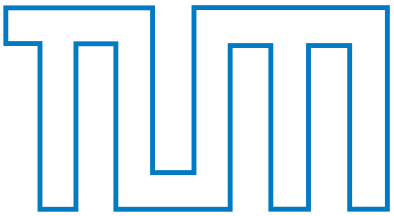
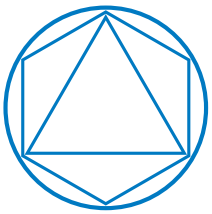




HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

HelmholtzZentrum münchen
German Research Center for Environmental Health

ICB Institute of Computational Biology

Bayesian Inference for Stochastic Differential Equation Models of Intracellular Processes

Susanne Pieschner

2021

TECHNISCHE UNIVERSITÄT MÜNCHEN

Fakultät für Mathematik
Lehrstuhl für Mathematische Modellierung biologischer Systeme

Bayesian Inference for Stochastic Differential Equation Models of Intracellular Processes

Susanne Pieschner

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

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzende: Prof. Dr. Silke Rolles

Prüfer der Dissertation:

1. Prof. Dr. Dr. Fabian J. Theis
2. Prof. Dr. Christiane Fuchs
3. Prof. Dr. Wilfried Grecksch

Die Dissertation wurde am 26.01.2021 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 01.06.2021 angenommen.

Acknowledgments

An dieser Stelle möchte ich mich bei allen bedanken, die in irgendeiner Form zur Entstehung dieser Doktorarbeit beigetragen und mich auf dem Weg zur Promotion unterstützt haben. Besonders bedanken möchte ich mich hierfür bei ...

... meinem Doktorvater Prof. Dr. Dr. Fabian Theis dafür, dass er diese Aufgabe übernommen hat, für hilfreiches Feedback während der Thesis Committee Meetings und für den Aufbau und die Leitung des ICBs, an dem eine so hervorragende Arbeitsatmosphäre herrscht.

... meiner Betreuerin Prof. Dr. Chrisitiane Fuchs dafür, dass sie mein Promotionsvorhaben ermöglicht und mit viel Geduld und Verständnis begleitet hat, für viel Freiraum, die fachlichen Diskussionen und die Zusammenstellung eines sehr netten Kollegenteams.

... Prof. Dr. Dr. h. c. Wilfried Grecksch für die Erstellung des weiteren Gutachtens und für die fachlichen Diskussionen und vorallem die moralische Unterstützung, die maßgeblich zum erfolgreichen Abschluss meines Promotionsvorhabens beigetragen haben.

... Prof. Dr. Jan Hasenauer für interessante Diskussionen und hilfreiches Feedback zu meiner Arbeit an den Daten aus dem mRNA-Transfektionsexperiment.

... Prof. Dr. Micheal Stumpf und seiner Forschungsgruppe an der University of Melbourne, bei denen ich einen wunderbaren Forschungsaufenthalt verbringen durfte.

... meinen Kollegen der Biostatistikgruppe in München und Bielefeld und des gesamten ICBs, insbesondere bei Hannah, Lisa, Caro und Hans, für die Hilfsbereitschaft, die sehr schöne Zeit und viel Spaß im Büro.

... der Helmholtz Graduate School (HELENA) und dem Global Challenges for Women in Math Science Programm der TUM für die finanzielle Unterstützung meiner Teilnahme an mehreren internationalen Konferenzen und des Forschungsaufenthaltes in Melbourne.

... meinen Freunden und meiner Familie, insbesondere bei meiner Mutter Corina, für die bedingungslose Unterstützung, die Fürsorge, die Ablenkung und den Zuspruch während des Studiums und der Promotion.

Abstract

Mathematical modeling is a powerful tool in many areas of science. In systems biology, mechanistic models are particularly useful to gain insights into biological processes as they can immediately disclose causal mechanisms. The parameters of such models, e. g. kinetic rate constants, usually cannot be measured directly but need to be inferred from experimental data. Continuous-time, discrete-space stochastic processes provide an adequate description of the amount of molecular species and their interactions within a cell. However, parameter inference for such processes is usually computationally intractable. Therefore, several approximation models have been developed. One approach that preserves the stochastic nature of the underlying process is the approximation by Itô diffusion processes. These continuous-time, continuous-space stochastic processes described by Itô-type stochastic differential equations (SDEs) are the focus of this thesis. Here, the goal is to enable leveraging the potential of diffusion processes to generate systems biological insights. To this end, we explore computationally efficient inference methods for diffusion processes and consider the application of diffusion processes to a real-world phenomenon in order to study their impact.

Also for diffusion processes, parameter inference is a very challenging problem, in particular because the corresponding likelihood function is usually intractable. Model parameters can be estimated from discretely observed data using e. g. Markov chain Monte Carlo (MCMC) methods that introduce auxiliary data. These methods typically approximate the transition densities of the process numerically based on the Euler-Maruyama scheme and are computationally expensive. Using higher-order approximations may accelerate them, but the specific implementation and benefit remain unclear. Hence, we investigate the utilization and usefulness of higher-order approximations in the example of the Milstein scheme and find that, in fact, the use of the Milstein scheme does improve the estimation accuracy for the parameters appearing in the diffusion coefficient. However, our study also shows that the applicability of the Milstein scheme is very limited in this context in the case of multi-dimensional processes.

Concerning the application to a real-world example, we use diffusion processes to model the translation kinetics after mRNA transfection and infer the model parameters from time-lapse

fluorescence microscopy data using the open source software Stan. We compare this SDE model to a corresponding deterministic ordinary differential equation (ODE) model in terms of parameter identifiability and find that the SDE model provides better identifiability of the kinetic parameters than the ODE model.

Finally, we provide a sound mathematical foundation for the SDE model of the translation kinetics by proving the existence and uniqueness of a strong solution of the SDE and that the Euler-Maruyama approximation of the SDE strongly converges to this solution, although the standard assumptions from stochastic analysis are not fulfilled.

In summary, this thesis addresses multiple important aspects that need to be considered in order to harness the capabilities of mathematical modeling to generate systems biological insights, including mathematical theory, computational efficiency, and consideration of the specific challenges that arise when working with experimental data. Thus, it provides several building blocks to pave the way towards a holistic understanding of biological systems.

Zusammenfassung

Mathematische Modellierung ist ein hilfreiches Werkzeug in vielen Wissenschaftsbereichen. In der Systembiologie sind insbesondere mechanistische Modelle nützlich um Erkenntnisse über biologische Prozesse zu gewinnen, da sie kausale Zusammenhänge unmittelbar erkennen lassen. Die Parameter solcher Modelle können meist nicht direkt gemessen werden und müssen deshalb anhand von experimentellen Daten geschätzt werden. Stochastische Prozesse in stetiger Zeit und mit diskretem Zustandsraum liefern eine adäquate Beschreibung der Anzahl von Molekülen und ihrer Interaktionen innerhalb einer Zelle. Da jedoch die Parameterschätzung für solche Prozesse oft rechnerisch nicht durchführbar ist, wurden verschiedene Approximationsmethoden entwickelt. Eine Vorgehensweise, die die Stochastik des zugrunde liegenden Prozesses beibehält, ist die Approximation durch Itô-Diffusionsprozesse. Diese stochastischen Prozesse in stetiger Zeit mit kontinuierlichem Zustandsraum, die durch stochastische Differentialgleichungen (SDEs) vom Itô-Typ beschrieben werden, stehen im Fokus dieser Dissertation. Ziel ist es, das Potenzial von Diffusionsprozessen für die Erkenntnisgewinnung in der Systembiologie nutzbar zu machen.

Parameterschätzung ist auch für Diffusionsprozesse eine Herausforderung, insbesondere da die zugehörige Likelihood-Funktion meist nicht analytisch zur Verfügung steht. Die Modellparameter können aus diskret beobachteten Daten z. B. mithilfe von Markov-Chain-Monte-Carlo Methoden geschätzt werden, die zusätzliche Datenpunkte einfügen. Diese Methoden approximieren die Übergangsdichten des Prozesses typischerweise numerisch basierend auf dem Euler-Maruyama-Schema und sind rechnerisch sehr aufwändig. Die Verwendung eines Schemas mit höherer Konvergenzordnung birgt das Potenzial, die Methoden zu verbessern, aber die genaue Implementierung und die Vorteile waren unklar. Deshalb untersuchen wir die Verwendung und den Nutzen eines Schemas höherer Ordnung am Beispiel des Milstein-Schemas und stellen fest, dass dieses die Schätzgenauigkeit für Parameter, die im Diffusionskoeffizienten vorkommen, verbessert. Jedoch zeigen unsere Untersuchungen auch, dass die Anwendbarkeit des Milstein-Schemas im Falle von mehrdimensionalen Prozessen sehr eingeschränkt ist.

Außerdem verwenden wir Diffusionsprozesse, um die Translationskinetik nach der Transfektion von mRNA zu modellieren, und schätzen die Modellparameter aus zeitaufgelösten Fluoreszenzmikroskopiedaten mithilfe der Open Source Software Stan. Ein Vergleich mit dem zugehörigen deterministischen Differentialgleichungsmodell zeigt, dass das SDE-Modell zu besserer Identifizierbarkeit der kinetischen Parameter führt.

Schließlich stellen wir das SDE-Modell auf ein sicheres mathematisches Fundament. Wir beweisen, dass eine eindeutige starke Lösung der SDE existiert und dass die Euler-Maruyama-Approximation der SDE stark gegen diese Lösung konvergiert, obwohl die Standardvoraussetzungen aus der stochastischen Analysis nicht erfüllt sind.

Zusammenfassend behandelt diese Dissertation damit mehrere wichtige Aspekte, die es zu beachten gilt, um mittels mathematischer Modellierung neue Erkenntnisse in der Systembiologie gewinnen zu können. Diese umfassen vor allem die mathematische Theorie, rechnerische Effizienz und die Beachtung der konkreten Herausforderungen, die die Arbeit mit experimentellen Daten mit sich bringt. Somit trägt diese Dissertation dazu bei, den Weg hin zu einem ganzheitlichen Verständnis von biologischen Systemen zu ebnen.

Contents

Chapter 1

Introduction

Mathematical modeling is a powerful tool in many areas of science, including natural and social sciences (see e. g. Humphreys, 2003, Müller & Kuttler, 2015, Neimark, 2003). By mathematical modeling, we mean the process of describing a real-world problem or phenomenon by a mathematical model, e. g. by a set of equations, then using mathematical tools to analyze and solve the mathematical problem and finally, interpreting and validating the results to gain a better understanding of the underlying real-world phenomenon. When deciding about the complexity of the employed model, one has to carefully weigh up the advantages of an elaborate model describing the real-world problem in detail and the disadvantages of such an elaborate model in terms of the mathematical tools necessary to analyze it. Therefore, usually only essential aspects are taken into account. An important step in developing and validating a mathematical model is to relate real-world data to the model by means of statistics.

Also in systems biology, mathematical models are widely used to gain insights into biological processes on a variety of different scales including e. g. whole organs or tissues on the macroscale and cell-to-cell interactions but also intracellular processes on the microscale (Kitano, 2002a,b). Mechanistic models are particularly useful in this field because they can immediately disclose causal mechanisms. Comparing the input-output relationship predicted by the model to experimental data allows to verify or falsify the biological hypothesis represented by the model even when some of the involved quantities are not accessible through experiments (Baker et al., 2018). Moreover, they can be employed for in-silico experiments of various experimental conditions that might be too difficult, too expensive, or even impossible to perform in real. One important type of mechanistic models are differential equation models. They can be used to describe the temporal evolution of the abundance of various biological species in a system. On the molecular level, the development of time-lapse fluorescence microscopy has enabled the collection of measurements for the same cells over time (Young et al., 2011). Besides,

experiments have shown that there is a vast amount of variability of outcomes of gene expression not only between different cell populations, but also within isogenic cell populations and even within individual cells which is due to the inherently stochastic nature of the underlying biological processes (Elowitz et al., 2002, Raj & van Oudenaarden, 2008). Using stochastic models, i. e. models that explicitly account for this stochasticity, can help improve our ability to determine model parameters based on experimental data (Munsky et al., 2009).

In general, there are several different sources of variability in data and of discrepancy between data and a considered model. Theses include (i) measurement error, (ii) uncertainty in model specification, (iii) intra-individual variability (also known as intrinsic noise), and (iv) inter-individual variability (also known as extrinsic noise) (see e. g. Kirk et al., 2016, Leander et al., 2015, Regan et al., 2002). Taking stochasticity into account and considering the uncertainty that arises from it is an important and useful aspect of the modeling process.

One adequate description of the amount of molecular species and their interactions within a cell is a continuous-time, discrete-space stochastic process such as a Markov jump process (MJP). However, parameter inference for MJPs is usually computationally intractable. Therefore, several approximation models have been developed. One approach that preserves the stochastic nature of the underlying process is the approximation by Itô diffusion processes. We also write diffusion processes or just diffusions for short. These continuous-time, continuous-space stochastic processes described by Itô-type stochastic differential equations (SDEs) are the focus of this thesis. Our goal is to enable harnessing the potential of diffusion processes to generate systems biological insights.

Parameter inference for diffusion processes is a very challenging problem, in particular because the corresponding likelihood function is usually intractable, and the existing inference methods are computationally expensive. In this thesis, we investigate one potential remedy to this problem, namely the use of a higher-order approximation scheme of the paths of a diffusion process. Further, we apply diffusion processes to model the translation kinetics after mRNA transfection, infer the model parameters from experimental data, and analyze the advantages in terms of parameter identifiability of this stochastic model compared to a deterministic ordinary differential equation (ODE) model. Moreover, if we want to leverage the capabilities of mathematical modeling in generating new insights, we also need to ensure that the underlying mathematical theory is well founded and sound. Therefore, this aspect is another pillar of this thesis and we develop essential theoretical results for the SDE model of the translation kinetics.

1.1 Contributions of this thesis

While the remainder of this thesis is written from the *we*-perspective referring to the author and the reader (and occasionally the supervisor as will become clear below); in this section, I describe the specific contributions of this thesis and what my role was in obtaining them. The contributions are delineated in detail in the two main chapters of this thesis, namely Chapters 4 and 5. Here, I only briefly highlight the main points.

The overall goal of this thesis is to enable leveraging the potential of diffusion processes to generate systems biological insights. To that end, three aims were targeted in the following way:

- **Aim 1:** *Exploring computationally efficient inference methods for diffusion processes*

  Parameter estimation for SDEs is a very challenging problem, especially when the diffusion coefficient depends on the process states. The available methods are computationally very expensive. The transition density of the diffusion process usually needs to be approximated which is commonly done by the Euler-Maruyama scheme.

  I investigated how the Milstein scheme, as an approximation scheme of higher convergence order, can be used in the context of Bayesian data augmentation for diffusions and analyzed whether due to the higher approximation accuracy, fewer imputed data points would be required such that overall for a fixed computational cost a higher estimation accuracy can be achieved. I described for what kind of SDEs the Milstein scheme can be applied in this context and developed an alternative (arguably more straight forward) derivation of the transition density based on the Milstein scheme. I implemented the estimation procedures for this study myself in R.

  Moreover, I have implemented inference procedures for SDE and ODE models in R and in the Stan software which provides an efficient C++ implementation of the Hamiltonian Monte Carlo based No-U-Turn Sampler.

- **Aim 2:** *Analyzing the benefits of an SDE model in terms of parameter identifiability*

  Using a model that explicitly accounts for the stochasticity inherent to intracellular processes holds the potential for better identifiability of kinetic parameters. I formulated an SDE model for the translation kinetics after mRNA transfection and compared it to an ODE model in terms of structural and practical parameter identifiability. I suggested two approaches to assess the structural parameter identifiability for the SDE model (by transforming the model and by simulation) and also implemented a recently suggested approach using the software DAISY. All three approaches suggested that the SDE model might yield better parameter identifiability.

Besides, I used the inference procedures implemented in R and Stan to infer parameters for both model types from simulated as well as (previously published) experimental data. The results show that the SDE model for the translation kinetics is clearly superior to the ODE model in terms of identifiability of the kinetic parameters.

- **Aim 3:** *Ensuring a sound mathematical foundation for the SDE model*

  The diffusion processes that are usually used to approximate biochemical processes do not fulfill the assumptions for standard results from stochastic analysis that ensure the existence of a unique solution of an SDE and that the Euler-Maruyama approximation converges to this solution. While the fundamental importance of these results is obvious, their derivation is generally neglected when diffusion approximations are applied in systems biology (and beyond).

  I proved the existence and uniqueness of a strong solution of the SDE model for the translation kinetics after mRNA transfection and that the Euler-Maruyama approximation of the SDE strongly converges to this solution. These proofs can easily be extended to further models as explained in Section 5.9.

Chapters 4 mainly addresses Aim 1. My work on this project was supervised by Prof. Dr. Christiane Fuchs and is published in the article
Pieschner, S. & Fuchs, C. (2020). Bayesian inference for diffusion processes: using higher-order approximations for transition densities. *Royal Society Open Science*, 7(10), 200270.

The code of the implementation of the investigated methods and the results of the simulation study are publicly available at `https://github.com/fuchslab/Inference_for_SDEs_with_the_Milstein_scheme`.

Chapter 5 addresses all 3 aims and contains two subprojects. The first is devoted to Aim 3 and is contained in Section 5.3. The process of developing the proofs was supervised by Prof. Dr. Wilfried Grecksch. The second subproject is covered in the remainder of the chapter and addresses Aims 1 and 2. My work on this part was supervised by Prof. Dr. Christiane Fuchs and some advice was provided by Prof. Dr. Jan Hasenauer.

1.2 Outline

This thesis is structured as follows: In Chapter 2, we provide background information on several topics that are relevant to follow the main chapters of this thesis. These include different ways to represent biochemical processes, Bayesian inference methods, in particular Markov chain Monte Carlo (MCMC) methods, and the concept of parameter identifiability. In

Chapter 3, we give a more detailed introduction to Itô diffusion processes which are the focus of this thesis. We define what Itô diffusion processes are and give some of their properties and some examples. As Itô diffusion processes are described by SDEs, we use the terms "diffusion (process)" and "SDE" interchangeably throughout this thesis. Moreover, we show how diffusion processes are commonly approximated and how parameter inference can be performed for them. In Chapter 4, we further study one of the inference methods for diffusion processes, that is based on Bayesian data augmentation, and investigate how to integrate the Milstein scheme, as an approximation scheme of higher convergence order, into this framework. We assess the effectiveness of this new combination in a simulation study and analyze whether due to the higher approximation accuracy, a higher estimation accuracy can be achieved for a fixed computational cost. In Chapter 5, we apply diffusion processes to model the translation kinetics after mRNA transfection. This application is motivated by the availability of time-lapse fluorescence microscopy data for single cells from an mRNA transfection experiment. For the SDE model of the translation kinetics, we proof the existence and uniqueness of a strong solution and that the Euler-Maruyama approximation of the SDE strongly converges to this solution. Moreover, we compare the SDE model to the corresponding ODE model in terms of structural and practical parameter identifiability. In Chapter 6, we provide a summary of the findings of this thesis and conclude with suggestions for further research.

Chapter 2

Background

This chapter provides a brief introduction to several topics that are relevant to follow the contents of this thesis. In Section 2.1, we present different representations of biochemical processes and shortly discuss their benefits and drawbacks. Furthermore, we give a primer on Bayesian statistics and Markov chain Monte Carlo (MCMC) methods in Section 2.2 and introduce the concept of parameter identifiability in Section 2.3.

2.1 Biochemical kinetic models

Biochemical kinetic models (BKMs) are used to describe the interactions and evolution of different species within a biological organism, e. g. a cell. They have been successfully applied to various problems and thus receive much attention as is apparent from several recent reviews on this topic such as Loskot et al. (2019), Schnoerr et al. (2017), Warne et al. (2019). In this section, we describe several common representations of BKMs.

2.1.1 Markov jump processes

The species within a biological system have a discrete nature. Moreover, random fluctuations of species numbers play a key role in biological systems. They can have a substantial effect on the system's behavior, in particular in the case of low species numbers. Therefore, a continuous-time, discrete-space Markov process, also called a *Markov jump process (MJP)*, for which the dynamics are described by the so-called chemical master equation (CME) is widely accepted to be an appropriate stochastic description of such a system (Gillespie, 1992b, Schnoerr et al., 2017).

Definition 2.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space and consider the measurable space $\left(\mathbb{N}_0^d, \mathcal{P}\left(\mathbb{N}_0^d\right)\right)$, where $\mathcal{P}\left(\mathbb{N}_0^d\right)$ denotes the power set of $\mathbb{N}_0^d$. The stochastic process $(\boldsymbol{X}(t))_{t\geq 0}$ with state space $\mathbb{N}_0^d$ is called a* Markov process *if for all $n \in \mathbb{N}$, all $0 \leq t_1 < \ldots < t_n$ and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{N}_0^d$, it holds that*

$$\mathbb{P}(\boldsymbol{X}(t_n) = \boldsymbol{x}_n \,|\, \boldsymbol{X}(t_1) = \boldsymbol{x}_1, \ldots, \boldsymbol{X}(t_{n-1}) = \boldsymbol{x}_{n-1}) = \mathbb{P}(\boldsymbol{X}(t_n) = \boldsymbol{x}_n \,|\, \boldsymbol{X}(t_{n-1}) = \boldsymbol{x}_{n-1}).$$

A Markov process is a memoryless process, i. e. future states $\boldsymbol{X}(t_n)$ conditioned on the current state $\boldsymbol{X}(t_{n-1})$ are independent of the past. The most common description of a BKM by a MJP assumes that the system is well mixed, in thermal equilibrium, and of constant size V, where V is a quantity that appropriately measures the size of the system such as the volume. Under these assumptions, spatial effects can be neglected and are, therefore, not included in the model (Schnoerr et al., 2017).

Suppose we consider a system of d species (e. g. d different types of molecules in a cell) denoted by $X_1, \ldots, X_d$ and of r reactions $R_1, \ldots, R_r$ which we define by the following notation:

$$R_j: \quad \sum_{i=1}^{d} s_{ij}^{-} X_i \xrightarrow{c_j} \sum_{i=1}^{d} s_{ij}^{+} X_i \qquad \text{for } j = 1, \ldots, r,$$

where $c_j \in \mathbb{R}_+$ is called the *reaction rate constant* of reaction R_j and $s_{ij}^{-}, s_{ij}^{+} \in \mathbb{N}_0$ are the *stoichiometric coefficients* denoting the number of reactants consumed and the number of products produced of species X_i by reaction R_j, respectively. For a total number $m = \sum_{i=1}^{d} s_{ij}^{-}$ of reactants involved in reaction R_j, we say that reaction R_j is of *order* m. Further, we define the stoichiometric matrix $\boldsymbol{S} \in \mathbb{N}_0^{d\times r}$ by

$$S_{ij} = s_{ij}^{+} - s_{ij}^{-} \qquad \text{for } i = 1, \ldots, d, \text{ and } j = 1, \ldots, r$$

and denote the columns of $\boldsymbol{S}$ by $\boldsymbol{S}_j$ which correspond to the net change of the system state when reaction R_j occurs. Throughout this thesis, we denote vectors and matrices by bold symbols, vectors are assumed to be column vectors, and the transpose of vector $\boldsymbol{x}$ (and matrix $\boldsymbol{M}$) is denoted by $\boldsymbol{x}^{\mathsf{Tr}}$ (and $\boldsymbol{M}^{\mathsf{Tr}}$, respectively).

For a stochastic representation of the kinetic model described above as a MJP, let $\boldsymbol{X}(t) = (X_1(t), \ldots, X_d(t))^{\mathsf{Tr}} \in \mathbb{N}_0^d$ be the random variable denoting the state of the system at time $t \geq 0$. Given that $\boldsymbol{X}(t) = \boldsymbol{x}$, the probability that reaction R_j occurs within an infinitesimal time step Δt and thus changes the system state by $\boldsymbol{S}_j$ is given by

$$\mathbb{P}(\boldsymbol{X}(t + \Delta t) = \boldsymbol{x} + \boldsymbol{S}_j \,|\, \boldsymbol{X}(t) = \boldsymbol{x}) = h_j(\boldsymbol{x}, c_j)\Delta t + o(\Delta t) \qquad \text{for } j = 1, \ldots, r,$$

where $o(\Delta t)/\Delta t \to 0$ for $\Delta t \to 0$, and $h_j(\boldsymbol{x}, c_j)$ is called the hazard or propensity function or simply reaction rate of reaction R_j. A common choice of the hazard function is the one of *mass-action kinetics type* (Wilkinson, 2019), i. e.

$$h_j(\boldsymbol{x}, c_j) = c_j \prod_{i=1}^{d} \frac{x_i!}{(x_i - s_{ij}^-)!\, s_{ij}^-!} \qquad \text{for } j = 1, \ldots, r.$$

We define $P(\boldsymbol{x}, t) = \mathbb{P}(\boldsymbol{X}(t) = \boldsymbol{x} \,|\, \boldsymbol{X}(t_0) = \boldsymbol{x}_0)$ for $t \geq t_0$ and all $\boldsymbol{x}, \boldsymbol{x}_0 \in \mathbb{N}_0^d$, where the dependence on the initial state $\boldsymbol{x}_0$ at time t_0 is suppressed for simplicity of notation. Then, a MJP described as above fulfills the following Kolmogorov's forward equations

$$\frac{\partial P(\boldsymbol{x}, t)}{\partial t} = \sum_{j=1}^{r} \left[h_j(\boldsymbol{x} - \boldsymbol{S}_j, c_j) P(\boldsymbol{x} - \boldsymbol{S}_j, t) - h_j(\boldsymbol{x}, c_j) P(\boldsymbol{x}, t) \right]. \tag{2.1}$$

In the context of BKMs, Equation (2.1) is known as the CME. It describes the temporal evolution of the probability distribution of the system states conditional on the initial state. While its structure is quite simple, analytical solutions of the CME only exist for the simplest examples of BKMs. Jahnke & Huisinga (2007), for example, derive an analytical solution for systems of only monomolecular reactions, i. e. $m = \sum_{i=1}^{d} s_{ij}^- \leq 1$ and $\sum_{i=1}^{d} s_{ij}^+ \leq 1$ for all $j = 1, \ldots, r$. Schnoerr et al. (2017) review further examples for which analytical solutions exist.

Despite the lack of an analytical solution to (2.1), exact simulation of sample paths is possible for all MJPs by means of the stochastic simulation algorithm (SSA) that was introduced to the field of biochemical kinetics by Gillespie (1976, 1977) and thus also came to be known as Gillespie's algorithm. The algorithm makes use of the fact that the time intervals between two successive reactions are exponentially distributed with intensity

$$h_0(\boldsymbol{x}, \boldsymbol{c}) = \sum_{j=1}^{r} h_j(\boldsymbol{x}, c_j),$$

where $\boldsymbol{c} = (c_1, \ldots, c_r)^{\mathsf{Tr}}$ denotes the vector of the reaction rate constants. It iteratively samples the time τ until the next reaction and the type k of reaction that occurs next. We summarize the steps of Gillespie's algorithm in Algorithm 2.1 in a similar way as Fuchs (2013) and Wilkinson (2019).

The computational cost of Gillespie's algorithm becomes cumbersome if many reactions occur within a short time. This is the case if there are many reaction types or the value of any hazard function is high, i. e. if the respective rate constant and/or the numbers of the involved reactants are large. Several more efficient implementation of the exact simulation algorithm

Algorithm 2.1: Gillespie's algorithm

Input: *An initial state $\boldsymbol{x}_0$ and initial time point t_0, maximal time point T, the hazard functions h_j and reaction rate constants c_j for $j = 1, \dots, r$.*

Set $t = t_0$ and $\boldsymbol{x}(t) = \boldsymbol{x}_0$.
while $t < T$ **do**
1. Calculate $h_j(\boldsymbol{x}, c_j)$ for $j = 1, \dots, r$ and $h_0(\boldsymbol{x}, \boldsymbol{c})$.
2. Draw $\tau \sim \mathrm{Exp}(h_0(\boldsymbol{x}(t), \boldsymbol{c}))$ and set $\tau^* = \min\{\tau, T - t\}$.
3. Draw the index $k \in \{1, \dots, r\}$ of the next reaction as discrete random variable with probabilities $h_j(\boldsymbol{x}, c_j)/h_0(\boldsymbol{x}, \boldsymbol{c})$ for $j = 1, \dots, r$.
4. Set $\boldsymbol{x}(s) = \boldsymbol{x}(t)$ for all $s \in (t, t + \tau^*)$ and $\boldsymbol{x}(t + \tau^*) = \boldsymbol{x}(t) + \boldsymbol{S}_k \mathbb{1}(\tau^* = \tau)$.
5. Set $t = t + \tau$

end

Output: *A sample path $\boldsymbol{x}(t)$ of MJP $(\boldsymbol{X}(t))_{t \geq 0}$ on time interval $[t_0, T]$.*

have been suggested such as the *next reaction method* by Gibson & Bruck (2000) and a refinement of this method by Anderson (2007). Despite these improvements, the computational cost of exact simulation becomes often intractable. Therefore, approximate simulation methods for MJPs have been developed such as the *tau-leaping method* (Gillespie, 2001), several improvements thereof (e. g. in Cao et al., 2006, Tian & Burrage, 2004), and also combinations of exact and approximate simulation (e. g. in Cao et al., 2005, Marchetti et al., 2016).

Due to the fact that there is generally no analytical solution to the CME (2.1), inference for MJPs is challenging because in this case the likelihood is not available. Therefore, only likelihood-free inference approaches are feasible. Those include for example pseudo-marginal MCMC methods as in Andrieu & Roberts (2009) for exact inference and approximate Bayesian computation (ABC) as e. g. in Toni et al. (2009) for approximate inference. Both approaches require several forward simulations of the considered MJP and are thus computationally very intense. See Warne et al. (2019) for an overview and further references for both classes of methods. Since inference for MJP representations of BKMs is usually computationally intractable, several approximations to MJPs have been developed some of which we will introduce in the next subsection.

2.1.2 Approximation methods

There are several other representations of the BKM introduced in the previous section. To some extent those can be considered as approximations to the corresponding MJP. The most commonly used representation is the *reaction rate equation (RRE)* which is a system of ordinary differential equations (ODEs) and thus provides a deterministic and state-continuous description of the kinetics. This approach commonly considers concentrations of the different species

relative to the system size V instead of absolute numbers. We denote the concentrations by

$$[\boldsymbol{X}] = \frac{\boldsymbol{X}}{V},$$

whereby $[\boldsymbol{X}] \in \mathbb{R}_+^d$. For a large system size V, the concentrations can thus be considered to be approximately continuous. Moreover, the propensity functions according to classical mass-action kinetics (Waage & Gulberg, 1986) are defined by

$$\tilde{h}_j([\boldsymbol{X}], \tilde{c}_j) = \tilde{c}_j \prod_{i=1}^{d} [X_i]^{s_{ij}^-} \qquad \text{for } j = 1, \ldots, r,$$

where the rate constants $\tilde{c}_j \in \mathbb{R}_+$ may differ from those introduced in the previous section due to the different units of $\boldsymbol{X}$ and $[\boldsymbol{X}]$. Wilkinson (2019, Chapter 6.7) gives some examples of how to convert between the different rate constants. For first order reactions, the rate constants of the stochastic and the deterministic description are equal. We denote the vector-valued function of all propensity functions by $\tilde{\boldsymbol{h}}([\boldsymbol{X}], \tilde{\boldsymbol{c}}) = (\tilde{h}_1([\boldsymbol{X}], \tilde{c}_1), \ldots, \tilde{h}_r([\boldsymbol{X}], \tilde{c}_r))^{\mathsf{Tr}}$. Then the RRE written in matrix notation reads as follows

$$\frac{\mathrm{d}[\boldsymbol{X}](t)}{\mathrm{d}t} = \boldsymbol{S}\tilde{\boldsymbol{h}}([\boldsymbol{X}](t), \tilde{\boldsymbol{c}}), \qquad [\boldsymbol{X}](0) = \frac{\boldsymbol{x}_0}{V}, \tag{2.2}$$

where $\boldsymbol{S}$ denotes the stoichiometric matrix as defined in the previous section.

The deterministic dynamics described by the RRE (2.2) do not capture the inherently stochastic nature of the underlying process. The RRE is only an appropriate description for a system that has large numbers of all species such that relative stochastic fluctuations become less prominent. Moreover, for non-linear dynamics, i. e. in the case of second order reactions and higher, the solution of the RRE does not necessarily describe the mean behavior of concentrations of the corresponding MJP. Nevertheless, the RRE has successfully been applied to many problems because it has "the advantage of being relatively straightforward to analyse" (Schnoerr et al., 2017) and several computationally highly efficient software tools are available for the RRE and other ODE-based approximation approaches (see e. g. Fröhlich et al., 2017, Kazeroonian et al., 2016, Raue et al., 2015, Stapor et al., 2018). This allows simulation, analysis, and inference even for large-scale BKMs, i. e. networks with many species and types of reactions (Fröhlich et al., 2018, Kapfer et al., 2019, Schmiester et al., 2019, Terje Lines et al., 2019).

Another approach to approximate the MJP from the previous section is the approximation by a diffusion process which is also state-continuous as the solution of the RRE but it preserves the stochastic nature of the process. We give a formal introduction to diffusion processes in Chapter 3 and here only briefly present their role as a representation of a BKM. There are

several different ways how to derive such a *diffusion approximation*. Fuchs (2013, Chapter 4) provides an overview of different diffusion approximation techniques and explains that under mild regularity conditions, all of them yield the same result. In the derivation, many but not all of the techniques also divide the process states by the system size V (as for the RRE) in order to obtain smaller jump sizes and thus justify the approximation by a continuous process. The resulting diffusion approximation, however, can then in most cases easily be scaled again by V to obtain the scaling of the original process. For simplicity of notation, we therefore keep the original scaling and state the diffusion approximation in a similar way as derived in Gillespie (2000).

The approximating diffusion process is described by the following stochastic differential equation (SDE) which is better known as the *chemical Langevin equation (CLE)* in this context:

$$\mathrm{d}\boldsymbol{X}(t) = \boldsymbol{\mu}(\boldsymbol{X}(t), \boldsymbol{c})\mathrm{d}t + \boldsymbol{\sigma}(\boldsymbol{X}(t), \boldsymbol{c})\mathrm{d}\boldsymbol{B}(t), \qquad \boldsymbol{X}(0) = \boldsymbol{x}_0, \tag{2.3}$$

with drift function

$$\boldsymbol{\mu}(\boldsymbol{X}(t), \boldsymbol{c}) = \boldsymbol{S}\boldsymbol{h}(\boldsymbol{X}(t), \boldsymbol{c})$$

and diffusion function

$$\boldsymbol{\sigma}(\boldsymbol{X}(t), \boldsymbol{c}) = \sqrt{\boldsymbol{S}\mathrm{diag}\{\boldsymbol{h}(\boldsymbol{X}(t), \boldsymbol{c})\}\boldsymbol{S}^{\mathsf{Tr}}},$$

and where $\boldsymbol{B}(t)$ denotes a q-dimensional standard Brownian motion, $\boldsymbol{S}$ denotes the stoichiometric matrix as defined in the previous section, and $\mathrm{diag}\{\boldsymbol{h}(\boldsymbol{X}(t), \boldsymbol{c})\}$ denotes the diagonal matrix with the elements of the vector $\boldsymbol{h}(\boldsymbol{X}(t), \boldsymbol{c}) = (h_1(\boldsymbol{X}(t), c_1), \ldots, h_r(\boldsymbol{X}(t), c_r))$ of the hazard functions on the main diagonal. Moreover, for a square matrix $\boldsymbol{A}$, the square root $\sqrt{\boldsymbol{A}}$ here denotes any matrix $\boldsymbol{B}$ that satisfies $\boldsymbol{A} = \boldsymbol{B}\boldsymbol{B}^{\mathsf{Tr}}$. Thus, one possible choice is $\boldsymbol{\sigma}(\boldsymbol{X}(t), \boldsymbol{c}) = \boldsymbol{S}\mathrm{diag}\{\sqrt{\boldsymbol{h}(\boldsymbol{X}(t), \boldsymbol{c})}\}$ such that $\boldsymbol{\sigma}(\boldsymbol{X}(t), \boldsymbol{c}) \in \mathbb{R}^{d\times r}$ and $q = r$. However, a square matrix obtained e. g. by Cholesky factorization such that $\boldsymbol{\sigma}(\boldsymbol{X}(t), \boldsymbol{c}) \in \mathbb{R}^{d\times d}$ and $q = d$ is usually preferred (Wilkinson, 2019, Chapter 8.3).

By preserving the stochastic nature of the process, the diffusion approximation can capture more information about the underlying process. For systems that only contain zeroth and first order reactions, the first and second moments of the diffusion approximation coincide with the moments of the corresponding MJP (Schnoerr et al., 2017); whereas the RRE can only capture the first moment. The diffusion approximation does not maintain the discrete nature of the MJP which on the one hand is a disadvantage of this representation, but on the other hand this allows for more efficient (approximate) simulation of sample paths as the computational cost only scales with the number of different species in the system instead of with the rate of

occurring reactions as for the MJP. Another advantage is that for the diffusion approximation a wider range of possible inference methods becomes available. We give an overview of these methods in Section 3.4.

There are ample of further approximation methods for BKMs. Schnoerr et al. (2017) provide an overview of many of them. Another prominent example is the linear noise approximation (LNA) which can be considered an intermediate result between the RRE and the CLE (Wallace et al., 2012). It is appealing because under certain (commonly fulfilled) conditions it has a tractable likelihood. However, it is also only appropriate for very large-sized systems.

Before we close this section, we want to briefly illustrate the different representations of BKMs on a small example. Therefore, we consider a linear birth-death process in a similar way as Wilkinson (2019, Chapter 1.3). This process models the population size of one species X. The initial size of the population at time $t = 0$ is $X(0) = x_0$. There are two possible events (or reactions, in the terminology that we have used so far):

- *birth* that occurs at rate $\theta_1 \in \mathbb{R}_+$ which means each unit of the population on average gives rise to θ_1 new units of the population per time unit, and
- *death* that occurs at rate $\theta_2 \in \mathbb{R}_+$ which means on average the population decreases by a proportion of θ_2 per time unit.

Hence, for a MJP representation of this process, it holds that

$$\begin{aligned}\mathbb{P}(X(t+\Delta t) = x+1 \,|\, X(t) = x) &= \theta_1 x \Delta t + o(\Delta t) \text{ and}\\ \mathbb{P}(X(t+\Delta t) = x-1 \,|\, X(t) = x) &= \theta_2 x \Delta t + o(\Delta t),\end{aligned}$$

and the process can be simulated using Gillespie's algorithm described in Algorithm 2.1.

The deterministic representation by an ODE (the RRE) reads

$$\frac{\mathrm{d}X(t)}{\mathrm{d}t} = (\theta_1 - \theta_2)X(t), \qquad X(0) = x_0,$$

where for the purpose of this illustration we use absolute units instead of concentrations. Its solution is the deterministic function

$$X(t) = x_0 e^{(\theta_1 - \theta_2)t}.$$

The SDE representing the same phenomenon is given by

$$\mathrm{d}X(t) = (\theta_1 - \theta_2)X(t)\,\mathrm{d}t + \sqrt{\theta_1 + \theta_2}\,\mathrm{d}B(t), \qquad X(0) = x_0,$$

where B(t) is a one-dimensional standard Brownian motion. Trajectories of the (approximate) solution of this SDE can be generated e. g. based on normally distributed random numbers as we will explain in Section 3.3.

As an illustration, we simulate trajectories for each of the three representations (MJP, ODE, SDE) of the linear birth-death process and for two different parameter combinations. We choose the parameters such that for both scenarios, the difference $\theta_1 - \theta_2$ is equal to -0.05. We set $X(0) = 30$ and simulate the processes on the time interval $[0, 40]$. Figure 2.1 shows the trajectories. We see that for each parameter combination, the stochastic representations MJP and SDE give rise to several quite different trajectories while the deterministic ODE model only yields one trajectory. Moreover, the trajectories of the ODE model are the same for both parameter combinations as they are determined solely by the difference $\theta_1 - \theta_2$ but not by the individual values of the parameters θ_1 and θ_2. Hence, it would not be possible to determine the parameters individually from one observed trajectory by means of the ODE model. Whereas, for the MJP and the SDE model, we clearly see that the higher values for θ_1 and θ_2 (in the lower row of the figure) lead to more variability. This indicates that these two representations capture information about both parameters individually, not only about their difference, which yields the potential to also determine both parameters from observed data - a property which we would like to harness.

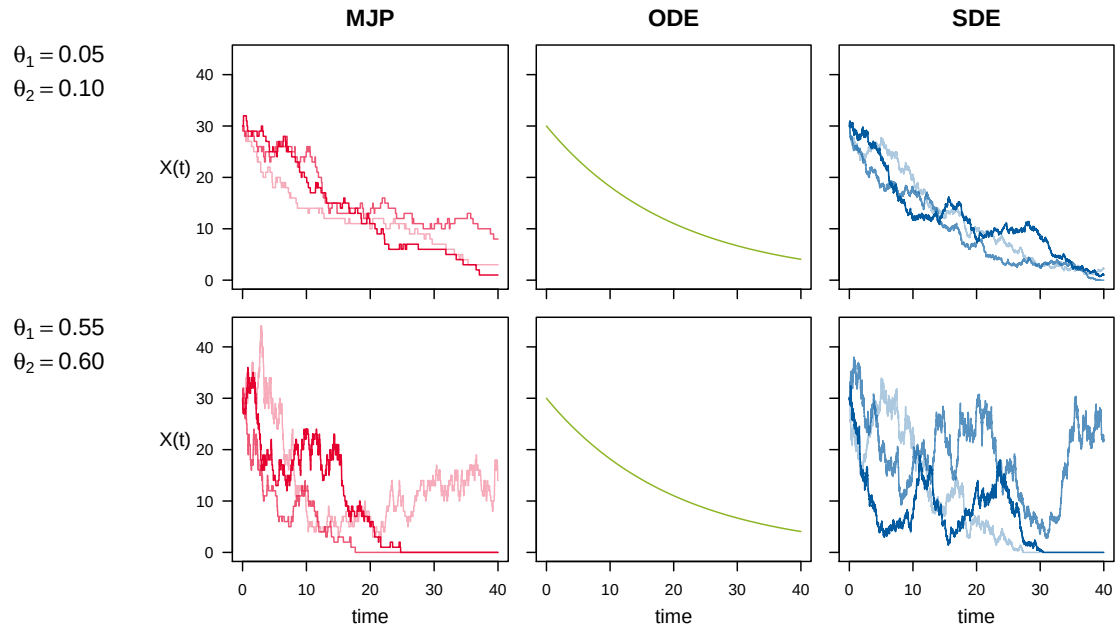


Figure 2.1: Example trajectories for three representations (MJP, ODE, SDE) of a linear birth-death process with starting value $X(0) = 30$ and for different values of the birth rate θ_1 and the death rate θ_2. For both parameter combinations, the difference $\theta_1 - \theta_2$ is equal to -0.05; and therefore, the ODE trajectories are identical.

2.2 Bayesian statistics and MCMC methods

Throughout this thesis, we mainly take a Bayesian approach to inference for parametric models. In Bayesian statistics, we can formulate our assumptions and general knowledge about the model parameters $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$ in terms of a prior distribution with probability density $p(\boldsymbol{\theta})$. After having observed data $\mathcal{D}$ about the phenomenon which we are trying to model, we update our knowledge about the parameter and describe it by the posterior distribution with density $p(\boldsymbol{\theta} \mid \mathcal{D})$. The relation between the prior and the posterior density is defined by Bayes' theorem:

$$p(\boldsymbol{\theta} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\mathcal{D})}, \tag{2.4}$$

where $p(\mathcal{D})$ denotes the density of the distribution of $\mathcal{D}$ and is sometimes called the marginal density of the evidence, and $p(\mathcal{D} \mid \boldsymbol{\theta})$ denotes the density of the distribution of $\mathcal{D}$ conditioned on $\boldsymbol{\theta}$ and is determined by the considered model. Viewed as a function of the parameter, $l(\boldsymbol{\theta} \mid \mathcal{D}) := p(\mathcal{D} \mid \boldsymbol{\theta})$ is called the likelihood (function), and often it is more convenient to consider the log-likelihood $\mathcal{L}(\boldsymbol{\theta} \mid \mathcal{D}) := \log l(\boldsymbol{\theta} \mid \mathcal{D})$. The posterior distribution $p(\boldsymbol{\theta} \mid \mathcal{D})$ represents our beliefs about how plausible different parameter values are, after we have observed data $\mathcal{D}$. It is often summarized e. g. by point estimates as its mean, median, and mode or based on credible intervals (which can be obtained in different ways). Yet, it is important to remember that in Bayesian statistics the entire distribution constitutes the Bayesian estimate. Comprehensive introductions to Bayesian statistics can be found e. g. in Lee (2012) and Gelman et al. (2013).

Two of the merits of Bayesian methods are the straightforward ways to include prior knowledge about the model parameter into the inference problem and to assess the uncertainty about a parameter estimate e. g. based on the posterior variance (see e. g. Berger, 1985, Chapter 4.1). On the other hand, it is often difficult or not possible to derive analytical results for the posterior distribution, especially as they tend to be very high-dimensional probability distributions. However, MCMC methods have proved very useful in this context. They can be used to draw samples from (almost) any probability density function $\pi(\boldsymbol{\theta})$. The idea behind MCMC methods is to construct a Markov chain, i. e. a discrete-time Markov process, whose stationary (sometimes also called invariant) distribution corresponds to the target density $\pi(\boldsymbol{\theta})$. Before we will introduce some commonly used MCMC methods, we shortly provide some basic theory of Markov chains in the following subsection.

2.2.1 A brief introduction to Markov chain theory

This short introduction to some basic concepts about Markov chains is based on the expositions in Geyer (1992), Chib & Greenberg (1995), and Robert & Casella (2002). For the following definitions, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space and $(S, \mathcal{S})$ a measurable space, which we will call the state space. S can be continuous, e.g. $(S, \mathcal{S}) = \left(\mathbb{R}^d, \mathcal{B}\left(\mathbb{R}^d\right)\right)$ where $\mathcal{B}\left(\mathbb{R}^d\right)$ is the Borel σ-algebra on $\mathbb{R}^d$, or discrete, e.g. $(S, \mathcal{S}) = \left(\mathbb{Z}^d, \mathcal{P}\left(\mathbb{Z}^d\right)\right)$ where $\mathcal{P}\left(\mathbb{Z}^d\right)$ denotes the power set of $\mathbb{Z}^d$. Analogously to Definition 2.1, we define a discrete-time Markov process as follows.

Definition 2.2. *A discrete-time stochastic process* $(\boldsymbol{X}_n)_{n\geq 0}$ *is called a* Markov chain *with respect to the probability measure* $\mathbb{P}$*, if for all* $n \in \mathbb{N}_0$ *and all* $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{n+1} \in S$*, it holds that*

$$\mathbb{P}\left(\boldsymbol{X}_{n+1} = \boldsymbol{x}_{n+1} | \boldsymbol{X}_0 = \boldsymbol{x}_0, \ldots, \boldsymbol{X}_n = \boldsymbol{x}_n\right) = \mathbb{P}\left(\boldsymbol{X}_{n+1} = \boldsymbol{x}_{n+1} | \boldsymbol{X}_n = \boldsymbol{x}_n\right) \tag{2.5}$$

given that the conditional probabilities are well defined.

The condition (2.5) is called *Markov property* and may be interpreted as stating that the conditional probability of any future state $\boldsymbol{X}_{n+1}$, given the past states $\boldsymbol{X}_0, \ldots, \boldsymbol{X}_{n-1}$ and the present state $\boldsymbol{X}_n$, only depends on the present state, but is independent of the past states. A Markov chain is called *homogeneous* if the conditional probability distribution $\mathbb{P}\left(\boldsymbol{X}_{n+1} = \boldsymbol{y} | \boldsymbol{X}_n = \boldsymbol{x}\right)$, representing the probability of moving from state $\boldsymbol{x}$ to state $\boldsymbol{y}$, is independent of n. The conditional probability distribution is also called transition kernel, and in the homogeneous case, it can simply be denoted by $P(\boldsymbol{x}, \boldsymbol{y}) := \mathbb{P}\left(\boldsymbol{X}_{n+1} = \boldsymbol{y} | \boldsymbol{X}_n = \boldsymbol{x}\right)$ for $\boldsymbol{x}, \boldsymbol{y} \in S$. The formal definition of a stochastic kernel is as follows:

Definition 2.3. *A* transition kernel *is a function* P *defined on* $(S, \mathcal{S})$ *such that*

- $\forall \boldsymbol{x} \in S$, $P(\boldsymbol{x}, .)$ *is a probability measure;*
- $\forall A \in \mathcal{S}$, $P(., A)$ *is* $\mathcal{S}$*-measurable.*

Two central questions when studying Markov chains are to find conditions for the existence of a so-called stationary distribution π, as well as conditions under which repeated iterations of the transition kernel of the Markov chain will converge to π.

Definition 2.4. *A probability measure* π *is* stationary *(or* invariant*) for the transition kernel* $P(.,.)$ *if it satisfies*

$$\pi(A) = \int_S P(\boldsymbol{x}, A)\pi(d\boldsymbol{x}), \quad \forall A \in \mathcal{S}.$$

In order to ensure that a stationary distribution exists, one of the prerequisites imposed on the transition kernel in the setup of MCMC methods is irreducibility.

Definition 2.5. *Given a measure φ on $\mathcal{S}$, the Markov chain $(\boldsymbol{X}_n)$ with transition kernel $P(\boldsymbol{x}, \boldsymbol{y})$ is called* φ-irreducible *if for every $A \in \mathcal{S}$ with $\varphi(A) > 0$, there exists n such that $P^n(\boldsymbol{x}, A) > 0$ for all $\boldsymbol{x} \in S$, where $P^n(\boldsymbol{x}, A)$ denotes the kernel for n transitions obtained by $P^1(\boldsymbol{x}, A) = P(\boldsymbol{x}, A)$ and $P^n(\boldsymbol{x}, A) = \int_S P^{n-1}(\boldsymbol{y}, A) P(\boldsymbol{x}, d\boldsymbol{y})$.*

The property of irreducibility means that the transition kernel allows the chain $(\boldsymbol{X}_n)$ to move across the entire state space S. It implies that the chain can reach any non-zero measure region of the state space regardless of what the starting value $\boldsymbol{X}_0$ was. Moreover, there are the following classification criteria for Markov chains:

Definition 2.6. *A Markov chain $(\boldsymbol{X}_n)$ is* recurrent *if*

- *there exists a measure ϕ such that $(\boldsymbol{X}_n)$ is ϕ-irreducible, and*
- *for every $A \in \mathcal{S}$ such that $\phi(A) > 0$, $\mathbb{E}_{\boldsymbol{x}}[\eta_A] = \infty$ for all $\boldsymbol{x} \in A$, where $\eta_A = \sum_{n=1}^{\infty} \mathbb{1}_A(\boldsymbol{X}_n)$ and $\mathbb{1}_A(.)$ denotes the indicator function.*

This signifies that a recurrent chain on average visits any set $A \in \mathcal{S}$ infinitely many times. Whereas the following property means that the expected return time to any state is finite.

Definition 2.7. *A Markov chain $(\boldsymbol{X}_n)$ is* positive recurrent *if for all states $\boldsymbol{x} \in S$, it holds that $\mathbb{E}_{\boldsymbol{x}}[\tau(\boldsymbol{x})] < \infty$, where $\tau(\boldsymbol{x}) = \inf\{n \in \mathbb{N} : \boldsymbol{X}_n = \boldsymbol{x}\}$.*

A key result from Markov chain theory states that for any irreducible and positive recurrent Markov chain (X_n), there exists a (up to a constant) unique stationary probability distribution π. Note that the mere existence and uniqueness of a stationary probability distribution does not guarantee that a Markov chain will actually converge to this distribution. This behavior can be ensured by further properties such as aperiodicity and ergodicity. See Robert & Casella (2002) for a derivation of these results.

For MCMC methods, the question is not whether an invariant measure exists but quite the opposite is the case: the density of the invariant measure is known as it is the target density $\pi(\cdot)$ from which we would like to sample. In order to generate samples from $\pi(\cdot)$, the crucial point is how to construct a transition kernel $P(\boldsymbol{x}, \boldsymbol{y})$ whose n^{th} iteration tends to $\pi(\cdot)$ as n grows large. That means the chain should be able to start at an arbitrary state $\boldsymbol{x} \in S$ and then after a large number of iterations, the distribution of the generated samples corresponds approximately to the target distribution. A useful property of a transition kernel in that regard is the so-called reversibility or detailed-balance condition.

Definition 2.8. *A Markov chain with transition kernel P satisfies the* detailed-balance condition *if there exists a function f satisfying*

$$f(\boldsymbol{x})P(\boldsymbol{x}, d\boldsymbol{y}) = f(\boldsymbol{y})P(\boldsymbol{y}, d\boldsymbol{x}), \quad \text{for every } \boldsymbol{x}, \boldsymbol{y} \in S.$$

If a Markov chain with transition kernel P satisfies the detailed-balance condition with a probability density function f, it can be shown that this density f is the invariant density of the chain (see e. g. Robert & Casella, 2002, Theorem 6.46). The detailed-balance condition is quite restrictive; however, it is not a necessary but a sufficient condition and usually easy to check.

2.2.2 Examples of MCMC methods

The Metropolis-Hastings algorithm

A basic but very versatile MCMC method that makes use of the detailed-balance condition and its consequence is the *Metropolis-Hastings algorithm*. The algorithm was originally suggested by Metropolis et al. (1953) to generate Markov chains according to a Boltzmann distribution and later generalized by Hastings (1970).

In each iteration of the algorithm, a new state $\boldsymbol{y}$ is generated from a proposal density $q(\boldsymbol{y}|\boldsymbol{x})$ that may depend on the current state $\boldsymbol{x}$ of the Markov chain and satisfies $\int_S q(\boldsymbol{y}|\boldsymbol{x})d\boldsymbol{y} = 1$ (in the case of a discrete state space, the integral is replaced by a sum). In practice, we will try to choose q, of course, such that it is easy to simulate. Apart from that, there are only two important requirements for q: Firstly, the ratio $\pi(\boldsymbol{y})/q(\boldsymbol{y}|\boldsymbol{x})$ must be known up to a constant which is independent of $\boldsymbol{x}$, and secondly, $q(\cdot|\boldsymbol{x})$ must have enough dispersion to ensure that the entire support of π can be explored. Therefore, a minimum requirement is that

$$\bigcup_{\boldsymbol{x} \,\in\, \mathrm{supp}(\pi)} \mathrm{supp}\left(q(\cdot|\boldsymbol{x})\right) \supset \mathrm{supp}(\pi),$$

where $\mathrm{supp}(g) = \{\boldsymbol{x} \in S \,|\, g(\boldsymbol{x}) > 0\}$ denotes the support of function g.

In order to ensure that the transition kernel of the algorithm satisfies the detailed balance condition, the proposed state $\boldsymbol{y}$ is only accepted with probability

$$\alpha(\boldsymbol{x}, \boldsymbol{y}) = \min\left[1, \frac{\pi(\boldsymbol{y})q(\boldsymbol{x}|\boldsymbol{y})}{\pi(\boldsymbol{x})q(\boldsymbol{y}|\boldsymbol{x})}\right]$$

which is called the Metropolis-Hastings acceptance probability. If the proposed state is rejected (which happens with probability $1 - \alpha(\boldsymbol{x}, \boldsymbol{y})$), the chain stays in the current state $\boldsymbol{x}$. The transition kernel of a Markov chain generated by the Metropolis-Hasting algorithm is thus:

$$P_{MH}(\boldsymbol{x}, \boldsymbol{y}) = q(\boldsymbol{y}|\boldsymbol{x})\alpha(\boldsymbol{x}, \boldsymbol{y}) + \left(1 - \int_{\mathbb{R}^d} q(\boldsymbol{y}|\boldsymbol{x})\alpha(\boldsymbol{x}, \boldsymbol{y})d\boldsymbol{y}\right)\delta_{\boldsymbol{x}}(\boldsymbol{y}),$$

where $\delta_{\boldsymbol{x}}(\cdot)$ denotes the Dirac mass in $\boldsymbol{x}$; and therefore, the second summand gives the probability of staying in state $\boldsymbol{x}$. We can easily verify that P_{MH} satisfies the detailed balance condition with the target density π. For the first summand, we have

$$\begin{aligned} q(\boldsymbol{y}|\boldsymbol{x})\alpha(\boldsymbol{x}, \boldsymbol{y})\pi(\boldsymbol{x}) &= q(\boldsymbol{y}|\boldsymbol{x}) \min\left[1, \frac{\pi(\boldsymbol{y})q(\boldsymbol{x}|\boldsymbol{y})}{\pi(\boldsymbol{x})q(\boldsymbol{y}|\boldsymbol{x})}\right]\pi(\boldsymbol{x}) = \min\left[\pi(\boldsymbol{x})q(\boldsymbol{y}|\boldsymbol{x}), \pi(\boldsymbol{y})q(\boldsymbol{x}|\boldsymbol{y})\right] \\ &= q(\boldsymbol{x}|\boldsymbol{y}) \min\left[\frac{\pi(\boldsymbol{x})q(\boldsymbol{y}|\boldsymbol{x})}{\pi(\boldsymbol{y})q(\boldsymbol{y}|\boldsymbol{x})}, 1\right]\pi(\boldsymbol{y}) = q(\boldsymbol{x}|\boldsymbol{y})\alpha(\boldsymbol{y}, \boldsymbol{x})\pi(\boldsymbol{y}), \end{aligned}$$

and for the second summand,

$$\begin{aligned} \left(1 - \int_{\mathbb{R}^d} q(\boldsymbol{y}|\boldsymbol{x})\alpha(\boldsymbol{x}, \boldsymbol{y})d\boldsymbol{y}\right)\delta_{\boldsymbol{x}}(\boldsymbol{y})\pi(\boldsymbol{x}) &= \left(1 - \int_{\mathbb{R}^d} q(\boldsymbol{x}|\boldsymbol{y})\alpha(\boldsymbol{y}, \boldsymbol{x})d\boldsymbol{x}\right)\delta_{\boldsymbol{y}}(\boldsymbol{x})\pi(\boldsymbol{y}) \\ &= 0, \text{ if } \boldsymbol{x} \neq \boldsymbol{y}. \end{aligned}$$

Hence, the Markov chain generated with the Metropolis-Hastings kernel P_{MH} has the target π as its stationary distribution. If we let the chain run for infinitely many iterations, we would obtain a sample distributed according to π. In practice, of course, it is not possible to let the chain run infinitely long. For a finite chain, we have to ensure that chain has converged to sampling from the stationary distribution. We will discuss how this can be assessed in Section 2.2.3. Moreover, in order to reduce the dependence on the initial state in which the chain was started and thus to reduce the bias this may cause in a finite sample compared to the stationary distribution, one usually discards several iterations at the beginning of the chain as a so-called *burn-in* or *warm-up* phase (see e. g. Brooks & Roberts, 1998, Gelman et al., 2013). We summarize the basic steps of the Metropolis-Hastings algorithm (including a burn-in phase) in Algorithm 2.2.

The acceptance/rejection decision in Step 2 can be implemented by generating a uniform random variable $u \sim \mathcal{U}(0, 1)$, accepting $\boldsymbol{y}$ if $u < \alpha(\boldsymbol{x}, \boldsymbol{y})$, and rejecting it otherwise. If the proposal density is symmetric, i.e. $q(\boldsymbol{y}|\boldsymbol{x}) = q(\boldsymbol{x}|\boldsymbol{y})$, then the acceptance probability reduces to $\alpha(\boldsymbol{x}, \boldsymbol{y}) = \pi(\boldsymbol{y})/\pi(\boldsymbol{x})$. In this case, the proposed state is always accepted if $\pi(\boldsymbol{y}) \geq \pi(\boldsymbol{x})$, i.e. if the new state increases the value of the target density, otherwise the process will move to $\boldsymbol{y}$ only with probability $\pi(\boldsymbol{y})/\pi(\boldsymbol{x}) < 1$. This property also serves as the basis for several (stochastic) optimization algorithms such as the simulated annealing algorithm (see Fouskakis & Draper, 2002, for a review).

Algorithm 2.2: Metropolis-Hastings algorithm

Input: *A target density $\pi(\cdot)$, the proposal density $q(\cdot|\cdot)$, an initial state $\boldsymbol{x}^{(0)}$, number of iterations n, and number of samples to discard as burn-in k.*

In each iteration $i = 1, \ldots, n$:

Step 1 Generate a new state $\boldsymbol{y}$ according to $q(\cdot|\boldsymbol{x}^{(i-1)})$.

Step 2 Accept $\boldsymbol{y}$ as $\boldsymbol{x}^{(i)}$ with probability $\alpha(\boldsymbol{x}^{(i-1)}, \boldsymbol{y}) = \min\left[1, \frac{\pi(\boldsymbol{y})q(\boldsymbol{x}^{(i-1)}|\boldsymbol{y})}{\pi(\boldsymbol{x}^{(i-1)})q(\boldsymbol{y}|\boldsymbol{x}^{(i-1)})}\right]$, if $\boldsymbol{y}$ is rejected $\boldsymbol{x}^{(i)} := \boldsymbol{x}^{(i-1)}$.

Output: *A sample $\{\boldsymbol{x}^{(k+1)}, \ldots, \boldsymbol{x}^{(n)}\}$ approximately distributed according to $\pi(\cdot)$.*

Moreover, it is important to note that in general, the calculation of the acceptance probability $\alpha(\boldsymbol{x}, \boldsymbol{y})$ does not require us to know the normalization constant of $\pi(\cdot)$. As $\pi(\cdot)$ appears in the numerator as well as in the denominator of $\alpha(\boldsymbol{x}, \boldsymbol{y})$, the normalization constant simply cancels out. So these are two of the features that make MCMC methods such a powerful tool in Bayesian statistics. They allow us to simulate Markov chains with a desired stationary distribution and a corresponding density which in Bayesian statistics will usually be the posterior distribution, even if it is not possible to sample directly from this distribution. Moreover, for many of these methods, we only need to be able to evaluate the posterior density up to a constant, i. e. it is sufficient to be able to evaluate the prior density and the likelihood function as by Bayes' theorem, we have

$$p(\boldsymbol{\theta} \mid \mathcal{D}) \propto p(\mathcal{D} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta}).$$

Gibbs-sampling

Another very general MCMC method is the *Gibbs sampling algorithm* named after the physicist Josia Willard Gibbs by Geman & Geman (1984) (see Brooks et al., 2011, Chapter 2. for a historical account). Suppose we want to sample from the p-dimensional distribution $p(\boldsymbol{\theta} \mid \mathcal{D})$ for parameter $\boldsymbol{\theta} \in \mathbb{R}^p$. Let $\boldsymbol{\theta}_{-j} = (\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_p)$ be the vector of all components of $\boldsymbol{\theta}$ except for θ_j and further suppose that we are able to sample from the (so-called *full*) conditional densities $p_j(\theta_j \mid \boldsymbol{\theta}_{-j}, \mathcal{D})$ for $j = 1, \ldots, p$. Then, in each iteration, the components

of $\boldsymbol{\theta}$ can be updated one-by-one as summarized in Algorithm 2.3 where we also include a burn-in phase.

Algorithm 2.3: Gibbs sampling algorithm

Input: *A target density $p(\cdot \mid \mathcal{D})$, full conditional densities $p_j(\cdot \mid \cdot, \mathcal{D})$, initial state $\boldsymbol{\theta}^{(0)}$, number of iterations n, and number of samples to discard as burn-in k.*

In each iteration $i = 1, \ldots, n$:

For each index $j = 1, \ldots, p$:

Step j Generate $\theta_j^{(i)}$ according to $p_j\left(\cdot \,\middle|\, \boldsymbol{\theta}_{-j}^{(i-1)}, \mathcal{D}\right)$

with $\boldsymbol{\theta}_{-j}^{(i-1)} = \left(\theta_1^{(i)}, \ldots, \theta_{j-1}^{(i)}, \theta_{j+1}^{(i-1)}, \ldots, \theta_p^{(i-1)}\right)$.

Output: *A sample $\{\boldsymbol{\theta}^{(k+1)}, \ldots, \boldsymbol{\theta}^{(n)}\}$ approximately distributed according to $p(\cdot \mid \mathcal{D})$.*

Gibbs sampling can be interpreted as a special case of the Metropolis-Hastings algorithm where in each iteration only one component (or a subset of the components) of the parameter is updated and every proposal is accepted because with the full conditional density as proposal density, the acceptance probability is always equal to one (see Lee, 2012, Chapter 9). If several components are updated at once within one step of the iterations, the procedure is also called *blocked* Gibbs sampling. Moreover, in case it is not possible to sample from the full conditional density directly, we can also use a Metropolis-Hastings draw in each step. This approach is used in Bayesian data augmentation for the inference for diffusion processes that we will describe in detail in Section 3.4.1 and investigate further in Chapter 4.

Hamiltonian Monte Carlo methods

The last class of MCMC methods that we want to describe are *Hamiltonian Monte Carlo (HMC)* methods (originally called *hybrid* Monte Carlo methods by Duane et al. (1987)). A concise description of these methods can be found in Gelman et al. (2013), whereas Betancourt (2018) gives a rather conceptual introduction, and Neal (2011) gives a detailed account. The computational cost in each iteration for HMC methods is higher than for basic Gibbs sampling or Metropolis-Hastings algorithms because HMC makes use of the derivative of the target distribution, but by that, transitions between the chain states can be generated that efficiently span the (with respect to the target distribution) important regions of the state space. By taking into account the information of the gradient, HMC avoids the random walk behavior and difficulties caused by distributions with high correlations that other MCMC methods exhibit.

Again, we want to sample from the p-dimensional distribution $\pi(\boldsymbol{\theta})$ for parameter $\boldsymbol{\theta} \in \mathbb{R}^p$. Motivated by the physical concept of Hamiltonian dynamics, HMC introduces an auxiliary

momentum variables $\boldsymbol{\rho} \in \mathbb{R}^p$ and draws from a joint density $p(\boldsymbol{\theta}, \boldsymbol{\rho}) = p(\boldsymbol{\rho} \,|\, \boldsymbol{\theta})\pi(\boldsymbol{\theta})$. The joint density defines the so-called Hamiltonian

$$H(\boldsymbol{\theta}, \boldsymbol{\rho}) = -\log p(\boldsymbol{\theta}, \boldsymbol{\rho}) = -\log p(\boldsymbol{\rho} \,|\, \boldsymbol{\theta}) - \log \pi(\boldsymbol{\theta}) = K(\boldsymbol{\theta}, \boldsymbol{\rho}) + V(\boldsymbol{\theta}) \tag{2.6}$$

that describes the total energy of the system and is equal to the sum of the kinetic energy K and the potential energy V. In HMC, the distribution of $\boldsymbol{\rho}$ is usually chosen to be independent of $\boldsymbol{\theta}$. A common choice is $\boldsymbol{\rho} \sim \mathcal{N}(\mathbf{0}_p, \boldsymbol{M})$, where $\mathcal{N}(\mathbf{0}_p, \boldsymbol{M})$ denotes the multivariate normal distribution with mean vector $\mathbf{0}_p$ and covariance matrix $\boldsymbol{M}$, and $\boldsymbol{M}$ is called the *design* or (by analogy to the physical model) *mass matrix* and often chosen to be a diagonal matrix. Thus, the kinetic energy becomes

$$K(\boldsymbol{\rho}) = \boldsymbol{\rho}^{\mathsf{Tr}} \boldsymbol{M}^{-1} \boldsymbol{\rho} / 2, \tag{2.7}$$

where $\boldsymbol{M}^{-1}$ denotes the inverse matrix of $\boldsymbol{M}$.

In each iteration of the HMC algorithm, a momentum $\boldsymbol{\rho}$ is sampled (e. g. from $\mathcal{N}(\mathbf{0}_p, \boldsymbol{M})$) and then by analogy to the physical model of the frictionless movement of a marble with position $\boldsymbol{\theta}$ and momentum $\boldsymbol{\rho}$ (describing the marble's mass and velocity) across a surface, the dynamics, i. e. the changes in position and momentum, that preserve the total energy are described by the Hamiltonian equations

$$\begin{aligned} \frac{d\rho_i}{dt} &= -\frac{\partial H}{\partial \theta_i}, \\ \frac{d\theta_i}{dt} &= \frac{\partial H}{\partial \rho_i} \end{aligned}$$

for $i = 1, \ldots, p$. With the choice of H, K, and V as in Equations (2.6) and (2.7), we have

$$\begin{aligned} \frac{d\boldsymbol{\rho}}{dt} &= -\nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta}) \quad &&= \nabla_{\boldsymbol{\theta}} \log \pi(\boldsymbol{\theta}), \\ \frac{d\boldsymbol{\theta}}{dt} &= \nabla_{\boldsymbol{\rho}} K(\boldsymbol{\rho}) \quad &&= \boldsymbol{M}^{-1} \boldsymbol{\rho}, \end{aligned} \tag{2.8}$$

where $\nabla_{\boldsymbol{x}}$ denotes the gradient with respect to $\boldsymbol{x}$. In each iteration, the Equations (2.8) are numerically integrated to obtain proposals $\boldsymbol{\theta}^*$ and $\boldsymbol{\rho}^*$. A common choice of the numerical integrator is the leap-frog method which first performs half a step for $\boldsymbol{\rho}$, then a full step for $\boldsymbol{\theta}$ using the new value of the momentum $\boldsymbol{\rho}_{\frac{1}{2}}$, and finally another half step for $\boldsymbol{\rho}$ using the new values $\boldsymbol{\rho}_{\frac{1}{2}}$ and $\boldsymbol{\theta}^*$. The used step size ϵ and the number L of steps taken in each iteration are tuning (or hyper) parameters of the algorithm. In a general HMC method, the momentum variable resulting after the L steps is negated to obtain the proposal $(\boldsymbol{\theta}^*, \boldsymbol{\rho}^*)$, but with the choice of the distribution of $\boldsymbol{\rho}$ as above, the negation does not make a difference due to the symmetry.

Finally an accept-reject step is performed analogously to the Metropolis-Hastings algorithm. That is the proposals are accepted with probability $\alpha(\boldsymbol{\theta}, \boldsymbol{\rho}, \boldsymbol{\theta}^*, \boldsymbol{\rho}^*) = \min(1, r)$ with

$$r = \frac{\pi(\boldsymbol{\theta}^*)p(\boldsymbol{\rho}^* \,|\, \boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta})p(\boldsymbol{\rho} \,|\, \boldsymbol{\theta})} = \exp\left(H(\boldsymbol{\theta}, \boldsymbol{\rho}) - H(\boldsymbol{\theta}^*, \boldsymbol{\rho}^*)\right).$$

Of the (accepted) proposals, only $\boldsymbol{\theta}^*$ needs to be saved, as a new value for the momentum is drawn right at the beginning of each iteration independent of previous values. We summarize these steps in Algorithm 2.4.

Algorithm 2.4: Hamiltonian Monte Carlo algorithm (with leap-frog integrator)

Input: *A target density $\pi(\cdot)$, an initial state $\boldsymbol{\theta}^{(0)}$, number of iterations n, mass matrix $\boldsymbol{M}$, and step size ϵ and number L of steps for numerical integration.*

In each iteration $i = 1, \ldots, n$:

Step 1 Generate $\boldsymbol{\rho} \sim \mathcal{N}(\mathbf{0}_p, \boldsymbol{M})$ and set $\boldsymbol{\theta}^* \leftarrow \boldsymbol{\theta}^{(i-1)}$ and $\boldsymbol{\rho}^* \leftarrow \boldsymbol{\rho}$.

Step 2 Repeat L leap-frog steps by setting:

$$\begin{aligned}
\boldsymbol{\rho}_{\frac{1}{2}} &\leftarrow \boldsymbol{\rho}^* + \frac{1}{2}\epsilon \, \nabla_{\boldsymbol{\theta}} \log \pi(\boldsymbol{\theta}^*) \\
\boldsymbol{\theta}^* &\leftarrow \boldsymbol{\theta}^* + \epsilon \, M^{-1} \boldsymbol{\rho}_{\frac{1}{2}} \\
\boldsymbol{\rho}^* &\leftarrow \boldsymbol{\rho}_{\frac{1}{2}} + \frac{1}{2}\epsilon \, \nabla_{\boldsymbol{\theta}} \log \pi(\boldsymbol{\theta}^*)
\end{aligned}$$

Step 3 Accept $\boldsymbol{\theta}^*$ as $\boldsymbol{\theta}^{(i)}$ with probability

$$\alpha(\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\rho}, \boldsymbol{\theta}^*, \boldsymbol{\rho}^*) = \min\left[1, \exp\left(H(\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\rho}) - H(\boldsymbol{\theta}^*, \boldsymbol{\rho}^*)\right)\right],$$

if $\boldsymbol{\theta}^*$ is rejected $\boldsymbol{\theta}^{(i)} := \boldsymbol{\theta}^{(i-1)}$.

Output: *A sample $\{\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(n)}\}$ approximately distributed according to $\pi(\cdot)$.*

Two of the limitations of this general HMC algorithm are on the one hand that due to the use of the derivative with respect to the parameter, it is only suitable for continuous distributions, and on the other hand, the choice of the tuning parameters is of crucial importance to the performance of the algorithm and can be cumbersome. The tuning parameters include the mass matrix M, and step size ϵ and number L of steps for numerical integration.

An extension of HMC, the *No-U-Turn Sampler (NUTS)*, introduced by Hoffman & Gelman (2014) includes a way to automatically determine the number L of steps for numerical integration using an recursive algorithm that grows a binary tree representing leap-frog steps forward and backward in time which is stopped as soon as further steps do no longer increase the distance between a newly explored point and the original starting point (i. e. as soon as the steps start to make a U-turn).

An efficient C++ implementation of NUTS is provided in the open-source Bayesian inference package called *Stan* (Carpenter et al., 2017) which we make use of several times in this thesis through its R interface `rstan` (Stan Development Team, 2019). In Stan, the gradient of the log-posterior distribution is calculated (exactly) by reverse-mode automatic differentiation (Carpenter et al., 2015). Moreover, Stan can automatically optimize the step size ϵ to match a (user-defined) acceptance-rate target based on dual averaging as proposed by Nesterov (2009) and it also estimates the mass matrix M during a warm-up phase consisting of several stages.

HMC methods that use a fixed symmetric, positive-definite mass matrix M throughout the algorithm are also called *Euclidean* HMC. Whereas methods that adapt the matrix M to the local structure are called *Riemann manifold* HMC methods and are supposed to allow even more "efficient convergence and exploration of the target density" (Girolami & Calderhead, 2011). However, there is evidence that numerical integration can be unstable for Riemann manifold HMC methods for different reasons including the geometry of the target density as well as the implementation of the integrator; and therefore, Riemann manifold HMC methods are not included in Stan and it is rather recommended to reparameterize the considered model to improve sampling efficiency (see e. g. Betancourt, 2019, Betancourt et al., 2015).

There is a plethora of other MCMC methods each having its advantages and disadvantages depending on the structure of the considered model. Here, we have only introduced those methods that will be used in this thesis. For an overview of further methods see e. g. Brooks et al. (2011). The need for a benchmarking framework to evaluate the performance of the many different algorithms is addressed by Ballnus et al. (2017) in the context of dynamical systems modeled by ODEs. We will give a short overview of how to assess the performance of an MCMC methods based on its output in the next subsection.

2.2.3 Evaluating MCMC output

While in theory, any MCMC method (for which convergence of the transition kernel is ensured) will give a sample from the target distribution if infinitely many iterations are executed; in practice, the sample size can only be finite. We have already mentioned this as a reason to cut off a burn-in/warm-up phase at the beginning of the iterations in order to reduce the chance of bias due to the starting value. In general, the finite sample size leads to the necessity to carefully evaluate the MCMC output.

Visual inspection to asses the obtained MCMC chains is still common practice, but also several so-called convergence tests can be found in literature (see e. g. Brooks & Roberts, 1998). They aim to assess whether an MCMC chain has already converged to the target distribution. However, similar to most hypothesis tests, they are not able to proof convergence but may only

suggest that there is substantial evidence against the hypothesis that a chain has converged. A quantity that can be used to quantify the degree of convergence when several chains have been simulated is the $\hat{R}$ value. The $\hat{R}$ convergence (or rather stationarity) diagnostic compares the between- and within-chain variance for individual model parameters and other univariate quantities of interest. Assume we are considering the scalar parameter ψ for which we have simulations $\psi_{i,j}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m$ and for m chains (after discarding the warm-up iterations and then splitting each simulated chain in half) of length n. Let

$$\widehat{\mathrm{var}}^{+}(\psi \mid \mathcal{D}) = \frac{n-1}{n} W + \frac{1}{n} B \tag{2.9}$$

be an estimate for the marginal posterior variance of ψ, where the *within-sequence variance* W is defined by

$$W = \frac{1}{m} \sum_{j=1}^{m} s_j^2 \qquad \text{with} \quad s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\psi_{ij} - \bar{\psi}_{\cdot j})^2,$$

and the *between-sequence variance* B is defined by

$$B = \frac{n}{m-1} \sum_{j=1}^{m} (\bar{\psi}_{\cdot j} - \bar{\psi}_{\cdot\cdot})^2 \qquad \text{with} \quad \bar{\psi}_{\cdot j} = \frac{1}{n} \sum_{i=1}^{n} \psi_{ij} \quad \text{and} \quad \bar{\psi}_{\cdot\cdot} = \frac{1}{m} \sum_{j=1}^{m} \bar{\psi}_{\cdot j}.$$

Then, $\hat{R}$ is defined as

$$\hat{R} = \sqrt{\frac{\hat{\mathrm{var}}^{+}(\psi \mid \mathcal{D})}{W}}.$$

Due to the splitting of chains in half, $\hat{R}$ calculated in this way is also known as split-$\hat{R}$ and was suggested in Gelman et al. (2013). The value can be interpreted as the factor by which the scale of the distribution of the current simulations for ψ can be reduced by continuing the number of iterations to infinity. If chains have mixed well, $\hat{R}$ is close to 1. Gelman et al. (2013) state that values up to 1.1 are acceptable. The $\hat{R}$ reported by Stan is calculated as the maximum of a so-called rank-normalized split-$\hat{R}$ and a rank-normalized folded-split-$\hat{R}$ which was recently suggested by Vehtari et al. (2020).

Another issue in MCMC sampling is the fact that the draws are not independent but may even be highly correlated. There are two main causes of high auto-correlation within a chain: the first may be that only small steps are proposed, so the consecutive chain states are very close to each other; and the second may be a low acceptance rate (which is the proportion of proposals that is accepted) such that many consecutive chain states are even equal to each other. Unfortunately, there is usually a trade-off between large steps and high acceptance rates as large steps tend to lead to lower acceptance probability. It is important to keep in mind that such a correlated sample from the parameter posterior distribution does not contain

the same amount of information as an independent and identically distributed sample. This issue is addressed by the notion of the effective sample size (ESS). The ESS of a sample of correlated draws quantifies the size of a corresponding independent and identically distributed sample that contains the same amount information.

The ESS for a sample of scalar parameter ψ consisting of m chains each of length n (again after discarding warm-up iterations but without splitting of the chains) can be defined as

$$n_{\text{eff}} = \frac{mn}{1 + 2\sum_{t=1}^{\infty} \rho_t},$$

where ρ_t is the autocorrelation of the sequence ψ at lag t. This quantity can be approximated in different ways. Here, we give the approximation that is presented in Gelman et al. (2013) and implemented in `rstan`. The estimated autocorrelations $\hat{\rho}_t$ are computed as

$$\hat{\rho}_t = 1 - \frac{V_t}{2\widehat{\text{var}}^{+}(\psi \mid \mathcal{D})}$$

for $t = 1, \ldots, T$ and, where the estimate $\widehat{\text{var}}^{+}$ for the marginal posterior variance is calculated as in (2.9) and the variogram V_t at lag t is calculated as

$$V_t = \frac{1}{m(n-t)} \sum_{j=1}^{m} \sum_{i=t+1}^{n} (\psi_{i,j} - \psi_{i-t,j})^2.$$

The maximal considered lag T is chosen to be the first odd positive integer for which $\hat{\rho}_{T+1} + \hat{\rho}_{T+2}$ is negative and finally, the ESS is approximated by

$$\hat{n}_{\text{eff}} = \frac{mn}{1 + 2\sum_{t=1}^{T} \hat{\rho}_t}.$$

Gelman et al. (2013) recommend that a minimum ESS of 10 per simulated chain is achieved. The between-chain information is taken into account in the calculation of $\hat{n}_{\text{eff}}$ by including the term $\widehat{\text{var}}^{+}(\psi \mid \mathcal{D})$. Thus, the ESS is affected when we try to sample from multi-modal distributions. In fact, in the case of well-separated modes and each chain sampling only from one of these modes, the ESS roughly equals to the number chains divided by the number of modes.

Vats et al. (2019) also introduced a way to calculate a multivariate ESS and provide an implementation in the R package `mcmcse` (Flegal et al., 2020).

Finally, we would like to point out that the common practice of thinning the output chains, i. e. only keeping every l^{th} element and discarding the rest of the chain before further analysis, in order to reduce autocorrelation within the sample is in fact advised against by many MCMC

experts (see e. g. Geyer, 1992, Maceachern & Berliner, 1994). Link & Eaton (2012) state that "Thinning is often unnecessary and always inefficient, reducing the precision with which features of the Markov chain are summarised." The few reasons that justify a use of thinning include on the one hand a limitation in memory or storage capacity which nowadays will rarely be the case. And on the other hand, if generating further samples is less computationally costly than the post-processing of the samples, thinning may also be justified. Owen (2017) investigates how to determine an optimal thinning frequency in such a case.

We describe some further diagnostics specific to HMC and NUTS in Appendix A.3.1.

2.3 Parameter identifiability

Another important aspect when considering parameter inference problems is the concept of parameter identifiability which is about the question whether the parameters of a model can be estimated from a given data set or type. There are two notions with respect to identifiability (Raue et al., 2009): *Structural identifiability* exclusively considers the structure of the model including the model of the observations (sometimes called the observable map) and answers the question whether the parameters can be uniquely determined if we are given perfect data, i. e. an infinite amount of data observed without measurement error and continuously in time. Whereas *practical identifiability* is concerned with the question whether the parameters can be determined from a specific data set (which is always finite and usually subject to measurement error).

Let $\mathcal{M}_\theta$ be a dynamical model parameterized by a p-dimensional parameter $\theta \in \Theta$ from an open set $\Theta \in \mathbb{R}^p$ and let $\Psi(\theta, t)$ be the "output" of $\mathcal{M}_\theta$ at time $t \geq t_0$. For a deterministic model, $\Psi(\theta, t)$ is a (vector-valued) function of the parameter θ for every time point $t \geq t_0$ and is also known as the observable(s). For a stochastic model, $\Psi(\theta, t)$ can be thought of as the probability measure induced by the random variable of the observable components of the process state at time point t. Then, we can define the following notions analogously to Chiș et al. (2011a) for the deterministic case and loosely following Reiersøl (1950) and Rothenberg (1971) for the stochastic case.

Definition 2.9. *A parameter component θ_i, $i = 1, \ldots, p$, is* structurally globally *(or* uniquely*)* identifiable *if for any $\theta, \theta' \in \Theta$, it holds that*

$$\Psi(\theta, t) = \Psi(\theta', t) \quad \Longrightarrow \quad \theta = \theta' \qquad \text{for all } t \geq t_0.$$

Definition 2.10. *A parameter component θ_i, $i = 1, \dots, p$, is* structurally locally identifiable *if for any $\theta \in \Theta$, there exists a neighborhood N_θ such that*

$$\theta' \in N_\theta \;\; \textit{and} \;\; \Psi(\theta, t) = \Psi(\theta', t) \quad \Longrightarrow \quad \theta = \theta' \qquad \textit{for all } t \geq t_0.$$

Definition 2.11. *A parameter component θ_i, $i = 1, \dots, p$, is* structurally non-identifiable *if for any $\theta \in \Theta$, there exists no neighborhood N_θ such that*

$$\theta' \in N_\theta \;\; \textit{and} \;\; \Psi(\theta, t) = \Psi(\theta', t) \quad \Longrightarrow \quad \theta = \theta' \qquad \textit{for all } t \geq t_0.$$

For ODE models, structural (as well as practical) identifiability analysis is a common step in the modeling process and design of experiments (Brouwer et al., 2017, Hengl et al., 2007, Janzén et al., 2016, Raue et al., 2010). And there exist several software tools to assess structural identifiability for ODE models such as DAISY that implements a differential algebra approach (Bellu et al., 2007), GenSSI that implements a generating series approach (Chiș et al., 2011b), and its extension GenSSI 2.0 that also allows for multi-experiment structural identifiability analysis (Ligon et al., 2017). In contrast with that, there exists hardly any literature on structural identifiability analysis for SDE models which is probably due to the complex nature of the problem. One would have to consider and compare infinite-dimensional probability distributions, and in addition to that, the distributions are not tractable for most models. One approach to structural identifiability analysis for SDE models has recently been suggested by Browning et al. (2020). It determines the moment equations of the SDE and uses them as a surrogate model which consists of ODEs. Thus, the tools for ODE models as mentioned above can be applied. Also, Komorowski et al. (2011) consider identifiability analysis for the LNA in the context of maximum likelihood estimation; however, this approach is not transferable to general SDEs.

Practical identifiability "depends more stongly on the inferential framework and is less clearly defined in the literature" (Simpson et al., 2020). It is concerned with the question whether for a given level of credibility or confidence, we can determine a finite or rather sufficiently narrow interval estimate for a parameter given real and thus imperfect data (i. e. a finite dataset which is usually subject to measurement error). Since we apply Bayesian inference methods, we are interested in *credible intervals (CIs)* and introduce them analogously to Held & Sabanés Bové (2020, Chapter 6):

For a given credible level $\alpha \in (0, 1)$, a credible interval $C_i^\alpha \subset \mathbb{R}$ for parameter component θ_i, $i = 1, \dots, p$, is an interval such that

$$\int_{\{\theta \in \Theta \,|\, \theta_i \in C_i^\alpha\}} p(\theta \,|\, \mathcal{D}) \, \mathrm{d}\theta = 1 - \alpha, \tag{2.10}$$

whereby C_i^α is not uniquely determined.

There are different approaches to calculate credible (or confidence) intervals. A Bayesian counterpart of the profile likelihood calculation from the context of maximum likelihood estimation (Raue et al., 2009) is the calculation of the profile posterior

$$\mathrm{PP}(\theta_i \mid \mathcal{D}) = \max_{\theta_{j \neq i}} [p(\theta \mid \mathcal{D})],$$

where for a given value of parameter component θ_i, $i = 1, \ldots, p$, the (unnormalized) posterior density $p(\theta \mid \mathcal{D})$ is maximized with respect to all other parameter components (Raue et al., 2013). However, for SDE models, it is not generally possible to calculate the profile posterior analytically, since the likelihood and thus the posterior density are usually not tractable.

An alternative is the calculation of credible intervals based on an MCMC sample (Hines et al., 2014, Simpson et al., 2020). There are different ways how to calculate CIs from a sample (McElreath, 2016, Chapter 3.2). Two commonly used ways are the following: For a posterior sample $\theta^1, \ldots, \theta^n$, the simplest approach to calculate the CI of parameter component θ_i, $i = 1, \ldots, p$, is to (implicitly) marginalize over all other parameter components and set

$$C_i^\alpha = \{\theta_i \mid \theta_i^{(\alpha/2)} \leq \theta_i \leq \theta_i^{(1-\alpha/2)}\},$$

where $\theta_i^{(\beta)}$ denotes the β-quantile of the sample. Another approach is to arrange the sample in descending order with respect to the corresponding posterior value $p(\theta^j \mid \mathcal{D})$, $j = 1, \ldots, n$, determine the $(1-\alpha)$-quantile $q_{1-\alpha}$ of the reordered posterior values, and then calculate the range of all values θ_i^k such that $p(\theta_i^k \mid \mathcal{D}) > q_{1-\alpha}$ as the CI C_i^α. The CI obtained in the latter way corresponds to a highest probability density interval (HPDI).

A parameter is considered practically identifiable if the obtained CI is sufficiently tight, but there is no general quantitative rule what "sufficiently tight" means. Moreover, Raue et al. (2013) demonstrate that the CIs obtained from the profile posterior do not necessarily agree with the CIs obtained based on MCMC samples in the presence of parameter non-identifiability. The MCMC CIs may indicate non-identifiabilities where the profile posterior CIs do not and one has to carefully check the MCMC diagnostics to ensure meaningful results. Then again, for the SDE models as we consider them in this thesis, only inference methods based on MCMC sampling are feasible as will be explained in detail in the next chapter and carefully assessing MCMC diagnostics is a prerequisite for meaningful inference in any case.

Chapter 3

Itô diffusion processes

In this chapter, we define what Itô diffusion processes are, give some of their properties and some examples, and show how they can be approximated and how parameter inference can be performed for them. Parts of this chapter, in particular Section 3.4.1, are similar or identical to the following article:

Pieschner, S. & Fuchs, C. (2020). Bayesian inference for diffusion processes: using higher-order approximations for transition densities. *Royal Society Open Science*, 7(10), 200270.

3.1 Definition and basic properties

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a given complete probability space with sample space Ω, σ-algebra $\mathcal{F}$, and probability measure $\mathbb{P}$ defined on $(\Omega, \mathcal{F})$ and let this space be equipped with a filtration $\mathbb{F} = \{\mathcal{F}_t\}_{t\in[0,T]}$ of σ-fields contained in $\mathcal{F}$.

In this thesis, we consider parametric time-homogeneous Itô diffusion processes and inference for such processes which we will simply call diffusion processes. A d-dimensional *time-homogeneous Itô diffusion process* $(\boldsymbol{X}_t)_{t\geq 0}$ is a stochastic process that fulfills the following stochastic differential equation (SDE):

$$\mathrm{d}\boldsymbol{X}_t = \boldsymbol{\mu}\left(\boldsymbol{X}_t, \boldsymbol{\theta}\right) \mathrm{d}t + \boldsymbol{\sigma}\left(\boldsymbol{X}_t, \boldsymbol{\theta}\right) \mathrm{d}\boldsymbol{B}_t, \qquad \boldsymbol{X}_0 = \boldsymbol{x}_0, \tag{3.1}$$

with state space $\mathcal{X} \subseteq \mathbb{R}^d$, starting value $\boldsymbol{x}_0 \in \mathcal{X}$, and an r-dimensional Brownian motion $(\boldsymbol{B}_t)_{t\geq 0}$. The model parameter $\boldsymbol{\theta} \in \Theta$ is from an open set $\Theta \subseteq \mathbb{R}^p$. The function $\boldsymbol{\mu} : \mathbb{R}^d \times \Theta \to \mathbb{R}^d$ is usually called the drift coefficient and $\boldsymbol{\sigma} : \mathbb{R}^d \times \Theta \to \mathbb{R}^{d\times r}$ the diffusion

coefficient. Equation (3.1) is a symbolic way of writing the stochastic integral equation

$$\boldsymbol{X}_t = \boldsymbol{x}_0 + \int_0^t \boldsymbol{\mu}\left(X_s, \boldsymbol{\theta}\right) \,\mathrm{d}s + \int_0^t \boldsymbol{\sigma}\left(\boldsymbol{X}_s, \boldsymbol{\theta}\right) \,\mathrm{d}\boldsymbol{B}_s \quad \text{for all } t \geq 0 \quad \mathbb{P}\text{-almost surely,}$$

where the first integral is an ordinary Riemann integral and the second integral is a stochastic integral in the Itô sense. In the remainder of this section, we omit the dependence of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ on the parameter $\boldsymbol{\theta}$ and focus on some general properties and results for diffusion processes. More elaborate and general introductions to SDEs can be found e. g. in Øksendal (2003), Klebaner (2005), Fuchs (2013), and Braumann (2019).

The following theorem provides conditions that ensure the existence and uniqueness of a solution for SDE (3.1).

Theorem 3.1 (Existence and uniqueness theorem, (Øksendal, 2003))**.** *Let $T > 0$ and $\boldsymbol{\mu} : \mathbb{R}^d \to \mathbb{R}^d$, $\boldsymbol{\sigma} : \mathbb{R}^d \to \mathbb{R}^{d\times r}$ be measurable functions satisfying*

$$\|\boldsymbol{\mu}(\boldsymbol{x}) - \boldsymbol{\mu}(\boldsymbol{y})\| + \|\boldsymbol{\sigma}(\boldsymbol{x}) - \boldsymbol{\sigma}(\boldsymbol{y})\| \leq C_1\|\boldsymbol{x} - \boldsymbol{y}\|, \qquad \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d \tag{3.2}$$

for some positive constant C_1, where $\|\boldsymbol{\sigma}\|^2 = \sum |\sigma_{ij}|^2$.
Let $\boldsymbol{Z}$ be a random variable which is independent of the σ-algebra $\mathcal{F}_\infty^{(r)}$ generated by the Brownian motion $\{\boldsymbol{B}_s\}_{s\geq 0}$ and such that $\mathbb{E}[\|Z\|^2] < \infty$.
Then the SDE

$$\mathrm{d}\boldsymbol{X}_t = \boldsymbol{\mu}(\boldsymbol{x})\,\mathrm{d}t + \boldsymbol{\sigma}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{B}_t, \qquad 0 \leq t \leq T,\, \boldsymbol{X}_0 = \boldsymbol{Z}$$

has a unique continuous solution $\boldsymbol{X}_t(\omega)$ with the property that $\boldsymbol{X}_t(\omega)$ is adapted to the filtration $\mathcal{F}_t^Z$ generated by $\boldsymbol{Z}$ and $\{\boldsymbol{B}_s\}_{0\leq s\leq t}$ and $\mathbb{E}\left[\int_0^T \|\boldsymbol{X}_t\|^2 \mathrm{d}t\right] < \infty$.

The solution $\boldsymbol{X}_t$ is called a *strong solution* as it is some functional $F(t, (\boldsymbol{B}_s, s \leq t))$ of a given Brownian motion $\boldsymbol{B}_t$. The *Lipschitz condition* (3.2) ensures the (pathwise) uniqueness of the solution, i. e. given two continuous solutions $\boldsymbol{X}_t$ and $\boldsymbol{X}_t^*$, we have

$$\mathbb{P}\left(\sup_{0\leq t\leq T} |\boldsymbol{X}_t - \boldsymbol{X}_t^*| = 0\right) = 1.$$

For our case of time-homogeneous coefficient functions $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$, Condition (3.2) also implies the *linear growth bound* which ensures that the solution does not explode on $[0, T]$. The linear growth bound for time-dependent coefficient functions $\tilde{\boldsymbol{\mu}}(\boldsymbol{x}, t)$ and $\tilde{\boldsymbol{\sigma}}(\boldsymbol{x}, t)$ is fulfilled if

$$\|\tilde{\boldsymbol{\mu}}(\boldsymbol{x}, t)\| + \|\tilde{\boldsymbol{\sigma}}(\boldsymbol{x}, t)\| \leq C_2(1 + \|\boldsymbol{x}\|), \qquad \boldsymbol{x} \in \mathbb{R}^d,\, t \in [0, T]$$

for some positive constant C_2 (cf. Øksendal, 2003, p. 70).

Lipschitz continuity and linear growth bounds are the standard conditions to ensure existence and uniqueness of the solution. For several of the models which are studied in this thesis, however, the Lipschitz condition does not hold due to the occurrence of the square root function in the diffusion coefficient. Consequently, we need to consider weaker assumptions that lead to existence and uniqueness. For this reason, let us define the notion of weak solutions.

Definition 3.2. *Given the coefficient functions $\boldsymbol{\mu} : \mathbb{R}^d \to \mathbb{R}^d$ and $\boldsymbol{\sigma} : \mathbb{R}^d \to \mathbb{R}^{d\times r}$, the process $\tilde{\boldsymbol{X}}_t$ is a called a* weak solution *if there exist a probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$, a filtration $\{\tilde{\mathcal{F}}_t\}_{t\geq 0}$, and a Brownian motion $\tilde{\boldsymbol{B}}_t$ with respect to $\tilde{\mathcal{F}}_t$ such that*

$$\tilde{\boldsymbol{X}}_t = \tilde{\boldsymbol{X}}_0 + \int_0^t \boldsymbol{\mu}\left(\tilde{\boldsymbol{X}}_s, \boldsymbol{\theta}\right) \mathrm{d}s + \int_0^t \boldsymbol{\sigma}\left(\tilde{\boldsymbol{X}}_s, \boldsymbol{\theta}\right) \mathrm{d}\tilde{\boldsymbol{B}}_s,$$

where $\tilde{\boldsymbol{X}}_t$ is $\tilde{\mathcal{F}}_t$-adapted and has continuous paths with probability one.

There are a number of theorems about the existence of weak solutions. Ikeda & Watanabe (1981, Theorem 2.3, p. 159) ensure the existence of weak solutions if the coefficient functions $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are continuous. Moreover, their following theorem gives conditions to ensure existence of weak solutions with finite second moments and that do not explode.

Theorem 3.3 (Ikeda & Watanabe (1981, Theorem 2.4, p. 163))**.** *If $\boldsymbol{\mu}(\boldsymbol{x})$ and $\boldsymbol{\sigma}(\boldsymbol{x})$ are continuous and satisfy the condition*

$$\|\boldsymbol{\mu}(\boldsymbol{x})\|^2 + \|\boldsymbol{\sigma}(\boldsymbol{x})\|^2 \leq C_3\left(1 + \|\boldsymbol{x}\|^2\right)$$

for some positive constant C_3, then for any solution of Equation (3.1) with $\mathbb{E}\left(\|\boldsymbol{X}(0)\|^2\right) < \infty$, we have $\mathbb{E}\left(\|\boldsymbol{X}(t)\|^2\right) < \infty$ for all $t > 0$. Thus, $\boldsymbol{X}(t)$ is almost surely non-explosive.

Remark 3.1. *Weak existence and pathwise uniqueness imply strong existence: If $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$ are weak solutions defined on the same probability space with the same initial condition and the same Brownian motion and pathwise uniqueness holds, i. e.*

$$\mathbb{P}\{\boldsymbol{X}_t = \boldsymbol{Y}_t \text{ for all } t \geq 0\} = 1,$$

then $\boldsymbol{X}_t$ (and $\boldsymbol{Y}_t$) are also the strong solution. This is proved in Karatzas & Shreve (1998, Chapter 5.3.D).

In the case of one-dimensional processes, pathwise uniqueness can be ensured based on the following corollary where $\boldsymbol{\sigma}$ is called *Hölder continuous with exponent* α if

$$\|\boldsymbol{\sigma}(\boldsymbol{x}) - \boldsymbol{\sigma}(\boldsymbol{y})\| \leq C_4\|\boldsymbol{x} - \boldsymbol{y}\|^\alpha, \qquad \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$$

for some positive constant C_4.

Corollary 3.4 (Ikeda & Watanabe (1981), Corollary, p. 168)**.** *If $\boldsymbol{\mu}$ is Lipschitz continuous and $\boldsymbol{\sigma}$ is Hölder continuous with exponent $1/2$, then the pathwise uniqueness of solutions holds for Equation (3.1) in the case $d = 1$.*

In the following, we list some basic properties of diffusion processes. In this thesis, we do not distinguish between stochastic processes which are stochastic modifications and assume that the considered processes are separable and continuous in t.

The *quadratic variation* of the sample paths of an Itô diffusion process $(\boldsymbol{X}_t)_{t\geq 0}$ on a time interval $[s,t]$ is

$$[\boldsymbol{X},\boldsymbol{X}]_{[s,t]} = \lim_{\delta_n\to 0}\sum_{i=1}^{n}\left(\boldsymbol{X}_{t_i^{(n)}} - \boldsymbol{X}_{t_{i-1}^{(n)}}\right)\left(\boldsymbol{X}_{t_i^{(n)}} - \boldsymbol{X}_{t_{i-1}^{(n)}}\right)^{\mathsf{Tr}} = \int_s^t \boldsymbol{\sigma}(\boldsymbol{X}_u)\boldsymbol{\sigma}(\boldsymbol{X}_u)^{\mathsf{Tr}}\,\mathrm{d}u, \tag{3.3}$$

where the limit is in probability taken over partitions $s = t_0^{(n)} < t_1^{(n)} < \ldots < t_n^{(n)} = t$ with $\delta_n = \max_{1\leq i\leq n}\left(t_i^{(n)} - t_{i-1}^{(n)}\right)$.

Moreover, the diffusion process $\boldsymbol{X}_t$ satisfies the *Markov property*, i. e. for any $0 \leq s \leq t \leq T$ and any Borel set $B \in \mathcal{B}(\mathcal{X})$, we have

$$\mathbb{P}\left(\boldsymbol{X}_t \in B \,|\, \boldsymbol{X}_u, 0 \leq u \leq s\right) = \mathbb{P}\left(\boldsymbol{X}_t \in B \,|\, \boldsymbol{X}_s\right). \tag{3.4}$$

The Markov property signifies that conditioned on the present state of the process, future states are independent of the past. Regarded as a Markov process, $\boldsymbol{X}_t$ can also be characterized by its initial distribution $\mathbb{P}(\boldsymbol{X}_0)$ and the transition probability distributions which are conditional distributions defined by the transition probability

$$\mathbb{P}\left(\boldsymbol{X}_t \in B \,|\, \boldsymbol{X}_s = \boldsymbol{x}\right) = \int_B p(s,\boldsymbol{x};t,\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}$$

for $0 \leq s < t$ and $B \in \mathcal{B}(\mathcal{X})$. The transition probability is the probability that the process will transition from state $\boldsymbol{x} \in \mathcal{X}$ at time s to state $\boldsymbol{y} \in \mathcal{X}$ at time $t > s$. $p(s,\boldsymbol{x};t,\boldsymbol{y})$ denotes the probability density function of the transition probability distribution and is called the *transition density*. For $s = t$, we define $p(s,\boldsymbol{x};s,\boldsymbol{y}) := \delta(\boldsymbol{y}-\boldsymbol{x})$, where δ is the Dirac delta function.

The Itô integral and thus also Itô diffusion processes do not adhere to the rules of classical calculus. Instead, the following theorem states the stochastic counterpart of the chain rule from classical calculus which is known as *Itô formula*. The formulation of the Itô formula

specific for Itô diffusion processes as we state it here follows directly from the general Itô formula as stated in Øksendal (2003, Chapter 4.2).

Theorem 3.5 (Itô formula)**.** *Let $\boldsymbol{X}_t$ be a d-dimensional Itô diffusion process described by an SDE as in (3.1). Let $g(t, \boldsymbol{x}) = (g_1(t, \boldsymbol{x}), \ldots, g_q(t, \boldsymbol{x}))$ be a map from $[0, T] \times \mathbb{R}^d$ into $\mathbb{R}^q$ with continuous first-order partial derivatives in t and continuous first- and second-order partial derivatives in $\boldsymbol{x}$. Then the process*

$$\boldsymbol{Y}(t, \omega) = g(t, \boldsymbol{X}_t)$$

is an Itô process whose k^{th} component $\boldsymbol{Y}^{(k)}$ is given by

$$\begin{aligned} \mathrm{d}\boldsymbol{Y}^{(k)} &= \frac{\partial g_k}{\partial t}(t, \boldsymbol{X})\,\mathrm{d}t + \sum_{i=1}^{q} \frac{\partial g_k}{\partial x^{(i)}}(t, \boldsymbol{X})\,\mathrm{d}\boldsymbol{X}^{(i)} + \frac{1}{2}\sum_{i=1}^{q}\sum_{j=1}^{q} \frac{\partial^2 g_k}{\partial x^{(i)} \partial x^{(j)}}(t, \boldsymbol{X})\,\mathrm{d}\boldsymbol{X}^{(i)} \cdot \mathrm{d}\boldsymbol{X}^{(j)}, \\ &= \left(\frac{\partial g_k}{\partial t}(t, \boldsymbol{X}) + \boldsymbol{\mu}(\boldsymbol{X})^{Tr} \nabla g_k(t, \boldsymbol{X}) + \frac{1}{2}\mathrm{trace}\left(\boldsymbol{\sigma}(\boldsymbol{X})\boldsymbol{\sigma}(\boldsymbol{X})^{Tr} \nabla \left(\nabla g_k(t, \boldsymbol{X}) \right) \right) \right) \mathrm{d}t \\ &\quad + \left(\nabla g_k(t, \boldsymbol{X}) \right)^{Tr} \boldsymbol{\sigma}(\boldsymbol{X})\,\mathrm{d}\boldsymbol{B}_t, \end{aligned} \tag{3.5}$$

where ∇g_k denotes the gradient of g_k with respect to the components of $\boldsymbol{x}$ and $\mathrm{d}\boldsymbol{X}^{(i)} \cdot \mathrm{d}\boldsymbol{X}^{(j)}$ is computed according to the rules $\mathrm{d}\boldsymbol{B}^{(i)} \cdot \mathrm{d}t = \mathrm{d}t \cdot \mathrm{d}\boldsymbol{B}^{(j)} = (\mathrm{d}t)^2 = 0$ and $\mathrm{d}\boldsymbol{B}^{(i)} \cdot \mathrm{d}\boldsymbol{B}^{(j)} = \delta_{ij}\,\mathrm{d}t$ with δ_{ij} denoting the Kronecker delta.

3.2 Example models

In this section, we briefly introduce two simple, well-known examples of diffusion processes that we use as illustrative examples and as benchmark models in Chapter 4. The first example is the *geometric Brownian motion (GBM)* which is described by the following one-dimensional SDE:

$$\mathrm{d}X_t = \alpha X_t\,\mathrm{d}t + \sigma X_t\,\mathrm{d}B_t, \quad X_0 = x_0, \tag{3.6}$$

with state space $\mathcal{X} = \mathbb{R}_+$, starting value $x_0 \in \mathcal{X}$ and the two-dimensional parameter $\boldsymbol{\theta} = (\alpha, \sigma)^T$, where $\alpha \in \mathbb{R}$ and $\sigma \in \mathbb{R}_+$, $\mathbb{R}_+$ being the set of all strictly positive real numbers. The GBM is famous in finance where it models asset prices and is also known as the Black-Scholes-model (see e. g. Black & Scholes, 1973, Merton, 1973). Braumann (2019) also considers the GBM to model population growth. For us, the GBM is especially suitable as a benchmark model because it has an explicit solution. The stochastic process

$$X_t = x_0 \exp\left(\left(\alpha - \frac{1}{2}\sigma^2 \right) t + \sigma B_t \right)$$

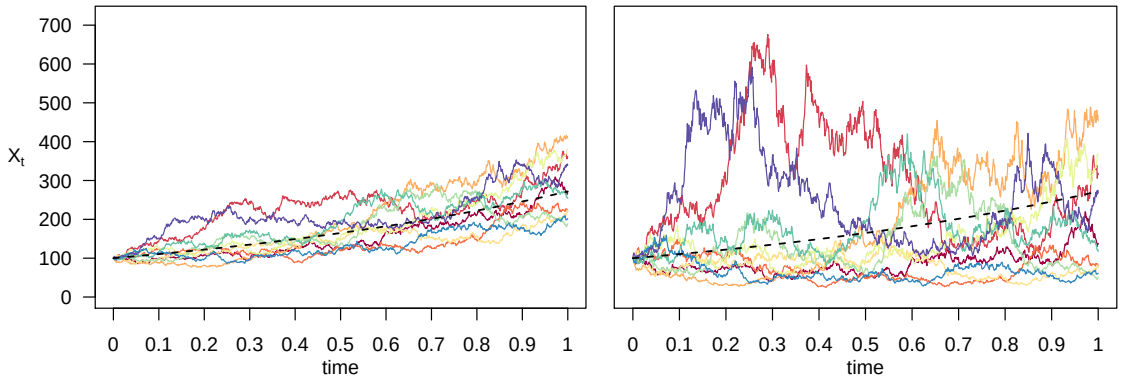
fulfills (3.6) for all $t \geq 0$. Hence, the multiplicative increments of the GBM are log-normally distributed as follows:

$$\frac{X_t}{X_s} \sim \mathcal{LN}\left(\left(\alpha - \frac{1}{2}\sigma^2\right)(t-s), \sigma^2 (t-s)\right)$$

for $t \geq s \geq 0$, and the transition density $P(X_t = y \,|\, X_s = x)$ which we denote by $p(s, x; t, y)$ is explicitly known as

$$p(s,x;t,y) = \frac{1}{\sqrt{2\pi(t-s)}\sigma y} \exp\left(-\frac{\left(\log y - \log x - \left(\alpha - \frac{1}{2}\sigma^2\right)(t-s)\right)^2}{2\sigma^2(t-s)}\right). \tag{3.7}$$

A derivation of the solution of the GBM and its transition density can be found in Iacus (2008). Figure 3.1 presents realizations of the GBM for two different parameter combinations.



(a) $\alpha = 1$, $\sigma^2 = 0.25$ **(b)** $\alpha = 1$, $\sigma^2 = 2$

Figure 3.1: Ten trajectories of the GBM described by SDE (3.6) with starting value $X_0 = 100$ and for different parameter combinations. The dashed black line represents the expected value of the GBM solution $\mathbb{E}[X_t] = X_0 \exp(\alpha t)$.

In some contexts, one considers the logarithm of the GBM, $\log X_t$, which is simply a normally distributed random variable for fixed t, with corresponding SDE

$$\mathrm{d}(\log X_t) = \left(\alpha - \frac{1}{2}\sigma^2\right)\mathrm{d}t + \sigma\,\mathrm{d}B_t, \quad \log X_0 = \log x_0. \tag{3.8}$$

However, we do not employ this transformation here because of the constant diffusion function in (3.8). For the log-transformed GBM, the approximation methods that we will introduce in the next section and that we wish to compare in Chapter 4 would yield an identical approximation.

Our second example model is the *Cox-Ingersoll-Ross (CIR) process* which fulfills the one-dimensional SDE

$$\mathrm{d}X_t = \alpha\left(\beta - X_t\right)\,\mathrm{d}t + \sigma\sqrt{X_t}\,\mathrm{d}B_t, \quad X_0 = x_0, \tag{3.9}$$

with starting value $x_0 \in \mathbb{R}_+$ and parameters $\alpha, \beta, \sigma \in \mathbb{R}_+$. If $2\alpha\beta > \sigma^2$, the process is strictly positive (i. e. $\mathcal{X} = \mathbb{R}_+$) otherwise it is non-negative (i. e. $\mathcal{X} = \mathbb{R}_0$). This model was originally introduced by Feller (1951a,b) to model population growth. Later, Cox, Ingersoll, and Ross used it to model the evolution of short-term interest rates in Cox et al. (1985) and it became well-known in finance under their names. The CIR process is also used as the volatility process in the Heston model introduced by Heston (1993). Etchegaray & Meunier (2019) employ the CIR process to model bond dynamics in cell adhesion.

The transition density of the CIR process is explicitly known as

$$p\left(s, x; t, y\right) = c\left(\frac{v}{u}\right)^{\frac{\eta}{2}} e^{-(u+v)} I_\eta(2\sqrt{uv}) \tag{3.10}$$

for $t > s \geq 0$, where

$$c = \frac{2\alpha}{\sigma^2\left(1 - e^{-\alpha(t-s)}\right)}, \quad u = cxe^{-\alpha(t-s)}, \quad v = cy, \quad \eta = \frac{2\alpha\beta}{\sigma^2} - 1,$$

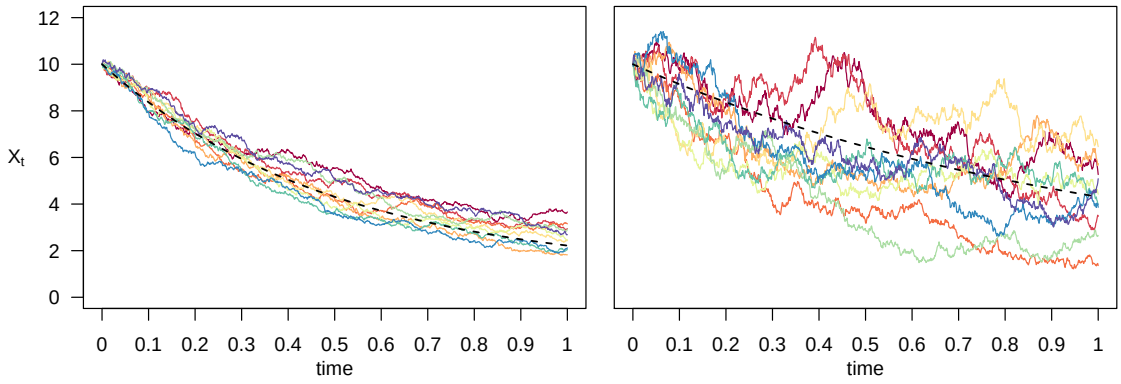
and I_η denotes the modified Bessel function of the first kind of order η, i. e.

$$I_\eta(z) = \sum_{k=0}^{\infty} \left(\frac{z}{2}\right)^{2k+\eta} \frac{1}{k!\,\Gamma(k+\eta+1)}$$

for $z \in \mathbb{R}$, where Γ is the Gamma function (see e. g. Fuchs, 2013). A derivation of the transition density is provided in Jeanblanc et al. (2009, Chapter 6.3). There are several algorithms to generate samples from this density (see e. g. Alfonsi, 2015, Glasserman, 2003). We use the algorithm presented in Glasserman (2003, p. 124) and state it in Appendix A.1. Figure 3.2 shows realizations of the CIR process for two different parameter combinations.

3.3 Approximation of the solution of an SDE

Unlike the example models from the previous section, most SDEs do not have an analytical solution and their transition densities are not explicitly known. Instead, numerical approximation schemes are used for the solution of the SDEs. Kloeden & Platen (1992) have provided a detailed description of these methods.



(a) $\alpha = 2$, $\beta = 1$, $\sigma^2 = 0.25$ **(b)** $\alpha = 1$, $\beta = 1$, $\sigma^2 = 2$

Figure 3.2: Ten trajectories of the CIR process described by SDE (3.9) with starting value $X_0 = 10$ and for different parameter combinations. The dashed black line represents the expected value of the CIR process $\mathbb{E}[X_t] = \beta - (\beta - X_0)\exp(-\alpha t)$.

The numerical approximation schemes are evaluated through their convergence property. A discrete-time approximation $\mathbf{Y}^\Delta$ with maximum time step $\Delta > 0$ *converges strongly* to the solution $\boldsymbol{X}_T$ of a given SDE at time T if

$$\lim_{\Delta \searrow 0} \mathbb{E}\left(\left\|\boldsymbol{X}_T - \mathbf{Y}_T^\Delta\right\|\right) = 0.$$

To compare different approximation schemes, one usually considers their rates of strong convergence. A discrete-time approximation $\mathbf{Y}^\Delta$ with maximum time step $\Delta > 0$ *converges with strong order* $\gamma > 0$ to the solution $\boldsymbol{X}_T$ of a given SDE at time T if there exists a positive constant C independent of Δ and a $\Delta_0 > 0$ such that

$$\mathbb{E}\left(\left\|\boldsymbol{X}_T - \mathbf{Y}_T^\Delta\right\|\right) \leq C\Delta^\gamma$$

for all $\Delta \in (0, \Delta_0)$. Strong convergence ensures a pathwise approximation of the solution process $(\boldsymbol{X}_t)_{t\geq 0}$ of the given SDE as shown in Kloeden & Neuenkirch (2007). The higher the order of strong convergence is, the faster the mean absolute error between the approximation and the solution decreases as the maximum time step size Δ decreases.

Several of the approximation schemes are based on the *Itô-Taylor expansion*. This stochastic counterpart of the Taylor expansion is obtained by iteratively applying the Itô formula (3.5) to the coefficient functions of SDE (3.1) which are assumed to be sufficiently smooth real-valued functions satisfying a linear growth bound. For the i^{th} component of the process $\boldsymbol{X}_t$, the

expansion reads

$$\begin{aligned}\boldsymbol{X}_t^{(i)} = \boldsymbol{X}_{t_0}^{(i)} + \boldsymbol{\mu}_i(\boldsymbol{X}_{t_0}) \int_{t_0}^{t} \mathrm{d}s + \sum_{l=1}^{r} \boldsymbol{\sigma}_{il}(\boldsymbol{X}_{t_0}) \int_{t_0}^{t} \mathrm{d}\boldsymbol{B}_s^{(l)} \\ + \sum_{l=1}^{r}\sum_{q=1}^{r}\sum_{j=1}^{d} \boldsymbol{\sigma}_{jq}\left(\boldsymbol{X}_{t_0}\right) \frac{\partial \boldsymbol{\sigma}_{il}}{\partial x^{(j)}}\left(\boldsymbol{X}_{t_0}\right) \int_{t_0}^{t}\int_{t_0}^{s} \mathrm{d}\boldsymbol{B}_u^{(q)}\mathrm{d}\boldsymbol{B}_s^{(l)} + \boldsymbol{R}^{(i)}\end{aligned}$$

for $i = 1, \ldots, d$ and with remainder term $\boldsymbol{R}$ that contains further multiple Itô integrals but with non-constant integrands.

The most commonly used approximation is the *Euler(-Maruyama) scheme* which contains only the time component and the stochastic integral of multiplicity one from the Itô-Taylor expansion and was first investigated by Maruyama (1955). It can be conveniently written in vector notation and approximates the d-dimensional solution $(\boldsymbol{X}_t)_{t\geq 0}$ of an SDE by setting $\boldsymbol{Y}_0 = \boldsymbol{x}_0$ and, then, successively calculating the following:

$$\boldsymbol{Y}_{k+1} = \boldsymbol{Y}_k + \boldsymbol{\mu}\left(\boldsymbol{Y}_k\right)\Delta t_k + \boldsymbol{\sigma}\left(\boldsymbol{Y}_k\right)\Delta \boldsymbol{B}_k, \tag{3.11}$$

where $\Delta t_k = t_{k+1} - t_k$, $\Delta \boldsymbol{B}_k = \boldsymbol{B}_{t_{k+1}} - \boldsymbol{B}_{t_k}$, and $\boldsymbol{Y}_k$ is the approximation of $\boldsymbol{X}_{t_k}$ for $k = 0, 1, 2, \ldots$. If the coefficient functions $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ fulfill the Lipschitz (and the linear growth) condition, the Euler scheme has strong order of convergence $\gamma = 0.5$.

By adding another term of the Itô-Taylor expansion to Equation (3.11), one obtains the Milstein scheme that approximates the d-dimensional process $(\boldsymbol{X}_t)_{t\geq 0}$ by setting $\boldsymbol{Y}_0 = \boldsymbol{x}_0$ and, then, successively calculating for the i^{th} component:

$$\begin{aligned}\boldsymbol{Y}_{k+1}^{(i)} =& \boldsymbol{Y}_k^{(i)} + \boldsymbol{\mu}_i\left(\boldsymbol{Y}_k\right)\Delta t_k + \sum_{l=1}^{r} \boldsymbol{\sigma}_{il}\left(\boldsymbol{Y}_k\right)\Delta B_k^{(l)} \\ &+ \sum_{l=1}^{r}\sum_{q=1}^{r}\sum_{j=1}^{d} \boldsymbol{\sigma}_{jq}\left(\boldsymbol{Y}_k\right) \frac{\partial \boldsymbol{\sigma}_{il}}{\partial y^{(r)}}\left(\boldsymbol{Y}_k\right) \int_{t_k}^{t_{k+1}}\int_{t_k}^{s} \mathrm{d}\boldsymbol{B}_u^{(q)}\mathrm{d}\boldsymbol{B}_s^{(l)}\end{aligned} \tag{3.12}$$

for $k = 0, 1, \ldots$ and $i = 1, \ldots, d$.

When $\boldsymbol{\sigma}\left(\boldsymbol{Y}_k\right)$ is constant in $\boldsymbol{Y}_k$, the last term vanishes and the Milstein scheme reduces to the Euler scheme. If $\boldsymbol{\mu}$ is once continuously differentiable and $\boldsymbol{\sigma}$ is twice continuously differentiable with uniformly Lipschitz continuous derivatives, then the Milstein scheme is strongly convergent of order 1.0, which is higher than that of the Euler scheme. An illustration of this difference in the convergence rates for the simulation of SDE trajectories is presented e. g. in Bayram et al. (2018).

Note that the stated orders of strong convergence for the two schemes assume Lipschitz continuity of the coefficient functions. As already mentioned this condition is not fulfilled for several of the models considered in this thesis. Gyöngy & Rásonyi (2011) provide convergence rates for the Euler scheme under weaker assumptions. For the case of $\frac{1}{2}$-Hölder continuous diffusion functions and a discretization time step of $\Delta = T/n$, they show in their Theorem 2.1 that the Euler scheme achieves only a slower strong convergence rate of $1/\ln n$ (instead of $\Delta^{\frac{1}{2}}$).

For the CIR process, Hefter & Herzwurm (2018) prove the strong convergence of a Milstein-type approximation scheme. They state that for values X_t "away" from zero their scheme coincides with the classical Milstein scheme as described above. Moreover, the scheme has a strong rate of convergence of $n^{-(\frac{1}{2}-\epsilon)}$ for some small $\epsilon > 0$ (instead of Δ^1) which converges to zero faster than $1/\ln n$ as n tends to infinity.

To prove the strong convergence under weaker assumptions, a different representation of the Euler scheme is usually considered in the literature. As in Gyöngy & Rásonyi (2011), we define the functions $\kappa_n : [0,T] \to [0,T]$ for $n \geq 1$ by setting $\kappa_n(T) := \frac{n-1}{n}T$ and

$$\kappa_n(t) = \frac{iT}{n} \qquad \text{for } \frac{iT}{n} \leq t \leq \frac{(i+1)T}{n} \tag{3.13}$$

and for $i = 0, \ldots, n-1$. Then the Euler approximations of the solution $\boldsymbol{X}(t)$, $t \in [0,T]$, of Equation (3.1) can be defined as the solutions of

$$\mathrm{d}\boldsymbol{X}_n(t) = \boldsymbol{\mu}(\boldsymbol{X}_n(\kappa_n(t)))\,\mathrm{d}t + \boldsymbol{\sigma}(\boldsymbol{X}_n(\kappa_n(t)))\,\mathrm{d}\boldsymbol{B}(t), \qquad \boldsymbol{X}_n(0) = \boldsymbol{x}_0, \tag{3.14}$$

for each $n \geq 1$.

For the Euler approximation of a one-dimensional diffusion process, Gyöngy & Rásonyi (2011) state the following bounds that can be derived from Krylov (1980, Corollary 12, p. 86).

Lemma 3.6 (Gyöngy & Rásonyi (2011), Lemma 1.2 and Remark 1.2.)**.** *Assume that $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ satisfy the linear growth condition, i. e.*

$$\|\boldsymbol{\mu}(x)\| + \|\boldsymbol{\sigma}(x)\| \leq K(1 + \|x\|), \qquad x \in \mathbb{R},$$

with some $K > 0$. For each $p > 0$, there is $C > 0$, independent of n, such that

$$\mathbb{E}\left[\sup_{0 \leq t \leq T} |X_n(t)|^p\right] \leq C(1 + x_0)^p \tag{3.15}$$

for all n, where C is a constant depending on T, p, and K. Moreover, there is $\tilde{C} > 0$, independent of n, such that

$$\mathbb{E}\left[\sup_{0\leq t\leq T}|X_n(t) - X_n(\kappa_n(t))|^p\right] \leq \frac{\tilde{C}}{n^{p/2}} \tag{3.16}$$

for all n, where $\tilde{C}$ is a constant depending on T, p, x_0, and K.

Corollary 3.7 (Gyöngy & Rásonyi (2011) Corollary 2.3)**.** *Let $\boldsymbol{\mu}$ be Lipschitz continuous and $\boldsymbol{\sigma}$ Hölder continuous and let $x_0 \in \mathbb{R}$. Then there is a constant C depending on K, T, and $(x_0)^2$ such that for all $n \geq 2$ we have*

$$\mathbb{E}\left[\sup_{0\leq t\leq T}|X(t) - X^n(t)|\right] \leq \frac{C}{\sqrt{\ln n}}. \tag{3.17}$$

We will use these results for the proof of convergence of the Euler scheme for the application model in Section 5.3.2.

3.4 Inference for SDEs

In this thesis, we assume the drift coefficient $\boldsymbol{\mu}$ and the diffusion coefficient $\boldsymbol{\sigma}$ of SDE (3.1) to be known in parametric form and summarize methods to infer the parameter $\boldsymbol{\theta} \in \mathbb{R}^p$. A more detailed overview of these methods can be found e. g. in Sørensen (2004), Fuchs (2013), and the references therein.

Inference would be straight forward in the hypothetical case of observing a trajectory of $\boldsymbol{X} = (\boldsymbol{X}_t)_{t\geq 0}$ continuously over a time interval $[r, s]$. In this case, the components of the parameter $\boldsymbol{\theta}$ that appear in the diffusion coefficient can be directly determined from Relationship (3.3) between the quadratic variation and the diffusion coefficient. Afterwards, the (log-)likelihood function of the remaining parameter components can be constructed using Girsanov's formula and then maximized to obtain an estimate for $\boldsymbol{\theta}$. This approach is presented for linear SDEs in Le Breton (1977) and described for general SDEs in Fuchs (2013, Chapter 6.1).

However, in practice, the continuous-time observation of a process is not possible. Therefore, the inference problem needs to be considered for states $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ of $\boldsymbol{X}$ observed at discrete time points $0 = t_0 < t_1 < \ldots < t_n$. Different asymptotic observation schemes have been investigated in the literature. Assuming that the states are observed with equidistant time step Δ between consecutive observations on a time interval $[0, T]$ with $T = t_n = n\Delta$, Iacus

(2008) uses the following terms for the different schemes: In the *large-sample scheme*, the time step Δ remains fixed and T tends to infinity as $n \to \infty$. In the *high-frequency scheme*, the upper bound T of the time interval is fixed and the time step $\Delta = \Delta_n$ goes to zero as $n \to \infty$. And finally, in the *rapidly increasing design*, $\Delta = \Delta_n$ goes to zero and T increases as $n \to \infty$. While the last two schemes are more convenient from a theoretical point of view as their limit corresponds to continuous-time observations, sometimes only the large-sample scheme is realistic in practice.

If the SDE yields an analytical solution, the transition densities of the corresponding diffusion process are explicitly known and parameter estimation can be easily performed through a maximum likelihood approach, as demonstrated e. g. in Dacunha-Castelle & Florens-Zmirou (1986). Due to the Markov property of diffusion process $\boldsymbol{X}$, the likelihood function of $\boldsymbol{\theta}$ factorizes as

$$\mathcal{L}_n(\boldsymbol{\theta}) = \prod_{k=0}^{n-1} p_{\boldsymbol{\theta}}(t_k, \boldsymbol{x}_k; t_{k+1}, \boldsymbol{x}_{k+1}) \tag{3.18}$$

and can be evaluated based on the transition densities $p_{\boldsymbol{\theta}}$ known up to the parameter $\boldsymbol{\theta}$. Yet again this procedure is only of little practical use as in the majority of applications, the SDE model does not have an analytical solution and the transition densities are intractable. Thus, exact maximum likelihood estimation is usually not possible.

Instead, *approximate* (sometimes also called *quasi* or *pseudo*) *maximum likelihood estimation* can be performed by appropriately approximating the likelihood function. The most naive and at the same time straight forward approach is to approximate the transition density based on the Euler scheme (3.11). Since the Euler scheme is a linear transformation of the normally-distributed increments $\Delta \boldsymbol{B}_k \sim \mathcal{N}(\boldsymbol{0}, \Delta t_k \boldsymbol{I}_r)$ of the Brownian motion, where $\boldsymbol{I}_r$ denotes the r-dimensional identity matrix, the transition density derived from the Euler scheme is also a multivariate Gaussian density:

$$p_{\boldsymbol{\theta}}(t_k, \boldsymbol{x}_k; t_{k+1}, \boldsymbol{x}_{k+1}) \approx \phi\left(\boldsymbol{x}_{k+1} \,|\, \boldsymbol{x}_k + \boldsymbol{\mu}(\boldsymbol{x}_k, \boldsymbol{\theta})\, \Delta t_k, \boldsymbol{\sigma}(\boldsymbol{x}_k, \boldsymbol{\theta})\, \boldsymbol{\sigma}^{\mathsf{Tr}}(\boldsymbol{x}_k, \boldsymbol{\theta})\, \Delta t_k\right)$$

for $k = 0, \ldots, n-1$ and where $\phi(\boldsymbol{y} \,|\, \boldsymbol{a}, \boldsymbol{b})$ denotes the multivariate Gaussian density with mean $\boldsymbol{a} \in \mathbb{R}^d$ and covariance matrix $\boldsymbol{b} \in \mathbb{R}^{d \times d}$ evaluated at $\boldsymbol{y} \in \mathbb{R}^d$. Hence, the approximated likelihood function becomes

$$\mathcal{L}_n^{Euler}(\boldsymbol{\theta}) = \prod_{k=0}^{n-1} \phi\left(\boldsymbol{x}_{k+1} \,|\, \boldsymbol{x}_k + \boldsymbol{\mu}(\boldsymbol{x}_k, \boldsymbol{\theta})\, \Delta t_k, \boldsymbol{\sigma}(\boldsymbol{x}_k, \boldsymbol{\theta})\, \boldsymbol{\sigma}^{\mathsf{Tr}}(\boldsymbol{x}_k, \boldsymbol{\theta})\, \Delta t_k\right). \tag{3.19}$$

For the large-sample scheme (where the time step between observations is fixed), this naive approximate maximum likelihood estimator has been shown to be biased and thus inconsistent as $n \to \infty$ e. g. in Florens-Zmirou (1989) and Lo (1988). For a given dataset, the time step

and the number of observations is always fixed of course in practice; therefore, when using this fairly simple approach one has to ensure that the time step between observations is small enough. However, there are no general rules what "small enough" means.

In many applications, the time step is rather large. In this case, one has to use more sophisticated approximations for the likelihood or turn to different inference approaches. One way to analytically approximate the transition density (and hence the likelihood) was proposed by Aït-Sahalia (2002) for one-dimensional diffusion processes and extended to the multi-dimensional case by Aït-Sahalia (2008). The approach involves an expansion based on modified Hermite polynomials and the transformation of the diffusion $\boldsymbol{X}$ into a diffusion $\boldsymbol{Y}$ whose diffusion matrix is the identity matrix. This transformation into a so-called *unit diffusion*, however, is not generally possible for multi-dimensional diffusions.

The local linearization method is another approach that potentially makes use of the transformation into a unit diffusion. It has been described in Shoji & Ozaki (1998b) for one-dimensional diffusion processes and extended to the multi-dimensional case in Shoji & Ozaki (1998a). After transformation into a unit diffusion (if this is even necessary), the drift function of the resulting process is linearized; and thus, the considered diffusion process is approximated by a linear one. For linear diffusions, explicit solutions exist (cf. e. g. Kloeden & Platen, 1992, Chapter 4.2). Hence, their transition density is available and can be used as an approximation of the transition density of the original process.

Another approach approximates the likelihood by numerically solving the Kolmogorov forward equation (also called Fokker-Planck equation) that is associated with the diffusion process and that is a deterministic partial differential equation for the transition density. This approach was used and further developed e. g. in Poulson (1999), Jensen & Poulsen (2002), Hurn et al. (2007) and Lux (2013).

The simulated maximum likelihood (SML) approach on the other hand uses Monte Carlo integration to approximate the unknown transition density. The approach was suggested independently by Pedersen (1995) and Santa-Clara (1995). The idea is to use the Chapman-Kolmogorov equation to reformulate the transition density as follows

$$\begin{aligned} p_{\boldsymbol{\theta}}(s, \boldsymbol{x}; t, \boldsymbol{y}) &= \int_{\mathcal{X}} p_{\boldsymbol{\theta}}(s, \boldsymbol{x}; t-\delta, \boldsymbol{z}) p_{\boldsymbol{\theta}}(t-\delta, \boldsymbol{z}; t, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{z} \\ &= \mathbb{E}_{\boldsymbol{\theta}} \left[p_{\boldsymbol{\theta}}(t-\delta, \boldsymbol{X}_{t-\delta}; t, \boldsymbol{y}) \mid \boldsymbol{X}_s = \boldsymbol{x} \right] \end{aligned} \tag{3.20}$$

for small δ with $0 < \delta \ll t - s$. One simulates M realizations $\boldsymbol{z}_j$, $j = 1, \ldots, M$, of $\boldsymbol{X}_{t-\delta}$ conditioned on $\boldsymbol{X}_s = \boldsymbol{x}$, e. g. by dividing the interval $[s, t-\delta]$ into $N-1$ steps and forward simulating based on the Euler scheme for these smaller time steps, and then approximates the

expectation in (3.20) by calculating

$$p_{\boldsymbol{\theta}}^{M,N}(s,\boldsymbol{x};t,\boldsymbol{y}) = \frac{1}{M}\sum_{m=1}^{M} p_{\boldsymbol{\theta}}^{Euler}(t-\delta,\boldsymbol{z}_j;t,\boldsymbol{y}), \tag{3.21}$$

where $p_{\boldsymbol{\theta}}^{Euler}(t-\delta,\boldsymbol{z}_j;t,\boldsymbol{y}) = \phi\left(\boldsymbol{z}_j \,|\, \boldsymbol{y} + \boldsymbol{\mu}(\boldsymbol{z}_j,\boldsymbol{\theta})\delta, \boldsymbol{\sigma}(\boldsymbol{z}_j,\boldsymbol{\theta})\boldsymbol{\sigma}^{\mathsf{Tr}}(\boldsymbol{z}_j,\boldsymbol{\theta})\delta\right)$ is again the transition density derived from the Euler scheme. Several refinements of the SML approach have been suggested in the literature. For example, Durham & Gallant (2002) propose to employ importance sampling when approximating (3.20) by conditioning not only on the initial state $\boldsymbol{X}_s = \boldsymbol{x}$ but also on the end point $\boldsymbol{X}_t = \boldsymbol{y}$ when simulating the realizations $\boldsymbol{z}_j$. This sampling method is called the *modified-bridge* and will be discussed in more detail in the context of Bayesian data augmentation in Section 3.4.1. Durham & Gallant (2002) also investigate the use of other variance reduction techniques and the application of a higher-order Itô-Taylor expansion. Elerian (1998) suggests to approximate the transition density for sampling the realizations $\boldsymbol{z}_j$ as well as in (3.21) based on the Milstein schemes instead of the Euler scheme. We will consider the transition density derived from the Milstein scheme in detail in Section 4.1.

Several other estimation techniques have been developed that do not rely on the approximation of the likelihood function but rather match certain statistics of the diffusion model with the corresponding statistics of the observed data. An overview of the class of methods based on so-called estimating functions and the related method of generalized moments can be found e. g. in Kessler et al. (2012, Chapter 1). Gourieroux et al. (1993) suggested the use of indirect inference and Gallant & Tauchen (1996) proposed the efficient methods of moments.

With the Exact Algorithm (EA), Beskos & Roberts (2005) introduced a method to exactly simulate diffusion paths at discrete time points and Beskos et al. (2006a, 2008) further developed the EA for more general cases. However, in order for the EA to be applicable, again the transformation into a unit diffusion must be possible and further conditions need to be satisfied. A variety of statistical inference methods based on the EA have been studied e. g. by Beskos et al. (2009, 2006b) and Sermaidis et al. (2013).

Donnet & Samson (2013) also review further parameter estimation methods for diffusion models in the context of pharmacokinetic/pharmacodynamic models including several methods based on the well known Kalman filter, but also for these methods, state dependence of the diffusion coefficient is precluded.

For SDEs that are derived as the diffusion approximation of a Markov jump process (MJP) as described in Section 2.1.2, the diffusion coefficient always depends on the state variable for non-trivial models. Indeed, for many models, the diffusion coefficient components appearing in the equation for one component of the process usually depend on several of the process

components. This renders the transformation into a unit diffusion usually impossible and thus, many of the inference methods mentioned so far are rarely applicable in this context. Besides, especially in biological applications, it is commonly not possible to observe all components of a process directly and one also has to account for measurement error. Therefore, inference methods are needed that can handle latent variables and noisy observations.

Bayesian inference methods and in particular Markov chain Monte Carlo (MCMC) methods have been applied very successfully in this context, especially because they also allow for the inclusion of prior knowledge and for an assessment of the uncertainty of the parameter estimates. Assume we have observations $\boldsymbol{y}_0, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ observed at discrete time points $0 = t_0 < t_1 < \ldots < t_n$ and with

$$\boldsymbol{y}_i \sim h(\boldsymbol{y}_i \,|\, \boldsymbol{X}_{t_i}, \boldsymbol{\theta}) \qquad \text{for } i = 0, \ldots, n,$$

where h denotes the distribution of $\boldsymbol{y}_i$ conditioned on (possibly a transformed subset of the components of) the diffusion process $\boldsymbol{X}$ and on the parameter vector $\boldsymbol{\theta}$ that may include parameters for the transformation of $\boldsymbol{X}$. Moreover, the distribution can represent the case of additive or multiplicative measurement error. In order to obtain the posterior distribution of the parameter $\boldsymbol{\theta}$ given the observed data, we have to marginalize over the states of the diffusion process:

$$\pi\left(\boldsymbol{\theta} \,|\, \{\boldsymbol{y}_k\}_{k=0,\ldots,n}\right) = \int_{\mathcal{X}^{n+1}} \pi\left(\boldsymbol{\theta}, \{\boldsymbol{X}_{t_k}\}_{k=0,\ldots,n} \,|\, \{\boldsymbol{y}_k\}_{k=0,\ldots,n}\right) \mathrm{d}(\boldsymbol{X}_{t_0}, \ldots, \boldsymbol{X}_{t_n}).$$

This marginalization can be achieved by Monte Carlo integration. Therefore, we need to draw samples from the following conditional density that is reformulated using Bayes' theorem:

$$\begin{aligned}
&\pi\left(\boldsymbol{\theta}, \{\boldsymbol{X}_{t_k}\}_{k=0,\ldots,n} \,|\, \{\boldsymbol{y}_k\}_{k=0,\ldots,n}\right) \\
&\quad \propto \pi\left(\{\boldsymbol{y}_k\}_{k=0,\ldots,n} \,|\, \boldsymbol{\theta}, \{\boldsymbol{X}_{t_k}\}_{k=0,\ldots,n}\right) \pi(\boldsymbol{\theta}, \{\boldsymbol{X}_{t_k}\}_{k=0,\ldots,n}) \\
&\quad = \pi\left(\{\boldsymbol{y}_k\}_{k=0,\ldots,n} \,|\, \boldsymbol{\theta}, \{\boldsymbol{X}_{t_k}\}_{k=0,\ldots,n}\right) \pi(\{\boldsymbol{X}_{t_k}\}_{k=0,\ldots,n} \,|\, \boldsymbol{\theta})\, p(\boldsymbol{\theta}) \\
&\quad = \left(\prod_{k=0}^{n} h(\boldsymbol{y}_i \,|\, \boldsymbol{X}_{t_i}, \boldsymbol{\theta})\right) \left(\prod_{k=0}^{n-1} p_{\boldsymbol{\theta}}(t_k, \boldsymbol{X}_{t_k}; t_{k+1}, \boldsymbol{X}_{t_{k+1}})\right) p_{\boldsymbol{\theta}}(\boldsymbol{X}_{t_0})\, p(\boldsymbol{\theta}),
\end{aligned} \tag{3.22}$$

where $p(\boldsymbol{\theta})$ denotes the density of the prior distribution for the parameter $\boldsymbol{\theta}$. The transition density $p_{\boldsymbol{\theta}}$ of the diffusion process $\boldsymbol{X}_t$ in the second factor of (3.22) needs to be approximated as mentioned before. When using the Euler scheme, one has to carefully check whether the time steps between observations are sufficiently small. Even in the case where the Euler scheme is eligible, drawing samples from (3.22) is a computationally intense task. This is due to the high dimension of the conditional density which is equal to $p + d(n+1)$, where p is the dimension of the parameter vector $\boldsymbol{\theta}$, d is the dimension of the diffusion process $\boldsymbol{X}$,

and $n+1$ is the number of observations. Therefore, highly efficient MCMC methods such as Hamiltonian Monte Carlo (HMC) sampling have to be used to draw samples from (3.22). For the case where the time step between observations is too large, methods have been developed that artificially augment the path of the diffusion process. We will discuss this approach in the next subsection.

3.4.1 Bayesian data augmentation

For parameter estimation from low-frequency observations, MCMC techniques have been developed that introduce imputed data points to reduce the time steps between data points. This concept of Bayesian data imputation goes back to Tanner & Wong (1987) and has been utilized for the inference of diffusions and developed further by many authors such as Elerian et al. (2001), Eraker (2001), Roberts & Stramer (2001), and Golightly & Wilkinson (2008). These methods are applicable to multidimensional processes and were extended for the case of latent process components as well as for the occurrence of measurement error. Thus, they are very promising for the use in real data applications (see e. g. Fuchs (2013) and Golightly & Wilkinson (2006b)).

With low-frequency observations $\boldsymbol{X}^{obs} = (\boldsymbol{X}_{\tau_0}, \dots, \boldsymbol{X}_{\tau_M})$ of the process $(\boldsymbol{X}_t)_{t\geq 0}$ described by the SDE (3.1), we wish to estimate parameter $\boldsymbol{\theta}$. In this section, we assume for simplicity that all observations are complete (i.e. there are no latent or unobserved components for all observations) and that there are no measurement errors. The approximation schemes for the solution of the SDE as introduced in Section 3.3 are only appropriate for small time steps. Therefore, we introduce additional data points $\boldsymbol{X}^{imp}$ at intermediate time points (as visualized in Figure 3.3 and explained in detail in the following subsections) and estimate the parameter $\boldsymbol{\theta}$ from the augmented path $\left\{\boldsymbol{X}^{obs}, \boldsymbol{X}^{imp}\right\}$. To this end, a two-step MCMC approach is used to construct the Markov chain $\left\{\boldsymbol{\theta}_{(i)}, \boldsymbol{X}^{imp}_{(i)}\right\}_{i=1,\dots,L}$, the elements of which are samples from the joint posterior distribution $\pi\left(\boldsymbol{\theta}, \boldsymbol{X}^{imp} \,|\, \boldsymbol{X}^{obs}\right)$ of the parameter and the imputed data points conditioned on the observations. This construction is achieved via a Gibbs sampling approach by alternately executing the following two steps:

Step (1) **Parameter update**: Draw $\boldsymbol{\theta}_{(i)} \sim \pi\left(\boldsymbol{\theta}_{(i)} \,|\, \boldsymbol{X}^{obs}, \boldsymbol{X}^{imp}_{(i-1)}\right)$,

Step (2) **Path update**: Draw $\boldsymbol{X}^{imp}_{(i)} \sim \pi\left(\boldsymbol{X}^{imp}_{(i)} \,|\, \boldsymbol{X}^{obs}, \theta_{(i)}\right)$.

In both steps, direct sampling from the corresponding conditional distribution is generally not possible; therefore, a Metropolis-Hastings algorithm is applied. The resulting MCMC chain $\left\{\boldsymbol{\theta}_{(i)}, \boldsymbol{X}^{imp}_{(i)}\right\}_{i=l+1,\dots,L}$, after discarding the first l elements as burn-in, can be considered

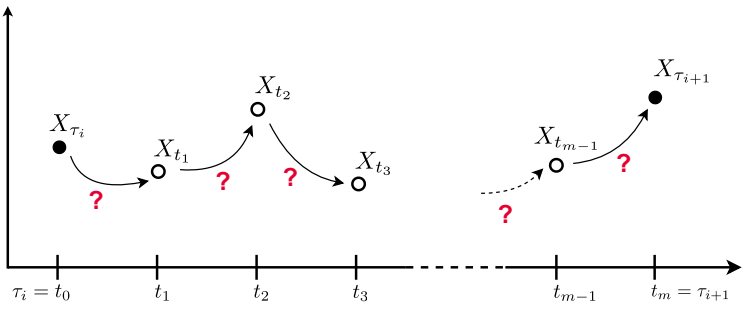


Figure 3.3: Augmented path segment: • represents observed data points and ∘ represents imputed points.

a sample drawn from the joint posterior distribution $\pi\left(\boldsymbol{\theta}, \boldsymbol{X}^{imp} \,|\, \boldsymbol{X}^{obs}\right)$ and can be used for a fully Bayesian analysis. The two steps of the algorithm are described in detail in the following two subsections. We use π to denote the exact densities of the process that is the (full conditional) posterior densities as well as the transition densities. The meaning becomes clear from the arguments. Approximated densities are indicated by a corresponding superscript.

Parameter update

In Step (1), a parameter proposal $\boldsymbol{\theta}^*$ is drawn from a proposal density $q\left(\boldsymbol{\theta}^* \,|\, \boldsymbol{\theta}, \boldsymbol{X}^{obs}, \boldsymbol{X}^{imp}\right)$ which may or may not depend on the imputed and observed data. If a proposal $\boldsymbol{\theta}^* = \boldsymbol{\theta} + \boldsymbol{u}$ with an update $\boldsymbol{u}$ that is independent of the current parameter value $\boldsymbol{\theta}$ is used, the proposal strategy is called a random walk proposal. Proposal $\boldsymbol{\theta}^*$ is accepted with the following probability:

$$\zeta\left(\boldsymbol{\theta}^*, \boldsymbol{\theta}\right) = 1 \wedge \frac{\pi\left(\boldsymbol{\theta}^* \,|\, \boldsymbol{X}^{obs}, \boldsymbol{X}^{imp}\right) q\left(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^*, \boldsymbol{X}^{obs}, \boldsymbol{X}^{imp}\right)}{\pi\left(\boldsymbol{\theta} \,|\, \boldsymbol{X}^{obs}, \boldsymbol{X}^{imp}\right) q\left(\boldsymbol{\theta}^* \,|\, \boldsymbol{\theta}, \boldsymbol{X}^{obs}, \boldsymbol{X}^{imp}\right)}.$$

Otherwise, the previous $\boldsymbol{\theta}$ value is kept.

Due to Bayes' theorem and the fact that a diffusion process has the Markov property, the (full conditional) posterior density can be represented as

$$\pi\left(\boldsymbol{\theta} \,|\, \boldsymbol{X}^{obs}, \boldsymbol{X}^{imp}\right) \propto \left(\prod_{k=0}^{n-1} \pi\left(\boldsymbol{X}_{t_{k+1}} \,|\, \boldsymbol{X}_{t_k}, \boldsymbol{\theta}\right)\right) p(\boldsymbol{\theta}),$$

where $\pi\left(\boldsymbol{X}_{t_{k+1}} \,|\, \boldsymbol{X}_{t_k}, \boldsymbol{\theta}\right)$ denotes the transition density of the process $(\boldsymbol{X}_t)_{t \geq 0}$, $n+1$ is the total number of data points in the augmented path, and p denotes the prior density of the parameter. We choose a random walk proposal where the r components of $\boldsymbol{\theta}^*$ that take values on the entire real line $\mathbb{R}$ are drawn from the normal distribution $\mathcal{N}(\theta_j, \gamma_j^2)$ for $j = 1, \ldots, r$ and some predefined $\gamma_j \in \mathbb{R}_+$. The (remaining) strictly positive components are drawn from

a log-normal distribution $\mathcal{LN}(\log\theta_j, \gamma_j^2)$, for $j = r+1, \ldots, p$. In this case, the acceptance probability reduces to

$$\zeta(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = 1 \wedge \left(\prod_{k=0}^{n-1} \frac{\pi\left(\boldsymbol{X}_{t_{k+1}} \,|\, \boldsymbol{X}_{t_k}, \boldsymbol{\theta}^*\right)}{\pi\left(\boldsymbol{X}_{t_{k+1}} \,|\, \boldsymbol{X}_{t_k}, \boldsymbol{\theta}\right)} \right) \frac{p(\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta})} \left(\prod_{j=r+1}^{p} \frac{\boldsymbol{\theta}_j^*}{\boldsymbol{\theta}_j} \right) \tag{3.23}$$

as derived in (Fuchs, 2013, Chapter 7.1.3).

The transition density $\pi\left(\boldsymbol{X}_{t_{k+1}} \,|\, \boldsymbol{X}_{t_k}, \boldsymbol{\theta}\right)$ is generally not explicitly known, but it can be approximated e. g. by the Euler scheme as described in Section 3.3.

Path update

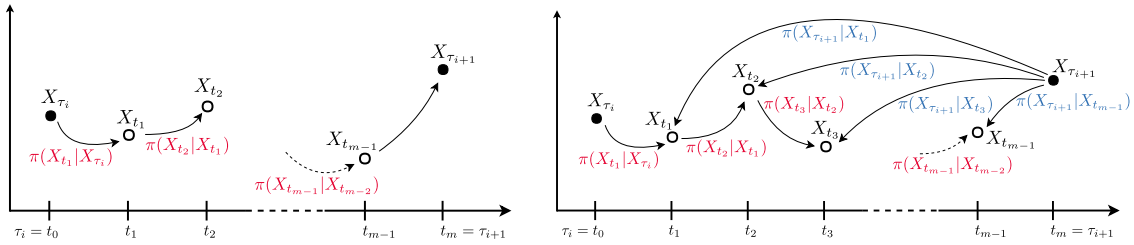
Since a diffusion process has the Markov property, the likelihood function of parameter $\boldsymbol{\theta}$ factorizes as

$$\pi\left(\boldsymbol{X}_{\tau_0}, \ldots, \boldsymbol{X}_{\tau_M} \,|\, \boldsymbol{\theta}\right) = \pi\left(\boldsymbol{X}_{\tau_0} \,|\, \boldsymbol{\theta}\right) \prod_{i=1}^{M} \pi\left(\boldsymbol{X}_{\tau_i} \,|\, \boldsymbol{X}_{\tau_{i-1}}, \boldsymbol{\theta}\right) \tag{3.24}$$

and the latent path segments between observations are conditionally independent given the observations. Hence, it is sufficient to consider the imputation problem in Step (2) only for one path segment between two consecutive observations $\boldsymbol{X}_{\tau_i}$ and $\boldsymbol{X}_{\tau_{i+1}}$. As Figure 3.3 illustrates, the time interval between the two observations is divided into m subintervals such that the end points of these intervals are $\tau_i = t_0 < t_1 < \cdots < t_m = \tau_{i+1}$ and the time steps are $\Delta t_k = t_{k+1} - t_k$ for $k = 0, \ldots, m-1$. We denote the observations by $\boldsymbol{X}^{obs}_{\{\tau_i, \tau_{i+1}\}} = \{\boldsymbol{X}_{\tau_i}, \boldsymbol{X}_{\tau_{i+1}}\}$ and the imputed data points by $\boldsymbol{X}^{imp}_{(\tau_i, \tau_{i+1})} = \{\boldsymbol{X}_{t_1}, \ldots, \boldsymbol{X}_{t_{m-1}}\}$.

After initializing the imputed data by linear interpolation, the path is updated using the Metropolis-Hastings algorithm. A proposal $\boldsymbol{X}^{imp*}_{(\tau_i, \tau_{i+1})}$ is drawn from a distribution with density q, which may depend on the observed data, current imputed data, and parameter $\boldsymbol{\theta}$. The proposal is accepted with the following probability:

$$\begin{aligned} &\zeta\left(\boldsymbol{X}^{imp*}_{(\tau_i, \tau_{i+1})}, \boldsymbol{X}^{imp}_{(\tau_i, \tau_{i+1})}\right) \\ &\quad = 1 \wedge \frac{\pi\left(\boldsymbol{X}^{imp*}_{(\tau_i, \tau_{i+1})} \,\middle|\, \boldsymbol{X}^{obs}_{\{\tau_i, \tau_{i+1}\}}, \boldsymbol{\theta}\right) q\left(\boldsymbol{X}^{imp}_{(\tau_i, \tau_{i+1})} \,\middle|\, \boldsymbol{X}^{imp*}_{(\tau_i, \tau_{i+1})}, \boldsymbol{X}^{obs}_{\{\tau_i, \tau_{i+1}\}}, \boldsymbol{\theta}\right)}{\pi\left(\boldsymbol{X}^{imp}_{(\tau_i, \tau_{i+1})} \,\middle|\, \boldsymbol{X}^{obs}_{\{\tau_i, \tau_{i+1}\}}, \boldsymbol{\theta}\right) q\left(\boldsymbol{X}^{imp*}_{(\tau_i, \tau_{i+1})} \,\middle|\, \boldsymbol{X}^{imp}_{(\tau_i, \tau_{i+1})}, \boldsymbol{X}^{obs}_{\{\tau_i, \tau_{i+1}\}}, \boldsymbol{\theta}\right)}. \end{aligned} \tag{3.25}$$



(a) Left-conditioned proposal **(b)** Bridge proposal

Figure 3.4: Illustration of the different proposal strategies.

Otherwise, the proposal is discarded and the previously imputed data $\boldsymbol{X}^{imp}_{(\tau_i,\tau_{i+1})}$ is kept. Due to the Markov property, we have:

$$\frac{\pi\left(\boldsymbol{X}^{imp*}_{(\tau_i,\tau_{i+1})} \,\middle|\, \boldsymbol{X}^{obs}_{\{\tau_i,\tau_{i+1}\}}, \boldsymbol{\theta}\right)}{\pi\left(\boldsymbol{X}^{imp}_{(\tau_i,\tau_{i+1})} \,\middle|\, \boldsymbol{X}^{obs}_{\{\tau_i,\tau_{i+1}\}}, \boldsymbol{\theta}\right)} = \prod_{k=0}^{m-1} \frac{\pi\left(\boldsymbol{X}^{*}_{t_{k+1}} \,|\, \boldsymbol{X}^{*}_{t_k}, \boldsymbol{\theta}\right)}{\pi\left(\boldsymbol{X}_{t_{k+1}} \,|\, \boldsymbol{X}_{t_k}, \boldsymbol{\theta}\right)},$$

where $\boldsymbol{X}^{*}_{t_0} = \boldsymbol{X}_{t_0} = \boldsymbol{X}_{\tau_i}$, $\boldsymbol{X}^{*}_{t_m} = \boldsymbol{X}_{t_m} = \boldsymbol{X}_{\tau_{i+1}}$, and $\pi\left(\boldsymbol{X}_{t_{k+1}} \,|\, \boldsymbol{X}_{t_k}, \boldsymbol{\theta}\right)$ denotes the transition density of process $(\boldsymbol{X}_t)_{t\geq 0}$.

The challenging aspect of the path update step involves determining how to propose new points. The simplest approach uses the (approximated) transition density to propose a new point by conditioning only on the point to the left of the new point. We call this proposal method the *left-conditioned (LC) proposal* and illustrate it in Figure 3.4a. The proposal density of an entire path segment is simply the product

$$q_{LC}\left(\boldsymbol{X}^{imp*}_{(\tau_i,\tau_{i+1})} \,\middle|\, \boldsymbol{X}_{\tau_i}, \theta\right) = \prod_{k=0}^{m-2} \pi\left(\boldsymbol{X}^{*}_{t_{k+1}} \,|\, \boldsymbol{X}^{*}_{t_k}, \boldsymbol{\theta}\right), \tag{3.26}$$

where $\boldsymbol{X}^{*}_{t_0} = \boldsymbol{X}_{\tau_i}$. Thus, the acceptance probability reduces to

$$\begin{aligned}\zeta\left(\boldsymbol{X}^{imp*}_{(\tau_i,\tau_{i+1})}, \boldsymbol{X}^{imp}_{(\tau_i,\tau_{i+1})}\right) &= 1 \wedge \left(\prod_{k=0}^{m-1} \frac{\pi\left(\boldsymbol{X}^{*}_{t_{k+1}} \,|\, \boldsymbol{X}^{*}_{t_k}, \boldsymbol{\theta}\right)}{\pi\left(\boldsymbol{X}_{t_{k+1}} \,|\, \boldsymbol{X}_{t_k}, \boldsymbol{\theta}\right)}\right)\left(\prod_{k=0}^{m-2} \frac{\pi\left(\boldsymbol{X}_{t_{k+1}} \,|\, \boldsymbol{X}_{t_k}, \boldsymbol{\theta}\right)}{\pi\left(\boldsymbol{X}^{*}_{t_{k+1}} \,|\, \boldsymbol{X}^{*}_{t_k}, \boldsymbol{\theta}\right)}\right) \\ &= 1 \wedge \frac{\pi\left(\boldsymbol{X}_{\tau_{i+1}} \,|\, \boldsymbol{X}^{*}_{t_{m-1}}, \boldsymbol{\theta}\right)}{\pi\left(\boldsymbol{X}_{\tau_{i+1}} \,|\, \boldsymbol{X}_{t_{m-1}}, \boldsymbol{\theta}\right)},\end{aligned}$$

where $\boldsymbol{X}^{*}_{t_m} = \boldsymbol{X}_{t_m} = \boldsymbol{X}_{\tau_{i+1}}$. Here, the transition density again needs to be approximated e. g. by the Euler scheme from Section 3.3.

This proposal strategy considers the information from the observation $\boldsymbol{X}_{\tau_i}$ on the left, while the proposed path segment is independent of the observation $\boldsymbol{X}_{\tau_{i+1}}$ on the right. This may lead to a large jump in the last step from $\boldsymbol{X}_{t_{m-1}}$ to $\boldsymbol{X}_{\tau_{i+1}}$, and hence, to an improbable transition. Therefore, the acceptance probability for the left-conditioned proposal $\boldsymbol{X}^{imp*}_{(\tau_i,\tau_{i+1})}$, and consequently, the acceptance rate of the MCMC sampler is usually low.

A number of more sophisticated proposal strategies have been suggested. Chapter 7.1 in Fuchs (2013) reviews some of these. Here, we consider the *modified bridge (MB) proposal*, which conditions on both the previous data point and the following observation on the right, as visualized in Figure 3.4b. This strategy was originally proposed by Durham & Gallant (2002) and first applied in the Bayesian framework in Chib & Shephard (2002). More recently, Whitaker et al. (2017) suggested improved bridge constructs, and van der Meulen & Schauer (2017) proposed so-called guided proposals.

For the MB proposal, the proposal density of an entire path segment factorizes again as follows:

$$q_{MB}\left(\boldsymbol{X}^{imp*}_{(\tau_i,\tau_{i+1})} \,|\, \boldsymbol{X}_{\tau_i,}, \boldsymbol{X}_{\tau_{i+1}}, \boldsymbol{\theta}\right) = \prod_{k=0}^{m-2} \pi\left(\boldsymbol{X}^*_{t_{k+1}} \,|\, \boldsymbol{X}^*_{t_k}, \boldsymbol{X}_{\tau_{i+1}}, \boldsymbol{\theta}\right),$$

where $\boldsymbol{X}^*_{t_0} = \boldsymbol{X}_{\tau_i}$. We apply Bayes' theorem and the Markov property to rewrite the left- and right-conditioned proposal density of one point as

$$\pi\left(\boldsymbol{X}^*_{t_{k+1}} \,|\, \boldsymbol{X}^*_{t_k}, \boldsymbol{X}_{\tau_{i+1}}, \theta\right) \propto \pi\left(\boldsymbol{X}^*_{t_{k+1}} \,|\, \boldsymbol{X}^*_{t_k}, \boldsymbol{\theta}\right) \pi\left(\boldsymbol{X}_{\tau_{i+1}} \,|\, \boldsymbol{X}^*_{t_{k+1}}, \boldsymbol{\theta}\right) \tag{3.27}$$

for $k = 0, \ldots, m-2$.

In Durham & Gallant (2002), it is suggested to approximate the two transition densities on the right-hand side by the Euler scheme and to further approximate $\boldsymbol{\mu}\left(\boldsymbol{X}^*_{t_{k+1}}, \boldsymbol{\theta}\right)$ and $\boldsymbol{\sigma}\left(\boldsymbol{X}^*_{t_{k+1}}, \boldsymbol{\theta}\right)$ by $\boldsymbol{\mu}\left(\boldsymbol{X}^*_{t_k}, \boldsymbol{\theta}\right)$ and $\boldsymbol{\sigma}\left(\boldsymbol{X}^*_{t_k}, \boldsymbol{\theta}\right)$, respectively. This way, they obtain that (3.27) is approximately proportional to a Gaussian density which we will use for the MB proposal based on the Euler scheme:

$$\begin{aligned} &\pi^{Euler}\left(\boldsymbol{X}^*_{t_{k+1}} \,|\, \boldsymbol{X}^*_{t_k}, \boldsymbol{X}_{\tau_{i+1}}, \boldsymbol{\theta}\right) \\ &\quad = \phi\left(\boldsymbol{X}^*_{t_{k+1}} \,\middle|\, \boldsymbol{X}^*_{t_k} + \left(\frac{\boldsymbol{X}_{\tau_{i+1}} - \boldsymbol{X}^*_{t_k}}{\tau_{i+1} - t_k}\right) \Delta t_k,\, \left(\frac{\tau_{i+1} - t_{k+1}}{\tau_{i+1} - t_k}\right) \boldsymbol{\Sigma}\left(\boldsymbol{X}^*_{t_k}, \boldsymbol{\theta}\right) \Delta t_k\right), \end{aligned} \tag{3.28}$$

where $\boldsymbol{\Sigma}\left(\boldsymbol{X}^*_{t_k}, \boldsymbol{\theta}\right) = \boldsymbol{\sigma}\left(\boldsymbol{X}^*_{t_k}, \boldsymbol{\theta}\right) \boldsymbol{\sigma}^{\mathsf{Tr}}\left(\boldsymbol{X}^*_{t_k}, \boldsymbol{\theta}\right)$ and $\phi\left(\cdot \,|\, \boldsymbol{a}, \boldsymbol{b}\right)$ denotes the multivariate Gaussian density with mean $\boldsymbol{a} \in \mathbb{R}^d$ and covariance matrix $\boldsymbol{b} \in \mathbb{R}^{d \times d}$.

Since the MB proposal takes into account information not only from the left data point but also from the observation on the right, it does not have a large jump in the last step as the left-conditioned proposal does.

Thus far, our path update has only been applied to imputed points between two observations. It can easily be extended to a case with several observations along the path by simply decomposing the path into independent path proposals, multiplying the respective acceptance probabilities and collectively accepting or rejecting the proposals. Moreover, the entire path does not have to be updated all at once, but can be divided into several path segments that are successively updated. Different algorithms for choosing the update interval are summarized in Fuchs (2013) and Appendix A.2.1 describes one of them.

3.4.2 Extensions of the basic data augmentation scheme and alternatives

The data augmentation scheme introduced in the previous section can be generalized for the case of (additive) measurement error and latent components of the diffusion process. The respective proposal procedures based on the modified bridge proposal and the corresponding acceptance probabilities are derived e. g. in Fuchs (2013, Chapter 7.2). Several other works also account for measurement error and latent components e. g. Golightly & Wilkinson (2006b, 2008), and Whitaker et al. (2017).

Another challenge in the context of Bayesian data augmentation and the MCMC scheme discussed in the previous subsection is the dependence between the parameter components included in the diffusion function and the missing path segments between two observations that results from Relationship (3.3) between the diffusion matrix and the quadratic variation of the process. Roberts & Stramer (2001) were the first to highlight that in the discretized setting (as we consider it here), this dependence leads to a slower convergence of the MCMC algorithm as the number of imputed points $m - 1$ increases. Several approaches have been developed to overcome this problem. Some of them are summarized in Fuchs (2013, Chapter 7.4). Here, we only mention the approaches that are generally applicable (e. g. that do not require transformation to a unit diffusion).

One approach was motivated by a reparametrization first used in Chib et al. (2004) and became to be known as the *innovation scheme*. The idea is to exploit the bijective relationship between the diffusion path and the driving Brownian motion conditional on the parameter, i. e. there is an invertible function that maps between the two processes. The acceptance probability in the parameter update is conditioned on the Brownian motion (which does not contain information about the parameter) rather than on the diffusion path and then, the diffusion path is obtained by transforming the Brownian motion once a new parameter is

accepted. Thus, the parameter and the imputed path segments are consistent at any step of the algorithm and the convergence problem is overcome. While Golightly & Wilkinson (2008) study this approach for general diffusion processes in the discrete-time setting only and refine it in Wilkinson & Golightly (2010), Fuchs (2013, Chapter 7.4.4) also considers the continuous-time framework, i. e. where the time step between the imputed data points tends to zero. For their guided proposals, van der Meulen & Schauer (2017) also make use of the idea of an innovation process and show that the results obtained in Fuchs (2013) are a special case of their work.

Golightly & Wilkinson (2006a) introduced a sequential MCMC algorithm (also referred to as particle filter) that simultaneously updates the parameter vector and the imputed process states and thus circumvents issues arising from their dependence. They also applied this algorithm to stochastic kinetic models in Golightly & Wilkinson (2006b). In this sequential approach, the observations are taken into account one after another to update the posterior distribution. Hence, it allows for on-line estimation of the parameter vector as additional data points becomes available (instead of restarting the whole MCMC procedure for every new data point) which is very useful for real-time data analysis. However, the algorithm suffers from other problems, e. g. a poor approximation of the posterior in one step of the algorithm will propagate to the next step. In particular, it may occur that the final approximation of the posterior distribution based on all currently available data concentrates only on very few distinct parameter values. This problem becomes the more severe the more observed time points are available.

Another class of algorithms that combine MCMC and sequential Monte Carlo methods is known as *particle MCMC (pMCMC)*, a term that was coined in Andrieu et al. (2010). One representative of the pMCMC methods is the so-called particle marginal Metropolis–Hastings (PMMH) algorithm. This algorithm can also be interpreted in the light of the pseudo-marginal approach as described in Beaumont (2003) and Andrieu & Roberts (2009). Golightly & Wilkinson (2011) use this approach for inference of SDE parameters. The idea of this algorithm is to construct a Metropolis–Hastings algorithm that targets the posterior distribution of the parameter vector $\boldsymbol{\theta}$, but since the marginal likelihood needed in the acceptance probability is intractable, it is approximated using a particle filter. Inside the particle filter, the stochastic model can either simply be forward simulated or Golightly & Wilkinson (2011) also consider the use of a bridge construct. In both cases, the parameter vector and the imputed path segment between observations are jointly updated and no issues from their dependence arise. Moreover, since the particle filter provides an unbiased estimate of the marginal likelihood, the PMMH algorithm targets the exact posterior distribution despite using an approximation in the acceptance probability. However, the algorithm is computationally extremely expensive. Within each iteration of the Metropolis–Hastings algorithm, $N \times M \times m$ simulation steps for

the diffusion process are required if N is the number of particles used in the particle filter, $M+1$ time points are observed, and $m-1$ points are imputed between every two observations.

There is also a non-Bayesian approach based on particle filtering called *iterated filtering* which is described in Ionides et al. (2006).

More recently, also variational inference approaches have been explored for SDEs e. g. in Ryder et al. (2018) and Opper (2019); however, there are no results available yet comparing these approaches to the other inference techniques mentioned in this section.

Chapter 4

Using higher-order approximations in Bayesian inference for diffusions

Having introduced several approaches to perform inference for diffusion processes; in this chapter, we further consider the Bayesian data augmentation method as described in Section 3.4.1 that is used to infer the parameter $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$ from low-frequency observations $\boldsymbol{X}^{obs} = (\boldsymbol{X}_{\tau_0}, \ldots, \boldsymbol{X}_{\tau_M})$ of a diffusion process $(\boldsymbol{X}_t)_{t \geq 0}$ described by SDE

$$\mathrm{d}\boldsymbol{X}_t = \boldsymbol{\mu}(\boldsymbol{X}_t, \boldsymbol{\theta})\, \mathrm{d}t + \boldsymbol{\sigma}(\boldsymbol{X}_t, \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{B}_t, \qquad \boldsymbol{X}_0 = \boldsymbol{x}_0, \tag{4.1}$$

as detailed in Section 3.1. In this approach, the numerical approximation of the transition densities of the process is necessary not only for calculating the posterior densities, but also for proposing the imputed data points. In both contexts, the Euler-Maruyama scheme is the standard approximation technique in the literature (see e. g. Elerian et al., 2001, Eraker, 2001, Golightly & Wilkinson, 2006a, 2008, Roberts & Stramer, 2001). To reduce the amount of imputed data and the number of necessary iterations for the computationally expensive estimation method, one possible solution is to employ higher-order approximation schemes.

Therefore, we investigate the utilization and usefulness of such higher-order approximations on the example of the Milstein scheme introduced in Section 3.3. A closed form of the transition density based on the Milstein scheme is derived in Elerian (1998). In Tse et al. (2004), this closed form is used to estimate the parameters of a hyperbolic diffusion process from high-frequency financial data, but not in the context of Bayesian data augmentation. For the latter, Elerian et al. (2001) propose the possible use of the Milstein scheme. However, the specific implementation and benefit of this framework, in particular when using sophisticated proposal methods, remain unclear, and therefore, are the focus of this work. For our investigation,

we first derive the transition density of the diffusion process approximated by the Milstein scheme, then explain how to integrate the Milstein scheme into the framework of Bayesian data augmentation, and finally assess the effectiveness of this new combination in a simulation study which is a common approach in the literature (see e.g. Whitaker et al. (2017) and Mrázek & Pospíšil (2017)). In the simulation study, we consider the GBM and the CIR process as introduced in Section 3.2. Parts of this chapter are similar or identical to the following article:

Pieschner, S. & Fuchs, C. (2020). Bayesian inference for diffusion processes: using higher-order approximations for transition densities. *Royal Society Open Science*, 7(10), 200270.

4.1 The transition density based on the Milstein scheme

For the reader's convenience, we restate the formula of the Milstein scheme from Section 3.3. It approximates the d-dimensional process $(\boldsymbol{X}_t)_{t\geq 0}$ by setting $\boldsymbol{Y}_0 = \boldsymbol{x}_0$ and, then, successively calculating for the i^{th} component:

$$\begin{aligned}\boldsymbol{Y}_{k+1}^{(i)} =& \boldsymbol{Y}_k^{(i)} + \boldsymbol{\mu}_i\left(\boldsymbol{Y}_k, \boldsymbol{\theta}\right)\Delta t_k + \sum_{l=1}^{r}\boldsymbol{\sigma}_{il}\left(\boldsymbol{Y}_k, \boldsymbol{\theta}\right)\Delta B_k^{(l)} \\ &+ \sum_{l=1}^{r}\sum_{q=1}^{r}\sum_{j=1}^{d}\boldsymbol{\sigma}_{jq}\left(\boldsymbol{Y}_k, \boldsymbol{\theta}\right)\frac{\partial \boldsymbol{\sigma}_{il}}{\partial y^{(r)}}\left(\boldsymbol{Y}_k, \boldsymbol{\theta}\right)\int_{t_k}^{t_{k+1}}\int_{t_k}^{s}\mathrm{d}\boldsymbol{B}_u^{(q)}\mathrm{d}\boldsymbol{B}_s^{(l)}\end{aligned} \tag{4.2}$$

for $k = 0, 1, \ldots$ and $i = 1, \ldots, d$. We have already pointed out that the convergence rate of the Milstein scheme is higher than that of the Euler scheme which is the reason for our investigation. However, there is a severe restriction on the practical applicability of the Milstein scheme because the stochastic double integral in the last term of (4.2) only yields an analytical solution for $j = l$. Although approximation techniques for the double integral exist (see e.g. Kloeden & Platen (1992)), they are unsuitable for our purposes. On the one hand, we wish to avoid adding yet another layer of approximation and, thus, additional computational time. On the other hand, we must find the distribution of $\boldsymbol{Y}_{k+1}$ based on approximation schemes (4.2), which is also not explicitly possible when adding another approximation. For this reason, we focus on models where the double integral appears exclusively for the same components of the Brownian motion. For example, this is the case when the process is driven by a one-dimensional Brownian motion (i. e. the diffusion function $\boldsymbol{\sigma}\left(\boldsymbol{Y}_k, \boldsymbol{\theta}\right)$ is of dimension $d \times 1$). Hence, the diffusion model includes only one source of noise that may affect

each of the components of the process. More generally, we require that

$$\boldsymbol{\sigma}_{rj}\left(\boldsymbol{Y}_k, \boldsymbol{\theta}\right) \frac{\partial \boldsymbol{\sigma}_{il}}{\partial y^{(r)}}\left(\boldsymbol{Y}_k, \boldsymbol{\theta}\right) \equiv 0 \tag{4.3}$$

for $j \neq l$ so that only $j = l$ is inside the double integral. Relation (4.3) implies the following:

- if an entry $\boldsymbol{\sigma}_{rj}\left(\boldsymbol{Y}_k, \boldsymbol{\theta}\right)$ is non-zero, then the entries of all *other* columns and *all* rows must not depend on $\boldsymbol{Y}_k^{(r)}$, and
- if an entry $\boldsymbol{\sigma}_{il}\left(\boldsymbol{Y}_k, \boldsymbol{\theta}\right)$ depends on $\boldsymbol{Y}_k^{(r)}$, then the entries of all *other* columns in row r must be zero.

In particular, this means that unless the r^{th} row of the diffusion function contains only zeros, component $\boldsymbol{Y}_k^{(r)}$ can only appear in *one* column of the diffusion function (and if it appears, then the entries of all *other* columns in row r must be zero). Moreover, each component of the diffusion process $(\boldsymbol{X}_t)_{t\geq 0}$ can only be directly affected by more than one component of the Brownian motion, if the size of all stochastic effects (i. e. *all* entries of the diffusion function) does not depend on the respective component of the diffusion process. Further, if all d components of the diffusion process appear in the diffusion function, then the process can be affected by at most d components of the Brownian motion. Besides, if all d components of the diffusion process appear in the diffusion function and the process shall be affect by d components of the Brownian motion, the diffusion function must be a (possibly column-wise permuted) diagonal matrix. In many applications, these are not realistic assumptions.

Assume that the i^{th} component of the diffusion process appears in the i^{th} row of the diffusion function and that the respective entry of the diffusion function does not depend on the remaining components $\boldsymbol{Y}_k^{(r)}$, $r \neq i$ (the contrary would impose restrictions on other rows, as described above). Then, the i^{th} component of the approximated process is

$$\begin{aligned}\boldsymbol{Y}_{k+1}^{(i)} =& \boldsymbol{Y}_k^{(i)} + \boldsymbol{\mu}_i\left(\boldsymbol{Y}_k, \boldsymbol{\theta}\right)\Delta t_k + \boldsymbol{\sigma}_{ij}\left(\boldsymbol{Y}_k, \boldsymbol{\theta}\right)\Delta B_k^{(j)} \\ &+ \boldsymbol{\sigma}_{ij}\left(\boldsymbol{Y}_k, \boldsymbol{\theta}\right)\frac{\partial \boldsymbol{\sigma}_{ij}}{\partial y^{(i)}}\left(\boldsymbol{Y}_k, \boldsymbol{\theta}\right)\frac{1}{2}\left(\left(\Delta \boldsymbol{B}_k^{(j)}\right)^2 - \Delta t_k\right)\end{aligned} \tag{4.4}$$

for $k = 0, 1, \ldots$ and where j is the column index of the one non-zero entry depending on $\boldsymbol{Y}_k^{(i)}$ in the i^{th} row of the diffusion function.

Moreover, note that if we consider the approximation $\boldsymbol{Y}_{k+1}^{(i)}$ in Equation (4.4) as a function $g\left(\Delta \boldsymbol{B}_k^{(j)}\right)$ of the increment of the Brownian motion, g is quadratic in $\Delta \boldsymbol{B}_k^{(j)}$. Therefore,

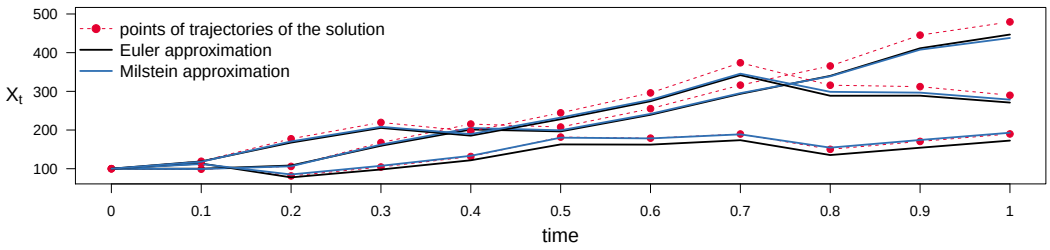


Figure 4.1: Three trajectories of a GBM (3.6) with $\alpha = 1$ and $\sigma^2 = 0.25$ and their approximations by the Euler and the Milstein scheme for time steps $\Delta t = 0.1$.

the function g has a global extremum with value

$$\begin{aligned} g^* = \boldsymbol{Y}_k^{(i)} - \frac{1}{2}\boldsymbol{\sigma}_{ij}\left(\boldsymbol{Y}_k, \boldsymbol{\theta}\right) \Bigg/ \left(\frac{\partial \boldsymbol{\sigma}_{ij}}{\partial y^{(i)}}\left(\boldsymbol{Y}_k, \boldsymbol{\theta}\right)\right) \\ + \left(\boldsymbol{\mu}_i\left(\boldsymbol{Y}_k, \boldsymbol{\theta}\right) - \frac{1}{2}\boldsymbol{\sigma}_{ij}\left(\boldsymbol{Y}_k, \boldsymbol{\theta}\right)\frac{\partial \boldsymbol{\sigma}_{ij}}{\partial y^{(i)}}\left(\boldsymbol{Y}_k, \boldsymbol{\theta}\right)\right)\Delta t_k. \end{aligned} \tag{4.5}$$

Hence, there is a bound on the range of possible values for $\boldsymbol{Y}_{k+1}^{(i)}$ resulting from the Milstein scheme which might exclude values that the solution process $\boldsymbol{X}_{t_k}$ could take. Whether this is a lower or upper bound depends on the sign of the diffusion function and its derivative. The second derivative of g is given by

$$\frac{\partial^2 g(\Delta \boldsymbol{B}_k^{(j)})}{\partial \left(\Delta \boldsymbol{B}_k^{(j)}\right)^2} = \boldsymbol{\sigma}_{ij}\left(\boldsymbol{Y}_k, \boldsymbol{\theta}\right)\frac{\partial \boldsymbol{\sigma}_{ij}}{\partial y^{(i)}}\left(\boldsymbol{Y}_k, \boldsymbol{\theta}\right) =: g''.$$

Thus, the extremum g^* is a maximum and puts an upper bound on the possible values of $\boldsymbol{Y}_{k+1}^{(i)}$ if $g'' < 0$, and g^* is a minimum and puts a lower bound on $\boldsymbol{Y}_{k+1}^{(i)}$ if $g'' > 0$. For the case where $g'' = 0$, the Milstein scheme reduces to the Euler scheme.

Since our examples, the GBM and the CIR process, are one-dimensional processes, the double integral in Equation (4.2) vanishes. The Milstein scheme for the GBM yields the following:

$$Y_{k+1} = Y_k + \alpha Y_k \Delta t_k + \sigma Y_k \Delta B_k + \frac{1}{2}\sigma^2 Y_k\left((\Delta B_k)^2 - \Delta t_k\right)$$

for $k = 0, 1, \ldots,$ where the first three summands also correspond to the Euler scheme. Figure 4.1 illustrates the two approximation schemes. It presents three trajectories of the GBM, which are represented by red points and which were simulated by setting a seed for the random number generator and, then, sampling from the exact transition density (3.7). The same seed was used to sample the increments of the Brownian motion from the normal density and then transform them by (3.11) and (4.4) to obtain the Euler (black) and the Milstein (blue) approx-

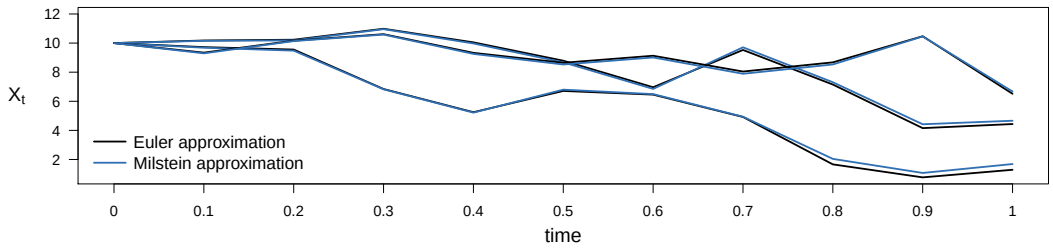


Figure 4.2: Three approximated trajectories of a CIR process (3.9) with $\alpha = 1$, $\beta = 1$, and $\sigma^2 = 2$ approximated by the Euler and the Milstein scheme for time steps $\Delta t = 0.1$.

imation of the trajectories. We observe that in almost all cases, the Milstein approximation is either closer to or as close to the points of the trajectories as the Euler approximation.

For the CIR process, the Milstein scheme yields the following:

$$Y_{k+1} = Y_k + \alpha\left(\beta - Y_k\right)\Delta t_k + \sigma\sqrt{Y_k}\Delta B_k + \frac{1}{4}\sigma^2\left(\left(\Delta B_k\right)^2 - \Delta t_k\right) \tag{4.6}$$

for $k = 0, 1, \ldots$, where the first three summands again correspond to the Euler scheme. A similar illustration as in Figure 4.1 where the approximations are compared to the trajectories of the true process is not easily possible for the CIR process because sampling from the transition density (3.10) is not achieved directly but e. g. by generating a normally distributed and a chi-square distributed random variable and their transformation. Therefore, Figure 4.2 only shows the approximated trajectories of the CIR process again obtained by sampling the increments of the Brownian motion and then transforming them by (3.11) and (4.4).

While sampling approximated diffusion paths is fairly straightforward for both approximation schemes as described above, determining the corresponding transition density is less apparent for the Milstein scheme. As already pointed out in Section 3.4, the transition density derived from the Euler scheme is simply a multivariate Gaussian density:

$$\pi^{Euler}\left(\boldsymbol{Y}_{k+1}|\boldsymbol{Y}_k, \boldsymbol{\theta}\right) = \phi\left(\boldsymbol{Y}_{k+1}|\boldsymbol{Y}_k + \boldsymbol{\mu}\left(\boldsymbol{Y}_k, \boldsymbol{\theta}\right)\Delta t_k, \boldsymbol{\sigma}\left(\boldsymbol{Y}_k, \boldsymbol{\theta}\right)\boldsymbol{\sigma}^{\mathsf{Tr}}\left(\boldsymbol{Y}_k, \boldsymbol{\theta}\right)\Delta t_k\right),$$

where $\phi\left(\boldsymbol{y}|\boldsymbol{a}, \boldsymbol{b}\right)$ denotes the multivariate Gaussian density with mean $\boldsymbol{a} \in \mathbb{R}^d$ and covariance matrix $\boldsymbol{b} \in \mathbb{R}^{d\times d}$ evaluated at $\boldsymbol{y}$.

For the Milstein scheme, deriving the transition density is more complicated, even in the case of a one-dimensional diffusion process, which we consider here. Elerian (1998) derived the transition density by first rearranging the Milstein scheme to obtain a transformation of a non-central chi-squared distributed variable for which the density is known, and then applying the random variable transformation theorem. Here, we present an alternative derivation

that directly applies the random variable transformation theorem to increments ΔB_k of the Brownian motion. Both approaches produce the same result. For simplicity of notation, we set $\mu_k := \mu\left(Y_k, \boldsymbol{\theta}\right)$, $\sigma_k := \sigma\left(Y_k, \boldsymbol{\theta}\right)$, and $\sigma_k' := \partial\sigma\left(y, \boldsymbol{\theta}\right)/\partial y\,|_{y=Y_k}$.

Theorem 4.1. *Given a one-dimensional diffusion process described by SDE (4.1), the approximated transition density based on the Milstein scheme (4.4) is as follows:*

$$\pi^{Mil}\left(Y_{k+1}|Y_k, \boldsymbol{\theta}\right) = \frac{\exp\left(-\dfrac{C_k(Y_{k+1})}{D_k}\right)}{\sqrt{2\pi}\sqrt{\Delta t_k}\sqrt{A_k(Y_{k+1})}} \cdot \left[\exp\left(-\frac{\sqrt{A_k(Y_{k+1})}}{D_k}\right) + \exp\left(\frac{\sqrt{A_k(Y_{k+1})}}{D_k}\right)\right]$$

with

$$\begin{aligned} A_k(Y_{k+1}) &= \left(\sigma_k\right)^2 + 2\sigma_k\sigma_k'\left(Y_{k+1} - Y_k - \left(\mu_k - \frac{1}{2}\sigma_k\sigma_k'\right)\Delta t_k\right), \\ C_k(Y_{k+1}) &= \sigma_k + \sigma_k'\left(Y_{k+1} - Y_k - \left(\mu_k - \frac{1}{2}\sigma_k\sigma_k'\right)\Delta t_k\right), \\ D_k &= \sigma_k\left(\sigma_k'\right)^2\Delta t_k \end{aligned}$$

and for

$$Y_{k+1} \geq Y_k - \frac{1}{2}\frac{\sigma_k}{\sigma_k'} + \left(\mu_k - \frac{1}{2}\sigma_k\sigma_k'\right)\Delta t_k, \qquad \text{if } \sigma_k\sigma_k' > 0, \tag{4.7}$$

and

$$Y_{k+1} \leq Y_k - \frac{1}{2}\frac{\sigma_k}{\sigma_k'} + \left(\mu_k - \frac{1}{2}\sigma_k\sigma_k'\right)\Delta t_k, \qquad \text{if } \sigma_k\sigma_k' < 0. \tag{4.8}$$

Proof. The Milstein scheme

$$Y_{k+1} = Y_k + \mu\left(Y_k, \boldsymbol{\theta}\right)\Delta t_k + \sigma\left(Y_k, \boldsymbol{\theta}\right)\Delta B_k + \frac{1}{2}\sigma\left(Y_k, \boldsymbol{\theta}\right)\frac{\partial\sigma}{\partial y}\left(Y_k, \boldsymbol{\theta}\right)\left(\left(\Delta B_k\right)^2 - \Delta t_k\right)$$

can be considered a variable transformation of the random variable $Z \sim \mathcal{N}(0,1)$ with density $\phi(z)$ using the transformation function

$$f(z) = az^2 + bz + c,$$

where the coefficients are defined as

$$\begin{aligned} a &= \frac{1}{2}\sigma\left(Y_k, \boldsymbol{\theta}\right)\frac{\partial\sigma}{\partial y}\left(Y_k, \boldsymbol{\theta}\right)\Delta t_k, \\ b &= \sigma\left(Y_k, \boldsymbol{\theta}\right)\sqrt{\Delta t_k}, \end{aligned}$$

$$c = Y_k + \left[\mu\left(Y_k, \boldsymbol{\theta}\right) - \frac{1}{2}\sigma\left(Y_k, \boldsymbol{\theta}\right)\frac{\partial \sigma}{\partial y}\left(Y_k, \boldsymbol{\theta}\right)\right]\Delta t_k,$$

and whose derivative and inverse function are

$$\begin{aligned} f'(z) &= 2az + b, \\ f^{-1}(y) &= -\frac{b}{2a} \pm \frac{\sqrt{b^2 + 4a\left(y - c\right)}}{2a} \text{ for } y \geq -\frac{b^2}{4a} + c. \end{aligned}$$

By applying the random variable transformation theorem as found in Schmidt (2009, p. 269) or Gillespie (1992a, p.27), the density ρ_Y of Y_{k+1} can be derived as follows:

$$\begin{aligned} \rho_Y(y) &= \sum_{\{z \in \mathbb{R}: f(z) = y\}} \frac{\phi(z)}{|f'(z)|} \\ &= \frac{\phi\left(-\frac{b}{2a} - \frac{\sqrt{b^2 + 4a(y-c)}}{2a}\right)}{\left|f'\left(-\frac{b}{2a} - \frac{\sqrt{b^2 + 4a(y-c)}}{2a}\right)\right|} + \frac{\phi\left(-\frac{b}{2a} + \frac{\sqrt{b^2 + 4a(y-c)}}{2a}\right)}{\left|f'\left(-\frac{b}{2a} + \frac{\sqrt{b^2 + 4a(y-c)}}{2a}\right)\right|} \\ &= \frac{\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(-\frac{b}{2a} - \frac{\sqrt{b^2 + 4a(y-c)}}{2a}\right)^2\right)}{\left|b + 2a\left(-\frac{b}{2a} - \frac{\sqrt{b^2 + 4a(y-c)}}{2a}\right)\right|} \\ &\quad + \frac{\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(-\frac{b}{2a} + \frac{\sqrt{b^2 + 4a(y-c)}}{2a}\right)^2\right)}{\left|b + 2a\left(-\frac{b}{2a} + \frac{\sqrt{b^2 + 4a(y-c)}}{2a}\right)\right|} \\ &= \frac{1}{\sqrt{2\pi}}\left(\frac{\exp\left(-\frac{1}{8a^2}\left(b^2 + 2b\sqrt{b^2 + 4a(y-c)} + b^2 + 4a(y-c)\right)\right)}{\left|-\sqrt{b^2 + 4a(y-c)}\right|}\right. \\ &\quad \left. + \frac{\exp\left(-\frac{1}{8a^2}\left(b^2 - 2b\sqrt{b^2 + 4a(y-c)} + b^2 + 4a(y-c)\right)\right)}{\left|\sqrt{b^2 + 4a(y-c)}\right|}\right) \end{aligned}$$

$$= \frac{\exp\left(-\frac{b^2+2a(y-c)}{4a^2}\right)}{\sqrt{2\pi}\sqrt{b^2+4a(y-c)}}\left(\exp\left(-\frac{b\sqrt{b^2+4a(y-c)}}{4a^2}\right)+\exp\left(\frac{b\sqrt{b^2+4a(y-c)}}{4a^2}\right)\right)$$

$$= \frac{\exp\left(-\frac{b^2+2a(y-c)}{4a^2}\right)}{\sqrt{2\pi}\sqrt{b^2+4a(y-c)}}\cdot 2\cosh\left(\frac{b\sqrt{b^2+4a(y-c)}}{4a^2}\right).$$

After substituting the coefficients a, b, and c and abbreviating $\mu_k := \mu\left(Y_k, \boldsymbol{\theta}\right)$, $\sigma_k := \sigma\left(Y_k, \boldsymbol{\theta}\right)$, and $\sigma_k' := \sigma'\left(Y_k, \boldsymbol{\theta}\right) = \partial\sigma\left(y, \boldsymbol{\theta}\right)/\partial y\big|_{y=Y_k}$, we obtain the transition density based on the Milstein scheme

$$\pi^{Mil}\left(Y_{k+1}|Y_k, \boldsymbol{\theta}\right) = \frac{\exp\left(-\frac{\left(\sigma_k\sqrt{\Delta t_k}\right)^2+2\frac{1}{2}\sigma_k\sigma_k'\Delta t_k\left(Y_{k+1}-Y_k-\left(\mu_k-\frac{1}{2}\sigma_k\sigma_k'\right)\Delta t_k\right)}{4\left(\frac{1}{2}\sigma_k\sigma_k'\Delta t_k\right)^2}\right)}{\sqrt{2\pi}\sqrt{\left(\sigma_k\sqrt{\Delta t_k}\right)^2+4\frac{1}{2}\sigma_k\sigma_k'\Delta t_k\left(Y_{k+1}-Y_k-\left(\mu_k-\frac{1}{2}\sigma_k\sigma_k'\right)\Delta t_k\right)}}$$

$$\cdot\left[\exp\left(-\frac{\sigma_k\sqrt{\Delta t_k}\sqrt{\left(\sigma_k\sqrt{\Delta t_k}\right)^2+4\frac{1}{2}\sigma_k\sigma_k'\Delta t_k\left(Y_{k+1}-Y_k-\left(\mu_k-\frac{1}{2}\sigma_k\sigma_k'\right)\Delta t_k\right)}}{4\left(\frac{1}{2}\sigma_k\sigma_k'\Delta t_k\right)^2}\right)\right.$$

$$\left.+\exp\left(\frac{\sigma_k\sqrt{\Delta t_k}\sqrt{\left(\sigma_k\sqrt{\Delta t_k}\right)^2+4\frac{1}{2}\sigma_k\sigma_k'\Delta t_k\left(Y_{k+1}-Y_k-\left(\mu_k-\frac{1}{2}\sigma_k\sigma_k'\right)\Delta t_k\right)}}{4\left(\frac{1}{2}\sigma_k\sigma_k'\Delta t_k\right)^2}\right)\right]$$

$$= \frac{\exp\left(-\frac{C_k(Y_{k+1})}{D_k}\right)}{\sqrt{2\pi}\sqrt{\Delta t_k}\sqrt{A_k(Y_{k+1})}}\cdot\left[\exp\left(-\frac{\sqrt{A_k(Y_{k+1})}}{D_k}\right)+\exp\left(\frac{\sqrt{A_k(Y_{k+1})}}{D_k}\right)\right]$$

with

$$A_k(Y_{k+1}) = \left(\sigma_k\right)^2+2\sigma_k\sigma_k'\left(Y_{k+1}-Y_k-\left(\mu_k-\frac{1}{2}\sigma_k\sigma_k'\right)\Delta t_k\right)$$

$$C_k(Y_{k+1}) = \sigma_k+\sigma_k'\left(Y_{k+1}-Y_k-\left(\mu_k-\frac{1}{2}\sigma_k\sigma_k'\right)\Delta t_k\right)$$

$$D_k = \sigma_k\left(\sigma_k'\right)^2\Delta t_k$$

and for

$$Y_{k+1} \geq Y_k - \frac{1}{2}\frac{\sigma_k}{\sigma_k'} + \left(\mu_k - \frac{1}{2}\sigma_k\sigma_k'\right)\Delta t_k, \qquad \text{if } \sigma_k\sigma_k' > 0, \text{ and}$$
$$Y_{k+1} \leq Y_k - \frac{1}{2}\frac{\sigma_k}{\sigma_k'} + \left(\mu_k - \frac{1}{2}\sigma_k\sigma_k'\right)\Delta t_k, \qquad \text{if } \sigma_k\sigma_k' < 0.$$

In the case of $\sigma_k = 0$, Y_{k+1} conditioned on Y_k is deterministic. For $\sigma_k' = 0$, the Milstein scheme reduces to the Euler scheme. □

The bounds in (4.7) and (4.8) coincide with the bound (4.5) on the range of possible values Y_{k+1} resulting from the Milstein scheme. For values of Y_{k+1} within the respective bound, $A_k(Y_{k+1})$ is non-negative and its square root takes real values; otherwise, the transition density is equal to zero. Hence, there is a lower or an upper bound on the support of π^{Mil}. Moreover, one can show that the value of the transition density tends to infinity as Y_{k+1} approaches the bound. However, the interval for which the density increases towards infinity may be arbitrarily narrow depending on the parameter setting.

For the GBM, we have $\sigma(X_t, \boldsymbol{\theta}) = \sigma X_t$ with parameter $\sigma > 0$, the process taking values in $\mathbb{R}_+$. Therefore, we obtain a lower bound for the possible values of Y_{k+1}:

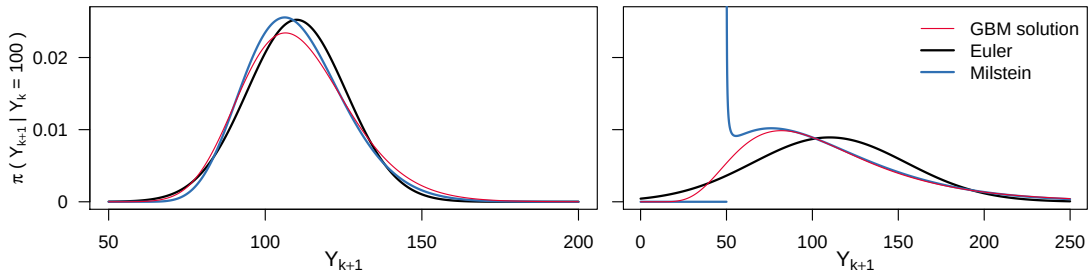
$$Y_{k+1} \geq Y_k\left(\frac{1}{2} + \left(\alpha - \frac{1}{2}\sigma^2\right)\Delta t_k\right). \tag{4.9}$$

Depending on the parameter combination $\boldsymbol{\theta} = (\alpha, \sigma)^T$, this lower bound may be negative, in which case the support of the transition density includes the entire state space of the GBM.

In Figure 4.3, we illustrate the transition densities based on the GBM solution, Euler scheme, and Milstein scheme for two different parameter settings. We observe that the Milstein transition density better approximates the mode of the transition density of the solution than the Euler transition density does. On the other hand, while the support of the Euler transition density is the set of all real numbers, the Milstein transition density puts zero weight on the values of Y_{k+1} that are below the lower bound (4.9), even though some of the values are feasible according to the transition density of the solution process.

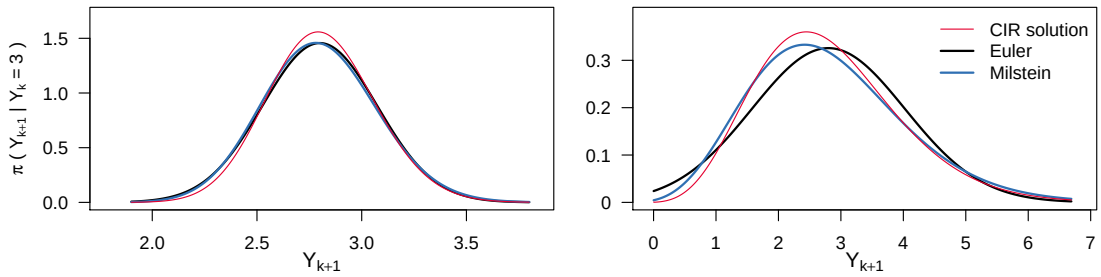
For the CIR process, we have $\sigma(X_t, \boldsymbol{\theta}) = \sigma\sqrt{X_t}$ with parameter $\sigma > 0$, the process taking values in $\mathbb{R}_0$. We therefore obtain a lower bound for the possible values of $X_{t_{k+1}}$ when applying the Milstein scheme:

$$X_{t_{k+1}} \geq \left(\alpha(\beta - X_{t_k}) - \frac{1}{4}\sigma^2\right)\Delta t_k. \tag{4.10}$$



(a) $\alpha = 1$ and $\sigma^2 = 0.25$ **(b)** $\alpha = 1$ and $\sigma^2 = 2$

Figure 4.3: Transition densities for a transition from $Y_k = 100$ to Y_{k+1} with a time step of $\Delta t_k = 0.1$ for two different parameter settings based on the GBM solution, Euler scheme, and Miltstein scheme, respectively.



(a) $\alpha = 1$, $\beta = 1$, and $\sigma^2 = 0.25$ **(b)** $\alpha = 1$, $\beta = 1$, and $\sigma^2 = 5$

Figure 4.4: Transition densities for a transition from $Y_k = 3$ to Y_{k+1} with time step $\Delta t_k = 0.1$ for two different parameter settings based on the solution of the CIR process, Euler scheme, and Miltstein scheme, respectively.

Again, depending on the parameter combination $\boldsymbol{\theta} = (\alpha, \beta, \sigma)^T$, this lower bound may be negative, in which case the support of the transition density includes the entire state space of the CIR process.

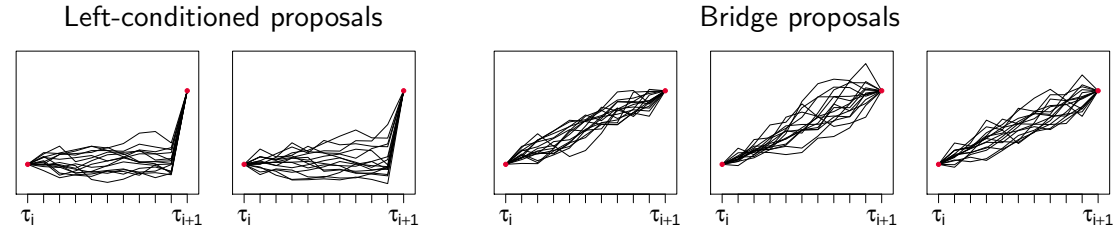
Figure 4.4 illustrates the transition densities based on the solution of the CIR process, Euler scheme, and Milstein scheme for two different parameter settings. For a small value of the diffusion parameter σ^2 as in Figure 4.4a, there is only little difference between the approximated transition densities based on the Euler and the Milstein scheme. This is also apparent from Equation (4.6). But for a larger value of σ^2 as in Figure 4.4b, we observe that the Milstein transition density again better approximates the mode of the transition density of the solution than the Euler transition density does.

Other approximation methods for the transition densities were developed for example in Aït-Sahalia (2002), Aït-Sahalia (2008), and Filipović et al. (2013). Here, we focus on the numerical approximation methods described in Section 3.3. Because for the Bayesian data augmentation

method for parameter estimation introduced in Section 3.4.1, it is crucial to not only be able to approximate the transition density, but also sampling from the resulting density needs to be possible and fast.

4.2 Path proposal methods based on the Milstein scheme

In this section, we explain how to incorporate the Milstein transition density derived in the previous section into the framework of Bayesian data augmentation as described in Section 3.4.1. The Milstein transition density can be used to approximate the likelihood of the diffusion path in both steps of the method, namely in the acceptance probability of the parameter update (3.23) and of the path update (3.25) as well as in the proposal density of the path update. For the LC proposal, we can simply plug the Milstein transition density into Equation (3.26). As already pointed out, this proposal method leads to a large jump in the last step from $X_{t_{m-1}}$ to $X_{\tau_{i+1}}$ which we illustrate by simulations for the GBM and both of the considered approximation schemes in Figures 4.5a and 4.5b.



(a) Left-cond. Euler proposal **(b)** Left-cond. Milstein proposal **(c)** Modified bridge Euler proposal **(d)** Modified bridge Milstein proposal **(e)** Diffusion bridge Milstein proposal

Figure 4.5: Realizations of the two proposal strategies using different approximation schemes. Fifteen proposed paths for the GBM with $\alpha = \sigma^2 = 0.1$ on the interval $[\tau_i, \tau_{i+1}] = [0, 1]$ with $X_{\tau_i} = 10$, $X_{\tau_{i+1}} = 25$, and $m = 10$ subintervals.

We now consider the Milstein approximation for the MB proposal, namely for the two factors on the right-hand side of (3.27). The first factor for the transition from X_{t_k} to $X^*_{t_{k+1}}$ resembles the Milstein transition density stated in Theorem 4.1. With the same notation, $\Delta_+ = t_m - t_{k+1}$, and $t_m = \tau_{i+1}$, the second factor for the transition from $X^*_{t_{k+1}}$ to X_{t_m} is as follows:

$$\pi^{Mil}\left(X_{t_m}|X^*_{t_{k+1}}, \boldsymbol{\theta}\right) = \frac{\exp\left(-\dfrac{F_m(X^*_{t_{k+1}})}{G_m(X^*_{t_{k+1}})}\right)}{\sqrt{2\pi}\sqrt{\Delta_+}\sqrt{E_m(X^*_{t_{k+1}})}} \quad \times$$
$$\left[\exp\left(-\frac{\sqrt{E_m(X^*_{t_{k+1}})}}{G_m(X^*_{t_{k+1}})}\right) + \exp\left(\frac{\sqrt{E_m(X^*_{t_{k+1}})}}{G_m(X^*_{t_{k+1}})}\right)\right]$$

with

$$E_m(X^*_{t_{k+1}}) = \left(\sigma^*_{k+1}\right)^2 + 2\sigma^*_{k+1}\sigma^{*\prime}{}_{k+1}\left(X_{t_m} - X^*_{t_{k+1}} - \left(\mu^*_{k+1} - \frac{1}{2}\sigma^*_{k+1}\sigma^{*\prime}{}_{k+1}\right)\Delta_+\right),$$
$$F_m(X^*_{t_{k+1}}) = \sigma^*_{k+1} + \sigma^{*\prime}{}_{k+1}\left(X_{t_m} - X^*_{t_{k+1}} - \left(\mu^*_{k+1} - \frac{1}{2}\sigma^*_{k+1}\sigma^{*\prime}{}_{k+1}\right)\Delta_+\right),$$
$$G_m(X^*_{t_{k+1}}) = \sigma^*_{k+1}\left(\sigma^{*\prime}{}_{k+1}\right)^2\Delta_+$$

for $E_m(X^*_{t_{k+1}}) \geq 0$ (which cannot be rearranged for $X^*_{t_{k+1}}$ in general); otherwise, the density is equal to zero. The terms μ^*_{k+1} and σ^*_{k+1} are similar to μ_{k+1} and σ_{k+1}, but $X_{t_{k+1}}$ is replaced by $X^*_{t_{k+1}}$. Here, we do not respectively approximate μ_{k+1} and σ_{k+1} by μ_k and σ_k because doing so does not lead to simplification. Moreover, there is no closed formula for the normalization constant needed to scale the product of the two transition densities to a proper density.

For the GBM, we have $X_t > 0$ and $\sigma^*_{k+1} = \sigma X^*_{t_{k+1}} > 0$ and thus, obtain the following bounds for $\pi^{Mil}\left(X_{t_m}|X^*_{t_{k+1}}, \boldsymbol{\theta}\right)$, the second factor in (3.27):

$$X^*_{t_{k+1}} \leq \frac{X_{t_m}}{\frac{1}{2} + \left(\alpha - \frac{1}{2}\sigma^2\right)\Delta_+} =: u_{2nd}, \qquad \text{if } \frac{1}{2} + \left(\alpha - \frac{1}{2}\sigma^2\right)\Delta_+ > 0 \quad \text{(Case I)},$$
$$X^*_{t_{k+1}} \geq \frac{X_{t_m}}{\frac{1}{2} + \left(\alpha - \frac{1}{2}\sigma^2\right)\Delta_+} =: l_{2nd}, \qquad \text{if } \frac{1}{2} + \left(\alpha - \frac{1}{2}\sigma^2\right)\Delta_+ < 0 \quad \text{(Case II)},$$
$$\text{and } X^*_{t_{k+1}} \geq 0, \qquad \text{if } \frac{1}{2} + \left(\alpha - \frac{1}{2}\sigma^2\right)\Delta_+ = 0 \quad \text{(Case III)}.$$

From (4.9), we obtain the following lower bound for $\pi^{Mil}\left(X^*_{t_{k+1}} \,|\, X^*_{t_k}, \boldsymbol{\theta}\right)$, the first factor in (3.27):

$$X^*_{t_{k+1}} \geq X^*_{t_k}\left(\frac{1}{2} + \left(\alpha - \frac{1}{2}\sigma^2\right)\Delta t_k\right) =: l_{1st}.$$

At the same time, proposals $X^*_{t_{k+1}}$ for the GBM should always be strictly positive to be in the state space. Let $l := \max\{0, l_{1st}\}$. The constraints on $X^*_{t_{k+1}}$ derived from the two factors in (3.27) lead to three cases for the set $\mathcal{D}$ of feasible points of $X^*_{t_{k+1}}$ for the GBM (assuming $X_{t_m} > 0$):

$$\mathcal{D} = \begin{cases} \emptyset, & \text{if (Case I) applies and } l_{1st} > u_{2nd}, \\ [l, u_{2nd}], & \text{if (Case I) applies and } l_{1st} \leq u_{2nd}, \\ [l, \infty), & \text{if (Case II) or (Case III) apply.} \end{cases}$$

Since the MB proposal takes into account information not only from the left data point but also from the observation on the right, it does not have a large jump in the last step as the left-conditioned proposal does. This is also apparent in the simulations for the GBM in Figures 4.5c and 4.5d. Therefore, the acceptance probability and acceptance rate are usually higher for the MB proposal than for the left-conditioned proposal. As Appendix A.2.2 demonstrates, the acceptance probability is even equal to 1 for the MB proposal if only one data point is imputed between two observations (i. e. the number of inter-observation intervals is $m = 2$). This holds when using the Milstein scheme to approximate the transition density for the likelihood function and proposal density, but also when using the Euler scheme without the approximation of μ_{k+1} and σ_{k+1} by μ_k and σ_k, respectively.

For the CIR process, we have obtained the lower bound in Equation (4.10) for the possible values of $X_{t_{k+1}}$ when applying the Milstein scheme:

$$l_{left} := \left(\alpha\left(\beta - X_{t_k}\right) - \frac{1}{4}\sigma^2\right)\Delta t_k.$$

The second bound that occurs when combining the MB proposal with the Milstein scheme is as follows:

$$X_{t_{k+1}} \geq \beta - \frac{1}{\alpha}\left(\frac{1}{\Delta_+}X_{t_m} + \frac{1}{4}\sigma^2\right) =: l_{right}.$$

The set $\mathcal{D}$ of feasible points $X_{t_{k+1}}$ for the CIR process when combining the MB proposal with the Milstein scheme is thus $\mathcal{D} = [l, \infty)$ with $l := \max\left(0, l_{left}, l_{right}\right)$.

The density of the MB proposal based on the Euler scheme in Equation (3.28) can also be interpreted as the density that results from applying the Euler scheme to the following diffusion process:

$$\mathrm{d}X_t = \left(\frac{X_{\tau_{i+1}} - X_t}{\tau_{i+1} - t}\right)\mathrm{d}t + \sqrt{\frac{\tau_{i+1} - t_{k+1}}{\tau_{i+1} - t}}\sigma\left(X_t, \boldsymbol{\theta}\right)\mathrm{d}B_t$$

for $t \in [t_k, t_{k+1}]$. See Whitaker et al. (2017) for a detailed discussion of the connection between the modified bridge and this continuous-time conditioned process. Applying the Milstein scheme to this process yields another proposal scheme to which we refer as the *diffusion bridge Milstein (DBM) proposal*. For the DBM proposal, the proposal density of a path segment also factorizes as:

$$q_{DBM}\left(X^{imp*}_{(\tau_i,\tau_{i+1})} \,\middle|\, X_{\tau_i,}, X_{\tau_{i+1}}, \boldsymbol{\theta}\right) = \prod_{k=0}^{m-2} \pi\left(X^*_{t_{k+1}} \,|\, X^*_{t_k}, X_{\tau_{i+1}}, \boldsymbol{\theta}\right),$$

where $X^*_{t_0} = X_{\tau_i}$, and each factor $\pi\left(X^*_{t_{k+1}} \,|\, X^*_{t_k}, X_{\tau_{i+1}}, \boldsymbol{\theta}\right)$ corresponds to the density based on the Milstein scheme from Theorem 4.1 where we replace

- μ_k by $(X_{\tau_{i+1}} - X_{t_k})/(\tau_{i+1} - t_k)$,
- σ_k by $\sqrt{(\tau_{i+1} - t_{k+1})/(\tau_{i+1} - t_k)}\sigma(X_{t_k}, \boldsymbol{\theta})$, and
- σ'_k by $\sqrt{(\tau_{i+1} - t_{k+1})/(\tau_{i+1} - t_k)}\, \partial\sigma\left(y, \boldsymbol{\theta}\right)/\partial y\,|_{y=X_{t_k}}$.

Like the MB proposal, the DBM proposal takes into account information from the observation on the right and; therefore, it does not have a large jump in the last step as illustrated in Figure 4.5e.

We have discussed another challenge in the context of Bayesian data augmentation and the MCMC scheme in Section 3.4.2: the dependence between the parameter components included in the diffusion function and the missing path segments between two observations that leads to a slower convergence of the MCMC algorithm as the number of imputed points $m - 1$ increases. However, since all estimation methods compared here are affected by this issue in the same way; we do not further consider it here.

To our knowledge, we are the first to utilize the Milstein scheme in the MCMC context described here.

4.2.1 Implementation

The implementation is relatively straightforward for the majority of the estimation procedures, and only the combination of the MB proposal and the Milstein approximation requires additional explanation. As mentioned, when approximating the two factors on the right-hand side of (3.27) by the transition density based on the Milstein scheme, there is no closed formula for the normalization constant to obtain a proper density. The normalization is necessary because the proposal density for a path segment is the product of several of the terms from (3.27), where the condition on the left point, $X^*_{t_k}$, differs between a newly proposed segment and the last accepted segment if several consecutive points are imputed. Therefore, the normalization constants differ and do not cancel out in the acceptance probability. Normalization is not necessary only in the case where just one point is imputed between two observations (i. e. $m = 2$ subintervals) because the left point, X_{t_k}, is always a (fixed) observed point that is not updated. Thus, the normalization constants cancel out in the acceptance probability. For $m > 2$, we numerically integrate the product (3.27) over $X_{t_{k+1}}$ to obtain the normalization constant. The product in (3.27) may be very small (but not zero everywhere in a non-empty feasible set $\mathcal{D}$) and may thus numerically integrate to zero, especially when the upper interval bound of the feasible set is infinite. To overcome this problem, we take two measures. First, we do not integrate over the entire set of feasible points but determine the maximum of the

product numerically and then integrate over the interval that includes all points with a function value of at least 10^{-20} times this maximum. Second, we rescale the product in (3.27) by dividing by the maximum before integrating.

To sample from the Milstein MB proposal density, we employ rejection sampling. For this, normalization of the product in (3.27) is not necessary. Again, we numerically determine the maximum d_{max} of the product, and the interval $\mathcal{I}$ that includes all points with a function value of at least 10^{-20} times this maximum. Then, we uniformly sample (u_1, u_2) from rectangle $\mathcal{I} \times (0, d_{max})$ and accept u_1 as a proposal $X^*_{t_{k+1}}$ if the unnormalized density value of (3.27) at u_1 is at most u_2.

For the combination of the MB proposal and the Milstein approximation, the set of feasible proposal points may be empty. In this case, our implementation shifts to the Euler approximation for this point, i. e. the point is proposed with the MB proposal based on the Euler scheme and also the corresponding factor of the proposal density in the acceptance probability is based on the Euler scheme. In addition, for all methods, a negative point may be proposed, which is not feasible for a GBM. Therefore, in this case, we propose a new point. For both cases, we count the number of times that they occur during the estimation procedure. In the following simulation study no cases of switching to the Euler scheme occurred and negative proposals occurred only very rarely (less than 1‰ of the number of iterations in the very worst case).

We implemented the described estimation procedures in R version 3.6.2 (R Core Team, 2019). The source code of our implementation and the following simulation study is publicly available at `https://github.com/fuchslab/Inference_for_SDEs_with_the_Milstein_scheme`.

4.3 Simulation study

Next, we study the computational performance of competing inference methods on the two benchmark models, the GBM and the CIR process. In Section 3.4.1 and Section 4.2, we have introduced a number of possible options for the choices to be made when constructing an estimation method in the framework of Bayesian data augmentation for diffusion processes:

- approximate the transition densities in the likelihood function based on the Euler or Milstein scheme,
- use the left-conditioned, the MB, or the DBM proposal, and
- use the Euler or Milstein scheme for the proposal densities (for the left-conditioned or MB proposal).

In the following, we will omit the left-conditioned proposal due to the inefficiency that we already pointed out. Instead, we will consider the following four combinations:

(MBE-E) MB proposal and transition density both based on the Euler scheme,

(MBE-M) MB proposal based on the Euler scheme and transition density based on the Milstein scheme,

(MBM-M) MB proposal and transition density both based on the Milstein scheme, and

(DBM-M) DBM proposal (which is based on the Milstein scheme) and transition density based on the Milstein scheme.

Combination MBE-M merges the Euler and Milstein scheme. We include it here because it combines the faster scheme for the proposals (where accuracy is less important) and the more accurate scheme for the acceptance probability.

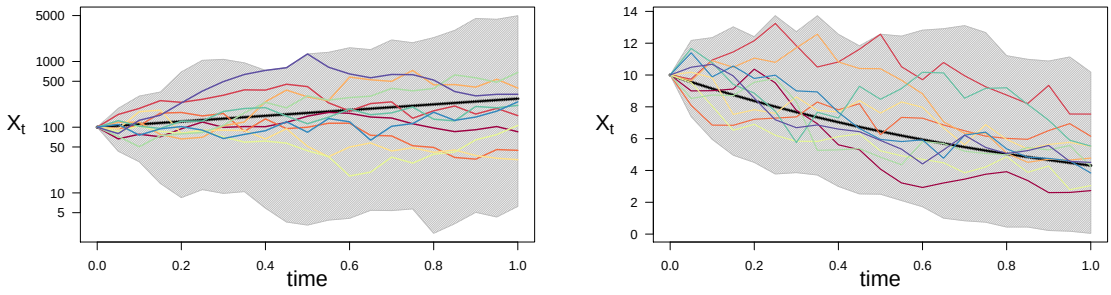
In this work, we focus on Bayesian inference by data augmentation and compare the four approaches listed above. Conceptually different inference procedures, as summarized in the beginning of Section 3.4 are not considered as competitors here as they would be employed in different data contexts. There are two aspects that are important to consider when we want to evaluate the different methods:

a) the accuracy with which the true posterior distribution is approximated based on one of the approximation schemes and a given number m and

b) the accuracy with which we are able to draw from this approximated posterior distribution.

We are interested in the overall accuracy, i. e. the combination of a) and b), achieved within a fixed amount of computational time.

For the simulation study, we generated 100 paths of both benchmark models in the time interval $[0, 1]$ using the exact transition densities stated in (3.7) for the GBM and in (3.10) for the CIR process. From each path, we took $M = 20$ points observed at equidistant time points (i. e. the inter-observation time Δt is 0.05) and applied each of the four described estimation methods once. We imputed data such that we got $m = 2$ and $m = 5$ inter-observation intervals. We also included the case $m = 1$, i. e. no data was imputed and only Step (1) from Section 3.4.1, the parameter update, was repeated in the estimation procedure where the likelihood of the path in the acceptance probability is approximated by the Euler or the Milstein scheme.



(a) GBM with $\boldsymbol{\theta} = (1, 2)^{\mathsf{Tr}}$ and $x_0 = 100$

(b) CIR process with $\boldsymbol{\theta} = (1, 1, 2)^{\mathsf{Tr}}$ and $x_0 = 10$

Figure 4.6: Trajectories used in the simulation study. The colored lines are 10 examples of the 100 trajectories used in the simulation study. Each trajectory consists of 20 points used as observations. The grey-shaded area shows the range of the 100 trajectories. The solid black line represents the expected value of (a) the GBM $\mathbb{E}[X_t] = X_0 \exp(\alpha t) = 100 \exp(t)$ and (b) the CIR process $\mathbb{E}[X_t] = \beta - (\beta - X_0) \exp(-\alpha t) = 1 + 9 \exp(-t)$.

Each of the estimation procedures performs the following steps:

1. Draw initial values for the parameter $\boldsymbol{\theta}$ from the prior distributions.
2. Initialize Y^{imp} by linear interpolation.
3. Repeat the following steps:
   (a) Parameter update: Apply random walk proposals.
      i. Draw a proposal for each component of the parameter $\boldsymbol{\theta}$.
      ii. Accept the proposals for all components or none.
   (b) Path update:
      i. Choose an update interval (t_a, t_b) as described in Appendix A.2.1 with $\lambda = 5$.
      ii. Draw a proposal $X^{imp*}_{(t_a,t_b)}$ according to the investigated method.
      iii. Accept or reject the proposal.

We let each procedure run for one hour and evaluate the overall accuracy of the obtained sample compared to a sample from the true posterior distribution (as described below).

For the GBM, the paths for the simulation study were generated with the parameter combination $\boldsymbol{\theta} = \left(\alpha, \sigma^2\right)^{\mathsf{Tr}} = (1, 2)^{\mathsf{Tr}}$ and initial value $x_0 = 100$. Figure 4.6a illustrates some of these paths. For the prior distribution of the parameters, we assumed that they were independently distributed with $\alpha \sim \mathcal{N}(0, 10)$ and $\sigma^2 \sim \mathrm{IG}(\kappa_0 = 2, \nu_0 = 2)$, where IG denotes the inverse gamma distribution with shape parameter κ_0 and scale parameter ν_0. The a priori expectations

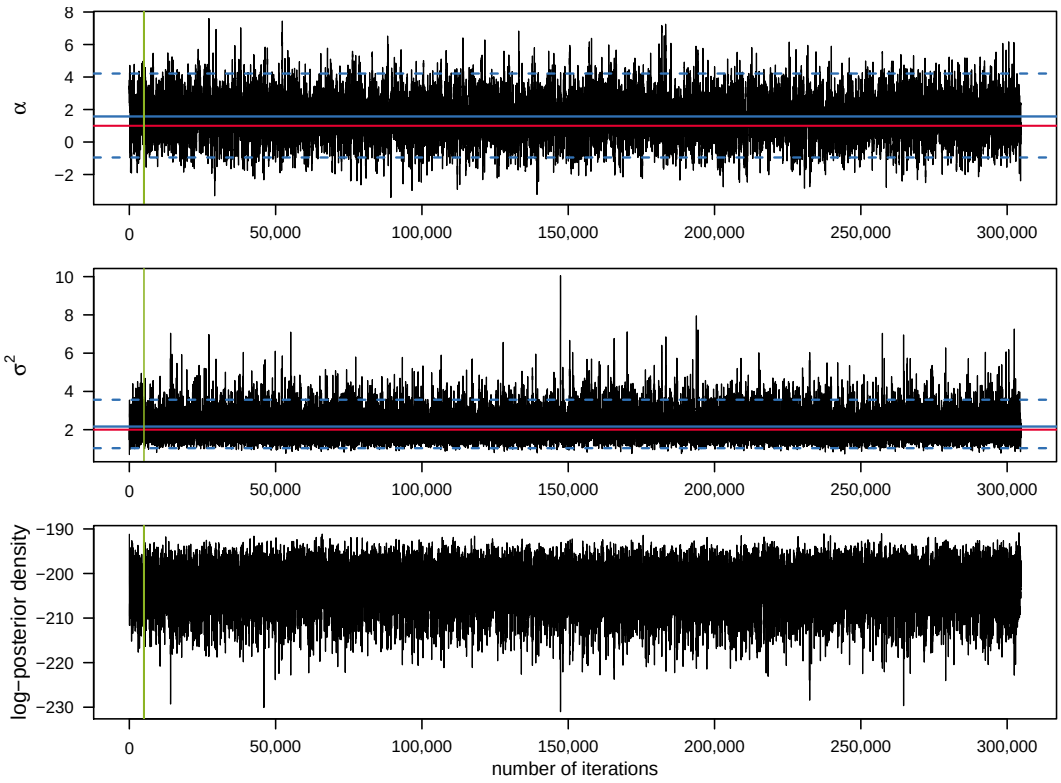


Figure 4.7: Trace plots of the MCMC chains for parameters α and σ^2 of the GBM (3.6) and of the log-posterior density values for one parameter estimation run using the combination MBM-M of the modified bridge proposal with $m = 2$ and the Milstein approximation for the proposal density and the likelihood function. The red lines represent the true values of parameters $\alpha = 1$ and $\sigma^2 = 2$, the blue solid lines represent the mean, and the blue dashed lines represent the lower and upper bounds of the highest-probability density interval of 95% after cutting off the first 5000 values of the chains as burn-in, which is represented by the green line.

of the parameters are thus $\mathbb{E}(\alpha) = 0$ and $\mathbb{E}\left(\sigma^2\right) = 2$. As proposal densities for the parameters in Step (3a), we used $\alpha^* \sim \mathcal{N}(\alpha_{i-1}, 0.25)$ and $\sigma^{2*} \sim \mathcal{LN}(\log \sigma^2_{i-1}, 0.25)$.

Figures 4.7 and 4.8 present the output from one estimation procedure for the GBM on the example of the combination MBM-M of the MB proposal and the Milstein approximation for the proposal density and the likelihood function. From each estimation procedure, we obtained an MCMC chain of dimension $n(m-1)+2$. For each chain, we used the two components for parameters α and σ^2 and calculated the mean, the median, and the variance after cutting off a burn-in phase of 5000 iterations.To justify our use of independent proposals for the parameter update, we show in Appendix A.2.3 that the parameters are not strongly correlated.

For the CIR process, we generated the 100 paths with the parameter combination $\boldsymbol{\theta} = \left(\alpha, \beta, \sigma^2\right)^{\mathsf{Tr}} = (1, 1, 2)^{\mathsf{Tr}}$ and initial value $x_0 = 10$. Some of the paths are illustrated in Fig-

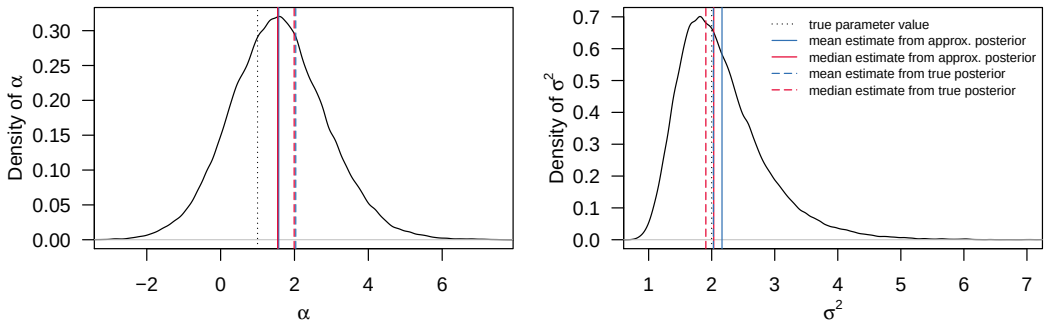


Figure 4.8: Estimated posterior densities for α and σ^2 from one parameter estimation run using the combination MBM-M of the modified bridge proposal and the Milstein approximation for the proposal and the transition density. Moreover, true values of the parameters, the mean and the median of the MCMC chains after 5000 iterations burn-in, and the mean and the median of a sample from the true posterior distribution of the sample path based on the solution of the GBM are shown.

ure 4.6b. We assumed α to be known and performed the inference methods for the parameters β and σ^2. For the prior distribution of the parameters, we assumed that they were independently distributed with $\beta \sim \mathrm{IG}\left(\kappa_b = 3, \nu_b = 3\right)$ and $\sigma^2 \sim \mathrm{IG}(\kappa_s = 3, \nu_s = 4)$. The a priori expectations of the parameters are thus $\mathbb{E}(\beta) = \frac{3}{2}$ and $\mathbb{E}\left(\sigma^2\right) = 2$. As proposal densities for the parameters in Step (3a), we used $\beta^* \sim \mathcal{LN}(\log \beta_{i-1}, 0.25)$ and $\sigma^{2*} \sim \mathcal{LN}(\log \sigma^2_{i-1}, 0.25)$.

As a benchmark, we also sampled from the true parameter posterior distribution based on the exact transition densities. We used the Stan software (Carpenter et al., 2017, Stan Development Team, 2019) which provides an efficient C++ implementation of the HMC-based No-U-turn sampler as briefly described in Section 2.2.2 to sample from the true parameter posterior distribution. For each posterior distribution corresponding to one of the 100 sample paths, we generated four HMC chains with 500,000 iterations each. The first half of the chains was discarded as warm-up and the remaining draws were combined to give a sample of size 10^6. We calculated the multivariate effective sample size (ESS) as defined in Vats et al. (2019) which provides the size of an independent and identically distributed sample equivalent to our samples in terms of variance and found that the ESS of the obtained samples from the true posterior distribution is well over 500,000. For each of these samples, we also calculated the mean, the median, and the variance and compared them to the respective summary statistic of the samples from the approximated posterior distribution.

The estimation procedures and time measurements were performed on a cluster of machines with the following specifications: AMD Opteron(TM) Processor 6376 (1.40GHz), 512GB DDR3-RAM.

4.3.1 Results for the GBM

Figures 4.9 and 4.10 and Tables 4.1 and 4.2 summarize the results of running each of the methods once for one hour for each of the 100 GBM trajectories. Figures 4.9 and 4.10 show the density plots of the difference between the respective statistic (mean, median, or variance) calculated for a sample from the approximated posterior distribution obtained by the respective method and the statistic for a sample from the true posterior distribution of the same sample path. Each density plot aggregates 100 such difference values, one for each of the 100 GBM trajectories.

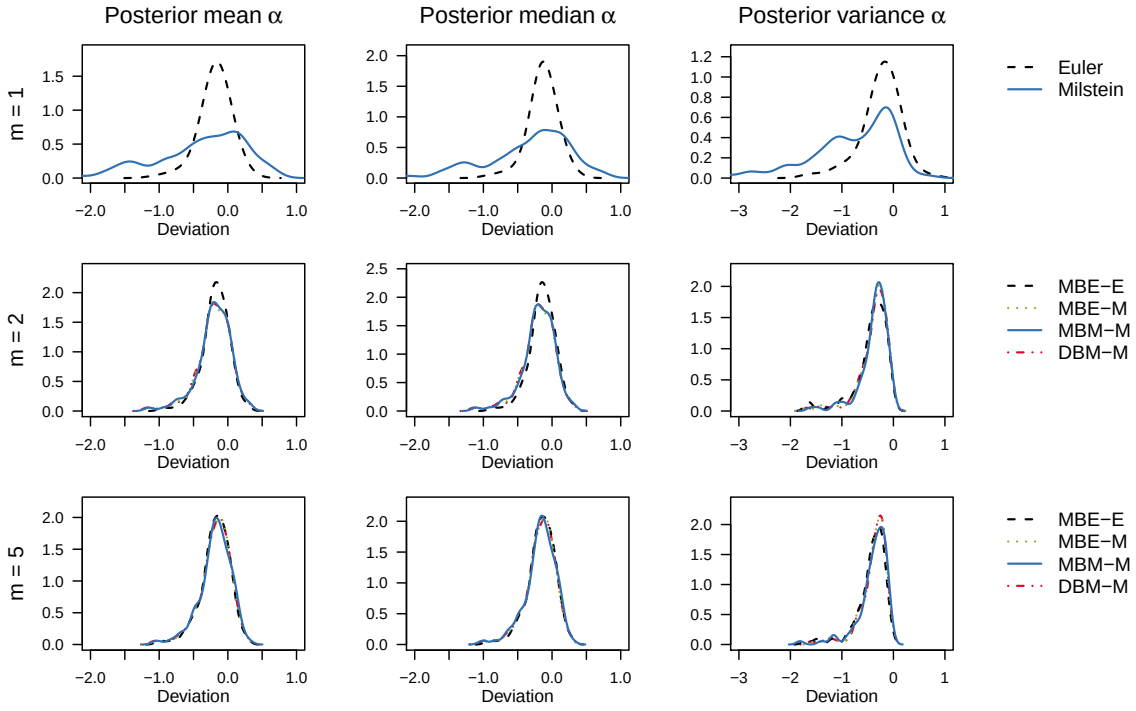


Figure 4.9: Sampling results for α obtained by each of the estimation procedures. Each density plot aggregates 100 deviations between the respective statistics (left: mean, middle: median, right: variance) calculated for the sample from the approximated posterior and for the sample from the true posterior distribution, one for each of the 100 sample paths of the GBM. The rows show results for different numbers m of subintervals between two observations. For $m = 1$, no data points were imputed and only Step (1) in Section 3.4.1, the parameter update, was repeated in the estimation procedure.

Table 4.1 tabulates the root mean square error (RMSE) based on these differences for each of the considered methods, discretization levels m, and statistics. We use the RMSE as the measure of the overall accuracy. The lower the RMSE is, the higher the accuracy of the respective method. Table 4.2 empirically evaluates the computational efficiency of the considered methods, including the number of iterations completed after one hour, the multivariate ESS based on the obtained sample after discarding a burn-in phase of 5000 iterations, and the acceptance

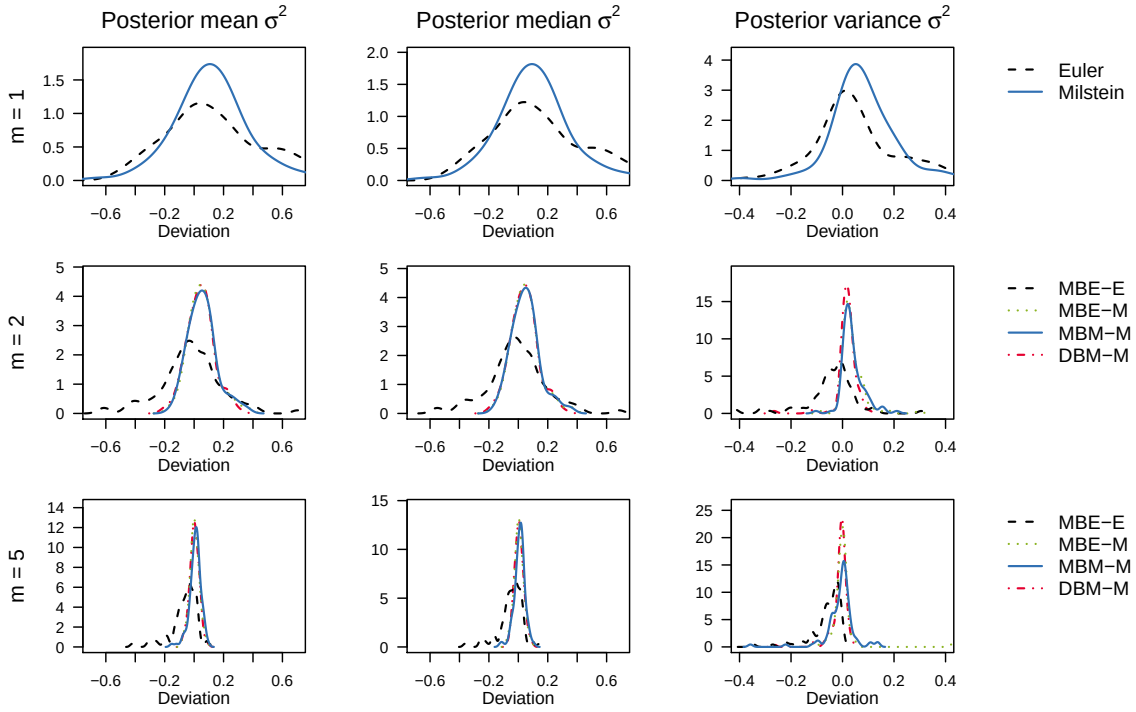


Figure 4.10: Sampling results for σ^2 of the GBM as described in Figure 4.9.

rates of the parameter and the path proposals. Each of these quantities is averaged over the 100 GBM trajectories and the coefficient of variation is also stated.

For the drift parameter α of the GBM, the four considered schemes perform comparably for $m = 2$ and $m = 5$. In particular, the use of the Milstein approximation does not improve the accuracy of the posterior mean and median for the same discretization level m. The accuracy of the posterior variance is slightly improved by the use of the Milstein approximation when data are imputed. Moreover, for MBE-E, the accuracy does not consistently improve as m is increased. Whereas, the accuracy for the methods including the Milstein scheme improves considerably when imputed data are introduced (i. e. $m > 1$) and it improves slightly when m is increased from 2 to 5.

For the diffusion parameter σ^2 of the GBM, we clearly see an improvement in overall accuracy for the methods involving the Milstein scheme. Combination DBM-M turns out to be the most accurate, closely followed by MBE-M in the case of the mean and median.

According to Table 4.2, the number of iterations completed within one hour varies substantially among the different estimation procedures. It is always higher for the procedures that use the Euler approximation, while especially Combination MBM-M is very time-consuming and thus completes fewer iterations. Similarly, the multivariate ESS varies substantially among the different estimation procedures. It is higher for $m = 2$ than for $m = 5$ for each of the

Table 4.1: Empirical characteristics for evaluating the overall accuracy of the parameter estimation procedures for different numbers m of subintervals between two observations aggregated over 100 deviations between the respective statistics calculated for the sample from the approximated posterior and for the sample from the true posterior distribution, one for each of the 100 sample paths of the GBM. The lowest RMSE per m and per statistic is printed in boldface.

	Method	RMSEs for α			RMSEs for σ^2		
		mean	median	variance	mean	median	variance
$m = 1$	Euler	**0.282**	**0.244**	**0.456**	0.638	0.600	0.471
	Milstein	0.851	0.780	1.158	**0.282**	**0.265**	**0.176**
$m = 2$	MBE-E	**0.266**	**0.238**	0.526	0.211	0.198	0.141
	MBE-M	0.311	0.302	0.476	0.109	0.106	0.057
	MBM-M	0.315	0.305	**0.470**	0.112	0.107	0.057
	DBM-M	0.318	0.308	0.485	**0.101**	**0.099**	**0.044**
$m = 5$	MBE-E	**0.277**	**0.254**	0.524	0.113	0.098	0.127
	MBE-M	0.288	0.274	0.474	0.031	0.031	0.050
	MBM-M	0.292	0.278	0.492	0.040	0.037	0.058
	DBM-M	0.291	0.275	**0.472**	**0.031**	**0.030**	**0.037**

RMSE denotes the root mean square error.

considered estimation procedures. The acceptance rate of the parameters is slightly lower when the Milstein scheme is used for the approximation of the likelihood function. In addition, the acceptance rate of the parameters decreases as the number of imputed points increases. The acceptance rate of the path is highest for Combination MBM-M. For MBE-E, it would be just as high if one did not substitute μ_{k+1} and σ_{k+1} by μ_k and σ_k. For MBE-E, MBE-M, and DBM-M, the acceptance rate of the path increases as the number of imputed points increases.

Table 4.2: Empirical characteristics for evaluating the computational efficiency of the parameter estimation procedures for different numbers m of subintervals between two observations aggregated over 100 trajectories of the GBM. Each of the procedures was run for one hour. Acceptance rates are defined to take values between 0 and 1. For $m = 1$, no data points were imputed and only Step (1) in Section 3.4.1, the parameter update, was repeated in the estimation procedure. Specifications for the computing power are stated in the main text.

	Method	Number of iterations after 1 hour		Multivariate effective sample size		Acceptance rate of the parameters		Acceptance rate of the path	
		mean	c.v.	mean	c.v.	mean	c.v.	mean	c.v.
$m = 1$	Euler	25134301	0.03	1273744	0.16	0.518	0.02	–	–
	Milstein	4454863	0.03	146362	0.41	0.425	0.14	–	–
$m = 2$	MBE-E	8583614	0.03	170827	0.19	0.442	0.01	0.842	0.04
	MBE-M	1816144	0.03	24090	0.38	0.417	0.03	0.799	0.05
	MBM-M	300870	0.03	6881	0.21	0.417	0.03	1.000	0.00
	DBM-M	1754024	0.10	28089	0.31	0.417	0.03	0.839	0.04
$m = 5$	MBE-E	6765054	0.10	49885	0.18	0.310	0.01	0.892	0.02
	MBE-M	892487	0.02	5033	0.24	0.304	0.01	0.844	0.03
	MBM-M	78215	0.04	573	0.20	0.304	0.01	0.978	0.01
	DBM-M	879227	0.03	5535	0.21	0.304	0.01	0.884	0.02

c.v. denotes the coefficient of variation.

4.3.2 Results for the CIR process

Figures 4.11 and 4.12 and Tables 4.3 and 4.4 summarize the results of the simulation study for the CIR process. Figures 4.11 and 4.12 show the density plots of the difference between the respective statistic (mean, median, or variance) calculated for a sample from the approximated posterior distribution obtained by the respective method and the statistic for a sample from the true posterior distribution of the same sample path. Table 4.3 tabulates the RMSE based on these differences for each of the considered methods, discretization levels m, and statistics, and Table 4.4 empirically evaluates the computational efficiency of the considered methods as explained in more detail in the beginning of the previous subsection.

Similar to the results for the GBM, the use of the Milstein approximation does not consistently improve the overall accuracy for the drift parameter β. The accuracy increases (i. e. the RMSE decreases) for increasing m for most of the methods. Only Combination MBM-M

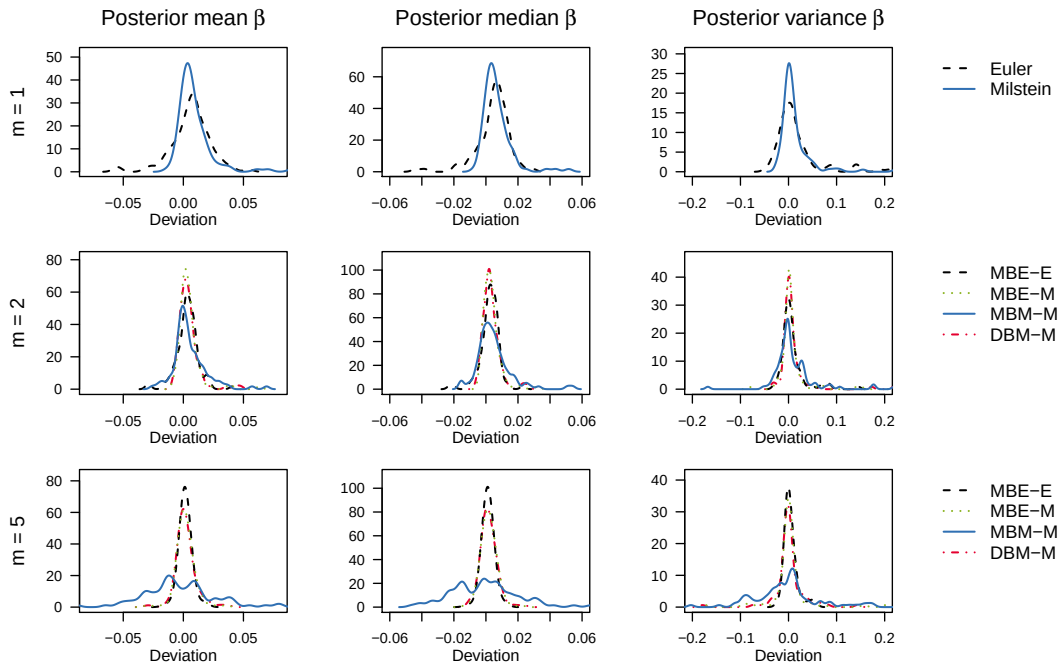


Figure 4.11: Sampling results for β obtained by each of the estimation procedures. Each density plot aggregates 100 deviations between the respective statistics (left: mean, middle: median, right: variance) calculated for the sample from the approximated posterior and for the sample from the true posterior distribution, one for each of the 100 sample paths of the CIR process. The rows show results for different numbers m of subintervals between two observations. For $m = 1$, no data points were imputed and only Step (1) from Section 3.4.1, the parameter update, was repeated in the estimation procedure.

has lower accuracy for $m = 5$ due to the low sampling efficiency and the resulting low ESS. For the diffusion parameter σ^2, the use of the Milstein approximation and increasing m both improve the overall accuracy. Again Combination DBM-M achieves the highest accuracy, closely followed by MBE-M.

Also for the CIR process, the number of iterations completed after one hour and the multivariate ESS of the obtained sample vary substantially between the different procedures. Both quantities are highest for Combination MBE-E, they are similar for MBE-M and DBM-M, and particularly low for MBM-M.

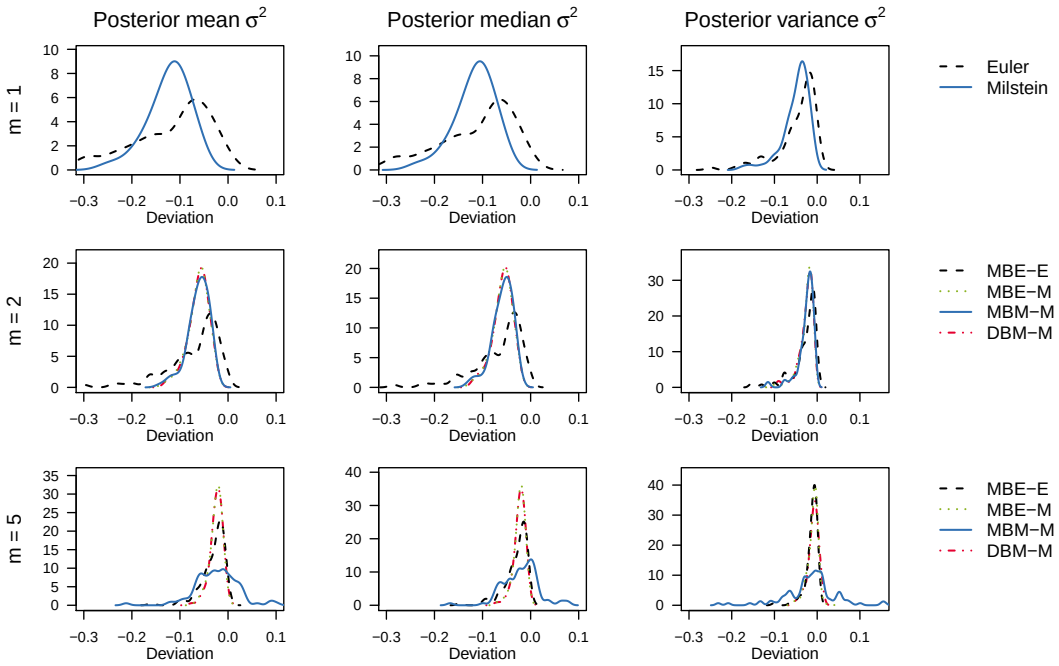


Figure 4.12: Sampling results for σ^2 of the CIR process as described in Figure 4.11.

Table 4.3: Empirical characteristics for evaluating the overall accuracy of the parameter estimation procedures for different numbers m of subintervals between two observations aggregated over 100 deviations between the respective statistics calculated for the sample from the approximated posterior and for the sample from the true posterior distribution, one for each of the 100 sample paths of the CIR process. The lowest RMSE per m and per statistic is printed in boldface.

	Method	RMSEs for β			RMSEs for σ^2		
		mean	median	variance	mean	median	variance
$m = 1$	Euler	0.0179	0.0115	**0.0478**	0.1603	0.1530	0.0673
	Milstein	**0.0174**	**0.0110**	0.0587	**0.1306**	**0.1233**	**0.0595**
$m = 2$	MBE-E	0.0099	0.0064	**0.0265**	0.0910	0.0865	0.0417
	MBE-M	0.0105	0.0063	0.0413	0.0656	0.0619	0.0309
	MBM-M	0.0151	0.0120	0.0462	0.0658	0.0625	0.0325
	DBM-M	**0.0097**	**0.0061**	0.0330	**0.0653**	**0.0617**	**0.0308**
$m = 5$	MBE-E	**0.0052**	**0.0036**	0.0144	0.0400	0.0380	0.0194
	MBE-M	0.0077	0.0049	0.0375	0.0271	0.0259	**0.0156**
	MBM-M	0.0307	0.0204	0.1103	0.0509	0.0420	0.0615
	DBM-M	0.0085	0.0052	**0.0321**	**0.0270**	**0.0256**	**0.0156**

RMSE denotes the root mean square error.

Table 4.4: Empirical characteristics for evaluating the computational efficiency of the parameter estimation procedures for different numbers m of subintervals between two observations aggregated over 100 trajectories of the CIR process. Each of the procedures was run for one hour. Acceptance rates are defined to take values between 0 and 1. For $m = 1$, no data points were imputed and only Step (1) from Section 3.4.1, the parameter update, was repeated in the estimation procedure. Specifications for the computing power are stated in the main text.

	Method	Number of iterations after 1 hour		Multivariate effective sample size		Acceptance rate of the parameters		Acceptance rate of the path	
		mean	c.v.	mean	c.v.	mean	c.v.	mean	c.v.
$m = 1$	Euler	23461023	0.11	2422521	0.14	0.443	0.03	–	–
	Milstein	4685450	0.03	480549	0.08	0.442	0.03	–	–
$m = 2$	MBE-E	8482241	0.06	422034	0.10	0.384	0.03	0.964	0.01
	MBE-M	1944229	0.05	94071	0.10	0.383	0.03	0.957	0.01
	MBM-M	186588	0.06	9429	0.13	0.383	0.03	1.000	0.00
	DBM-M	1905354	0.04	95262	0.10	0.383	0.03	0.968	0.01
$m = 5$	MBE-E	6851197	0.05	114344	0.10	0.272	0.03	0.976	0.01
	MBE-M	966579	0.04	15599	0.13	0.272	0.03	0.965	0.01
	MBM-M	37648	0.12	574	0.25	0.272	0.03	0.993	0.00
	DBM-M	906791	0.08	14881	0.14	0.272	0.03	0.975	0.01

c.v. denotes the coefficient of variation.

4.4 Summary and discussion

We have demonstrated how to implement an algorithm for the parameter estimation of SDEs from low-frequency data using the Milstein scheme to approximate the transition density of the underlying process. Our motivation was to improve numerical accuracy and thus reduce the amount of imputed data and computational overhead. However, our findings are rather discouraging: We found that this method can be applied to multidimensional processes only with impractical restrictions. Moreover, we showed that the combination of the MB proposal with the Milstein scheme for the proposal density may lead to an empty set of possible proposal points, which would require switching to the Euler scheme in order to proceed. One of the strengths of the original (Euler-based) MCMC scheme is its generic character and applicability. Through this, it possesses a practical advantage over otherwise more sophisticated methods such as the Exact Algorithm (Beskos et al., 2008). This strength does not translate to the Milstein-based MCMC scheme due to the limited applicability of the Milstein approximation especially in the multidimensional setting. Thus, methods like the Exact Algorithm may be a reasonable alternative. The limited applicability of the Milstein approximation would also persist for advanced forms of the discussed MCMC scheme like the innovation scheme in Golightly & Wilkinson (2008) or for even more generic algorithms like particle MCMC as studied in Golightly & Wilkinson (2011).

In our simulation study, we found that the overall accuracy for the estimates for the drift parameter of the GBM does not necessarily improve when the Milstein scheme is used. Fewer iterations are completed for the methods involving the Milstein scheme and also the ESS is substantially lower. Thus, the poor sampling efficiency might outweigh the (potential) increase in accuracy of the approximation of the posterior distribution. Especially the combination MBM-M results in a particularly low number of iterations and a low ESS. Due to the already quite low ESS achieved by the Milstein-based methods for $m = 5$ subintervals between two observations, we did not consider higher discretization levels. Moreover, note that tuning the variance hyperparameters for the random walk proposals of the parameters in Step 3a in the simulation study to reach an optimal acceptance rate might lead to a higher ESS. However, since the acceptance rates achieved in the simulation study lie in a range where the sampling efficiency is rather robust to changes in the acceptance rate as shown in Roberts & Rosenthal (2001) (in the high-dimensional limit), we do not expect the change in the ESS after tuning to be substantial.

For the estimates for the GBM diffusion parameter, the overall accuracy is increased by the use of the Milstein scheme. DBM-M turns out to be the most effective combination in terms of overall accuracy.

The results of the simulation study for the CIR process are very similar as for the GBM. The use of the Milstein approximation does not consistently improve the overall accuracy for the drift parameter; however, it does improve the accuracy for the diffusion parameter. Again Combination DBM-M achieves the highest accuracy, closely followed by MBE-M.

It was expected that the use of the Milstein scheme would make a difference for the estimates for the diffusion parameters because the additional term added by the Milstein scheme compared to the Euler scheme involves the diffusion function and its derivative. Nevertheless, the general applicability of the Euler scheme remains a great advantage and the search for different proposal schemes such as in Whitaker et al. (2017) and van der Meulen & Schauer (2017) rather than for different numerical discretization schemes may be a more promising way towards more efficient estimation algorithms for diffusion processes.

Chapter 5

Application: Modeling translation kinetics after mRNA transfection using diffusion processes

In this chapter, we apply Itô diffusion processes to model the translation kinetics after mRNA transfection and perform parameter inference for this model based on single-cell data from time-lapse fluorescence microscopy. mRNA transfection is the process of introducing mRNA into a living cell. mRNA delivery has become increasingly interesting for biomedical applications because it enables treatment of diseases by means of targeted expression of proteins and it is transient, avoiding the risk of permanently integrating into the genome (see e. g. Sahin et al., 2014). One of the most prominent applications of mRNA transfection at the moment are the mRNA-based vaccine candidates that are already in use or currently under investigation to prevent COVID-19 infections (DeFrancesco, 2020). In such a context, it is, of course, very import to have a precise understanding of the dynamics of the underlying processes in order to be able to control them. Yet, many aspects and the determinants of the mRNA delivery process and the translation kinetics are difficult to measure and therefore poorly understood.

One of the few ways to measure quantities within a living cell over time (i. e. keeping it alive is necessary) is the use of fluorescence reporters and fluorescence microscopy. Single-cell fluorescence data from transfection experiments has been analyzed based on ordinary differential equation (ODE) modeling in several previous studies e. g. Ligon et al. (2014), Leonhardt et al. (2014), Fröhlich et al. (2018), and Reiser et al. (2019). Here, we use the experimental data from Fröhlich et al. (2018) and investigate a stochastic differential equation (SDE) modeling approach. Our main interest lies in the question whether an SDE model allows

to identify more model parameters from experimental data compared to the corresponding ODE model. Moreover, we provide essential theoretical results for the SDE model.

Inference from fluorescence data for SDE models has also been conducted e. g. in Heron et al. (2007), Finkenstädt et al. (2008), and Komorowski et al. (2009), however for an experimental setup that also included the transcription process. Finkenstädt et al. (2008) even considered an SDE and an ODE model in one of their case studies, but their results did not directly show any differences in the parameter identifiability and the study was not focused on this aspect. When the main part of our study was conducted there was no published work that systematically compared the parameter identifiability between an SDE and an ODE model. Meanwhile, Browning et al. (2020) have recently published a study on this topic investigating four different example models with simulated data. We include their approach to structural identifiability analysis for SDE models in our study.

This chapter is composed as follows: We first describe the experimental data in Section 5.1 and formulate the reaction network which we want to consider for the translation kinetics and its ODE and diffusion approximation in Section 5.2. Then in Section 5.3, we prove the existence and uniqueness of the solution for the SDE formulation and the convergence of the Euler approximation to this solution which is one of the most notable contributions of this chapter. After stating the model of the observations in Section 5.4, we try to assess and compare the structural identifiability of the parameters for both modeling approaches in Section 5.5. We define the parameter posteriors for both modeling approaches in Section 5.6, study the practical identifiability of the parameters based on simulated and experimental data in Sections 5.7 and 5.8, respectively, and conclude with a summary and discussion of our findings in Section 5.9.

5.1 Experimental data

We consider data that was collected in the lab of Prof. Joachim Rädler at LMU Munich and has previously been analyzed (based on ODE modeling) and published in Fröhlich et al. (2018). The data was generated in an experiment where cells (human hepatoma epithelial cell line HuH7) were transfected with mRNA encoding a green fluorescence protein (GFP). The cells were fixed on micro patterned protein arrays and time lapse microscopy images of the cells were taken every 10 minutes over the course of at least 30 hours (i.e. there are at least 180 measurements per cell). For the first hour, mRNA lipoplexes were added. Afterwards, the cells were washed with cell culture medium such that no further lipoplex uptake occurs. The time point at which the lipoplexes were taken up, dissolved and released the mRNA as well as the number of mRNA molecules released are unknown.

The released mRNA was translated into a fluorescence protein which caused the cells to fluoresce. For each image taken during the experiment, the fluorescence intensity is integrated over squares occupied by one cell in order to obtain one value for the mean fluorescence intensity per cell and time point (see Fröhlich et al., 2018, for further details about the image analysis).

The experiment was conducted with two different types of GFP that differ in their protein lifetime: enhanced GFP (eGFP) and a destabilized enhanced GFP (d2eGFP). For each type of GFP, three replications of the experiment were conducted. We use the data from the experiment on April 27, 2016. It contains measurements for more than 800 cells for each type of GFP. Some trajectories of the mean fluorescence intensity are displayed in Figure 5.1.

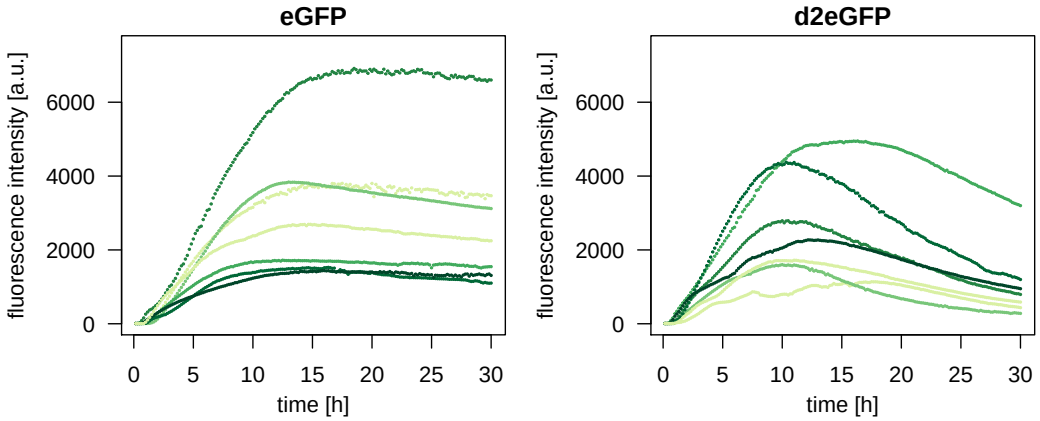


Figure 5.1: Trajectories of the mean fluorescence intensity for seven cells from the mRNA transfection experiment in Fröhlich et al. (2018) for eGFP and d2eGFP (April 27, 2016), respectively.

As will become clear in the course of this chapter, ODE models of the translation kinetics of an individual cell are not globally identifiable with the available experimental data as described above. Several of the ODE model parameters cannot be uniquely determined based on one observed fluorescence trajectory. Fröhlich et al. (2018) use a mixed-effect ODE model in order to incorporate the translation kinetics of several cells and data for both different types of GFP (eGFP and d2eGFP). Through this approach, they are able to improve parameter identifiability (by breaking the symmetry between the degradation rate constants); however, their approach is computationally very intense, required conducting the experiment with two types of GFP, and still leaves several parameters non-identifiable. Here, we are interested in the question whether the use of an SDE model can improve the parameter identifiability even when only one fluorescence trajectory is observed.

5.2 Modeling the translation kinetics

While Fröhlich et al. (2018) use a mixed-effect ODE model in order to incorporate the translation kinetics of several cells, we will focus on modeling the translation kinetics of one cell in order to study parameter identifiability based on one observed fluorescence trajectory.

We consider the basic model configuration that models only the (released) mRNA and the GFP molecules explicitly. Therefore, our model is a dynamic process with two components:

$$\boldsymbol{X}(t) = \begin{pmatrix} X_1(t) \\ X_2(t) \end{pmatrix} \quad \begin{matrix} \text{amount of mRNA molecules,} \\ \text{amount of GFP molecules.} \end{matrix}$$

We assume that all mRNA molecules (within one cell) are released at once at the initial time point denoted by t_0. Before t_0, there are neither mRNA nor GFP molecules, and at t_0, an amount of m_0 mRNA molecules is released, i.e.

$$\boldsymbol{X}(t) \equiv \begin{pmatrix} 0 \\ 0 \end{pmatrix} \text{ for } t < t_0 \quad \text{and } \boldsymbol{X}(t_0) = \begin{pmatrix} m_0 \\ 0 \end{pmatrix}.$$

Conceivable extensions of this basic model configuration are e. g. to include enzymatic degradation of the mRNA and/or the protein, ribosomal binding to the mRNA for translation, and a maturation step of the protein. However, we will only consider the basic configuration as described above.

5.2.1 Markov Jump Process

Assuming that the matter within the cell is well-stirred and in thermodynamic equilibrium, an Markov jump process (MJP) is regarded to be the most adequate representation of this system after t_0. In the basic model configuration, there are three possible reactions:

$$\begin{aligned} X_1 &\xrightarrow{\theta_1} \emptyset && \text{degradation of mRNA,} \\ X_1 &\xrightarrow{\theta_2} X_1 + X_2 && \text{translation,} \\ X_2 &\xrightarrow{\theta_3} \emptyset && \text{degradation of GFP.} \end{aligned}$$

The three reactions change the state of the system in the following way and occur with the following reaction rates:

$$\begin{aligned}
\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} &\longrightarrow \begin{pmatrix} X_1 - 1 \\ X_2 \end{pmatrix} && \text{with rate } \theta_1 X_1, \\
\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} &\longrightarrow \begin{pmatrix} X_1 \\ X_2 + 1 \end{pmatrix} && \text{with rate } \theta_2 X_1, \\
\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} &\longrightarrow \begin{pmatrix} X_1 \\ X_2 - 1 \end{pmatrix} && \text{with rate } \theta_3 X_2.
\end{aligned}$$

If we denote the probability distribution of the random variable $\boldsymbol{X}(t)$ by

$$P_{i,j}(t) = \mathbb{P}(X_1(t) = i, X_2(t) = j),$$

the corresponding chemical master equation (CME) reads

$$\frac{\partial P_{i,j}(t)}{\partial t} = \theta_1(i+1)P_{i+1,j}(t) + \theta_2 i P_{i,j-1}(t) + \theta_3(j+1)P_{i,j+1}(t) - (\theta_1 i + \theta_2 i + \theta_3 j)P_{i,j}(t).$$

Although the system contains only first-order reactions, there is no closed-form solution to the CME. Thus, there is no explicit formula for the transition probability distribution $p(\boldsymbol{X}(t)|\boldsymbol{X}(s), \theta)$ for $s < t$.

5.2.2 ODE model

The following system of ODEs is a deterministic approximation of the MJP modeling the dynamics as described above:

$$\frac{\mathrm{d}\boldsymbol{X}(t)}{\mathrm{d}t} = \begin{pmatrix} -\theta_1 X_1(t) \\ \theta_2 X_1(t) - \theta_3 X_2(t) \end{pmatrix} \quad \text{for } t \geq t_0. \tag{5.1}$$

This system admits the solution

$$\begin{aligned}
X_1(t) &= m_0 \exp\left(-\theta_1(t - t_0)\right), \\
X_2(t) &= \begin{cases} \frac{\theta_2 m_0}{\theta_3 - \theta_1}\left(e^{-\theta_1(t-t_0)} - e^{-\theta_3(t-t_0)}\right) & \text{, for } \theta_1 \neq \theta_3, \\ \theta_2 m_0 (t - t_0) e^{-\theta_3(t-t_0)} & \text{, for } \theta_1 = \theta_3. \end{cases}
\end{aligned} \tag{5.2}$$

Note that the solution for $X_2(t)$ is symmetric in the parameters θ_1 and θ_3.

5.2.3 SDE model

A stochastic but state-continuous approximation to the MJP in Section 5.2.1 is given by an Itô diffusion process that is described by the following SDE:

$$\mathrm{d}\boldsymbol{X}(t) = \begin{pmatrix} -\theta_1 X_1(t) \\ \theta_2 X_1(t) - \theta_3 X_2(t) \end{pmatrix} \mathrm{d}t + \begin{pmatrix} \sqrt{\theta_1 X_1(t)} & 0 \\ 0 & \sqrt{\theta_2 X_1(t) + \theta_3 X_2(t)} \end{pmatrix} \mathrm{d}\boldsymbol{B}(t) \quad (5.3)$$

for $t \geq t_0$ and where $\boldsymbol{B}(t)$ is a 2-dimensional standard Brownian motion.

Note that for a diffusion approximation (as well as for the ODE approximation), the size of the system can play an important role. However, since the model that we consider here contains only first-order reactions, the size of the system does not affect the interpretation of the kinetic parameters and does not need to be considered here.

5.3 Essential theoretical results for the SDE model

Before we can further consider the inference problem for the models introduced in the previous section, we need to ensure that SDE (5.3) is meaningful, i. e. that it admits a unique solution. Moreover, since there is no known explicit solution of SDE (5.3), we want to apply the Euler approximation. Hence, we need to show that this approximation scheme converges to the solution as the time step decreases. While the fundamental importance of both results should be obvious, their derivations are usually neglected when diffusion approximations are applied in systems biology. Due to the square root, the diffusion coefficient is not Lipschitz continuous and general SDE results such as Theorem 3.1 do not apply. Therefore, the proof of existence and uniqueness of the solution for SDE (5.3) and the proof of convergence of the Euler approximation are the subject of the following two subsections.

5.3.1 Existence and uniqueness of the solution

We first consider the following modified version of SDE (5.3):

$$\mathrm{d}X_1(t) = -\theta_1 X_1(t)\,\mathrm{d}t + \sqrt{\theta_1 X_1(t) \vee 0}\,\mathrm{d}B_1(t), \quad (5.4a)$$

$$\mathrm{d}X_2(t) = (\theta_2 X_1(t) - \theta_3 X_2(t))\,\mathrm{d}t + \sqrt{(\theta_2 X_1(t) + \theta_3 X_2(t)) \vee 0}\,\mathrm{d}B_2(t), \quad (5.4b)$$

$$X_1(t_0) = m_0, \quad (5.4c)$$

$$X_2(t_0) = 0, \quad (5.4d)$$

where $\vee$ denotes the max operator, i. e. $a \vee b := \max(a, b)$ for $a, b \in \mathbb{R}$.

Therefore, we consider the drift coefficient function

$$\boldsymbol{\mu}(\boldsymbol{x}) := \begin{pmatrix} \mu_1(x_1, x_2) \\ \mu_2(x_1, x_2) \end{pmatrix} = \begin{pmatrix} -\theta_1 x_1 \\ \theta_2 x_1 - \theta_3 x_2 \end{pmatrix}$$

and diffusion coefficient function

$$\boldsymbol{\sigma}(\boldsymbol{x}) := \begin{pmatrix} \sigma_1(x_1, x_2) & 0 \\ 0 & \sigma_2(x_1, x_2) \end{pmatrix} = \begin{pmatrix} \sqrt{\theta_1(x_1 \vee 0)} & 0 \\ 0 & \sqrt{(\theta_2 x_1 + \theta_3 x_2) \vee 0} \end{pmatrix}$$

for $\boldsymbol{x} = (x_1, x_2) \in \mathbb{R}^2$.

In the following, we assume without loss of generality that $t_0 = 0$.

Evidently, $\mu_1(x_1, x_2)$ and $\mu_2(x_1, x_2)$ are (Lipschitz) continuous functions. The square root function and thus also $\sigma_1(x_1, x_2)$ and $\sigma_2(x_1, x_2)$ are Hölder continuous with exponent $1/2$. Moreover, the squared norm of the coefficient functions can be estimated as follows

$$\begin{aligned} \left\| \begin{pmatrix} \mu_1(x_1, x_2) \\ \mu_2(x_1, x_2) \end{pmatrix} \right\|^2 &= \theta_1^2 x_1^2 + (\theta_2 x_1 - \theta_3 x_2)^2 \\ &\leq \theta_1^2 x_1^2 + \theta_2^2 x_1^2 + \theta_3^2 x_2^2 + \theta_2 \theta_3 (x_1^2 + x_2^2) \\ &\leq 3 \cdot \max\left\{\theta_1^2, \theta_2^2, \theta_3^2\right\} (x_1^2 + x_2^2 + 1) \\ &= C(\|\boldsymbol{x}\|^2 + 1) \end{aligned}$$

and

$$\begin{aligned} \left\| \begin{pmatrix} \sigma_1(x_1, x_2) & 0 \\ 0 & \sigma_2(x_1, x_2) \end{pmatrix} \right\|^2 &= (\sigma_1(x_1, x_2))^2 + (\sigma_2(x_1, x_2))^2 \\ &= \theta_1(x_1 \vee 0) + (\theta_2 x_1 + \theta_3 x_2) \vee 0 \\ &\leq \theta_1 |x_1| + \theta_2 |x_1| + \theta_3 |x_2| \\ &\leq 8 \cdot \max\left\{\theta_1, \theta_2, \theta_3\right\} ((x_1^2 + x_2^2 + 1) \\ &= \tilde{C}(\|\boldsymbol{x}\|^2 + 1). \end{aligned}$$

Hence the prerequisites of Theorem 3.3 are fulfilled; and therefore, we know that weak solutions of SDE (5.4) with finite second moments exist.

Equation (5.4a) describing the evolution of $X_1(t)$ does not depend on the second component $X_2(t)$. Thus, we can consider it as a one-dimensional process which fulfills the prerequisites for Corollary 3.4 and obtain that the pathwise uniqueness of solutions holds. Due

to Remark 3.1, we know that Equation (5.4a) has a unique strong solution. Moreover, the solution takes non-negative values (cf. Ikeda & Watanabe, 1981, Example 8.2, p.221), so we can omit the max operator. From Lamberton & Lapeyre (1996, Proposition 6.2.4.), it follows that for the stopping time

$$\tau_{X_1,0} := \inf\{t > 0 : X_1(t) = 0\}$$

it holds that

$$\mathbb{P}\{\tau_{X_1,0} < \infty\} = 1.$$

This property that almost all trajectories reach zero in finite time goes well with the interpretation of the solution process as the approximation of a stochastic (discrete) decay process.

Next, we show that Equation (5.4b) describing the evolution of $X_2(t)$ has with probability 1 a pathwise unique solution. The idea of the proof is similar to the proof of Theorem 3.2 in Ikeda & Watanabe (1981, p.168). Then, again due to Remark 3.1, we obtain that this unique solution is a strong solution.

Theorem 5.1. *For the weak solutions of Equation (5.4b), pathwise uniqueness holds.*

Proof. As in Ikeda & Watanabe (1981, pp. 168), let $1 > a_1 > a_2 > \ldots > a_n > \ldots > 0$ be defined by

$$\int_{a_1}^{1} \frac{1}{u}\,\mathrm{d}u = 1, \quad \int_{a_2}^{a_1} \frac{1}{u}\,\mathrm{d}u = 2, \quad \ldots, \quad \int_{a_n}^{a_{n-1}} \frac{1}{u}\,\mathrm{d}u = n, \quad \ldots.$$

Clearly, $a_n \longrightarrow 0$ as $n \to \infty$. Let $\psi_n(u)$, $n = 1, 2, \ldots$, be a continuous function such that its support is contained in $]a_n, a_{n-1}[$,

$$0 \le \psi_n(u) \le \frac{2}{nu} \quad \text{and} \quad \int_{a_n}^{a_{n-1}} \psi_n(u)\,\mathrm{d}u = 1.$$

Set

$$\varphi_n(x) = \int_0^{|x|} \int_0^y \psi_n(u)\,\mathrm{d}u\,\mathrm{d}y, \qquad x \in \mathbb{R}^1.$$

It is easy to see that $\varphi_n \in C^2(\mathbb{R}^1)$, $|\varphi_n'(x)| \le 1$, $\varphi_n(x) \uparrow |x|$ as $n \to \infty$, and

$$\begin{aligned}\varphi_n(x) &= \int_{a_{n-1}}^{|x|} \underbrace{\int_0^y \psi_n(u)\,\mathrm{d}u}_{=1 \text{ for } y \ge a_{n-1}}\,\mathrm{d}y + \int_{a_n}^{a_{n-1}} \underbrace{\int_0^y \psi_n(u)\,\mathrm{d}u}_{\ge 0 \text{ for } a_n \le y \le a_{n-1}}\,\mathrm{d}y + \int_0^{a_n} \underbrace{\int_0^y \psi_n(u)\,\mathrm{d}u}_{=0 \text{ for } y \le a_n}\,\mathrm{d}y \\ &\ge \int_{a_{n-1}}^{|x|} 1\,\mathrm{d}y = |x| - a_{n-1}. \end{aligned} \tag{5.5}$$

Suppose Y_1, Y_2 are two weak solutions of Equation (5.4b) with $Y_1(0) = Y_2(0) = 0$, the same Brownian motion $B_2(t)$, and X_1 is the strong solution of Equation (5.4a). We introduce the following stopping times

$$\begin{aligned}\tau_{X_1,r} &:= \inf\{t > 0 : X_1(t) > r\}\\ \tau_{Y_1,r} &:= \inf\{t > 0 : |Y_1(t)| > r\}\\ \tau_{Y_2,r} &:= \inf\{t > 0 : |Y_2(t)| > r\},\end{aligned}$$

for $r = 1, 2, 3, \ldots$. Let r be arbitrary but fixed and define $\tau_r := \tau_{X_1,r} \wedge \tau_{Y_1,r} \wedge \tau_{Y_2,r}$.

Then we have

$$\begin{aligned}&Y_1(t \wedge \tau_r) - Y_2(t \wedge \tau_r) =\\ &\quad\int_0^{t\wedge\tau_r} [(\theta_2 X_1(s) - \theta_3 Y_1(s)) - (\theta_2 X_1(s) - \theta_3 Y_2(s))]\,\mathrm{d}s\\ &\quad+ \int_0^{t\wedge\tau_r} \left[\sqrt{(\theta_2 X_1(s) + \theta_3 Y_1(s)) \vee 0} - \sqrt{(\theta_2 X_1(s) + \theta_3 Y_2(s)) \vee 0}\right] \mathrm{d}B_2(s)\end{aligned}$$

and by Itô's formula

$$\begin{aligned}&\varphi_n(Y_1(t \wedge \tau_r) - Y_2(t \wedge \tau_r)) =\\ &\quad\int_0^{t\wedge\tau_r} \varphi_n'(Y_1(s) - Y_2(s))(\theta_3(Y_2(s) - Y_1(s))\,\mathrm{d}s\\ &\quad+ \frac{1}{2}\int_0^{t\wedge\tau_r} \varphi_n''(Y_1(s) - Y_2(s))\cdot\\ &\qquad\qquad\left[\sqrt{(\theta_2 X_1(s) + \theta_3 Y_1(s)) \vee 0} - \sqrt{(\theta_2 X_1(s) + \theta_3 Y_2(s)) \vee 0}\right]^2 \mathrm{d}s\\ &\quad+ \int_0^{t\wedge\tau_r} \varphi_n'(Y_1(s) - Y_2(s))\cdot\\ &\qquad\qquad\left[\sqrt{(\theta_2 X_1(s) + \theta_3 Y_1(s)) \vee 0} - \sqrt{(\theta_2 X_1(s) + \theta_3 Y_2(s)) \vee 0}\right] \mathrm{d}B_2(s).\end{aligned}$$

Since the expectation of the last term on the right-hand side is zero, we have

$$\begin{aligned}&\mathbb{E}\left[\varphi_n(Y_1(t \wedge \tau_r) - Y_2(t \wedge \tau_r))\right] =\\ &\quad\mathbb{E}\left[\int_0^{t\wedge\tau_r} \varphi_n'(Y_1(s) - Y_2(s))(\theta_3(Y_2(s) - Y_1(s))\,\mathrm{d}s\right]\\ &\quad+ \frac{1}{2}\,\mathbb{E}\left[\int_0^{t\wedge\tau_r} \varphi_n''(Y_1(s) - Y_2(s))\cdot\right.\\ &\qquad\qquad\left.\left[\sqrt{(\theta_2 X_1(s) + \theta_3 Y_1(s)) \vee 0} - \sqrt{(\theta_2 X_1(s) + \theta_3 Y_2(s)) \vee 0}\right]^2 \mathrm{d}s\right]\\ &\quad= I_{1,r} + I_{2,r}.\end{aligned}$$

Now, we estimate both summands by

$$
\begin{aligned}
|I_{1,r}| &= \left| \mathbb{E}\left[\int_0^t \varphi_n'(Y_1(s\wedge\tau_r) - Y_2(s\wedge\tau_r))(\theta_3(Y_2(s\wedge\tau_r) - Y_1(s\wedge\tau_r)))\,\mathrm{d}s \right] \right| \\
&\leq \mathbb{E}\left[\int_0^t \left|\varphi_n'(Y_1(s\wedge\tau_r) - Y_2(s\wedge\tau_r))\right| |\theta_3(Y_2(s\wedge\tau_r) - Y_1(s\wedge\tau_r))|\,\mathrm{d}s \right] \\
&\leq \mathbb{E}\left[\int_0^t \theta_3 |(Y_2(s\wedge\tau_r) - Y_1(s\wedge\tau_r)|\,\mathrm{d}s \right] \\
&= \theta_3 \int_0^t \mathbb{E}\left[|(Y_2(s\wedge\tau_r) - Y_1(s\wedge\tau_r)|\right]\,\mathrm{d}s
\end{aligned}
$$

and

$$
\begin{aligned}
|I_{2,r}| &= \frac{1}{2}\left| \mathbb{E}\left[\int_0^t \varphi_n''(Y_1(s\wedge\tau_r) - Y_2(s\wedge\tau_r)) \cdot \right.\right. \\
&\qquad \left.\left. \left[\sqrt{(\theta_2 X_1(s\wedge\tau_r) + \theta_3 Y_1(s\wedge\tau_r)) \vee 0} - \sqrt{(\theta_2 X_1(s\wedge\tau_r) + \theta_3 Y_2(s\wedge\tau_r)) \vee 0} \right]^2 \mathrm{d}s \right] \right| \\
&\leq \frac{1}{2}\mathbb{E}\left[\int_0^t \left|\varphi_n''(Y_1(s\wedge\tau_r) - Y_2(s\wedge\tau_r))\right| \cdot \right. \\
&\qquad \left. \left[\sqrt{(\theta_2 X_1(s\wedge\tau_r) + \theta_3 Y_1(s\wedge\tau_r)) \vee 0} - \sqrt{(\theta_2 X_1(s\wedge\tau_r) + \theta_3 Y_2(s\wedge\tau_r)) \vee 0} \right]^2 \mathrm{d}s \right] \\
&\leq \frac{1}{2}\mathbb{E}\left[\int_0^t \frac{2}{n} \frac{1}{|Y_1(s\wedge\tau_r) - Y_2(s\wedge\tau_r))|} \cdot \right. \\
&\qquad \left. \left[\sqrt{(\theta_2 X_1(s\wedge\tau_r) + \theta_3 Y_1(s\wedge\tau_r)) \vee 0} - \sqrt{(\theta_2 X_1(s\wedge\tau_r) + \theta_3 Y_2(s\wedge\tau_r)) \vee 0} \right]^2 \mathrm{d}s \right].
\end{aligned}
$$

It follows from the Hölder continuity of the square root function that

$$
\begin{aligned}
|I_{2,r}| &\leq \frac{1}{2}\mathbb{E}\left[\int_0^t \frac{2}{n} \frac{1}{|Y_1(s\wedge\tau_r) - Y_2(s\wedge\tau_r))|} \cdot \right. \\
&\qquad \left. \left[\sqrt{((\theta_2 X_1(s\wedge\tau_r) + \theta_3 Y_1(s\wedge\tau_r)) \vee 0) - ((\theta_2 X_1(s\wedge\tau_r) + \theta_3 Y_2(s\wedge\tau_r)) \vee 0)} \right]^2 \mathrm{d}s \right] \\
&\leq \frac{1}{2}\mathbb{E}\left[\int_0^t \frac{2}{n} \frac{1}{|Y_1(s\wedge\tau_r) - Y_2(s\wedge\tau_r))|} \cdot \right. \\
&\qquad \left. \left[\sqrt{(\theta_2 X_1(s\wedge\tau_r) + \theta_3 Y_1(s\wedge\tau_r)) - (\theta_2 X_1(s\wedge\tau_r) + \theta_3 Y_2(s\wedge\tau_r))} \right]^2 \mathrm{d}s \right] \\
&= \frac{1}{2}\mathbb{E}\left[\int_0^t \frac{2}{n} \frac{1}{|Y_1(s\wedge\tau_r) - Y_2(s\wedge\tau_r))|} \left[\sqrt{\theta_3(Y_1(s\wedge\tau_r) - Y_2(s\wedge\tau_r))} \right]^2 \mathrm{d}s \right] \\
&= \frac{t\theta_3}{n} \longrightarrow 0 \text{ as } n \to \infty \text{ and for every } r > 0.
\end{aligned}
$$

Consequently, by letting $n \to \infty$, we have $\varphi_n(x) \to |x|$ and

$$\mathbb{E}\left[|Y_1(t \wedge \tau_r) - Y_2(t \wedge \tau_r)|\right] \leq \theta_3 \int_0^t \mathbb{E}\left[|(Y_1(s \wedge \tau_r) - Y_2(s \wedge \tau_r)|\right] \mathrm{d}s.$$

By Gronwall's Lemma (see Theorem A.1 in the appendix), we have

$$\mathbb{E}\left[|Y_1(t \wedge \tau_r) - Y_2(t \wedge \tau_r)|\right] \leq 0 \text{ for all } t, r > 0$$

and thus

$$\mathbb{E}\left[|Y_1(t \wedge \tau_r) - Y_2(t \wedge \tau_r)|\right] = 0 \text{ for all } t, r > 0.$$

The last relation implies

$$\mathbb{P}\left\{Y_1(t \wedge \tau_r) = Y_2(t \wedge \tau_r) \text{ for all } t, r > 0\right\} = 1$$

which ensure the pathwise uniqueness of the solution. □

Next, we show that the term $\theta_2 X_1(t) + \theta_3 X_2(t)$ in the radicand of σ_2 is non-negative. We introduce the stopping time

$$\tau_{-\varepsilon} := \inf\left\{t > 0 : \theta_2 X_1(t) + \theta_3 X_2(t) = -\varepsilon\right\}.$$

Suppose that $\mathbb{P} = \{\tau_{-\varepsilon} < \infty\} > 0$. Then, since the trajectories of $\boldsymbol{X}(t)$ are continuous and $\theta_2 X_1(0) + \theta_3 X_2(0) = \theta_2 m_0 > 0$, there exists an $r < \tau_{-\varepsilon}$ such that

$$\theta_2 X_1(t) + \theta_3 X_2(t) \geq 0 \text{ for } t \in]0, r]$$

and

$$\theta_2 X_1(t) + \theta_3 X_2(t) < 0 \text{ for } t \in]r, \tau_{-\varepsilon}[.$$

In this case, for the second component, $X_2(t) < -\frac{\theta_2}{\theta_3} X_1(t)$ for $t \in]r, \tau_\varepsilon[$ must hold because $X_1(t)$ is non-negative. Moreover, the diffusion coefficient of the second component vanishes on this time interval and we have

$$\mathrm{d}X_2(t) = \theta_2 X_1(t) - \theta_3 X_2(t)\,\mathrm{d}t \text{ for } t \in]r, \tau_{-\varepsilon}[.$$

As $X_2(t)$ is negative for $t \in]r, \tau_{-\varepsilon}[$, the right hand side is positive for $t \in]r, \tau_{-\varepsilon}[$. Therefore, $X_2(t)$ is increasing for $t \in]r, \tau_{-\varepsilon}[$; and thus, the radicand cannot further decrease and cannot reach $-\varepsilon$. This insight contradicts the assumption that $\tau_{-\varepsilon}$ is finite. From this, it follows that

$\theta_2 X_1(t) + \theta_3 X_2(t)$ stays non-negative and consequently, we can omit taking the maximum of zero and $\theta_2 X_1(t) + \theta_3 X_2(t)$.

We have shown that the modified 2-dimensional SDE (5.4) has a unique strong solution. Moreover, we know that the radicands in the diffusion terms for both components of the solution process stay non-negative; and consequently, we can omit the max operator in the respective diffusion coefficient. In conclusion, we obtain that also for the original SDE (5.3) there exists a unique strong solution.

5.3.2 Convergence of the Euler-Maruyama scheme

In this section, we show that the Euler approximation of the solution $\boldsymbol{X}(t)$ of Equation (5.3) strongly converges to $\boldsymbol{X}(t)$. We consider the representation of the Euler approximation as introduced in Equations (3.13) and (3.14) in Section 3.3. For $n \geq 1$, let $\kappa_n : [0,T] \to [0,T]$ be defined by $\kappa_n(T) := \frac{n-1}{n}T$ and

$$\kappa_n(t) = \frac{iT}{n} \quad \text{for } \frac{iT}{n} \leq t \leq \frac{(i+1)T}{n}, \quad \text{for } i = 0, \ldots, n-1. \tag{5.6}$$

We define the Euler approximation $\boldsymbol{X}^n(t) = (X_1^n(t), X_2^n(t))^{\mathsf{Tr}}$ of the solution $\boldsymbol{X}(t)$ of Equation (5.3) by

$$\mathrm{d}X_1^n(t) = -\theta_1 X_1^n(\kappa_n(t))\,\mathrm{d}t + \sqrt{\theta_1 X_1^n(\kappa_n(t))}\,\mathrm{d}B_1(t), \tag{5.7a}$$

$$\mathrm{d}X_2^n(t) = \left(\theta_2 X_1^n(\kappa_n(t)) - \theta_3 X_2^n(\kappa_n(t))\right)\mathrm{d}t + \sqrt{\theta_2 X_1^n(\kappa_n(t)) + \theta_3 X_2^n(\kappa_n(t))}\,\mathrm{d}B_2(t), \tag{5.7b}$$

$$X_1^n(t_0) = m_0, \tag{5.7c}$$

$$X_2^n(t_0) = 0, \tag{5.7d}$$

for $t \in [t_0, T]$. As in the previous subsection, we assume without loss of generality that $t_0 = 0$.

Theorem 5.2 (Convergence of the Euler scheme)**.** *For the solution $\boldsymbol{X}(t)$ of Equation (5.3) and its Euler approximation $\boldsymbol{X}^n(t)$, the solution of Equation (5.7), it holds that*

$$\lim_{n\to\infty} \sup_{0\leq t\leq T} \mathbb{E}\,\|\boldsymbol{X}(t) - \boldsymbol{X}^n(t)\| = 0.$$

The idea of the proof is again similar to that in Ikeda & Watanabe (1981, pp. 168) and based on the functions φ_k as introduced in the proof of Theorem 5.1. This idea has also been used

for proofs of convergence for different types of SDEs e. g. in Gyöngy & Rásonyi (2011), Ngo & Raguchi (2016), and Yang et al. (2019).

Proof. We can again consider the proof for the first component $X_1(t)$ independently as a one-dimensional equation as it does not depend on $X_2(t)$. Therefore, we first show that

$$\lim_{n\to\infty} \sup_{0\le t\le T} \mathbb{E}|X_1(t) - X_1^n(t)| = 0.$$

With Corollary 3.7, we have

$$\mathbb{E}\left[\sup_{0\le t\le T} |X_1(t) - X_1^n(t)|\right] \le \frac{C_1}{\sqrt{\ln n}}, \tag{5.8}$$

where C_1 is a constant depending on K, T, and $(m_0)^2$, but independent of n.

In particular, we obtain

$$\sup_{0\le t\le T} \mathbb{E}|X_1(t) - X_1^n(t)| \le \mathbb{E}\left[\sup_{0\le t\le T} |X_1(t) - X_1^n(t)|\right] \le \frac{C_1}{\sqrt{\ln n}} \longrightarrow 0 \quad \text{for } n\to\infty.$$

Next, we consider the Euler scheme (5.7b) for the second component $X_2(t)$ of Equation (5.3). We have

$$\begin{aligned} X_2(t) - X_2^n(t) = \theta_2 \int_0^t X_1(s) - X_1^n(\kappa_n(s))\,\mathrm{d}s + \theta_3 \int_0^t -X_2(s) + X_2^n(\kappa_n(s))\,\mathrm{d}s \\ + \int_0^t \sqrt{\theta_2 X_1(s) + \theta_3 X_2(s)} - \sqrt{\theta_2 X_1^n(\kappa_n(s)) + \theta_3 X_2^n(\kappa_n(s))}\,\mathrm{d}B_2(s). \end{aligned}$$

We again use the functions φ_k, $k = 1, 2, \dots$, as constructed in the proof of Theorem 5.1 and apply the Itô formula to obtain

$$\begin{aligned} \mathbb{E}\left[\varphi_k\left(X_2(t) - X_2^n(t)\right)\right] &= \theta_2 \mathbb{E}\left[\int_0^t \varphi_k'\left(X_2(s) - X_2^n(s)\right)\left(X_1(s) - X_1^n(\kappa_n(s))\right)\,\mathrm{d}s\right] \\ &\quad + \theta_3 \mathbb{E}\left[\int_0^t \varphi_k'\left(X_2(s) - X_2^n(s)\right)\left(X_2(s) - X_2^n(\kappa_n(s))\right)\,\mathrm{d}s\right] \\ &\quad + \frac{1}{2}\mathbb{E}\left[\int_0^t \varphi_k''\left(X_2(s) - X_2^n(s)\right)\left(\sqrt{\theta_2 X_1(s) + \theta_3 X_2(s)}\right.\right. \\ &\qquad \left.\left. - \sqrt{\theta_2 X_1^n(\kappa_n(s)) + \theta_3 X_2^n(\kappa_n(s))}\right)^2 \mathrm{d}s\right] \\ &=: R_3(t) + R_4(t) + R_5(t). \end{aligned}$$

Due to the properties of φ_k (in particular $|\varphi_k'(x)| \leq 1$), we have

$$\begin{aligned}
R_3(t) &\leq \theta_2 \mathbb{E}\left[\int_0^t |X_1(s) - X_1^n(\kappa_n(s))| \, \mathrm{d}s\right] \\
&= \theta_2 \mathbb{E}\left[\int_0^t |X_1(s) - X_1^n(s) + X_1^n(s) - X_1^n(\kappa_n(s))| \, \mathrm{d}s\right] \\
&\leq \theta_2 \mathbb{E}\left[\int_0^t |X_1(s) - X_1^n(s)| \, \mathrm{d}s\right] + \theta_2 \mathbb{E}\left[\int_0^t |X_1^n(s) - X_1^n(\kappa_n(s))| \, \mathrm{d}s\right] \\
&\leq \theta_2 \mathbb{E}\left[\int_0^T \sup_{0 \leq s \leq T} |X_1(s) - X_1^n(s)| \, \mathrm{d}s\right] + \theta_2 \mathbb{E}\left[\int_0^T \sup_{0 \leq s \leq T} |X_1^n(s) - X_1^n(\kappa_n(s))| \, \mathrm{d}s\right].
\end{aligned}$$

We apply Relation (5.8) to the first term and Relation (3.16) to the second term and obtain

$$R_3(t) \leq \theta_2 T \frac{C_1}{\sqrt{\ln n}} + \theta_2 T \frac{C_2}{\sqrt{n}}. \tag{5.9}$$

Similarly, we obtain

$$\begin{aligned}
R_4(t) &\leq \theta_3 \mathbb{E}\left[\int_0^t |X_2(s) - X_2^n(s)| \, \mathrm{d}s\right] + \theta_3 \mathbb{E}\left[\int_0^t |X_2^n(s) - X_2^n(\kappa_n(s))| \, \mathrm{d}s\right] \\
&\overset{(3.16)}{\leq} \theta_3 \mathbb{E}\left[\int_0^t |X_2(s) - X_2^n(s)| \, \mathrm{d}s\right] + \theta_3 T \frac{C_3}{\sqrt{n}}.
\end{aligned} \tag{5.10}$$

Due to the Hölder continuity of the square root function, we have

$$\begin{aligned}
R_5(t) &\leq \frac{1}{2} \mathbb{E}\left[\int_0^t \varphi_k''\left(X_2(s) - X_2^n(s)\right) \right. \\
&\qquad \left. \left(\sqrt{\theta_2(X_1(s) - X_1^n(\kappa_n(s))) + \theta_3(X_2(s) - X_2^n(\kappa_n(s)))}\right)^2 \mathrm{d}s\right] \\
&\leq \frac{1}{2} \mathbb{E}\left[\int_0^t \varphi_k''\left(X_2(s) - X_2^n(s)\right) \theta_2 |X_1(s) - X_1^n(\kappa_n(s))| \, \mathrm{d}s\right] \\
&\quad + \frac{1}{2} \mathbb{E}\left[\int_0^t \varphi_k''\left(X_2(s) - X_2^n(s)\right) \theta_3 |X_2(s) - X_2^n(\kappa_n(s))| \, \mathrm{d}s\right] \\
&=: R_6(t) + R_7(t).
\end{aligned}$$

With the properties of φ_k (in particular $\varphi_k''(x) = \psi_k(x)$, $0 \leq \psi_k(x) \leq \frac{2}{kx}$ and the support of ψ_k is contained in the interval $]a_k, a_{k-1}[$), it follows that

$$\begin{aligned}
R_6(t) &\leq \mathbb{E}\left[\int_0^t \frac{1}{k a_k} \theta_2 (X_1(s) - X_1^n(\kappa_n(s))) \, \mathrm{d}s\right] \\
&\leq \frac{1}{k a_k} \theta_2 \left(\mathbb{E}\left[\int_0^t |X_1(s) - X_1^n(s)| \, \mathrm{d}s\right] + \mathbb{E}\left[\int_0^t |X_1^n(s) - X_1^n(\kappa_n(s))| \, \mathrm{d}s\right]\right).
\end{aligned}$$

We again apply Relation (5.8) to the first summand and Relation (3.16) to the second summand and obtain

$$R_6(t) \leq \frac{1}{ka_k}\theta_2 T\left(\frac{C_1}{\sqrt{\ln n}} + \frac{C_2}{\sqrt{n}}\right). \tag{5.11}$$

Moreover, we have

$$\begin{aligned} R_7(t) \leq \frac{\theta_3}{2}\mathbb{E}\left[\int_0^t \varphi_k''\left(X_2(s) - X_2^n(s)\right)\left|X_2(s) - X_2^n(s)\right| \mathrm{d}s\right] \\ + \frac{\theta_3}{2}\mathbb{E}\left[\int_0^t \varphi_k''\left(X_2(s) - X_2^n(s)\right)\left|X_2^n(s) - X_2^n(\kappa_n(s))\right| \mathrm{d}s\right] \end{aligned}$$

and again use the properties of φ_k to obtain

$$\begin{aligned} R_7(t) &\leq \frac{\theta_3 T}{k} + \frac{\theta_3}{ka_k}\mathbb{E}\left[\int_0^t \left|X_2^n(s) - X_2^n(\kappa_n(s))\right| \mathrm{d}s\right] \\ &\leq \frac{\theta_3 T}{k} + \frac{\theta_3 T}{ka_k}\mathbb{E}\left[\sup_{0\leq t\leq T}\left|X_2^n(t) - X_2^n(\kappa_n(t))\right|\right] \\ &\overset{(3.16)}{\leq} \frac{\theta_3 T}{k} + \frac{\theta_3 T}{ka_k}\cdot\frac{C_3}{\sqrt{n}}. \end{aligned} \tag{5.12}$$

By construction (see (5.5)), it holds that $\varphi_k(x) \geq |x| - a_{k-1}$. With this and the Relations (5.9), (5.10), (5.11), and (5.12), we obtain

$$\begin{aligned} \mathbb{E}|X_2(t) - X_2^n(t)| &\leq a_{k-1} + \mathbb{E}\left[\varphi_k\left(X_2(t) - X_2^n(t)\right)\right] \\ &\leq a_{k-1} + \theta_2 T\frac{C_1}{\sqrt{\ln n}} + \theta_2 T\frac{C_2}{\sqrt{n}} + \theta_3\mathbb{E}\left[\int_0^t \left|X_2(s) - X_2^n(s)\right| \mathrm{d}s\right] \\ &\quad + \theta_3 T\frac{C_3}{\sqrt{n}} + \frac{1}{ka_k}\theta_2 T\left(\frac{C_1}{\sqrt{\ln n}} + \frac{C_2}{\sqrt{n}}\right) + \frac{\theta_3 T}{k} + \frac{\theta_3 T}{ka_k}\cdot\frac{C_3}{\sqrt{n}}. \end{aligned}$$

By Gronwall's Lemma (see Theorem A.1 in the appendix), it follows that

$$\begin{aligned} \mathbb{E}|X_2(t) - X_2^n(t)| &\leq \left(a_{k-1} + \theta_2 T\frac{C_1}{\sqrt{\ln n}} + \theta_2 T\frac{C_2}{\sqrt{n}} + \theta_3 T\frac{C_3}{\sqrt{n}}\right. \\ &\quad \left. + \frac{1}{ka_k}\theta_2 T\left(\frac{C_1}{\sqrt{\ln n}} + \frac{C_2}{\sqrt{n}}\right) + \frac{\theta_3 T}{k} + \frac{\theta_3 T}{ka_k}\cdot\frac{C_3}{\sqrt{n}}\right)\cdot e^{\theta_3 T} \\ &\leq \left(a_{k-1} + \frac{\theta_3 T}{k} + C_{max}T\left(\frac{1}{ka_k} + 1\right)\left(\frac{\theta_2}{\sqrt{\ln n}} + \frac{\theta_2}{\sqrt{n}} + \frac{\theta_3}{\sqrt{n}}\right)\right)\cdot e^{\theta_3 T}, \end{aligned} \tag{5.13}$$

where $C_{max} := \max\{C_1, C_2, C_3\}$.

Let $\varepsilon > 0$. As $a_k \to 0$ and $\frac{\theta_3 T}{k} \to 0$ for $k \to \infty$, there exists a $k_0(\varepsilon)$ such that for $k \geq k_0$, it holds that

$$\left(\frac{\theta_3 T}{k} + a_{k-1}\right) e^{\theta_3 T} \leq \frac{\varepsilon}{2}.$$

We choose $k = k_0$. As $\frac{1}{\sqrt{\ln n}} \to 0$ and $\frac{1}{\sqrt{n}} \to 0$ for $n \to 0$, there exists an $n_0(\varepsilon)$ such that for $n \geq n_0(\varepsilon)$, it holds that

$$C_{max} T \left(\frac{1}{k_0 a_{k_0}} + 1\right) \left(\frac{\theta_2}{\sqrt{\ln n}} + \frac{\theta_2}{\sqrt{n}} + \frac{\theta_3}{\sqrt{n}}\right) \cdot e^{\theta_3 T} \leq \frac{\varepsilon}{2}.$$

With Relation (5.13), it follows that for $n \geq n_0(\varepsilon)$

$$\mathbb{E}|X_2(t) - X_2^n(t)| \leq \varepsilon \quad \text{for all } t \in [0, T],$$

and thus,

$$\sup_{0 \leq t \leq T} \mathbb{E}|X_2(t) - X_2^n(t)| \leq \varepsilon.$$

This concludes the proof of Theorem 5.2. □

Remark 5.1. *From Gyöngy & Rásonyi (2011, Theorem 2.1), we know that the convergence rate for the Euler approximation $X_1^n(t)$ of the first component is $1/\ln n$, and from Relation (5.13), we see that the convergence rate for the Euler approximation $X_2^n(t)$ of the second component is $1/\sqrt{\ln n}$. Therefore, overall we obtain a convergence rate of $1/\sqrt{\ln n}$ for $\boldsymbol{X}^n(t)$.*

5.4 Model of the observations

In the experiment described in Section 5.1, neither the amount of mRNA molecules nor that of GFP molecules can be measured directly. Instead, a fluorescence signal is observed which is assumed to be a linear transformation of the amount of GFP molecules. Moreover, Fröhlich et al. (2018) state that "Analysis of processed data suggested a constant offset and multiplicative measurement noise in the recorded fluorescence trajectories." Therefore, denoting a trajectory of mean fluorescence intensity observed at time points t_k, for $k = 1, \ldots, K$, by $\{y_k\}_{k=1,\ldots,K}$, we assume that

$$\log(y_k) = \log\left(\text{scale} \cdot X_2(t_k) + \text{offset}\right) + \epsilon_k, \quad \epsilon_k \sim \mathcal{N}(0, \sigma^2),$$

where the random variables ϵ_k are independent.

Note that the observations depend only on the amount X_2 of GFP molecules, but not directly on the amount X_1 of mRNA molecules.

Based on the observations $\{y_k\}_{k=1,\dots,K}$, we want to infer the following unknown parameters:

- the three kinetic parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$ that denote the rate constants for mRNA degradation, translation, and GFP degradation,
- the initial amount m_0 of mRNA molecules and the time point t_0 at which it is released,
- the scaling factor scale and the offset for the fluorescence signal,
- and the standard deviation σ of the measurement errors.

5.5 Structural identifiability analysis

Our main interest lies in the question which of the model parameters for our two model types (ODE and SDE) can be inferred from the experimental data as described in Sections 5.1 and 5.4. Here, we first focus on the parameters $\boldsymbol{\theta}$, m_0, scale, and offset that drive the dynamics of the process and the fluorescence signal. We analyze the structural identifiability which only considers the model equations of the process dynamics and the observation equation (not the actual data) and assumes that we are in a perfect data situation, i. e. we have an infinite amount of data observed without measurement error. Plainly speaking, structural identifiability analysis answers the question whether different parameter combinations can lead to the same model output. While for ODE models, there are analytical methods to assess structural identifiability, no such methods exit for SDE models. Therefore, we use several different approaches to heuristically assess structural identifiability for our SDE model. In the following subsections, we consider a transformed version of both model types, we make use of the open source software DAISY as has recently been suggested by Browning et al. (2020), and finally we also study simulations of both model types.

5.5.1 Transformed models

We can reformulate the differential equations for both model types by setting

$$\boldsymbol{Z}(t) = \begin{pmatrix} Z_1(t) \\ Z_2(t) \end{pmatrix} := \begin{pmatrix} \frac{X_1(t)}{m_0} \\ \mathrm{scale} \cdot X_2(t) + \mathrm{offset} \end{pmatrix},$$

which means that

$$\boldsymbol{Z}(t) \equiv \begin{pmatrix} 0 \\ \mathrm{offset} \end{pmatrix} \text{ for } t < t_0, \quad \text{and } \boldsymbol{Z}(t_0) = \begin{pmatrix} 1 \\ \mathrm{offset} \end{pmatrix}.$$

Hence, the second component of the transformed process models the fluorescence signal which we assume to be observed.

Transformed ODE model

For the ODE model in Equation (5.1), we obtain the transformed model

$$\frac{\mathrm{d}\boldsymbol{Z}(t)}{\mathrm{d}t} = \begin{pmatrix} -\theta_1 Z_1(t) \\ \mathrm{scale}\,\theta_2 m_0 Z_1(t) - \theta_3(Z_2(t) - \mathrm{offset}) \end{pmatrix} \text{ for } t \geq t_0, \tag{5.14}$$

and the corresponding solution

$$Z_1(t) = \exp\left(-\theta_1(t - t_0)\right),$$

$$Z_2(t) = \begin{cases} \frac{\mathrm{scale}\,\theta_2 m_0}{\theta_3 - \theta_1}\left(e^{-\theta_1(t-t_0)} - e^{-\theta_3(t-t_0)}\right) + \mathrm{offset} & \text{, for } \theta_1 \neq \theta_3, \\ \mathrm{scale}\,\theta_2 m_0 (t - t_0) e^{-\theta_3(t-t_0)} + \mathrm{offset} & \text{, for } \theta_1 = \theta_3. \end{cases}$$

The parameters scale, m_0, and θ_2 appear only as a product. Thus, we can already deduce from this equation that at most the product of the three parameters will be identifiable but not the three parameters individually. Moreover, since only $Z_2(t)$ is observed and it is symmetric in the parameters θ_1 and θ_3 (i. e. switching their values will lead to the same model output), these two parameters can at most be locally identifiable.

Transformed SDE model

For the SDE model in Equation (5.3), we apply the Itô formula from Theorem 3.5 to obtain the transformed model

$$\begin{aligned} \mathrm{d}\boldsymbol{Z}(t) = &\begin{pmatrix} -\theta_1 Z_1(t) \\ \mathrm{scale}\,\theta_2 m_0 Z_1(t) - \theta_3(Z_2(t) - \mathrm{offset}) \end{pmatrix} \mathrm{d}t \\ &+ \begin{pmatrix} \sqrt{\frac{\theta_1}{m_0} Z_1(t)} & 0 \\ 0 & \sqrt{\mathrm{scale}}\sqrt{\mathrm{scale}\,\theta_2 m_0 Z_1(t) + \theta_3(Z_2(t) - \mathrm{offset})} \end{pmatrix} \mathrm{d}B_t \text{ for } t \geq t_0. \end{aligned} \tag{5.15}$$

Note that here, the parameters scale and m_0 also appear outside the product $\mathrm{scale}\,\theta_2 m_0$. Therefore, we hope to gain more information about the individual parameters from data for the SDE model than for the ODE model.

5.5.2 Using a surrogate model and existing software tools

The open source software DAISY (Differential Algebra for Identifiability of SYstems) was introduced by Bellu et al. (2007). It is a software tool that implements a differential algebra algorithm to perform structural identifiability analysis for systems of polynomial or rational ODEs and that also allows to include unknown initial conditions. Mathematically, the problem translates into checking the solvability of a very large system of nonlinear algebraic equations. However, the use of the DAISY software does not require an in-depth understanding of the underlying theory.

Here, we want to use DAISY to assess the structural identifiability of the parameters in the two models of the translation kinetics. In order to include the parameters scale and offset, we use the transformed models from the previous subsection for the identifiability analysis. For the ODE model in Equation (5.14), the analysis with DAISY is straight forward since it is intended for the use for ODE models. After applying DAISY, the obtained output shows that when considering the set of parameters $\{\boldsymbol{\theta}, m_0, \mathrm{scale}, \mathrm{offset}\}$, the model is non-identifiable. The DAISY output also reveals that this non-identifiability is due to the fact that the parameters θ_1 and θ_3 are only locally identifiable and the parameters θ_2, m_0, and scale are not individually identifiable, but only their product is. This confirms our assertions from the previous subsection. Moreover, we obtain that the remaining parameter offset is structurally identifiable.

For SDE models, Browning et al. (2020) suggest to formulate a surrogate model based on the moment equations of the diffusion process. The moment equations are a system of ODEs, and thus, DAISY can be applied to this system. For the SDE (5.15), let $m_{ij}(t) = \mathbb{E}\left[(Z_1(t))^i (Z_2(t))^j\right]$ be the (mixed) moment of the diffusion process of order i and j. The moments are obtained by applying the Itô formula in Theorem 3.5 to $(Z_1(t))^i (Z_2(t))^j$ and then taking the expectation. Considering the first and the second moments of the process states results in the following system of ODEs:

$$
\begin{aligned}
\frac{\mathrm{d}m_{10}(t)}{\mathrm{d}t} &= -\theta_1 m_{10}(t), && m_{10}(t_0) = 1,\\
\frac{\mathrm{d}m_{01}(t)}{\mathrm{d}t} &= \mathrm{scale}\,\theta_2 m_0\, m_{10}(t) - \theta_3\, m_{01}(t) + \theta_3 \mathrm{offset}, && m_{01}(t_0) = \mathrm{offset},\\
\frac{\mathrm{d}m_{20}(t)}{\mathrm{d}t} &= \frac{\theta_1}{m_0}\, m_{10}(t) - 2\theta_1\, m_{20}(t), && m_{20}(t_0) = 1,\\
\frac{\mathrm{d}m_{02}(t)}{\mathrm{d}t} &= \mathrm{scale}^2 \theta_2 m_0\, m_{10}(t) + \theta_3(\mathrm{scale} + 2\mathrm{offset})\, m_{01}(t) - 2\theta_3\, m_{02}(t)\\
&\quad + 2\mathrm{scale}\,\theta_2 m_0\, m_{11}(t) - \mathrm{scale}\,\theta_3 \mathrm{offset}, && m_{02}(t_0) = \mathrm{offset}^2,\\
\frac{\mathrm{d}m_{11}(t)}{\mathrm{d}t} &= \theta_3 \mathrm{offset}\, m_{10}(t) + \mathrm{scale}\,\theta_2 m_0\, m_{20}(t) - (\theta_1 + \theta_3)\, m_{11}(t), && m_{11}(t_0) = \mathrm{offset},
\end{aligned}
$$

where the equations for the two first moments m_{10} and m_{01} coincide with the ODE model in Equation (5.14). Since in the experiment, only the fluorescence signal is observed, we consider the moments that only depend on the second component of the process, i. e. m_{01} and m_{02}, as output states for the identifiability analysis. Using DAISY, we obtain that the surrogate model is globally identifiable, i. e. all six parameter values could be uniquely determined if we were able to observe the moments m_{01} and m_{02} directly, infinitely long over time, and without measurement error. However, this property of structural identifiability (in particular when using a surrogate model) is only a necessary, but not a sufficient condition for practical identifiability. From this result, we cannot conclude that the parameters will be identifiable from the actual experimental data.

5.5.3 Simulating from the models

Another attempt to assess parameter identifiability is to simulate from both model types for different parameter settings and compare whether we see differences in the simulation output. To obtain simulations from the ODE model, we use its solution in Equation (5.2). Since the ODE model is deterministic, each parameter setting yields one unique output trajectory while for the SDE model, we simulate several trajectories for each parameter setting using the Euler-Maruyama scheme with a time step of 0.01 hours.

Keeping the product $\mathrm{scale}\,\theta_2 m_0$ constant

As already pointed out in Subsection 5.5.1, the trajectories of the fluorescence intensity for the ODE model are identical if the product $\mathrm{scale}\,\theta_2 m_0$ and the remaining parameters are fixed, even when the individual factors scale, θ_2, and m_0 vary. Here, we use (approximately) the mean values for the parameters estimated from the data in Fröhlich et al. (2018), and therefore, set $\mathrm{scale}\,\theta_2 m_0 = 350$, $\theta_1 = 0.11$, $\theta_3 = 0.03$, $\mathrm{offset} = 8.9$, and $t_0 = 0$. For the SDE model, we simulate several trajectories with different values for scale, θ_2, and m_0 while keeping their product constant. For each parameter setting, we set the same random seed at the beginning of the simulation. Figure 5.2 displays the simulated trajectories.

It is evident that the SDE trajectories behave differently for different combinations of scale, θ_2, and m_0. For example, when we keep m_0 fixed while increasing scale and decreasing θ_2, the variation between but also within the trajectories increases. When we keep scale fixed while decreasing m_0 and increasing θ_2, especially the variation between trajectories seems to increase. And finally, when we keep θ_2 fixed while decreasing m_0 and increasing scale, the variation between and within the trajectories increases. Our focus is on estimating the

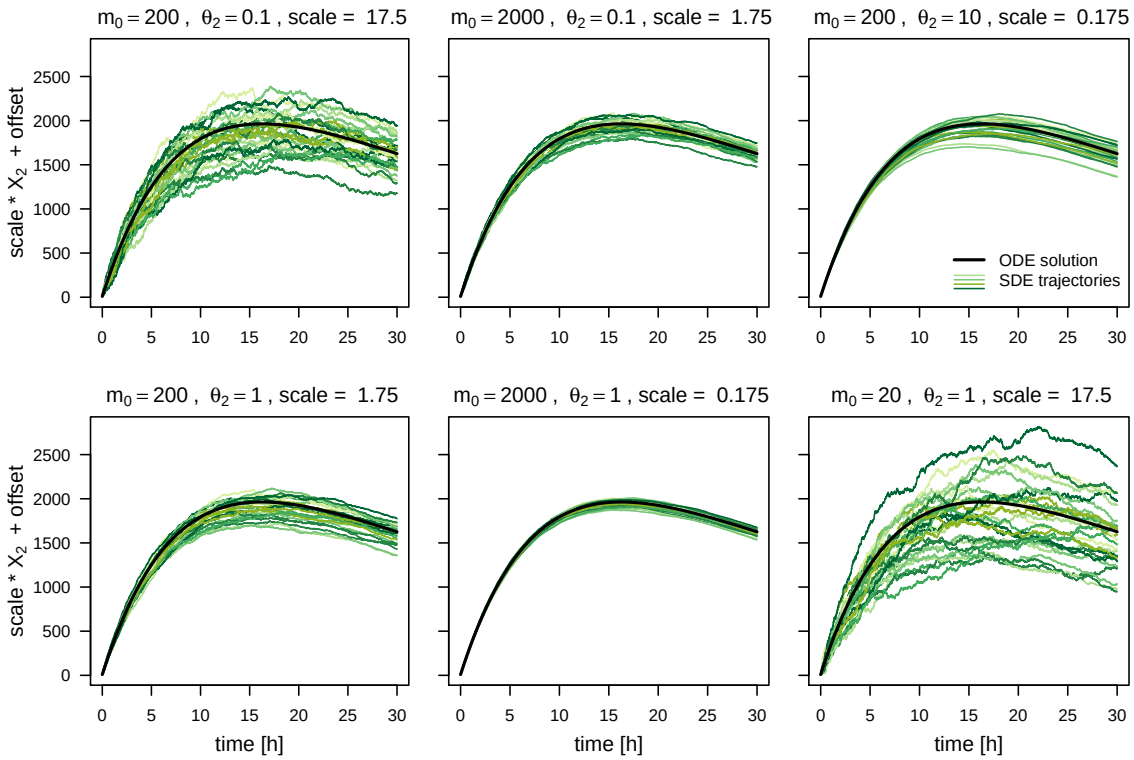


Figure 5.2: The ODE trajectory and 20 SDE trajectories of the fluorescence intensity simulated for different values of m_0, θ_2, and scale, while keeping their product constant at $\mathrm{scale}\,\theta_2 m_0 = 350$. The remaining parameters are set to $\theta_1 = 0.11$, $\theta_3 = 0.03$, $\mathrm{offset} = 8.9$, and $t_0 = 0$.

parameters from individual observed trajectories. In this context, especially the difference in the variation within the trajectories is relevant.

Swapping the degradation rate constants θ_1 and θ_3

The trajectories of the fluorescence intensity for the ODE model are also identical if the values for θ_1 and θ_3 are swapped while the remaining parameters are fixed. We simulate trajectories for the parameter combinations $(\theta_1, \theta_3) = (0.11, 0.03)$ and $(\theta_1^*, \theta_3^*) = (0.03, 0.11)$, respectively, while setting the remaining parameters to $\mathrm{scale} = 17.5$, $\theta_2 = 0.1$, $m_0 = 200$, $\mathrm{offset} = 8.9$, and $t_0 = 0$. For the SDE model, we again simulate several trajectories for both parameter settings and set the same random seed at the beginning of the simulation.

Figure 5.3 shows the ODE trajectory and several SDE trajectories in one panel for each of the two parameter combinations separately. Whereas, Figure 5.4 presents one SDE trajectory for each of the two parameter combinations together in one panel. Again, the SDE trajectories do behave differently for the different parameter combinations. While there seems to be only

little difference in the variation between the trajectories, the variation within the trajectories is clearly higher for lower θ_1 and higher θ_3. This indicates that it may be possible to uniquely determine the values of θ_1 and θ_3 even when estimating from only one observed trajectory.

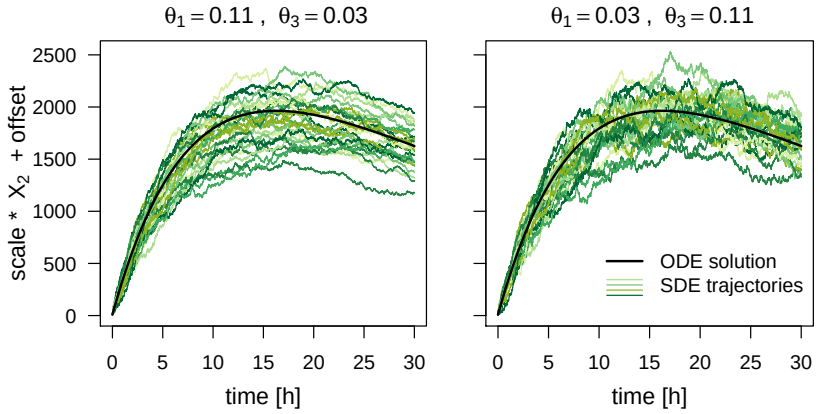


Figure 5.3: The ODE trajectory and 20 SDE trajectories of the fluorescence intensity simulated for two parameter combinations where the values of θ_1 and θ_3 are swapped. The remaining parameters are set to $\text{scale} = 17.5$, $\theta_2 = 0.1$, $m_0 = 200$, $\text{offset} = 8.9$, and $t_0 = 0$.

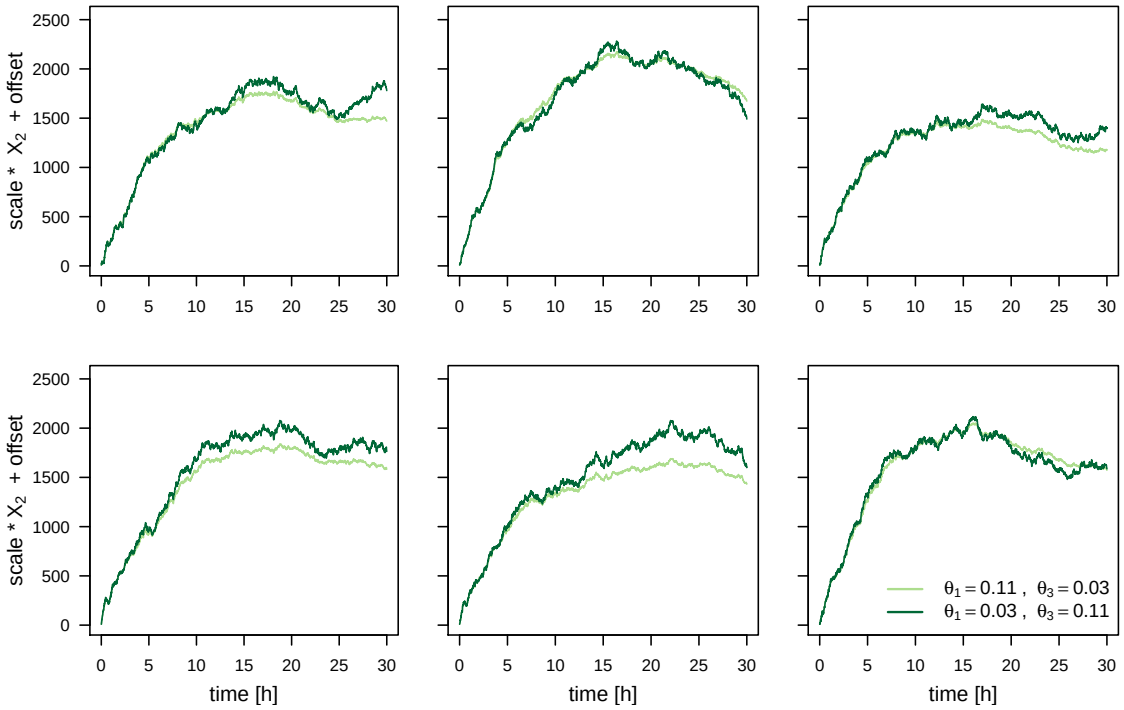


Figure 5.4: One trajectory of the fluorescence intensity for the SDE model simulated for each of the two parameter combinations where the values of θ_1 and θ_3 are swapped. The remaining parameters are set to $\text{scale} = 17.5$, $\theta_2 = 0.1$, $m_0 = 200$, $\text{offset} = 8.9$, and $t_0 = 0$.

5.6 Definition of the parameter posteriors

Next, we would like to assess the practical parameter identifiability by trying to estimate the parameters from observed data as described in Section 5.4. We take a Bayesian approach to parameter estimation as motivated in Section 2.2 because it allows for uncertainty assessment of the parameter estimates and also for handling unobserved process components and measurement error by using Markov chain Monte Carlo (MCMC) methods to sample from the parameter posterior distribution as explained in Section 3.4. Therefore, in this section, we define the parameter posterior densities for the two model types.

5.6.1 ODE model

For the ODE model, there is a deterministic relationship between the process values $\boldsymbol{X}(t)$ and the parameters $\boldsymbol{\theta}$, m_0 and t_0 (or between the fluorescence signal and the parameters including scale and offset, respectively).

Define the index $k^* := \min\{k \in \{1, \ldots, K\} | t_k \geq t_0\}$ of the first observation time point after the mRNA molecules are released, then the posterior density π from which we would like to sample is proportional to

$$\begin{aligned}
&\pi\left(\boldsymbol{\theta}, m_0, \mathrm{scale}, \mathrm{offset}, \sigma^2, t_0 | \{y_k\}_{k=1,\ldots,K}\right) \\
&\quad \propto \left(\prod_{k=k^*}^{K} \phi\left(\log(y_k) \middle| \log\left(\mathrm{scale} \frac{\theta_2 m_0}{\theta_3 - \theta_1} \left(e^{-\theta_1 (t - t_0)} - e^{-\theta_3 (t - t_0)} \right) + \mathrm{offset} \right), \sigma^2 \right) \right) \\
&\qquad \cdot \left(\prod_{k=1}^{k^*-1} \phi\left(\log(y_k) \middle| \log\left(\mathrm{offset} \right), \sigma^2 \right) \right) \\
&\qquad \cdot p(\theta_1) p(\theta_2) p(\theta_3) p(m_0) p(t_0) p(\mathrm{scale}) p(\mathrm{offset}) p(\sigma^2),
\end{aligned} \tag{5.16}$$

where $\phi(\cdot | \mu, \eta^2)$ denotes the density of the normal distribution with mean μ and variance η^2 and the $p(\cdot)$ denote the parameter prior densities.

If the priors $p(\theta_1)$ and $p(\theta_3)$ are symmetric to each other, then the posterior is also symmetric with respect to the two degradation rate constants.

The scaling factor scale, the translation rate constant θ_2, and the initial amount of mRNA m_0 appear only as a product in the likelihood function, therefore, as pointed out before, at most their product $\mathrm{scale}\theta_2 m_0$ is identifiable.

5.6.2 SDE model

For the SDE model, the states $\boldsymbol{X}(t_k)$, for $k = 1, \ldots, K$, of the process conditioned on the parameters $\boldsymbol{\theta}$, m_0 and t_0 are random numbers (for $t_k \geq t_0$). Hence, we have to marginalize over the process states to obtain the posterior density of the parameters which we want to infer:

$$\begin{aligned}
&\pi\left(\boldsymbol{\theta}, m_0, \text{scale}, \text{offset}, \sigma^2, t_0 | \{y_k\}_{k=1,\ldots,K}\right) \\
&\quad = \int_{\mathbb{R}_+^{2\times K}} \pi\left(\boldsymbol{\theta}, m_0, \text{scale}, \text{offset}, \sigma^2, t_0, \{\boldsymbol{X}(t_k)\}_{k=1,\ldots,K} | \{y_k\}_{k=1,\ldots,K}\right) \, \mathrm{d}\boldsymbol{X}(t_1) \ldots \mathrm{d}\boldsymbol{X}(t_K).
\end{aligned}$$

Therefore, again defining $k^* := \min\{k \in \{1, \ldots, K\} | t_k \geq t_0\}$, we would need to sample from

$$\begin{aligned}
&\pi\left(\boldsymbol{\theta}, m_0, \text{scale}, \text{offset}, \sigma^2, t_0, \{\boldsymbol{X}(t_k)\}_{k=1,\ldots,K} | \{y_k\}_{k=1,\ldots,K}\right) \\
&\quad \propto \left(\prod_{k=1}^{K} \phi\left(\log(y_k) \left| \log\left(\text{scale} \cdot X_2(t_k) + \text{offset}\right), \sigma^2\right.\right) \right) \\
&\qquad \cdot \left(\prod_{k=k^*}^{K-1} \pi\left(\boldsymbol{X}(t_{k+1}) | \boldsymbol{X}(t_k), \boldsymbol{\theta}\right) \right) \pi\left(\boldsymbol{X}(t_{k^*}) | \boldsymbol{\theta}, m_0, t_0\right) \left(\prod_{k=1}^{k^*-1} \delta(\|\boldsymbol{X}(t_k) - (0,0)^T\|) \right) \\
&\qquad \cdot p(\boldsymbol{\theta}) p(m_0) p(t_0) p(\text{scale}) p(\text{offset}) p(\sigma^2),
\end{aligned}$$

where $\phi(\cdot | \mu, \eta^2)$ denotes the density of the normal distribution with mean μ and variance η^2, $\delta(\cdot)$ denotes the Dirac delta function, $\|\cdot\|$ denotes a norm (e. g. the l_2-norm), and the factors $\pi\left(\boldsymbol{X}(t_{k+1}) | \boldsymbol{X}(t_k), \boldsymbol{\theta}\right)$, $k = k^*, \ldots, K-1$, denote the transition density of the process. However, the fact that the process $\boldsymbol{X}$ switches from a deterministic regime before t_0 to a stochastic one after t_0 complicates the estimation of t_0 together with the remaining parameters. Therefore, we will assume that t_0 is determined beforehand, e. g. based on the estimates for the ODE model. Consequently, we sample from

$$\begin{aligned}
&\pi\left(\boldsymbol{\theta}, m_0, \text{scale}, \text{offset}, \sigma^2, \{\boldsymbol{X}(t_k)\}_{k=1,\ldots,K} | \{y_k\}_{k=1,\ldots,K}, t_0\right) \\
&\quad \propto \left(\prod_{k=1}^{K} \phi\left(\log(y_k) \left| \log\left(\text{scale} \cdot X_2(t_k) + \text{offset}\right), \sigma^2\right.\right) \right) \\
&\qquad \cdot \left(\prod_{k=k^*}^{K-1} \pi\left(\boldsymbol{X}(t_{k+1}) | \boldsymbol{X}(t_k), \boldsymbol{\theta}\right) \right) \pi\left(\boldsymbol{X}(t_{k^*}) | \boldsymbol{\theta}, m_0, t_0\right) \left(\prod_{k=1}^{k^*-1} \delta(\|\boldsymbol{X}(t_k) - (0,0)^T\|) \right) \\
&\qquad \cdot p(\boldsymbol{\theta}) p(m_0) p(\text{scale}) p(\text{offset}) p(\sigma^2). \qquad (5.17)
\end{aligned}$$

While for the ODE model, the posterior distribution is only 8-dimensional and can be sampled from directly; for the SDE model, we need to sample from a $(7+2K)$-dimensional distribution and then marginalize over the $2K$ dimensions of the process states to obtain the posterior distribution of the parameters of interest. Moreover, there is no explicit exact expression for the transition density $\pi\left(\boldsymbol{X}(t_{k+1})|\boldsymbol{X}(t_k), \boldsymbol{\theta}\right)$; wherefore, it will be approximated by a normal density based on the Euler-Maruyama scheme. For this approximation to be appropriate, we have to ensure that the time steps between observations are small enough. The Milstein scheme, which we investigated in Chapter 4, cannot be used here since two components of the diffusion coefficient in Equation (5.3) depend on $X_1(t)$. Hence, Relation (4.3) does not hold. Consequently, the Milstein scheme to approximate the solution of Equation (5.3) contains stochastic double integrals for which no analytical solution is known; and therefore, the transition density of the process based on the Milstein scheme is also intractable.

5.7 Estimation based on simulated data

We want to use the open source software Stan that implements the Hamiltonian Monte Carlo (HMC)-based algorithm No-U-Turn Sampler (NUTS) as described in Section 2.2.2 to sample from the parameter posteriors as defined in the previous section. In order to assess how well we can recover the model parameters for both model types from individually observed trajectories, we first work with simulated data that is generated with Gillespie's algorithm. For the SDE model, we first need to check whether it is reasonable to assume that the time steps between the observations are sufficiently small for the Euler-Maruyama scheme to be appropriate.

5.7.1 Investigating the need for data augmentation

In this section, we focus on the SDE model and investigate whether data augmentation is necessary for the amount of data that we have available ($K = 181$ observations per cell with time step $\Delta t = 1/6$ hours). We simulate one trajectory of the MJP described in Section 5.2.1 with parameters $\theta_1 = 0.11$, $\theta_2 = 0.3$, $\theta_3 = 0.09$, and $m_0 = 200$ using Gillespie's algorithm. We assume for now that the amount X_2 of GFP is directly observed without error and that for the amount X_1 of mRNA, we only observe the initial value $m_0 = 200$. All observations are without measurement error and we assume $t_0 = 0$ to be known. Thus, we only estimate the kinetic parameters $\boldsymbol{\theta}$ for the SDE model, and to this end, use Stan and Bayesian data augmentation with different numbers of inter-observation intervals. A number of inter-observation intervals of 1 means that we do not impute any points between observations. A number of 2 inter-observation intervals means that we impute one point between every two observations and

so on. We generated 4 HMC chains with 1000 iterations after warm-up each. Figure 5.5 shows the median of the obtained posterior sample as the point estimates and the credible intervals (CIs) for the three kinetic parameters and for different numbers of inter-observation intervals. Evidently, the estimation results do not improve when increasing the number of inter-observation intervals. Therefore, we conclude that data augmentation is not necessary and do not make use of data augmentation in the remaining sections of this chapter.

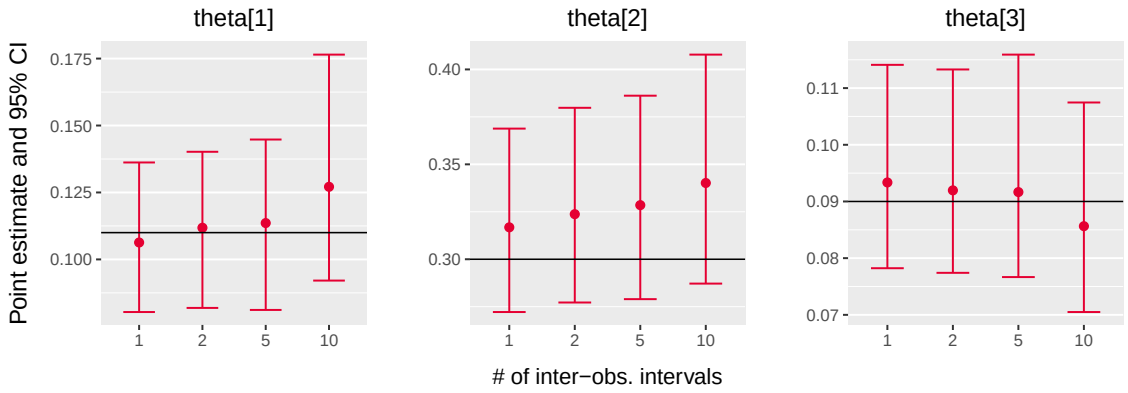


Figure 5.5: Point estimates (median of the posterior sample) and 95% CIs for the kinetic parameters estimated with Stan and Bayesian data augmentation for different numbers of inter-observations intervals. The black line represents the true parameter values with which the data was generated.

5.7.2 Simulated data without measurement error

For now, we assume the fluorescence intensity to be observed without measurement error. The data was simulated with Gillespie's algorithm with parameters $\boldsymbol{\theta} = (0.2, 0.32, 0.01)$, $m_0 = 240$, $t_0 = 0.96$, $\mathrm{scale} = 1.8$, and $\mathrm{offset} = 8.5$. The simulated fluorescence intensity (without measurement error) is depicted by the blue dotted line on the right hand side of Figure 5.6. We use Stan to sample from the posterior distributions of the ODE model and the SDE model given the simulated data. Since we assume the data to be observed without measurement error, the parameter offset can be determined directly from the first observation. Therefore, we do not include measurement error (and thus the parameter σ) and the parameter offset in the posterior distribution of the SDE model. Whereas for the ODE model, deviations of the observed data from the deterministic ODE trajectory have to be attributed to measurement error; therefore, the parameter σ has to be included in the posterior distribution of the ODE model. We also include the parameter offset for the ODE model in order to avoid degeneracy of the posterior. We use the following prior distributions: $\theta_i \sim \mathcal{N}_{\geq 0}(0, 5^2)$ for $i = 1, 2, 3$, $m_0 \sim \mathcal{N}_{\geq 0}(300, 300^2)$, $\mathrm{scale} \sim \mathcal{U}(0, 30)$, where $\mathcal{N}_{\geq a}(\mu, \eta^2)$ denotes the normal distribution truncated from below by a, and additionally for the ODE model, $\mathrm{offset} \sim \mathcal{U}(0, 30)$, $\sigma \sim \mathcal{U}(0.001, 10)$, and $t_0 \sim \mathcal{U}(0, 30)$.

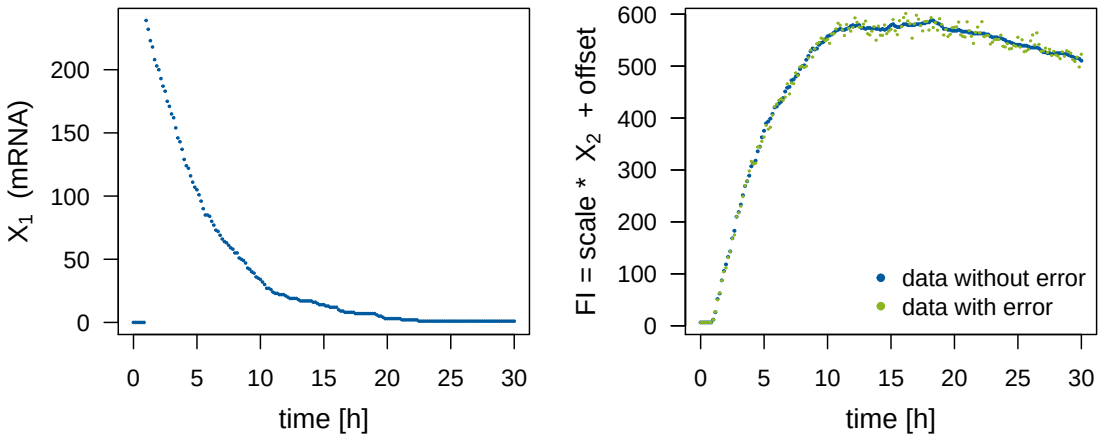


Figure 5.6: One trajectory used in the simulation study that was simulated with Gillespie's algorithm with parameters $\boldsymbol{\theta} = (0.2, 0.32, 0.01)$, $m_0 = 240$, $t_0 = 0.96$, $\text{scale} = 1.8$, and $\text{offset} = 8.5$, and for the green dotted line, multiplicative measurement error with $\sigma = 0.02$ was added to the fluorescence intensity (FI).

For both model types, we generate 8 HMC chains of 5000 iterations and discard the first half of the iterations as warm-up. Thus, we use a posterior sample of size 20,000 for each model type in the subsequent analysis. Tables 5.1 and 5.2 summarize the Stan output of the posterior samples for the ODE and the SDE model, respectively, and also include the true parameter values that were used to simulate the data for comparison. The tables also contain the 2.5%-, 50%-, and 97.5%-quantiles of the samples. We use the interval between the 2.5%- and the 97.5%-quantile as an estimate of the 95%-CI. For the ODE model, we see that the parameters offset and t_0 are well estimated since mean and median of the sample correspond to the true value, the CIs are very narrow, the effective sample size (ESS) n_{eff} is high and $\hat{R}$ is equal to 1. As expected, the measurement error parameter σ is estimated to be higher than the true value of zero. Of greater interest are the remaining parameters as we can compare the results for them between the two model types.

We first focus on the two degradation rate constants θ_1 and θ_3. Our analysis in Section 5.5 already showed that for the ODE model, these two parameters are only locally identifiable and the posterior distribution is symmetric with respect to them in the case of identical priors for both parameters. This is also apparent in the density plots in Figure 5.7. The density estimates of the posterior sample for the ODE model are clearly bimodal. The reason that the two modes are not exactly symmetric here is that HMC chains usually are only able to explore one mode and in our example 5 out of the 8 chains happen to end up in the mode where θ_1 is higher than θ_3 while only 3 chains converge to the other mode. The fact that each chain only samples from one of the modes is also the reason for the extremely low ESSs and the very high values of $\hat{R}$ for θ_1 and θ_3 in Table 5.1. Moreover, note that neither of the modes

Table 5.1: Summary of the Stan output for the ODE model given simulated data without measurement error and the true parameter values that were used to simulate the data. c.v. denotes the coefficient of variation and the columns headed by percentages contain the quantiles of the respective percentage value.

	true value	mean	c.v.	2.5%	50%	97.5%	n_{eff}	$\hat{R}$
θ_1	0.20	0.11	0.634	0.02	0.16	0.17	4	26.65
θ_2	0.32	1.52	1.370	0.02	0.64	7.56	12168	1.00
θ_3	0.01	0.07	0.943	0.02	0.02	0.16	4	33.00
m_0	240.00	204.62	1.017	2.26	135.79	724.38	10984	1.00
scale	1.80	7.02	1.137	0.07	3.46	27.28	9806	1.00
offset	6.50	6.50	0.011	6.37	6.50	6.64	17113	1.00
t_0	0.96	0.96	0.002	0.96	0.96	0.97	18718	1.00
σ	0.00	0.03	0.054	0.02	0.03	0.03	16091	1.00
$\theta_2 m_0$	76.80	213.08	2.487	4.57	35.99	1668.06	11087	1.00
$\theta_2 \mathrm{scale}$	0.58	6.46	2.875	0.17	0.92	55.00	7704	1.00
$m_0 \mathrm{scale}$	432.00	1033.30	2.181	16.48	195.96	7975.22	7899	1.00
$\theta_2 m_0 \mathrm{scale}$	138.24	124.67	0.007	122.96	124.66	126.38	13299	1.00

and not even the ranges of all values in the posterior sample cover the true parameter values of θ_1 and θ_3. For the SDE model on the other hand, Figure 5.7 and Table 5.2 show that the posterior density is clearly unimodal with respect to θ_1 and θ_3, the 95% CI are narrow and cover the true parameter values, mean and median of the sample are close or equal to the true values, and high ESSs and $\hat{R}$ values equal to 1 are achieved. Thus, we can conclude that the parameters θ_1 and θ_3 are identifiable for the SDE model here.

Next, we consider the translation rate constant θ_2, the initial amount m_0 of mRNA molecules, and the factor scale. For the ODE, at most the product $\theta_2 m_0 \mathrm{scale}$ is identifiable. This is also apparent from the results presented in Table 5.1 and Figure 5.8. For the individual parameters and also for all products of two out of the three parameters, the 95% CIs are extremely broad and the mean and median as point estimates are not at all close to the true values. The reason why there are nevertheless quite high ESSs and $\hat{R}$ values equal to 1 achieved is that the variation within each of the HMC chains is very high and thus does not differ substantially from the variation between the chains for these parameters. For the product $\theta_2 m_0 \mathrm{scale}$ of all three parameters, the 95% CI is very narrow for posterior sample of the ODE model and also the ESS is high and $\hat{R}$ equal to 1. Without knowing the true parameter values, one would assume that this product is well estimated. However, the 95% CI and even the whole range of the sample do not cover the true value. For the SDE model, the 95% CI for the product $\theta_2 m_0 \mathrm{scale}$ is broader, but it covers the true parameter value and also the mean and the median as point estimates are closer to the true value than the mean and the median for the ODE model. Moreover, the ESS is quite high and $\hat{R}$ is equal to 1 for the SDE model. We

Table 5.2: Summary of the Stan output for the SDE model given simulated data without measurement error and the true parameter values that were used to simulate the data.

	true value	mean	c.v.	2.5%	50%	97.5%	n_{eff}	$\hat{R}$
θ_1	0.20	0.19	0.108	0.15	0.19	0.23	3120	1.00
θ_2	0.32	0.39	0.999	0.09	0.26	1.48	206	1.04
θ_3	0.01	0.01	0.167	0.01	0.01	0.02	2514	1.00
m_0	240.00	344.37	0.589	57.21	313.25	800.67	184	1.05
scale	1.80	1.66	0.172	1.19	1.62	2.30	3062	1.00
$\theta_2 m_0$	76.80	82.60	0.178	55.68	81.69	113.53	5296	1.00
$\theta_2\mathrm{scale}$	0.58	0.61	0.923	0.17	0.43	2.21	178	1.05
$m_0\mathrm{scale}$	432.00	576.69	0.634	86.08	511.66	1440.41	232	1.04
$\theta_2 m_0\mathrm{scale}$	138.24	133.18	0.083	112.47	132.74	156.40	5829	1.00

therefore conclude that the product $\theta_2 m_0\mathrm{scale}$ is identifiable. The generally lower ESSs are due to the fact that for the SDE model, we sample from a distribution of much larger dimension as explained in Section 5.6.2. Additionally, for the SDE model, the parameters scale and $\theta_2 m_0$ have narrow 95% CIs (especially compared to those for the ODE model) that include the true parameter values, high ESSs, and $\hat{R}$ values of 1 and can therefore be considered identifiable. The remaining parameters θ_2, m_0, $\theta_2\mathrm{scale}$, and $m_0\mathrm{scale}$ have rather broad 95% CIs and only achieve low ESSs and $\hat{R}$ values higher than 1.02. Hence, they seem to be non-identifiable. Notice, however, that at least for the parameters θ_2 and $\theta_2\mathrm{scale}$, the 95% CIs are substantially more narrow for the SDE model compared to the ODE model.

We have simulated another 99 trajectories with the same parameters and performed Stan sampling in the same way as described in the beginning of this subsection. For each model type and each posterior sample of the different simulated trajectories, we calculate the length of the 95% CI and determine the median and the coefficient of variation (c.v.) over these lengths for each model type. Also, we rescale the lengths of the 95% CI by dividing by the true parameter value and again determine the median of the normalized quantities. The rescaling is done to transfer the values to a more similar scale. Note, however, that the values are nevertheless not directly comparable between different parameters. Moreover, we check whether the true parameter value that was used to simulate the data is included in the 95% CI. Table 5.3 shows the aggregated results for the posterior samples of all 100 trajectories and also includes the length of the interval between the 2.5%- and the 97.5%-quantile of the prior distributions. Except for the parameters m_0 and $\theta_2 m_0\mathrm{scale}$, the median length of the 95% CIs for the SDE model is always smaller than for the ODE model. For parameter $\theta_2 m_0\mathrm{scale}$, the CI lengths are a lot smaller for the ODE model; however, the CIs cover the true parameter value only 13 out of 100 times while for the SDE model, the true value is covered 93 times. For the other parameters that we classified as identifiable for the SDE model in the analysis

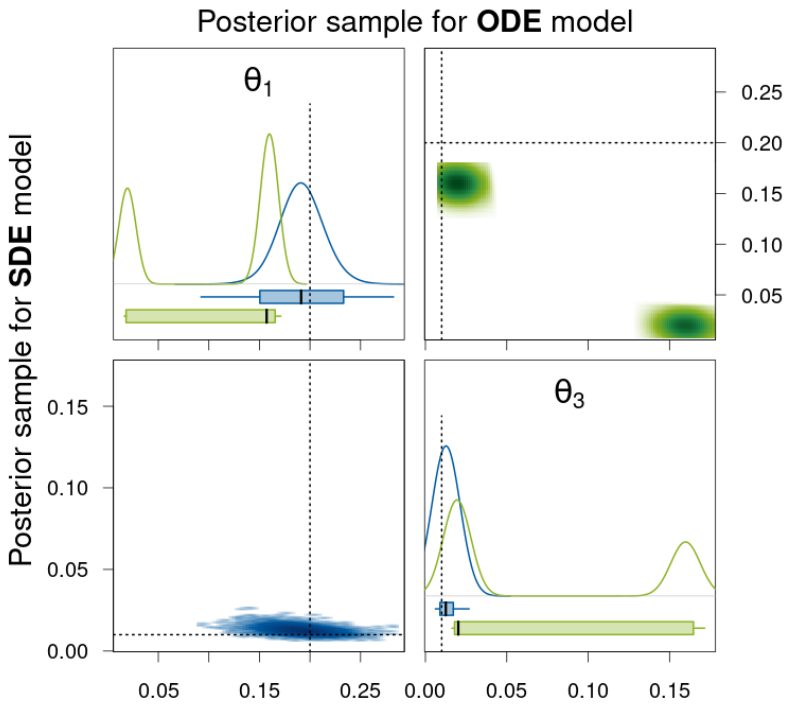


Figure 5.7: Density estimates of the posterior samples for parameters θ_1 and θ_3 for the SDE (blue, lower triangle) and ODE (green, upper triangle) model given simulated data without measurement error. *Diagonal panels:* Marginal densities for the respective parameter and boxplots showing the 95% CI as box, the range of the sample as whiskers, and the median as thick black line. *Off-diagonal panels:* Smoothed scatter plots of the two-dimensional projections of the samples where darker hues signify higher density values. The dotted lines represent the true parameter values that were used to simulate the data.

of the individual trajectory (i. e. θ_1, θ_3, scale, and $\theta_2 m_0$), the median length of the 95% CIs is clearly smaller for the SDE model than for the ODE model and the true parameter value is covered at least 91 out of 100 times for the SDE model. For parameter m_0, the CI lengths are high for both model types because the parameter is not identifiable for either model type. For the other parameters that we classified as not identifiable for both model types in the analysis of the individual trajectory (i. e. θ_2, $\theta_2 \mathrm{scale}$, and $m_0 \mathrm{scale}$), the median length of 95% CIs is clearly smaller for the SDE model than for the ODE model, at least by a factor of 4.

The last two columns of Table 5.3 are visualized in Figure 5.9 where we plot the median of the rescaled CI lengths against the number of CIs that cover the true parameter value. The desirable region of value combinations is in the bottom right corner of the graph where the number of CIs covering the true value is high and the median rescaled CI length is small. Note that, clearly, more importance should be given to high numbers of CIs covering the true value as it is useless to be very certain about a parameter estimate (indicated by a short CI) while the correct value is not included in the CI. However, even for parameters that are identifiable, we

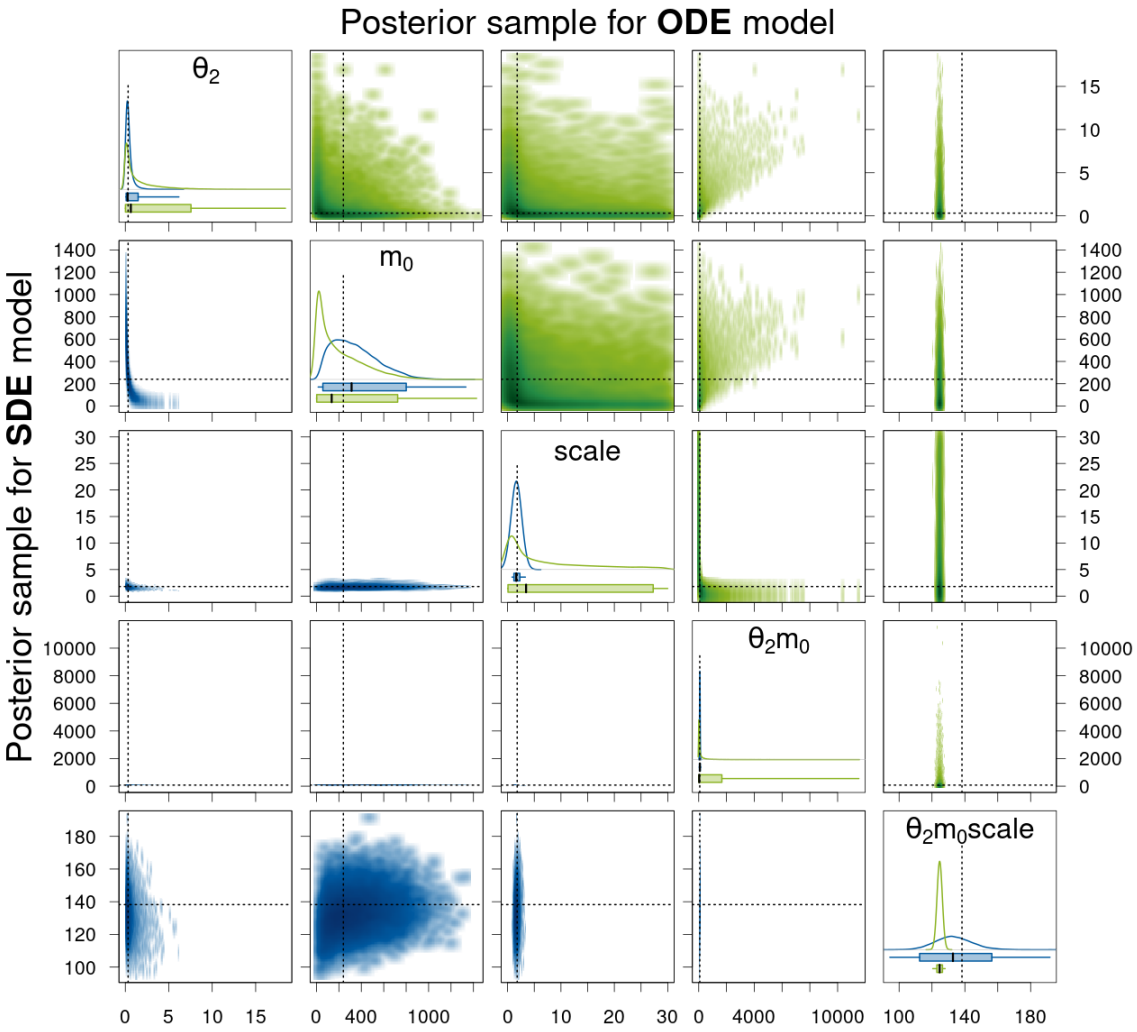


Figure 5.8: Density estimates of the posterior samples for parameters θ_2, m_0, scale, and their products for the SDE (blue, lower triangle) and ODE (green, upper triangle) model given simulated data without measurement error. For a detailed description of the figure's elements, see Figure 5.7.

do not expect to obtain a coverage of the true value of 100% since we are considering 95% CIs. Therefore, values of 100 rather tend to hint at non-identifiability. In Figure 5.9, we can see that for the majority of the parameters, the triangles representing the value combinations for the SDE model are closer to the desirable region. Only for parameter m_0 (which is not identifiable for either model type), the value combinations are almost the same for both model types. And as we already pointed out for the product $\theta_2 m_0 \text{scale}$, the median length of the 95% CIs is smaller for the ODE model; however, a lot fewer CIs cover the true parameter value for the ODE model than for the SDE model. Thus, the result obtained for the SDE model is to be preferred.

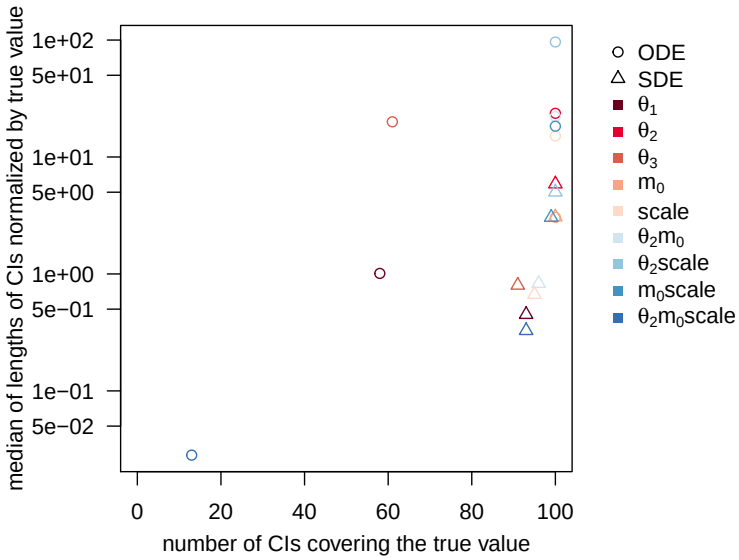


Figure 5.9: Statistics of posterior samples for the two model types aggregated over 100 simulated trajectories without measurement error. The desirable region of value combinations is in the bottom right corner of the graph.

We provide further Stan-specific diagnostics in Appendix A.3.2. Those mostly show poorer values for the sampling output for the SDE model than for the ODE model. This is not surprising as we sample from a much higher-dimensional distribution for the SDE model. We do not consider the poor diagnostics as a disadvantage of the procedure as they provide information that we do not even have for other MCMC algorithms and thus cannot compare to them.

Table 5.3: Statistics of posterior samples for the two model types aggregated over 100 simulated trajectories without measurement error. We also include the length of the interval between the 2.5%- and the 97.5%-quantile of the prior distribution.

		length of prior 95% center interval	median length of 95% CIs	c.v. of lengths of 95% CIs	median of length of CIs rescaled by true value	number of CIs covering true value
θ_1	ODE	11.05	0.20	0.009	1.01	58
	SDE	11.05	0.09	0.002	0.45	93
θ_2	ODE	11.05	7.56	0.002	23.63	100
	SDE	11.05	1.88	0.712	5.87	100
θ_3	ODE	11.05	0.20	0.010	20.00	61
	SDE	11.05	0.01	0.000	0.80	91
m_0	ODE	884.82	730.41	0.057	3.04	100
	SDE	884.82	735.98	9.868	3.07	100
scale	ODE	28.50	27.27	0.001	15.15	100
	SDE	28.50	1.20	0.158	0.67	95
$\theta_2 m_0$	ODE	6056.48	1701.92	2.524	22.16	100
	SDE	6056.48	63.60	2.768	0.83	96
$\theta_2 \mathrm{scale}$	ODE	228.08	55.47	0.164	96.29	100
	SDE	228.08	2.89	0.762	5.01	100
$m_0 \mathrm{scale}$	ODE	19271.13	7923.38	8.603	18.34	100
	SDE	19271.13	1315.63	86.806	3.05	99
$\theta_2 m_0 \mathrm{scale}$	ODE	113232.70	3.90	41.595	0.03	13
	SDE	113232.70	45.28	0.624	0.33	93

5.7.3 Simulated data with measurement error

In this section, we use the same simulated data as in the previous section, but for each of the 100 trajectories, we add multiplicative measurement error with parameter $\sigma = 0.02$. Again, we use Stan to sample from the posterior distributions of the ODE model (5.16) and the SDE model (5.17) for each of the simulated trajectories and use the same priors as stated in the previous section. We generate 8 HMC chains of 5000 iterations, discard the first half of the iterations as warm-up, and thus use a posterior samples of size 20,000 in the subsequent analysis.

At first, we again focus on the results for one of the trajectories, namely the trajectory represented by the green dotted line in Figure 5.6. Tables 5.4 and 5.5 summarize the Stan output of the posterior samples for the ODE and the SDE model, respectively. The parameter t_0 is estimated very accurately based on the posterior sample for the ODE model. Also, the parameter offset is well estimated for both model types but with a more narrow 95% CI for the SDE model. The parameter σ is accurately determined for the SDE model as well. For the ODE model, σ is again overestimated. Figure 5.10 visualizes the components of the posterior samples for parameters θ_1, θ_3, offset, and σ. Again, the bimodality of the posterior with respect to θ_1 and θ_3 is apparent for the ODE model and neither the 95% CIs nor the ranges of the sample cover the true parameter values. For the SDE model on the other hand, the distribution is unimodal and the 95% CIs do cover the true parameter values for θ_1 and θ_3. However, their 2-dimensional smoothed scatter plot in Figure 5.10 is not a simple elliptic shape (as for the simulated data without measurement error) but almost a banana-like shape. This may also be the reason for the deteriorated sampling efficiency discernible from the low ESS and higher $\hat{R}$-values in Table 5.5.

Figure 5.10 visualizes the components of the posterior samples for parameters θ_2, m_0, scale, and their products. For the ODE model, again only the product $\theta_2 m_0 \mathrm{scale}$ is identifiable in the sense that the corresponding 95% CI is very narrow, the ESS is high, and the $\hat{R}$-value is equal to 1. However, the 95% CI again does not cover the true parameter value. For the SDE model, the 95% CI for $\theta_2 m_0 \mathrm{scale}$ is broader but it does contain the true value. Also, the ESS is high and the $\hat{R}$-value is close to 1. Moreover, the parameters scale and $\theta_2 m_0$ have narrow 95% CIs, high ESSs, and $\hat{R}$-values close to 1 for the SDE model, and thus, we conclude that they are identifiable. Note that also for θ_2, $m_0 \mathrm{scale}$, and $\theta_2 \mathrm{scale}$, the 95% CIs are much narrower for the SDE model than for the ODE model.

In Appendix A.3.3, we include further figures of the sampling output for the trajectory displayed in Figure 5.6. They present the same posterior samples as used in this and the previous subsection. But instead of comparing the posterior samples between the two model types, the

Table 5.4: Summary of the Stan output for the ODE model given simulated data with measurement error and the true parameter values that were used to simulate the data.

	true value	mean	c.v.	2.5%	50%	97.5%	n_{eff}	$\hat{R}$
θ_1	0.20	0.11	0.632	0.02	0.15	0.17	4	20.94
θ_2	0.32	1.54	1.364	0.02	0.65	7.63	11619	1.00
θ_3	0.01	0.07	0.938	0.02	0.02	0.16	4	26.81
m_0	240.00	205.41	1.024	2.25	133.96	738.49	10649	1.00
scale	1.80	6.84	1.152	0.07	3.29	27.06	8778	1.00
offset	6.50	6.50	0.013	6.34	6.50	6.67	14496	1.00
t_0	0.96	0.96	0.003	0.96	0.96	0.97	17129	1.00
σ	0.02	0.03	0.053	0.03	0.03	0.04	13581	1.00
$\theta_2 m_0$	76.80	217.59	2.441	4.56	37.40	1683.06	9479	1.00
$\theta_2 \mathrm{scale}$	0.58	6.31	2.841	0.17	0.92	54.94	7209	1.00
$m_0 \mathrm{scale}$	432.00	983.07	2.151	16.20	189.75	7514.28	8054	1.00
$\theta_2 m_0 \mathrm{scale}$	138.24	123.47	0.009	121.43	123.46	125.58	12035	1.00

posterior samples are compared between the simulated data without and with measurement error for each model type separately. In summary, we find that for the SDE model, the 95% CIs increase for almost all parameters except m_0 for data with measurement error. Whereas for the ODE model, there is hardly any difference for most of the parameters between the posterior samples for the data without and with measurement error since the majority of the parameters is not identifiable anyway. The marginal posterior samples for the parameters offset, t_0, and $\theta_2 m_0 \mathrm{scale}$ are only slightly affected by the measurement error. Only the marginal posterior sample of the measurement error parameter σ is substantially affected and, as expected, consists of higher values for data with measurement error.

Table 5.6 and Figure 5.12 display the statistics of the posterior samples aggregated over the 100 simulated trajectories. Similar to the results for the simulated data without measurement error, the median length of the 95% CIs for the SDE model is always smaller than for the ODE model, except for the parameters m_0 and $\theta_2 m_0 \mathrm{scale}$ and additionally σ (which was not included for the SDE model in the previous subsection). Again, for the majority of the parameters, the results for the SDE model represented by triangles in Figure 5.12 are closer to the desirable region of value combinations in the bottom right corner of the graph, except for the parameters m_0, $\theta_2 m_0 \mathrm{scale}$, and offset. For the parameter offset, the median CI length is slightly higher for the ODE model than for the SDE model, however, the CIs for the ODE model also contain the true parameter value more often. So for this parameter, the ODE model, for once, shows the preferable result.

Table 5.5: Summary of the Stan output for the SDE model given simulated data with measurement error and the true parameter values that were used to simulate the data.

	true value	mean	c.v.	2.5%	50%	97.5%	n_{eff}	$\hat{R}$
θ_1	0.20	0.16	0.201	0.08	0.17	0.22	304	1.03
θ_2	0.32	0.73	1.448	0.07	0.37	3.93	296	1.02
θ_3	0.01	0.02	0.660	0.01	0.02	0.05	176	1.04
m_0	240.00	274.72	0.742	28.30	224.46	777.96	225	1.04
scale	1.80	1.67	0.430	0.65	1.54	3.36	415	1.02
offset	6.50	6.50	0.007	6.41	6.50	6.60	15599	1.00
σ	0.02	0.02	0.069	0.02	0.02	0.02	2106	1.00
$\theta_2 m_0$	76.80	89.78	0.473	35.58	80.76	191.24	349	1.02
$\theta_2$scale	0.58	0.93	1.145	0.16	0.56	4.15	247	1.03
$m_0$scale	432.00	477.62	0.944	29.55	338.17	1697.86	216	1.04
$\theta_2 m_0$scale	138.24	124.39	0.093	102.85	123.92	148.63	2119	1.01

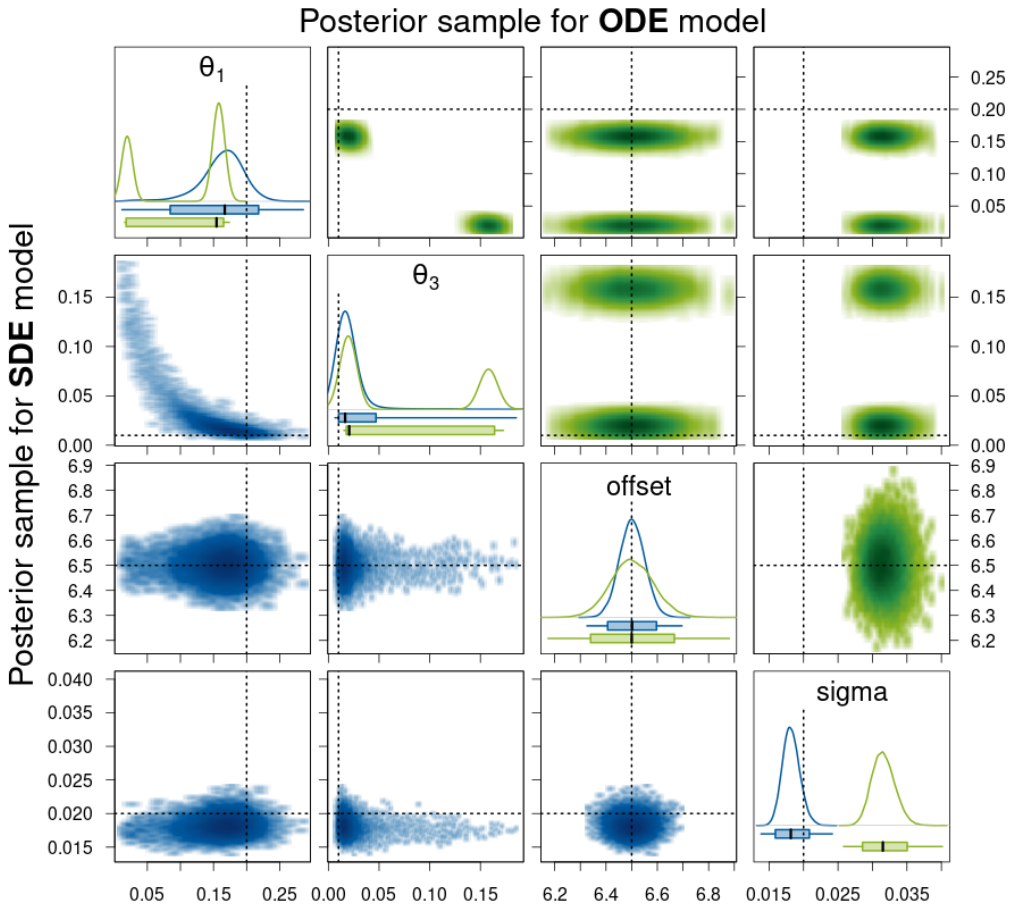


Figure 5.10: Density estimates of the posterior samples for parameters θ_1, θ_3, offset, and σ for the SDE (blue, lower triangle) and ODE (green, upper triangle) model given simulated data with measurement error. For a detailed description of the figure's elements, see Figure 5.7.

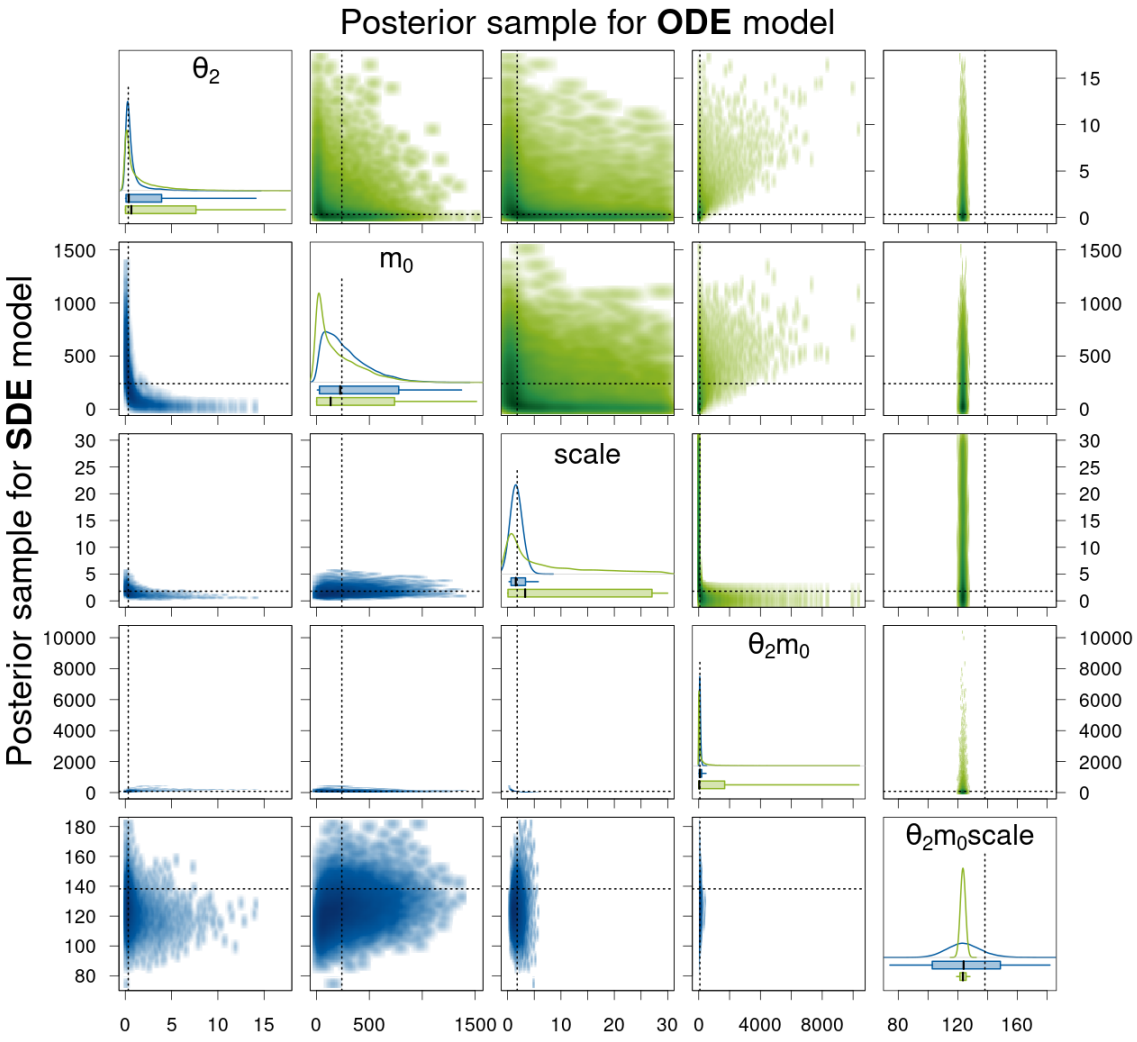


Figure 5.11: Density estimates of the posterior samples for parameters θ_2, m_0, scale, and their products for the SDE (blue, lower triangle) and ODE (green, upper triangle) model given simulated data with measurement error. For a detailed description of the figure's elements, see Figure 5.7.

Table 5.6: Statistics of posterior samples for the two model types aggregated over 100 simulated trajectories with measurement error. We also include the length of the interval between the 2.5%- and the 97.5%-quantile of the prior distribution.

		length of prior 95% center interval	median length of 95% CIs	c.v. of lengths of 95% CIs	median of length of CIs rescaled by true value	number of CIs covering true value
θ_1	ODE	11.05	0.20	0.008	1.01	60
	SDE	11.05	0.11	0.016	0.55	90
θ_2	ODE	11.05	7.55	0.002	23.59	100
	SDE	11.05	3.69	0.996	11.52	99
θ_3	ODE	11.05	0.20	0.008	20.35	63
	SDE	11.05	0.02	0.094	1.65	89
m_0	ODE	884.82	733.85	3.584	3.06	100
	SDE	884.82	746.74	5.909	3.11	100
scale	ODE	28.50	27.25	0.002	15.14	100
	SDE	28.50	2.54	0.255	1.41	91
$\theta_2 m_0$	ODE	6056.48	1702.37	2.714	22.17	100
	SDE	6056.48	223.22	428.350	2.91	92
$\theta_2\mathrm{scale}$	ODE	228.08	55.70	0.125	96.70	100
	SDE	228.08	3.55	0.639	6.16	100
$m_0\mathrm{scale}$	ODE	19271.13	7896.07	667.494	18.28	100
	SDE	19271.13	1479.13	253.530	3.42	98
$\theta_2 m_0\mathrm{scale}$	ODE	113232.70	4.96	38.038	0.04	15
	SDE	113232.70	45.56	1.492	0.33	92
offset	ODE	28.50	0.33	0.987	0.05	96
	SDE	28.50	0.21	0.016	0.03	84
σ	ODE	9.50	0.01	0.325	0.35	0
	SDE	9.50	0.01	0.000	0.25	87

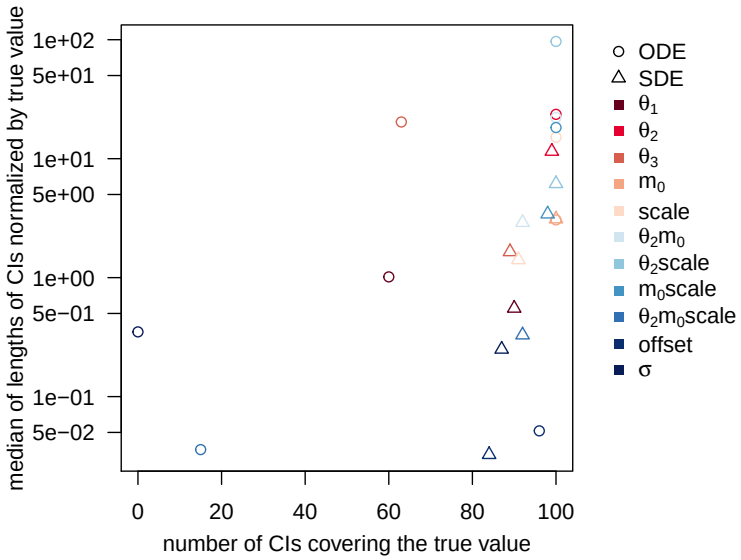


Figure 5.12: Statistics of posterior samples for the two model types aggregated over 100 simulated trajectories with measurement error. The desirable region of value combinations is in the bottom right corner of the graph.

5.8 Estimation based on experimental data

In this section, we use the experimental data published in Fröhlich et al. (2018) and described in Section 5.1. For each type of GFP (eGFP and d2eGFP), we randomly select 100 observed trajectories for our analysis and again use Stan to sample from the posterior distributions of the ODE model (5.16) and the SDE model (5.17) for each of the trajectories using the same priors as stated in Section 5.7.2. We generate 8 HMC chains of 5000 iterations, discard the first half of the iterations as warm-up, and thus use a posterior samples of size 20,000 in the subsequent analysis. For each type of GFP, we first analyze the sampling output for one observed trajectory in detail and then summarize results for all 100 observed trajectories. Moreover, we provide further Stan-specific diagnostics in Appendix A.3.2.

5.8.1 Experimental dataset 1 (for eGFP)

Tables 5.7 and 5.8 present a summary of the Stan output for the posterior sample of one observed trajectory for the ODE and the SDE model, respectively, and Figures 5.13 and 5.14 compare the density estimates of these two posterior samples. The results look qualitatively *very* similar (almost identical) to those obtained for the simulated data with measurement error in Section 5.7.3. Therefore, we do not repeat the detailed description but only point out that the range of values sampled for the parameters θ_1 and θ_3 for the SDE model is slightly smaller for the experimental trajectory here. Thus, we do not see the banana-like shape in the two-dimensional smoothed scatter plot of the two parameters for the SDE model in Figure 5.13 as for the simulated trajectory in Figure 5.10 and the sampling efficiency increases as indicated by higher ESSs and lower $\hat{R}$ values for the two parameters in Table 5.8.

The statistics of posterior samples aggregated for 100 experimental trajectories for eGFP in Table 5.9 are also qualitatively similar to those for the simulated trajectories in Table 5.6. For the majority of the parameters, the median length of the 95% CI is smaller for the posterior samples for the SDE model than for those for the ODE model. Only for parameters θ_1, θ_2, and $\theta_2 m_0 \mathrm{scale}$, this is not the case. Note in particular that for the parameters $\theta_2 m_0$ and scale, which are non-identifiable for the ODE (also apparent from the very long CIs here), the median length of the 95% CI for the SDE model is again much narrower compared to that of the ODE and to that of the prior. This indicates that these two parameters are identifiable for the SDE model also for the experimental data. That the uncertainty of the parameter estimate for $\theta_2 m_0 \mathrm{scale}$ is greater for the SDE than for the ODE model is consistent with our results for the simulated data. The parameter θ_2 is considered to be non-identifiable for both model types and the difference between the median CI lengths is relatively small. Finally, for parameter θ_1, we see that the result is more or less the same as for θ_3 for the ODE model due

Table 5.7: Summary of the Stan output for the ODE model given experimental data for eGFP.

	mean	c.v.	2.5%	50%	97.5%	n_{eff}	$\hat{R}$
θ_1	0.11	0.617	0.02	0.16	0.18	4	13.60
θ_2	1.44	1.407	0.01	0.56	7.39	12661	1.00
θ_3	0.08	0.903	0.02	0.03	0.18	4	16.98
m_0	198.99	1.046	1.75	127.89	731.84	12008	1.00
scale	6.64	1.186	0.05	3.01	27.00	10321	1.00
offset	7.18	0.017	6.94	7.18	7.42	17800	1.00
t_0	1.46	0.004	1.44	1.46	1.47	15996	1.00
σ	0.05	0.054	0.04	0.05	0.05	16833	1.00
$\theta_2 m_0$	200.04	2.654	3.17	28.49	1627.75	11121	1.00
$\theta_2$scale	5.53	3.043	0.12	0.67	48.94	8535	1.00
$m_0$scale	942.49	2.307	11.63	153.30	7610.72	9270	1.00
$\theta_2 m_0$scale	85.74	0.014	83.35	85.73	88.19	12577	1.00

Table 5.8: Summary of the Stan output for the SDE model given experimental data for eGFP.

	mean	c.v.	2.5%	50%	97.5%	n_{eff}	$\hat{R}$
θ_1	0.20	0.152	0.14	0.20	0.26	946	1.01
θ_2	0.33	1.366	0.04	0.19	1.52	305	1.02
θ_3	0.02	0.269	0.01	0.02	0.03	894	1.01
m_0	298.35	0.705	34.46	250.62	809.42	218	1.03
scale	2.12	0.309	1.14	2.02	3.69	772	1.01
offset	7.18	0.012	7.01	7.18	7.35	20589	1.00
σ	0.03	0.058	0.03	0.03	0.04	17224	1.00
$\theta_2 m_0$	47.92	0.327	23.41	46.00	83.31	857	1.01
$\theta_2$scale	0.60	1.182	0.11	0.37	2.54	256	1.02
$m_0$scale	650.64	0.831	57.98	498.32	2049.93	283	1.02
$\theta_2 m_0$scale	92.71	0.115	73.32	92.15	115.16	2711	1.00

to the symmetry of the posterior distribution with respect to these two parameters. Whereas for the SDE model there is no symmetry and there is more variance in the posterior samples with respect to θ_1 than to θ_3 (indicated by a greater median CI length). The smaller median CI length of θ_1 for the ODE model compared to the SDE model is due to the fact that for many of the observed trajectories the values of θ_1 and θ_3 seem to be very close together. In this case, the posterior distribution of the ODE model appears to be unimodal and the posterior variance with respect to the two parameters is small (and equal due to the symmetry). Thus, overall this variance is smaller than the posterior variance with respect to θ_1 for the SDE model.

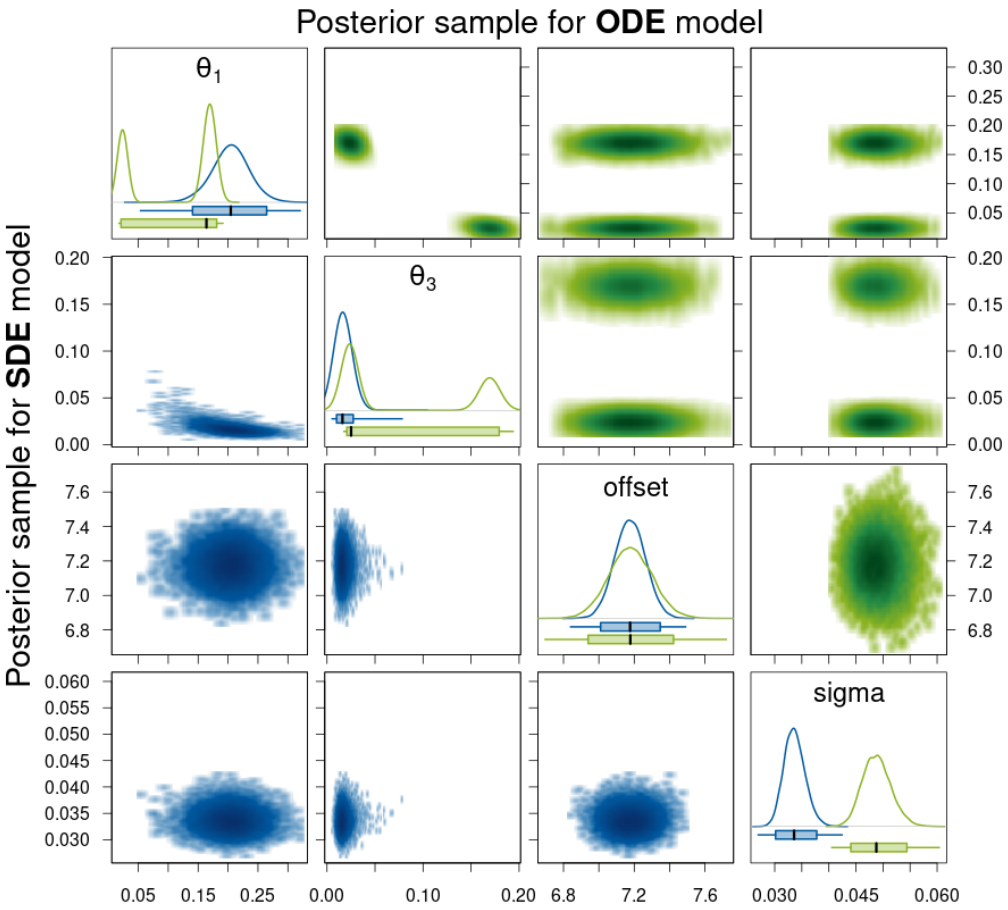


Figure 5.13: Density estimates of the posterior samples for parameters θ_1, θ_3, offset, and σ for the SDE (blue, lower triangle) and ODE (green, upper triangle) model given experimental data for eGFP. *Diagonal panels:* Marginal densities for the respective parameter and boxplots showing the 95% CI as box, the range of the sample as whiskers, and the median as thick black line. *Off-diagonal panels:* Smoothed scatter plots of the two-dimensional projections of the samples where darker hues signify higher density values.

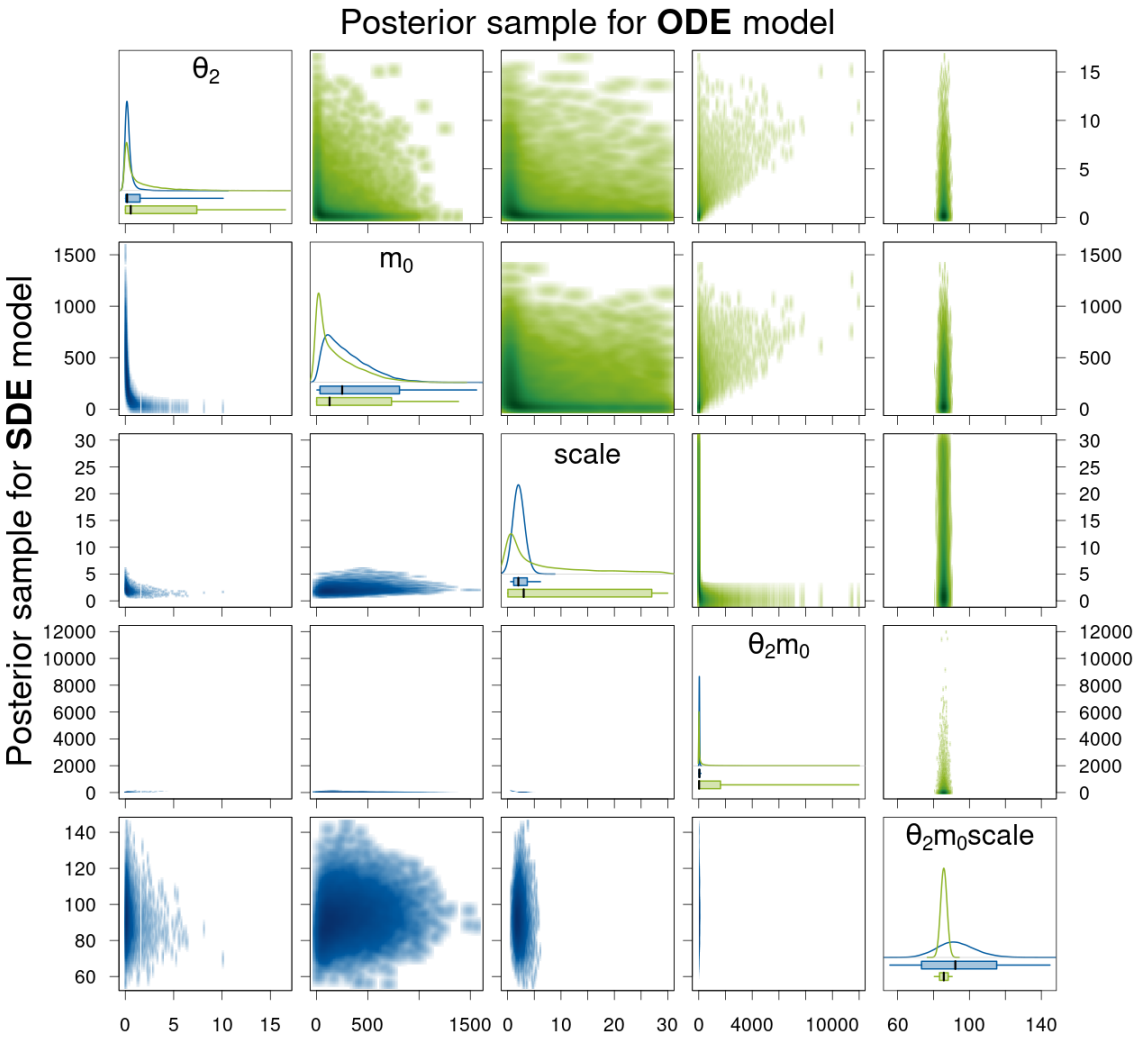


Figure 5.14: Density estimates of the posterior samples for parameters θ_2, m_0, scale, and their products for the SDE (blue, lower triangle) and ODE (green, upper triangle) model given experimental data for eGFP. For a detailed description of the figure's elements, see Figure 5.13.

Table 5.9: Statistics of posterior samples aggregated for 100 experimental trajectories for eGFP.

		length of prior 95% center interval	median length of 95% CIs	c.v. of lengths of 95% CIs
θ_1	ODE	11.05	0.11	0.058
	SDE	11.05	0.14	0.012
θ_2	ODE	11.05	7.91	0.016
	SDE	11.05	9.91	0.998
θ_3	ODE	11.05	0.11	0.057
	SDE	11.05	0.06	0.039
m_0	ODE	884.82	747.71	0.212
	SDE	884.82	456.80	172.338
scale	ODE	28.50	27.50	0.004
	SDE	28.50	4.61	5.288
$\theta_2 m_0$	ODE	6056.48	2032.46	23.287
	SDE	6056.48	230.89	219.006
$\theta_2\mathrm{scale}$	ODE	228.08	68.38	1.879
	SDE	228.08	22.01	33.157
$m_0\mathrm{scale}$	ODE	19271.13	9093.22	69.292
	SDE	19271.13	1392.46	4601.374
$\theta_2 m_0\mathrm{scale}$	ODE	113232.70	24.10	57.316
	SDE	113232.70	138.03	78.659
offset	ODE	28.50	0.96	2.661
	SDE	28.50	0.38	1.086
σ	ODE	9.50	0.01	0.244
	SDE	9.50	0.01	0.004

5.8.2 Experimental dataset 2 (for d2eGFP)

Tables 5.10 and 5.11 present a summary of the Stan output for the posterior sample of one observed trajectory for d2eGFP for the ODE and the SDE model, respectively, and Figures 5.15 and 5.16 compare the density estimates of these two posterior samples. Here, while of course still being symmetric, the posterior sample for the ODE model seems to be unimodal with respect to the parameters θ_1 and θ_3. This is due to the fact that the values of the two parameters are likely to be quite close to each other for this trajectory as can also be seen from the overlapping 95% CIs and the similar mean and median estimates for the SDE model. For the parameter offset, the mean and median estimates from the posterior samples are very similar for the ODE and SDE model, but the 95% CI is a lot wider for the ODE model. For the measurement error parameter σ, the 95% CI for the SDE model is a lot narrower than that for the ODE model and the locations of the samples are quite far apart with a difference in the median estimates of 0.16.

Table 5.10: Summary of the Stan output for the ODE model given experimental data for d2eGFP.

	mean	c.v.	2.5%	50%	97.5%	n_{eff}	$\hat{R}$
θ_1	0.09	0.079	0.08	0.09	0.11	11585	1.00
θ_2	2.03	1.114	0.08	1.17	8.33	12371	1.00
θ_3	0.09	0.078	0.08	0.09	0.11	11200	1.00
m_0	244.67	0.868	9.82	187.79	761.08	12127	1.00
scale	9.21	0.901	0.35	6.44	28.11	9845	1.00
offset	8.72	0.073	7.52	8.69	10.04	17557	1.00
t_0	0.94	0.011	0.92	0.94	0.96	15806	1.00
σ	0.17	0.053	0.15	0.17	0.18	16365	1.00
$\theta_2 m_0$	367.06	1.893	27.97	121.84	2279.87	11547	1.00
$\theta_2 \mathrm{scale}$	12.37	1.907	1.03	4.19	80.33	7681	1.00
$m_0 \mathrm{scale}$	1756.24	1.569	94.74	672.69	10085.29	8815	1.00
$\theta_2 m_0 \mathrm{scale}$	786.93	0.026	746.72	786.79	828.01	22688	1.00

For the parameters θ_2, m_0, scale, and their products, the results look somewhat different from those for the eGFP trajectory and those for the simulated data. For the product $\theta_2 m_0 \mathrm{scale}$, the 95% CI for the SDE model is again a lot wider than for the ODE model, but here, the CIs do not overlap. For the parameters scale and $\theta_2 m_0$, the 95% CI for the SDE model are again a lot narrower than for the ODE model, and we consider them as practically identifiable for the SDE model but not the ODE model. But here, also for the parameters m_0, $\theta_2 m_0$, and $m_0 \mathrm{scale}$, the 95% CI for the SDE model are much narrower than for the ODE model, and

Table 5.11: Summary of the Stan output for the SDE model given experimental data for d2eGFP.

	mean	c.v.	2.5%	50%	97.5%	n_{eff}	$\hat{R}$
θ_1	0.11	0.244	0.06	0.10	0.17	1494	1.01
θ_2	10.36	0.292	5.14	10.18	16.77	1226	1.01
θ_3	0.09	0.095	0.08	0.09	0.11	674	1.02
m_0	13.45	0.317	7.06	12.73	23.72	954	1.01
scale	4.93	0.212	3.21	4.83	7.27	785	1.01
offset	8.65	0.005	8.57	8.65	8.74	22392	1.00
σ	0.01	0.067	0.01	0.01	0.01	13124	1.00
$\theta_2 m_0$	130.52	0.229	80.35	127.18	196.70	897	1.01
$\theta_2 \mathrm{scale}$	49.67	0.282	27.16	48.09	82.52	838	1.01
$m_0 \mathrm{scale}$	65.47	0.363	34.66	60.26	125.01	1343	1.01
$\theta_2 m_0 \mathrm{scale}$	615.77	0.092	509.06	614.02	733.51	17424	1.00

the parameters seem to be practically identifiable. For parameter θ_2 the 95% CI for the SDE model is slightly wider than for the ODE model, however, the distribution looks different.

The statistics of posterior samples aggregated for 100 experimental trajectories for d2eGFP in Table 5.12 are qualitatively very similar to those for eGFP in Table 5.9. Therefore, we do not repeat the detailed description. We only point out that again unlike for the ODE model, the parameters scale and $\theta_2 m_0$ are identifiable for the SDE model which is indicated by the much narrower median length of the 95% CIs. We also want to mention that here, the median CI lengths for both degradation rate constants θ_1 *and* θ_3 are smaller for the ODE model than those for the SDE model. This is again due to the fact that for the majority of the observed trajectories the parameter values seem to be very close to each other; and therefore, the two modes of the ODE posterior distribution with respect to these parameters simply overlap. This leads to very narrow CIs which is consistent with our results for the simulated data if we consider the width of the individual modes there. However, we would like to remind the reader that the simulated data also showed that often neither of the modes (and sometimes not even the range of sampled values) covered the true parameter. So assuming that an MJP is the most appropriate description for the generating process of the experimental data, the low uncertainty suggested by narrow CIs for the ODE model might be misleading.

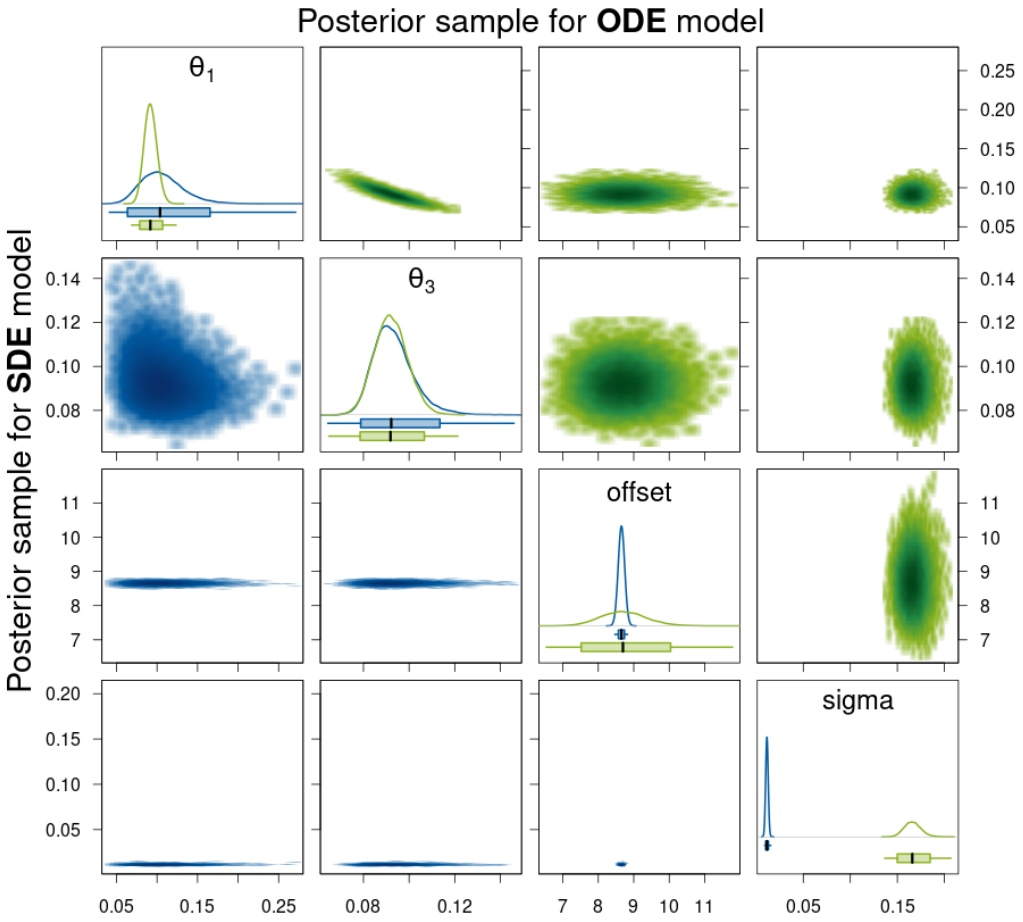


Figure 5.15: Density estimates of the posterior samples for parameters θ_1, θ_3, offset, and σ for the SDE (blue, lower triangle) and ODE (green, upper triangle) model given experimental data for d2eGFP. For a detailed description of the figure's elements, see Figure 5.13.

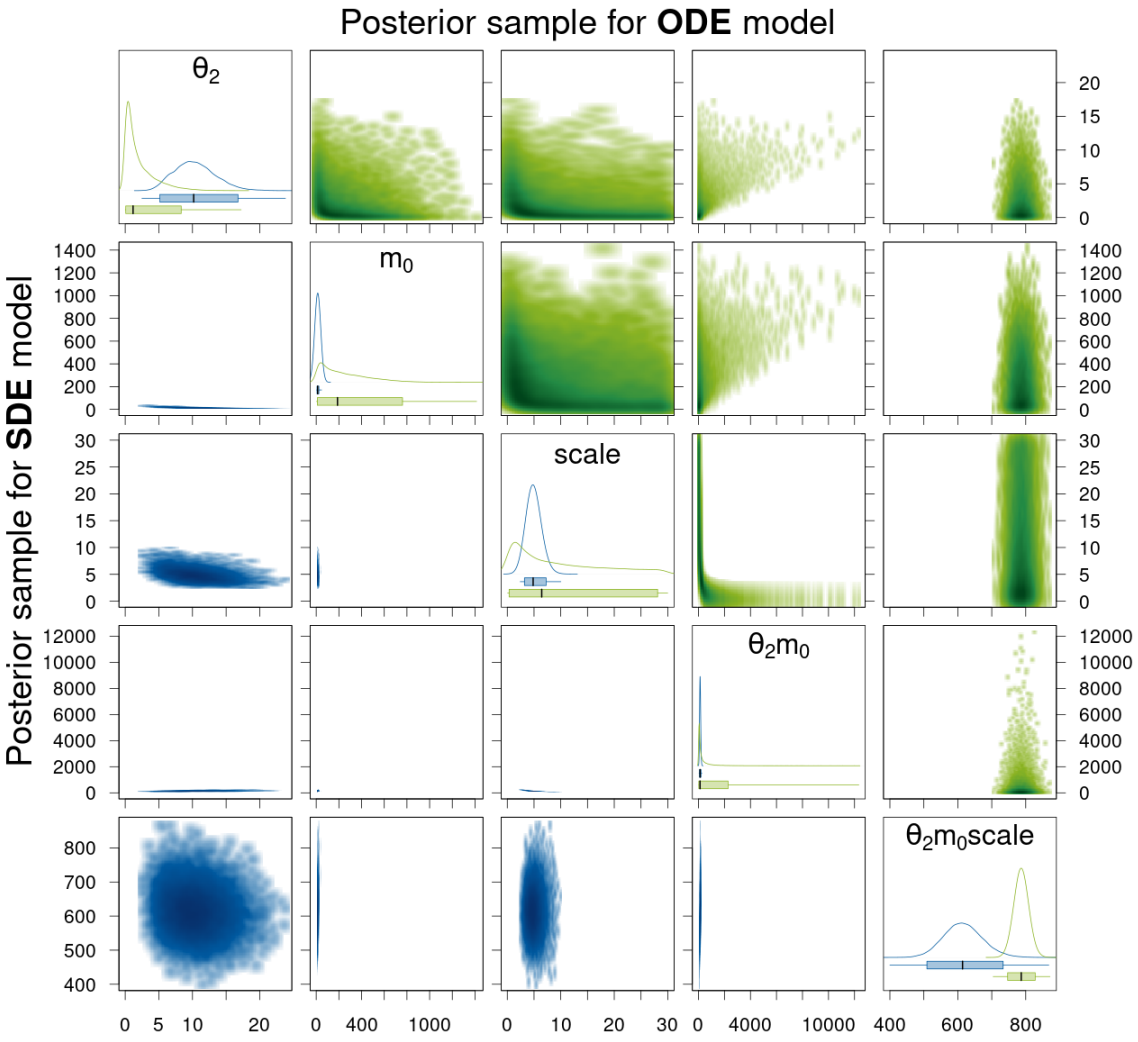


Figure 5.16: Density estimates of the posterior samples for parameters θ_2, m_0, scale, and their products for the SDE (blue, lower triangle) and ODE (green, upper triangle) model given experimental data for d2eGFP. For a detailed description of the figure's elements, see Figure 5.13.

Table 5.12: Statistics of posterior samples aggregated for 100 experimental trajectories for d2eGFP.

		length of prior 95% center interval	median length of 95% CIs	c.v. of lengths of 95% CIs
θ_1	ODE	11.05	0.03	0.061
	SDE	11.05	0.12	0.012
θ_2	ODE	11.05	7.88	0.006
	SDE	11.05	11.61	0.269
θ_3	ODE	11.05	0.03	0.062
	SDE	11.05	0.07	0.025
m_0	ODE	884.82	749.92	0.181
	SDE	884.82	76.56	224.650
scale	ODE	28.50	27.55	0.001
	SDE	28.50	5.51	5.550
$\theta_2 m_0$	ODE	6056.48	2048.33	15.537
	SDE	6056.48	172.02	173.641
$\theta_2\mathrm{scale}$	ODE	228.08	67.95	0.843
	SDE	228.08	34.96	44.862
$m_0\mathrm{scale}$	ODE	19271.13	9085.56	56.326
	SDE	19271.13	482.87	3408.541
$\theta_2 m_0\mathrm{scale}$	ODE	113232.70	40.80	30.553
	SDE	113232.70	145.18	83.283
offset	ODE	28.50	2.02	1.751
	SDE	28.50	0.79	1.029
σ	ODE	9.50	0.03	0.006
	SDE	9.50	0.01	0.005

5.9 Summary and discussion

In this chapter, we have modeled the translation kinetics after mRNA transfection using a two-dimensional Itô diffusion process described by an SDE and compared this modeling approach to one using ODEs. For the SDE model, we have proved the existence and uniqueness of the solution for the SDE and the convergence of the Euler scheme to this solution. The proof of these essential results should be a prerequisite to the application of SDE modeling for obvious reasons. Although SDE approximations by now have received some considerable attention in systems biology (usually termed chemical Langevin equation (CLE) in this field) and this type of SDEs generally does not fulfill the assumptions for standard existence and convergence results, this aspect is usually neglected. In our case, the proofs rely on the fact that the first component of the process only depends on itself (and the parameters) but not on the second process component. Nevertheless, the proofs are not straightforward, and in particular, proofs of convergence of the Euler scheme for SDEs with non-Lipschitz continuous diffusion coefficient functions are still an active field of research as can be seen from recently published work, e. g. Yang et al. (2019) where each component of the diffusion coefficient is only allowed to depend on the respective process component (whereas in our case the diffusion coefficient of the second component depends on *both* process components). Our proofs can be analogously extended for an SDE model of the translation kinetics after mRNA transfection when including a maturation step for the protein molecules before they start to glow as has been considered in the context of ODEs in Reiser et al. (2019). More generally speaking, the idea of the proofs can be sequentially applied to any diffusion approximation

- whose process components can be ordered in a way such that each process component $X_i(t)$ depends at most on the process components $X_j(t)$ with $j \leq i$; and
- that only includes first order reactions, i. e. the terms inside the square root functions of the diffusion coefficient are linear.

An extension to SDE models such as considered in case study 3 in Finkenstädt et al. (2008) where a deterministic, continuous function for the transcription process is included for the first process component is also feasible.

Moreover, we have studied the parameter identifiability for both modeling approaches (SDE vs. ODE) for the case that we observe a fluorescence signal which we assume to be a linear transformation of the amount of protein molecules (corrupted by multiplicative measurement error). For the ODE model, previous studies had already shown that the degradation rate constants θ_1 and θ_3 for the mRNA and the protein are only locally identifiable, and only the product $\theta_2 m_0 \mathrm{scale}$ of the translation rate constant, the initial amount of mRNA molecules

transfected, and the scaling factor of the fluorescence signal is identifiable but the three parameters individually are not identifiable. In order to try to assess structural identifiability of the SDE model, we transformed the model, used the DAISY software, and also simulated from the model. Each of the approaches indicated that the SDE model might lead to better parameter identifiability. The most systematic approach is the one based on the surrogate model and DAISY as suggested by Browning et al. (2020); however, it only provides a necessary condition (even) for structural identifiability of the SDE model parameters. While checking this necessary condition is certainly useful especially e. g. when designing an experiment, it cannot help us *confirm* a difference in the parameter identifiability between the SDE and the ODE model. Especially because we are interested in the parameter identifiability based on one observed trajectory and the DAISY-based approach assumes that we were able to observe the first and the second moment of the fluorescence signal. Even when we take into account that we have several observed trajectories available from the experiment, these do not provide information about the moments because the initial time point t_0 of mRNA release is different for every trajectory and also for the other parameters, in particular for m_0, assuming that they are equal for all observed cells does not seem reasonable. By simulating from the SDE model, we were able to assess the differences in the variation within individual trajectories for different parameter combinations. We saw that the variation within trajectories was clearly higher for lower θ_1 and higher θ_3 which suggest that they are structurally globally identifiable. The variation within trajectories was also higher for higher values of scale and lower values of the product $\theta_2 m_0$. Whereas there did not seem to be much difference in the variation within trajectories when the values of scale and $\theta_2 m_0$ were kept constant and only the individual values of θ_2 and m_0 varied. Therefore, scale and $\theta_2 m_0$ seem to be structurally identifiable, but θ_2 and m_0 do not. While this simple simulation approach worked out well for the model considered here, one of its weak points is, of course, the somewhat subjective visual assessment of the variation within trajectories. A more quantitative approach to this would be to simulate a large number of trajectories (with very small time step) for every considered parameter combination, to approximate the quadratic variation for each trajectory, and then, to compare these values between individual trajectories started with the same seed for different parameter combinations and to compare also the distributions of these values for different parameter combinations. Another drawback of both simulation-based approaches is the fact that the analysis is based on a finite set of parameter combinations that can be considered; and thus, drawing general conclusions for the entire parameter space may be problematic.

Finally, we have assessed the practical parameter identifiability for both model types by sampling from the parameter posterior distribution given simulated data without and with measurement error and the experimental data published in Fröhlich et al. (2018). We found that the parameters θ_1 and θ_3 are indeed globally identifiable for the SDE model given individual trajectories, unlike for the ODE model. And not only the product $\theta_2 m_0 \mathrm{scale}$ but also the

parameter scale and the product $\theta_2 m_0$ are globally identifiable for the SDE model. Moreover, for the simulated data, the 95% CIs for the identifiable parameters for the SDE model covered the true parameter value adequately many times. Whereas for the ODE model, the true parameter values for the parameters θ_1, θ_3, and $\theta_2 m_0 \mathrm{scale}$ were not covered by the 95% CIs for many of the posterior samples and were sometimes not even included in the range of values in the sample. The fact that the parameters θ_1 and θ_3 can be adequately determined using the SDE modeling approach given an individual trajectory renders the multi-experiment approach with different mRNA constructs and the computationally intense hierarchical optimization algorithm used in Fröhlich et al. (2018) unnecessary in the case that the determination of these parameters is the main objective. Besides, assuming that an MJP is the most appropriate description of the underlying dynamics, we saw that the estimated parameter values for a single cell trajectory based on the ODE model cannot be trusted even when narrow 95% CIs suggest low uncertainty. While the SDE model is clearly superior in terms of the information that we are able to extract from a single trajectory about the parameters that determine the dynamics of the underlying process, it has nevertheless several disadvantages. First of all, we were not able to include the estimation of the initial time point t_0 of mRNA release into the Stan sampling procedure. We believe that this is not easily possible due to the fact that for the SDE model, the process switches from a deterministic evolution to a stochastic one at t_0 and including t_0 as a parameter in the posterior distribution leads to non-smoothness of the posterior distribution which cannot be handled by HMC sampling as it makes use of the derivative of the log-posterior. Other sampling approaches such as particle MCMC might alleviate this problem, but to our knowledge, no examples of inferring a random time point for SDE models have been investigated so far and would thus require further work. Another drawback of the SDE model are the higher computational costs as we need to sample from a higher-dimensional distribution (due to the random process values) than for the ODE model. For the SDE model, the sampling in our study takes on average almost 5.5 hours while for the ODE model, it averages at about 20 minutes. In general, estimation procedures for SDE models are more complex and unlike for ODE models, publicly available software tools are rare and usually not generally applicable. There is definitely a need to further develop such tools for SDE models in order to harness their full potential, especially with regard to better identifiability of kinetic parameters. On the other hand, combining both modeling approaches as we have done here by first determining t_0 based on the ODE model and then estimating the kinetic parameters based on the SDE model is clearly also meaningful.

Chapter 6

Summary and conclusion

In this thesis, we considered inference methods for diffusion processes described by SDEs, applied diffusions to model the translation kinetics after mRNA transfection, proved important theoretical results of this model, and compared this SDE model to an ODE model in terms of parameter identifiability.

We investigated the use of a higher-order approximation scheme in the context of Bayesian data augmentation for inference for diffusion processes with the aim of improving computational efficiency and thus obtaining more accurate estimation results within a given computational time. We found that, in fact, the use of the Milstein scheme does improve the estimation accuracy for the parameters appearing in the diffusion coefficient. However, our study also shows that the applicability of the Milstein scheme is very limited in this context in the case of multi-dimensional processes. This is a major drawback compared to the generally applicable Euler scheme. Even for the comparatively small reaction network that we considered in Chapter 5 and that represents two species and three reactions, the methods based on the Milstein scheme considered in Chapter 4 cannot be applied as two components of the diffusion coefficient depend on the first process component. Yet, our analysis answers a natural question that had not been addressed in the literature previously.

For the application in Chapter 5, we instead use the open source software Stan that provides an efficient implementation of a general state-of-the-art MCMC method and achieve good sampling results. Even though some of the diagnostics that are specific to Stan are not perfect for the sampling output for the SDE model, the diagnostics that one would commonly look at for general MCMC output are satisfying and inference from simulated data shows that most parameters can be adequately recovered. We do not consider the poor Stan-specific diagnostics as a disadvantage of the procedure, as they provide information that we do not even have for

other MCMC algorithms. One major advantage of using Stan is that hardly any hand-tuning is required unlike for other MCMC algorithms. Comparing its performance to other MCMC methods in the context of SDE inference is a direction for future work. Also, a recent review and a comprehensive benchmark study for a wide range of MCMC methods for SDE inference represents relevant future work.

Moreover, our results for the application example in Chapter 5 showed that the SDE model provides better identifiability of the kinetic parameters than the ODE model. We found that the degradation rate constants of mRNA and GFP are indeed globally identifiable for the SDE model given individual trajectories, while for the ODE model they are only locally identifiable due to symmetry. Besides, not only the product $\theta_2 m_0 \mathrm{scale}$ of the translation rate constant, the initial amount of the mRNA, and the scaling factor of the fluorescence signal is identifiable as for the ODE model, but also the parameter scale and the product $\theta_2 m_0$ are globally identifiable for the SDE model. Not all model parameters could be determined solely from the fluorescence signal of the GFP molecules. Additional experiments to gain information about one of the two unidentifiable parameters (the initial amount m_0 of mRNA and the translation rate constant θ_2) would be necessary in order to be able to also estimate the other. Also, we combined both modeling approaches to predetermine the initial time point t_0. Still, our results once again underline that using a model that explicitly accounts for inherent stochasticity can lead to additional parameter identifiability. Despite this potential, SDE models are not that commonly applied for parameter inference from experimental data, neither in method articles nor in application articles. From the application side, one crucial reason for this gap is probably the lack in the available software tools for inference for SDE models (and other stochastic kinetic models). Even method articles that simply publish the code used to generate the results for the article are rare. To facilitate the intelligibility of this thesis, all relevant code used to obtain the results included in this work is made publicly available. Developing widely applicable tools is an enormous task but it is a necessary step to make stochastic models more usable and to leverage their potential. Therefore, future research should focus on this development, at best by an interdisciplinary team in order to make sustainable and rapid progress.

We have also proved essential theoretical results for the considered SDE model and pointed out several important further examples for which our proofs can be easily extended. However, an extension to general diffusion approximations would require substantial further work. Nevertheless these results are crucial. Especially for inference from individual time-lapse trajectories of which more and more are becoming available with the advancement of single-cell experimental methods, ensuring *strong* uniqueness of the solution and *strong* convergence of the approximation scheme is necessary. Furthermore, it would be very interesting to see further work on mathematical results about structural identifiability of SDE model parameters based on individual observed trajectories.

As pointed out before, in order to harness the capabilities of mathematical modeling to generate practical insights, ensuring a sound mathematical foundation is key. Moreover, it is important to develop tools that render analyzing and solving the corresponding mathematical problem (computationally) feasible within an acceptable amount of time. Likewise, one has to find ways how to deal with the challenges that arise when working with experimental data (e. g. finite amount of data, unobservable components). This thesis addresses all three of these aspects in the context of SDE models for intracellular processes and thus provides further building blocks to pave the way towards a holistic understanding of biological systems.

Appendix A

Appendix

A.1 Mathematical basics

Gronwall's lemma

An important tool to obtain estimates in the context of ordinary as well as stochastic differential equations is Gronwall's lemma (also known as Gronwall's inequality). There are different formulations of Gronwall's lemma. Here, we state the most simple formulation that is suitable for our purposes.

Theorem A.1. *(Gronwall's lemma) Let $T > 0$ and $c \geq 0$. Let $u(\cdot)$ be a non-negative continuous function on $[0, T]$, and let $\beta(\cdot)$ be a non-negative continuous and integrable function on $[0, T]$. If*

$$u(t) \leq c + \int_0^t \beta(s)u(s)\,\mathrm{d}s \qquad \text{for all } t \in [0, T],$$

then

$$u(t) \leq c \exp\left(\int_0^t \beta(s)\,\mathrm{d}s\right) \qquad \text{for all } t \in [0, T].$$

A proof of Theorem A.1 and more general formulations of Gronwall's lemma can be found e. g. in Pachpatte (1998).

Algorithm for the exact simulation of the Cox-Ingersoll-Ross (CIR) process

For the exact simulation of the CIR process described by SDE (3.9) at time points $0 = t_0 < t_1 \ldots < t_n$ and with $d = 4\alpha\beta/\sigma^2$, we use the following algorithm as stated in Glasserman (2003, p. 124):

Case 1: $d > 1$

For $i = 0, \ldots, n-1$

- $c \leftarrow \sigma^2 \left(1 - e^{-\alpha(t_{i+1} - t_i)}\right) / (4\alpha)$
- $\lambda \leftarrow X_{t_i} e^{-\alpha(t_{i+1} - t_i)} / c$
- generate $Z \sim \mathcal{N}(0, 1)$
- generate $Y \sim \chi^2_{d-1}$
- $X_{t_{i+1}} \leftarrow c \left[\left(Z + \sqrt{\lambda} \right)^2 + Y \right]$

Case 2: $d \leq 1$

For $i = 0, \ldots, n-1$

- $c \leftarrow \sigma^2 \left(1 - e^{-\alpha(t_{i+1} - t_i)}\right) / (4\alpha)$
- $\lambda \leftarrow X_{t_i} e^{-\alpha(t_{i+1} - t_i)} / c$
- generate $N \sim \mathrm{Po}(\lambda/2)$
- generate $Y \sim \chi^2_{d+2N}$
- $X_{t_{i+1}} \leftarrow cY$

where χ^2_k denotes the central chi-square distribution with k degrees of freedom and $\mathrm{Po}(\lambda)$ denotes the Poisson distribution with parameter λ.

A.2 Details for Bayesian data augmentation for diffusion processes

A.2.1 Choice of path update interval

For choosing the update interval in the simulation study in Section 4.3, we use the random block size algorithm as suggested in (Elerian et al., 2001). Assuming that the augmented path contains a total of $n + 1$ data points $Y_0, \ldots, Y_n$, it is divided into update segments $Y_{(c_0, c_1)}, Y_{(c_1, c_2)}, \ldots$ by the following algorithm:

1. Set $c_0 = 0$ and $j = 1$.
2. While $c_{j-1} < n$:
   (a) Draw $Z \sim \mathrm{Po}(\lambda)$ and set $c_j = \min\{c_{j-1} + Z, n\}$.
   (b) Increment j.

Here, $\mathrm{Po}(\lambda)$ denotes the Poisson distribution with parameter λ.

Such a random choice of the path update interval is a simple way to vary the set of points that are updated together within one iteration.

A.2.2 Derivation of the acceptance probability for the modified bridge (MB) proposal for $m = 2$ inter-observation intervals

As stated in Section 3.4.1, the acceptance probability for the path update between two consecutive observations X_{τ_i} and $X_{\tau_{i+1}}$ with the MB proposal is

$$\begin{aligned}\zeta\left(X^{imp*}_{(\tau_i,\tau_{i+1})}, X^{imp}_{(\tau_i,\tau_{i+1})}\right) &= 1 \wedge \frac{\pi\left(X^{imp*}_{(\tau_i,\tau_{i+1})} \,|\, X^{obs}_{\{\tau_i,\tau_{i+1}\}}, \theta\right) q_{MB}\left(X^{imp}_{(\tau_i,\tau_{i+1})} \,|\, X_{\tau_i,}, X_{\tau_{i+1}}, \theta\right)}{\pi\left(X^{imp}_{(\tau_i,\tau_{i+1})} \,|\, X^{obs}_{\{\tau_i,\tau_{i+1}\}}, \theta\right) q_{MB}\left(X^{imp*}_{(\tau_i,\tau_{i+1})} \,|\, X_{\tau_i,}, X_{\tau_{i+1}}, \theta\right)} \\ &= 1 \wedge \prod_{k=0}^{m-1} \frac{\pi\left(X^*_{t_{k+1}} \,|\, X^*_{t_k}, \theta\right)}{\pi\left(X_{t_{k+1}} \,|\, X_{t_k}, \theta\right)} \prod_{k=0}^{m-2} \frac{\pi\left(X_{t_{k+1}} \,|\, X_{t_k}, X_{\tau_{i+1}}, \theta\right)}{\pi\left(X^*_{t_{k+1}} \,|\, X^*_{t_k}, X_{\tau_{i+1}}, \theta\right)}\end{aligned}$$

where $X^*_{t_0} = X_{t_0} = X_{\tau_i}$ and $X^*_{t_m} = X_{t_m} = X_{\tau_{i+1}}$. For the case where only one data point is imputed between two observations (i.e. $m = 2$) this reduces to

$$\begin{aligned}\zeta\left(X^{imp*}_{(\tau_i,\tau_{i+1})}, X^{imp}_{(\tau_i,\tau_{i+1})}\right) &= 1 \wedge \frac{\pi\left(X^*_{t_1} \,|\, X_{\tau_i}, \theta\right) \pi\left(X_{\tau_{i+1}} \,|\, X^*_{t_1}, \theta\right)}{\pi\left(X_{t_1} \,|\, X_{\tau_i}, \theta\right) \pi\left(X_{\tau_{i+1}} \,|\, X_{t_1}, \theta\right)} \frac{\pi\left(X_{t_1} \,|\, X_{\tau_i}, X_{\tau_{i+1}}, \theta\right)}{\pi\left(X^*_{t_1} \,|\, X_{\tau_i}, X_{\tau_{i+1}}, \theta\right)} \\ &= 1 \wedge \left[\frac{\pi\left(X^*_{t_1} \,|\, X_{\tau_i}, \theta\right) \pi\left(X_{\tau_{i+1}} \,|\, X^*_{t_1}, \theta\right)}{\pi\left(X_{t_1} \,|\, X_{\tau_i}, \theta\right) \pi\left(X_{\tau_{i+1}} \,|\, X_{t_1}, \theta\right)} \right. \\ &\qquad \left. \frac{\pi\left(X_{t_1} \,|\, X_{\tau_i}, \theta\right) \pi\left(X_{\tau_{i+1}} \,|\, X_{t_1}, \theta\right) / \pi\left(X_{\tau_{i+1}} \,|\, X_{\tau_i}, \theta\right)}{\pi\left(X^*_{t_1} \,|\, X_{\tau_i}, \theta\right) \pi\left(X_{\tau_{i+1}} \,|\, X^*_{t_1}, \theta\right) / \pi\left(X_{\tau_{i+1}} \,|\, X_{\tau_i}, \theta\right)}\right] \\ &= 1.\end{aligned}$$

This relation holds for any (approximated) transition density $\pi\left(X_{t_{k+1}} \,|\, X_{t_k}, \theta\right)$.

A.2.3 Analysis of the correlation between the parameters

In this section, we provide several plots (see Figures A.1, A.2, A.3, and A.4) showing that the parameters of the two benchmark models are not strongly correlated in order to justify our use of independent parameter proposals in the simulation study in Section 4.3.

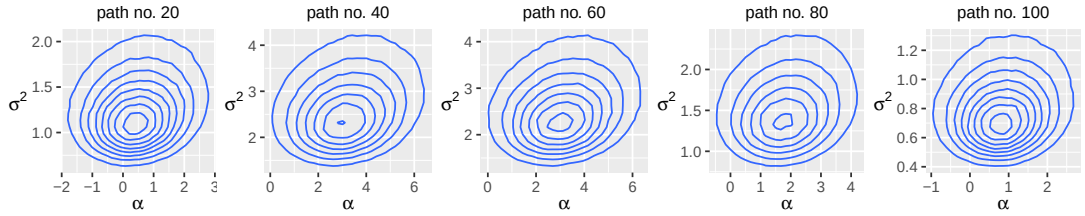


Figure A.1: Two-dimensional density plots of the parameter samples from the true posterior distribution for exemplary paths of the GBM.

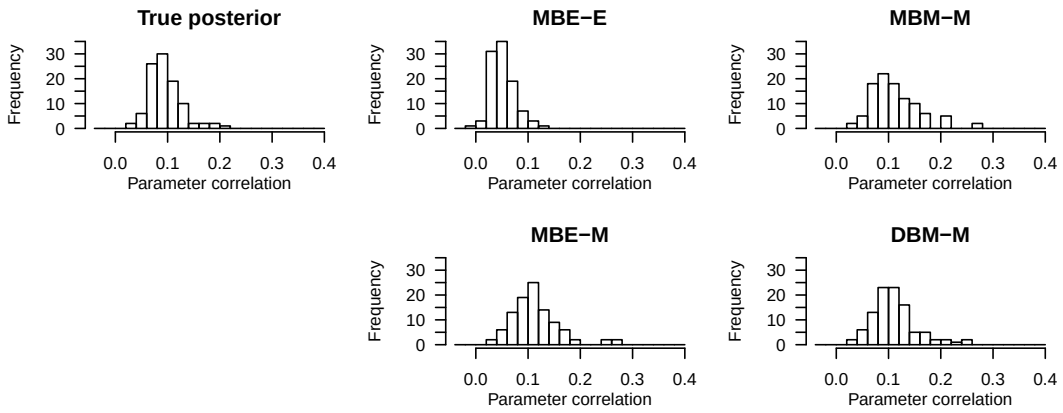


Figure A.2: Histograms of the values of Pearson's correlation coefficient calculated for each of the 100 sample paths of the GBM for the parameter samples from the true posterior distributions and the parameter samples from the approximated posterior distributions obtained with one of the four considered methods for $m = 5$.

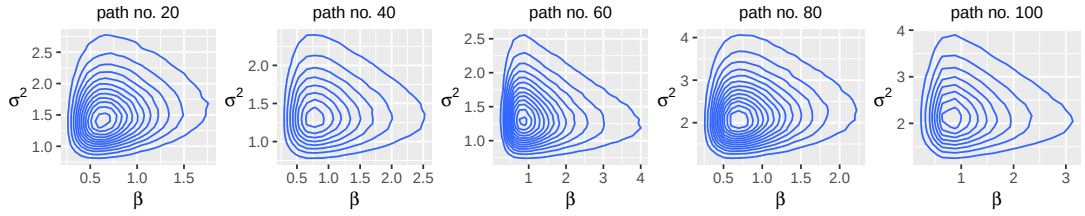


Figure A.3: Two-dimensional density plots of the parameter samples from the true posterior distribution for exemplary paths of the CIR process.

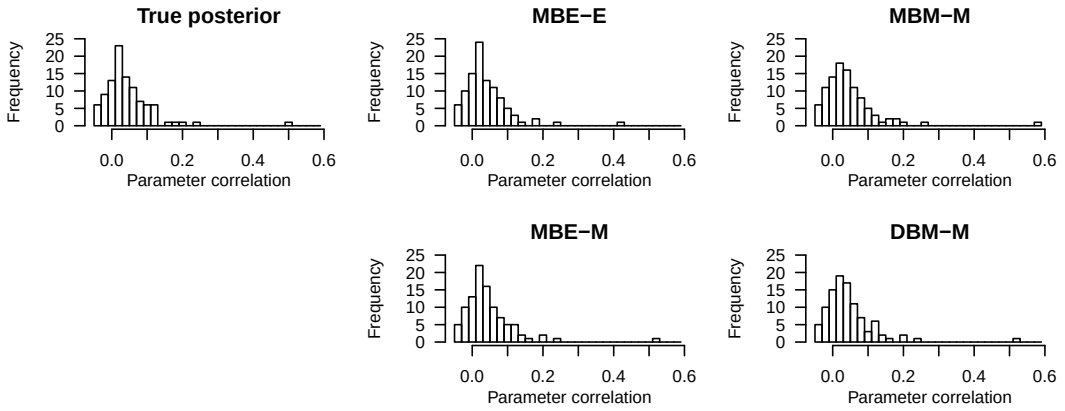


Figure A.4: Histograms of the values of Pearson's correlation coefficient calculated for each of the 100 sample paths of the CIR process for the parameter samples from the true posterior distributions and the parameter samples from the approximated posterior distributions obtained with one of the four considered methods for $m = 5$.

A.3 Details of the parameter estimation for the translation kinetics models

In this section, we provide some additional information about the sampling diagnostics and the estimation results for the models of the translation kinetics in Sections 5.7 and 5.8.

A.3.1 Further diagnostics of MCMC output specific to HMC and NUTS

In addition to the quality indicators for MCMC output mentioned in Section 2.2.3, Stan reports further quantities that are specific to HMC and NUTS and are of interest to assess sampling efficiency. These include the number of divergent transitions, the tree depth, and the (energy) Bayesian fraction of missing (BFMI) which we briefly describe below. See the Stan reference manual for more detailed explanations (Stan Development Team, 2019).

Integrating the Hamiltonian equations (2.8) in Section 2.2.2 analytically would preserve the value of the Hamiltonian $H(\boldsymbol{\theta}, \boldsymbol{\rho})$; however, since analytical integration is not possible for most problems of interest, the equations are numerically integrated which leads to numerical errors. If the difference between $H(\boldsymbol{\theta}, \boldsymbol{\rho})$ of the starting point and $H(\boldsymbol{\theta}^*, \boldsymbol{\rho}^*)$ of the proposed point at the end of the simulated Hamiltonian trajectory becomes too large (where the default threshold is 10^3), Stan will classify the starting point as one of a *divergent transition*. If many of such starting points of divergent transitions are concentrated within a region of parameter

space, this may be an indication that the curvature of the posterior is very high in this region and that the step size ϵ is too large to adequately explore this region.

As briefly mentioned in Section 2.2.2, NUTS builds up a binary tree when determining the number L of leapfrog steps to take before a U-turn would occur. Stan records the depth of this tree for each iteration and thus also the corresponding starting point. Moreover, the user can specify a maximum tree depth d to avoid long execution times due too many steps; as at most 2^{d-1} leapfrog steps are taken in each iteration. The default value is $d = 10$. Hitting this maximum means that NUTS is terminated prematurely (i. e. more steps would have been possible before a U-turn) and Stan counts how many times this occurs. Reasons for having to take many steps may be a too small step size due to poor adaptation to a posterior of varying curvature or targeting a very high acceptance rate.

According to Betancourt et al. (2015), the *BFMI* indicates how well the energy sets of the Hamiltonian are explored. Let $E = H(\boldsymbol{\theta}, \boldsymbol{\rho})$ be the total energy, $\pi(E|\rho)$ the energy transition distribution, and $\pi(E)$ the marginal energy distribution. If $\pi(E|\rho)$ is substantially more narrow than $\pi(E)$, then a HMC chain may not be able to completely explore the tails of the target distribution. The BFMI quantifies the mismatch between the two distributions and is defined and approximated by

$$BFMI := \frac{\mathbb{E}_\pi\left[Var_{\pi_{E|\rho}}[E|\rho]\right]}{Var_{\pi_E}[E]} \approx \frac{\sum_{n=1}^{N}(E_n - E_{n-1})^2}{\sum_{n=0}^{N}(E_n - \bar{E})^2} =: \widehat{BFMI}.$$

The Stan development team recommends to ensure that the value of $\widehat{BFMI}$ is greater than 0.2.

A.3.2 Stan specific diagnostics for the sampling output for the translation kinetics models

Here, we summarize the Stan specific diagnostics described in A.3.1 for the HMC output from Sections 5.7 and 5.8. Tables A.1 and A.2 present the statistics of the number of divergent transition, Tables A.3 and A.4 the statistics of the number of times that the user-specified maximal tree depth was exceeded, and Tables A.5 and A.6 that statistics of the BFMI.

Overall, all three diagnostics show poorer values for the sampling output for the SDE model than for the ODE model. This is not surprising as we sample from a much higher-dimensional distribution for the SDE model. We do not consider the poor diagnostics as a disadvantage of the procedure as they provide information that we do not even have for other MCMC algorithms and thus cannot compare to them.

Table A.1: Statistics for the Stan diagnostic of the number of divergent transitions for the SDE model. The 100 sampling outputs per dataset are categorized by the number of divergent transitions that occurred after warm-up, i. e. during a total of 20,000 iterations. Hence, the values in columns 1 to 4 sum to 100. Column 5 gives the maximum number of divergent transitions that occurred after warm-up for one sampling output.

dataset	none	$1-10$	$11-100$	>100	maximum
simulated data without error	37	10	25	28	1644
simulated data with error	88	4	5	3	568
experimental data for eGFP	93	4	3	0	39
experimental data for d2eGFP	90	3	6	1	540

Table A.2: Statistics for the Stan diagnostic of the number of divergent transitions for the ODE model. See Table A.1 for a detailed description.

dataset	none	$1-10$	$11-100$	>100	maximum
simulated data without error	100	0	0	0	0
simulated data with error	100	0	0	0	0
experimental data for eGFP	99	1	0	0	1
experimental data for d2eGFP	92	8	0	0	2

Table A.3: Statistics for the Stan diagnostic of the number of times that the maximal tree depth was exceeded for the SDE model. The user-defined maximal tree depth was set to a value of 15 prior to sampling. The 100 sampling outputs per dataset are categorized by the number of times that the maximal tree depth was exceeded after warm-up, i. e. during a total of 20,000 iterations. Hence, the values in columns 1 to 4 sum to 100. Column 5 gives the maximum number of times that the maximal tree depth was exceeded after warm-up for one sampling output.

dataset	none	$1-10$	$11-100$	>100	maximum
simulated data without error	99	0	1	0	11
simulated data with error	10	26	21	43	7126
experimental data for eGFP	25	19	31	25	1976
experimental data for d2eGFP	95	2	3	0	59

Table A.4: Statistics for the Stan diagnostic of the number of times that the maximal tree depth was exceeded for the ODE model. See Table A.3 for a detailed description.

dataset	none	$1-10$	$11-100$	>100	maximum
simulated data without error	91	6	0	3	2500
simulated data with error	96	0	0	4	2500
experimental data for eGFP	97	0	0	3	2500
experimental data for d2eGFP	100	0	0	0	0

Table A.5: Statistics for the Stan diagnostic $\widehat{BFMI}$ for the SDE model. Each of the 100 sampling outputs per dataset consists of 8 HMC chains for each of which $\widehat{BFMI}$ is calculated. Then, we determine the minimum and the mean over the 8 chains. The table presents the mean and the standard deviation (s.d.) of these minima and means aggregated over the 100 sampling outputs per dataset.

dataset	mean of minima	s.d. of minima	mean of means	s.d. of means
simulated data without error	0.03	0.01	0.05	0.01
simulated data with error	0.05	0.02	0.07	0.01
experimental data for eGFP	0.05	0.04	0.08	0.04
experimental data for d2eGFP	0.07	0.05	0.09	0.05

Table A.6: Statistics for the Stan diagnostic $\widehat{BFMI}$ for the ODE model. See Table A.5 for a detailed description.

dataset	mean of minima	s.d. of minima	mean of means	s.d. of means
simulated data without error	0.95	0.19	1.03	0.06
simulated data with error	0.95	0.15	1.03	0.06
experimental data for eGFP	0.94	0.19	1.03	0.05
experimental data for d2eGFP	0.90	0.23	1.02	0.05

A.3.3 Additional figures of the estimation results

Results for simulated data

Figures A.5, A.6, A.7, and A.8 show the same sampling output (the four posterior samples for the two simulated data sets depicted in Figure 5.6) as Figures 5.7, 5.8, 5.10, and 5.11 in Section 5.7; however here, the results are not compared between the ODE and the SDE model but between simulated data with and without measurement error.

For the SDE, we see in Figure A.5 that the occurrence of measurement error substantially impacts the distribution of the posterior sample with respect to the parameters θ_1 and θ_3. The shape of the two dimensional projection changes from an elliptic shape to a banana-like shape. Especially for θ_3, the 95% CI and the range of values in the posterior sample increase a lot and the true parameter value is only barely covered by the 95% CI for simulated data with measurement error.

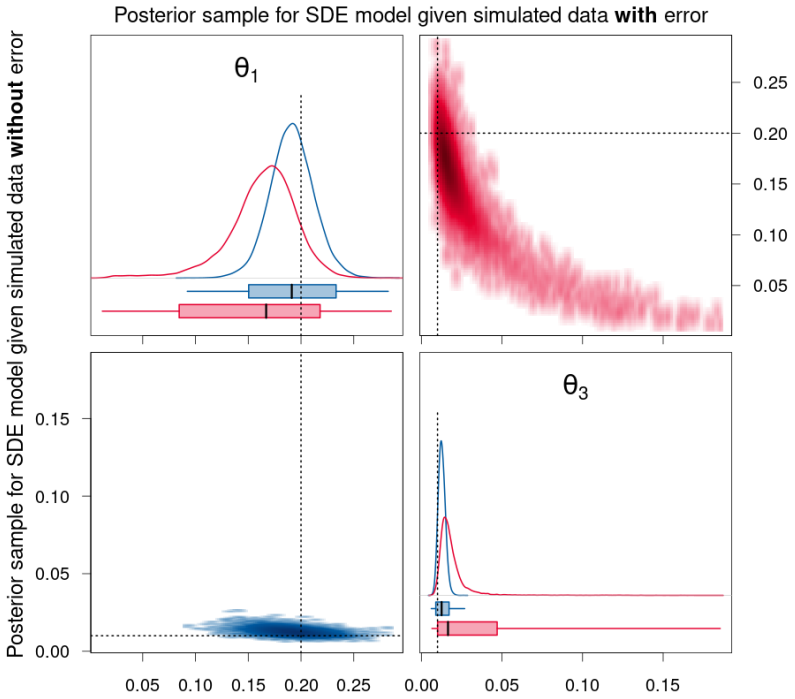


Figure A.5: Density estimates of the posterior samples for parameters θ_1 and θ_3 for the SDE model given simulated data without (blue, lower triangle) and with (red, upper triangle) measurement error. *Diagonal panels:* Marginal densities for the respective parameter and boxplots showing the 95% CI as box, the range of the sample as whiskers, and the median as thick black line. *Off-diagonal panels:* Smoothed scatter plots of the two-dimensional projections of the samples where darker hues signify higher density values. The dotted lines represent the true parameter values that were used to simulate the data.

Similarly for the parameters θ_2, m_0, scale and their products, Figure A.6 shows that there is quite a difference between the distributions of the posterior samples for the simulated data without and with measurement error. In particular for the parameters scale and $\theta_2 m_0$ which we consider to be identifiable, the 95% CIs increase substantially for data with measurement error, and also the appearance of the two-dimensional projections with respect to these two parameters changes a lot, from a slightly bent ellipse to a clear banana shape. For the product $\theta_2 m_0 \mathrm{scale}$, the dispersion of the posterior samples changes only slightly which is apparent from the similar lengths of the 95% CIs in Figure A.6 and also from the similar c.v. in Tables 5.2 and 5.5 (0.083 for data without measurement error and 0.093 for data with measurement error). The location of the sample measured e. g. by the median slightly shifts away from the true parameter value for the data with measurement error; however, the true value is still included in the 95% CIs. Only for parameter m_0 for which we also did not see much difference in the posterior samples for the ODE vs. SDE model, the occurrence of measurement error does not seem to affect the posterior sample much. For the remaining parameters θ_2, $\theta_2 \mathrm{scale}$, and $m_0 \mathrm{scale}$ which we do not consider to be identifiable but for which the 95% CIs of the posterior samples for the SDE model were clearly more narrow than the 95% CIs of the corresponding posterior sample for the ODE model, the 95% CIs and ranges of values of the posterior sample for the SDE model for data with measurement error are broader than for data without measurement error.

For the ODE model, Figures A.7 and A.8 show that there is hardly any difference for most of the parameters between the posterior sample for the data without and with measurement error since the majority of the parameters are not identifiable anyway. For the parameters offset and t_0, there is a slight difference. For the measurement error parameter σ, the posterior sample consists of higher values for data with measurement error as expected. Note that for both simulated datasets, the range of the posterior sample does not include the true parameter value for σ. Finally for the product $\theta_2 m_0 \mathrm{scale}$, the dispersion of the posterior sample increases only slightly for data with measurement error and the location of the sample shifts away from the true parameter value. Also for this parameter, the range of the posterior sample does not include the true parameter value for both simulated datasets.

Figure A.9 shows the statistics of the posterior samples for the simulated data without and with measurement error aggregated over 100 simulated trajectories. It visualizes the last two columns of Tables 5.3 and 5.6 and compares the results of the posterior samples for the simulated data without to those with measurement error separately for the SDE and the ODE model within each plot, instead of comparing the two model types separately for each kind of data as in Figures 5.9 and 5.12.

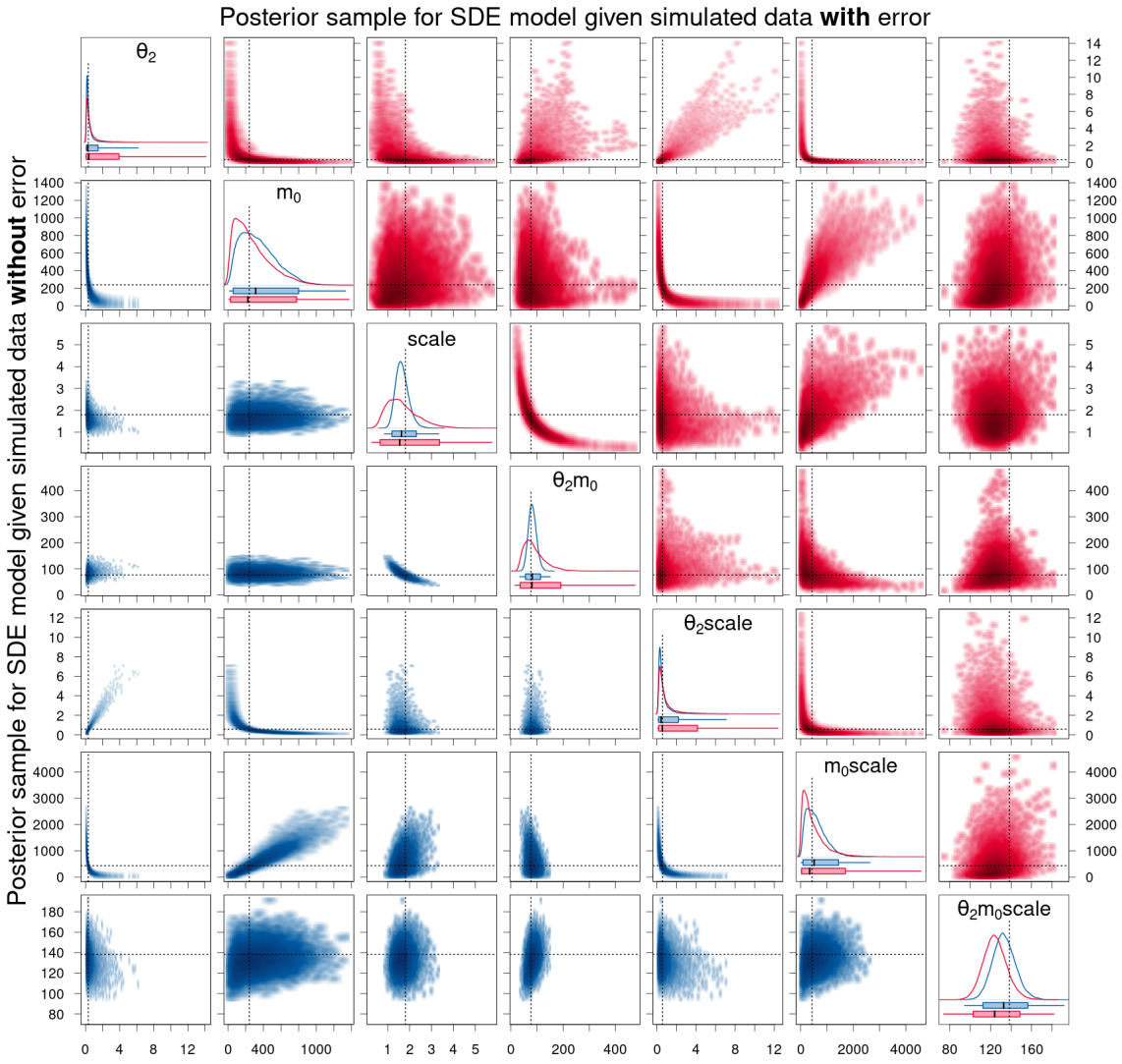


Figure A.6: Density estimates of the posterior samples for parameters θ_2, m_0, scale, and their products for the SDE model given simulated data without (blue, lower triangle) and with (red, upper triangle) measurement error. *Diagonal panels:* Marginal densities for the respective parameter and boxplots showing the 95% CI as box, the range of the sample as whiskers, and the median as thick black line. *Off-diagonal panels:* Smoothed scatter plots of the two-dimensional projections of the samples where darker hues signify higher density values. The dotted lines represent the true parameter values that were used to simulate the data.

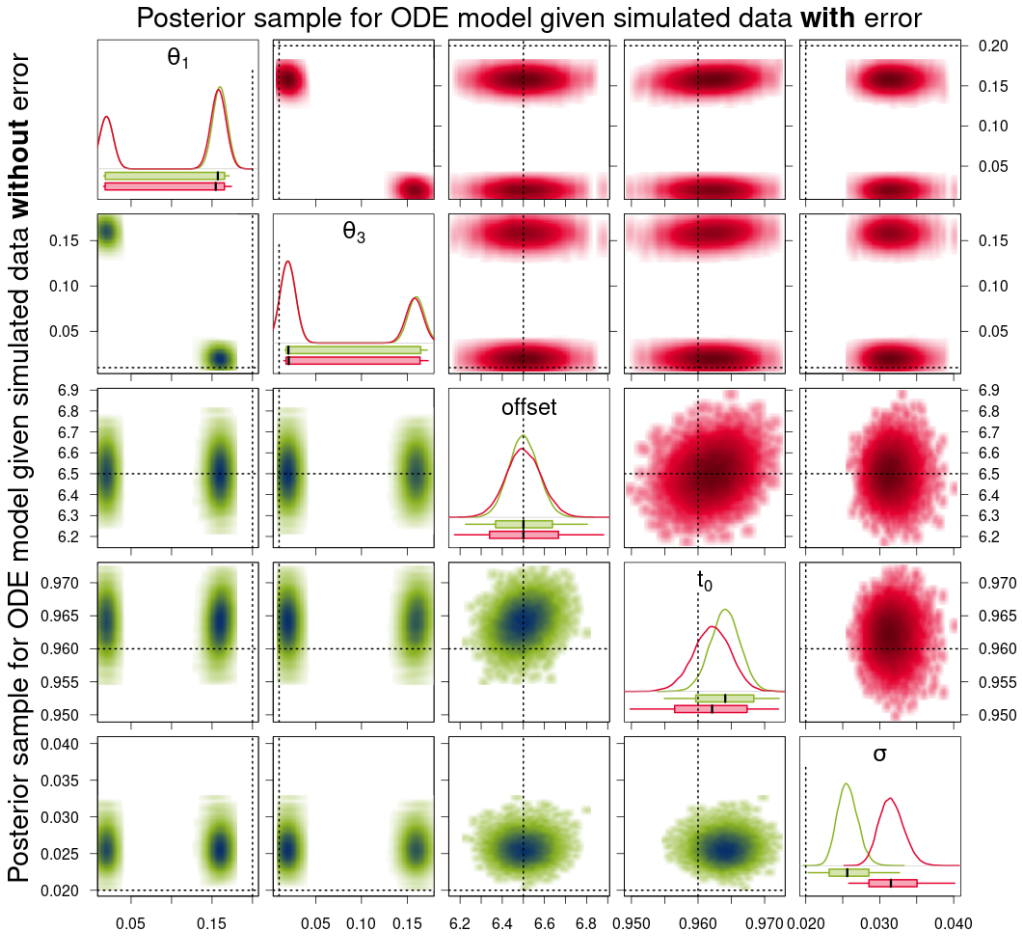


Figure A.7: Density estimates of the posterior samples for parameters θ_1 and θ_3 for the ODE model given simulated data without (green, lower triangle) and with (red, upper triangle) measurement error. *Diagonal panels:* Marginal densities for the respective parameter and boxplots showing the 95% CI as box, the range of the sample as whiskers, and the median as thick black line. *Off-diagonal panels:* Smoothed scatter plots of the two-dimensional projections of the samples where darker hues signify higher density values. The dotted lines represent the true parameter values that were used to simulate the data. For the parameter σ, the dotted line only represents the true value for the data with measurement error. For the data without measurement error, σ is equal to 0.

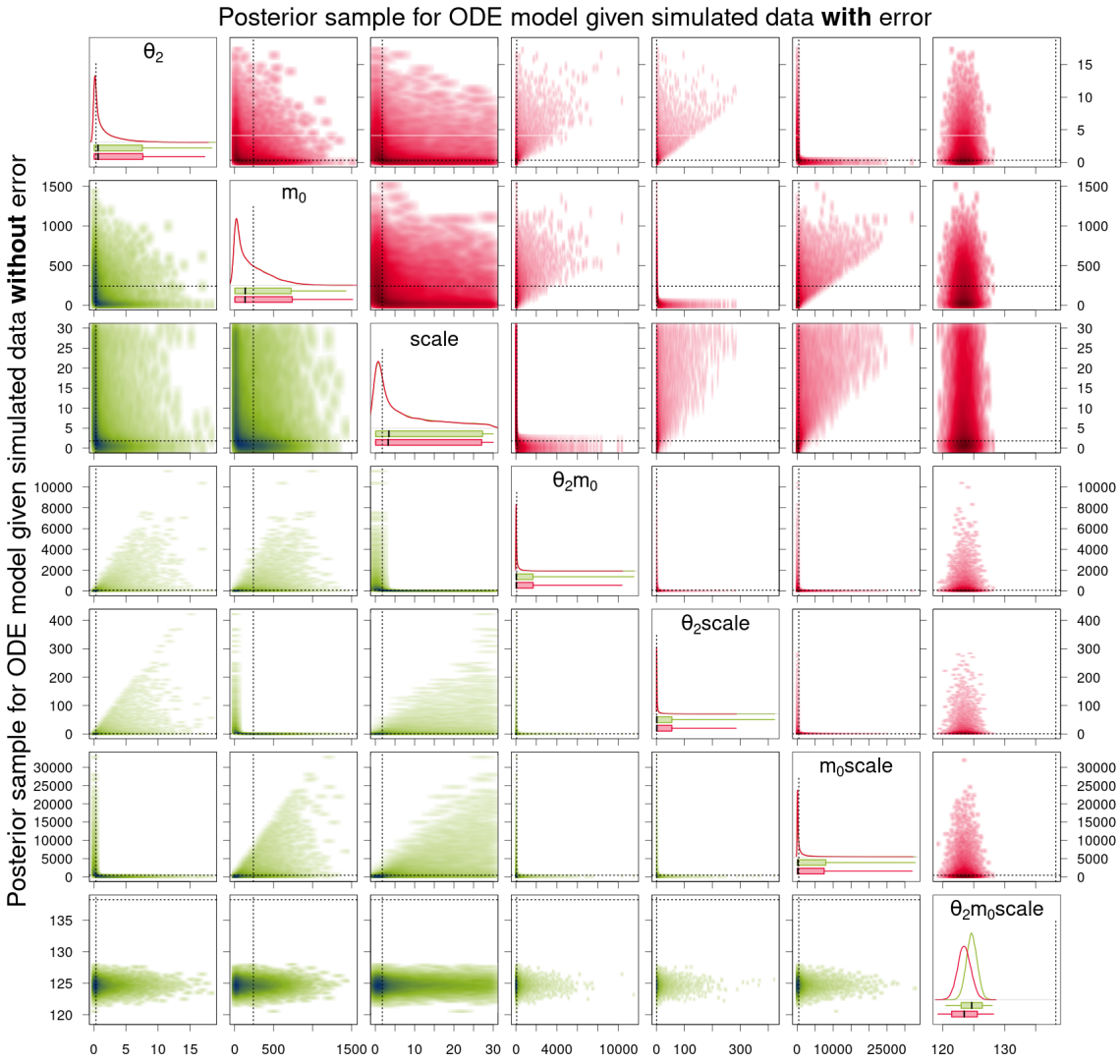
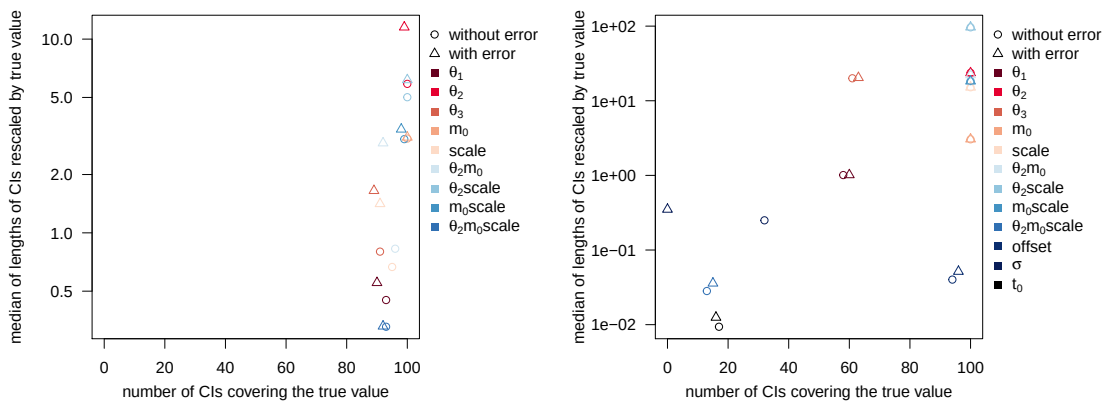


Figure A.8: Density estimates of the posterior samples for parameters θ_2, m_0, scale, and their products for the ODE model given simulated data without (green, lower triangle) and with (red, upper triangle) measurement error. *Diagonal panels:* Marginal densities for the respective parameter and boxplots showing the 95% CI as box, the range of the sample as whiskers, and the median as thick black line. *Off-diagonal panels:* Smoothed scatter plots of the two-dimensional projections of the samples where darker hues signify higher density values. The dotted lines represent the true parameter values that were used to simulate the data.



(a) SDE results **(b)** ODE results

Figure A.9: Statistics of posterior samples for the simulated data without and with measurement error aggregated over 100 simulated trajectories. The desirable region of value combinations is in the bottom right corner of each graph.

Bibliography

Aït-Sahalia, Y. (2002). Maximum likelihood estimation of discretely sampled diffusions: A closed-form approximation approach. *Econometrica*, 70(1), 223–262.

Aït-Sahalia, Y. (2008). Closed-form likelihood expansions for multivariate diffusions. *The Annals of Statistics*, 36(2), 906–937.

Alfonsi, A. (2015). *Affine Diffusions and Related Processes: Simulation, Theory and Applications*. Bocconi & Springer Series. Springer International Publishing.

Anderson, D. F. (2007). A modified next reaction method for simulating chemical systems with time dependent propensities and delays. *The Journal of Chemical Physics*, 127(21), 214107.

Andrieu, C., Doucet, A., & Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3), 269–342.

Andrieu, C. & Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *Annals of Statistics*, 37(2), 697–725.

Baker, R. E., Peña, J.-M., Jayamohan, J., & Jérusalem, A. (2018). Mechanistic models versus machine learning, a fight worth fighting for the biological community? *Biology Letters*, 14(5), 20170660.

Ballnus, B., Hug, S., Hatz, K., Görlitz, L., Hasenauer, J., & Theis, F. J. (2017). Comprehensive benchmarking of Markov chain Monte Carlo methods for dynamical systems. *BMC Systems Biology*, 11(1), 63.

Bayram, M., Partal, T., & Buyukoz, G. O. (2018). Numerical methods for simulation of stochastic differential equations. *Advances in Difference Equations*, 2018(1), 17.

Beaumont, M. A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164(3), 1139–1160.

Bellu, G., Saccomani, M. P., Audoly, S., & D'Angiò, L. (2007). DAISY: A new software tool to test global identifiability of biological and physiological systems. *Computer Methods and Programs in Biomedicine*, 88(1), 52 – 61.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics. New York, NY: Springer, 2nd edition.

Beskos, A., Papaspiliopoulos, O., & Roberts, G. (2009). Monte Carlo maximum likelihood estimation for discretely observed diffusion processes. *Annals of Statistics*, 37(1), 223–245.

Beskos, A., Papaspiliopoulos, O., & Roberts, G. O. (2006a). Retrospective Exact Simulation of Diffusion Sample Paths with Applications. *Bernoulli*, 12(6), 1077–1098.

Beskos, A., Papaspiliopoulos, O., & Roberts, G. O. (2008). A Factorisation of Diffusion Measure and Finite Sample Path Constructions. *Methodology and Computing in Applied Probability*, 10(1), 85–104.

Beskos, A., Papaspiliopoulos, O., Roberts, G. O., & Fearnhead, P. (2006b). Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3), 333–382.

Beskos, A. & Roberts, G. O. (2005). Exact simulation of diffusions. *Annals of Applied Probability*, 15(4), 2422–2444.

Betancourt, M. (2018). A conceptual introduction to Hamiltonian Monte Carlo. *Preprint*, arXiv:1701.02434.

Betancourt, M. (2019). Incomplete reparameterizations and equivalent metrics. *Preprint*, arXiv:1910.09407.

Betancourt, M. J., Byrne, S., & Girolami, M. (2015). Optimizing the integrator step size for Hamiltonian Monte Carlo. *Preprint*, arXiv:1411.6669.

Black, F. & Scholes, M. (1973). The Pricing of Options and Corporate Liabilities. *Journal of Political Economy*, 81(3), 637–54.

Braumann, C. A. (2019). *Introduction to Stochastic Differential Equations with Applications to Modelling in Biology and Finance*. John Wiley & Sons, Ltd.

Brooks, S., Gelman, A., Jones, G., & Meng, X. (2011). *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press.

Brooks, S. P. & Roberts, G. O. (1998). Convergence assessment techniques for Markov chain Monte Carlo. *Statistics and Computing*, 8(4), 319–335.

Brouwer, A. F., Meza, R., & Eisenberg, M. C. (2017). Parameter estimation for multistage clonal expansion models from cancer incidence data: A practical identifiability analysis. *PLOS Computational Biology*, 13(3), 1–18.

Browning, A. P., Warne, D. J., Burrage, K., Baker, R. E., & Simpson, M. J. (2020). Identifiability analysis for stochastic differential equation models in systems biology. *Journal of The Royal Society Interface*, 17(173), 20200652.

Cao, Y., Gillespie, D. T., & Petzold, L. R. (2005). The slow-scale stochastic simulation algorithm. *The Journal of Chemical Physics*, 122(1), 014116.

Cao, Y., Gillespie, D. T., & Petzold, L. R. (2006). Efficient step size selection for the tau-leaping simulation method. *The Journal of Chemical Physics*, 124(4), 044109.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, Articles*, 76(1), 1–32.

Carpenter, B., Hoffman, M. D., Brubaker, M., Lee, D., Li, P., & Betancourt, M. (2015). The Stan math library: Reverse-mode automatic differentiation in C++. *Preprint*, arXiv:1509.07164.

Chib, S. & Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4), 327–335.

Chib, S., Pitt, M. K., & Shephard, N. (2004). *Likelihood based inference for diffusion driven models*. Working paper, Nuffield College, University of Oxford.

Chib, S. & Shephard, N. (2002). [Numerical Techniques for Maximum Likelihood Estimation of Continuous-Time Diffusion Processes]: Comment. *Journal of Business & Economic Statistics*, 20(3), 325–327.

Chiș, O., Banga, J., & Balsa-Canto, E. (2011a). Structural identifiability of systems biology models: A critical comparison of methods. *PLoS ONE*, 6.

Chiș, O., Banga, J. R., & Balsa-Canto, E. (2011b). GenSSI: a software toolbox for structural identifiability analysis of biological models. *Bioinformatics*, 27(18), 2610–2611.

Cox, J. C., Ingersoll, J. E., & Ross, S. A. (1985). A Theory of the Term Structure of Interest Rates. *Econometrica*, 53(2), 385–407.

Dacunha-Castelle, D. & Florens-Zmirou, D. (1986). Estimation of the coefficients of a diffusion from discrete observations. *Stochastics*, 19(4), 263–284.

DeFrancesco, L. (2020). Whither COVID-19 vaccines? *Nature Biotechnology*, 38(10), 1132–1145.

Donnet, S. & Samson, A. (2013). A review on estimation of stochastic differential equations for pharmacokinetic/pharmacodynamic models. *Advanced Drug Delivery Reviews*, 65(7), 929–939.

Duane, S., Kennedy, A., Pendleton, B. J., & Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B*, 195(2), 216 – 222.

Durham, G. B. & Gallant, A. R. (2002). Numerical Techniques for Maximum Likelihood Estimation of Continuous-Time Diffusion Processes. *Journal of Business & Economic Statistics*, 20(3), 297–316.

Elerian, O. (1998). *A note on the existence of a closed form conditional transition density for the Milstein scheme*. Working paper, Nuffield College, University of Oxford.

Elerian, O., Chib, S., & Shephard, N. (2001). Likelihood Inference for Discretely Observed Nonlinear Diffusions. *Econometrica*, 69(4), 959–993.

Elowitz, M. B., Levine, A. J., Siggia, E. D., & Swain, P. S. (2002). Stochastic gene expression in a single cell. *Science*, 297(5584), 1183–1186.

Eraker, B. (2001). MCMC analysis of diffusion models with application to finance. *Journal of Business & Economic Statistics*, 19(2), 177–191.

Etchegaray, C. & Meunier, N. (2019). A stochastic model for cell adhesion to the vascular wall. *Journal of Mathematical Biology*, 79(5), 1665–1697.

Feller, W. (1951a). Diffusion Processes in Genetics. In J. Neyman (Ed.), *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability* (pp. 227–246). University of California Press, Berkeley.

Feller, W. (1951b). Two singular diffusion problems. *Annals of Mathematics*, 54(1), 173–182.

Filipović, D., Mayerhofer, E., & Schneider, P. (2013). Density approximations for multivariate affine jump-diffusion processes. *Journal of Econometrics*, 176(2), 93 – 111.

Finkenstädt, B., Heron, E. A., Komorowski, M., Edwards, K., Tang, S., Harper, C. V., Davis, J. R. E., White, M. R. H., Millar, A. J., & Rand, D. A. (2008). Reconstruction of transcriptional dynamics from gene reporter data using differential equations. *Bioinformatics*, 24(24), 2901–2907.

Flegal, J. M., Hughes, J., Vats, D., & Dai, N. (2020). *mcmcse: Monte Carlo Standard Errors for MCMC*. R package version 1.4-1.

Florens-Zmirou, D. (1989). Approximate discrete-time schemes for statistics of diffusion processes. *Statistics*, 20(4), 547–557.

Fouskakis, D. & Draper, D. (2002). Stochastic optimization: a review. *International Statistical Review*, 70(3), 315–349.

Fröhlich, F., Reiser, A., Fink, L., Woschée, D., Ligon, T., Theis, F. J., Rädler, J. O., & Hasenauer, J. (2018). Multi-experiment nonlinear mixed effect modeling of single-cell translation kinetics after transfection. *npj Systems Biology and Applications*, 4(1), 42.

Fröhlich, F., Kaltenbacher, B., Theis, F. J., & Hasenauer, J. (2017). Scalable parameter estimation for genome-scale biochemical reaction networks. *PLOS Computational Biology*, 13(1), 1–18.

Fröhlich, F., Kessler, T., Weindl, D., Shadrin, A., Schmiester, L., Hache, H., Muradyan, A., Schütte, M., Lim, J.-H., Heinig, M., Theis, F. J., Lehrach, H., Wierling, C., Lange, B., & Hasenauer, J. (2018). Efficient parameter estimation enables the prediction of drug response using a mechanistic pan-cancer pathway model. *Cell Systems*, 7(6), 567 – 579.e6.

Fuchs, C. (2013). *Inference for Diffusion Processes*. Berlin Heidelberg: Springer.

Gallant, A. R. & Tauchen, G. (1996). Which Moments to Match? *Econometric Theory*, 12(4), 657–681.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC.

Geman, S. & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6), 721–741.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo. *Statistical Science*, 7(4), 473–483.

Gibson, M. A. & Bruck, J. (2000). Efficient exact stochastic simulation of chemical systems with many species and many channels. *The Journal of Physical Chemistry A*, 104(9), 1876–1889.

Gillespie, D. T. (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4), 403 – 434.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25), 2340–2361.

Gillespie, D. T. (1992a). *Markov Processes: An Introduction for Physical Scientists*. Boston: Academic Press.

Gillespie, D. T. (1992b). A rigorous derivation of the chemical master equation. *Physica A: Statistical Mechanics and its Applications*, 188(1), 404 – 425.

Gillespie, D. T. (2000). The chemical Langevin equation. *The Journal of Chemical Physics*, 113(1), 297–306.

Gillespie, D. T. (2001). Approximate accelerated stochastic simulation of chemically reacting systems. *The Journal of Chemical Physics*, 115(4), 1716–1733.

Girolami, M. & Calderhead, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2), 123–214.

Glasserman, P. (2003). *Monte Carlo Methods in Financial Engineering*. Stochastic Modelling and Applied Probability. New York: Springer.

Golightly, A. & Wilkinson, D. J. (2006a). Bayesian sequential inference for nonlinear multivariate diffusions. *Statistics and Computing*, 16(4), 323–338.

Golightly, A. & Wilkinson, D. J. (2006b). Bayesian sequential inference for stochastic kinetic biochemical network models. *Journal of Computational Biology*, 13(3), 838–851.

Golightly, A. & Wilkinson, D. J. (2008). Bayesian inference for nonlinear multivariate diffusion models observed with error. *Computational Statistics & Data Analysis*, 52(3), 1674–1693.

Golightly, A. & Wilkinson, D. J. (2011). Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo. *Interface focus*, 1(6), 807–820.

Gourieroux, C., Monfort, A., & Renault, E. (1993). Indirect Inference. *Journal of Applied Econometrics*, 8, S85–S118.

Gyöngy, I. & Rásonyi, M. (2011). A note on Euler approximations for SDEs with Hölder continuous diffusion coefficients. *Stochastic Processes and their Applications*, 121(10), 2189–2200.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1), 97–109.

Hefter, M. & Herzwurm, A. (2018). Strong convergence rates for Cox–Ingersoll–Ross processes — Full parameter range. *Journal of Mathematical Analysis and Applications*, 459(2), 1079–1101.

Held, L. & Sabanés Bové, D. (2020). *Likelihood and Bayesian Inference: With Applications in Biology and Medicine*. Statistics for Biology and Health. Berlin, Heidelberg: Springer, 2nd edition.

Hengl, S., Kreutz, C., Timmer, J., & Maiwald, T. (2007). Data-based identifiability analysis of non-linear dynamical models. *Bioinformatics*, 23(19), 2612–2618.

Heron, E. A., Finkenstädt, B., & Rand, D. A. (2007). Bayesian inference for dynamic transcriptional regulation; the Hes1 system as a case study. *Bioinformatics*, 23(19), 2596–2603.

Heston, S. L. (1993). A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options. *The Review of Financial Studies*, 6(2), 327–343.

Hines, K. E., Middendorf, T. R., & Aldrich, R. W. (2014). Determination of parameter identifiability in nonlinear biophysical models: A Bayesian approach. *Journal of General Physiology*, 143(3), 401–416.

Hoffman, M. D. & Gelman, A. (2014). The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(47), 1593–1623.

Humphreys, P. (2003). *Mathematical Modeling in the Social Sciences*, chapter 7, (pp. 166–184). John Wiley & Sons, Ltd.

Hurn, A. S., Jeisman, J. I., & Lindsay, K. A. (2007). Seeing the Wood for the Trees: A Critical Evaluation of Methods to Estimate the Parameters of Stochastic Differential Equations. *Journal of Financial Econometrics*, 5(3), 390–455.

Iacus, S. (2008). *Simulation and Inference for Stochastic Differential Equations*. New York: Springer-Verlag.

Ikeda, N. & Watanabe, W. (1981). *Stochastic Differential Equations and Diffusion Processes*. John Wiley & Sons, Ltd.

Ionides, E. L., Bretó, C., & King, A. A. (2006). Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 103(49), 18438–18443.

Jahnke, T. & Huisinga, W. (2007). Solving the chemical master equation for monomolecular reaction systems analytically. *Journal of Mathematical Biology*, 54, 1–26.

Janzén, D. L. I., Bergenholm, L., Jirstrand, M., Parkinson, J., Yates, J., Evans, N. D., & Chappell, M. J. (2016). Parameter identifiability of fundamental pharmacodynamic models. *Frontiers in Physiology*, 7, 590.

Jeanblanc, M., Chesney, M., & Yor, M. (2009). *Mathematical methods for financial markets*. Springer Finance SpringerLink. Dordrecht: Springer.

Jensen, B. & Poulsen, R. (2002). Transition Densities of Diffusion Processes: Numerical Comparison of Approximation Techniques. *The Journal of Derivatives*, 9(4), 18–32.

Kapfer, E.-M., Stapor, P., & Hasenauer, J. (2019). Challenges in the calibration of large-scale ordinary differential equation models. *IFAC-PapersOnLine*, 52(26), 58 – 64. 8th Conference on Foundations of Systems Biology in Engineering FOSBE 2019.

Karatzas, I. & Shreve, S. E. (1998). *Brownian Motion and Stochastic Calculus*. New York, NY: Springer.

Kazeroonian, A., Fröhlich, F., Raue, A., Theis, F. J., & Hasenauer, J. (2016). CERENA: ChEmical REaction Network Analyzer—A toolbox for the simulation and analysis of stochastic chemical kinetics. *PLOS ONE*, 11(1), 1–15.

Kessler, M., Lindner, A., & Sørensen, M., Eds. (2012). *Statistical Methods for Stochastic Differential Equations*, volume 124 of *Monographs on statistics and applied probability*. Boca Raton: CRC Press.

Kirk, P., Silk, D., & Stumpf, M. P. (2016). Reverse engineering under uncertainty. In L. Geris & D. Gomez-Cavrero (Eds.), *Uncertainty in Biology, A computational modeling approach*. Springer International Publishing.

Kitano, H. (2002a). Computational systems biology. *Nature*, 420(6912), 206–210.

Kitano, H. (2002b). Systems biology: A brief overview. *Science*, 295(5560), 1662–1664.

Klebaner, F. C. (2005). *Introduction to Stochastic Calculus with Applications*. London: Imperial College Press, 2nd edition.

Kloeden, P. & Neuenkirch, A. (2007). The pathwise convergence of approximation schemes for stochastic differential equations. *LMS Journal of Computation and Mathematics*, 10, 235–253.

Kloeden, P. E. & Platen, E. (1992). *Numerical Solution of Stochastic Differential Equations*. Berlin Heidelberg: Springer.

Komorowski, M., Costa, M. J., Rand, D. A., & Stumpf, M. P. H. (2011). Sensitivity, robustness, and identifiability in stochastic chemical kinetics models. *Proceedings of the National Academy of Sciences*, 108(21), 8645–8650.

Komorowski, M., Finkenstädt, B., Harper, C. V., & Rand, D. A. (2009). Bayesian inference of biochemical kinetic parameters using the linear noise approximation. *BMC Bioinformatics*, 10(1), 343.

Krylov, N. V. (1980). *Controlled Diffusion Processes*. Stochastic Modelling and Applied Probability. Berlin Heidelberg: Springer-Verlag.

Lamberton, D. & Lapeyre, B. (1996). *Introduction to stochastic calculus applied to finance*. London: Chapman & Hall, 1st edition.

Le Breton, A. (1977). Parameter Estimation in a Linear Stochastic Differential Equation. In J. Kožešnik (Ed.), *Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions, Random Processes and of the 1974 European Meeting of Statisticians: held at Prague, from August 18 to 23, 1974*, Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions, Random Processes and of the 1974 European Meeting of Statisticians (pp. 353–366). Dordrecht: Springer Netherlands.

Leander, J., Almquist, J., Ahlström, C., Gabrielsson, J., & Jirstrand, M. (2015). Mixed effects modeling using stochastic differential equations: Illustrated by pharmacokinetic data of nicotinic acid in obese zucker rats. *The AAPS Journal*, 17(3), 586–596.

Lee, P. M. (2012). *Bayesian Statistics: An Introduction*. Wiley Publishing, 4th edition.

Leonhardt, C., Schwake, G., Stögbauer, T. R., Rappl, S., Kuhr, J.-T., Ligon, T. S., & Rädler, J. O. (2014). Single-cell mRNA transfection studies: delivery, kinetics and statistics by numbers. *Nanomedicine*, 10(4), 679–688.

Ligon, T. S., Fröhlich, F., Chiș, O. T., Banga, J. R., Balsa-Canto, E., & Hasenauer, J. (2017). GenSSI 2.0: multi-experiment structural identifiability analysis of SBML models. *Bioinformatics*, 34(8), 1421–1423.

Ligon, T. S., Leonhardt, C., & Rädler, J. O. (2014). Multi-level kinetic model of mRNA delivery via transfection of lipoplexes. *PLoS One*, 9(9), e107148.

Link, W. A. & Eaton, M. J. (2012). On thinning of chains in MCMC. *Methods in Ecology and Evolution*, 3(1), 112–115.

Lo, A. W. (1988). Maximum Likelihood Estimation of Generalized Itô Processes with Discretely Sampled Data. *Econometric Theory*, 4(2), 231–247.

Loskot, P., Atitey, K., & Mihaylova, L. (2019). Comprehensive review of models and methods for inferences in bio-chemical reaction networks. *Frontiers in Genetics*, 10, 549.

Lux, T. (2013). Inference for systems of stochastic differential equations from discretely sampled data: a numerical maximum likelihood approach. *Annals of Finance*, 9(2), 217–248.

Maceachern, S. N. & Berliner, L. M. (1994). Subsampling the Gibbs sampler. *The American Statistician*, 48(3), 188–190.

Marchetti, L., Priami, C., & Thanh, V. H. (2016). HRSSA – efficient hybrid stochastic simulation for spatially homogeneous biochemical reaction networks. *Journal of Computational Physics*, 317, 301 – 317.

Maruyama, G. (1955). Continuous Markov processes and stochastic equations. *Rendiconti del Circolo Matematico di Palermo*, 4(1), 48.

McElreath, R. (2016). *Statistical rethinking: a Bayesian course with examples in R and Stan*. Texts in statistical science series. Boca Raton: CRC Press.

Merton, R. C. (1973). Theory of Rational Option Pricing. *The Bell Journal of Economics and Management Science*, 4(1), 141–183.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., & Teller, E. (1953). Equations of State Calculations by Fast Computing Machines. *Journal of Chemical Physics*, 21, 1087–1092.

Mrázek, M. & Pospíšil, J. (2017). Calibration and simulation of Heston model. *Open Mathematics*, 15(1), 679–704.

Munsky, B., Trinh, B., & Khammash, M. (2009). Listening to the noise: random fluctuations reveal gene network parameters. *Molecular systems biology*, 5, 318–318.

Müller, J. & Kuttler, C. (2015). *Methods and models in mathematical biology*. Lecture notes on mathematical modelling in the life sciences. Berlin Heidelberg: Springer.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. Jones, & X. Meng (Eds.), *Handbook of Markov Chain Monte Carlo*. CRC Press.

Neimark, J. I. (2003). *Mathematical Models in Natural Science and Engineering*. Berlin Heidelberg: Springer.

Nesterov, Y. (2009). Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1), 221–259.

Ngo, H.-L. & Raguchi, D. (2016). Strong rate of convergence for the Euler-Maruyama approximation of stochastic differential equations with irregular coefficients. *Mathematics and Computers in Simulation*, 85(300), 1793–1819.

Øksendal, B. K. (2003). *Stochastic Differential Equations: An Introduction with Applications*. Universitext. Berlin: Springer, 6th edition.

Opper, M. (2019). Variational inference for stochastic differential equations. *Annalen der Physik*, 531(3), 1800233.

Owen, A. B. (2017). Statistically efficient thinning of a Markov chain sampler. *Journal of Computational and Graphical Statistics*, 26(3), 738–744.

Pachpatte, B. G. (1998). *Inequalities for differential and integral equations*. Number 197 in Mathematics in science and engineering. San Diego: Academic Press.

Pedersen, A. R. (1995). A New Approach to Maximum Likelihood Estimation for Stochastic Differential Equations Based on Discrete Observations. *Scandinavian Journal of Statistics*, 22(1), 55–71.

Pieschner, S. & Fuchs, C. (2020). Bayesian inference for diffusion processes: using higher-order approximations for transition densities. *Royal Society Open Science*, 7(10), 200270.

Poulson, R. (1999). *Approximate maximum likelihood estimation of discretely observed diffusion processes*. Working paper, Centre for Analytical Finance, Aarhus.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raj, A. & van Oudenaarden, A. (2008). Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*, 135(2), 216–226.

Raue, A., Becker, V., Klingmüller, U., & Timmer, J. (2010). Identifiability and observability analysis for experimental design in nonlinear dynamical models. *Chaos*, 20(4), 045105.

Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U., & Timmer, J. (2009). Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15), 1923–1929.

Raue, A., Kreutz, C., Theis, F. J., & Timmer, J. (2013). Joining forces of Bayesian and frequentist methodology: a study for inference in the presence of non-identifiability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1984), 20110544.

Raue, A., Steiert, B., Schelker, M., Kreutz, C., Maiwald, T., Hass, H., Vanlier, J., Tönsing, C., Adlung, L., Engesser, R., Mader, W., Heinemann, T., Hasenauer, J., Schilling, M., Höfer, T., Klipp, E., Theis, F., Klingmüller, U., Schöberl, B., & Timmer, J. (2015). Data2dynamics: a modeling environment tailored to parameter estimation in dynamical systems. *Bioinformatics*, 31(21), 3558–3560.

Regan, H. M., Colyvan, M., & Burgman, M. A. (2002). A taxonomy and treatment of uncertainty for ecology and conservation biology. *Ecological Applications*, 12(2), 618–628.

Reiersøl, O. (1950). Identifiability of a linear relation between variables which are subject to error. *Econometrica*, 18(4), 375–389.

Reiser, A., Woschée, D., Mehrotra, N., Krzysztoń, R., Strey, H. H., & Rädler, J. O. (2019). Correlation of mRNA delivery timing and protein expression in lipid-based transfection. *Integrative Biology*, 11(9), 362–371.

Robert, C. P. & Casella, G. (2002). *Monte Carlo statistical methods*. Springer texts in statistics. New York, NY: Springer, corr. 3rd edition.

Roberts, G. O. & Rosenthal, J. S. (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16(4), 351–367.

Roberts, G. O. & Stramer, O. (2001). On inference for partially observed nonlinear diffusion models using the Metropolis-Hastings algorithm. *Biometrika*, 88(3), 603–621.

Rothenberg, T. J. (1971). Identification in parametric models. *Econometrica*, 39(3), 577–591.

Ryder, T., Golightly, A., McGough, A. S., & Prangle, D. (2018). Black-box variational inference for stochastic differential equations. *Proceedings of the 35th International Conference on Machine Learning Research*, 80, 4423–4432.

Sahin, U., Karikó, K., & Türeci, Ö. (2014). mRNA-based therapeutics – developing a new class of drugs. *Nature Reviews Drug Discovery*, 13(10), 759–780.

Santa-Clara, P. (1995). *Simulated Likelihood Estimation of Diffusions With an Application to the Short-Term Interest R*. Working paper, Department of Economics, Havard Univesity.

Schmidt, K. D. (2009). *Maß und Wahrscheinlichkeit [Measure and Probability]*. Berlin Heidelberg: Springer.

Schmiester, L., Schälte, Y., Fröhlich, F., Hasenauer, J., & Weindl, D. (2019). Efficient parameterization of large-scale dynamic models based on relative measurements. *Bioinformatics*, 36(2), 594–602.

Schnoerr, D., Sanguinetti, G., & Grima, R. (2017). Approximation and inference methods for stochastic biochemical kinetics—a tutorial review. *Journal of Physics A: Mathematical and Theoretical*, 50(9), 093001.

Sermaidis, G., Papaspiliopoulos, O., Roberts, G. O., Beskos, A., & Fearnhead, P. (2013). Markov Chain Monte Carlo for Exact Inference for Diffusions. *Scandinavian Journal of Statistics*, 40(2), 294–321.

Shoji, I. & Ozaki, T. (1998a). A statistical method of estimation and simulation for systems of stochastic differential equations. *Biometrika*, 85(1), 240–243.

Shoji, I. & Ozaki, T. (1998b). Estimation for nonlinear stochastic differential equations by a local linearization method. *Stochastic Analysis and Applications*, 16(4), 733–752.

Simpson, M. J., Baker, R. E., Vittadello, S. T., & Maclaren, O. J. (2020). Practical parameter identifiability for spatio-temporal models of cell invasion. *Journal of The Royal Society Interface*, 17(164), 20200055.

Sørensen, H. (2004). Parametric inference for diffusion processes observed at discrete points in time: a survey. *International Statistical Review*, 72(3), 337–354.

Stan Development Team (2019). RStan: the R interface to Stan. R package version 2.19.1, `http://mc-stan.org/`.

Stapor, P., Weindl, D., Ballnus, B., Hug, S., Loos, C., Fiedler, A., Krause, S., Hroß, S., Fröhlich, F., & Hasenauer, J. (2018). PESTO: Parameter EStimation TOolbox. *Bioinformatics*, 34(4), 705–707.

Tanner, M. A. & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398), 528–540.

Terje Lines, G., Łukasz Paszkowski, Schmiester, L., Weindl, D., Stapor, P., & Hasenauer, J. (2019). Efficient computation of steady states in large-scale ode models of biochemical reaction networks. *IFAC-PapersOnLine*, 52(26), 32 – 37. 8th Conference on Foundations of Systems Biology in Engineering FOSBE 2019.

Tian, T. & Burrage, K. (2004). Binomial leap methods for simulating stochastic chemical kinetics. *The Journal of Chemical Physics*, 121(21), 10356–10364.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., & Stumpf, M. P. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of The Royal Society Interface*, 6(31), 187–202.

Tse, Y. K., Zhang, X., & Yu, J. (2004). Estimation of hyperbolic diffusion using the Markov chain Monte Carlo method. *Quantitative Finance*, 4(2), 158–169.

van der Meulen, F. & Schauer, M. (2017). Bayesian estimation of discretely observed multi-dimensional diffusion processes using guided proposals. *Electronic Journal of Statistics*, 11(1), 2358–2396.

Vats, D., Flegal, J. M., & Jones, G. L. (2019). Multivariate output analysis for Markov chain Monte Carlo. *Biometrika*, 106(2), 321–337.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2020). Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC. *Bayesian Analysis*. To appear.

Waage, P. & Gulberg, C. M. (1986). Studies concerning affinity. *Journal of Chemical Education*, 63(12), 1044.

Wallace, E. W. J., Gillespie, D. T., Sanft, K. R., & Petzold, L. R. (2012). Linear noise approximation is valid over limited times for any chemical system that is sufficiently large. *IET Systems Biology*, 6(4), 102–115.

Warne, D. J., Baker, R. E., & Simpson, M. J. (2019). Simulation and inference algorithms for stochastic biochemical reaction networks: from basic concepts to state-of-the-art. *Journal of The Royal Society Interface*, 16(151), 20180943.

Whitaker, G. A., Golightly, A., Boys, R. J., & Sherlock, C. (2017). Improved bridge constructs for stochastic differential equations. *Statistics and Computing*, 27(4), 885–900.

Wilkinson, D. J. (2019). *Stochastic modelling for systems biology*. Mathematical & computational biology. Boca Raton, Fla.: Taylor & Francis, 3rd edition.

Wilkinson, D. J. & Golightly, A. (2010). Markov chain Monte Carlo algorithms for SDE parameter estimation. In N. Lawrence, M. Girolami, M. Rattray, & G. Sanguinetti (Eds.), *Learning and inference in computational systems biology*. Cambridge, MA: MIT Press.

Yang, H., Wu, F., & Kloeden, P. E. (2019). Existence and approximation of strong solutions of SDEs with fractional diffusion coefficients. *Discrete & Continuous Dynamical Systems - B*, 24(10).

Young, J. W., Locke, J. C. W., Altinok, A., Rosenfeld, N., Bacarian, T., Swain, P. S., Mjolsness, E., & Elowitz, M. B. (2011). Measuring single-cell gene expression dynamics in bacteria using fluorescence time-lapse microscopy. *Nature Protocols*, 7(1), 80–88.

Abbreviations

ABC approximate Bayesian computation.

BFMI Bayesian fraction of missing.

BKM biochemical kinetic model.

c.v. coefficient of variation.

CI credible interval.

CIR Cox-Ingersoll-Ross.

CLE chemical Langevin equation.

CME chemical master equation.

DBM diffusion bridge Milstein.

ESS effective sample size.

GBM geometric Brownian motion.

GFP green fluorescence protein.

HMC Hamiltonian Monte Carlo.

HPDI highest probability density interval.

LC left-conditioned.

LNA linear noise approximation.

MB modified bridge.

MCMC Markov chain Monte Carlo.

MJP Markov jump process.

mRNA messenger ribonucleic acid.

NUTS No-U-Turn Sampler.

ODE ordinary differential equation.

RMSE root mean square error.

RRE reaction rate equation.

s.d. standard deviation.

SDE stochastic differential equation.

SSA stochastic simulation algorithm.

Symbols

$\mathbb{N}$	The natural numbers.
$\mathbb{N}_0$	The non-negative integers, i. e. $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$.
$\mathbb{Z}$	The integers.
$\mathbb{R}$	The real numbers.
$\mathbb{R}_+$	The non-negative real numbers, i. e. $\mathbb{R}_+ = \{x \in \mathbb{R} \,\vert\, x \geq 0\}$.
$\mathbb{R}_+^*$	The positive real numbers, i. e. $\mathbb{R}_+^* = \{x \in \mathbb{R} \,\vert\, x > 0\}$.
$\vee$	The max operator, i. e. $a \vee b := \max(a, b)$ for $a, b \in \mathbb{R}$.
$\boldsymbol{a}, \boldsymbol{A}$	Vectors and matrices are denoted by bold symbols.
$\boldsymbol{a}^{\mathsf{Tr}}, \boldsymbol{A}^{\mathsf{Tr}}$	The transpose of vector $\boldsymbol{a}$ and matrix $\boldsymbol{A}$, respectively.
$\boldsymbol{I}_d$	The $d \times d$-dimensional identity matrix.
$\mathcal{U}(a, b)$	The uniform distribution on the interval $[a, b] \subset \mathbb{R}$.
$\mathrm{Exp}(\lambda)$	The exponential distribution with intensity parameter $\lambda \in \mathbb{R}_+^*$.
$\mathrm{Po}(\lambda)$	The Poisson distribution with parameter $\lambda \in \mathbb{R}_+^*$.
$\phi(\cdot \,\vert\, \mu, \eta^2)$	The density of the normal distribution with mean μ and variance η^2.
$\mathcal{N}(\mu, \eta^2)$	The normal distribution with mean μ and variance η^2.
$\mathcal{N}_{\geq a}(\mu, \eta^2)$	The truncated normal distribution with mean μ and variance η^2 truncated from below by a.
$\mathcal{LN}(\mu, \eta^2)$	The lognormal distribution, i.e. $X \sim \mathcal{LN}(\mu, \eta^2) \Leftrightarrow \log(X) \sim \mathcal{N}(\mu, \eta^2)$.