parameters, given experimental data, is encoded in the corresponding posterior distribution. As complexity of data and mechanistic models in computational biology steadily increases, the feasibility and competitiveness of methods which apply to those models and data – in particular parameter estimation methods – are subjected to constant development. While optimization algorithms have been actively developed and compared, sampling methods as MCMC were neglected regarding broad quantitative evaluation. In this thesis, MCMC methods have been critically investigated, compared and improved in order to make their application more compatible for modern science and to support further development. For this purpose, the thesis employed tools from applied statistics, Markov theory, non-linear optimization, ODE modeling and machine learning.

In principle, MCMC methods are capable to provide a representative sample from the posterior distribution. This facilitates the assessment of uncertainties of parameters and model predictions. In particular, the identification of non-identifiabilities and the search of meaningful dependency structures between biological model parameter distributions proved valuable (Raue et al., 2009, 2013b; Hug, 2015). However, often a finite MCMC sample is not (yet) representative of the posterior and drawn conclusions may be wrong or misleading. Unfortunately, while in theory MCMC convergence is guaranteed by terms as ergodicity, law of large numbers and central limit theorems, for finite samples this is no longer necessarily true. The assessment of sample quality, however, is not straightforward, as passing existing statistical tests and convergence criteria is just necessary but not sufficient for an MCMC result in order to be representative.

To improve the robustness and reliability of MCMC convergence diagnostics, in Chapter 3 an analysis pipeline for the assessment of sample representativity was introduced. It includes multiple established tests for this purpose, e.g. Gelman-Rubin-Brooks, and accounts for burn-in, auto-correlation and proper exploration of the Markov chains. In

comparison to standard assessments, the advantage of the pipeline is a lower chance of false positives meaning it is less prone for overestimating sample quality. Furthermore, its semi-automated nature increases objectivity during the assessment as standard approaches often rely on manual observations.

In computational biology, ODE models are among the most important types of mechanistic models. Their development is often driven by prior knowledge or hypothesized biological mechanisms, e.g. a certain signaling cascade of proteins in a cell.

As mechanistic models in biology nowadays tend to become more and more high dimensional and typically include feedback loops, this practice of model construction often leads to mathematical properties such as multi-stability, oscillations, bifurcations or even chaotic dynamics in the model. Analyzing these properties analytically is usually not straightforward while they directly impact the shape of the posterior distribution and, thus, influence the performance of parameter estimation methods as MCMC. However, the connection between ODE model properties and sampling performance has never been assessed thoroughly. In order to close this gap, in Chapter 4 a collection of ODE based parameter problems is established covering a variety of different ODE properties to study.

Beside the quantitative assessment of the connection between ODE model and MCMC method performance, there has never been a comprehensive evaluation of state-of-the-art MCMC methods. This is in contrast to other classes of parameter estimation methods as optimization where extensive benchmarks exist. In order to close this gap and to derive an unbiased picture of the performance of state-of-the-art MCMC methods in ODE-constraint parameter estimation problems, in Chapter 4 a stocktaking of multiple MCMC methods has been provided. The corresponding analysis is based on the self developed analysis pipeline from Chapter 3. In total, the study includes 100 MCMC runs of 23 algorithms, tuning or initialization schemes in 7 benchmark problems summing up to more than 300.000 hours of CPU time. The results revealed that multi-modality and sharp rims in the posterior driven by structural non-identifiabilities, oscillatory dynamics and multi-stability led to insufficient performances. Chaotic regimes in the model led to complete failure of all tested methods and suggest that standard MCMC is not suited for this kind of problems. Correlation structures were only challenging for the algorithms if very pronounced and non-linear, e.g. depending on how pronounced a banana shaped posterior mode was chosen, the performance of the algorithms decreased. On absolute scale, the benchmark revealed overall high auto-correlations in the MCMC samples which in some cases could be significantly increased by using multi-chain samplers and multi-start local optimization for chain initialization. Overall, these unique results suggest (i)

the importance of careful design of models and (ii) highlight the danger of generating non-representative samples in even fairly low dimensional problems while mechanistic models in computational biology tend to be much more high dimensional. Furthermore, (iii) the results highlight the benefits of a robust analysis framework and the need for broad quantitative benchmarking of MCMC methods.

Based on the results of the benchmark study in Chapter 4, in Chapter 5 a novel MCMC method, RAmPART, was developed. RAmPART combines multiple advantages of existing methods into one tool while maintaining strong self-tuning capabilities to enable the frictionless application to diverse parameter estimation problems. In Chapter 5 a proof for ergodicity is provided. Furthermore, novel heuristics for auto-tuning, e.g. maximum temperature selection which is a common problem for tempering algorithms are proposed. RAmPART outperformed similar algorithms, which were used as ingredients in its design, by a factor 2 to 100 and has little computational overhead compared to standard PT.

In conclusion, this thesis contributed to the analysis, comparison and improvement of MCMC sampling for ODE constraint parameter estimation problems in computational biology and beyond. The proposed analysis pipeline provides a framework for highly robust sample quality assessment. The MCMC benchmark study of multiple state-of-the-art methods in diverse ODE-based parameter estimation problems is the first of its kind and will support future decision making in applications and development of novel MCMC methods. The proposed MCMC method, RAmPART, resolves some of the issues of standard methods while being able to adapt to a vast variety of parameter estimation problems.

6.2 Outlook

In the following, potential research direction arising from the results in this thesis are shortly discussed.

6.2.1 Further improvements of sample analysis

The presented sampling analysis pipeline exploits several statistical tests. Those tests may get computationally expensive with increasing sample sizes or parameter dimensions. Unfortunately, limiting the maximum number of parameters constraints the problems which the pipeline could be applied to while limiting the sample sizes is typically done by thinning the sample, which corresponds to an increase of the variance of a-posterior-

estimates (Geyer, 1992) and is expected to lower the robustness of statistical tests. Both bottlenecks could be resolved by proper parallelization within the statistical tests. All elements of the pipeline admit parallelization, e.g. the calculation of eigenvalues λ in the Gelman-Rubin-Brooks test or each interval calculation in the sequential Geweke test. However, in comparison to such intra-test parallelization, inter-test parallelization would be more straight-forward to implement. For example, the pair-wise similarity assessment step of the pipeline could be easily distributed. Another potential improvement of the analysis pipeline could be the integration of additional filter steps in order to further increase the strictness of the pipeline. While this could decrease the rate of sampling results classified as well performing, this would also require larger numbers of MCMC results and thus would increase the overall computational demand in to run the pipeline. Instead, improving existing tests (e.g. (Mengersen et al., 1999)), could make the pipeline steps more sensible and could decrease the number of required MCMC results.

Currently, the pipeline reduces the subjective judgment by shifting the usual visual inspection of standard analysis methods from a single MCMC sample to a visual inspection of similar groups of sampling results. It would be convenient to entirely eliminate the need of visual inspection from the framework. One way of approaching could be the usage of finite state machines (Middleton, 1996), which would make decisions based on certain characteristic values. For example, one could construct a scheme where each mode in a sample is mapped to a center position point and a weight representing its volume estimate. While the ground truth is unknown, these characteristic values could be used for comparisons between runs, as a human would judge similarity of modes, e.g. in a bivariate parameter sample plot. The same could be done for each group of similar chains as obtained by the pipeline. Hereby, a finite state machine could provide judgments based on the value pairs for each group. For example, group A has two modes with certain weights but group B sampled only one of these two modes. In this case, the finite state machine should judge group B as certainly not converged as it misses important parts of the posterior density. Adding more sophisticated characteristic values, e.g. curvature of the mode tails via local derivatives or posterior values could make the decisions more human like. However, in general this approach leads to an unsupervised machine learning problem which is not straight-forward to tackle. In particular, deciding at which point two characteristic values differ significantly so that the algorithm distinguishes two modes is highly problem dependent in general.

A complimentary approach to improve the sampling analysis could be an improvement of the similarity grouping of MCMC results, as the pipeline heavily relies on it. Currently, the number of available MCMC results impacts the overall robustness of the framework.

Larger numbers are expected to increase the overall robustness of the pipeline decision but the chance of false positive testing results are expected to increase as well. Thus, adding a careful multiple testing appraisal to the similarity pipeline elements, e.g. by using Bonferroni-Holm adaptation (Holm, 1979), could be valuable. In order to quantify the increase in robustness with increasing number of MCMC results and to further support objectivity of the framework, one could derive p-values for the overall pipeline decisions. This, however, is not straight-forward, as the pipeline consists of multiple tests and heuristics, whose inputs depend on the previous pipeline steps.

The presented benchmark collection, while covering many common model properties, may be used as a basis for future more vast collections of problems, e.g. for large benchmarks similar to optimization, e.g. (Villaverde et al., 2015). Recently, first steps were taken in this direction (Hass et al., 2018). In order to critically evaluate the performance of MCMC methods under realistic conditions, one has to reevaluate existing methods similar to Chapter 4 in a broader range of problems. Due to “no-free-lunch” limitations (Wolpert and Macready, 1997) each class of problem may require a different class of algorithm to be solved efficiently. However, prominent sampling algorithms such as HMC (Neal, 2011) have not been tested in quantitative benchmarks, yet. Thus, it is not clear where they provide peak performances or may under-perform in comparison to other sampling methods. Furthermore, novel methods claim to achieve a better performance than existing algorithms are worth including in benchmark studies as well. In order to draw a comprehensive picture of the method landscape, benchmarks as the one presented in Chapter 4 have to become standard tools for MCMC, similar to optimization routines. For that, it is important to provide implementations of methods and standard analysis frameworks easily assessable and maintained by the scientific community.

6.2.2 Parameter estimation and mathematical modeling

The results in Chapter 4 revealed, that certain model properties will impact the performance of MCMC methods. In particular, chaotic parameter regimes were found to be highly challenging, as they typically imply a posterior landscape consisting of an extremely high number of sharp peaks. Unfortunately, there are some problems in nature, where model properties like chaos are required in order to describe a process mechanistically (Braxenthaler et al., 1997; Oladyshkin et al., 2011; Glass et al., 1983). However, models often possess oscillatory or chaotic behavior without deliberately decided mechanistic design choice, e.g. by introducing oscillatory behavior via feedback loops in a chemical reaction network (Feinberg and Horn, 1977). In all these cases, approximate methods as ABC (Toni and Stumpf, 2010; Buzbas and Rosenberg, 2013; Jennings and Madigan,

2017) similar to the sampling in stochastic differential equation settings may be required. Furthermore, ideas from multi resolution samplers (Latz et al., 2018) or early rejection schemes (Solonen et al., 2012) could be beneficial for parameter estimation in ODE models with chaotic regimes, as they allow for a sequential evaluation of the ODE integration while the difference of slightly differently parametrized chaotic orbits increases over time (and is typically small for small integration times). Generally, depending on the Lyapunov exponents (Wolf et al., 1985) and simulation times, the posterior gets more “rough” possessing more local modes. Thus, in order to obtain a posterior with low “resolution”, only chronologically first few data points and the corresponding ODE solution would be compared within the likelihood while for the highest resolution all data points would be used. Such a scheme could be complemented by the application of SMC type algorithms. However, instead of applying a suited algorithm, either parameter constraints or modifications of the model may help to prevent running into problematic parameter regimes. Unfortunately, such model properties (e.g. a feedback causing oscillatory dynamics) are not straight-forward to detect and may occur even in simple models possessing feedback loops. However, in contrast to point estimates of the posterior generated by optimization methods, posterior samples facilitate the identification of posterior landscapes which are driven by chaotic behavior. A similar usage of posterior samples was already successfully employed to identify non-identifiabilities (Hug et al., 2013).

Another important and interesting link between MCMC sampling and mathematical modeling is model selection, in particular Bayesian model selection (Schmidl, 2012; Hug et al., 2016b,a). There exists a variety of different approaches for model selections, e.g. via information theoretical approaches (Chehreghani et al., 2012). However, these methods do not consider the uncertainty of parameters as Bayes factors do (Gelman and Meng, 1998; Hug, 2015). For the calculation of Bayes factors, the marginals of two potentially high dimensional models are required. This is a challenging problem in general because the calculation of the marginals requires the evaluation of high-dimensional integrals. One state-of-the-art sampling based approach is thermodynamic integration (Calderhead, 2007; Gelman and Meng, 1998; Lartillot and Philippe, 2006; Friel and Pettitt, 2008) which is closely connected to MCMC methods as presented in this thesis and has been frequently used in computational biology problems (e.g. (Calderhead, 2007; Eydgahi et al., 2013)). Providing an improvement regarding the calculation of Bayes factors would greatly support the model selection process. For example, it would be interesting to see if ideas of novel methods as RAMPART could be used to improve efficiency of thermodynamical integration by replacing standard sampling methods. Following this paradigm, executing benchmark studies, similar to the one in this thesis, for the calculation of Bayes factors would be very interesting.

6.2.3 Further hybrid MCMC methods

In high dimensional problems, random walk type MCMC algorithms can be inefficient if there is a pronounced non-linear correlation structure between parameters with medium to high co-dimension. In these cases, derivative based methods as HMC are expected to be valuable. However, the requirement of derivatives of the posterior density often make them computationally demanding. Thus, it would be interesting to study how HMC and RAmPART scale with increasing problem complexity (e.g. by using benchmark problems possessing correlation structures of increasing dimensions, non-linearity and shaping) and eventually combine them as another hybrid method. Such a hybrid method would use HMC movements in a certain fraction of iterations in order to improve mixing in presence of pronounced parameter correlations. This would be particularly useful for the warm-up phase of RAmPART, as the training of regions depends on it. Once that proper regions suiting the local correlation structures were found, RAmPART is expected to perform well – even in high dimensions – without the need of derivatives. Another potentially valuable approach could be the inclusion of early rejections (Solonen et al., 2012) or resolution samplers (Latz et al., 2018) into the RAmPART framework in order to deal with computationally expensive model evaluations.

6.2.4 Model based data integration

The process of a model formulation enforces a precise and executable formulation of a biological hypotheses (Stelling, 2004) and allows to directly falsify them (Hug et al., 2016a; Hross, 2016; Schilling et al., 2009; Toni et al., 2012). When combined with the flexibility of parameter estimation, mechanistic models allow to integrate data from multiple sources (Cotten and Reed, 2013; Link et al., 2014; Yizhak et al., 2010; Kuepfer et al., 2012). Multi-omic data integration allows to include relationships between geno-, pheno- and environmental-types facilitating a full picture across complex biological systems. By using multi-omic approaches one hopes to reveal novel insights, e.g. detecting relevant biomarkers in a human population expressing a certain disease, with much smaller sample sizes compared to non-multi-omic approaches.

Today, despite the decreasing costs of experiments (Palsson and Zengler, 2010) and the high potential of comprehensive data sets, e.g. (1000Genomes Project Consortium and others, 2010; Rozenblatt-Rosen et al., 2017), for human medicine technology the possibility of model based multi-omic integration is still under-used for a couple of reasons. One important reason is that mechanistic models often require large computational resources. The more different data sources one wants to incorporate, the more complex the

corresponding models typically get, e.g. by including different time scales or additional biological hypotheses – resulting in slow models at the brink of feasibility. As the connections between different omics can not be easily disentangled prior analysis, multi-omic models require a comprehensive mapping of the whole biological system and thus tend to become genome-scale models. In order to expand the computational feasibility of comprehensive biological models, one needs highly efficient, swift and scalable algorithms, e.g. for numeric ODE solvers (Fröhlich, 2018) and parameter estimation methods as presented in this thesis, complemented by proper analysis methods for genome-scale models, e.g. (Terzer et al., 2009). For example, finding smart ways to parallelize model evaluations and parameter estimation methods as RAmPART (Chapter 5) would likely push the usage of model based multi-omic integration.

Another reason for under-usage of model based multi-omics integration is the challenge of proper data preparation (Palsson and Zengler, 2010). Ensuring homogeneous assumptions and normalizations across different sets of data is complicated. Often different labs use slightly different techniques to obtain, store and process data, even for the same omic type. Unfortunately, the details regarding these steps are often treated rather implicit. Even for data which is accumulated in large public databases, the assumptions and normalizations are often not revealed explicitly. This can eventually lead to inconsistencies in the data and training models with such data may provide misleading implications or interpretations when considering model parameter distributions or output predictions. Fortunately, by investigating uncertainty and consistency across different sets of data, e.g. based on sampling techniques as presented in this thesis, one can minimize the risk of false conclusions.

While there is a trend towards the usage of multi-omics data in computational biology, an often neglected approach is to integrate homogeneous-omic data from multiple studies. Mathematical modeling in combination with parameter estimation offers a powerful tool to integrate data, even for different experimental conditions across multiple studies. In particular, by employing Bayesian statistics and MCMC the posterior parameter distributions of study A could be used as prior knowledge for study B, effectively integrating the information from both studies. Unfortunately, studies often treat normalization or experimental nuances differently, causing inhomogeneity of data a challenge. Therefore, carefully selected data and robust sampling methods, as RAmPART, could be used for a proof of concept in this direction enabling a new paradigm of learning from multiple sources of data.

Bibliography

- 1000 Genomes Project Consortium and others. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061, 2010.
- J. H. Albert. Computational methods using a Bayesian hierarchical generalized linear model. *Journal of the American Statistical Association*, 83(404):1037–1044, 1988.
- R. Alexander. Diagonally implicit Runge–Kutta methods for stiff O.D.E.’s. *SIAM J. Numer. Anal.*, 14(6):1006–1021, 1977. doi: 10.1137/0714068.
- U. Alon. An introduction to systems biology: design principles of biological circuits, 2006.
- C. Andrieu, A. Doucet, and R. Holenstein. Particle Markov chain Monte Carlo methods. *J. R. Statist. Soc. B*, 72(3):269–342, June 2010.
- C. Andrieu and J. Thoms. A tutorial on adaptive MCMC. *Stat. Comp.*, 18(4):343–373, 2008.
- C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan. An introduction to MCMC for machine learning. *Machine learning*, 50(1-2):5–43, 2003.
- J. F. Apgar, D. K. Witmer, F. M. White, and B. Tidor. Sloppy models, parameter uncertainty, and the role of experimental design. *Mol. BioSyst.*, 6(10):1890–1900, 2010. doi: 10.1039/B918098B. URL <http://pubs.rsc.org/en/Content/ArticleLanding/2010/MB/b918098b>.
- J. Bachmann, A. Raue, M. Schilling, M. E. Böhm, C. Kreutz, D. Kaschek, H. Busch, N. Gretz, W. D. Lehmann, J. Timmer, and U. Klingmüller. Division of labor by dual feedback regulators controls JAK2/STAT5 signaling over broad ligand range. *Mol. Syst. Biol.*, 7(1):516, July 2011.
- Y. Bai, R. Craiu, and A. Di Narzo. A mixture-based approach to regional adaptation for MCMC. *J. Comput. Graph. Statist.*, 2010.
- B. Ballnus, S. Hug, K. Hatz, L. Görlitz, J. Hasenauer, and F. J. Theis. Comprehensive benchmarking of Markov chain Monte Carlo methods for dynamical systems. *BMC Syst. Biol.*, 11(63), June 2017. doi: 10.1186/s12918-017-0433-1.
- B. Ballnus, S. Schaper, F. J. Theis, and J. Hasenauer. Bayesian parameter estimation for biochemical reaction networks using region-based adaptive parallel tempering. *Bioinformatics*, 34(13):i494–i501, 2018.
- E. Balsa-Canto, A. A. Alonso, and J. R. Banga. An iterative identification procedure for dynamic modeling of biochemical networks. *BMC Syst. Biol.*, 4(11), Feb. 2010. doi: 10.1186/1752-0509-4-11. URL <http://www.biomedcentral.com/1752-0509/4/11>.
- M. J. Bayarri and J. O. Berger. The interplay of bayesian and frequentist analysis. *Statist. Sci.*, 19(1):58–80, 02 2004. doi: 10.1214/088342304000000116. URL <https://doi.org/10.1214/088342304000000116>.

- M. Bédard. Hierarchical models: Local proposal variances for RWM-within-Gibbs and MALA-within-Gibbs. *Computational Statistics & Data Analysis*, 109:231–246, 2017.
- C. H. Beentjes and R. E. Baker. Quasi-Monte Carlo methods applied to tau-leaping in stochastic biological systems. *Bulletin of mathematical biology*, 81(8):2931–2959, 2019.
- J. E. Bennett, A. Racine-Poon, and J. C. Wakefield. *MCMC for nonlinear hierarchical models*. London, UK: Chapman and Hall, 1996.
- A. Beskos, N. Pillai, G. Roberts, J.-M. Sanz-Serna, A. Stuart, et al. Optimal tuning of the hybrid monte carlo algorithm. *Bernoulli*, 19(5A):1501–1534, 2013.
- M. Betancourt, S. Byrne, and M. Girolami. Optimizing the integrator step size for hamiltonian monte carlo. *arXiv preprint arXiv:1411.6669*, 2014.
- W. Betz, I. Papaioannou, and D. Straub. Transitional Markov Chain Monte Carlo: Observations and Improvements. *Journal of Engineering Mechanics*, 142(5):04016016, 2016.
- U. S. Bhalla and R. Iyengar. Emergent properties of networks of biological signaling pathways. *Science*, 283(5400):381–387, 1999.
- R. Boiger, J. Hasenauer, S. Hross, and B. Kaltenbacher. Integration based profile likelihood calculation for PDE constrained parameter estimation problems. *Inverse Prob.*, 32(12):125009, Dec. 2016. doi: 10.1088/0266-5611/32/12/125009.
- G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.
- M. Braxenthaler, R. Unger, D. Auerbach, J. A. Given, and J. Moul. Chaos in protein dynamics. *Proteins Structure Function and Genetics*, 29(4):417–425, 1997.
- P. J. Brockwell and R. A. Davis. *Introduction to time series and forecasting*. springer, 2016.
- S. P. Brooks and G. O. Roberts. Assessing convergence of Markov chain Monte Carlo algorithms. *Stat. Comp.*, 8(4):319–335, 1998.
- S. P. Brooks and A. Gelman. General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, 7(4):434–455, 1998.
- S. Brooks, A. Gelman, G. Jones, and X.-L. Meng. *Handbook of markov chain monte carlo*. CRC press, 2011.
- A. Buchholz and N. Chopin. Improving approximate bayesian computation via quasi-monte carlo. *Journal of Computational and Graphical Statistics*, 28(1):205–219, 2019.
- J. C. Butcher. Implicit Runge-Kutta processes. *Math. Comp.*, 18(85):50–64, 1964. ISSN 00255718, 10886842. URL <http://www.jstor.org/stable/2003405>.
- E. O. Buzbas and N. A. Rosenberg. AABC: approximate approximate Bayesian computation when simulating a large number of data sets is computationally infeasible.

- arXiv:1301.6282v1 [stat.CO]*, Jan. 2013.
- B. Calderhead. A study of population MCMC for estimating Bayes factors over nonlinear ODE models. Master thesis, University of Glasgow, 2007.
- B. Calderhead. *Differential geometric MCMC methods and applications*. PhD thesis, University of Glasgow, 2011.
- B. Calderhead. A general construction for parallelizing Metropolis- Hastings algorithms. *Proceedings of the National Academy of Sciences*, 111(49):17408–17413, 2014.
- B. Calderhead and M. Girolami. Statistical analysis of nonlinear dynamical systems using differential geometric sampling methods. *Interface focus*, 1(6):821–835, 2011.
- B. P. Carlin, A. E. Gelfand, and A. F. Smith. Hierarchical Bayesian analysis of changepoint problems. *Applied statistics*, pages 389–405, 1992.
- G. Casella and E. I. George. Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.
- R. J. Casey. Periodic orbits in neural models: Sensitivity analysis and algorithms for parameter estimation. 2004.
- L. Chaari, H. Batatia, N. Dobigeon, and J.-Y. Tournet. A hierarchical sparsity-smoothness Bayesian model for $l_0 + l_1 + l_2$ regularization. In *IEEE International Conference on Acoustics, Speech, and Signal Processing-ICASSP 2014*, pages pp–1901, 2014.
- C. Chatfield. *The analysis of time series: an introduction*. CRC press, 2016.
- M. Chaves, T. Eissing, and F. Allgöwer. Bistable biological systems: A characterization through local compact input-to-state stability. *IEEE Trans. Autom. Control*, 53:87–100, Jan. 2008.
- M. H. Chehreghani, A. G. Busetto, and J. M. Buhmann. Information theoretic model validation for spectral clustering. In *Proceedings of 15th International Conference on Artificial Intelligence and Statistics (AISTATS), La Palma, Canary Islands*, pages 495–503, 2012.
- M.-H. Chen and Q.-M. Shao. Monte Carlo estimation of Bayesian credible and HPD intervals. *J. Comput. Graphical Statist.*, 8(1):69–92, Mar. 1999.
- Y. Chen, R. Dwivedi, M. J. Wainwright, and B. Yu. Fast MCMC sampling algorithms on polytopes. *ArXiv e-prints*, October 2017.
- O.-T. Chis, J. R. Banga, and E. Balsa-Canto. Structural identifiability of systems biology models: A critical comparison of methods. *PLoS ONE*, 6(11):e27755, Nov. 2011. doi: 10.1371/journal.pone.0027755.
- N. Chopin, S. S. Singh, et al. On particle gibbs sampling. *Bernoulli*, 21(3):1855–1883, 2015.

- G. Claeskens, N. L. Hjort, et al. Model selection and model averaging. *Cambridge Books*, 2008.
- B. L. Clarke. Stability of complex reaction networks. *Advances in chemical physics*, pages 1–215, 1980.
- E. A. Coddington and N. Levinson. *Theory of ordinary differential equations*. Tata McGraw-Hill Education, 1955.
- C. Cotten and J. L. Reed. Mechanistic analysis of multi-omics datasets to generate kinetic parameters for constraint-based metabolic models. *BMC bioinformatics*, 14(1):32, 2013.
- M. W. Covert, N. Xiao, T. J. Chen, and J. R. Karr. Integrating metabolic, transcriptional regulatory and signal transduction models in escherichia coli. *Bioinformatics*, 24(18):2044–2050, 2008.
- M. K. Cowles and B. P. Carlin. Markov chain monte carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, 91(434):883–904, 1996a.
- M. K. Cowles and B. P. Carlin. Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, 91(434):883–904, 1996b.
- R. V. Craiu, J. Rosenthal, and C. Yang. Learn from thy neighbor: Parallel-chain and regional adaptive mcmc. *J. Am. Stat. Assoc.*, 104(488):1454–1466, 2009.
- R. V. Craiu, L. Gray, K. Łatuszyński, N. Madras, G. O. Roberts, J. S. Rosenthal, et al. Stability of adversarial markov chains, with an application to adaptive mcmc algorithms. *The Annals of Applied Probability*, 25(6):3592–3623, 2015.
- J. D. Crawford. Introduction to bifurcation theory. *Reviews of Modern Physics*, 63(4):991, 1991.
- K.-D. Dang, M. Quiroz, R. Kohn, M.-N. Tran, and M. Villani. Hamiltonian monte carlo with energy conserving subsampling. *arXiv preprint arXiv:1708.00955*, 2017.
- F. Dercole and S. Rinaldi. Dynamical systems and their bifurcations. *Advanced Methods of Biomedical Signal Processing, IEEE-Wiley Press, New York, USA*, pages 291–325, 2011.
- T. J. DiCiccio and B. Efron. Bootstrap confidence intervals. *Statist. Sci.*, 11(3):189–228, 1996. doi: 10.1214/ss/1032280214.
- P. M. Djuric and J.-H. Chun. An mcmc sampling approach to estimation of nonstationary hidden markov models. *IEEE Transactions on Signal Processing*, 50(5):1113–1123, 2002.
- A. Doucet, S. Godsill, and C. Andrieu. On sequential Monte Carlo sampling methods for Bayesian filtering. *Stat. Comp.*, 10(3):197–208, July 2000. doi: 10.1023/A:1008935410038.

- A. Doucet, N. de Freitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice*. Statistics for Engineering and Information Science. Springer-Verlag New York, 1 edition, 2001. ISBN 978-1-4419-2887-0,978-1-4757-3437-9. URL <http://gen.lib.rus.ec/book/index.php?md5=c41b6bc5edb09235541114197e97534c>.
- H. Du and L. A. Smith. Rising Above Chaotic Likelihoods. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):246–258, 2017.
- J. A. Egea, D. Henriques, T. Cokelaer, A. F. Villaverde, A. MacNamara, D. P. Danciu, J. R. Banga, and J. Saez-Rodriguez. MEIGO: An open-source software suite based on metaheuristics for global optimization in systems biology and bioinformatics. *BMC Bioinf.*, 15(136), 2014. doi: 10.1186/1471-2105-15-136.
- M. C. Eisenberg and M. A. L. Hayashi. A computational approach to determining structurally identifiable parameter combinations using subset profiling. Technical Report arXiv:1307.2298v1 [q-bio.QM], arXiv, 2013.
- S. Engblom. Computing the moments of high dimensional solutions of the master equation. *Appl. Math. Comp.*, 180:498–515, 2006. doi: 10.1016/j.amc.2005.12.032.
- H. Eydgahi, W. W. Chen, J. L. Muhlich, D. Vitkup, J. N. Tsitsiklis, and P. K. Sorger. Properties of cell death models calibrated and compared using bayesian approaches. *Molecular systems biology*, 9(1):644, 2013.
- M. Feinberg and F. J. Horn. Chemical mechanism structure and the coincidence of the stoichiometric and kinetic subspaces. *Archive for Rational Mechanics and Analysis*, 66(1):83–97, 1977.
- J. d. Freitas, M. Niranjana, A. H. Gee, and A. Doucet. Sequential Monte Carlo methods to train neural network models. *Neural computation*, 12(4):955–993, 2000.
- B. Friedlander and K. Sharman. Performance evaluation of the modified yule-walker estimator. *IEEE transactions on acoustics, speech, and signal processing*, 33(3):719–725, 1985.
- N. Friel and A. N. Pettitt. Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(3):589–607, 2008.
- F. Fröhlich, S. Hross, F. J. Theis, and J. Hasenauer. Radial basis function approximation of Bayesian parameter posterior densities for uncertainty analysis. In P. Mendes, J. O. Dada, and K. O. Smallbone, editors, *Proc. 12th Int. Conf. Comp. Meth. Syst. Biol.*, Lecture Notes in Bioinformatics, pages 73–85. Springer International Publishing Switzerland, Nov. 2014a.
- F. Fröhlich, F. J. Theis, and J. Hasenauer. Uncertainty analysis for non-identifiable dynamical systems: Profile likelihoods, bootstrapping and more. In P. Mendes, J. O. Dada, and K. O. Smallbone, editors, *Proc. 12th Int. Conf. Comp. Meth. Syst. Biol.*, Lecture

- Notes in Bioinformatics, pages 61–72. Springer International Publishing Switzerland, Nov. 2014b.
- F. Fröhlich, B. Kaltenbacher, F. J. Theis, and J. Hasenauer. Scalable parameter estimation for genome-scale biochemical reaction networks. *PLoS Comput. Biol.*, 13(1):e1005331, Jan. 2017a. doi: 10.1371/journal.pcbi.1005331.
- F. Fröhlich, F. J. Theis, J. O. Rädler, and J. Hasenauer. Parameter estimation for dynamical systems with discrete events and logical operations. *Bioinformatics*, 33(7): 1049–1056, Apr. 2017b. doi: 10.1093/bioinformatics/btw764.
- F. Fröhlich. *Scalable Simulation and Optimization Methods for Differential Equation Models describing Biochemical Reaction Networks*. PhD thesis, 2018.
- F. Fröhlich, D. Weindl, P. Stapor, and J. Hasenauer. Icb-dcm/amici: Amici 0.4.0 (version v0.4.0). Zenodo, May 2017c. <http://doi.org/10.5281/zenodo.579891>.
- A. Funahashi, Y. Matsuoka, A. Jouraku, M. Morohashi, N. Kikuchi, and H. Kitano. CellDesigner 3.5: A versatile modeling tool for biochemical networks. *Proceedings of the IEEE*, 96(8):1254–1265, Aug. 2008. doi: 10.1109/JPROC.2008.925458. URL <http://dx.doi.org/10.1109/JPROC.2008.925458>.
- A. Gábor and J. R. Banga. Robust and efficient parameter estimation in dynamic models of biological systems. *BMC Syst Biol*, 9:74, 2015. doi: 10.1186/s12918-015-0219-2.
- A. Gandhi, S. Levin, and S. Orszag. Moment expansions in spatial ecological models and moment closure through gaussian approximation. *Bull. Math. Biol.*, 62(4):595–632, July 2000.
- T. Gardner, C. Cantor, and J. Collins. Construction of a genetic toggle switch in escherichia coli. *Nature*, 403(6767):242–339, Jan. 2000.
- A. Gelman and X.-L. Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical science*, pages 163–185, 1998.
- S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741, 1984.
- I. Gesteira Costa Filho. *Mixture models for the analysis of gene expression*. Phd. thesis, Freie Universität Berlin, Berlin, Germany, May 2008.
- J. Geweke. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In J. M. Bernardo, A. F. M. Smith, A. P. Dawid, and J. O. Berger, editors, *Bayesian Statistics*, pages 169–193. University Press, Oxford, UK, 1992.
- C. J. Geyer. Practical markov chain monte carlo. *Statistical science*, pages 473–483, 1992.
- D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, 81(25):2340–2361, Dec. 1977. doi: 10.1021/j100540a008.

- D. T. Gillespie. A rigorous derivation of the chemical master equation. *Physica A*, 188(1):404–425, Sept 1992. doi: 10.1016/0378-4371(92)90283-V.
- D. T. Gillespie. The chemical Langevin equation. *J. Chem. Phys.*, 113(1):297–306, July 2000.
- M. Girolami and B. Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. R. Statist. Soc. B*, 73(2):123–214, Mar. 2011. doi: 10.1111/j.1467-9868.2010.00765.x.
- L. Glass, M. R. Guevara, A. Shrier, and R. Perez. Bifurcation and chaos in a periodically stimulated cardiac oscillator. *Physica D: Nonlinear Phenomena*, 7(1):89–101, 1983.
- P. Gonnet, S. Dimopoulos, L. Widmer, and J. Stelling. A specialized ODE integrator for the efficient computation of parameter sensitivities. *BMC Syst. Biol.*, 6(46), May 2012. doi: 10.1186/1752-0509-6-46.
- M. M. Graham and A. J. Storkey. Continuously tempered hamiltonian monte carlo. *arXiv preprint arXiv:1704.03338*, 2017.
- P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995. doi: 10.1093/biomet/82.4.711.
- P. J. Green, K. Latuszyński, M. Pereyra, and C. P. Robert. Bayesian computation: a summary of the current state, and samples backwards and forwards. *Statistics and Computing*, 25(4):835–862, 2015.
- R. Grima. A study of the accuracy of moment-closure approximations for stochastic chemical kinetics. *J. Chem. Phys.*, 136(15):154105, Apr. 2012. doi: 10.1063/1.3702848.
- Y. Guan and S. M. Krone. Small-world mcmc and convergence to multi-modal distributions: From slow mixing to fast mixing. *The Annals of Applied Probability*, 17(1):284–304, 2007.
- M. R. Guevara. Bifurcations involving fixed points and limit cycles in biological systems, 2003.
- H. Haario, E. Saksman, and J. Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223–242, 2001.
- H. Haario, M. Laine, A. Mira, and E. Saksman. DRAM: Efficient adaptive MCMC. *Stat. Comp.*, 16(4):339–354, 2006. doi: 10.1007/s11222-006-9438-0.
- J. Hasenauer. *Modeling and parameter estimation for heterogeneous cell populations*. PhD thesis, 2013.
- H. Hass, C. Loos, E. R. Alvarez, J. Timmer, J. Hasenauer, and C. Kreutz. Benchmark problems for dynamic modeling of intracellular processes. *bioRxiv*, page 404590, 2018.
- W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applica-

- tions. *Biometrika*, 51(1):97–109, April 1970.
- T. Heldt, E. B. Shim, R. D. Kamm, and R. G. Mark. Computational modeling of cardiovascular response to orthostatic stress. *Journal of Applied Physiology*, 92(3):1239–1254, 2002.
- J. Hespanha. Moment closure for biochemical networks. In *Proc. Int. Symp. on Communications, Control and Signal Processing*, pages 42–147, 2008. doi: 10.1109/ISCCSP.2008.4537208.
- A. L. Hodgkin and A. F. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.*, 117(4):500–544, Aug. 1952.
- M. D. Hoffman and A. Gelman. The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.
- M. Hofmann and C. Czado. Assessing the var of a portfolio using d-vine copula based multivariate garch models. *Preprint*, 2010.
- S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.
- S. Hoops, S. Sahle, R. Gauges, C. Lee, J. Pahle, N. Simus, M. Singhal, L. Xu, P. Mendes, and U. Kummer. COPASI – a COMplex PATHway SIMulator. *Bioinformatics*, 22(24):3067–3074, 2006. doi: 10.1093/bioinformatics/btl485.
- S. Hross. *Parameter estimation and uncertainty quantification for image based systems biology*. Ph.d. thesis, Technische Universität München, Oct. 2016.
- S. Hross and J. Hasenauer. Analysis of CFSE time-series data using division-, age- and label-structured population models. *Bioinformatics*, 32(15):2321–2329, Aug. 2016. doi: 10.1093/bioinformatics/btw131.
- M. Hucka, A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, A. P. Arkin, B. J. Bornstein, D. Bray, A. Cornish-Bowden, A. A. Cuellar, S. Dronov, E. D. Gilles, M. Ginkel, V. Gor, I. I. Goryanin, W. J. Hedley, T. C. Hodgman, J.-H. Hofmeyr, P. J. Hunter, N. S. Juty, J. L. Kasberger, A. Kremling, U. Kummer, N. Le Novère, L. M. Loew, D. Lucio, P. Mendes, E. Minch, E. D. Mjolsness, Y. Nakayama, M. R. Nelson, P. F. Nielsen, T. Sakurada, J. C. Schaff, B. E. Shapiro, T. S. Shimizu, H. D. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, and J. Wang. The systems biology markup language (SBML): A medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531, 2003. doi: 10.1093/bioinformatics/btg015.
- J. P. Huelsenbeck and F. Ronquist. MRBAYES: Bayesian inference of phylogenetic trees.

- Bioinformatics*, 17(8):754–755, 2001.
- S. Hug, A. Raue, J. Hasenauer, J. Bachmann, U. Klingmüller, J. Timmer, and F. J. Theis. High-dimensional Bayesian parameter estimation: Case study for a model of JAK2/STAT5 signaling. *Math. Biosci.*, 246(2):293–304, Nov. 2013. doi: 10.1016/j.mbs.2013.04.002.
- S. Hug, M. Schwarzfischer, J. Hasenauer, C. Marr, and F. J. Theis. An adaptive scheduling scheme for calculating Bayes factors with thermodynamic integration using simpson’s rule. *Stat. Comput.*, 26(3):663–677, May 2016a. doi: 10.1007/s11222-015-9550-0.
- S. Hug, D. Schmidl, W. B. Li, M. B. Greiter, and F. J. Theis. Bayesian model selection methods and their application to biological ode systems. In *Uncertainty in Biology*, pages 243–268. Springer, 2016b.
- S. C. Hug. *From low-dimensional model selection to high-dimensional inference: tailoring Bayesian methods to biological dynamical systems*. PhD thesis, Technische Universität München, 2015.
- P. Jacob, C. P. Robert, and M. H. Smith. Using parallel computation to improve independent Metropolis–Hastings based estimation. *Journal of Computational and Graphical Statistics*, 20(3):616–635, 2011.
- P. E. Jacob, J. O’Leary, and Y. F. Atchadé. Unbiased Markov chain Monte Carlo with couplings. *arXiv preprint arXiv:1708.03625*, 2017.
- N. Jagiella, B. Müller, M. Müller, I. E. Vignon-Clementel, and D. Drasdo. Inferring growth control mechanisms in growing multi-cellular spheroids of NSCLC cells from spatial-temporal image data. *PLoS Comput. Biol.*, 12(2):e1004412, Feb. 2016. doi: 10.1371/journal.pcbi.1004412.
- N. Jagiella, D. Rickert, F. J. Theis, and J. Hasenauer. Parallelization and high-performance computing enables automated statistical inference of multi-scale models. *Cell Systems*, 4(2):194–206, Feb. 2017. doi: 10.1016/j.cels.2016.12.002.
- A. Jasra, C. C. Holmes, and D. A. Stephens. Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, pages 50–67, 2005.
- E. Jennings and M. Madigan. astroabc: an approximate bayesian computation sequential monte carlo sampler for cosmological parameter estimation. *Astron. Comput.*, 19:16–22, April 2017. doi: doi.org/10.1016/j.ascom.2017.01.001.
- G. Jia, G. N. Stephanopoulos, and R. Gunawan. Parameter estimation of kinetic models from metabolic profiles: Two-phase dynamic decoupling method. *Bioinformatics*, 27(4):196–1970, 2011. doi: 10.1093/bioinformatics/btr293.
- I. T. Jolliffe. *Principal component analysis*. Springer Series in Statistics. Springer, 2nd

- edition, 2002.
- G. L. Jones et al. On the markov chain central limit theorem. *Probability surveys*, 1 (299-320):5–1, 2004.
- M. Joshi, A. Seidel-Morgenstern, and A. Kremling. Exploiting the bootstrap method for quantifying parameter confidence intervals in dynamical systems. *Metabolic Eng.*, 8: 447–455, May 2006.
- B. N. Kholodenko. Negative feedback and ultrasensitivity can bring about oscillations in the mitogen-activated protein kinase cascades. *Eur. J. Biochem.*, 267(6):1583–1588, 2000. URL <http://onlinelibrary.wiley.com/doi/10.1046/j.1432-1327.2000.01197.x/pdf>.
- P. D. Kirk, T. Toni, and M. P. Stumpf. Parameter inference for biochemical systems that undergo a Hopf bifurcation. *Biophysical Journal*, 95(2):540–549, 2008.
- H. Kitano. Systems biology: A brief overview. *Science*, 295(5560):1662–1664, Mar. 2002.
- S. Klamt and E. D. Gilles. Minimal cut sets in biochemical reaction networks. *Bioinformatics*, 20(2):226–234, 2004.
- E. Klipp, R. Herwig, A. Kowald, C. Wierling, and H. Lehrach. *Systems biology in practice*. Wiley-VCH, Weinheim, 2005a. ISBN 978-3-527-31078-4.
- E. Klipp, B. Nordlander, R. Krüger, P. Gennemark, and S. Hohmann. Integrative model of the response of yeast to osmotic shock. *Nature biotechnology*, 23(8):975, 2005b.
- R. Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada, 1995.
- I. Kontoyiannis and S. P. Meyn. Geometric Ergodicity and the Spectral Gap of Non-Reversible Markov Chains. *ArXiv e-prints*, June 2009.
- S. Kosuta, S. Hazledine, J. Sun, H. Miwa, R. J. Morris, J. A. Downie, and G. E. Oldroyd. Differential and chaotic calcium signatures in the symbiosis signaling pathway of legumes. *Proceedings of the National Academy of Sciences*, 105(28):9823–9828, 2008.
- A. Kramer and N. Radde. Towards experimental design using a bayesian framework for parameter identification in dynamic intracellular network models. *Procedia Comput. Sci.*, 1(1):1639–1647, May 2010.
- A. Kramer. *Stochastic Methods for Parameter Estimation and Design of Experiments in Systems Biology*. Ph.d. thesis, 2016.
- M. Krauss, S. Schaller, S. Borchers, R. Findeisen, J. Lippert, and L. Kuepfer. Integrating cellular metabolism into a multiscale whole-body model. *PLoS Comput. Biol.*, 8(10): e1002750, Oct. 2012. doi: 10.1371/journal.pcbi.1002750.
- M. Krauss, R. Burghaus, J. Lippert, M. Niemi, P. Neuvonen, A. Schuppert, S. Willmann,

- L. Kuepfer, and L. Görlitz. Using Bayesian-PBPK modeling for assessment of inter-individual variability and subgroup stratification. *In Silico Pharmacology*, 1(6), Apr. 2013. doi: 10.1186/2193-9616-1-6.
- C. Kreutz, A. Raue, D. Kaschek, and J. Timmer. Profile likelihood in systems biology. *FEBS J.*, 280(11):2564–2571, June 2013.
- M. Kronfeld, H. Planatscher, and A. Zell. The EvA2 optimization framework. In C. Blum and R. Battiti, editors, *Proceedings of the 4th International Conference, LION 4, Venice, Italy, January 18-22,*, volume 6073 of *Lecture Notes in Computer Science*, pages 247–250. Springer Berlin Heidelberg, Jan. 2010.
- L. Kuepfer, J. Lippert, and T. Eissing. Multiscale mechanistic modeling in pharmaceutical research and development. In *Advances in Systems Biology*, pages 543–561. Springer, 2012.
- A. N. Kuhn, M. Diken, S. Kreiter, B. Vallazza, Ö. Türeci, and U. Sahin. Determinants of intracellular RNA pharmacokinetics: Implications for RNA-based immunotherapeutics. *RNA biology*, 8(1):35–43, 2011.
- Y. A. Kuznetsov. *Elements of applied bifurcation theory*, 2013.
- M. K. Lacki and B. Miasojedow. State-dependent swap strategies and automatic reduction of number of temperatures in adaptive parallel tempering algorithm. *Stat. Comput.*, 26(5):951–964, June 2015. doi: 10.1007/s11222-015-9579-0.
- S. Lan, J. Streets, and B. Shahbaba. Wormhole Hamiltonian Monte Carlo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 2014, page 1953, 2014.
- B. Larget and D. L. Simon. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Molecular biology and evolution*, 16(6):750–759, 1999.
- N. Lartillot and H. Philippe. Computing bayes factors using thermodynamic integration. *Systematic biology*, 55(2):195–207, 2006.
- J. Latz, I. Papaioannou, and E. Ullmann. Multilevel Sequential-2 Monte Carlo for Bayesian inverse problems. *Journal of Computational Physics*, 368:154 – 178, 2018. ISSN 0021-9991. doi: <https://doi.org/10.1016/j.jcp.2018.04.014>. URL <http://www.sciencedirect.com/science/article/pii/S0021999118302286>.
- A. Lee, C. Yau, M. B. Giles, A. Doucet, and C. C. Holmes. On the utility of graphics cards to perform massively parallel simulation of advanced Monte Carlo methods. *Journal of computational and graphical statistics*, 19(4):769–789, 2010.
- C. J. Lee, C. R. Sugimoto, G. Zhang, and B. Cronin. Bias in peer review. *Journal of the American Society for Information Science and Technology*, 64(1):2–17, 2013.
- C. Leonhardt, G. Schwake, T. R. Stögbauer, S. Rappl, J. T. Kuhr, T. S. Ligon, and J. O. Rädler. Single-cell mRNA transfection studies: Delivery, kinetics and statistics by

- numbers. *Nanomedicine: Nanotechnology, Biology, and Medicine*, 10(4):679–688, May 2014. doi: 10.1016/j.nano.2013.11.008.
- G. Leonov, N. Kuznetsov, and V. Vagaitsev. Localization of hidden Chua’s attractors. *Physics Letters A*, 375(23):2230–2233, 2011.
- N. Le Novère. Quantitative and logic modelling of molecular and gene networks. *Nat. Rev. Genet.*, 16(3):146–58, Mar 2015. doi: 10.1038/nrg3885.
- M. A. Levenstien, Y. Yang, and J. Ott. Statistical significance for hierarchical clustering in genetic association and microarray expression studies. *BMC Bioinf.*, 4(62), Dec. 2003.
- T. S. Ligon, F. Fröhlich, O. T. Chiş, J. R. Banga, E. Balsa-Canto, and J. Hasenauer. Genssi 2.0: multi-experiment structural identifiability analysis of sbml models. *Bioinformatics*, 34(8):1421–1423, 2017.
- G. Lillacci and M. Khammash. The signal within the noise: efficient inference of stochastic gene regulation models using fluorescence histograms and stochastic simulations. *Bioinformatics*, 29(18):2311–2319, July 2013. doi: 10.1093/bioinformatics/btt380.
- E. Lindelöf. Sur l’application de la méthode des approximations successives aux équations différentielles ordinaires du premier ordre. *Comptes rendus hebdomadaires des séances de l’Académie des sciences*, 116(3):454–457, 1894.
- H. Link, D. Christodoulou, and U. Sauer. Advancing metabolic models with kinetic information. *Current opinion in biotechnology*, 29:8–14, 2014.
- W. A. Link and M. J. Eaton. On thinning of chains in mcmc. *Methods in ecology and evolution*, 3(1):112–115, 2012.
- D. Lu, M. Ye, and M. C. Hill. Analysis of regression confidence intervals and bayesian credible intervals for uncertainty quantification. *Water resources research*, 48(9), 2012.
- D. MacKay. *Information theory, inference, and learning algorithms*. Cambridge University Press, 7.2 edition, Mar 2005. URL <http://www.inference.phy.cam.ac.uk/mackay/itila/>.
- M. J. Mahoney. Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive therapy and research*, 1(2):161–175, 1977.
- C. Maier, C. Loos, and J. Hasenauer. Robust parameter estimation for dynamical systems from outlier-corrupted data. *Bioinformatics*, 33(5):718–725, Mar. 2017. doi: 10.1093/bioinformatics/btw703.
- T. Marwala. Bayesian training of neural networks using genetic programming. *Pattern Recognition Letters*, 28(12):1452–1458, 2007.
- H. McAdams and L. Shapiro. Circuit simulation of genetic networks. *Science*, 269(5224):650–656, 1995. ISSN 0036-8075. doi: 10.1126/science.7624793.

- K. L. Mengersen, C. P. Robert, and C. Guihenneuc-Jouyaux. Mcmc convergence diagnostics: a review. *Bayesian statistics*, 6:415–440, 1999.
- N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.
- R. Mettin, U. Parlitz, and W. Lauterborn. Bifurcation structure of the driven Van der Pol oscillator. *International Journal of Bifurcation and Chaos*, 3(06):1529–1555, 1993.
- S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.
- S. P. Meyn, R. L. Tweedie, et al. Computable bounds for geometric convergence rates of markov chains. *The Annals of Applied Probability*, 4(4):981–1011, 1994.
- B. Miasojedow, E. Moulines, and M. Vihola. An adaptive parallel tempering algorithm. *J. Comput. Graph. Stat.*, 22(3):649–664, 2013.
- D. Middleton. *An introduction to statistical communication theory*. IEEE press Piscataway, NJ, 1996.
- C. G. Moles, P. Mendes, and J. R. Banga. Parameter estimation in biochemical pathways: A comparison of global optimization methods. *Genome Res.*, 13:2467–2474, global optimization 2003. doi: 10.1101/gr.1262503.
- J. Monod. The technique of continuous culture. *Ann. Inst. Pasteur*, 79:390–410, 1950.
- B. Munsky and M. Khammash. The finite state projection algorithm for the solution of the chemical master equation. *J. Chem. Phys.*, 124(4):044104, Jan. 2006. doi: 10.1063/1.2145882.
- K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.
- R. M. Neal. *Handbook of Markov Chain Monte Carlo*, chapter MCMC using Hamiltonian dynamics. Chapman & Hall / CRC Press, London, United Kingdom, 2011.
- R. M. Neal. Slice sampling. *Annals of statistics*, pages 705–741, 2003.
- C. N. Ngonghala, M. I. Teboh-Ewungkem, and G. A. Ngwa. Observance of period-doubling bifurcation and chaos in an autonomous ode model for malaria with vector demography. *Theoretical ecology*, 9(3):337–351, 2016.
- S. Oladyshkin, H. Class, R. Helmig, and W. Nowak. An integrative approach to robust design and probabilistic risk assessment for CO2 storage in geological formations. *Comp. Geosciences*, 15(3):565–577, Feb. 2011. doi: 10.1007/s10596-011-9224-8.
- A. N. Olsen. *When Infinity is Too Long to Wait: On the Convergence of Markov Chain Monte Carlo Methods*. PhD thesis, The Ohio State University, 2015.
- A. B. Owen and S. D. Tribble. A quasi-monte carlo metropolis algorithm. *Proceedings of*

- the National Academy of Sciences*, 102(25):8844–8849, 2005.
- E. M. Ozbudak, M. Thattai, H. N. Lim, B. I. Shraiman, and A. van Oudenaarden. Multistability in the lactose utilization network of *Escherichia coli*. *Nature*, 427(6976):737–740, Feb. 2004.
- B. Palsson and K. Zengler. The challenges of integrating multi-omic data sets. *Nature chemical biology*, 6(11):787, 2010.
- P. Papastamoulis and G. Iliopoulos. On the Convergence Rate of Random Permutation Sampler and ECR Algorithm in Missing Data Models. *Methodology and Computing in Applied Probability*, 15(2):293–304, 2013. ISSN 1573-7713.
- U. Parlitz and W. Lauterborn. Period-doubling cascades and devil’s staircases of the driven van der Pol oscillator. *Physical Review A*, 36(3):1428, 1987.
- G. Peano. Démonstration de l’intégrabilité des équations différentielles ordinaires. In *Arbeiten zur Analysis und zur mathematischen Logik*, pages 76–126. Springer, 1890.
- D. B. Percival and A. T. Walden. *Spectral analysis for physical applications*. cambridge university press, 1993.
- D. Poland. Cooperative catalysis and chemical chaos: a chemical model for the Lorenz equations. *Physica D: Nonlinear Phenomena*, 65(1):86–99, 1993.
- A. Polynikis, S. Hogan, and M. di Bernardo. Comparing different ode modelling approaches for gene regulatory networks. *Journal of theoretical biology*, 261(4):511–530, 2009.
- A. Raue. *Quantitative Dynamic Modeling: Theory and Application to Signal Transduction in the Erythropoietic System*. Phd. thesis, Albert-Ludwigs-Universität Freiburg im Breisgau, 2013.
- A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, and J. Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(25):1923–1929, May 2009.
- A. Raue, V. Becker, U. Klingmüller, and J. Timmer. Identifiability and observability analysis for experimental design in nonlinear dynamical models. *Chaos*, 20(045105), Dec. 2010. doi: 10.1063/1.3528102.
- A. Raue, C. Kreutz, T. Maiwald, U. Klingmüller, and J. Timmer. Addressing parameter identifiability by model-based experimentation. *IET Syst. Biol.*, 5(2):120–130, Mar. 2011. doi: 10.1049/iet-syb.2010.0061.
- A. Raue, C. Kreutz, F. J. Theis, and J. Timmer. Joining forces of Bayesian and frequentist methodology: A study for inference in the presence of non-identifiability. *Philos T Roy Soc A*, 371(1984), Jan. 2013a. doi: 10.1098/rsta.2011.0544.
- A. Raue, M. Schilling, J. Bachmann, A. Matteson, M. Schelke, D. Kaschek, S. Hug,

- C. Kreutz, B. D. Harms, F. J. Theis, U. Klingmüller, and J. Timmer. Lessons learned from quantitative dynamical modeling in systems biology. *PLoS ONE*, 8(9):e74335, Sept. 2013b. doi: 10.1371/journal.pone.0074335.
- H. Resat, L. Petzold, and M. F. Pettigrew. Kinetic modeling of biological systems. *Methods Mol. Biol.*, 541:311–335, 2009. doi: 10.1007/978-1-59745-243-4{_}14. URL <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2877599/>.
- D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti. Detecting novel associations in large data sets. *Science*, 334(6062):1518–1524, Dec. 2011.
- S. I. Resnick. *A probability path*. Springer Science & Business Media, 2013.
- F. Rigat and A. Mira. Parallel hierarchical sampling: a general-purpose class of multiple-chains MCMC algorithms. *Comp. Stat. Data Anal.*, 56(6):1450–1467, June 2012. doi: 10.1016/j.csda.2011.11.020.
- H. Risken. *The Fokker-Planck equation: Methods of solution and applications*. Springer, Berlin / Heidelberg, 2nd edition, 1996.
- C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 2004.
- C. Robert and G. Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.
- G. O. Roberts and J. S. Rosenthal. Coupling and ergodicity of adaptive markov chain monte carlo algorithms. *Journal of applied probability*, 44(2):458–475, 2007.
- G. O. Roberts and J. S. Rosenthal. Examples of adaptive MCMC. *J. Comput. Graph. Stat.*, 18(2):349–367, 2009.
- G. O. Roberts and O. Stramer. On inference for partially observed nonlinear diffusion models using the Metropolis–Hastings algorithm. *Biometrika*, 88(3):603–621, 2001.
- G. O. Roberts, O. Papaspiliopoulos, and P. Dellaportas. Bayesian inference for non-Gaussian Ornstein–Uhlenbeck stochastic volatility processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(2):369–393, 2004.
- J. L. Rodgers and W. A. Nicewander. Thirteen ways to look at the correlation coefficient. *Am. Stat.*, 42(1):59–66, Feb. 1988.
- H. H. Rosenbrock. Some general implicit processes for the numerical solution of differential equations. *Comput. J.*, 5(4):329–330, 1963. doi: 10.1093/comjnl/5.4.329. URL [+http://dx.doi.org/10.1093/comjnl/5.4.329](http://dx.doi.org/10.1093/comjnl/5.4.329).
- J. S. Rosenthal. Minorization conditions and convergence rates for markov chain monte carlo. *Journal of the American Statistical Association*, 90(430):558–566, 1995.
- J. S. Rosenthal and J. Yang. Ergodicity of combocontinuous adaptive mcmc algorithms.

- Methodology and Computing in Applied Probability*, pages 1–17, 2017.
- O. Rozenblatt-Rosen, M. J. Stubbington, A. Regev, and S. A. Teichmann. The human cell atlas: from vision to reality. *Nature News*, 550(7677):451, 2017.
- M. D. Sacchi, T. J. Ulrych, and C. J. Walker. Interpolation and extrapolation using a high-resolution discrete Fourier transform. *IEEE Transactions on Signal Processing*, 46(1):31–38, 1998a.
- M. D. Sacchi, T. J. Ulrych, and C. J. Walker. Interpolation and extrapolation using a high-resolution discrete fourier transform. *IEEE Transactions on Signal Processing*, 46(1):31–38, 1998b.
- M. Sambridge. A parallel tempering algorithm for probabilistic sampling and multimodal optimization. *Geophys. J. Int.*, page ggt342, 2013.
- M. Schilling, T. Maiwald, S. Hengl, D. Winter, C. Kreutz, W. Kolch, W. D. Lehmann, J. Timmer, and U. Klingmüller. Theoretical and experimental analysis links isoform-specific ERK signalling to cell fate decisions. *Mol. Syst. Biol.*, 5(334), Dec. 2009.
- D. Schmidl. *Bayesian model inference in dynamic biological systems using Markov Chain Monte Carlo methods*. Ph.d. thesis, Technische Universität München, München, Feb. 2012.
- D. Schmidl, C. Czado, S. Hug, F. J. Theis, et al. A vine-copula based adaptive MCMC sampler for efficient inference of dynamical systems. *Bayesian Analysis*, 8(1):1–22, 2013.
- S. L. Scott, A. W. Blocker, F. V. Bonassi, H. A. Chipman, E. I. George, and R. E. McCulloch. Bayes and big data: The consensus Monte Carlo algorithm. *International Journal of Management Science and Engineering Management*, 11(2):78–88, 2016.
- R. Serban and A. C. Hindmarsh. CVODES: An ODE solver with sensitivity analysis capabilities. *ACM Math. Software*, 31(3):363–396, 2005.
- A. E. Sgro, D. J. Schwab, J. Noorbakhsh, T. Mestler, P. Mehta, and T. Gregor. From intracellular signaling to population oscillations: bridging size- and time-scales in collective behavior. *Molecular Systems Biology*, 11(1):779, 2015.
- A. Singh and J. Hespanha. A derivative matching approach to moment closure for the stochastic logistic model. *Bull. Math. Biol.*, 69(6):1909–1925, Aug. 2007.
- R. Smith. Peer review: a flawed process at the heart of science and journals. *Journal of the royal society of medicine*, 99(4):178–182, 2006.
- A. Sokal. Monte carlo methods in statistical mechanics: foundations and new algorithms. In *Functional integration*, pages 131–192. Springer, 1997.
- A. Solonen, P. Ollinaho, M. Laine, H. Haario, J. Tamminen, H. Järvinen, et al. Efficient MCMC for climate model parameter estimation: Parallel adaptive chains and early rejection. *Bayesian Analysis*, 7(3):715–736, 2012.

- P. Stapor, D. Weindl, B. Ballnus, S. Hug, C. Loos, A. Fiedler, S. Krause, S. Hross, F. Fröhlich, and J. Hasenauer. PESTO: Parameter ESTimation TOolbox. *Bioinformatics*, btx676, 2017. doi: 10.1093/bioinformatics/btx676.
- J. Stelling. Mathematical models in microbial systems biology. *Current opinion in microbiology*, 7(5):513–518, 2004.
- A. H. Stram, P. Marjoram, and G. K. Chen. al3c: high-performance software for parameter inference using approximate bayesian computation. *Bioinformatics*, 31(21):3549–3551, November 2015. doi: 10.1093/bioinformatics/btv393.
- I. Strid. Efficient parallelisation of Metropolis–Hastings algorithms using a prefetching approach. *Computational Statistics & Data Analysis*, 54(11):2814–2835, 2010.
- M. A. Suchard, Q. Wang, C. Chan, J. Frelinger, A. Cron, and M. West. Understanding GPU programming for statistical computation: Studies in massively parallel massive mixtures. *Journal of computational and graphical statistics*, 19(2):419–438, 2010.
- I. Swameye, T. G. Müller, J. Timmer, O. Sandra, and U. Klingmüller. Identification of nucleocytoplasmic cycling as a remote sensor in cellular signaling by databased modeling. *Proc. Natl. Acad. Sci. USA*, 100(3):1028–1033, Feb 2003. URL <http://www.pnas.org/content/100/3/1028.abstract>.
- M. Terzer, N. D. Maynard, M. W. Covert, and J. Stelling. Genome-scale metabolic networks. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 1(3):285–297, 2009.
- D. J. Thomson. Spectrum estimation and harmonic analysis. *Proceedings of the IEEE*, 70(9):1055–1096, 1982.
- T. Toni and M. P. H. Stumpf. Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics*, 26(1):104–110, Oct. 2010.
- T. Toni, Y.-i. Ozaki, P. Kirk, S. Kuroda, and M. P. H. Stumpf. Elucidating the in vivo phosphorylation dynamics of the ERK MAP kinase using quantitative proteomics data and Bayesian model selection. *Mol. Biosyst.*, 8:1921–1929, May 2012. doi: 10.1039/C2MB05493K.
- T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. P. Stumpf. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6(31):187–202, 2009.
- M. Tsatsos. Theoretical and numerical study of the Van der Pol equation. Technical Report [arXiv:0803.1658](https://arxiv.org/abs/0803.1658), arXiv, July 2006.
- A. M. Turing. The chemical basis of morphogenesis. *Phil. Trans. Roy. Soc. Lon. B*, 237(641):37–72, Aug. 1952.
- J. J. Tyson. Modeling the cell division cycle: cdc2 and cyclin interactions. *Proc. Nati.*

- Acad. Sci. USA*, 88:7328–7332, Aug 1991.
- N. G. van Kampen. *Stochastic processes in physics and chemistry*. North-Holland, Amsterdam, 3rd edition, 2007.
- J. Vanlier, C. A. Tiemann, P. A. J. Hilbers, and N. A. W. van Riel. An integrated strategy for prediction uncertainty analysis. *Bioinformatics*, 28(8):1130–1135, 2012. doi: 10.1093/bioinformatics/bts088.
- M. Vihola. Robust adaptive metropolis algorithm with coerced acceptance rate. *Statistics and Computing*, 22(5):997–1008, 2012. ISSN 1573-1375.
- A. F. Villaverde, D. Henriques, K. Smallbone, S. Bongard, J. Schmid, D. Cicin-Sain, A. Crombach, J. Saez-Rodriguez, K. Mauch, E. Balsa-Canto, P. Mendes, J. Jaeger, and J. R. Banga. BioPreDyn-bench: A suite of benchmark problems for dynamic modelling in systems biology. *BMC Syst. Biol.*, 9(8), Feb. 2015. doi: 10.1186/s12918-015-0144-4.
- W. Voudsen, W. M. Farr, and I. Mandel. Dynamic temperature selection for parallel tempering in Markov chain Monte Carlo simulations. *Mon. Not. R. Astron. Soc.*, 455(2):1919–1937, 2016.
- V. Vyshemirsky and M. A. Girolami. Bayesian ranking of biochemical system models. *Bioinformatics*, 24(6):833–839, 2008.
- P. Weber, J. Hasenauer, F. Allgöwer, and N. Radde. Parameter estimation and identifiability of biological networks using relative data. In S. Bittanti, A. Cenedese, and S. Zampieri, editors, *Proc. of the 18th IFAC World Congress*, volume 18, pages 11648–11653, Milano, Italy, Aug. 2011. doi: 10.3182/20110828-6-IT-1002.01007.
- D. Wegmann, C. Leuenberger, and L. Excoffier. Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics*, 2009.
- P. Welch. The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics*, 15(2):70–73, 1967.
- T. Wilhelm. The smallest chemical reaction system with bistability. *BMC Systems Biology*, 3(1):90, 2009.
- D. J. Wilkinson. Bayesian methods in bioinformatics and computational systems biology. *Briefings in Bioinf.*, 8(2):109–116, Mar 2007.
- D. J. Wilkinson. *Stochastic modelling for systems biology*. Chapman and Hall/CRC, 2006.
- A. Wolf, J. B. Swift, H. L. Swinney, and J. A. Vastano. Determining lyapunov exponents from a time series. *Physica D: Nonlinear Phenomena*, 16(3):285–317, 1985.
- D. H. Wolpert and W. G. Macready. No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.*, 1(1):67–82, Apr 1997. ISSN 1089-778X. doi: 10.1109/4235.585893.

- D. B. Woodard. Detecting poor convergence of posterior samplers due to multimodality. In *Discussion Paper 2008–05*. Citeseer, 2007.
- T.-R. Xu, V. Vyshemirsky, A. Gormand, A. von Kriegsheim, M. Girolami, G. S. Baillie, D. Ketley, A. J. Dunlop, G. Milligan, M. D. Houslay, and W. Kolch. Inferring signaling pathway topologies from multiple perturbation measurements of specific biochemical species. *Sci. Signal.*, 3(113):ra20, Mar. 2010. doi: 10.1126/scisignal.2000517.
- T. Yanagita and Y. Iba. Exploration of order in chaos using the replica exchange Monte Carlo method. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(02):P02043, 2009.
- J. Yang, R. V. Craiu, and J. S. Rosenthal. Adaptive component-wise multiple-try metropolis sampling. *arXiv preprint arXiv:1603.03510*, 2016.
- K. Yizhak, T. Benyamini, W. Liebermeister, E. Ruppin, and T. Shlomi. Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model. *Bioinformatics*, 26(12):i255–i260, 2010.
- C. Zechner, M. Unger, S. Pelet, M. Peter, and H. Koepl. Scalable inference of heterogeneous reaction kinetics from pooled single-cell recordings. *Nat. Methods*, 11:197–202, Jan. 2014. doi: 10.1038/nmeth.2794.
- Z. Zhang, B. T. Dai, and A. K. Tung. Estimating local optimums in em algorithm over gaussian mixture model. In *Proceedings of the 25th international conference on Machine learning*, pages 1240–1247. ACM, 2008.
- C. Zimmer, S. Sahle, and J. Pahle. Exploiting intrinsic fluctuations to identify model parameters. *IET Systems Biology*, 9(2):64–73, 2015.