Computational Inference in Mathematical Biology: Methodological Developments and Applications

David James Warne

BMath/BInfoTech (Hons.)

under the supervision of

Professor Matthew J. Simpson

and

Professor Scott W. McCue

School of Mathematical Sciences

Science and Engineering Faculty

Queensland University of Technology

and external associate supervision of

Professor Ruth E. Baker

Mathematical Institute

University of Oxford

2020

SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

To Sarah, the love of my life.

Copyright in Relation to This Thesis

Statement of Original Authorship

I hereby declare that this submission is my own work and to the best of my knowledge it contains no material previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any other degree or diploma at QUT or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by colleagues, with whom I have worked at QUT or elsewhere, during my candidature, is fully acknowledged.

I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the projects design and conception or in style, presentation and linguistic expression is acknowledged.

Signature: David Warne

Date: 19/01/2020

Abstract

Biological systems are complex and composed of many interacting components. This complexity occurs on multiple spatial and temporal scales. The behaviour and function of tissues depends on the complex interactions of cells, and in turn, cell dynamics depends on intercellular and intracellular biochemical reaction networks.

Depending on the scale, a diverse range of mathematical modelling frameworks are applied. On the macroscale of tissues and tumors, deterministic models based on differential equations are often used to describe collective cell behaviour. However, at the microscale of individual cell interactions or the nanoscale of biochemical processes, discrete stochastic models based on discrete state, continuous time Markov processes are employed to capture essential stochastic phenomena.

The effective application of models in practice depends upon reliable computational inference methods for experimental design, model calibration and model selection. There are many computational challenges in all these aspects, especially when stochastic models are applied. Stochastic models typically yield intractable likelihood functions, therefore likelihood-free methods for computational inference are often required. However, for models with computationally expensive simulation processes, likelihood-free inference is rarely feasible.

In this thesis, both applications for inference and the development of inference methodologies are considered. Key outcomes include: new results in the quantification of contact inhibition and cell motility mechanisms; and the development of novel computational inference algorithms suited for expensive stochastic models typical in the study of biochemical systems.

Acknowledgments

First, I would like to thank my principle supervisor, **Professor Matthew Simpson**, for all of his guidance and support over the duration of my candidature. Thank you for all of the fruitful discussions, guidance, technical assistance, feedback and career advice you have provided over the years. It has been instrumental to my development, both academically and personally. I especially want to acknowledge your support and encouragement during some of the more difficult times in my PhD journey. Thank you also for all of the time you have dedicated to reading drafts, discussing technical challenges, listening to my mathematical rants, and reigning in my tendency to fall down mathematical rabbit holes. I am grateful for all of the meetings and coffee chats, and I look forward future opportunities to collaborate.

Second, I want to acknowledge my external associate supervisor, **Professor Ruth Baker**. Thank you for finding time to regularly meet via Skype to provide guidance and useful discussions of technical ideas throughout the past four years. Thank you for hosting me at Oxford in July 2016 and introducing me to so many interesting people. This time enabled a number of essential revisions to my early work, and this thesis is far better off for it. I appreciate all the time you have taken over the years to read drafts, provide feedback and suggestions. I look forward to further opportunities to collaborate in the future.

I am also grateful for the support and encouragement from my associate supervisor, **Professor Scott McCue**.

In addition to my supervisors, there are many others who have been very helpful throughout my candidature. Thanks to all of my colleagues and fellow students at QUT, particularly, **Dr Wang Jin**, and **Mr Alexander Browning** for many interesting discussions and coffee chats. I also appreciate discussions with **Dr Christopher Lester** and **Dr Thomas Prescott** on multilevel and

multifidelity methods. Thanks to **Professor Michael Giles** and **Assistant Professor Abdul-Lateef Haji-Ali** for welcoming me into the Oxford MLMC special interest group, and enabling me to contribute to meetings via Skype throughout 2016 to 2018. These were very helpful meetings for my professional development.

There are also many to thank for funding that enabled this work to be performed. Thanks to the **Australian Government** for the scholarship through the Research Training Program. Thank you to **Professor Matthew Simpson** for the supervisor top-up scholarship. I also acknowledge all of the travel funding support from the **School of Mathematical Sciences**, the **Institute of Health and Biomedical Innovation**, the **Society for Mathematical Biology**, the **Australian Mathematical Society**, and the **Australian Mathematical Sciences Institute**. I also want to thank **Dr Joseph Young**, **Dr Neil Kelson** and **Professor Matthew Bellgard** for providing me with employment and invaluable experiences in the eResearch department (formerly High Performance Computing and Research Support) during my leave of absence from my PhD studies. Also thanks to **Associate Professor Christopher Drovandi** and the **Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers** for part-time Research Assistant work, and the **School of Mathematical Sciences** for the many sessional academic roles.

On a personal level, I want to thank my family. My dear wife, **Sarah**, thank you for your enduring love and support, your endless patience, and your blind confidence throughout this journey. Thanks to my son, **Nathanael**, and daughter, **Grace**, for your intellectually simulating discussion and teaching me to take time for play. Thanks to my brother, **Andrew**, for the many phone and coffee conversations discussing life, the universe and everything. Thanks to my parents, **Gary** and **Julie**, for making me the person I am today.

Finally, the spiritual and emotional journey of a PhD is something that can be understated. I can honestly say this journey, with all its heights of success and troughs of failure, has radically changed me on an internal level. Finding the language to express these things can be terribly difficult since terms are loaded with so many connections and connotations that I do not want to draw here. It will suffice to thank **all that cannot be described with words**.

"To realise that our knowledge is ignorance,

This is a noble insight.

To regard our ignorance as knowledge,

This is mental sickness."

~ Lao Tzu (Tao Te Ching, Chapter 71)

Table of Contents

Chapter 1

Introduction

1.1 Overview

Living organisms are immensely complex systems of interacting components. The proper functioning of macroscale components, such as organs and tissues, depend upon the collective behaviour and microscale interactions between motile cells that undergo coordinated proliferation and death. Furthermore, cell motility and proliferation are also fundamentally important in tissue development and repair (Puliafito et al., 2012). Aberrations in the intracellular biochemical processes that regulate normal cell motility and proliferation can lead to abnormal cell-cell interactions and result in pathological cases including life threatening diseases such as cancer (Abercrombie, 1979; Gerlee, 2013). These intracellular processes are a themselves complex nanoscale networks of reacting biomolecules including proteins, enzymes, DNA and RNA, that regulate gene expression leading to normal cellular function. To gain insight into biological function, both normal and pathological, requires mathematical modelling on multiple scales.

Mathematical models increase our understanding of biological systems in three main ways. Firstly, they enable the exploration of possible dynamics of a system under a proposed theoretical mechanism. Secondly, model predictions can be compared with experimental data to compare and validate theories (Silk et al., 2014). Lastly, key biological parameters can be

quantified using data interpreted through the lens of establish mathematical models (Johnston et al., 2015). These tasks require simulation of the model, the *forwards problem*, to explore dynamics and make predictions, and statistical inference, the *inverse problem*, to calibrate models, perform model selection, and to estimate parameters along with their uncertainties and sensitivities (Marino et al., 2008; Toni et al., 2009). The choice of modelling framework heavily affects the accuracy of predictions and inferences, though it is not always obvious how to determine the correct level of modelling detail *a priori*. Furthermore, some modelling frameworks will lead to computationally intractable inverse problems (Sunnåker et al., 2013). This results in a challenging trade-off problem between model accuracy, complexity and inference tractability.

Modern high resolution experimental techniques enable the observation of biological phenomena in unprecedented detail (Chen et al., 2014; Bajar et al., 2016; Leung and Chou, 2011; Sahl et al., 2017). Many experimental protocols, such as *in vitro* cell culture assays, use time-lapse optical microscopy to image the collective behaviour of cells under different environmental conditions or drug treatments (Delarue et al., 2014; Kramer et al., 2013; Liang et al., 2007). Intracellular biochemical information, such as cell cycle state and gene expression levels, can additionally be obtained through fluorescent reporters (Finkenstädt et al., 2008; Hass et al., 2014; Iofalla et al., 2008; Vittadello et al., 2018; Wilkinson, 2009). Despite high spatial resolutions, *in vitro* protocols often involve very few observations in time and assay duration is typically short (Chen et al., 2017; Huang et al., 2017), furthermore only a small number of fluorescent reporters may be used simultaneously (Wilkinson, 2009). The resulting data have low temporal resolution and partial biochemical state information (Golightly and Wilkinson, 2011; Toni et al., 2009), therefore experiments need to be designed specifically to maximise information content relevant for a particular study. Experimental design depends crucially on the model parameters of interest within the modelling framework that is applied, and optimal experimental design continues to be a challenging and unsolved problem. Consequently, experimental design is an active area of research (Browning et al., 2017; Drovandi and Pettitt, 2013; Liepe et al., 2013; Ryan et al., 2016).

A wide array of mathematical modelling frameworks are applied to improve our understanding of biological processes (Edelstein-Keshet, 2005; Murray, 2002; Wilkinson, 2012). Macroscale biological phenomena, such as wound healing and tumor growth, are often modelled using

deterministic continuum models based on differential equations. For example, ordinary differential equations are ubiquitous in the study of tumor growth and treatment (Enderling and Chaplain, 2014; Gerlee, 2013; Scott et al., 2013; Treloar et al., 2013), and reaction-diffusion partial differential equations are frequently applied to tissue engineering (Sengers et al., 2007), tumor growth and invasion (Swanson et al., 2003), cancer treatment (Jackson et al., 2015; Swanson et al., 2002), would healing (Flegg et al., 2010; Landman et al., 2007; Maini et al., 2004a,b; Nardini et al., 2016; Simpson et al., 2013), and embryonic development (Simpson et al., 2007b). In many cases, deterministic models of collective cell behaviour are based on the classical logistic growth model of Verhulst (1838) that describes population growth subject to a carrying capacity that limits growth (Edelstein-Keshet, 2005; Murray, 2002), and the Fisher-Kolmogorov-Petrovsky-Piscounov (Fisher-KPP) model (Fisher, 1937; Kolmogorov et al., 1937) that describes the spreading of a diffusive population that grows logistically (Edelstein-Keshet, 2005; Murray, 2002). Many extensions and generalisations are also studied and applied (Gurney and Nisbet, 1975; King and McCabe, 2003; Tsoularis and Wallace, 2002; Witelski, 1995).

If the spatial scale is sufficiently small, such that individual microscale cell-cell interactions must be characterised, then discrete stochastic models are more accurate descriptions of collective cell behaviour. These are often random walk models (Codling et al., 2008), in which individual cell motility and proliferation is governed by probability laws that depend on the local environment (Callaghan et al., 2006; Simpson et al., 2010). Many complex microscale cell behaviours and dynamics can be described within the framework (Binny et al., 2016; Böttger et al., 2015; Johnston et al., 2013; Treloar et al., 2014). Often deterministic continuum models can be considered as large population limits of discrete random walk models under mean-field assumptions (Cai et al., 2007; Callaghan et al., 2006; Jin et al., 2016a; Simpson et al., 2010). Compared to continuum models, the random walk framework is extremely computationally expensive, since a large number of simulations is required to determine expected behaviours. Therefore, it is common to apply mean-field models for the purpose of inference, despite the bias that will be incurred in doing so.

Within an individual cell, biochemical systems, such as gene regulatory networks, govern cell function and behaviour. At this nanoscale, regulatory networks are highly stochastic (Abkowitz et al., 1996; Arkin et al., 1998; McAdams and Arkin, 1997), subject to extrinsic and intrinsic noise (Elowitz et al., 2002; Keren et al., 2015) that is particularly prevalent when populations of

key molecules are low (Fedoroff and Fontana, 2002). This stochasticity can not be ignored when developing models since stochastic effects have been shown to be fundamental to normal cellular function (Elowitz et al., 2002) and have macroscale consequences (Smith and Grima, 2018). As a result, the most common modelling tools are continuous-time Markov processes (Gillespie, 1977), based on the chemical master equation (Gillespie, 1992), or stochastic differential equations (Higham, 2001), based on the chemical Langevin equation (Gillespie, 2000; Higham, 2008). Spatially heterogeneous biochemical processes can also be charactised using stochastic models (Gillespie et al., 2013; Smith and Grima, 2016) based on the reaction-diffusion master equation (Isaacson, 2008). The computational burden for simulation of stochastic biochemical reaction networks is also high.

Regardless of the modelling framework selected for a particular application, all models are an approximation of reality (Box, 1976). All models are built on underlying assumptions, and the validity of these assumptions can be difficult to test (Higham, 2008). However, reducing the number of assumptions is not always advantageous. For a given application and data, a complex model with fewer assumptions and more detailed mechanisms may provide little additional information over a simpler model with strict assumptions and fewer mechanisms. Furthermore, detailed models will often require inference on a larger number of parameters that can lead to overfitting (Akaike, 1974; Johnson and Omland, 2004; Schwarz, 1978; Stoica and Selen, 2004; Slezak et al., 2010; Spiegelhalter et al., 2002).

Additional model complexity typically comes with computational challenges, and model simplifications may be pragmatic choices that enable solutions to be obtained within practical computational constraints. A variety of options exist for model simulation, each of which holds implications for the solution accuracy and computational cost. For deterministic models, the spatial discretisation scheme and time stepping scheme needs to be selected (Fehlberg, 1969; Press et al., 1997; Sloan and Abbo, 1999). For stochastic models, exact simulation schemes (Anderson, 2007; Cao et al., 2004; Gibson and Bruck, 2000; Gillespie, 1977; Voliotis et al., 2016) are possible along with a multitude of approximate methods (Cao et al., 2005a,b, 2006; E et al., 2005; Gillespie, 2001; Rathinam et al., 2003; Tian and Burrage, 2004), including discretisation of stochastic differential equations (Higham, 2001; Kloeden and Platen, 1999). In general, more accurate numerical solutions have higher computational requirements. However,

stochastic models are generally significantly more computationally expensive than their deterministic counterparts. This is largely because a deterministic model need only be solved once for a given parameter set, but stochastic models require repeated simulations and Monte Carlo methods to obtain information on the distribution of possible model behaviours for a given fixed set of parameters. While significant advances have been made such as multilevel Monte Carlo methods (Giles, 2015; Higham, 2015), deterministic models require orders of magnitude less computation time than stochastic models. Therefore, it is common to apply deterministic mean-field continuum models in place of ensemble averages of many stochastic simulations.

The choice of model and simulation scheme also impacts the performance of algorithms for statistical inference. Bayesian inference typically requires Monte Carlo sampling schemes to quantify model or parameter uncertainties. In particular, rejection sampling (Beaumont et al., 2002; Pritchard et al., 1999; Tavaré et al., 1997), Markov chain Monte Carlo (Hastings, 1970; Marjoram et al., 2003; Metropolis et al., 1953) and sequential Monte Carlo (Del Moral et al., 2006; Sisson et al., 2007) are popular sampling approaches for Bayesian inference. These sampling schemes rely on the evaluation of a likelihood function that quantifies the probability of the observed data given the assumed model and parameter set. For deterministic models, the likelihood function typically has a closed-form that allows exact Bayesian inferences to be computed. However, this is rarely the case for stochastic models since the likelihood function is either computationally or analytically intractable, especially when the experimental data include only partial observation of system state (Golightly and Wilkinson, 2011). These inverse problems are very challenging and require likelihood-free methods (Sisson et al., 2018; Andrieu et al., 2010), that are extremely computationally expensive, since likelihood function evaluations are replaced by a large number of stochastic simulations. Just as with model simulation, the choice of computational inference scheme is also a trade-off between computational cost and statistical bias arising either from the choice of model or from the inference method.

The various options for experimental design, model development, model simulation and statistical inference, result in difficult decisions that involve trade-offs. In experimental design, the trade-off is between information obtained on one quantity of interest at the expense of another, another trade-off is information versus cost of materials and time. Model development must consider trade-offs between model complexity, fitness and computational cost. The choice of simulation method is also a trade-off between solution accuracy and computational cost.

Lastly, the choice of statistical inference technique is a trade-off between the bias or variance of an estimate versus the computational cost. None of these trade-offs can be considered in isolation: experimental design necessarily needs to consider the model and statistical methods to be applied; model development and selection needs to consider the kind of data that will be available and the computational requirements of the available simulation and inference schemes for the model; the choice of simulation scheme needs to be informed by the data and model resolution to determine appropriate required accuracy levels, but also the accuracy of the simulation scheme has implications on the possible accuracy of the inference method; and finally, the choice of inference scheme must take into account the combined bias of the model, simulation scheme, and inference method. Within the mathematical biology community, there are few research projects that consider the multifaceted interactions between various design choices on the final scientific conclusions.

In this thesis, we explore various aspects of experimental design, model selection, stochastic simulation and computational inference algorithms in the context of applications in cell biology and biochemistry. We demonstrate the advantage of the Bayesian approach over frequentist point estimates to resolve previously unanswered questions from cell biology. We also develop new efficient computational schemes that dramatically reduce the number of expensive stochastic simulations required for accurate Bayesian parameter inference for models with intractable likelihoods.

1.2 Research questions

In this section, we present the research questions that are designed to close gaps in the current research literature. Each research question motivates a chapter of this thesis. A short background highlighting the state of the literature at the time of each study is presented along with how the study contributes to the body of knowledge. For a detailed literature review surrounding each research question, see the introduction and background sections in each of the corresponding chapters.

1. **What are the design principles for proliferation assays to optimally quantify contact inhibition?**

   Contact inhibition of proliferation refers to the tendency for cells to become non-proliferative in regions of high local cell density (Abercrombie, 1970). This process is fundamental to the normal functioning and repair of tissues (Liu et al., 2011; Puliafito et al., 2012), and many pathological cases, such as cancer, arise from deleterious down-regulation of contact inhibition (Abercrombie, 1979). Unfortunately, standard proliferation assay protocols are not well suited for accurate quantification of contact inhibition (Jin et al., 2017) due largely to short assay durations (Chen et al., 2017; Huang et al., 2017). This is highlighted in a study by Sarapata and de Pillis (2014) that discusses the challenges in calibration of tumor growth models with parameters related to contact inhibition (Gerlee, 2013), such as the carrying capacity density within the logistic growth model.

   The Bayesian approach can be applied to determine the optimal experimental design in order to minimise inference uncertainty (Liepe et al., 2013; Silk et al., 2014; Vanlier et al., 2012, 2014). For example, a recent study by Browning et al. (2017) applied a Bayesian approach to experimental design to optimise proliferation assays duration for the purpose of estimating proliferation rates. We apply a detailed and rigorous Bayesian analysis of the proliferation assay protocol to quantify the effect that assay design elements have on the uncertainty in the carrying capacity density, a key parameter quantifying contact inhibition for the logistic growth model (Verhulst, 1838; Murray, 2002; Jin et al., 2017). In particular, we consider various design options for assay duration, the number of replicates, initial seeding strategies, and temporal spacing of observations. New design principles and recommendations are provided for minimisation of the uncertainty in carrying capacity estimates.

2. **To what extent can the Bayesian approach improve model comparison and selection results for reaction–diffusion models in cell biology?**

   In the study of collective cell behaviour in mathematical biology, reaction–diffusion partial differential equations are a commonly utilised modelling framework. Diverse flux and source terms are applied to model various mechanisms for cell motility and proliferation (Browning et al., 2017; Jin et al., 2016a; Tsoularis and Wallace, 2002; Witelski, 1995). However, it can often be difficult to determine the most important mechanisms to

include for any particular application. In many cases, little justification is provided for particular choices (Simpson et al., 2011). When model selection is performed, the most common approach is to compare residuals between competing models and data after first calibrating the competing models using the principle of maximum likelihood (Bianchi et al., 2016; Sengers et al., 2007; Sherratt and Murray, 1990). This approach, however, is known to be very sensitive to model miss-specification and will tend to favor over-parameterised models (Akaike, 1974; Box, 1976) that can lead to biologically unrealistic results (Jin et al., 2016b; Sarapata and de Pillis, 2014; Slezak et al., 2010).

We analyse the scratch assay data obtained by Jin et al. (2016b) for prostate cancer cells, and consider a Bayesian approach to model selection between different mechanisms for motility. We compare, for each model, Bayesian parameter inferences for various scratch assay initial cell population densities (Jin et al., 2016b, 2017). We also apply various information criteria that penalise the likelihood of more complex models (Akaike, 1974; Schwarz, 1978; Spiegelhalter et al., 2002). As a result, we find motility that is linearly dependent on local population density is the most parsimonious model for the experiments of Jin et al. (2016b). Furthermore, we provide an exemplar of Bayesian model selection in cell biology.

3. **What are the relationships and connections between the state-of-the-art methods for simulation and inference for biochemical systems?**

   For intracellular biochemical processes, it has been demonstrated that stochasticity is fundamental to the normal operation biochemical systems, such as gene regulation (Abkowitz et al., 1996; Arkin et al., 1998; Eldar and Elowitz, 2010; Elowitz et al., 2002; Kærn et al., 2005; Keren et al., 2015; McAdams and Arkin, 1997; Raj and van Oudenaarden, 2008; Soltani et al., 2016; Taniguchi et al., 2010). As a result, stochastic models of biochemical reaction networks are required to capture stochastic effects that cannot be replicated with deterministic chemical rate equations (Bressloff, 2017; Moss and Pei, 1995; Paulsson et al., 2000; Thattai and van Oudenaarden, 2001; Tian and Burrage, 2006). The are many algorithmic options to consider for practical applications that involve simulation for the *forwards problem* and computational Bayesian inference for the *inverse problem*. These two problems are interrelated, however, there has been no survey article that discusses both these topics together. Rather, reviews on stochastic simulation (Gillespie et al., 2013; Higham, 2008; Schnoerr et al., 2017), have been distinct to those on computational

inference (Golightly and Wilkinson, 2011; Green et al., 2015; Sunnåker et al., 2013; Toni et al., 2009).

In the context of stochastic biochemical reaction network models, we provide a detailed survey of stochastic simulation for generating model sample paths (Anderson, 2007; Gibson and Bruck, 2000; Gillespie, 1977, 2001; Voliotis et al., 2016) and the computation of summary statistics (Erban et al., 2007; Gillespie, 2000), along with a survey of inference algorithms that depend on Bayesian sampling techniques (Del Moral et al., 2006; Hastings, 1970; Metropolis et al., 1953) with a focus on likelihood-free approaches (Marjoram et al., 2003; Sisson et al., 2007) that rely on the model forwards problem. This review is unique as dependence between the inverse and forwards problems are highlighted, and, for the first time, an accessible introduction is presented on state-of-the-art multilevel Monte Carlo methods for the accelerated computation of expectations related to the forwards problem (Anderson and Higham, 2012; Giles, 2008; Lester et al., 2015; Wilson and Baker, 2016) and the inverse problem (Guha and Tan, 2017; Jasra et al., 2019) (Chapter 6). Furthermore, example implementations of all discussed algorithms are provided using the MATLAB® programming environment.

4. **What are the practical implementation considerations for pseudo-marginal likelihood-free inference?**

   Approximate Bayesian computation (Sisson et al., 2018; Sunnåker et al., 2013) is a method for likelihood-free inference that is becoming increasingly popular is biological applications (Liepe et al., 2014; Ross et al., 2017; Toni et al., 2009; Vo et al., 2015a). However, alternatives such as pseudo-marginal methods (Andrieu and Roberts, 2009; Andrieu et al., 2010; Beaumont, 2003) are not commonly used despite having some advantageous properties for inference on partially observed stochastic processes (Golightly and Wilkinson, 2008, 2011).

   We provide a practical guide to the application of pseudo-marginal methods for inference of chemical Langevin description of biochemical reaction network models (Gillespie, 2000; Higham, 2008). We compare and contrast the pseudo-marginal approach with approximate Bayesian computation methods. Key design choices and practical guides for tuning these methods are also demonstrated through examples, and example implementations are provided using the Julia programming language.

5. **To what extent can multilevel Monte Carlo be exploited to accelerate approximate Bayesian computation?**

   Multilevel Monte Carlo is a state-of-the-art numerical scheme for efficiently estimating expectations with low variance and bias (Giles, 2008; Higham, 2015). The method relies upon a sequence of approximations to the stochastic process of interest (Giles, 2015), with low-fidelity approximations that are computationally inexpensive and high-fidelity approximations that are computationally expensive. The key idea is that linearity of expectation enables the construction of a new Monte Carlo estimator using a telescoping sum of bias corrections that estimate the bias incurred from decreasing the simulation fidelity. The multilevel Monte Carlo estimator can be several orders of magnitude more efficient in mean square error, and has been successfully demonstrated for various applications involving stochastic differential equations (Giles, 2008; Giles et al., 2015; Higham, 2015) and discrete-state continuous time Markov processes (Anderson and Higham, 2012; Lester et al., 2015, 2016; Lester et al., 2017; Wilson and Baker, 2016).

   The multilevel approach has also be applied to Bayesian inference problems (Beskos et al., 2017; Dodwell et al., 2015; Efendiev et al., 2015) and particle filters Gregory et al. (2016); Jasra et al. (2017). However, it is unclear how to extend these approaches for likelihood-free inference methods, such as approximate Bayesian computation (Sisson et al., 2018). Here we apply a multilevel estimator for the cumulative distribution function of the Bayesian posterior distribution. Through application of ideas from Giles et al. (2015) and Wilson and Baker (2016), combined with a novel coupling scheme, we are able to accelerate rejection sampling for approximate Bayesian computation by an order of magnitude.

6. **How can approximate models be used to accelerate Bayesian inference for expensive stochastic models?**

   While continuum reaction–diffusion models are often applied to describe collective cell behaviour, there are many cases where these mean-field assumptions are invalid, particularly when clustering occurs (Agnew et al., 2014; Simpson et al., 2013). In such cases, stochastic discrete random walk models are applied to more accurately describe individual cell–cell interactions (Callaghan et al., 2006; Jin et al., 2016a; Johnston et al., 2014; Simpson et al., 2010). Unfortunately, the forwards problem for these discrete

models are computationally expensive to the extent that accurate likelihood-free Bayesian inference is completely intractable. The challenge of computational inference is not only an problem in biology, but also arises in many areas of science including astronomy (The Event Horizon Telescope Collaboration et al., 2019), climate science (Holden et al., 2018), and engineering (Rynn et al., 2019).

Consequently, there is a growing interest in using computationally cheap approximations to inform Bayesian sampling for computationally expensive stochastic models. Examples include, surrogate models (Rynn et al., 2019), emulators (Buzbas and Rosenberg, 2015), delayed rejection sampling (Banterle et al., 2019; Everitt and Rowińska, 2017), multilevel Monte Carlo (Jasra et al., 2019; Warne et al., 2018), and multifidelity methods (Prescott and Baker, 2018). We develop two new likelihood-free sequential Monte Carlo samplers that accelerate Bayesian inference for expensive stochastic models. In particular, we exploit approximate models to inform parameter proposals, and moment-matching transforms to emulate the expensive stochastic model using approximations. Our methods can accurately compute inferences consistent with the expensive stochastic model while using an order of magnitude less stochastic simulations.

1.3 Outcomes of this thesis

This thesis is presented by published papers and consists of six manuscripts in total. Four of these manuscripts have been peer-reviewed and published in high quality journals, the remaining two manuscripts have been submitted to high quality journals and are under review. The PhD candidate is the first author in all six of the manuscripts. This fulfills the requirements for the Queensland University of Technology to award a thesis by published papers.

The following six manuscripts form the chapters of this thesis:

- **David J. Warne**, Ruth E. Baker and Matthew J. Simpson (2017). Optimal quantification of contact inhibition in cell populations. *Biophysical Journal*, 113:1920–1924. DOI:10.1016/j.bpj.2017.09.016 (Chapter 2)
- **David J. Warne**, Ruth E. Baker and Matthew J. Simpson (2019). Using experimental data and information criteria to guide model selection for reaction–diffusion problems in

mathematical biology. *Bulletin of Mathematical Biology*, 81:1760–1804.
DOI:10.1007/s11538-019-00589-x (Chapter 3)

- **David J. Warne**, Ruth E. Baker and Matthew J. Simpson (2019). Simulation and inference algorithms for stochastic biochemical reaction networks: from basic concepts to state-of-the-art. *Journal of the Royal Society Interface*, 16:20180943.
DOI:10.1098/rsif.2018.0943 (Chapter 4)

- **David J. Warne**, Ruth E. Baker and Matthew J. Simpson (2019). A practical guide to pseudo-marginal methods for computational inference in systems biology. (In review, submitted Jan 2020) *Journal of Theoretical Biology*.
`arXiv:1912.12404 [q-bio.MN]` (Chapter 5)

- **David J. Warne**, Ruth E. Baker and Matthew J. Simpson (2018). Multilevel rejection sampling for approximate Bayesian computation. *Computational Statistics & Data Analysis*, 124:71–86.
DOI:10.1016/j.csda.2018.02.009 (Chapter 6)

- **David J. Warne**, Ruth E. Baker and Matthew J. Simpson (2019). Rapid Bayesian inference for expensive stochastic models. (In review, submitted Jan 2020) *Journal of Computational and Graphical Statistics*.
`arXiv:1909.06540 [stat.CO]` (Chapter 7)

1.4 Structure of this thesis

The main content of this thesis is divided into two parts. Part I consists of two chapters (Chapters 2 and 3) that address research questions related to applications of Bayesian methods in cell biology. Part II consists of four chapters (Chapters 4–7) that address research questions related to computational schemes for stochastic simulation and likelihood-free Bayesian inference. The remainder of this section lists the chapters of this thesis, their contents and novel contributions.

Each of the main chapters within this thesis corresponds to a journal publication (Chapters 2,3,4, and 6) or a manuscript under review (Chapters 5 and 7). Consequently, each chapter may be read independently of the others, however, the chapters are also organised in such a way that

they form a consistent whole that may be read in order sensibly. The formatting and notation has been altered, compared with published versions, for the sake of consistency, otherwise each chapter is exactly as published. Because each chapter can be read independently, there is some overlap in background content across different chapters. While the specific structure of individual chapters vary, each chapter contains an introduction section that describes the problem of interest, background information, and the contributions of work to the field. Each chapter also includes a supplementary material section that contains appendix information for the chapter, including extended analysis, additional results and data tables. Some chapters also contain links to publicly available code repositories for software developed for that chapter. The remainder of this section lists the chapters of this thesis, their contents and novel contributions.

Chapter 2 addresses the problem of reliable inference of contact inhibition using proliferation assays. The logistic growth model is applied as a characteristic model for cell population growth limited by contact inhibition. Through the application of a Bayesian approach, several guidelines for the optimal proliferation assay design are provided to ensure the carrying capacity density can be estimated with minimal uncertainty.

In Chapter 3 the problem of model selection in the reaction–diffusion framework is applied to answer an unresolved problem in cell biology. The problem is that of whether linear diffusion is an appropriate model to describe the mechanism of collective cell spreading. This question is addressed by comparing the Fisher-KPP model, the Porous Fisher model, and the Generalised Porous Fisher model for the purposes of inference using a detailed scratch assay data set. The results demonstrate that the Porous Fisher model, in which cell motility is linearly dependent on local cell population density, provides the best trade-off between model complexity and model fit.

Chapter 4 introduces Part II of this thesis with a detailed review of the various simulation and inference methods that are relevant to studies of biochemical processes. This is the first detailed, yet accessible, review including both algorithms for stochastic simulation, advanced multilevel Monte Carlo methods, approximate Bayesian computation sampling methods, and the connections between all these topics. Another significant contribution is a library of example implementation using the MATLAB® programming language.

Chapter 5 provides a more detailed guide to an alternative likelihood-free inference method,

called the pseudo-marginal approach. This chapter is designed to highlight the advantages and limitations of this approach compared with approximate Bayesian computation for stochastic models of biochemical reaction networks based on the chemical Langevin equation. Various important and practical design choices are discussed and demonstrated for efficient application of the, so called, particle marginal Metropolis-Hastings algorithm. Several examples are provided including parameter inference for the challenging Schlögl biochemical reaction network that exhibits stochastic bi-stability, and the replessilator gene regulatory network that exhibits stochastic oscillatory dynamics. Example algorithm implementations are provided using the high performance, open source Julia programming language.

Chapter 6 demonstrates that multilevel Monte Carlo methods can be applied in the context of likelihood-free inference using approximate Bayesian computation. The derivation of the algorithm is presented, including a novel method for coupling nested rejection samplers that enables a variance reduction on bias correction estimates. The new approach requires very little tuning, is significantly more efficient than standard rejection sampling, and is also competitive with state-of-the-art sequential Monte Carlo schemes.

Chapter 7 considers the challenge of practical application of approximate Bayesian computation when an expensive stochastic model is the target of study. Here, two novel extensions to sequential Monte Carlo are presented that utilise computationally inexpensive approximate models to inform sampling of the stochastic model. Numerical demonstrations are provided using lattice-based random walk models and mean-field models for approximation. The first extension results in unbiased estimates with respect to the expensive stochastic model with two–four times fewer expensive stochastic simulations. The second method introduces some bias, but requires an order of magnitude less stochastic simulations.

Chapter 8 provides a summary of the contributions to the literature from this thesis, and discusses various avenues for future research.

1.5 Statement of joint authorship

This section provides the contributions of the PhD candidate and the co-authors to each of the six manuscripts contained in this thesis. All co-authors have consented to the inclusion of each

manuscript.

Chapter 2: Optimal quantification of contact inhibition in cell populations

The associated manuscript for this chapter is:

David J. Warne, Ruth E. Baker and Matthew J. Simpson (2017). Optimal quantification of contact inhibition in cell populations. *Biophysical Journal*, 113:1920–1924.
DOI:10.1016/j.bpj.2017.09.016

Abstract Contact inhibition refers to a reduction in the rate of cell migration and/or cell proliferation in regions of high cell density. Under normal conditions contact inhibition is associated with the proper functioning tissues, whereas abnormal regulation of contact inhibition is associated with pathological conditions, such as tumor spreading. Unfortunately, standard mathematical modeling practices mask the importance of parameters that control contact inhibition through scaling arguments. Furthermore, standard experimental protocols are insufficient to quantify the effects of contact inhibition because they focus on data describing early time, low-density dynamics only. Here we use the logistic growth equation as a caricature model of contact inhibition to make recommendations as to how to best mitigate these issues. Taking a Bayesian approach we quantify the trade-off between different features of experimental design and estimates of parameter uncertainty so that we can re-formulate a standard cell proliferation assay to provide estimates of both the low-density intrinsic growth rate, λ, and the carrying capacity density, K, which is a measure of contact inhibition.

Statement of contribution:

- **David J. Warne (Candidate)** designed the research, performed the research, contributed analytic tools, analysed the data, wrote the manuscript, edited and approved the final manuscript.
- Ruth E. Baker designed the research, contributed analytic tools, analysed the data, wrote the manuscript, edited and approved the final manuscript.

- Matthew J. Simpson designed the research, contributed analytic tools, analysed the data, wrote the manuscript, edited and approved the final manuscript, acted as corresponding author.

Chapter 3: Using experimental data and information criteria to guide model selection for reaction–diffusion problems in mathematical biology

The associated manuscript for this chapter is:

David J. Warne, Ruth E. Baker and Matthew J. Simpson (2019). Using experimental data and information criteria to guide model selection for reaction–diffusion problems in mathematical biology. *Bulletin of Mathematical Biology*, 81:1760–1804.
DOI:10.1007/s11538-019-00589-x

Abstract Reaction–diffusion models describing the movement, reproduction and death of individuals within a population are key mathematical modelling tools with widespread applications in mathematical biology. A diverse range of such continuum models have been applied in various biological contexts by choosing different flux and source terms in the reaction–diffusion framework. For example, to describe the collective spreading of cell populations, the flux term may be chosen to reflect various movement mechanisms, such as random motion (diffusion), adhesion, haptotaxis, chemokinesis and chemotaxis. The choice of flux terms in specific applications, such as wound healing, is usually made heuristically, and rarely is it tested quantitatively against detailed cell density data. More generally, in mathematical biology, the questions of model validation and model selection have not received the same attention as the questions of model development and model analysis. Many studies do not consider model validation or model selection, and those that do often base the selection of the model on residual error criteria after model calibration is performed using nonlinear regression techniques. In this work, we present a model selection case study, in the context of cell invasion, with a very detailed experimental data set. Using Bayesian analysis and information criteria, we demonstrate that model selection and model validation should account for both residual errors and model complexity. These considerations are often overlooked in the mathematical

biology literature. The results we present here provide a straightforward methodology that can be used to guide model selection across a range of applications. Furthermore, the case study we present provides a clear example where neglecting the role of model complexity can give rise to misleading outcomes.

Statement of contribution:

- **David J. Warne (Candidate)** designed the research, performed the research, implemented numerical schemes, analysed the data, wrote the manuscript, edited and approved the final manuscript.
- Ruth E. Baker designed the research, contributed analytic tools, wrote the manuscript, edited and approved the final manuscript.
- Matthew J. Simpson designed the research, contributed analytic tools, analysed the data, wrote the manuscript, edited and approved the final manuscript, acted as corresponding author.

Chapter 4: Simulation and inference algorithms for stochastic biochemical reaction networks: from basic concepts to state-of-the-art

The associated manuscript for this chapter is:

David J. Warne, Ruth E. Baker and Matthew J. Simpson (2019). Simulation and inference algorithms for stochastic biochemical reaction networks: from basic concepts to state-of-the-art. *Journal of the Royal Society Interface*, 16:20180943.
DOI:10.1098/rsif.2018.0943

Abstract Stochasticity is a key characteristic of intracellular processes such as gene regulation and chemical signalling. Therefore, characterising stochastic effects in biochemical systems is essential to understand the complex dynamics of living things. Mathematical idealisations of biochemically reacting systems must be able to capture stochastic phenomena. While robust

theory exists to describe such stochastic models, the computational challenges in exploring these models can be a significant burden in practice since realistic models are analytically intractable. Determining the expected behaviour and variability of a stochastic biochemical reaction network requires many probabilistic simulations of its evolution. Using a biochemical reaction network model to assist in the interpretation of time course data from a biological experiment is an even greater challenge due to the intractability of the likelihood function for determining observation probabilities. These computational challenges have been subjects of active research for over four decades. In this review, we present an accessible discussion of the major historical developments and state-of-the-art computational techniques relevant to simulation and inference problems for stochastic biochemical reaction network models. Detailed algorithms for particularly important methods are described and complemented with MATLAB® implementations. As a result, this review provides a practical and accessible introduction to computational methods for stochastic models within the life sciences community.

Statement of contribution:

- **David J. Warne (Candidate)** designed the review, performed the review, contributed analytical tools, developed algorithm implementations, wrote the manuscript, edited and approved the final manuscript.

- Ruth E. Baker designed the review, contributed analytical tools, wrote the manuscript, edited and approved the final manuscript.

- Matthew J. Simpson designed the review, contributed analytical tools, wrote the manuscript, edited and approved the final manuscript, acted as corresponding author.

Chapter 5: A practical guide to pseudo-marginal methods for computational inference in systems biology

The associated manuscript for this chapter is:

David J. Warne, Ruth E. Baker and Matthew J. Simpson (2019). A practical guide to pseudo-marginal methods for computational inference. (In review, submitted Jan 2020) *Journal of Theoretical Biology*.

Abstract For many stochastic models of interest in mathematical biology, such as stochastic biochemical reaction networks, exact quantification of parameter uncertainty through statistical inference is intractable. Likelihood-free computational inference techniques enable parameter inferences when the likelihood function for the model is intractable, but the generation of many sample paths is feasible through stochastic simulation of the forward problem. A ubiquitous likelihood-free method in mathematical biology is approximate Bayesian computation that accepts parameters that result in low discrepancy between stochastic simulations and data. However, it can be difficult to assess how the resulting inferences are affected by the acceptance threshold and discrepancy function. The pseudo-marginal approach is an alternative likelihood-free inference method that applies a Monte Carlo estimate to the likelihood function. This approach has several advantages, particularly in the context of noisy, partially observed, time-course data typical in biochemical studies. Specifically, the pseudo-marginal approach enables exact inferences and uncertainty quantification, and may be efficiently combined with particle filters for low variance likelihood estimation. In this work, we provide a practical introduction to the pseudo-marginal approach using inference for biochemical reaction networks as a case study. Implementations of key algorithms and examples are provided using the Julia programming language; a high performance, open source programming language for scientific computing.

Statement of contribution:

- **David J. Warne (Candidate)** designed the research, performed the research, implemented numerical schemes, contributed analytic tools, wrote the manuscript, edited and approved the final manuscript, acted as corresponding author.

- Ruth E. Baker designed the research, contributed analytic tools, edited and approved the final manuscript.

- Matthew J. Simpson designed the research, contributed analytic tools, edited and approved the final manuscript.

Chapter 6: Multilevel rejection sampling for approximate Bayesian computation

The associated manuscript for this chapter is:

David J. Warne, Ruth E. Baker and Matthew J. Simpson (2018). Multilevel rejection sampling for approximate Bayesian computation. *Computational Statistics & Data Analysis*, 124:71–86. DOI:10.1016/j.csda.2018.02.009

Abstract Likelihood-free methods, such as approximate Bayesian computation, are powerful tools for practical inference problems with intractable likelihood functions. Markov chain Monte Carlo and sequential Monte Carlo variants of approximate Bayesian computation can be effective techniques for sampling posterior distributions in an approximate Bayesian computation setting. However, without careful consideration of convergence criteria and selection of proposal kernels, such methods can lead to very biased inference or computationally inefficient sampling. In contrast, rejection sampling for approximate Bayesian computation, despite being computationally intensive, results in independent, identically distributed samples from the approximated posterior. An alternative method is proposed for the acceleration of likelihood-free Bayesian inference that applies multilevel Monte Carlo variance reduction techniques directly to rejection sampling. The resulting method retains the accuracy advantages of rejection sampling while significantly improving the computational efficiency.

Statement of contribution:

- **David J. Warne (Candidate)** designed the research, performed the research, implemented numerical schemes, contributed analytic tools, wrote the manuscript, edited and approved the final manuscript, acted as corresponding author.
- Ruth E. Baker designed the research, contributed analytic tools, wrote the manuscript, edited and approved the final manuscript.

- Matthew J. Simpson designed the research, contributed analytic tools, wrote the manuscript, edited and approved the final manuscript.

Chapter 7: Rapid Bayesian inference for expensive stochastic models

The associated manuscript for this chapter is:

David J. Warne, Ruth E. Baker and Matthew J. Simpson (2019). Rapid Bayesian inference for expensive stochastic models. (In review, submitted Jan 2020) *Journal of Computation and Graphical Statistics*.
arXiv:1909.06540 [stat.CO]

Abstract Almost all fields of science rely upon statistical inference to estimate unknown parameters in theoretical and computational models. While the performance of modern computer hardware continues to grow, the computational requirements for the simulation of models are growing even faster. This is largely due to the increase in model complexity, often including stochastic dynamics, that is necessary to describe and characterise phenomena observed using modern, high resolution, experimental techniques. Such models are rarely analytically tractable, meaning that extremely large numbers of expensive stochastic simulations are required for parameter inference. In such cases, parameter inference can be practically impossible. In this work, we present new computational Bayesian techniques that accelerate inference for expensive stochastic models by using computationally cheap approximations to inform feasible regions in parameter space, and through learning transforms that adjust the biased approximate inferences to closer represent the correct inferences under the expensive stochastic model. Using topical examples from ecology and cell biology, we demonstrate a speed improvement of an order of magnitude without any loss in accuracy.

Statement of contribution:

- **David J. Warne (Candidate)** designed the research, performed the research, implemented numerical schemes, contributed analytic tools, wrote the manuscript, edited and

approved the final manuscript, acted as corresponding author.

- Ruth E. Baker designed the research, contributed analytic tools, wrote the manuscript, edited and approved the final manuscript.
- Matthew J. Simpson designed the research, contributed analytic tools, wrote the manuscript, edited and approved the final manuscript.

Part I

Applications for Computational Inference in Cell Biology

Chapter 2

Optimal Quantification of Contact Inhibition in Cell Populations

A paper published in *Biophysical Journal*.

David J. Warne, Ruth E. Baker and Matthew J. Simpson (2017). Optimal quantification of contact inhibition in cell populations. *Biophysical Journal*, 113:1920–1924.
DOI:10.1016/j.bpj.2017.09.016

Abstract Contact inhibition refers to a reduction in the rate of cell migration and/or cell proliferation in regions of high cell density. Under normal conditions contact inhibition is associated with the proper functioning tissues, whereas abnormal regulation of contact inhibition is associated with pathological conditions, such as tumor spreading. Unfortunately, standard mathematical modeling practices mask the importance of parameters that control contact inhibition through scaling arguments. Furthermore, standard experimental protocols are insufficient to quantify the effects of contact inhibition because they focus on data describing early time, low-density dynamics only. Here we use the logistic growth equation as a caricature model of contact inhibition to make recommendations as to how to best mitigate these issues. Taking a Bayesian approach we quantify the trade-off between different features of experimental design and estimates of parameter uncertainty so that we can re-formulate a standard cell proliferation

assay to provide estimates of both the low-density intrinsic growth rate, λ, and the carrying capacity density, K, which is a measure of contact inhibition.

2.1 Introduction

Contact inhibition is the tendency of cells to become non-migratory and/or non-proliferative in regions of high cell density (Abercrombie, 1970). The phenomena of contact inhibition of migration, involving processes such as adhesion, paralysis and contraction (Abercrombie, 1970), is distinct to contact inhibition of proliferation, driven by cell-cell signaling and adhesion (Levine et al., 1965; Liu et al., 2011); both phenomena are essential for the regulation of structure and function of multicellular organisms.

Down-regulation of contact inhibition of proliferation enhances tumor spreading (Abercrombie, 1979), while wound healing and tissue regeneration also depend crucially on contact inhibition of proliferation (Puliafito et al., 2012). While contact inhibition of proliferation is ubiquitous in both normal and pathological processes, it is difficult to quantify the impact of such contact inhibition in complex biological systems despite the availability of experimental data. Therefore, mathematical models have an important role in informing our understanding of how contact inhibition of proliferation affects collective cell behavior.

The most fundamental mathematical model describing contact inhibition of cell proliferation is the logistic growth model (Maini et al., 2004a,b; Treloar et al., 2014; Tsoularis and Wallace, 2002),

$$\frac{\mathrm{d}C(t)}{\mathrm{d}t} = \underbrace{\lambda C(t)}_{\textbf{proliferation}} \times \underbrace{\left[1 - \frac{C(t)}{K}\right]}_{\textbf{contact inhibition}}, \tag{2.1}$$

where $C(t) > 0$ is the cell density at time t, $\lambda > 0$ is the proliferation rate, and $K > 0$ is the carrying capacity density. The carrying capacity density is the density at which contact inhibition decreases the net growth rate to zero. The logistic growth model is used ubiquitously in the study of development, repair and tissue regeneration, including for the modeling of tumor growth (Enderling and Chaplain, 2014; Gerlee, 2013; Scott et al., 2013; Treloar et al., 2013) and wound healing (Maini et al., 2004b; Sherratt and Murray, 1990; Simpson et al., 2013). The

solution of Equation (2.1),

$$C(t) = \frac{C(0)K}{[(K - C(0))\exp(-\lambda t) + C(0)]}, \tag{2.2}$$

is a sigmoid curve that increases from $C(t) = C(0)$ to $C(t) = K$ as $t \to \infty$, provided that $C(0)/K \ll 1$.

In vitro cell proliferation assays are used routinely to examine mechanisms that control cell proliferation, such as the application of various drugs and other treatments on the rate of cell proliferation (Chen et al., 2017; Delarue et al., 2014). *In vitro* assays are routinely used to inform the development and interpretation of *in vivo* assays describing pathological situations, such as tumor growth (Beaumont et al., 2014). Therefore, improving the design and interpretation of *in vitro* assays will have an indirect influence on the way that we design and interpret *in vivo* assays.

A cell proliferation assay typically involves placing cells, at low density, $C(0)/K \ll 1$, onto a two-dimensional surface and re-examining the increased monolayer density at a later time, $t = T$. Typical data from a cell proliferation assay are given in Figure 2.1(A)-(B).

To use Equation (2.1) to quantitatively inform our understanding of a particular biological system, we must be able to estimate the initial density, $C(0)$, and the model parameters, λ and K. Obtaining an accurate estimate of K is crucial to understand how contact inhibition controls the net proliferation rate at modest to high densities. However, despite the importance of K, most theoretical studies work with a non-dimensionalised model by setting $c(t) = C(t)/K$. This leads to (Maini et al., 2004a,b; Sherratt and Murray, 1990)

$$\frac{\mathrm{d}c(t)}{\mathrm{d}t} = \lambda c(t)\left[1 - c(t)\right],$$

which completely masks the importance of being able to accurately estimate the carrying capacity, K.

Not only do standard mathematical approaches prevent quantitative assessment of the impact of contact inhibition, in addition, standard experimental protocols for *in vitro* proliferation assays are also insufficient to estimate K. Very recently, Jin et al. (2017) used Equation (2.1) to analyse a set of cell proliferation assays performed with a prostate cancer cell line, and concluded that

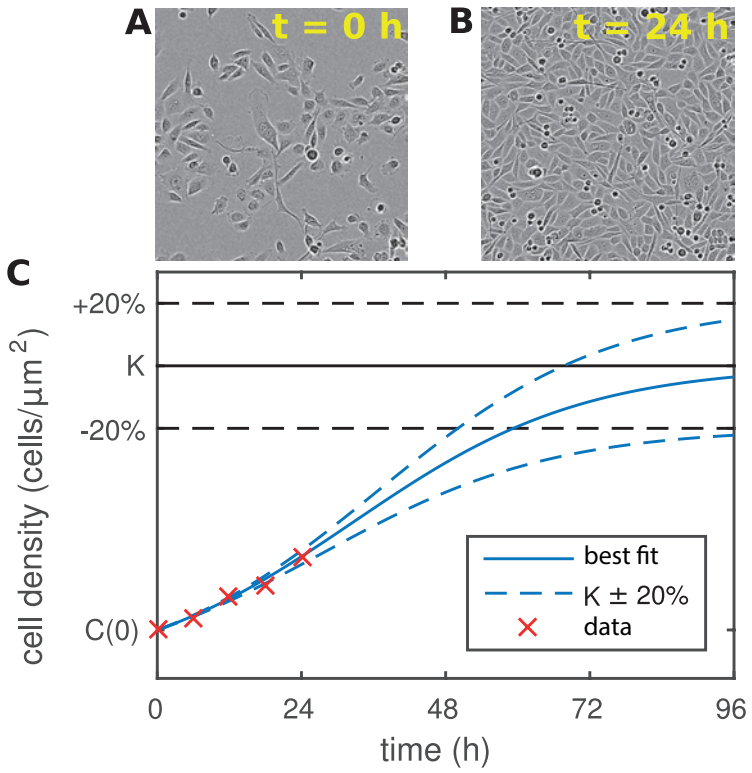

Figure 2.1: Proliferation assay using a prostate cancer cell line. Images of area 450 μm^2 are captured at (A) $t = 0$ h and (B) $t = 24$ h. Images are reproduced from Jin et al. (2017) with permission. (C) Logistic growth curve with $C(0) = 3.1 \times 10^{-4}$ (cells/μm^2), $\lambda = 0.052$ (1/h) and $K = 2 \times 10^{-3}$ (cells/μm^2) (solid black). Additional solutions with $K \pm 20\%$ (dashed black) also fit the short time experimental data (red crosses).

standard experimental data do not lead to robust, biologically relevant estimates of K. This is consistent with the work of Sarapata and de Pillis (2014) who also find that standard *in vitro* experimental data are insufficient to estimate K using best-fit, non-linear least-squares methods (Sarapata and de Pillis, 2014). The fundamental issue, as illustrated in Figure 2.1(C), is that cell proliferation assays are initiated with a small density and performed for a relatively short duration. This strategy is sufficient for estimating the low-density intrinsic proliferation rate, λ, but completely inadequate for estimating K which is associated with longer time, higher density data. Given the importance of K, we are motivated to re-consider the design of proliferation assays so that we can quantitatively estimate both λ and K from a single experiment.

The typical duration of a proliferation assay is less than 24 h (Chen et al., 2017), with some assays as short as 4 h (Huang et al., 2017), and the cell density is recorded once, at the end point of the experiment. This timescale is sufficient to estimate the low-density intrinsic proliferation rate, λ, since the doubling time of the majority of cell lines is approximately 24 hours (Maini

et al., 2004a,b; Treloar et al., 2013; Simpson et al., 2013). However, this standard timescale is far too short to robustly quantify contact inhibition effects given the low initial densities that are routinely used. For example, the PC-3 prostate cancer cell line examined by Jin et al. (2017) proliferates with $\lambda \approx 0.05$ /h, but on the timescale of the experiment the observed growth of the population is approximately exponential. This is consistent with Equation (2.1) since we have $C(t) \sim C(0)\exp(\lambda t)$ provided that $C(0)/K \ll 1$ and t is sufficiently small. Figure 2.1(C) shows that the early time growth dynamics are effectively independent of K, confirming that it is impossible to obtain robust estimates of K using standard data (Jin et al., 2017; Sarapata and de Pillis, 2014).

The aim of this work is to provide guidance about how to overcome these standard experimental and theoretical limitations by taking a Bayesian approach to experimental design (Gelman et al., 2014; Liepe et al., 2013; Silk et al., 2014; Vanlier et al., 2012, 2014; Browning et al., 2017). The advantage of taking a Bayesian approach is that we have a platform to quantitatively examine how the uncertainty in our estimate of K depends on the experimental design: we aim to provide guidance for experimental design that minimises the uncertainty in our estimate of K. To achieve this we consider a proliferation assay of duration T, with n observations of cell density, $C_{\text{obs}}^{1:n} = [C_{\text{obs}}(t_1), \ldots, C_{\text{obs}}(t_n)]$, at times $t_1, \ldots, t_n$ with $0 < t_i \leq T$ for all $i = [1, \ldots, n]$. These data could represent a single experiment observed at multiple time points, $t_1 < t_2 < \cdots < t_n \leq T$, or n identically prepared experiments, each of which is observed once, $t_1 = t_2 = \cdots = t_n = T$. We assume that cells proliferate according to Equation (2.1) with known λ and $C(0)$. Furthermore, we assume $C(0)/K \ll 1$. Our assumption that λ can be determined is reasonable since $C(t) \sim C(0)\exp(\lambda t) = C(0)[1 + \lambda t + \mathcal{O}(t^2)]$ for $C(0)/K \ll 1$. This means that fitting a simple straight line or exponential curve to typical experimental data will provide a reasonable estimate of λ. The assumption that $C(0)$ is known precisely is less realistic. For example, estimates of $C(0)$ are affected profoundly by fluctuations in the initialization of the experiment as proliferation assays are performed by placing a known number of cells onto a tissue culture plate. However, images of the experiments are obtained over a much smaller spatial scale. This means that the role of stochastic fluctuations can be significant (Jin et al., 2017).

2.2 Uncertainty quantification in estimation of carrying capacity

To make progress we assume that observations are made, and are subject to Gaussian-distributed experimental measurement error with zero mean and variance Σ^2. Under these conditions our knowledge of the carrying capacity, K, given such observations is represented by the probability density function

$$p\left(K \,\middle|\, C_{\text{obs}}^{1:n}\right) = A \prod_{i=1}^{n} \phi\left(C_{\text{obs}}(t_i); C(t_i), \Sigma^2\right), \tag{2.3}$$

where A is a normalization constant and $\phi\left(\cdot; C(t_i), \Sigma^2\right)$ denotes a Gaussian probability density with mean $C(t_i)$ and variance Σ^2 (Section 2.5.1). This probability density represents knowledge obtained from the data when no prior assumptions are made on K. The point of maximum density in Equation (2.3) corresponds to the maximum likelihood estimator or best-fit estimate, $\hat{K}$. Importantly, Equation (2.3) also enables the quantification of uncertainty in this estimate through calculation of the variance,

$$\sigma_n^2 = \int_0^\infty \left(K - \hat{K}\right)^2 p\left(K \,\middle|\, C_{\text{obs}}^{1:n}\right) \, \mathrm{d}K.$$

Figure 2.2(A) shows the probability density of K (Equation (2.3)) for several values of T for the typical assay protocol where only a single observation is made ($n = 1$). Here, our estimates of λ, $C(0)$ and Σ are based on reported values (Jin et al., 2017). The spread of these curves indicates the degree of uncertainty in any estimate of K. In particular, note the red line, which indicates the probability density for K for a measurement taken at the standard duration of $T = 24$ h. The relatively flat, disperse nature of the profile confirms that standard proliferation assay designs are completely inappropriate to estimate K since the profile lacks a well-defined maximum. This result provides a formal explanation for the observations of both Sarapata and de Pillis (2014) and Jin et al. (2017). In response to this issue, here we provide quantitative guidelines about how the experimental design can be chosen to facilitate accurate quantification of the effects of contact inhibition.

One optimistic assumption in Equation (2.3) is the supposition that $C(0)$ is known precisely. In reality, $C(0)$ is subject to both measurement errors and systematic errors owing to stochastic fluctuations (Jin et al., 2017). We extend our analysis to incorporate uncertainty in the estimate

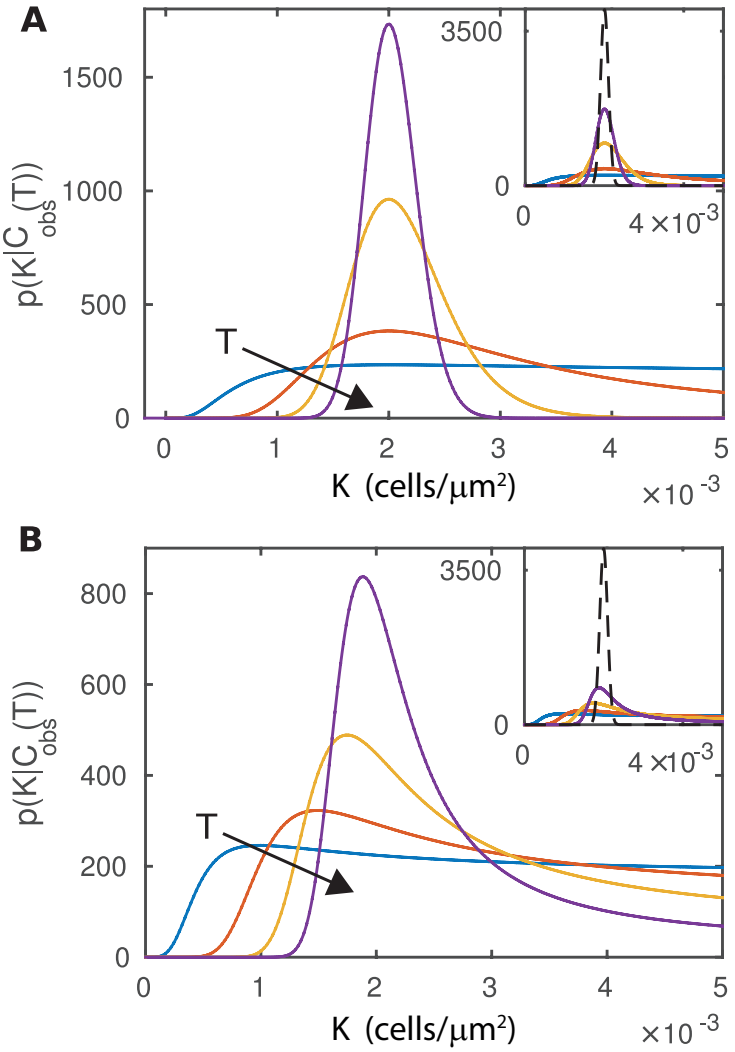

Figure 2.2: The probability density, $p\left(K \mid C_{\text{obs}}^{1:n}\right)$, with $n = 1$ plotted for $T = 12$ h (blue), 24 h (red), 36 h (yellow) and 48 h (purple), where $\lambda = 5.2 \times 10^{-2}$ (1/h), $\sigma = 10^{-4}$ (cells/μm^2) and the the true carrying capacity is $K = 2 \times 10^{-3}$ (cells/μm^2). (A) The initial density, $C(0)$, is assumed to be known precisely, $C(0) = 3.1 \times 10^{-4}$ (cells/μm^2). (B) Including uncertainty in the initial condition with $\mu_0 = 3.1 \times 10^{-4}$ (cells/μm^2) and $\Sigma_0 = 1.02 \times 10^{-4}$ (cells/μm^2). The inset panels in both (A) and (B) show the main plot in the context of the limiting density as $T \to \infty$ (black dashed line)

.

of $C(0)$ by also assuming $C(0)$ to be Gaussian-distributed with mean μ_0 and variance Σ_0^2, that is $p(C(0)) = \phi\left(C(0); \mu_0, \Sigma_0^2\right)$. In this case Equation (2.3) generalises to (Section 2.5.1)

$$p\left(K \,\middle|\, C_{\text{obs}}^{1:n}, \mu_0, \Sigma_0^2\right) = \int_0^\infty p\left(K \,\middle|\, C_{\text{obs}}^{1:n}\right) p(C(0))\, \mathrm{d}C(0). \tag{2.4}$$

The integral in Equation (2.4) is intractable, so numerical integration is required to evaluate $p\left(K \mid C_{\text{obs}}^{1:n}, \mu_0, \Sigma_0^2\right)$.

Since $C(t_i) \to K$ as $t_i \to \infty$ for all $i = 1, 2, \ldots, n$, then, for both Equation (2.3) and Equation (2.4), we obtain (Section 2.5.2)

$$\lim_{(t_1,\ldots,t_n)\to(\infty,\ldots,\infty)} p\left(K \,\middle|\, C_{\text{obs}}^{1:n}\right) = \phi\left(K; \frac{1}{n}\sum_{i=1}^{n} C_{\text{obs}}^{i}, \frac{\Sigma^2}{n}\right). \tag{2.5}$$

Figure 2.2(A) demonstrates that the probability density function for K, Equation (2.3), tightens toward the limiting density, Equation (2.5), as T increases. Note that for the typical assay duration, $T < 24$ h, the density is approximately uniform. Again, this reiterates the fact that standard experimental designs are inadequate to characterise the effects of contact inhibition, and that different approaches are required. The effect of uncertainty in $C(0)$ is clear in Figure 2.2(B). Even with $T = 48$ h there is a very large region of non-zero, near-constant probability density, indicating very wide confidence intervals for the estimate of K. Equation (2.5) also provides a lower bound on the uncertainty in the estimate of K,

$$\sigma_n^2 \geq \frac{\Sigma^2}{n}, \tag{2.6}$$

for any choice of $C(0)$ and λ. This result is independent of the treatment of the uncertainty in $C(0)$, but requires observations to be made after an infinite duration of time.

2.3 Experimental design challenges and resolutions

Equation (2.5) and Equation (2.6) tell us two important things about assay design in the study of contact inhibition. First, there is a fundamental lower bound on the uncertainty in our estimate of K for a fixed number, n, of observations of the density. However, increasing the assay duration, T, always provides more information. Second, increasing the number of observations, n, always provides more information and, further, it decreases the long time lower bound on the uncertainty in the estimate of K. Hence, Equation (2.6) informs the minimum number of observations required to estimate K accurately.

Clearly, practical experimental designs require finite T, and so we require methods to determine T such that the uncertainty in K is sufficiently close to the lower bound. We can quantify, for the standard choice of $n = 1$, the approximate uncertainty in our estimate of K (Section 2.5.3) (Ang

and Tang, 2007), given by

$$\sigma_1^2 = \frac{C(0)^4(1-\exp(-\lambda T))^2}{[C(0)-C_{\text{obs}}(T)\exp(-\lambda T)]^4}\Sigma^2. \tag{2.7}$$

This estimate of the uncertainty in K is accurate provided $C_{\text{obs}}(T)/\Sigma^2 \gg 1$. This expression provides a practical tool to assess the information content of data, but also enables one to estimate whether T is large enough that the uncertainty in K is sufficiently close to the lower bound (Section 2.5.3).

Equation (2.6) gives us a method of quantifying the information gained by introducing more observations in the idealised case that T is sufficiently large. For example, doubling n will half the uncertainty. However, since practical limitations mean that T is finite, this result does not always hold. Figure 2.3(A) shows that the effect of doubling n varies significantly depending on the choice of T. Here, the n observations are taken at regular intervals. For example, if $T < 1/\lambda$, increasing n has almost no effect on the uncertainty, σ_n^2. There is a similar negligible effect for $T > -(1/\lambda)\log[C(0)/(K-C(0))]$, which is the time corresponding to the point of inflection of Equation (2.2).

These results highlight several subtle, but immensely important considerations. For example, at sufficiently short times increasing n has very little benefit. Similarly, at sufficiently large times increasing T has little benefit. The most useful result is that for intermediate times there is more value in increasing T than increasing n. Furthermore, if we wish to go beyond the standard experimental design where a single measurement is made at the end of the experiment, $t = T$, we might want to quantify the benefit of making a second observation at an earlier time, $t < T$, during the same experiment. In this scenario, results in Figure 2.3(B) show that the choice of time at which the second measurement is made can be important. Comparing Figure 2.3(A) with Figure 2.3(B) shows that a poor choice of the time for the second observation might not lead to any change than taking a single observation at T. However, selecting two well-placed observation times can be as informative as making four equally spaced observations.

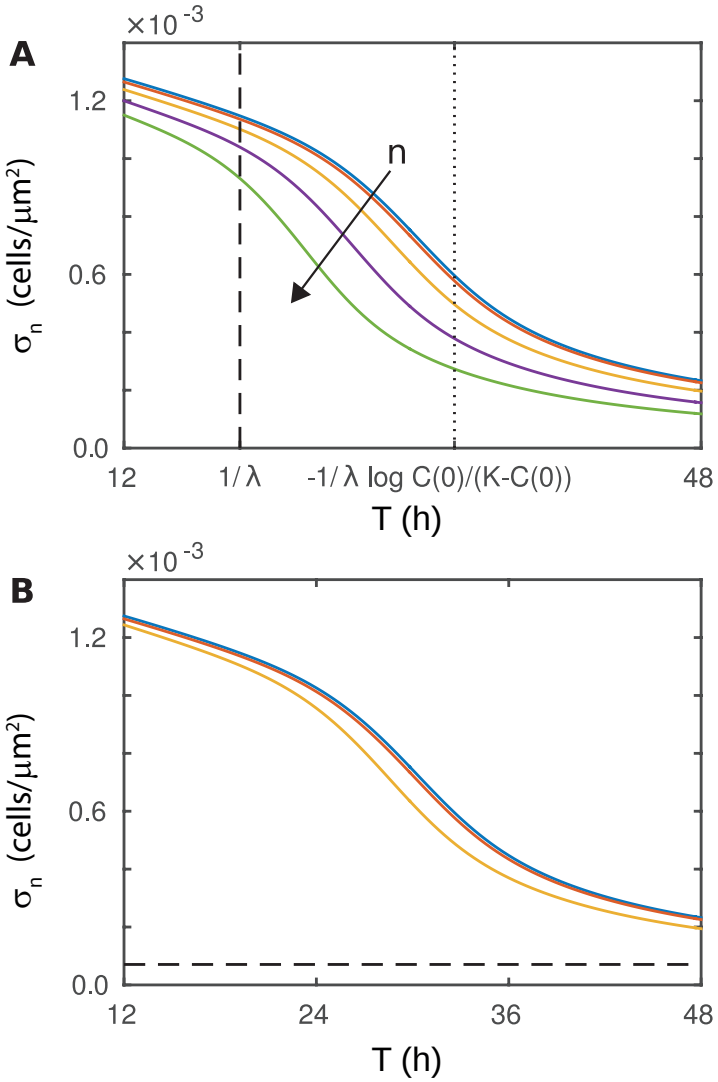

Figure 2.3: (A) Uncertainty in K as a function of T for $n = 1$ (blue), $n = 2$ (red), $n = 4$ (yellow), $n = 8$ (purple) and $n = 16$ (green); here observations are taken at regular intervals, that is $t_i = Ti/n$ for $i = 1, \ldots, n$. (B) Effect of observation placement for $n = 2$ with $t_2 = T$; $t_1 = T/5$ (blue), $t_1 = T/2$ (red) and $t_1 = 4T/5$ (yellow). The lower bound on uncertainty, σ_n, for $n = 2$ (dashed). The uncertainty, σ_n, is calculated using the trapezoid rule with 10^5 equally spaced panels over the interval $0 < K < 5 \times 10^{-3}$ (cells/μm^2).

We conclude by providing several guidelines for the design of cell proliferation assays:

1. Reducing the uncertainty in $C(0)$ is crucial, especially if $T < 48$ h;

2. Equation (2.7) should be used with short timescale data to estimate the smallest value of T that will result in acceptable uncertainty in K;

3. On a short timescale, increasing T is more informative than increasing n. However, if increasing T is infeasible, the optimal strategy is to repeat the experiment n times and make n observations at time T. In many situations, n will need to be large to account

for the short timescale. Figure 2.3(A) shows that even a 16-fold increase in n is still unacceptable for $T < 12$ as it is effectively still the same as a single observation.

2.4 Conclusion

In this work we highlight certain aspects of assay design that are often neglected and unreported. However, these features are critical if we are to quantify the role of crowding and contact inhibition of proliferation in populations of cells. The guidelines we propose allow us to provide the best estimates of λ and K using a single experiment, whereas standard experimental designs allow us to confidently estimate λ only. Our results confirm that standard *in vitro* experimental designs are well suited for estimating λ, but poorly suited for estimating K. A different approach is required to overcome this limitation, and here we provide quantitative guidance about designing *in vitro* proliferation assays that can be used to estimate both λ and K. The situation is even more complex for *in vivo* assays in which there are more experimental constraints and unknowns. However, improving the way that we design and interpret *in vitro* assays is relevant because these simpler experiments are routinely used in tandem with *in vivo* assays due to the fact that they are cheaper and faster to perform than working in live tissues.

2.5 Supplementary material

2.5.1 Posterior probability density functions

In Bayesian statistics, knowledge of unobserved model parameters, $\boldsymbol{\theta}$, given the results of some experiment that results in data, $\mathcal{D}$, is represented through a probability density function (PDF) (Gelman et al., 2014). This PDF, $p(\boldsymbol{\theta} \mid \mathcal{D})$, is called the *posterior* and can be interpreted as "the probability density of $\boldsymbol{\theta}$ given observation $\mathcal{D}$". The posterior is derived through Bayes' Theorem,

$$p(\boldsymbol{\theta} \mid \mathcal{D}) = \frac{\mathcal{L}(\boldsymbol{\theta}; \mathcal{D}) p(\boldsymbol{\theta})}{p(\mathcal{D})}, \tag{2.8}$$

where the *prior* PDF, $p(\boldsymbol{\theta})$, represents knowledge before the experiment, the *likelihood* function, $\mathcal{L}(\boldsymbol{\theta}; \mathcal{D})$, determines the probability density of the experimental results, $\mathcal{D}$, for a given set of parameter values and the *evidence*, $p(\mathcal{D})$, is the likelihood taken across all parameter values. In effect, $\mathcal{L}(\boldsymbol{\theta}; \mathcal{D})$ encodes assumptions of the model, $p(\boldsymbol{\theta})$ encodes the assumptions on the parameters and $p(\mathcal{D})$ acts as a normalization constant. Our task is to derive the posterior PDF for the carrying capacity density, K, given noisy observations of an assumed logistic growth curve.

First we derive the likelihood. We represent the process of taking a cell density measurement, $C_{\text{obs}}(t)$, at time, t, as a Gaussian random variable with mean around the true cell density, $C(t)$ (Equation (2.2)), and variance σ^2. Therefore, the probability density of a single observation is a Gaussian PDF,

$$\begin{aligned} \phi(C_{\text{obs}}(t); C(t), \sigma^2) &= \frac{1}{\sigma\sqrt{2\pi}} e^{-(C_{\text{obs}}(t) - C(t))^2/(2\sigma^2)} \\ &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(C_{\text{obs}}(t) - \frac{C(0)K}{(K - C(0))e^{-\lambda t} + C(0)}\right)^2/(2\sigma^2)}. \end{aligned} \tag{2.9}$$

Assuming $C(0)$ and λ are known, the exact logistic growth curve is fully determined for any proposed value of K. As a result, the observations made at time t_1 and t_2, $C_{\text{obs}}(t_1)$ and $C_{\text{obs}}(t_2)$, are independent given this value of K. Therefore the likelihood function for n observations,

$C_{\text{obs}}^{1:n} = [C_{\text{obs}}(t_1), \ldots, C_{\text{obs}}(t_n)]$, is given by

$$\mathcal{L}(K; C_{\text{obs}}^{1:n}) = \prod_{i=1}^{n} \phi(C_{\text{obs}}(t_i); C(t_i), \sigma^2), \tag{2.10}$$

where $\phi(C_{\text{obs}}(t_i); C(t_i), \sigma^2)$ is given by Equation (2.9).

For the prior, we assume $0 \leq K \leq K_{\text{max}}$ with equal probability for some $K_{\text{max}} < \infty$, that is K is uniformly distributed on the interval $(0, K_{\text{max}})$. The PDF is

$$p(K) = \begin{cases} \dfrac{1}{K_{\text{max}}}, & \text{if } K \in [0, K_{\text{max}}], \\ 0, & \text{otherwise.} \end{cases} \tag{2.11}$$

This uniform prior distribution imposes minimal assumptions, as we impose no prior knowledge about the value of K outside of some upper bound that could be arbitrarily large. For uninformative priors, the modes of the posterior correspond to the maximum likelihood estimator (MLE).

The evidence acts as a normalizing constant to ensure the product of Equation (2.10) and Equation (2.11) is a true PDF. Therefore, we have

$$\begin{aligned} p(C_{\text{obs}}^{1:n}) &= \int_{-\infty}^{\infty} \mathcal{L}(K; C_{\text{obs}}^{1:n}) p(K) \, \mathrm{d}K \\ &= \frac{1}{K_{\text{max}}} \int_{0}^{K_{\text{max}}} \prod_{i=1}^{n} \phi(C_{\text{obs}}(t_i); C(t_i), \sigma^2) \, \mathrm{d}K, \end{aligned} \tag{2.12}$$

which converges since Equation (2.9) is bounded and continuous on the closed interval $K \in [0, K_{\text{max}}]$.

Finally, we arrive at the posterior through substitution of Equation (2.10), Equation (2.11), and Equation (2.12) into Bayes' Theorem, Equation (2.8),

$$p(K \mid C_{\text{obs}}^{1:n}) = \begin{cases} A \prod_{i=1}^{n} \phi(C_{\text{obs}}(t_i); C(t_i), \sigma^2), & \text{if } K \in [0, K_{\text{max}}], \\ 0, & \text{otherwise.} \end{cases} \tag{2.13}$$

where

$$\frac{1}{A} = \int_0^{K_{\max}} \prod_{i=1}^{n} \phi(C_{\text{obs}}(t_i); C(t_i), \sigma^2)\, \mathrm{d}K, \tag{2.14}$$

which is Equation (2.3).

Equation (2.13) assumes $C(0)$ is fixed. To extend the model to capture uncertainty in $C(0)$ we define $C(0)$ to be a Gaussian random variable with mean μ_0 and variance σ_0^2. The PDF is $p(C(0)) = \phi(C(0); \mu_0, \sigma_0^2)$. In this case, the logistic growth curve, forming the means of the observations, depends on this random variable. Therefore, consider the joint distribution

$$\begin{aligned} p(K, C(0) \mid C_{\text{obs}}^{1:n}, \mu_0, \sigma_0^2) &= p(K \mid C(0), C_{\text{obs}}^{1:n}) p(C(0) \mid C_{\text{obs}}^{1:n}, \mu_0, \sigma_0^2) \\ &= p(K \mid C(0), C_{\text{obs}}^{1:n}) \phi(C(0); \mu_0, \sigma_0^2), \end{aligned} \tag{2.15}$$

where $p(K \mid C(0), C_{\text{obs}}^{1:n})$ is simply Equation (2.13) with the dependence on $C(0)$ made explicit. The desired posterior is a marginal density of Equation (2.15), that is,

$$p(K \mid C_{\text{obs}}^{1:n}, \mu_0, \sigma_0^2) = \int_{-\infty}^{\infty} p(K, C(0) \mid C_{\text{obs}}^{1:n}, \mu_0, \sigma_0^2)\, \mathrm{d}C(0). \tag{2.16}$$

Substitution of Equation (2.15) into Equation (2.16) results in Equation (2.4). If expanded, the posterior is

$$p(K \mid C_{\text{obs}}^{1:n}, \mu_0, \sigma_0^2) = A \int_0^{K_{\max}} \phi(C(0); \mu_0, \sigma_0^2) \prod_{i=1}^{n} \phi(C_{\text{obs}}(t_i); C(t_i), \sigma^2)\, \mathrm{d}C(0),$$

for $K \in [0, K_{\max}]$ and $p(K \mid C_{\text{obs}}^{1:n}, \mu_0, \sigma_0^2) = 0$ otherwise.

2.5.2 Limiting uncertainty

We now consider the posteriors Equation (2.13) and Equation (2.16) in the limit as the observation times become infinitely large, that is $t_i \to \infty$ for $i \in [1, \ldots, n]$. First, note that since $\lim_{t_i \to \infty} C(t_i) = K$, it is also true that,

$$\lim_{t_i \to \infty} \phi(C_{\text{obs}}(t_i); C(t_i), \sigma^2) = \phi(C_{\text{obs}}^i; K, \sigma^2). \tag{2.17}$$

That is, all observations are independent, identically distributed Gaussian random variables with mean K and variance σ^2 in the limit. Since $\phi(C^i_{\text{obs}}; K, \sigma^2) = \phi(K; C^i_{\text{obs}}, \sigma^2)$, then we extend the domain of K to $K \in (-\infty, \infty)$ and obtain

$$\lim_{(t_1,\ldots,t_n)\to(\infty,\ldots,\infty)} p(K \mid C^{1:n}_{\text{obs}}) = \frac{\prod_{i=1}^n \phi(K; C^i_{\text{obs}}, \sigma^2)}{\int_{-\infty}^{\infty} \prod_{i=1}^n \phi(K; C^i_{\text{obs}}, \sigma^2)\,\mathrm{d}K}. \tag{2.18}$$

Note that for any two Gaussian PDFs $\phi(X; \mu_1, \sigma_1)$ and $\phi(X; \mu_2, \sigma_2)$ it can be shown that

$$\phi(X; \mu_1, \sigma_1)\phi(X; \mu_2, \sigma_2) \propto \phi\left(X; \frac{\mu_1\sigma_2^2 + \mu_2\sigma_1^2}{\sigma_1^2 + \sigma_2^2}, \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}\right). \tag{2.19}$$

Through some tedious algebraic manipulations, we apply Equation (2.19) to Equation (2.18) to obtain

$$\lim_{(t_1,\ldots,t_n)\to(\infty,\ldots,\infty)} p(K \mid C^{1:n}_{\text{obs}}) = \phi\left(K; \frac{1}{n}\sum_{i=1}^n C^i_{\text{obs}}, \frac{\sigma^2}{n}\right). \tag{2.20}$$

Note that equality holds in Equation (2.20) because the constants of proportionality in Equation (2.19) cancel out through Equation (2.18).

Similar steps are applied to Equation (2.16) to obtain

$$\begin{aligned}
\lim_{(t_1,\ldots,t_n)\to(\infty,\ldots,\infty)} p(K \mid C^{1:n}_{\text{obs}}, \mu_0, \sigma_0^2) &= \int_{-\infty}^{\infty} \phi\left(K; \frac{1}{n}\sum_{i=1}^n C^i_{\text{obs}}, \frac{\sigma^2}{n}\right) p(C(0))\,\mathrm{d}C(0) \\
&= \phi\left(K; \frac{1}{n}\sum_{i=1}^n C^i_{\text{obs}}, \frac{\sigma^2}{n}\right) \int_{-\infty}^{\infty} p(C(0))\,\mathrm{d}C(0) \\
&= \phi\left(K; \frac{1}{n}\sum_{i=1}^n C^i_{\text{obs}}, \frac{\sigma^2}{n}\right) \times 1 \\
&= \lim_{(t_1,\ldots,t_n)\to(\infty,\ldots,\infty)} p(K \mid C^{1:n}_{\text{obs}}).
\end{aligned} \tag{2.21}$$

This results in Equation (2.5).

It is important to note that in the derivation of Equation (2.20) and Equation (2.21) we have extended the domain of K to be the entire real number line. The prior in this case is actually improper, thus, for $K \in (-\infty, \infty)$, the posteriors Equation (2.13) and Equation (2.16) are not guaranteed to be PDFs. However, in the limiting case when $t_i \to \infty$ for $i \in [1, \ldots, n]$, the result is a PDF. Therefore, this extension is justified so long as $K/\sigma \gg 1$. If this is not satisfied, then this analysis could be completed using a truncated Gaussian prior.

2.5.3 Short time uncertainty quantification

Here we derive Equation (2.7). The approximation uses the delta method to give an approximate variance for Equation (2.13), that is the uncertainty in K for $T < \infty$ and $n = 1$.

The posterior for $n = 1$ is

$$p(K \mid C_{\text{obs}}(T)) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(C_{\text{obs}}(T)-C(T))^2/(2\sigma^2)}. \tag{2.22}$$

Thus, we can treat $C(T)$ as a Gaussian random variable with mean $C_{\text{obs}}(T)$ and variance σ^2. Given a realization of $C(T)$ we can obtain from the logistic growth solution (Equation (2.2))

$$K = f(C(T)) = \frac{C(T)C(0)(1 - e^{-\lambda T})}{C(0) - C(T)e^{-\lambda T}}. \tag{2.23}$$

The delta method is an approximation technique for obtaining approximate moments of a distribution of a random variable that is a function of another (Ang and Tang, 2007). The idea is to consider the Taylor expansion about the mean of $C(T)$,

$$f(C(T)) = f(C_{\text{obs}}(T) + \delta) = \sum_{k=0}^{\infty} \frac{\delta^k}{k!} \frac{\mathrm{d}^k f}{\mathrm{d}C^k}[C_{\text{obs}}(T)].$$

From this expansion we obtain

$$\mathbb{V}[K] = \mathbb{V}[f(C(T))] = \mathbb{V}\left[\sum_{k=0}^{\infty} \frac{\delta^k}{k!} \frac{\mathrm{d}^k f}{\mathrm{d}C^k}[C_{\text{obs}}(T)]\right].$$

For small δ we have

$$\mathbb{V}[K] = \mathbb{V}\left[f(C_{\text{obs}}(T)) + \delta \frac{\mathrm{d}f}{\mathrm{d}C}[C_{\text{obs}}(T)] + \mathcal{O}(\delta^2)\right].$$

Recall that $\mathbb{V}[a] = 0$, $\mathbb{V}[X + a] = \mathbb{V}[X]$ and $\mathbb{V}[aX] = a^2\mathbb{V}[X]$ for any random variable, X,

and constant a. Therefore,

$$\mathbb{V}\left[f(C_{\text{obs}}(T)) + \delta\frac{\mathrm{d}f}{\mathrm{d}C}[C_{\text{obs}}(T)] + \mathcal{O}(\delta^2)\right] = \mathbb{V}\left[\delta\frac{\mathrm{d}f}{\mathrm{d}C}[C_{\text{obs}}(T)] + \mathcal{O}(\delta^2)\right]$$
$$= \left(\frac{\mathrm{d}f}{\mathrm{d}C}[C_{\text{obs}}(T)]\right)^2 \mathbb{V}\left[C(T)\right] + \mathcal{O}(\mu_4), \quad (2.24)$$

where μ_4 is the fourth central moment of the distribution of $C(T)$. The approximation is appropriate provided $|\delta|$ is sufficiently small, that is $|C(T) - C_{\text{obs}}(T)|$ is sufficiently small. This is achieved if $C_{\text{obs}}(T)/\sigma \gg 1$. The derivative of f with respect to C is

$$\frac{\mathrm{d}f}{\mathrm{d}C} = \frac{C(0)^2(1-e^{-\lambda T})}{[C(0) - C(T)e^{-\lambda T}]^2}. \quad (2.25)$$

Therefore, by substitution of Equation (2.25) into Equation (2.24), we arrive at the approximation

$$\mathbb{V}\left[K\right] \approx \frac{C(0)^4(1-e^{-\lambda T})^2}{[C(0) - C_{\text{obs}}(T)e^{-\lambda T}]^4}\sigma^2. \quad (2.26)$$

There are two practical uses for Equation (2.26). First, given an population density observation, $C_{\text{obs}}(T)$, taken at time T, one can calculate

$$\epsilon = \frac{C(0)^4(1-e^{-\lambda T})^2}{[C(0) - C_{\text{obs}}(T)e^{-\lambda T}]^4}.$$

Here, ϵ provides a relative uncertainty compared to the limiting best case σ. Therefore, the test provides a measure of how close to optimal the assay is. Furthermore, if the observation error is known then this sample approach provides the uncertainty in K after the observation. Second, given the upper limit of the prior K_{max}, as in Equation (2.11), one may consider the function

$$h(T) = \frac{C(0)^4(1-e^{-\lambda T})^2}{[C(0) - K_{\text{max}}e^{-\lambda T}]^4},$$

to identify how large T must be to be close enough to the limiting posterior, i.e., $\epsilon \approx 1$. If this value of T is too large for practical purposes, then it indicates more observations should be taken. It is important to note that the approximation does require $C_{\text{obs}}(T)/\sigma \gg 1$. In practice, this may not hold for early time, $T \ll 1/\lambda$. In such a case, Equation (2.26) tends to be an *underestimate*. Because of this, the result is still useful to decide if increasing n is valuable or

not.

2.5.4 Proliferation data

Table 2.1 presents the data used to inform Figure 2.1, Figure 2.2 and Figure 2.3. The data are derived from Jin et al. (2017).

Table 2.1: Cell density data

time (h)	0	6	12	18	24
cell density (cells/μm^2)	3.1×10^{-4}	3.8×10^{-4}	5.2×10^{-4}	5.9×10^{-4}	7.8×10^{-4}

Chapter 3

Using Experimental Data and Information Criteria to Guide Model Selection for Reaction–Diffusion Problems in Mathematical Biology

A paper published in *Bulletin of Mathematical Biology*.

David J. Warne, Ruth E. Baker and Matthew J. Simpson (2019). Using experimental data and information criteria to guide model selection for reaction–diffusion problems in mathematical biology. *Bulletin of Mathematical Biology*, 81:1760–1804. DOI:10.1007/s11538-019-00589-x

Abstract Reaction–diffusion models describing the movement, reproduction and death of individuals within a population are key mathematical modelling tools with widespread applications in mathematical biology. A diverse range of such continuum models have been applied in various biological contexts by choosing different flux and source terms in the reaction–diffusion framework. For example, to describe the collective spreading of cell populations, the flux term may be chosen to reflect various movement mechanisms, such as random motion (diffusion), adhesion, haptotaxis, chemokinesis and chemotaxis. The choice of flux terms

in specific applications, such as wound healing, is usually made heuristically, and rarely is it tested quantitatively against detailed cell density data. More generally, in mathematical biology, the questions of model validation and model selection have not received the same attention as the questions of model development and model analysis. Many studies do not consider model validation or model selection, and those that do often base the selection of the model on residual error criteria after model calibration is performed using nonlinear regression techniques. In this work, we present a model selection case study, in the context of cell invasion, with a very detailed experimental data set. Using Bayesian analysis and information criteria, we demonstrate that model selection and model validation should account for both residual errors and model complexity. These considerations are often overlooked in the mathematical biology literature. The results we present here provide a straightforward methodology that can be used to guide model selection across a range of applications. Furthermore, the case study we present provides a clear example where neglecting the role of model complexity can give rise to misleading outcomes.

3.1 Introduction

The development and testing of new theories to explain observations are keystones of the scientific method. In the biological sciences, mathematical models have become increasingly important to develop and test various hypotheses about putative mechanisms that drive biological processes (Silk et al., 2014). Furthermore, the interpretation of new biological data and the development of new experimental protocols is increasingly being enhanced through the use of mathematical models to explore questions of optimal experimental design (Browning et al., 2017; Drovandi and Pettitt, 2013; Liepe et al., 2013; Ryan et al., 2016) (see Chapter 2). However, for many applications in mathematical biology, there is a diverse range of valid modelling approaches, and it is not always obvious which model is most appropriate for a particular application. Since all models are, by definition, a simplification of reality (Box, 1976), the choice of modelling approach strongly depends on the purpose of the model and the set of modelling tools available (Gelman et al., 2014).

An unresolved question in the field of mathematical biology is how to best determine when a model is sufficient. In many biological applications, there can be multiple competing theories

about the particular phenomenon that we might wish to study. In these situations, mathematical models can be used to objectively evaluate the validity of these theories by comparing model predictions with a set of experimental observations. Such an evaluation requires a robust, reproducible and objective framework for choosing the model, out of the set of competing candidates, that best explains the data. It is standard practice throughout the field of mathematical biology to judge the suitability of a model on its ability to match the experimental data after calibration using the maximum likelihood estimator (MLE) (Bianchi et al., 2016; Jin et al., 2016b; Johnston et al., 2016; Maini et al., 2004a,b; Sherratt and Murray, 1990). However, this approach is known to favour model complexity (Akaike, 1974). We suggest that it is preferable instead to favour simple models unless additional complexity is warranted. In our work, we investigate and demonstrate techniques to select models, such as Bayesian analysis and information criteria, and give a practical illustration of the trade-off between consistency, fitness and complexity. We focus on the question of model selection between different reaction–diffusion continuum models that describe collective cell motility and proliferation. However, reaction–diffusion models are also of broad interest to the mathematical biology community since they frequently arise in many contexts that involve population dynamics, such as ecology (King and McCabe, 2003; Sherratt, 2015, 2016; Skellam, 1951) and evolutionary biology (Cohen and Galiano, 2013; Fortelius et al., 2015).

3.1.1 Continuum models in cell biology applications

Continuum models of collective cell spreading and proliferation are used to study cancer, wound healing, embryonic development and tissue engineering (Edelstein-Keshet, 2005; Murray, 2002). However, the details of the models are diverse (Jin et al., 2016b; Simpson et al., 2011). The motility of cells can be modelled by: linear diffusion (Jackson et al., 2015; Maini et al., 2004a,b; Sengers et al., 2007; Simpson et al., 2007b; Swanson et al., 2003); nonlinear diffusion where the diffusivity increases with density (Bianchi et al., 2016; Flegg et al., 2009; Sengers et al., 2007), and nonlinear diffusion where the diffusivity decreases with density (Cai et al., 2007; King and McCabe, 2003); nonlinear advection to describe directed motility, such as chemotaxis (Bianchi et al., 2016; Flegg et al., 2010), haptotaxis (Marchant et al., 2001), and cell–cell adhesion (Armstrong et al., 2009). Similarly, the proliferation of cells is often modelled using a logistic source term (Maini et al., 2004a,b; Sengers et al., 2007; Simpson et al., 2007b), but

there are also many other options for modelling carrying capacity-limited proliferation (Browning et al., 2017; Gerlee, 2013; Tsoularis and Wallace, 2002). Furthermore, the motility and proliferation of cells may also be coupled to diffusing chemical factors (Bianchi et al., 2016; Nardini et al., 2016; Savla et al., 2004; Sherratt and Murray, 1990). Despite this diverse range of modelling possibilities, few studies in the mathematical biology literature evaluate model uncertainty (Liepe et al., 2013; Silk et al., 2014) (see Chapter 2) or model selection (Bianchi et al., 2016; Jin et al., 2016b; Sengers et al., 2007; Sherratt and Murray, 1990). Therefore, a relevant question for us to address is: given complex biological data with many sources of uncertainty, how can we select models that provide a balance between model simplicity and agreement with data?

3.1.2 Challenges for model calibration

The study of the temporal growth, and the spatiotemporal spreading of cell populations is a canonical area within the mathematical biology literature where there are many different types of continuum models available to interpret experimental data. For example, Sarapata and de Pillis (2014) catalog the most commonly used temporal growth models of tumors (Gerlee, 2013) and calibrate against data from both *in vitro* and *in vivo* assays. Interestingly, Sarapata and de Pillis (2014) note that some of the MLE solutions lead to nonphysical predictions of the carrying capacity density. Through application of a Bayesian approach, Chapter 2 demonstrates that parameter uncertainty in such temporal models may be used to inform experimental design of proliferation assays. Sarapata and de Pillis (2014) and Chapter 2 focus on temporal growth dynamics only, and they note that data obtained through standard experimental protocols may not contain sufficient information to resolve accurate and realistic estimates for all unknown parameters. Thus we expect that experimental design and model calibration requires detailed spatial data to compensate for increased model complexity when considering spatial models that describe the spatiotemporal spreading of cell populations. However, very few studies that calibrate spatial models of collective cell spreading use detailed cell density data, instead, secondary quantities, such as the location of the moving cell front, are often used (Maini et al., 2004a,b; Sherratt and Murray, 1990). In a more recent study, Jin et al. (2016b) use detailed cell density profiles to identify parameter estimates based on the MLE solution for two commonly

used models of collective cell spreading. The main point of Jin et al. (2016b) is to show that parameter estimates appear to depend strongly upon the initial density of cells in the experiments, and this result is at odds with our intuitive expectation because typical models implicitly assume that the parameters are constants. This indicates that a cautionary approach must be taken to reliably estimate parameters and compare models, and provide a partial explanation about why cell biology experiments are notoriously difficult to reproduce (Jin et al., 2016b).

The examples of Jin et al. (2016b) and Sarapata and de Pillis (2014) highlight a problem in model calibration and parameter estimation. That is, the MLE-based approach: 1) can lead to biologically unrealistic parameter estimates or model behaviour (Jin et al., 2016b; Sarapata and de Pillis, 2014; Slezak et al., 2010); 2) is biased towards complex models that essentially overfit through an overabundance of free parameters (Box, 1976; Gelman et al., 2014; Johnson and Omland, 2004; Stoica and Selen, 2004); and 3) fails to capture uncertainty (Gelman et al., 2014) (see Chapter 2). There are five key sources of uncertainty when applying a mathematical model to interpret experimental data: 1) unknown model parameters that require statistical estimation; 2) uncertainty in the choice of model; 3) uncertainty that arises from stochastic fluctuations in the system dynamics; 4) uncertainty introduced through numerical error arising from computational methods that are often required to simulate models or sample probability distributions; and 5) uncertainty resulting from systematic or measurement error in experimental work. Careful treatment of all of these sources of uncertainty is important to reliably validate theory and analyse data. Bayesian inference techniques are promising alternatives to MLE-based methods, since they can account for all relevant sources of uncertainty. Bayesian frameworks have been demonstrated to be highly effective at determining optimal experimental designs to minimise parameter uncertainty under the presence of systematic and measurement error (Browning et al., 2017; Johnston et al., 2016; Liepe et al., 2013; Parker et al., 2018; Vanlier et al., 2012) (see Chapter 2).

The Bayesian view, however, has its challenges, such as, potential subjectivity of inference and lack of a clear decision process (Berger, 2006; Consonni et al., 2018; Efron, 1986; Gelman, 2008a,b; Kass and Wasserman, 1996; Lambert, 2018). As a result, extensive statistical research to address these problems is an ongoing endeavour (Akaike, 1974; Gelman et al., 2014; Schwarz, 1978; Spiegelhalter et al., 2002), and there are now many alternative statistical techniques to traditional MLE-based methods or null hypothesis testing. Johnson and Omland

(2004) review a number of common approaches in the context of ecological and evolutionary research, and Gelman et al. (2014) provide an accessible discussion on model selection techniques from a statistical theory perspective.

3.1.3 Contribution

We demonstrate how to apply a Bayesian framework to quantitatively compare and select a reaction–diffusion model of collective cell behaviour using detailed *in vitro* assay data. We show that the Bayesian view of data-driven model selection provides significantly more insight than traditional MLE approaches. A family of continuum reaction–diffusion models that describe cell motility and cell proliferation are evaluated from a Bayesian perspective through parameter uncertainty quantification. This exercise reveals important aspects of reliable model selection that are often not easily identified otherwise. We also evaluate a number of widely used decision-making processes for model selection and compare the results with the intuition gained from the full Bayesian approach. In particular, we demonstrate, using detailed experimental data, that a trade-off between model fit, complexity and consistency can be obtained using a Bayesian framework. Furthermore, the model comparison techniques presented here are valid for a wide class of models and are relevant for model comparison in stochastic settings such as those considered by Johnston et al. (2016) and Matsiaka et al. (2019), although technical details of model simulation would change considerably. Since reaction–diffusion models are ubiquitous in mathematical biology (Edelstein-Keshet, 2005; Murray, 2002), we expect this work to be an exemplar of the Bayesian approach that is also broadly relevant to the wider mathematical biology community [1].

3.2 Cell culture protocols

In cell biology, *in vitro* cell culture assays are commonly used to measure and observe the behaviour of cell populations in different environments. Typical examples are proliferation assays (Browning et al., 2017), scratch assays (Jin et al., 2016b) and invasion assays (Haridas, 2017). In this work, we will focus on the scratch assay (Liang et al., 2007), however, it is

[1] Code available from GitHub https://github.com/ProfMJSimpson/Warne2019_BulletinofMathematicalBiology

important to note that our methods are widely applicable to other assay types. Jin et al. (2016b) provide a particularly detailed scratch assay data set using the PC-3 prostate cancer cell line. Each well in a 96-well tissue culture plate is identically populated with a particular initial number of PC-3 cells. Cells are left to attach to the substrate for a small amount of time so that uniform monolayers have formed. Then, identical scratches, approximately 0.5 mm in width, are made in the monolayers. Images of the scratched region are then captured at regular time intervals for a total duration of 48 hours. A particularly insightful feature of the protocol used by Jin et al. (2016b) is that they performed multiple experiments to demonstrate how variation in the initial density of cells in the monolayer affects cell invasion. Experiments were performed by initially placing either $10,000$, $12,000$, $14,000$, $16,000$, $18,000$ or $20,000$ cells into the wells of the 96-well plate. For brevity, in our study, we focus on data from the experiments initialised with $12,000$, $16,000$ and $20,000$ cells per well as representatives of, respectively, low, medium and high initial cell densities(see supplementary material of Jin et al. (2016b) to access data). Some example images from these data sets are summarised in Figure 3.1.

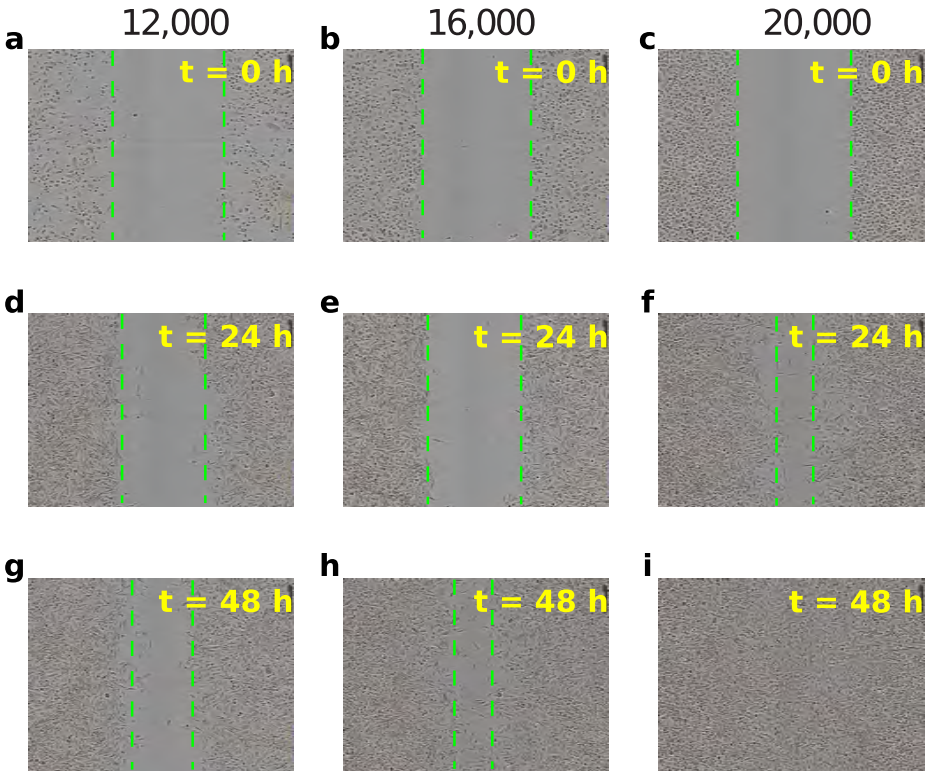

Figure 3.1: Example scratch assay data using the PC-3 prostate cancer cell line. Each column presents one experiment using an initial population of: (a),(d) and (g) $12,000$ cells; (b),(e) and (h) $16,000$ cells; (c),(f) and (i) and $20,000$ cells. The dimensions of each image are 1.43×1.97 [mm]. Images are shown at 24 [h] time intervals, however, the data is captured at 12 [h] intervals. Images are reproduced from Jin et al. (2016b), with permission.

The images in Figure 3.1 demonstrate that the scratch closure rate depends on the initial density of cells. For the low density initial condition of 12,000 cells per well, the scratch remains more than half its original size after 48 hours (Figure 3.1(a),(d) and (g)). In contrast, for the medium initial density of 16,000 cells per well, the scratch area is noticeably smaller at 48 hours, but it is still not closed (Figure 3.1(b),(e) and (g)). Only the high initial density initial condition of 20,000 cells per well leads to complete scratch closure after 48 hours (Figure 3.1(c),(f) and (i)). Most scratch assay protocols do not consider varying the initial density of cells, and those studies that use mathematical models to interpret experimental results from a scratch assay focus only on temporal data describing the position of the leading cell front (Sherratt and Murray, 1990; Maini et al., 2004a,b), indicated by the dashed green line in Figure 3.1(a)–(h). The study of Jin et al. (2016b) is unique since they provide detailed cell density profiles from three experimental replicates for each initial condition considered. These high resolution data enable us to pose and explore new questions about the applicability of some commonly used continuum reaction–diffusion models to describe this data set. We utilise a subset of the original PC-3 scratch assay data (see supplementary material of Jin et al. (2016b) to access data) to continue this line of reasoning using Bayesian analysis.

3.3 Continuum models of cell motility and proliferation

Fundamental mechanisms associated with cell invasion processes are the motility and proliferation of cells. Many different intracellular and intercellular mechanisms are relevant to both motility and proliferation, depending on the specific biological process. In the biological literature, it is not always clear which mechanisms are most important or relevant in a particular situation. Furthermore, it may be difficult to identify the most appropriate mathematical model, especially when a variety of models fit the experimental data both qualitatively and quantitatively. This is particularly true in the area of modelling of epidermal wound healing (Edelstein-Keshet, 2005; Maini et al., 2004a,b; Murray, 2002; Sherratt and Murray, 1990; Simpson et al., 2011).

3.3.1 Modelling cell populations with reaction–diffusion equations

Continuum models are routinely used to describe the evolution of a population of cells that undergo collective cell spreading and proliferation. Such models are often based on partial differential equations (PDEs) that are reaction–diffusion equations of the form

$$\frac{\partial C(\mathbf{x},t)}{\partial t} = -\nabla \cdot \mathbf{J}(\mathbf{x},t) + S(\mathbf{x},t). \tag{3.1}$$

Here, $C(\mathbf{x},t)$ is the cell density at position $\mathbf{x}$ and time t, $\mathbf{J}(\mathbf{x},t)$ is the cell population density flux vector and $S(\mathbf{x},t)$ is a source term representing cell proliferation and loss. Both $\mathbf{J}(\mathbf{x},t)$ and $S(\mathbf{x},t)$ are often functions of the cell density, $C(\mathbf{x},t)$, the cell density gradient vector, $\nabla C(\mathbf{x},t)$, or both; there is also usually a dependence on model parameters that must be calibrated using experimental data. The functional forms of $\mathbf{J}(\mathbf{x},t)$ and $S(\mathbf{x},t)$ used for modelling vary across diverse applications in tumor growth/invasion, wound healing, and embryology. However, there are some common choices. For the growth function, $S(\mathbf{x},t)$, Gerlee (2013) and Sarapata and de Pillis (2014) describe many of the key growth models in the context of tumor growth. A variety of options for the flux, $\mathbf{J}(\mathbf{x},t)$, are discussed by Murray (2002), and Simpson et al. (2006) perform simulations of these options and compare qualitatively with experimental data.

Of the available growth models, the most fundamental is the logistic growth model, $S(\mathbf{x},t) = \lambda C(\mathbf{x},t)\,(1 - C(\mathbf{x},t)/K)$, where $\lambda > 0$ is the rate of cell proliferation, $K > 0$ is the carrying capacity density, that is, the cell population density at which contact inhibition reduces the net population growth to zero. The logistic growth model is frequently used to describe cell growth as it is the most fundamental model that describes the effect of contact inhibition of proliferation (Chapter 2), however, generalisations of the logistic growth model have also be considered in cell biology applications (Browning et al., 2017; Sarapata and de Pillis, 2014; Tsoularis and Wallace, 2002). For the flux, $\mathbf{J}(\mathbf{x},t)$, the most common choice is Fickian diffusion (Maini et al., 2004a,b; Sherratt and Murray, 1990), $\mathbf{J}(\mathbf{x},t) = -D_0 \nabla C(\mathbf{x},t)$, where $D_0 > 0$ is a constant cell diffusivity. This formulation of the flux models cells for which motility is not affected by cell density, that is, cells are behaving like Brownian particles.

When logistic growth and Fickian diffusion are substituted into Equation (3.1) we obtain

$$\frac{\partial C(\mathbf{x},t)}{\partial t} = D_0 \nabla^2 C(\mathbf{x},t) + \lambda C(\mathbf{x},t)\left(1 - \frac{C(\mathbf{x},t)}{K}\right), \tag{3.2}$$

which, in one dimension, is known as the Fisher–Kolmogorov–Petrovsky–Piscounov model (Fisher–KPP) (Edelstein-Keshet, 2005; Murray, 2002). The Fisher–KPP model has been applied in many biological contexts (Edelstein-Keshet, 2005; Murray, 2002). In cell biology, applications include the modelling of wound healing (Maini et al., 2004a,b; Nardini et al., 2016; Savla et al., 2004), tissue engineering (Sengers et al., 2007), tumor growth (Swanson et al., 2002), cancer treatment (Jackson et al., 2015), and embryonic development (Simpson et al., 2007b).

A very common modification to the standard Fisher–KPP model is to incorporate density dependent diffusion of the form $\mathbf{J}(\mathbf{x},t) = -D(C(\mathbf{x},t))\nabla C(\mathbf{x},t)$, where $D(C(\mathbf{x},t))$ is the nonlinear diffusivity function. Often, $D(C(\mathbf{x},t))$ is chosen to be a monotonically increasing function with $D(0) = 0$ (Flegg et al., 2010; Simpson et al., 2011; Sengers et al., 2007). It is also intriguing that other studies choose to focus on monotonically decreasing nonlinear diffusivity functions (Cai et al., 2007). If $D(C(\mathbf{x},t)) = D_0 C(\mathbf{x},t)/K$, where D_0 is the cell diffusivity, we obtain the Porous Fisher model (Edelstein-Keshet, 2005; Gurney and Nisbet, 1975; Murray, 2002):

$$\frac{\partial C(\mathbf{x},t)}{\partial t} = D_0 \nabla \cdot \left(\frac{C(\mathbf{x},t)}{K}\nabla C(\mathbf{x},t)\right) + \lambda C(\mathbf{x},t)\left(1 - \frac{C(\mathbf{x},t)}{K}\right). \tag{3.3}$$

Unlike the Fisher–KPP model (Equation (3.2)), diffusion is not solely the result of the random movement of cells. Instead, cells exhibit movements that are directed away from crowded areas (Gurney and Nisbet, 1975), with a direct linear relationship between motility and density. Some studies have also considered $D(C(\mathbf{x},t)) = D_0(C(\mathbf{x},t)/K)^r$, which can be thought of as a Generalised Porous Fisher equation (Sherratt and Murray, 1990; Witelski, 1995),

$$\frac{\partial C(\mathbf{x},t)}{\partial t} = D_0 \nabla \cdot \left[\left(\frac{C(\mathbf{x},t)}{K}\right)^r \nabla C(\mathbf{x},t)\right] + \lambda C(\mathbf{x},t)\left(1 - \frac{C(\mathbf{x},t)}{K}\right), \tag{3.4}$$

where r is a constant that controls the density avoidance/attraction behaviour of cells. Here, the Fisher–KPP and Porous Fisher models are recovered with $r = 0$ and $r = 1$, respectively. In applications, r is often selected quite arbitrarily (Jin et al., 2016b; Sherratt and Murray, 1990) and there is little theory enabling its biological interpretation (Simpson et al., 2011).

The interpretation of the exponent, r, deserves further discussion. In effect, it models a nonlinear relationship between the motility of cells and the cell density (Simpson et al., 2011). For $r > 1$, the relationship is superlinear; the cell motility slowly increases with cell density at lower densities (Sherratt and Murray, 1990), but then rapidly increases at higher densities. For $0 < r < 1$, we have a sublinear relationship, that is, cells increase in motility faster at low densities (Jin et al., 2016b). Some have also considered the case of "fast diffusion" ($r < 0$) where cells become increasingly motile as the density decreases (King and McCabe, 2003).

For certain special choices of boundary conditions and initial conditions, the Fisher–KPP, Porous Fisher and Generalised Porous Fisher models are known to have travelling wave solutions. These solutions are of general mathematical interest and have been extensively studied (Harris, 2004; Witelski, 1995). However, since travelling waves only occur in the long-time limit and require rather special initial conditions, travelling waves are rarely observed experimentally (Jin et al., 2016b; Vittadello et al., 2018). Therefore, we do not consider connecting any kind of travelling wave solutions with experimental data in this work.

Different values of r also result in qualitatively different wave fronts (Edelstein-Keshet, 2005; Murray, 2002). Figure 3.2(c), (g), (k) and (o) compares the evolution of the Generalised Porous Fisher model in one spatial dimension for initial and boundary conditions known to lead to travelling waves; parameters are selected to correspond to similar wave speeds. The travelling wave of the Fisher–KPP model (Figure 3.2(c)) has no distinct leading edge, since $C(x,t) > 0$ for all x. However, the Porous Fisher model (Figure 3.2(k)) exhibits a distinct interface, sometimes called the contact point, separating regions of zero and nonzero density. This is the case for any $r > 0$. The shape of the wave front also changes with r. In particular, the wave front is concave upward in Figure 3.2(a)–(h) and concave downward in Figure 3.2(i)–(p) (see Section 3.7.1). Figure 3.2(a), (e), (i) and (m) are computed exactly (see Section 3.7.1) and all other solutions in Figure 3.2 are computed numerically (see Section 3.7.2).

As stated previously, formal travelling wave solutions are almost never observed experimentally. Typical scratch assays initial configurations lead to two opposingly-directed fronts, such as the profiles in Figure 3.2(d), (h), (l) and (p) showing the evolution of the Generalised Porous Fisher model for several values of r. Note that, for typical scratch widths, wound closure takes place before travelling waves have an opportunity to form (Jin et al., 2016b; Vittadello et al., 2018). However, the value of r still impacts the wound closure rate and the shape of the moving front.

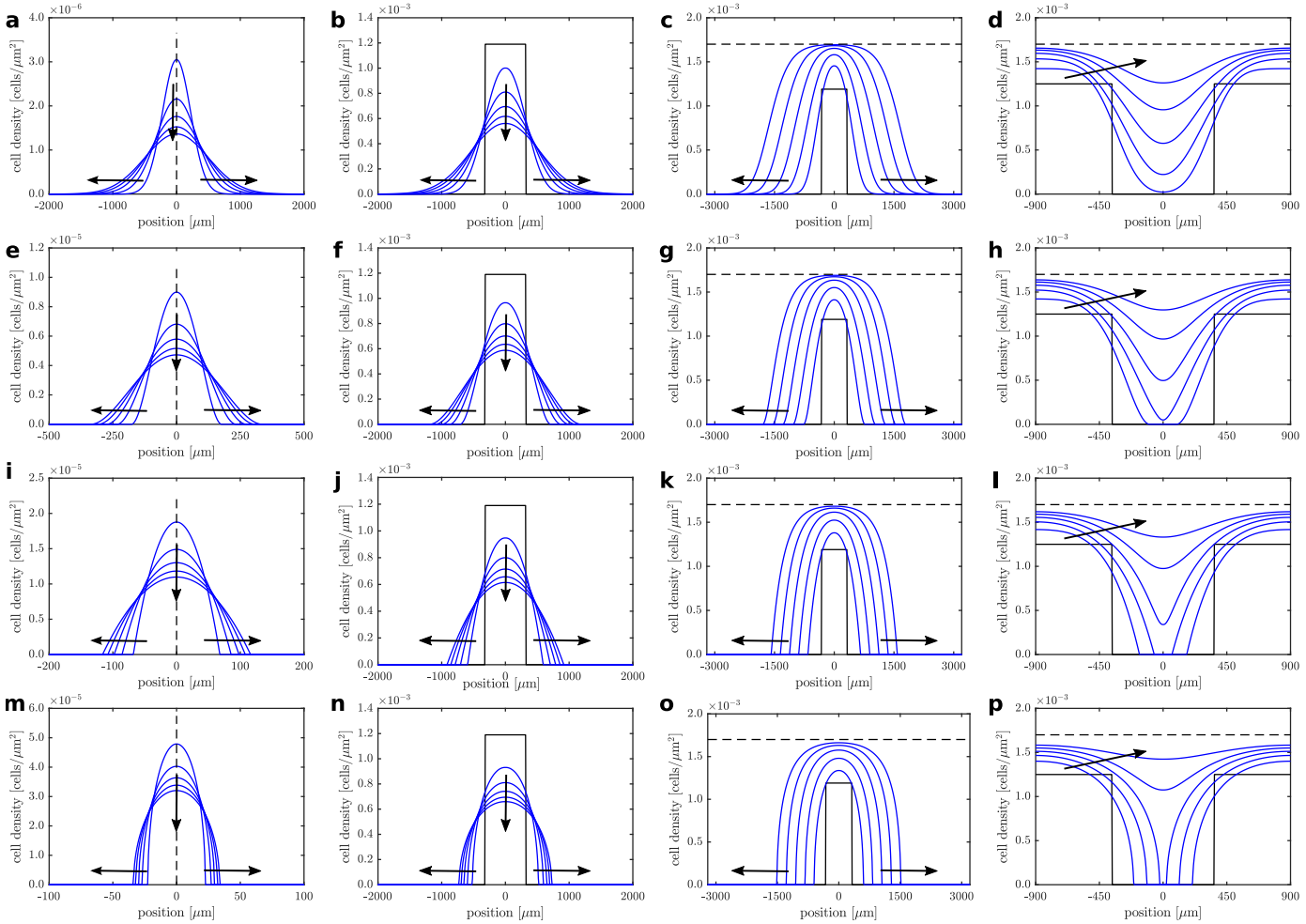

Figure 3.2: Evolution of the solution of the Generalised Porous Fisher equation (solid blue) for various parameter values and initial conditions. The various initial conditions are shown (solid black). Arrows indicate the direction of increasing t and profiles are shown at regular time intervals. Each row highlights the role of the exponent r: (a)–(d) $r = 0$; (e)–(h) $r = 1/2$; (i)–(l) $r = 1$; and (m)–(p) $r = 2$. Each column corresponds to a different initial conditions: (a), (e), (i) and (m) show the exact solution to the diffusion only problem with a delta function initial condition and $\lambda = 0$ shown at 24-hour time intervals; (b), (f), (j) and (n) show the solution of the diffusion-only problem for a spatially-extended initial condition, $\lambda = 0$ shown at 24-hour time intervals; (c), (g), (k) and (o) show the same initial condition and time intervals, but with logistic proliferation with proliferation rate $\lambda > 0$ and carrying capacity density K (dashed black); (d), (h), (l) and (p) show a simplified, but typical, wound healing configuration shown at 12-hour time intervals.

3.3.2 Model comparison

Traditional approaches to model calibration are usually based on regression (Johnson and Omland, 2004). Suppose we have data, $\mathcal{D}$, that consists of n experimental observations, $\mathcal{D} = \left\{Y_{\text{obs}}^{(1)}, Y_{\text{obs}}^{(2)}, \ldots, Y_{\text{obs}}^{(n)}\right\}$ and a mathematical model, $y = M(x;\boldsymbol{\theta})$, parameterised by $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ is a k-dimensional space of valid parameter combinations. The model makes predictions, $\mathcal{P}(\boldsymbol{\theta}) = \left\{y(\boldsymbol{\theta})^{(1)}, y(\boldsymbol{\theta})^{(2)}, \ldots, y(\boldsymbol{\theta})^{(n)}\right\}$, with $y(\boldsymbol{\theta})^{(i)} = M(x^{(i)};\boldsymbol{\theta})$ for $i = 1, 2, \ldots, n$ where $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$ are model inputs; for example, a model input may include initial conditions, or boundary conditions. The regression parameters, $\hat{\boldsymbol{\theta}}$, are obtained through minimisation of the residual error, that is,

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmin}}\, E(\boldsymbol{\theta}), \tag{3.5}$$

where $E(\boldsymbol{\theta})$ is the residual error,

$$E(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left(y(\boldsymbol{\theta})^{(i)} - Y_{\text{obs}}^{(i)}\right)^2.$$

In Equation (3.5), $\hat{\boldsymbol{\theta}}$ corresponds to the MLE under the assumption of Gaussian observational error. Nonlinear optimisation techniques are applied to obtain numerical solutions to Equation (3.5). One of the limitations of the MLE is that only a point estimate of the parameters is obtained, although bootstrapping may be applied to obtain confidence intervals (Gelman et al., 2004). While this can be sufficient, without more effective handling of uncertainty in modelling assumptions or experimental setup, the estimate can be biologically unrealistic (Slezak et al., 2010).

Based on detailed density data from a scratch assay (Figure 3.3(a)), it is unclear whether linear or nonlinear diffusion is most relevant. Sherratt and Murray (1990), using mammalian epidermal wound closure data, find that linear Fickian diffusion ($r = 0$) with chemically regulated proliferation provides a lower residual error than nonlinear diffusion ($r = 4$) without chemical regulation. Due to the scarcity of biological data at the time, Sherratt and Murray (1990) were unable compare results across different initial cell densities, and they were unable to construct spatial density profiles. More recently, through multiple model calibrations using scratch assay data with a range of initial densities, Jin et al. (2016b) demonstrate that estimates

of D_0 are not constant under changes in initial cell density, suggesting that the diffusion of PC-3 prostate cancer cells is density dependent.

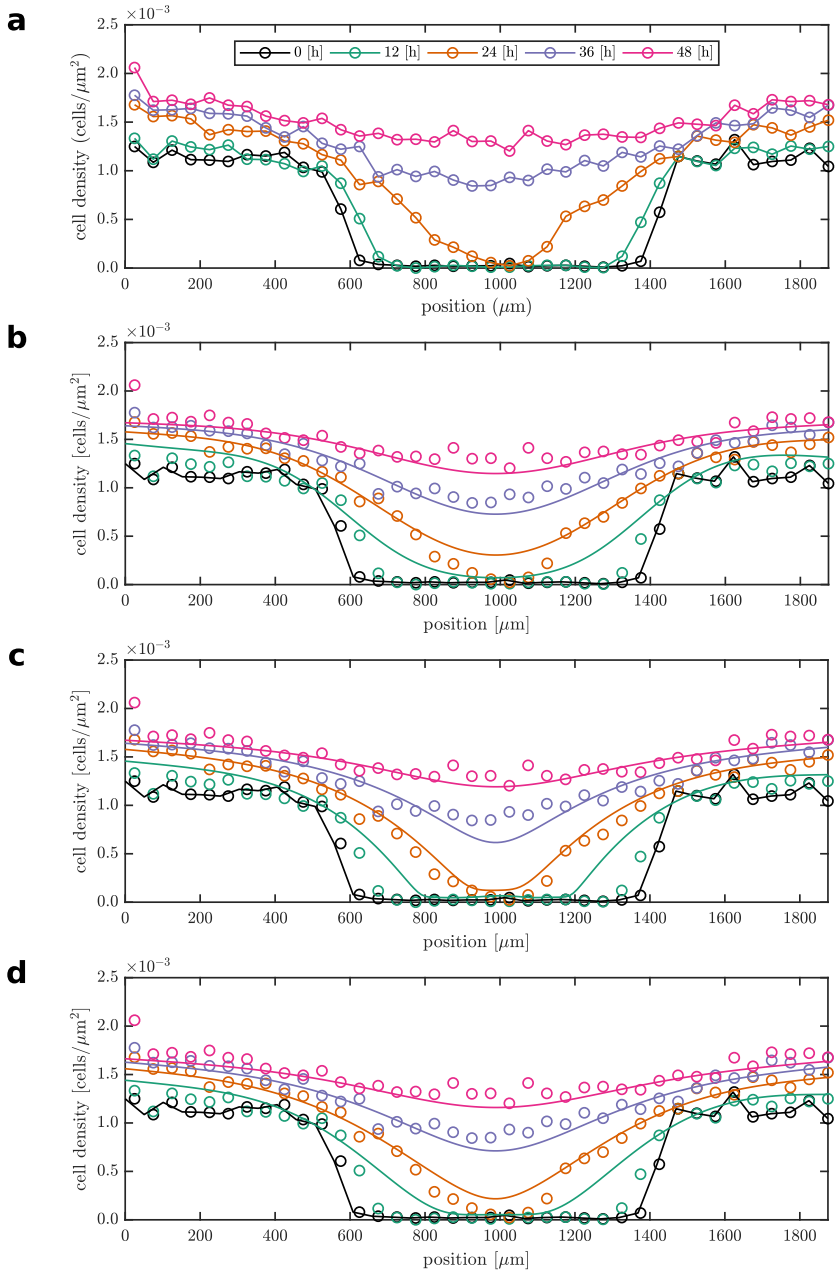

Figure 3.3: Calibration of the (b) Fisher–KPP, (c) Porous Fisher, and (d) Generalised Porous Fisher models against (a) PC-3 scratch assay data seeded with $20,000$ cells (as obtained by Jin et al. (2016b)). (a)–(d) The experimental data are shown at times $t = 0$ [h] (black circles), $t = 12$ [h] (green circles), $t = 24$ [h] (orange circles), $t = 36$ [h] (light purple circles), and $t = 48$ [h] (magenta circles); (a) linear interpolation of data are also shown. The obtained MLE parameter estimates are: (b) $D_0 = 1030$ [μm^2/h], $\lambda = 6.4 \times 10^{-2}$ [1/h] for the Fisher–KPP model; (c) $D_0 = 2900$ [μm^2/h], $\lambda = 6.4 \times 10^{-2}$ [1/h] for the Porous Fisher model; and (d) $D_0 = 2160$ [μm^2/h], $\lambda = 5.8 \times 10^{-2}$ [1/h], $r = 5.2 \times 10^{-1}$ for the Generalised Porous Fisher model. The carrying capacity density is $K = 1.7 \times 10^{-3}$ [cells/μm^2], determined using cell densities at $t = 48$ [h] within 200 [μm] from the left and right boundaries. The initial density profile (solid black) is determined through linear interpolation of the data at time $t = 0$ [h].

The work of Jin et al. (2016b) highlights the need to calibrate models over multiple data sets to effectively compare them. Figure 3.3(b) and (c) show scratch assay density profiles measured at 12-hour intervals superimposed on plots of the solutions of the Fisher–KPP model and the Porous Fisher model using the parameter estimates reported by Jin et al. (2016b). They used mathematical optimisation (Jin et al., 2016b) to find the MLE for the parameters, $\boldsymbol{\theta} = \{D_0, \lambda\}$, assuming K is fixed at a value that they estimate independently by counting cell densities in regions far from the scratch area at late time, $t = 48$ [h], so that the packing density they observed was close to the maximum possible packing density. Results in Figure 3.3 show that both models fit the data well. The minimised residual error, $E(\hat{\boldsymbol{\theta}})$, respectively, is $E(\hat{\boldsymbol{\theta}}) = 2.48 \times 10^{-6}$ for the Fisher–KPP model, and $E(\hat{\boldsymbol{\theta}}) = 2.58 \times 10^{-6}$ for the Porous Fisher model. If model selection were to be based on the minimum residual error, the conclusion would be that the Fisher–KPP model explains the data the best, even if by a small margin. However, we will show that this is an overly simplistic conclusion in this case. Since the Fisher–KPP model (Equation (3.2)) and the Porous Fisher model (Equation (3.3)) are both special cases of the Generalised Porous Fisher model (Equation (3.4)), the Generalised Porous Fisher model cannot have a higher residual error than either of the two special cases. For example, the calibrated Generalised Porous Fisher model, shown in Figure 3.3(d), has a residual error of $E(\hat{\boldsymbol{\theta}}) = 2.47 \times 10^{-6}$. However, we demonstrate that decreased residual error comes at a cost that is totally obscured by taking this standard approach to model selection. As we show, the trade off between model fitness and model complexity can be made very clear using Bayesian techniques.

It should be noted that there are other mechanisms that could be added to enhance the data fit. For example, diffusion of growth factors and/or chemotaxis mechanisms could be included in the suite of potential models that we apply to the experimental data (Bianchi et al., 2016; Nardini et al., 2016; Sherratt and Murray, 1990). Just as with the Generalised Porous Fisher model, this always leads to extra parameters that will enable the model to fit the data better. We suggest this improvement is meaningless without carefully considering the uncertainty in the parameter estimates and the increased model complexity, and we provide further discussion on this point in the Conclusions section.

3.4 A Bayesian framework for model comparison

In this section, we analyse the PC-3 scratch assay density profiles from a Bayesian perspective. We demonstrate that, despite the MLE approach giving preference to the standard Fisher–KPP model, there are other reasons to consider the Porous Fisher model as preferable in this case. This demonstration indicates that it is of benefit to include Bayesian uncertainty quantification as a standard technique for model calibration and validation in biological applications.

3.4.1 Fundamentals of Bayesian analysis

The Bayesian approach is to consider unknown model parameters as random variables with their respective probability distributions representing what is known about the parameters (Efron, 1986; Gelman et al., 2014; Lambert et al., 2018). The conditional probabilities of the parameters given experimental observations represent the new knowledge obtained from an experiment under the assumption of a given model.

Mathematically, this is expressed though Bayes' Theorem (Gelman et al., 2014),

$$p(\boldsymbol{\theta} \mid \mathcal{D}) = \frac{\mathcal{L}(\boldsymbol{\theta}; \mathcal{D}) p(\boldsymbol{\theta})}{p(\mathcal{D})}, \tag{3.6}$$

where $\boldsymbol{\theta}$ is the vector of unknown parameters that exist in some parameter space, $\boldsymbol{\Theta}$, and $\mathcal{D}$ is the set of observations within some space of possible outcomes, $\mathbb{D}$. The *prior* probability density, $p(\boldsymbol{\theta})$, represents any *a priori* knowledge preceding observations, the likelihood, $\mathcal{L}(\boldsymbol{\theta}; \mathcal{D})$, is the probability density of the observations, $\mathcal{D}$, and $p(\boldsymbol{\theta} \mid \mathcal{D})$ is the resulting *posterior* probability density representing new knowledge of the parameters after including observations. The *evidence*, $p(\mathcal{D})$, is a probability density function (PDF) for the observations over all parameters, that is, $p(\mathcal{D}) = \int_{\boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\theta}; \mathcal{D}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}$; from a practical perspective, the evidence is a normalisation constant.

Conceptually, the prior and the likelihood encode assumptions; the former is related to assumed knowledge of parameters and the latter to the underlying mathematical model. One criticism of the Bayesian approach is that the requirement of a prior leads to subjectivity and the definition of appropriate objective rules for "zero-information" priors remains an open problem (Berger,

2006; Consonni et al., 2018; Efron, 1986; Kass and Wasserman, 1996). On the other hand, the Bayesian approach is capable of dealing with arbitrarily complex models and priors, thus providing a very general and consistent analysis framework (Berger, 2006; Efron, 1986).

The Bayesian posterior PDF provides a natural way to describe uncertainty in parameter estimates. In this context, the uncertainty in the ith parameter estimate, θ_i, is defined as the variance of θ_i with respect to the posterior distribution. Bayesian methods have been shown to be highly effective at informing experimental design and parameter inference (Browning et al., 2017; Johnston et al., 2016; Lambert et al., 2018; Silk et al., 2014; Vanlier et al., 2012) (see Chapter 2). Through the Bayesian formulation, experiments can be designed to minimise the level of uncertainty in estimates of a given parameter set (Browning et al., 2019; Drovandi and Pettitt, 2013; Ryan et al., 2016). Furthermore, routine experimental protocols can be analysed to identify areas for potential improvement (Browning et al., 2017) (see Chapter 2).

3.4.2 Scratch assay data informs continuum model comparison

To interpret scratch assay data we let $C_{\text{obs}}(x, t)$ be the observed cell density at position x (along the one-dimensional profile as in Figure 3.3) and time t. We assume that observations are subject to additive noise,

$$C_{\text{obs}}(x, t) = C(x, t; \boldsymbol{\theta}) + \eta, \tag{3.7}$$

where $C(x, t; \boldsymbol{\theta})$ is the true density as determined through the assumed continuum model with parameters, $\boldsymbol{\theta}$, as discussed in Section 3.3, and η represents the combination of measurement error, systematic error and stochastic fluctuations. For simplicity, we treat the error, η, as Gaussian with mean zero, and known variance σ^2, that is, $\eta \sim \mathcal{N}(0, \sigma^2)$. Furthermore, such an assumption ensures that our likelihood formulation corresponds to the likelihood implied by the MLE interpretation of the nonlinear regression approach to model calibration (Equation (3.5)). However, it should be noted that the Bayesian techniques we apply here do not require this assumption, nor is it a requirement that σ be known. In Section 3.6, we discuss potential advantages of considering σ as an unknown parameter.

We also specify, for ease of description, that $C_{\text{obs}}(x, 0) = C(x, 0; \boldsymbol{\theta})$, as this is a typical requirement in the calibration of PDE-based models for simulation purposes (Jin et al., 2016b;

Maini et al., 2004a,b; Sherratt and Murray, 1990). However, it should be noted that our framework can be extended to deal with more realistic cases with observation error in the initial conditions. The treatment of the initial condition observation error is an interesting point for discussion, so we provide further results and details in Section 3.7.4. Importantly, parameter uncertainty is amplified in this case. See also Jin et al. (2016b) and Chapter 2 for further details. Processed scratch assay data are of the form $\mathcal{D} = \mathbf{C}_{\text{obs}}^{1:N,1:M}$ where $\mathbf{C}_{\text{obs}}^{1:N,1:M}$ is an $N \times M$ matrix with elements that are observations, as given by Equation (3.7), at NM position-time pairs taken from the Cartesian product of N spatial points $x_1, x_2, \ldots, x_N$ with M temporal points $t_1, t_2, \ldots, t_M$. The PC-3 cell line scratch assay data sets, described in Section 3.2, are derived from microscopy images that are taken at times $t_1 = 0$ [h], $t_2 = 12$ [h], $t_3 = 24$ [h], $t_4 = 36$ [h], and $t_5 = 48$ [h]. Densities are computed, for each point in time, using rectangular areas with centerlines at $x_1 = 25, x_2 = 75, x_3 = 125, \ldots, x_{38} = 1925$ [μm]. That is, $N = 38$ and $M = 5$. These derived data are provided by Jin et al. (2016b), and are more detailed than the data used by Maini et al. (2004a,b) and Sherratt and Murray (1990). However, the data used by Maini et al. (2004a,b) and Sherratt and Murray (1990) are still more detailed than many studies in which microscopy images alone, without any quantitative measurements of density or spatial position, are used to illustrate the outcomes of a scratch assay.

First, we consider the Fisher–KPP (Equation (3.2)) and Porous Fisher models (Equation (3.3)). The task is to construct a Bayesian posterior PDF for the parameters $\boldsymbol{\theta} = [D_0, \lambda, K]$ given each of the initial conditions and under the assumption of each model. The resulting posterior PDFs may be compared visually to get intuition on the appropriateness of each model given the PC-3 data. Most analyses of cell motility and proliferation assume the carrying capacity density K is a known parameter (Chapter 2), however, this is usually an approximation that is required due to short assay timescales (Chapter 2). In the case of the data of Jin et al. (2016b), K is more appropriately considered as unknown since the data captures more of the long-time effects of contact inhibition (Sarapata and de Pillis, 2014) (Chapter 2).

To keep subjective bias to a minimum, the aim is to assume as little as possible within the prior distributions, that is we wish them to be uninformative. To this end, we select uniform prior distributions for each of the three parameters such that the support extends well beyond biologically viable ranges. In the literature, typical ranges for each parameter are:

$D_0 = 155$–$6500[\mu\text{m}^2/\text{h}]$; $\lambda = 0.01$–$0.07[1/\text{h}]$; and $K = 1.5\times10^{-3}$–$2.0\times10^{-3}[\text{cells}/\mu\text{m}^2]$ (Browning et al., 2017; Maini et al., 2004a,b; Jin et al., 2016b). Therefore, we assume as priors $D_0 \sim \mathcal{U}(0, D_{\text{max}})$, $\lambda \sim \mathcal{U}(0, \lambda_{\text{max}})$, and $K \sim \mathcal{U}(0, K_{\text{max}})$ where $D_{\text{max}} = 10^5$ $[\mu\text{m}^2/\text{h}]$, $\lambda_{\text{max}} = 1$ $[1/\text{h}]$ and $K_{\text{max}} = 7 \times 10^{-3}$ $[\text{cells}/\mu\text{m}^2]$. As a joint PDF, we have

$$p(D_0, \lambda, K) = p(D_0)p(\lambda)p(K)$$

where

$$\begin{aligned} p(D_0) &= \mathbb{1}_{[0,D_{\text{max}}]}(D_0)/D_{\text{max}}, \\ p(\lambda) &= \mathbb{1}_{[0,\lambda_{\text{max}}]}(\lambda)/\lambda_{\text{max}}, \\ p(K) &= \mathbb{1}_{[0,K_{\text{max}}]}(K)/K_{\text{max}}. \end{aligned}$$

Here, $\mathbb{1}_A(x) = 1$ if $x \in A$, otherwise $\mathbb{1}_A(x) = 0$. It is important to note, however, that uniform priors are not always uninformative and care must be taken (Efron, 1986).

Under the aforementioned assumption of independent Gaussian observation error on the data (Equation (3.7)), the likelihood function is

$$\mathcal{L}(\mathbf{C}_{\text{obs}}^{1:N,1:M}; D_0, \lambda, K) = \frac{1}{(\sigma\sqrt{2\pi})^{NM}} \prod_{i=1}^{N}\prod_{j=1}^{M} \exp\left(-\frac{(C_{\text{obs}}(x_i,t_j) - C(x_i,t_j;D_0,\lambda,K))^2}{2\sigma^2}\right), \tag{3.8}$$

where $C(x_i, t_j; D_0, \lambda, K)$ is the solution of the continuum model of interest, computed numerically (see Section 3.7.2) at position x_i and time t_j, given values for D_0, λ and K. The model will be either the Fisher–KPP model (Equation (3.2)) or the Porous Fisher model (Equation (3.3)). The observation error is such that $\sigma \approx 10^{-5}$ (cells/μm^2), as obtained from previous studies (Jin et al., 2016b) (see Chapter 2). Comparing the nonlinear regression problem in Equation (3.5) with this likelihood (Equation (3.8)) reveals that the MLE parameter set in Equation (3.5) corresponds to the mode of the likelihood as a function of the parameter vector, $\boldsymbol{\theta} = [D_0, \lambda, K]$.

Substitution of the prior PDF, $p(D_0, \lambda, K)$, and the likelihood (Equation (3.8)) into the right-hand side of Bayes' Theorem (Equation (3.6)) yields the posterior PDF, $p(D_0, \lambda, K \mid \mathbf{C}_{\text{obs}}^{1:N,1:M})$.

From this PDF, we can construct the posterior marginal PDFs,

$$p(D_0 \mid \mathbf{C}_{\text{obs}}^{1:N,1:M}) = \iint_{\mathbb{R}^2} p(D_0, \lambda, K \mid \mathbf{C}_{\text{obs}}^{1:N,1:M}) \,\mathrm{d}K \,\mathrm{d}\lambda,$$
$$p(\lambda \mid \mathbf{C}_{\text{obs}}^{1:N,1:M}) = \iint_{\mathbb{R}^2} p(D_0, \lambda, K \mid \mathbf{C}_{\text{obs}}^{1:N,1:M}) \,\mathrm{d}K \,\mathrm{d}D_0,$$
$$p(K \mid \mathbf{C}_{\text{obs}}^{1:N,1:M}) = \iint_{\mathbb{R}^2} p(D_0, \lambda, K \mid \mathbf{C}_{\text{obs}}^{1:N,1:M}) \,\mathrm{d}\lambda \,\mathrm{d}D_0.$$

The posterior marginal PDFs represent the uncertainty in a single parameter taken over all possibilities of the remaining parameters. The integrals are best computed using Monte Carlo integration. Specifically, we apply approximate Bayesian computation (ABC) rejection sampling (Sisson et al., 2018; Sunnåker et al., 2013) (see Chapter 4) to obtain samples from the joint posterior density, then apply Monte Carlo integration with kernel smoothing to estimate the posterior marginal densities (Silverman, 1986). We leave the computational details to Section 3.7.3, however, it is important to note that, for deterministic models with additive noise, the ABC rejection sampler can be shown to be exact (Wilkinson, 2013).

We compare the posterior marginal PDFs obtained under the assumption of the Fisher–KPP model (Equation (3.2)), conditioned on data as given in Equation (3.7), against those obtained under the assumption of the Porous Fisher model (Equation (3.3)). Posterior marginal PDFs are computed using ABC rejection sampling to generate $n = 50,000$ samples from the joint posterior distribution as described in Section 3.7.3. The results are presented in Figure 3.4. Leaving more quantitative analysis for Section 3.5, we discuss here the qualitative aspects that are essential for, not only model selection and comparison, but also experimental design.

The first point of interest is the trade-off between uncertainty in the proliferation rate, λ, and the carrying capacity density, K. The uncertainty in λ increases as the initial cell density increases (Figure 3.4(b) and (e)), however, the reverse is true for the uncertainty in K (Figure 3.4(c) and (f)). This behaviour is observed regardless of whether the Fisher–KPP model (Figure 3.4(a)–(c)) or the Porous Fisher model (Figure 3.4(d)–(f)) is assumed. Further insight is obtained through the posterior correlation coefficient matrix (see Section 3.7.4, Table 3.6), as λ and K are negatively correlated for all initial densities. This result is consistent with the results of Chapter 2 and demonstrates that different experimental designs may be required to target different parameters. Lower initial densities result in data that only captures transient dynamics which maximises information related to λ. On the other hand, higher initial densities enable

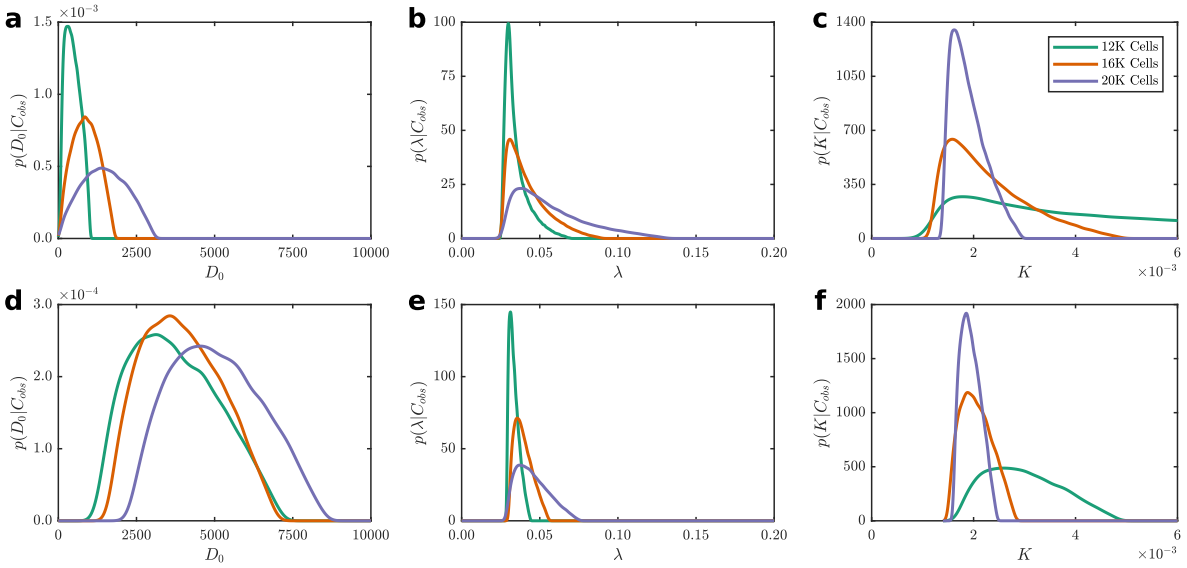

Figure 3.4: Marginal posterior probability densities obtained through Bayesian inference on the PC-3 scratch assay data under the (a)–(c) Fisher–KPP model and (d)–(f) Porous Fisher model. The spread in marginal densities demonstrate the degree of uncertainty in the diffusivity, D_0, the proliferation rate, λ, and the carrying capacity, K, for each model under the three different initial conditions; $12,000$ cells (solid green), $16,000$ cells (solid orange), $20,000$ cells (solid purple).

more precise estimation of limiting dynamics, that is, the effect of K can be observed. This feature would be difficult to elicit using traditional MLE-based methods.

The modes of the posterior marginal PDFs for λ and K are in agreement across both models (Figure 3.4(b)–(c) and (e)–(f)). However, the Porous Fisher model leads to lower uncertainty than the Fisher–KPP model in both of these parameters, regardless of initial density (see covariance matrices in Table 3.5). This indicates that the uncertainty in the diffusion parameter, D_0, has more of an impact on the other parameters under the Fisher–KPP model.

Comparing the posterior marginal PDFs for D_0 requires some care. An initial inspection reveals the uncertainty in D_0 looks significantly larger in the Porous Fisher model (Figure 3.4(b)) compared with the Fisher–KPP model (Figure 3.4(a)). However, the role of D_0 in the density dependent diffusion of the Porous Fisher model (Equation (3.3)) is not the same as that in the Fickian diffusion of the Fisher–KPP model (Equation (3.2)). In the Fisher–KPP model, D_0 is a constant cell diffusivity whereas in the Porous Fisher model, D_0 is the maximum diffusivity. Perhaps it would be more appropriate to give these two quantities different variables to make this point of distinction clear. Here, however, we have chosen to use the same variable to denote both quantities to be consistent with previous literature Jin et al. (2016b).

The most important aspect for the purposes of model comparison is the qualitative change in the diffusion parameter, D_0, for different initial densities. For the Fisher–KPP model, not only does the uncertainty in D_0 increase as the initial density increases, the mode also increases: $D_0 = 305.3$ [μm^2/h] for low initial density; $D_0 = 850.9$ [μm^2/h] for medium initial density; and $D_0 = 1371.4$ [μm^2/h] for high initial density (Figure 3.4(a)). This suggests that D_0 depends on cell density, and this observation directly contradicts the implicit assumption made in invoking the Fisher–KPP model which treats D_0 as a constant. This analysis would indicate that PC-3 cells exhibit density dependent motility where the diffusivity increases with the density. In contrast, the posterior marginal PDFs of D_0 under the Porous Fisher model are very similar across initial densities (Figure 3.4(d)). Furthermore, the variance is consistent across all three initial densities considered. There is an increase in the posterior mode to $D_0 = 4,933$ [μm^2/h] for high initial density. This is consistent with the Porous Fisher model since:
1) $D(C) = D_0$ only if $C = K$; 2) there is a significant decrease in the uncertainty in K for high initial density (see Table 3.5); and 3) the high initial density profiles achieve values closer to K at later times (Jin et al., 2016b). These results are in agreement with the observations of Jin et al. (2016b) who use the MLE to show that the Fisher–KPP model is inconsistent with the data. However, our Bayesian approach provides more detail through the reconstruction of the posterior PDF. Despite the fact that MLE model comparison, as presented in Figure 3.3, selects the Fisher–KPP model as the preferred model, direct visualisation of the parameter uncertainty indicates otherwise.

3.4.3 Inference on the Generalised Porous Fisher model

The model comparison analysis performed in Section 3.4.2 is naturally extended by encoding model comparison as a single Bayesian inference problem applied to the Generalised Porous Fisher model (Equation (3.4)) with the exponent in the nonlinear diffusion term, r, also treated as an unknown parameter. This is essentially model comparison of a continuous population of models. This is another advantage of the Bayesian approach: model uncertainty can be treated identically to parameter uncertainty, even when competing models cannot be obtained through special cases of a single general model (Clyde and George, 2004).

The inference problem must now be slightly modified based on Section 3.4.2. The prior PDF becomes $p(D_0, \lambda, K, r) = p(D_0)p(\lambda)p(K)p(r)$ where $p(r) = \mathbb{1}_{[r_{\min}, r_{\max}]}(r) / (r_{\max} - r_{\min})$. That

is, $r \sim \mathcal{U}(r_{\text{min}}, r_{\text{max}})$. The limits that should be placed on r are unclear since r is has no physical interpretation and various values of r are used with little justification (Edelstein-Keshet, 2005; Murray, 2002; Sherratt and Murray, 1990; Simpson et al., 2011). Since the most common values used in applications are $r = 0$ and $r = 1$, with the maximum known value used being $r = 4$, we take $r_{\text{max}} = 8$ so that we conservatively consider twice the largest value used in the mathematical biology literature. We also set $r_{\text{min}} = -1$ to allow for the possibility of $r < 0$, resulting in, so-called, "fast nonlinear diffusion" which is also thought to have some relevance to biological and ecological applications (King and McCabe, 2003).

Using the same assumptions for observation error as given in Section 3.4.2, the resulting likelihood is

$$\mathcal{L}(\mathbf{C}_{\text{obs}}^{1:N,1:M}; D_0, \lambda, K, r) = \frac{1}{(\sigma\sqrt{2\pi})^{NM}} \prod_{i=1}^{N} \prod_{j=1}^{M} \exp\left(-\frac{(C_{\text{obs}}(x_i, t_j) - C(x_i, t_j; D_0, \lambda, K, r))^2}{2\sigma^2}\right),$$

where $C(x_i, t_j; D_0, \lambda, K, r)$ is the numerical solution to the Generalised Porous Fisher equation, computed as per Section 3.7.2. Just as in Section 3.4.2, we apply Bayes' Theorem (Equation (3.6)) to obtain the posterior PDF, $p(D_0, \lambda, K, r \mid \mathbf{C}_{\text{obs}}^{1:N,1:M})$. The posterior marginal PDFs are

$$\begin{aligned}
p(D_0 \mid \mathbf{C}_{\text{obs}}^{1:N,1:M}) &= \iiint_{\mathbb{R}^3} p(D_0, \lambda, K, r \mid \mathbf{C}_{\text{obs}}^{1:N,1:M}) \,\mathrm{d}r \,\mathrm{d}K \,\mathrm{d}\lambda, \\
p(\lambda \mid \mathbf{C}_{\text{obs}}^{1:N,1:M}) &= \iiint_{\mathbb{R}^3} p(D_0, \lambda, K, r \mid \mathbf{C}_{\text{obs}}^{1:N,1:M}) \,\mathrm{d}r \,\mathrm{d}K \,\mathrm{d}D_0, \\
p(K \mid \mathbf{C}_{\text{obs}}^{1:N,1:M}) &= \iiint_{\mathbb{R}^3} p(D_0, \lambda, K, r \mid \mathbf{C}_{\text{obs}}^{1:N,1:M}) \,\mathrm{d}r \,\mathrm{d}\lambda \,\mathrm{d}D_0, \\
p(r \mid \mathbf{C}_{\text{obs}}^{1:N,1:M}) &= \iiint_{\mathbb{R}^3} p(D_0, \lambda, K, r \mid \mathbf{C}_{\text{obs}}^{1:N,1:M}) \,\mathrm{d}K \,\mathrm{d}\lambda \,\mathrm{d}D_0.
\end{aligned}$$

Computationally, low acceptance rates in the ABC rejection sampler render the technique ineffective for sampling the posterior distribution in this case. As a result, we apply an ABC variant of a Markov chain Monte Carlo (MCMC) sampler (Marjoram et al., 2003) (see Section 3.7.3). While we find that the ABC MCMC sampler works well for this problem, other advanced ABC-based Monte Carlo schemes are also possible, such as sequential Monte Carlo (Sisson et al., 2007) and multilevel Monte Carlo (see Chapter 6). Using the ABC MCMC sampler, the four parameter posterior marginal PDFs are estimated using the same data and observation error as in Section 3.4.2. Results are presented in Figure 3.5.

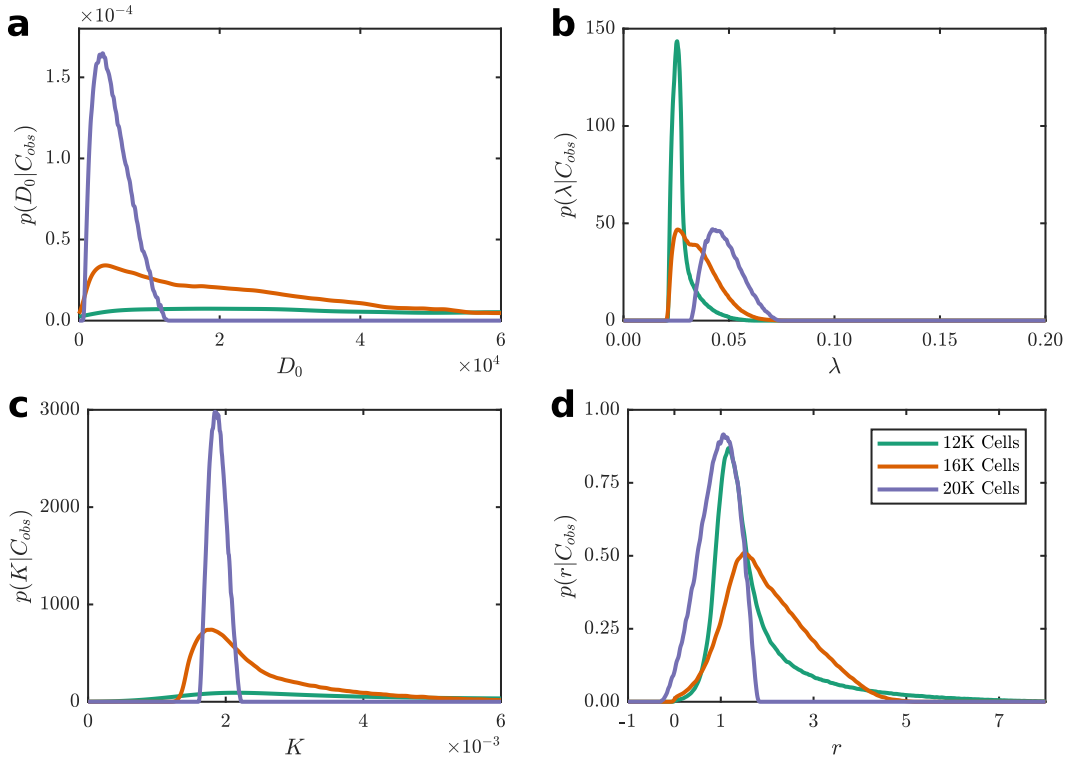

Figure 3.5: Marginal posterior probability densities obtained through Bayesian inference on the PC-3 scratch assay data under the Generalised Porous Fisher model. The spread in marginal densities demonstrate the degree of uncertainty in the diffusivity, D_0, the proliferation rate, λ, the carrying capacity, K, and the exponent of the nonlinear diffusion, r, under the three different initial conditions; $12,000$ cells (solid green), $16,000$ cells (solid orange), $20,000$ cells (solid purple).

The marginal posterior PDF for r, $p(r \mid \mathbf{C}_{\text{obs}}^{1:N,1:M})$, displays (Figure 3.5(d)) an overall higher probability density around $r = 1$ (corresponding to the Porous Fisher model) and very low probability density around $r = 0$ (corresponding to the Fisher–KPP model). However, this analysis also shows that other values of r may be justifiable. The uncertainty in r initially increases with increased initial density, but then decreases again with further increases in initial density. The same pattern occurs for the mode, as it transitions from $r \approx 1$ to $r \approx 2$ then back to $r \approx 1$.

The previously identified trade-off still exists between uncertainty in K and λ (Figure 3.5(b) and (c)) as the initial density increases. However, there is a qualitative difference in the marginal posterior PDFs compared to those in Figure 3.4(b),(c),(e) and (f). There is less consistency in the estimates across initial conditions, especially with the mode of λ apparently increasing as the initial density increases. For D_0, almost no information is provided through the posterior marginal PDF except for data with high initial cell densities. Further analysis of the multivariate

posterior marginal PDFs or the full joint posterior PDF would be required to obtain more information for this purpose. Such analysis is significantly more complex, and it may still yield a large degree of uncertainty in D_0. In Section 3.7.4 an analysis is performed using bivariate marginal PDFs. It is clear the interactions between r and the other parameters are quite complex (see Figure 3.8). Furthermore, as seen in Table 3.6, r and λ are positively correlated for low initial density, but negatively correlated for high initial density; similarly, r and K are negatively correlated for low initial density, but positively correlated for high initial density. This kind of behaviour is difficult to interpret biologically, and we conclude that it is an artifact of using an overly complex model.

We demonstrate that a full Bayesian approach can easily incorporate comparison of a population of models. However, significantly more detailed analysis is required to interpret the results. Conversely, comparison of two distinct models though individual Bayesian inferences resulted in reasonable conclusions with minimal detailed analysis. A generalised model will never provide a lower fit in MLE (Stoica and Selen, 2004), however, there must be a point where the improved MLE is negligible (or nonexistent) compared with the overall increase in uncertainty that must come with the generalisation. This raises the question whether the increased information obtained through a generalised model is worth the additional complexity.

3.5 Information criteria to balance model fit and complexity

In Sections 3.4.2 and 3.4.3, we observed that information gains from model generalisations may come at the cost of increased complexity and parameter uncertainty. We demonstrated in Sections 3.3 and 3.4.2 that the definition of a good model must be based on more than the MLE alone. The theoretical underpinnings, explanatory power, biological feasibility, verifiability and complexity of a model are all important factors to consider when performing model selection (Box, 1976; Jin et al., 2016b; Sarapata and de Pillis, 2014; Slezak et al., 2010; Spiegelhalter et al., 2002). Overparameterised, complex models may fit the data well, however, the principle of Occam's razor dictates that a simple model should be preferred wherever possible. In this section, we demonstrate the use of statistical measures, known as information criteria, that are designed to deal with trade-off between complexity and model fit (Gelman et al., 2014; Johnson and Omland, 2004).

3.5.1 Information criteria

Information criteria can be considered as methods for the ranking of models. Of the wide variety of information criteria available, many are based on rewarding models for lower residual error and penalising models for parameterisation (Gelman et al., 2014). Most information criteria are derived from decision theory and are based on the minimisation of some measure of information loss (Gelman et al., 2014; Johnson and Omland, 2004; Stoica and Selen, 2004). The resulting criteria, under suitable assumptions, are asymptotically proportional to information loss relative to an unknown true model (Gelman et al., 2014). That is, the model with the lowest information criterion value has the lowest information loss asymptotically. This is often taken to be a superior model from a decision theoretic perspective (Yang, 2005). However, caution must be taken when applying such measures because they are only guaranteed to inform correct decisions in the large sample limit (Gelman et al., 2014).

The three most fundamental information criteria are considered here, each of which have distinct properties, advantages and disadvantages. These information criteria have seen use in a limited set of biological applications (Johnson and Omland, 2004), and we are unaware of any application of these measures in the study of collective cell migration. It is important to note that there are many variants of the aforementioned criteria, however, we restrict ourselves the standard formulations in this work; for further information, see Gelman et al. (2014). We compare and contrast the information criteria results for the scratch assay data against the full Bayesian analysis performed in Section 3.4.

3.5.2 The Akaike information criterion

Akaike (1974) was the first to propose a solution to the key problem with MLE-based approaches, that is, overparameterised models will always be preferred (Akaike, 1974). Akaike considered the Kullback–Leibler (KL) information measure (Gelman et al., 2014; Kullback and Leibler, 1951),

$$H_p(\boldsymbol{\theta}) = \mathbb{E}\left[\ln\left(\frac{p_{\text{true}}(\mathcal{D})}{p(\mathcal{D} \mid \boldsymbol{\theta})}\right)\right] = \int_{\mathbb{D}} \ln\left(\frac{p_{\text{true}}(\mathcal{D})}{p(\mathcal{D} \mid \boldsymbol{\theta})}\right) p_{\text{true}}(\mathcal{D})\, \mathrm{d}\mathcal{D}, \tag{3.9}$$

where $\ln(\cdot)$ denotes the natural logarithm. The KL information measures the information loss,

in the Shannon entropy sense (Shannon, 1948), incurred by assuming a model, $p(\mathcal{D} \mid \boldsymbol{\theta})$, of the data, $\mathcal{D} \in \mathbb{D}$, instead of the hypothetical true model[2] $p_{\text{true}}(\mathcal{D})$. Therefore, given two candidate models, $p(\mathcal{D} \mid \boldsymbol{\theta})$ and $q(\mathcal{D} \mid \boldsymbol{\theta})$, comparing $H_p(\boldsymbol{\theta})$ and $H_q(\boldsymbol{\theta})$ would reveal the model that minimises information loss. However, $p_{\text{true}}(\mathcal{D})$ is unknown, so it is not possible to evaluate Equation (3.9) in practice. By analysing an asymptotic expansion of the KL information about the true parameter set, Akaike (1974) derives a penalty on the MLE based on the number of parameters that must be estimated. The result is the Akaike information criterion (AIC),

$$\text{AIC} = -2 \ln \left(\hat{\mathcal{L}}(\boldsymbol{\theta}; \mathcal{D}) \right) + 2k, \tag{3.10}$$

where k is the dimensionality of $\boldsymbol{\theta}$ and $\hat{\mathcal{L}}(\boldsymbol{\theta}; \mathcal{D})$ is the maximum likelihood estimate, that is,

$$\hat{\mathcal{L}}(\boldsymbol{\theta}; \mathcal{D}) = \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\theta}; \mathcal{D}). \tag{3.11}$$

Due to the penalty incurred by the number of model parameters, a more complex model must improve the agreement with the data sufficiently to outperform a simpler model with an inferior maximum likelihood estimate. However, for models with the same number of parameters, the AIC is equivalent to the maximum likelihood estimate. When models have different numbers of parameters, the AIC favours simpler models. Unfortunately, the AIC is not an asymptotically consistent estimator, that is, as $n \to \infty$, the AIC is not guaranteed to converge to a unique model (Yang, 2005). However, the AIC will select the model with optimal residual error, since it may be viewed as the MLE with a bias correction to compensate for overfitting (Yang, 2005).

3.5.3 The Bayesian information criterion

An alternative approach, the Bayesian information criterion (BIC), has different theoretical foundations (Schwarz, 1978). Schwarz (1978) considered Bayes estimators instead of the KL information, thereby defining the model with the maximum *a posteriori* probability to be the optimal choice. The resulting BIC is

$$\text{BIC} = -2 \ln \left(\hat{\mathcal{L}}(\boldsymbol{\theta}; \mathcal{D}) \right) + k \ln (n), \tag{3.12}$$

[2]Many reject the notion that a true model exists (Box, 1976; Spiegelhalter et al., 2014). However, the concept is a useful one for the purposes of deriving information criteria (Akaike, 1974; Spiegelhalter et al., 2002).

where k is the dimensionality of $\boldsymbol{\theta}$, n is the dimensionality of $\mathcal{D}$ and $\hat{\mathcal{L}}(\boldsymbol{\theta};\mathcal{D})$ is the maximum likelihood estimate as given in Equation (3.11). The BIC is a consistent approximation to the maximum *a posteriori* estimate (i.e., the mode of the posterior PDF) and is independent of model priors provided $k/n \ll 1$. Compared with the AIC, the BIC attributes a larger penalty for model complexity when $n \geq 8$, thus the BIC favours simplicity more than the AIC. The BIC also has the advantage of being consistent as $n \to \infty$. In addition, the BIC is asymptotically equivalent to the comparison of Bayes factors, which are considered more correct by some (Pooley and Marion, 2018), however, others note that the usage of Bayes factors assumes that the model prior covers the correct model (Gelman et al., 2004). The BIC, however, will not always select the model with the optimal residual error compared with the AIC (Yang, 2005).

3.5.4 The deviance information criterion

The prevalence of Monte Carlo sampling in practical Bayesian applications was the motivation for Spiegelhalter et al. (2002) to develop the deviance information criterion (DIC). The DIC has particular computational advantages when posterior samples, as obtained through MCMC, are available (Gelman et al., 2014; Spiegelhalter et al., 2002). However, since the AIC and BIC do not require posterior sampling, they can be more computationally efficient when samples are not readily available. Like the AIC, the DIC has its theoretical foundation in minimisation of the KL information loss (Gelman et al., 2004). However, unlike the AIC and BIC, the DIC does not utilise the maximum likelihood estimate, but is based on expectation calculations. As a result, the DIC is ideal for Monte Carlo integration schemes (Gelman et al., 2014). The DIC is given by

$$\mathrm{DIC} = \mathbb{E}\left[f(\boldsymbol{\theta})\right] + p_{\text{eff}}, \tag{3.13}$$

where $f(\cdot)$ is the deviance function,

$$f(\boldsymbol{\theta}) = -2\ln\left(\mathcal{L}(\boldsymbol{\theta};\mathcal{D})\right), \tag{3.14}$$

and p_{eff} is the effective number of parameters,

$$p_{\text{eff}} = \mathbb{E}\left[f(\boldsymbol{\theta})\right] - f(\mathbb{E}\left[\boldsymbol{\theta}\right]). \tag{3.15}$$

Importantly, the expectations are evaluated with respect to the posterior probability measure. The DIC is conceptually quite different to the AIC and BIC. Firstly, the DIC uses an averaged likelihood rather than point estimates like the AIC and BIC. Secondly, the DIC effective parameter term, p_{eff}, attempts to distinguish between information obtained through the prior distribution rather than the data (Gelman et al., 2004).

Because the averaged likelihood is utilised, it can be considered more closely aligned with a Bayesian viewpoint that aims to use information from the entire posterior distribution (Gelman et al., 2014). However, the use of the DIC as a reliable information criterion has been debated in the literature, in particular there is concern that the DIC is inconsistent with Bayes factors, that is, the DIC may fail to select the true model even if it is among the set of candidates (Pooley and Marion, 2018; Spiegelhalter et al., 2014). However, due to its simplicity of calculation via Monte Carlo integration and applicability to hierarchical models, the DIC has been widely adopted for practical applications (Gelman et al., 2014; Spiegelhalter et al., 2014).

3.5.5 Evaluation of continuum models using information criteria

We compute the AIC, BIC and DIC for the Fisher–KPP, Porous Fisher and Generalised Porous Fisher models given the data derived from Jin et al. (2016b). The results are presented for each initial condition in Table 3.1.

Table 3.1: Information criteria for initial conditions of $12,000$ cells, $16,000$ cells and $20,000$ cells.

	$12,000$ cells			$16,000$ cells			$20,000$ cells		
Model	DIC	BIC	AIC	DIC	BIC	AIC	DIC	BIC	AIC
Fisher–KPP	-2386.08	-2377.06	-2386.13	-2367.03	-2358.00	-2367.07	-2268.60	-2259.54	-2268.61
Porous Fisher	-2398.30	-2389.87	-2398.94	-2392.87	-2383.59	-2392.66	-2278.58	-2269.41	-2278.49
Generalised Porous Fisher	-2399.07	-2387.92	-2400.01	-2390.88	-2378.84	-2390.93	-2279.36	-2267.59	-2279.69

Across all initial conditions, the BIC consistently selects the Porous Fisher model. The consistency of the BIC is expected due to its theoretical basis of Bayes estimators (Schwarz, 1978; Spiegelhalter et al., 2014; Yang, 2005), that is, if the true model was in the set of candidates, then BIC would asymptotically select that model. The BIC results are also consistent with the full Bayesian analysis performed in Section 3.4.

The AIC and DIC are in agreement, preferring the Generalised Porous Fisher model for both

low and high density initial conditions (Table 3.1), but favouring the Porous Fisher Model for the intermediate initial density. The agreement between the AIC and DIC is also expected theoretically since they are both derived from the KL information loss. In fact, if $p_{\text{eff}} \approx k$ and $\mathbb{E}\left[f(\boldsymbol{\theta})\right] \approx -2\ln\left(\hat{\mathcal{L}}(\boldsymbol{\theta};\mathcal{D})\right)$ then Equation (3.10) and Equation (3.13) are approximately equivalent. The results in Table 3.1 indicate these relationships likely hold in our case.

Overall, the information criteria provide clear indications that the Fisher–KPP model does not sufficiently describe the collective behaviour of the PC-3 cells for any initial density. In Table 3.1 we see there is good agreement between the rankings suggested by the AIC and DIC, and some disagreement in the ranking suggested by the BIC. However, the improvement in the AIC and DIC for the Generalised Porous Fisher model over the Porous Fisher model is negligible compared with the improvement in the AIC and DIC for the Porous Fisher model over Fisher–KPP model. Furthermore, the BIC consistently selects the Porous Fisher model for each initial density. Therefore, we conclude that the Porous Fisher model represents the best trade-off between model fit and complexity.

3.6 Discussion and outlook

We have demonstrated in Section 3.3 that traditional MLE methods of model calibration and model selection do not provide a satisfactory method to compare continuum models of collective cell spreading and proliferation since MLE methods favour overparameterised models. The Bayesian approach presented in Section 3.4 and the analysis using information criteria that is presented in Section 3.5 provide a significantly more robust methodology to evaluate the ability of continuum models to explain collective cell behaviour. This methodology has enabled us to present a clear example of when model generalisation leads to less consistent and more uncertain parameter estimation using the PC-3 prostate cancer cell line. The Bayesian analysis presented in Section 3.4 indicates that the Porous Fisher model provides the most consistent parameter estimates, and information criteria demonstrated in Section 3.5 suggest that the Porous Fisher model represents the optimal trade-off between model fit and complexity. It is important to note our methodology can be more generally applied to other cell lines, however, the model selection results presented here are specific to PC-3 cells only.

Information criteria provide an objective approach to model selection that take into account

model fitness and complexity. However, reducing model comparison down to a single scalar comparison is bound to disregard some important aspects of the model comparison problem (Gelman et al., 2014; Spiegelhalter et al., 2014). On the other hand, Bayesian posterior distributions provide a rich source of information. For example, the shifts in modes and support in the diffusion parameter, D_0, for both the Fisher–KPP model (Figure 3.4(a)) and Generalised Porous Fisher model (Figure 3.5(a)) are not directly identifiable using information criteria. If an objective decision rule is required, then a clear understanding of the assumptions and asymptotic properties of the criterion used is essential. We conclude, along with Yang (2005), that the BIC is most appropriate to select the best overall model, whereas the AIC and DIC are better suited if the goal is prediction over a short time-scale. Bayes factors are another alternative, however, their results can be highly sensitive to the choice of model prior distributions, especially when model parameters are continuous (Gelman et al., 2014; Pooley and Marion, 2018).

The Bayesian approach we take in Section 3.4 is the most direct and intuitive approach to model selection. Model inconsistency across data sets is apparent, through comparing marginal posterior PDFs for different data sets. Parameters that the data provides little information about are also highlighted this way. Of course, there are many extensions to the analysis that could be performed. For example, we have assumed the standard deviation of the observation error, σ, is a known quantity. However, this is not a necessary assumption and more information may be revealed through considering an extended parameter space that includes σ. In particular, the posterior marginal in σ could provide another measure of model fitness. We have also assumed observation errors are independent both temporally and spatially. While this assumption is reasonable to ensure consistency with previous work, more general models of observation error that may include spatial or temporal correlation can also be treated in a Bayesian setting. Future work should consider both these cases of observation error modelling. It is also useful to consider correlation structures of the full joint posterior PDF to elicit connections between model parameters that cannot be identified from the marginal posterior PDFs alone. Visualisation of bivariate marginal posterior PDFs are also useful to see more details on these interactions. We have provided extended results in Section 3.7.4 that were excluded from the main chapter for clarity in our key points: 1) increasing model complexity increases uncertainty in parameter estimates; and 2) model consistency must be evaluated using multiple data sets with different initial conditions.

Jin et al. (2016b) and Chapter 2 highlight the importance of modelling the uncertainty in the initial density of cell culture assays. In particular, Jin et al. (2016b, 2017) demonstrate that variability in the initial density is not negligible across identically prepared replicates. Chapter 2 shows that this variation, if properly modelled as a random variable, greatly impacts the uncertainty in the estimates of carrying capacity, K. We extend this analysis in the context of the continuum models considered in this work and include results in Section 3.7.4. The key result is that parameter uncertainty is amplified in this more realistic, but rarely considered, case.

While we have primarily focused on model selection across different cell motility mechanisms, others have proposed models including generalisations of the source term in Equation (3.1) (Browning et al., 2017; Jin et al., 2016a; Tsoularis and Wallace, 2002), the inclusion of growth factors (Jin et al., 2016b; Sherratt and Murray, 1990), or chemotaxis (Bianchi et al., 2016). Such extensions are of interest and should be the subject of future research. We do not specifically investigate these here, though our analysis can be repeated in such cases at an increased computational expense. However, given the significant increase in parameter uncertainty incurred by the Generalised Porous Fisher model, that included only a fourth parameter, it is highly likely that scratch assay data is insufficient to provide any model certainty for these more complex extensions involving many more parameters. For example, as discussed by Johnston et al. (2015), chemotaxis amounts to setting $\mathbf{J}(\mathbf{x},t) = -D(C(\mathbf{x},t))\nabla C(\mathbf{x},t) + \chi C(\mathbf{x},t)\nabla G(\mathbf{x},t)$ in Equation (3.1), leading to

$$\frac{\partial C(\mathbf{x},t)}{\partial t} = -\nabla\cdot[-D(C(\mathbf{x},t))\nabla C(\mathbf{x},t) + \chi C(\mathbf{x},t)\nabla G(\mathbf{x},t)] + \lambda C(\mathbf{x},t)\left(1-\frac{C(\mathbf{x},t)}{K}\right),$$
$$\frac{\partial G(\mathbf{x},t)}{\partial t} = D_g\nabla^2 G(\mathbf{x},t) + k_1 C(\mathbf{x},t) - k_2 G(\mathbf{x},t),$$

where χ is the chemotatic sensitivity coefficient, $G(\mathbf{x},t) > 0$ is the concentration of a diffusive chemical signal, $D_g > 0$ is the diffusivity of the chemical, $k_1 > 0$ and $k_2 > 0$ are kinetic rate parameters for chemical production (by cells) and degradation, respectively. If $\chi < 0$, then cells are repelled by the diffusive signal, and if $\chi > 0$, then cells are attracted to it. Therefore, we have a minimum of seven parameters, with $\boldsymbol{\theta} = [D_0, \lambda, K, \chi, D_g, k_1, k_2]$. Furthermore, the model includes a single chemical species only and a realistic model would need to account for many more interacting chemical factors. Standard experimental protocols of cell culture assays do not measure this information. Cell density data alone will not be sufficient to calibrate

this model without significant levels of uncertainty. The increased number of parameters also impacts the convergence time of MCMC sampling (Gelman et al., 2014).

We do not advocate against the validity or utility of more complex models of collective cell motility and proliferation. There are many biologically based rationales for including extra biophysical and biochemical factors in a given model (Bianchi et al., 2016; Nardini et al., 2016). However, our results indicate that current *in vitro* cell culture assay data are not informative enough to distinguish between these models in practice, a point that is rarely discussed in the literature. This work is intended to motivate more detailed, Bayesian model selection within the mathematical biology community, and provide evidence that higher quality experimental methods and image analysis tools are required to validate and compare the biological hypotheses of the future.

Our results have broad implications for the mathematical biology community. Specifically, if a complex model is to be applied, then sufficient data must be collected in order to produce meaningful calibrations. Studies that compare hypotheses should also take model complexity and parameter uncertainty into account when making conclusions. The Bayesian framework, as presented here, provides tools that are designed to assist in these aspects. We suggest such techniques should be widely adopted.

3.7 Supplementary Material

3.7.1 Analysis of wave front concavity

The Generalised Porous Fisher model (Equation (3.4) in the main text) with $\lambda = 0$, in one dimension is

$$\frac{\partial C}{\partial t} = \frac{\partial}{\partial x}\left[D_0\left(\frac{C}{K}\right)^r \frac{\partial C}{\partial x}\right], \quad -\infty < x < \infty, \tag{3.16}$$

where D_0 is the free diffusivity and K the cell carrying capacity density. For the initial condition $C(x,0) = C_0\delta(x)$, Equation (3.16) has an exact solution,

$$C(x,t) = \begin{cases} \dfrac{K}{h(t)}\left[1 - \left(\dfrac{x}{d_0 h(t)}\right)^2\right]^{1/r}, & |x| \le d_0 h(t), \\ 0, & |x| > d_0 h(t), \end{cases}$$

where $d_0 = C_0\Gamma(1/r + 3/2)/(K\sqrt{\pi}\Gamma(1/r+1))$, $t_0 = d_0^2 r/(2D_0(r+2))$, $h(t) = (t/t_0)^{1/(r+2)}$ and $\Gamma(x)$ is the Gamma function. This solution, often called the source solution for the porous media equation, has compact support, $x \in [-d_0h(t), d_0h(t)]$. Here, $|x| = d_0h(t)$ are the contact points. This solution is very different to the source solution for the linear diffusion equation, $r = 0$, which is a Gaussian function without compact support (Barenblatt, 2003; Crank, 1975).

Without loss of generality, we now only consider the positive real line $x \ge 0$. The cell density is always decreasing as we approach the contact point, that is, $\partial C/\partial x < 0$ for $0 < x < d_0h(t)$. Specifically, we have

$$\frac{\partial C}{\partial x} = \frac{-2Kx}{d_0^2 h(t)^3 r}\left[1 - \left(\frac{x}{d_0 h(t)}\right)^2\right]^{1/r-1}. \tag{3.17}$$

From Equation (3.17) three different front properties are possible. As $x \to d_0h(t)$ we observe: (i) a sharp decreasing function with non-negative gradient, for $0 < r < 1$, as in Figure 3.2(e)–(h); (ii) a sharp front with finite negative slope, for $r = 1$, as in Figure 3.2(i)–(l), with $\partial C/\partial x \to -2K/(d_0^2 h(t)^3)$; and (iii) a sharp front with unbounded negative slope, for $r > 1$, as in Figure 3.2(m)–(p), with $\partial C/\partial x \to -\infty$.

To explore the concavity of the density profile, $C(x,t)$, at the contact point, it is sufficient to show how the sign of $\partial^2 C/\partial x^2$ at the contact point depends on r. The second derivative with

respect to x, for $r > 0$, is

$$\frac{\partial^2 C}{\partial x^2} = -\frac{2K}{d_0^2 h(t)^3 r}\left[1-\left(\frac{x}{d_0 h(t)}\right)^2\right]^{1/r} \times \left\{\left[1-\left(\frac{x}{d_0 h(t)}\right)^2\right]^{-1} - \frac{2x^2(1-r)}{d_0^2 h(t)^2 r}\left[1-\left(\frac{x}{d_0 h(t)}\right)^2\right]^{-2}\right\}.$$

We have, for $0 < r < 1$, that $\partial^2 C/\partial x^2 > 0$ as $x \to d_0 h(t)$. For $r \geq 1$, $\partial^2 C/\partial x^2 < 0$ as $x \to d_0 h(t)$. Hence, at the contact point, $x = d_0 h(t)$, the solution is concave down for $r \geq 1$, and concave up otherwise.

3.7.2 Numerical scheme

Here we describe our numerical scheme for the computational solution to the following reaction–diffusion equation:

$$\frac{\partial C}{\partial t} = \frac{\partial}{\partial x}\left[D(C)\frac{\partial C}{\partial x}\right] + S(C), \quad 0 < t < T, \quad 0 < x < L, \tag{3.18}$$

with initial condition,

$$C(x,t) = C_0(x), \quad t = 0,$$

and boundary conditions,

$$\frac{\partial C}{\partial x} = 0, \quad x = 0 \text{ and } x = L.$$

Consider N points in space, $\{x_i\}_{i=1}^N$, with $x_1 = 0$, $x_N = L$ and $\Delta x = x_{i+1} - x_i$ for all $i = [1, 2, \ldots, N]$. Similarly, define M temporal points, $\{t_j\}_{j=1}^M$, with $t_1 = 0$, $t_M = T$ and $\Delta t = t_{j+1} - t_j$ for all $j = [1, 2, \ldots, T]$. Next, define the notation, $C_{i+k} = C(x_i + k\Delta x, t)$, and $C_{i+k}^{j+s} = C(x_i + k\Delta x, t_j + s\Delta t)$.

Let $J(C) = -D(C)\partial C/\partial x$ and substitute into Equation (3.18) to yield

$$\frac{\partial C}{\partial t} = -\frac{\partial J}{\partial x} + S(C).$$

At the ith point, apply a first order central difference to $\partial J/\partial x$ with step $\Delta x/2$. The result is

the system of ODEs

$$\frac{\mathrm{d}C_i}{\mathrm{d}t} = -\frac{1}{\Delta x}\left[J(C_{i+1/2}) - J(C_{i-1/2})\right] + S(C_i), \quad i = 1, 2, \ldots N. \tag{3.19}$$

Similarly, a first order central difference is applied to $J(C_{i+1/2})$ and $J(C_{i-1/2})$ using the step $\Delta x/2$ yields

$$J(C_{i+1/2}) = -\frac{1}{\Delta x} D(C_{i+1/2})\left(C_{i+1} - C_i\right), \tag{3.20}$$

$$J(C_{i-1/2}) = -\frac{1}{\Delta x} D(C_{i-1/2})\left(C_i - C_{i-1}\right). \tag{3.21}$$

It is important to note that we will only obtain a solution for C_{i+k} at integer values of k, therefore the evaluation of the diffusion terms in Equation (3.20) and Equation (3.21) cannot be directly computed since $k = \pm 1/2$. We thus approximate with an averaging scheme,

$$D(C_{i+1/2}) = \frac{1}{2}\left(D(C_{i+1}) + D(C_i)\right), \tag{3.22}$$

$$D(C_{i-1/2}) = \frac{1}{2}\left(D(C_i) + D(C_{i-1})\right). \tag{3.23}$$

After substitution of Equation (3.20), Equation (3.21), Equation (3.22) and Equation (3.23) into Equation (3.19), we have the coupled system of nonlinear ODEs defined in terms of our spatial discretisation,

$$\begin{aligned}\frac{\mathrm{d}C_i}{\mathrm{d}t} = \frac{1}{2\Delta x^2}\big[&\left(D(C_{i+1}) + D(C_i)\right)\left(C_{i+1} - C_i\right) \\ &- \left(D(C_i) + D(C_{i-1})\right)\left(C_i - C_{i-1}\right)\big] + S(C_i), \quad i = 1, 2, \ldots, N.\end{aligned}$$

The no-flux boundaries are enforced using first order forward differences

$$\frac{C_1 - C_0}{\Delta x} = 0 \quad \text{and} \quad \frac{C_{N+1} - C_N}{\Delta x} = 0,$$

where C_0 and C_{N+1} represent the solution at "ghost nodes" that are not a part of the domain.

The ODEs are discretised in time using a first order backward difference method leading to the

backward-time, centered-space (BTCS) scheme,

$$\begin{aligned}
\frac{C_1^{j+1} - C_0^{j+1}}{\Delta x} &= 0, \\
\frac{C_i^{j+1} - C_i^{j}}{\Delta t} &= \frac{1}{2\Delta x^2}\left[\left(D(C_{i+1}^{j+1}) + D(C_i^{j+1})\right)\left(C_{i+1}^{j+1} - C_i^{j+1}\right)\right. \\
&\quad \left. - \left(D(C_i^{j+1}) + D(C_{i-1}^{j+1})\right)\left(C_i^{j+1} - C_{i-1}^{j+1}\right)\right] + S(C_i^{j+1}), \quad i = 1, \ldots, N, \\
\frac{C_{N+1}^{j+1} - C_N^{j+1}}{\Delta x} &= 0.
\end{aligned} \tag{3.24}$$

While this scheme is first order in time and space, it has the advantage of unconditional stability.

Since the scheme is implicit, a nonlinear root finding solver is required to compute solution at t_{j+1} given a previously computed solution at time t_j. To achieve this we apply fixed-point iteration. We re-arrange the system to be of the form $\mathbf{C}^{j+1} = \mathbf{G}(\mathbf{C}^{j+1})$ where $\mathbf{C}^{j+1} = \left[C_0^{j+1}, C_1^{j+1}, \ldots, C_{N+1}^{j+1}\right]^{\mathrm{T}}$. That is,

$$\mathbf{G}(\mathbf{C}^{j+1}) = \left[g_0(\mathbf{C}^{j+1}), g_1(\mathbf{C}^{j+1}), \ldots, g_{N+1}(\mathbf{C}^{j+1})\right],$$

where

$$\begin{aligned}
g_0(\mathbf{C}^{j+1}) &= C_1^{j+1}, \\
g_i(\mathbf{C}^{j+1}) &= C_i^{j} + \frac{\Delta t}{2\Delta x^2}\left[\left(D(C_{i+1}^{j+1}) + D(C_i^{j+1})\right)\left(C_{i+1}^{j+1} - C_i^{j+1}\right)\right. \\
&\quad \left. - \left(D(C_i^{j+1}) + D(C_{i-1}^{j+1})\right)\left(C_i^{j+1} - C_{i-1}^{j+1}\right)\right] + S(C_i^{j+1}), \quad i = 1, \ldots, N, \\
g_{N+1}(\mathbf{C}^{j+1}) &= C_N^{j+1}.
\end{aligned} \tag{3.25}$$

We then define the sequence $\{\mathbf{X}^k\}_{k\geq 0}$, generated through the nonlinear recurrence relation $\mathbf{X}^{k+1} = \mathbf{G}(\mathbf{X}^k)$ with $\mathbf{X}^0 = \mathbf{C}^j$. This sequence is iterated until $\|\mathbf{X}^{k+1} - \mathbf{X}^k\|_2 < \tau$, where τ is the error tolerance and $\|\cdot\|_2$ is the Euclidean vector norm. Once the sequence has converged, we set $\mathbf{C}^{j+1} = \mathbf{X}^{k+1}$ and continue to solve for the next time step.

For a given set of model parameters, the spatial and temporal step sizes, Δx and Δt, need to be selected. In particular, the following condition must hold to ensure accuracy, $\max_{C\in[0,K]} D(C) < \Delta x^2/\Delta t$. We then refine Δx and Δt together to ensure solutions are independent of the discretisation. Note that as r increases, higher values of $D(C)$ become

valid, therefore particular attention is required to generate Figure 3.5. The values of Δx, Δt and τ used for the simulations in this work are shown in Table 3.2. Note, that in all cases the discretisation is more refined than required to solve the given problem accurately.

Table 3.2: Discretisation and tolerance for numerical simulations.

Application	Δx [μm]	Δt [h]	τ
Example evolutions (Figure 3.2)	12.80	6.0×10^{-3}	1.0×10^{-4}
Inference on Fisher–KPP (Figure 3.4)	9.95	3.0×10^{-3}	1.0×10^{-6}
Inference on Porous Fisher (Figure 3.4)	9.95	1.5×10^{-3}	1.0×10^{-6}
Inference on Generalised Porous Fisher (Figure 3.5)	8.22	7.5×10^{-4}	1.0×10^{-6}

3.7.3 Computational inference

The Bayesian inference problems described in the main text all require the computation of the posterior PDF. Up to a normalisation constant, the posterior PDF is given by

$$p(\boldsymbol{\theta} \mid \mathcal{D}) \propto \mathcal{L}(\boldsymbol{\theta}; \mathcal{D}) p(\boldsymbol{\theta}). \tag{3.26}$$

If the posterior distribution can be sampled, the posterior PDF may be determined by using Monte Carlo integration. Thus, the main requirement is a method of generating N independent, identically distributed (i.i.d.) samples from the posterior distribution.

For many applications of practical interest, Equation (3.26) cannot be used directly to generate the samples required since the likelihood is often intractable. Approximate Bayesian computation (ABC) techniques resolve this complexity through the approximation (Sunnåker et al., 2013)

$$p(\boldsymbol{\theta} \mid \rho(\mathcal{D}, \mathcal{D}_s) < \epsilon) \propto \mathbb{P}(\rho(\mathcal{D}, \mathcal{D}_s) < \epsilon \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}), \tag{3.27}$$

where $\rho(\mathcal{D}, \mathcal{D}_s)$ is a discrepancy metric between the true data, $\mathcal{D}$, and simulated data, $\mathcal{D}_s \sim \mathcal{L}(\boldsymbol{\theta}; \mathcal{D}_s)$ and ϵ is the discrepancy threshold. ABC methods have the property that $p(\boldsymbol{\theta} \mid \rho(\mathcal{D}, \mathcal{D}_s) < \epsilon) \to p(\boldsymbol{\theta} \mid \mathcal{D})$ as $\epsilon \to 0$. This leads directly to the ABC rejection sampling algorithm (Algorithm 3.1). For deterministic models, under the assumption of Gaussian observation errors, $\epsilon/\sigma \ll 1$, and $\rho(\mathcal{D}, \mathcal{D}_s)$ taken as the sum of the squared errors, it can be shown that ABC methods are equivalent to exact posterior sampling (Wilkinson, 2013).

Algorithm 3.1 ABC rejection sampling

for $i = 1, \ldots, N$ **do**
  repeat
    Sample prior, $\boldsymbol{\theta}^* \sim p(\boldsymbol{\theta})$.
    Generate data, $\mathcal{D}_s \sim \mathcal{L}(\boldsymbol{\theta}^*; \mathcal{D}_s)$.
  until $\rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon$
  Set $\boldsymbol{\theta}^i \leftarrow \boldsymbol{\theta}^*$.
end for

In some cases, the acceptance probability in Algorithm 3.1 is computationally prohibitive for small ϵ. In such situations, an ABC extension to Markov Chain Monte Carlo sampling may be applied (Marjoram et al., 2003). The resulting ABC MCMC sampling method (Algorithm 3.2), under reasonable conditions on the proposal kernel $q(\boldsymbol{\theta}^i \mid \boldsymbol{\theta}^{i-1})$, simulates a Markov Chain with $p(\boldsymbol{\theta} \mid \rho(\mathcal{D}, \mathcal{D}_s) < \epsilon)$ (Equation (3.27)) as its stationary distribution. It is essential to simulate the Markov Chain for a sufficiently long time such that the N_T dependent samples are effectively equivalent to the required N i.i.d. samples.

Algorithm 3.2 ABC Markov chain Monte Carlo sampling

Given initial sample $\boldsymbol{\theta}^1 \sim p(\boldsymbol{\theta} \mid \rho(\mathcal{D}, \mathcal{D}_s) < \epsilon)$.
for $i = 2, \ldots, N_T$ **do**
  Sample transition kernel, $\boldsymbol{\theta}^* \sim q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{i-1})$.
  Generate data, $\mathcal{D}_s \sim \mathcal{L}(\boldsymbol{\theta}^*; \mathcal{D}_s)$.
  if $\rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon$ **then**
    Set $h \leftarrow \min\left(p(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^{i-1} \mid \boldsymbol{\theta}^*)/p(\boldsymbol{\theta}^{i-1})q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{i-1}), 1\right)$.
    Sample uniform distribution, $u \sim \mathcal{U}(0, 1)$.
    Set $\boldsymbol{\theta}^i \leftarrow \begin{cases} \boldsymbol{\theta}^*, & u \leq h, \\ \boldsymbol{\theta}^{i-1}, & u > h. \end{cases}$
  else
    Set $\boldsymbol{\theta}^i \leftarrow \boldsymbol{\theta}^{i-1}$.
  end if
end for

Using either ABC rejection sampling (Algorithm 3.1) or ABC MCMC sampling (Algorithm 3.2), we can apply Monte Carlo integration to compute the posterior PDF as given in Equation (3.27). For simplicity, we focus on the approximation of the jth marginal posterior PDF (Silverman, 1986),

$$p(\theta_j \mid \mathcal{D}) \approx \frac{1}{Nb} \sum_{i=1}^{N} K\left(\frac{\theta_j - \theta_j^{(i)}}{b}\right), \tag{3.28}$$

where θ_j is the jth element of $\boldsymbol{\theta}$, $\theta_j^{(i)}$ are the jth elements of $\boldsymbol{\theta}^{(i)} \overset{\text{i.i.d}}{\sim} p(\boldsymbol{\theta} \mid \mathcal{D})$, b is the smoothing parameter and $K(x)$ is the smoothing kernel with property $\int_{-\infty}^{\infty} K(x)\,\mathrm{d}x = 1$.

3.7.4 Additional results

In this section, we present extended results that are excluded from the main text for brevity. We provide more detailed information on the Bayesian analysis presented in Sections 3.4.2 and 3.4.3. Furthermore, we extend the Bayesian inference problem, as provided in Section 3.4.2, to account for the treatment of uncertainty in the initial condition.

3.7.4.1 Joint posterior features

Here we report various descriptive statistics for the joint posterior PDFs computed in Section 3.4. For each posterior distribution we report the posterior mode, the posterior mean, the variance/covariance matrix and the correlation coefficient matrix.

Given cell density data, $\mathcal{D}$, a set of continuum model parameters, $\boldsymbol{\theta}$, in parameter space $\boldsymbol{\Theta} \subseteq \mathbb{R}^k$ with $k > 0$, and a model implied through a likelihood function, $\mathcal{L}(\boldsymbol{\theta}; \mathcal{D})$, then summary statistics can be computed from the joint posterior, $p(\boldsymbol{\theta} \mid \mathcal{D})$, to obtain estimates and uncertainties on the true parameters. The maximum *a posteriori* (MAP) parameter estimate is the parameter set with the greatest posterior probability density as given by the posterior mode,

$$\hat{\boldsymbol{\theta}}_{\text{mode}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmax}}\, p(\boldsymbol{\theta} \mid \mathcal{D}).$$

The posterior mean is the central tendency of the parameters,

$$\bar{\boldsymbol{\theta}} = \mathbb{E}\left[\boldsymbol{\theta}\right] = \int_{\boldsymbol{\Theta}} \boldsymbol{\theta} p(\boldsymbol{\theta} \mid \mathcal{D}) \, \mathrm{d}\boldsymbol{\theta}.$$

The variance/covariance, matrix $\Sigma \in \mathbb{R}^{k \times k}$, provides information on the multivariate uncertainties, that is the spread of parameters. The (i, j)th element of Σ, denoted by $\sigma_{i,j}$, is given by

$$\sigma_{i,j} = \mathbb{C}\left[\theta_i, \theta_j\right] = \mathbb{E}\left[(\theta_i - \mathbb{E}\left[\theta_j\right])(\theta_j - \mathbb{E}\left[\theta_i\right])\right] = \int_{\boldsymbol{\Theta}} (\theta_i - \mathbb{E}\left[\theta_j\right])(\theta_j - \mathbb{E}\left[\theta_i\right]) p(\boldsymbol{\theta} \mid \mathcal{D}) \, \mathrm{d}\boldsymbol{\theta},$$

where θ_i and θ_j are the ith and jth elements of $\boldsymbol{\theta}$. Note that $\mathbb{C}\left[\theta_i, \theta_i\right] = \mathbb{V}\left[\theta_i\right]$ and $\sigma_{i,j} = \sigma_{j,i}$. Lastly, the correlation coefficient matrix $R \in \mathbb{R}^{k \times k}$ measures the linear dependence between

parameter pairs. The (i, j)th element of R, denoted by $\rho_{i,j}$, is given by

$$\rho_{i,j} = \frac{\sigma_{i,j}}{(\sigma_{i,i}, \sigma_{j,j})^{1/2}}.$$

Note $\rho_{i,i} = 1$ for all $i \in [1, k]$, and $\rho_{i,j} = \rho_{j,i}$. The results of all these statistics, for the inference problems considered in the main text, are presented in Tables 3.3, 3.4, 3.5, and 3.6.

Table 3.3: MAP parameter estimates (posterior modes) from posterior PDFs using initial conditions of $12,000$ cells, $16,000$ cells and $20,000$ cells.

Inference problem		MAP parameter estimate (posterior mode)			
Model	Initial condition	D_0 [μm^2/h]	λ [1/h]	K [cells/μm^2]	r [$-$]
Fisher–KPP	$12,000$	325	3.05×10^{-2}	1.85×10^{-3}	-
Fisher–KPP	$16,000$	1,010	3.14×10^{-2}	1.49×10^{-3}	-
Fisher–KPP	$20,000$	1,354	3.56×10^{-2}	1.65×10^{-3}	-
Porous Fisher	$12,000$	3,425	3.17×10^{-2}	2.58×10^{-3}	-
Porous Fisher	$16,000$	3,134	3.77×10^{-2}	1.56×10^{-3}	-
Porous Fisher	$20,000$	4,933	5.32×10^{-2}	1.69×10^{-3}	-
Generalised Porous Fisher	$12,000$	7,857	2.66×10^{-2}	1.01×10^{-2}	1.37
Generalised Porous Fisher	$16,000$	90,234	2.43×10^{-2}	7.18×10^{-3}	1.58
Generalised Porous Fisher	$20,000$	11,990	3.46×10^{-2}	2.10×10^{-3}	1.65

Table 3.4: Joint posterior means for posterior PDFs using initial conditions of $12,000$ cells, $16,000$ cells and $20,000$ cells.

Inference problem		Joint posterior mean			
Model	Initial condition	D_0 [μm^2/h]	λ [1/h]	K [cells/μm^2]	r [$-$]
Fisher–KPP	$12,000$	457	3.58×10^{-2}	3.21×10^{-3}	-
Fisher–KPP	$16,000$	847	4.29×10^{-2}	2.20×10^{-3}	-
Fisher–KPP	$20,000$	1,444	5.78×10^{-2}	1.88×10^{-3}	-
Porous Fisher	$12,000$	3,249	3.44×10^{-2}	2.86×10^{-3}	-
Porous Fisher	$16,000$	3,279	4.23×10^{-2}	1.95×10^{-3}	-
Porous Fisher	$20,000$	3,347	6.13×10^{-2}	1.75×10^{-3}	-
Generalised Porous Fisher	$12,000$	103,682	2.76×10^{-2}	2.40×10^{-2}	1.80
Generalised Porous Fisher	$16,000$	23,556	3.52×10^{-2}	2.55×10^{-3}	2.02
Generalised Porous Fisher	$20,000$	4,845	4.77×10^{-2}	1.89×10^{-3}	0.94

Table 3.5: Variance/covariance matricies for posterior PDFs using initial conditions of $12,000$ cells, $16,000$ cells and $20,000$ cells.

Inference problem			Variance/Covariance Matrix			
Model	Initial condition		D_0	λ	K	r
Fisher–KPP	$12,000$	D_0	5.32×10^{4}	-5.60×10^{-1}	8.68×10^{-2}	-
		λ	-5.60×10^{-1}	5.93×10^{-5}	-8.22×10^{-6}	-
		K	8.68×10^{-2}	-8.22×10^{-6}	1.84×10^{-6}	-
		r	-	-	-	-
Fisher–KPP	$16,000$	D_0	1.48×10^{5}	-3.26	1.57×10^{-1}	-
		λ	-3.26	1.51×10^{-4}	-6.93×10^{-6}	-
		K	1.57×10^{-1}	-6.93×10^{-6}	4.93×10^{-7}	-
		r	-	-	-	-
Fisher–KPP	$20,000$	D_0	4.41×10^{5}	-1.23×10^{1}	1.47×10^{-1}	-
		λ	-1.23×10^{1}	4.72×10^{-4}	-5.44×10^{-6}	-
		K	1.47×10^{-1}	-5.44×10^{-6}	9.29×10^{-8}	-
		r	-	-	-	-
Porous Fisher	$12,000$	D_0	1.14×10^{6}	-3.03	5.95×10^{-1}	-
		λ	-3.03	1.48×10^{-5}	-2.55×10^{-6}	-
		K	5.95×10^{-1}	-2.55×10^{-6}	5.90×10^{-7}	-
		r	-	-	-	-
Porous Fisher	$16,000$	D_0	1.05×10^{6}	-5.71	2.29×10^{-1}	-
		λ	-5.71	5.25×10^{-5}	-1.96×10^{-6}	-
		K	2.29×10^{-1}	-1.96×10^{-6}	1.04×10^{-7}	-
		r	-	-	-	-
Porous Fisher	$20,000$	D_0	1.14×10^{6}	-1.51×10^{1}	1.23×10^{-1}	-
		λ	-1.51×10^{1}	2.61×10^{-4}	-2.16×10^{-6}	-
		K	1.23×10^{-1}	-2.16×10^{-6}	2.89×10^{-8}	-
		r	-	-	-	-
Generalised Porous Fisher	$12,000$	D_0	6.22×10^{9}	-1.15×10^{2}	6.24×10^{2}	6.46×10^{3}
		λ	-1.15×10^{2}	3.05×10^{-5}	-6.53×10^{-5}	4.87×10^{-3}
		K	6.24×10^{2}	-6.53×10^{-5}	6.80×10^{-4}	-1.45×10^{-2}
		r	6.46×10^{3}	4.87×10^{-3}	-1.45×10^{-2}	1.32
Generalised Porous Fisher	$16,000$	D_0	3.20×10^{8}	-8.04×10^{1}	6.54	7.23×10^{3}
		λ	-8.04×10^{1}	8.55×10^{-5}	-7.50×10^{-6}	2.40×10^{-3}
		K	6.54	-7.50×10^{-6}	1.12×10^{-6}	-3.82×10^{-4}
		r	7.23×10^{3}	2.40×10^{-3}	-3.82×10^{-4}	7.72×10^{-1}
Generalised Porous Fisher	$20,000$	D_0	5.96×10^{6}	-1.54×10^{1}	1.87×10^{-1}	8.97×10^{2}
		λ	-1.54×10^{1}	6.69×10^{-5}	-8.99×10^{-7}	-1.96×10^{-3}
		K	1.87×10^{-1}	-8.99×10^{-7}	1.46×10^{-8}	2.04×10^{-5}
		r	8.97×10^{2}	-1.96×10^{-3}	2.04×10^{-5}	1.66×10^{-1}

Table 3.6: Correlation coefficient matricies for posterior PDFs using initial conditions of $12,000$ cells, $16,000$ cells and $20,000$ cells.

Inference problem			Correlation coefficient Matrix			
Model	Initial condition		D_0	λ	K	r
Fisher–KPP	$12,000$	D_0	1.00	-3.15×10^{-1}	2.77×10^{-1}	-
		λ	-3.15×10^{-1}	1.00	-7.86×10^{-1}	-
		K	2.77×10^{-1}	-7.86×10^{-1}	1.00	-
		r	-	-	-	-
Fisher–KPP	$16,000$	D_0	1.00	-6.90×10^{-1}	5.83×10^{-1}	-
		λ	-6.90×10^{-1}	1.00	-8.04×10^{-1}	-
		K	5.83×10^{-1}	-8.04×10^{-1}	1.00	-
		r	-	-	-	-
Fisher–KPP	$20,000$	D_0	1.00	-8.50×10^{-1}	7.27×10^{-1}	-
		λ	-8.50×10^{-1}	1.00	-8.21×10^{-1}	-
		K	7.27×10^{-1}	-8.21×10^{-1}	1.00	-
		r	-	-	-	-
Porous Fisher	$12,000$	D_0	1.00	-7.36×10^{-1}	7.25×10^{-1}	-
		λ	-7.36×10^{-1}	1.00	-8.61×10^{-1}	-
		K	7.25×10^{-1}	-8.61×10^{-1}	1.00	-
		r	-	-	-	-
Porous Fisher	$16,000$	D_0	1.00	-7.68×10^{-1}	6.90×10^{-1}	-
		λ	-7.68×10^{-1}	1.00	-8.36×10^{-1}	-
		K	6.90×10^{-1}	-8.36×10^{-1}	1.00	-
		r	-	-	-	-
Porous Fisher	$20,000$	D_0	1.00	-8.76×10^{-1}	6.80×10^{-1}	-
		λ	-8.76×10^{-1}	1.00	-7.87×10^{-1}	-
		K	6.80×10^{-1}	-7.87×10^{-1}	1.00	-
		r	-	-	-	-
Generalised Porous Fisher	$12,000$	D_0	1.00	-2.64×10^{-1}	3.03×10^{-1}	7.12×10^{-2}
		λ	-2.64×10^{-1}	1.00	-4.53×10^{-1}	7.67×10^{-1}
		K	3.03×10^{-1}	-4.53×10^{-1}	1.00	-4.83×10^{-1}
		r	7.12×10^{-2}	7.67×10^{-1}	-4.83×10^{-1}	1.00
Generalised Porous Fisher	$16,000$	D_0	1.00	-4.86×10^{-1}	3.45×10^{-1}	4.60×10^{-1}
		λ	-4.86×10^{-1}	1.00	-7.65×10^{-1}	2.96×10^{-1}
		K	3.45×10^{-1}	-7.65×10^{-1}	1.00	-4.10×10^{-1}
		r	4.60×10^{-1}	2.96×10^{-1}	-4.10×10^{-1}	1.00
Generalised Porous Fisher	$20,000$	D_0	1.00	-7.72×10^{-1}	6.35×10^{-1}	9.03×10^{-1}
		λ	-7.72×10^{-1}	1.00	-9.10×10^{-1}	-5.88×10^{-1}
		K	6.35×10^{-1}	-9.10×10^{-1}	1.00	4.15×10^{-1}
		r	9.03×10^{-1}	-5.88×10^{-1}	4.15×10^{-1}	1.00

3.7.4.2 Bivariate marginal posterior PDFs

In the main text we computed only univariate marginal posterior PDFs, we extend this analysis by providing bivariate marginal PDFs here. For the Fisher–KPP and Porous Fisher models we have three bivariate marginal posterior PDFs,

$$
\begin{aligned}
p(D_0, \lambda \mid \mathbf{C}_{\text{obs}}^{1:N,1:M}) &= \int_{\mathbb{R}} p(D_0, \lambda, K \mid \mathbf{C}_{\text{obs}}^{1:N,1:M})\, \mathrm{d}K, \\
p(\lambda, K \mid \mathbf{C}_{\text{obs}}^{1:N,1:M}) &= \int_{\mathbb{R}} p(D_0, \lambda, K \mid \mathbf{C}_{\text{obs}}^{1:N,1:M})\, \mathrm{d}D_0, \\
p(D_0, K \mid \mathbf{C}_{\text{obs}}^{1:N,1:M}) &= \int_{\mathbb{R}} p(D_0, \lambda, K \mid \mathbf{C}_{\text{obs}}^{1:N,1:M})\, \mathrm{d}\lambda.
\end{aligned}
$$

Similarly, for the Generalised Porous Fisher Model, we have six bivariate marginal posterior PDFs,

$$
\begin{aligned}
p(D_0, \lambda \mid \mathbf{C}_{\text{obs}}^{1:N,1:M}) &= \iint_{\mathbb{R}^2} p(D_0, \lambda, K, r \mid \mathbf{C}_{\text{obs}}^{1:N,1:M})\, \mathrm{d}r\, \mathrm{d}K, \\
p(D_0, K \mid \mathbf{C}_{\text{obs}}^{1:N,1:M}) &= \iint_{\mathbb{R}^2} p(D_0, \lambda, K, r \mid \mathbf{C}_{\text{obs}}^{1:N,1:M})\, \mathrm{d}r\, \mathrm{d}\lambda, \\
p(D_0, r \mid \mathbf{C}_{\text{obs}}^{1:N,1:M}) &= \iint_{\mathbb{R}^2} p(D_0, \lambda, K, r \mid \mathbf{C}_{\text{obs}}^{1:N,1:M})\, \mathrm{d}K\, \mathrm{d}\lambda, \\
p(\lambda, K \mid \mathbf{C}_{\text{obs}}^{1:N,1:M}) &= \iint_{\mathbb{R}^2} p(D_0, \lambda, K, r \mid \mathbf{C}_{\text{obs}}^{1:N,1:M})\, \mathrm{d}r\, \mathrm{d}D_0, \\
p(\lambda, r \mid \mathbf{C}_{\text{obs}}^{1:N,1:M}) &= \iint_{\mathbb{R}^2} p(D_0, \lambda, K, r \mid \mathbf{C}_{\text{obs}}^{1:N,1:M})\, \mathrm{d}K\, \mathrm{d}D_0, \\
p(K, r \mid \mathbf{C}_{\text{obs}}^{1:N,1:M}) &= \iint_{\mathbb{R}^2} p(D_0, \lambda, K, r \mid \mathbf{C}_{\text{obs}}^{1:N,1:M})\, \mathrm{d}\lambda\, \mathrm{d}D_0.
\end{aligned}
$$

The resulting PDFs using the three initial density conditions are shown for: the Fisher–KPP model (Figure 3.6); the Porous Fisher model (Figure 3.7); and the Generalised Porous model (Figure 3.8).

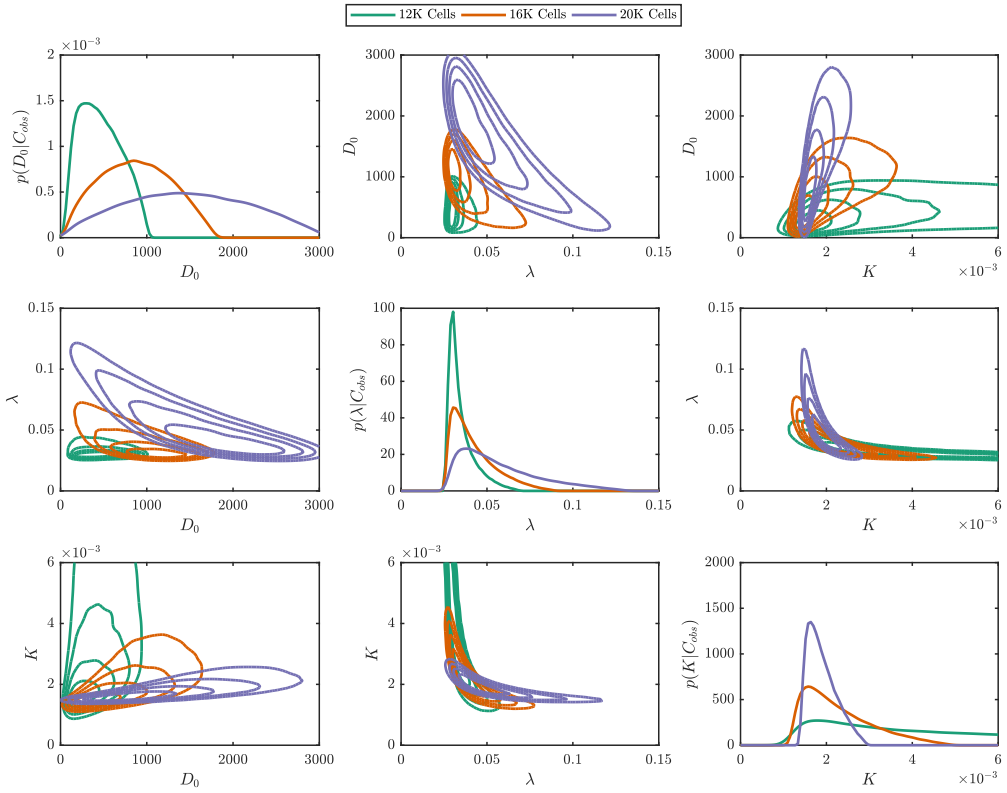

Figure 3.6: Plot matrix of univariate and bivariate marginal posterior probability densities obtained through Bayesian inference on the PC-3 scratch assay data under the Fisher–KPP model for the three different initial conditions; $12,000$ cells (solid green), $16,000$ cells (solid orange), $20,000$ cells (solid purple). Univariate marginal densities, on the main plot matrix diagonal, demonstrate the degree of uncertainty in the diffusivity, D_0, the proliferation rate, λ, and the carrying capacity, K. Off diagonals are contour plots of the pairwise bivariate posterior PDFs, these demonstrate the relationships between parameters.

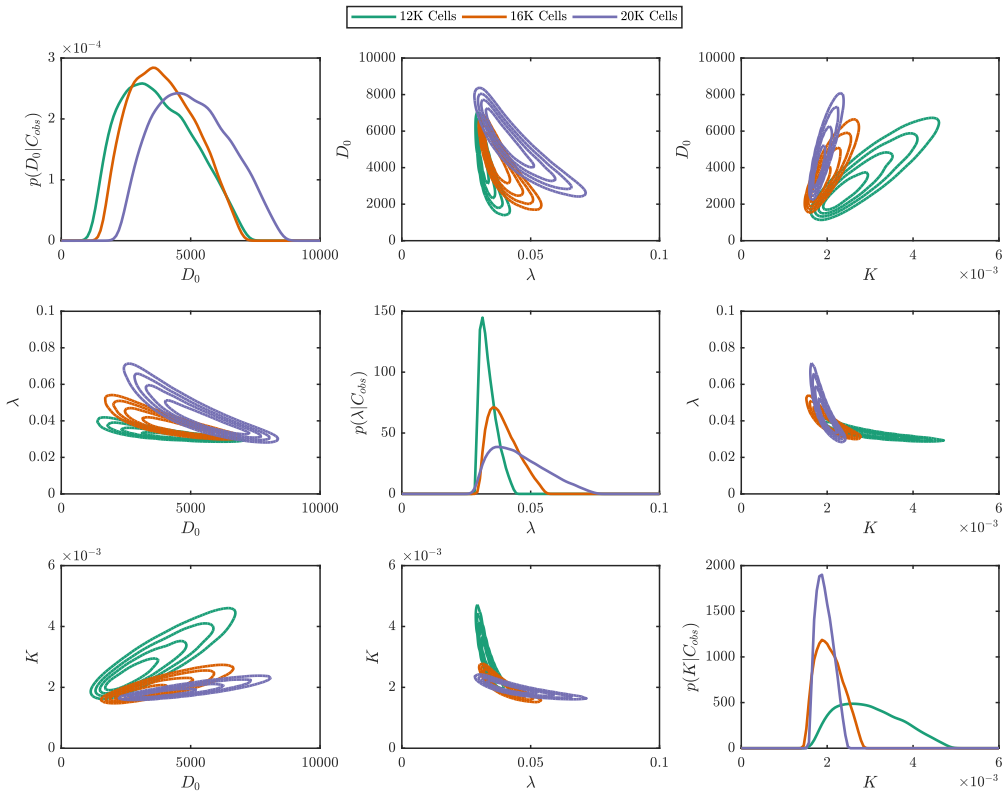

Figure 3.7: Plot matrix of univariate and bivariate marginal posterior probability densities obtained through Bayesian inference on the PC-3 scratch assay data under the Porous Fisher model for the three different initial conditions; $12,000$ cells (solid green), $16,000$ cells (solid orange), $20,000$ cells (solid purple). Univariate marginal densities, on the main plot matrix diagonal, demonstrate the degree of uncertainty in the diffusivity, D_0, the proliferation rate, λ, and the carrying capacity, K. Off diagonals are contour plots of the pairwise bivariate posterior PDFs, these demonstrate the relationships between parameters.

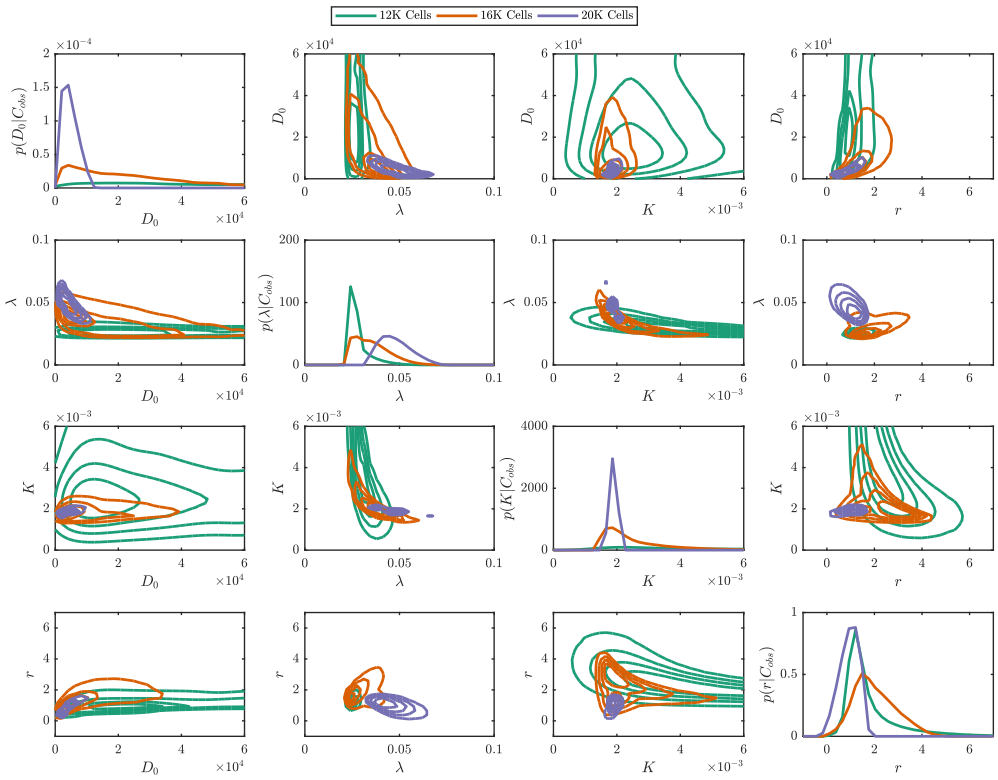

Figure 3.8: Plot matrix of univariate and bivariate marginal posterior probability densities obtained through Bayesian inference on the PC-3 scratch assay data under the Generalised Porous Fisher model for the three different initial conditions; $12,000$ cells (solid green), $16,000$ cells (solid orange), $20,000$ cells (solid purple). Univariate marginal densities, on the main plot matrix diagonal, demonstrate the degree of uncertainty in the diffusivity, D_0, the proliferation rate, λ, the carrying capacity, K, and the power, r. Off diagonals are contour plots of the pairwise bivariate posterior PDFs, these demonstrate the relationships between parameters.

3.7.4.3 Uncertainty in initial condition

In the main text, the assumption was made that $C_{\text{obs}}(x,0) = C(x,0;\boldsymbol{\theta})$. That is, we use initial observations as the initial density profile to simulate the model given parameters $\boldsymbol{\theta}$. Since the model is deterministic, the final form of the likelihood is a multivariate Gaussian distribution, which simplifies calculations considerably. Both Jin et al. (2016b) and Chapter 2 indicate that such an assumption could result in underestimation of the uncertainties in parameter estimates.

Following from Chapter 2, we take $C_{\text{obs}}(x,0) = C(x,0;\boldsymbol{\theta}) + \eta_0$, where η_0 is a Gaussian random variable with mean $C(x,0;\boldsymbol{\theta})$ and variance σ_0^2. Note that we do not require $\sigma_0 = \sigma$, in fact, there are reasons to consider $\sigma_0 > \sigma$, for example, experimental protocols for seeding cell culture plates can be an additional source of variation in initial cell densities (Jin et al., 2016b) (see Chapter 2). Since $C_{\text{obs}}(x,0) \sim \mathcal{N}(C(x,0;\boldsymbol{\theta}), \sigma_0^2)$, it is also true that $C(x,0;\boldsymbol{\theta}) \sim \mathcal{N}(C_{\text{obs}}(x,0), \sigma_0^2)$. Therefore, our models are to be treated as random PDEs with deterministic dynamics, but random initial conditions.

Since the initial conditions are random, the initial condition is a latent variable that must be integrated out. Thus, the likelihood becomes

$$\mathcal{L}(\mathbf{C}_{\text{obs}}^{1:N,1:M};\boldsymbol{\theta}) = \int_{\mathbb{R}^N} \prod_{i=1}^{N}\prod_{j=1}^{M} \frac{e^{-(C_{\text{obs}}(x_i,t_j)-C(x_i,t_j;\boldsymbol{\theta}))^2/(2\sigma^2)}}{\sigma\sqrt{2\pi}} p(C(x_i,0;\boldsymbol{\theta}) \mid \sigma_0)\,\mathrm{d}C(x_i,0;\boldsymbol{\theta}),$$

where σ_0 is assumed to be known and $p(C(x_i,0;\boldsymbol{\theta}) \mid \sigma_0)$ is a Gaussian PDF with mean $C_{\text{obs}}(x_i,0)$ and variance σ_0. This likelihood integral must be computed using Monte Carlo methods. Computationally, we apply directly the ABC MCMC method as given in Algorithm 3.2. The only algorithmic difference being that simulated data, $\mathcal{D}_s$, is generated though solving the model PDE after a realisation of the initial density profile has been generated. Overall, this leads to slower convergence in the Markov chain and hence longer computation times.

The inference problem using random initial density profiles was solved using ABC MCMC under the Fisher–KPP model and the Porous Fisher model for initial densities based on $16,000$ initial cells only. We take $\sigma_0 = 2\sigma$. Univariate and bivariate marginal posterior PDFs are shown in Figures 3.9 and 3.10. In the Fisher–KPP model, the additional uncertainty seems to have a significant effect on the uncertainty in the carrying capacity, K, in agreement with Chapter 2. However, the diffusion coefficient, D_0, and proliferation rate, λ, are not affected as significantly.

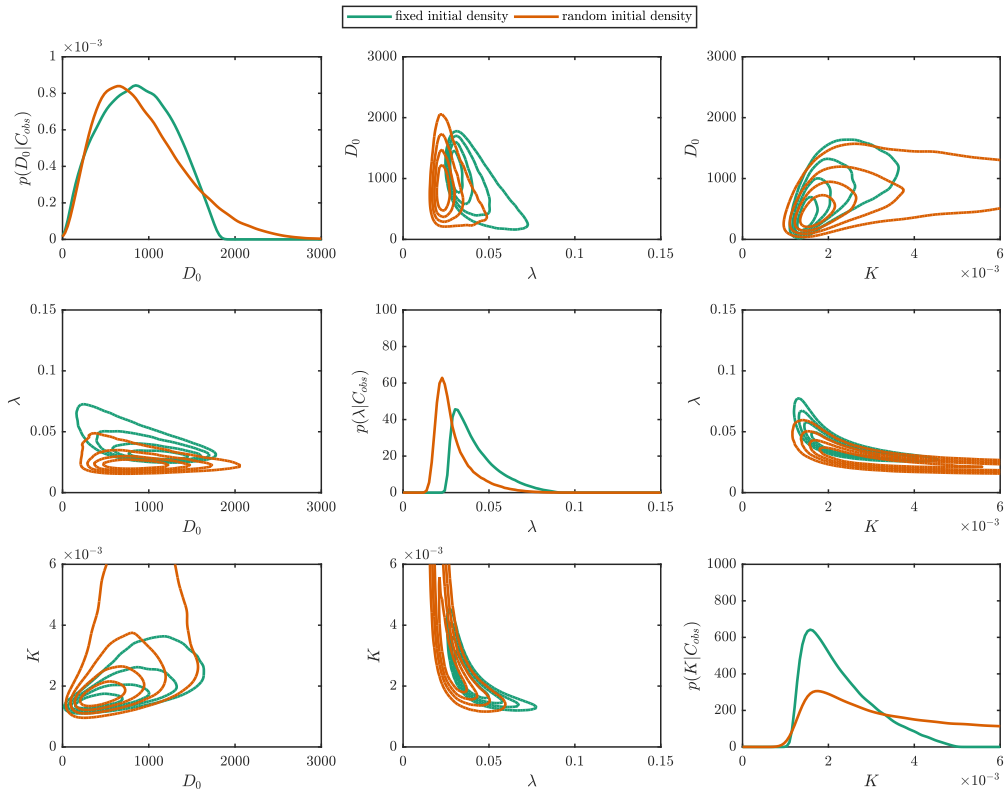

Figure 3.9: Plot matrix of univariate and bivariate marginal posterior probability densities obtained through Bayesian inference on the PC-3 scratch assay data under the Fisher–KPP model for the fixed initial conditions (solid green) and random initial conditions (solid orange). Univariate marginal densities, on the main plot matrix diagonal, demonstrate the degree of uncertainty in the diffusivity, D_0, the proliferation rate, λ, and the carrying capacity, K. Off diagonals are contour plots of the pairwise bivariate posterior PDFs, these demonstrate the relationships between parameters.

For the Porous Fisher model, both D_0 and K are greatly affected. This is not surprising, since motility is density dependent for the Porous Fisher model. By contrast the Fisher–KPP model is almost unaffected in the marginal posterior PDF of D_0, since it is independent of initial cell density. We did not apply this analysis to the Generalised Porous Fisher model.

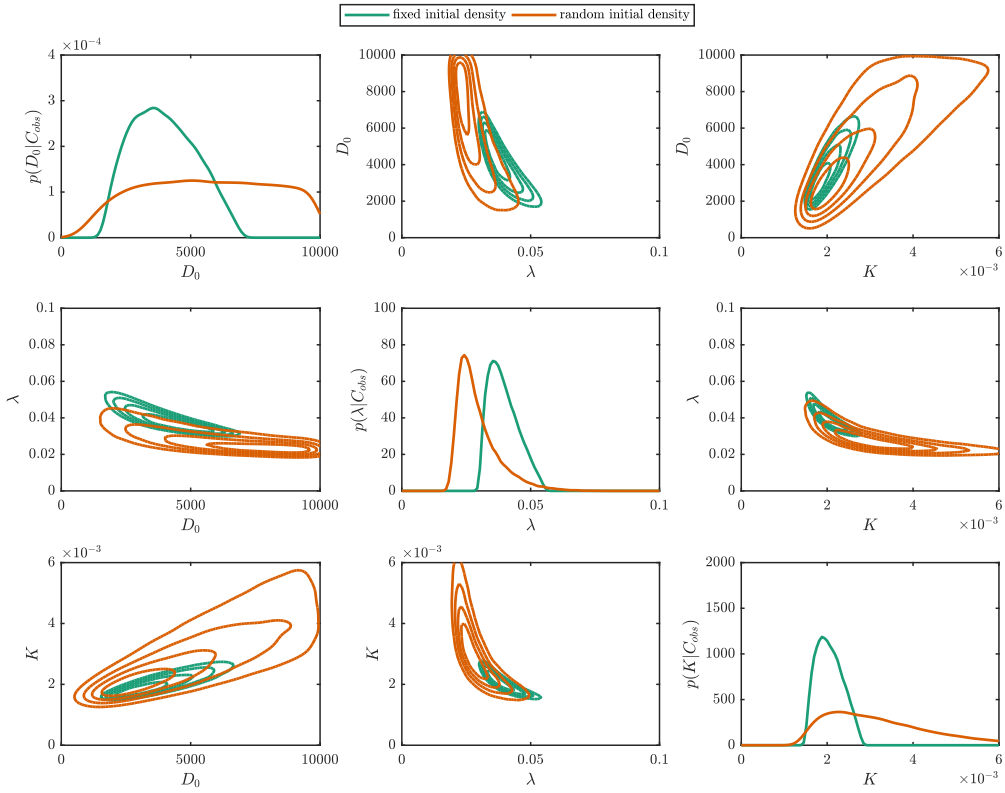

Figure 3.10: Plot matrix of univariate and bivariate marginal posterior probability densities obtained through Bayesian inference on the PC-3 scratch assay data under the Porous Fisher model for the fixed initial conditions (solid green) and random initial conditions (solid orange). Univariate marginal densities, on the main plot matrix diagonal, demonstrate the degree of uncertainty in the diffusivity, D_0, the proliferation rate, λ, and the carrying capacity, K. Off diagonals are contour plots of the pairwise bivariate posterior PDFs, these demonstrate the relationships between parameters.

Part II

Methodological Developments in Computational Inference for Stochastic Models

Chapter 4

Simulation and Inference Algorithms for Stochastic Biochemical Reaction Networks: From Basic Concepts to State-of-the-Art

A paper published in *Journal of the Royal Society Interface*.

David J. Warne, Ruth E. Baker and Matthew J. Simpson (2019). Simulation and inference algorithms for stochastic biochemical reaction networks: from basic concepts to state-of-the-art. *Journal of the Royal Society Interface*, 16:20180943. DOI:10.1098/rsif.2018.0943

Abstract Stochasticity is a key characteristic of intracellular processes such as gene regulation and chemical signalling. Therefore, characterising stochastic effects in biochemical systems is essential to understand the complex dynamics of living things. Mathematical idealisations of biochemically reacting systems must be able to capture stochastic phenomena. While robust theory exists to describe such stochastic models, the computational challenges in exploring these models can be a significant burden in practice since realistic models are analytically intractable. Determining the expected behaviour and variability of a stochastic biochemical

reaction network requires many probabilistic simulations of its evolution. Using a biochemical reaction network model to assist in the interpretation of time course data from a biological experiment is an even greater challenge due to the intractability of the likelihood function for determining observation probabilities. These computational challenges have been subjects of active research for over four decades. In this review, we present an accessible discussion of the major historical developments and state-of-the-art computational techniques relevant to simulation and inference problems for stochastic biochemical reaction network models. Detailed algorithms for particularly important methods are described and complemented with MATLAB® implementations. As a result, this review provides a practical and accessible introduction to computational methods for stochastic models within the life sciences community.

4.1 Introduction

Many biochemical processes within living cells, such as regulation of gene expression, are stochastic (Abkowitz et al., 1996; Arkin et al., 1998; Kærn et al., 2005; McAdams and Arkin, 1997; Raj and van Oudenaarden, 2008); that is, randomness or noise is an essential component of living things. Internal and external factors are responsible for this randomness (Elowitz et al., 2002; Keren et al., 2015; Soltani et al., 2016; Taniguchi et al., 2010), particularly within systems where low copy numbers of certain chemical species greatly affect the system dynamics (Fedoroff and Fontana, 2002). Intracellular stochastic effects are key components of normal cellular function (Eldar and Elowitz, 2010) and have a direct influence on the heterogeneity of multicellular organisms (Smith and Grima, 2018). Furthermore, stochasticity of biochemical processes can play a role in the onset of disease (Feinberg, 2014; Gupta et al., 2011) and immune responses (Satija and Shalek, 2014). Stochastic phenomena, such as resonance (Moss and Pei, 1995), focussing (Paulsson et al., 2000), and bistability (Bressloff, 2017; Thattai and van Oudenaarden, 2001; Tian and Burrage, 2006), are not captured by traditional deterministic chemical rate equation models. These stochastic effects must be captured by appropriate theoretical models. A standard approach is to consider a biochemical reaction network as a well-mixed population of molecules that diffuse, collide and react probabilistically. The stochastic law of mass action is invoked to determine the probabilities of reaction events over time (Gillespie, 1977; Wilkinson, 2012). The resulting time-series of biochemical populations may be analysed to determine both the average behaviour and variability (Schnoerr et al., 2017). This powerful

approach to modelling biochemical kinetics can be extended to deal with more biologically realistic settings that include spatial heterogeneity with molecular populations being well-mixed only locally (Erban and Chapman, 2009; Fange et al., 2010; Isaacson, 2008; Turner et al., 2004).

In practice, stochastic biochemical reaction network models are analytically intractable meaning that most approaches are entirely computational. Two distinct, yet related, computational problems are of particular importance: (i) the *forwards* problem that deals with the simulation of the evolution of a biochemical reaction network forwards in time; and (ii) the *inverse* problem that deals with the inference of unknown model parameters given time-course observations. Over the last four decades, significant attention has been given to these problems. Gillespie et al. (2013) describe the key algorithmic advances in the history of the forwards problem and Higham (2008) provides an accessible introduction connecting stochastic approaches with deterministic counterparts. Recently, Schnoerr et al. (2017) provide a detailed review of the forwards problem with a focus on analytical methods. Golightly and Wilkinson (2011), Toni et al. (2009) and Sunnåker et al. (2013) review techniques relevant to the inverse problem.

Given the relevance of stochastic computational methods to the life sciences, the aim of this review is to present an accessible summary of computational aspects relating to efficient simulation for both the forwards and inverse problems. Practical examples and algorithmic descriptions are presented and aimed at applied mathematicians and applied statisticians with interests in the life sciences. However, we expect the techniques presented here will also be of interest to the wider life sciences community. Supplementary material provides clearly documented code examples (available from GitHub https://github.com/ProfMJSimpson/Warne2018) using the MATLAB® programming language.

4.2 Biochemical reaction networks

We provide an algorithmic introduction to stochastic biochemical reaction network models. In the literature, rigorous theory exists for these stochastic modelling approaches (Gillespie, 1992). However, we focus on an informal definition useful for understanding computational methods in practice. Relevant theory on the chemical master equation, Markov processes and stochastic differential equations is not discussed in any detail (See Erban et al. (2007), Higham (2001) and Wilkinson (2012) for accessible introductions to these topics).

Consider a domain, for example, a cell nucleus, that contains a number of *chemical species*. The population count for a chemical species is a non-negative integer called its *copy number*. A biochemical reaction network model is specified by a set of chemical formulae that determine how the chemical species interact. For example, $X+2Y \rightarrow Z+W$, states "one X molecule and two Y molecules react to produce one Z molecule and one W molecule". If a chemical species is involved in a reaction, then the number of molecules required as reactants or produced as products are called *stoichiometric coefficients*. In the example, Y has a reactant stoichiometric coefficient of two, and Z has a product stoichiometric coefficient of one.

4.2.1 A computational definition

Consider a set of M reactions involving N chemical species with copy numbers $X_1(t), \ldots, X_N(t)$ at time t. The state vector is an $N \times 1$ vector of copy numbers, $\mathbf{X}(t) = [X_1(t), \ldots, X_N(t)]^T$. This represents the state of the population of chemical species at time t. When a reaction occurs, the copy numbers of the reactants and products are altered according to their respective stoichiometric coefficients. The net state change caused by a reaction event is called its stoichiometric vector. If reaction j occurs, then a new state is obtained by adding its stoichiometric vector, $\boldsymbol{\nu}_j$, that is,

$$\mathbf{X}(t) = \mathbf{X}(t^-) + \boldsymbol{\nu}_j, \tag{4.1}$$

where t^- denotes the time immediately preceding the reaction event. The vectors $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_M$ are obtained through $\boldsymbol{\nu}_j = \boldsymbol{\nu}_j^+ - \boldsymbol{\nu}_j^-$, where $\boldsymbol{\nu}_j^-$ and $\boldsymbol{\nu}_j^+$ are, respectively, vectors of the reactant and product stoichiometric coefficients of the chemical formula of reaction j. Equation (4.1) describes how reaction j affects the system.

Gillespie (1977, 1992) presents the fundamental theoretical framework that provides a probabilistic rule for the occurrence of reaction events. We shall not focus on the details here, but the essential concept is based on the stochastic law of mass action. Informally,

$$\mathbb{P}(\text{Reaction } j \text{ occurs in } [t, t+\mathrm{d}t)) \propto \mathrm{d}t \times \#\text{ of possible reactant combinations.} \tag{4.2}$$

The tacit assumption is that the system is well-mixed with molecules equally likely to be found anywhere in the domain. The right hand side of Equation (4.2) is typically expressed

as $a_j(\mathbf{X}(t))\mathrm{d}t$, where $a_j(\mathbf{X}(t))$ is the *propensity function* of reaction j. That is,

$$a_j(\mathbf{X}(t)) = \text{constant } \times \text{ total combinations in } \mathbf{X}(t) \text{ for reaction } j, \tag{4.3}$$

where the positive constant is known as the *kinetic rate parameter*[1]. Equations (4.1)–(4.3) are the main concepts needed to consider computational methods for the forwards problem. Importantly, Equations (4.1) and (4.2) indicate that the possible model states are discrete, but state changes occur in continuous time.

4.2.2 Two examples

We now provide some representative examples of biochemical reaction networks that will be used throughout this chapter.

4.2.2.1 Mono-molecular chain

Consider two chemical species, A and B, and three reactions that form a mono-molecular chain,

$$\underbrace{\emptyset \xrightarrow{k_1} A}_{\substack{\text{external production} \\ \text{of } A \text{ molecules}}} \quad , \qquad \underbrace{A \xrightarrow{k_2} B}_{\substack{\text{decay of } A \text{ molecules} \\ \text{into } B \text{ molecules}}} \quad , \qquad \underbrace{B \xrightarrow{k_3} \emptyset}_{\substack{\text{external consumption} \\ \text{of } B \text{ molecules}}} \quad , \tag{4.4}$$

with kinetic rate parameters k_1, k_2, and k_3. We adopt the convention that $\emptyset$ indicates the reactions are part of an open system involving external chemical processes that are not explicitly represented in Equation (4.4). Given the state vector, $\mathbf{X}(t) = [A(t), B(t)]^T$, the respective propensity functions are

$$a_1(\mathbf{X}(t)) = k_1, \quad a_2(\mathbf{X}(t)) = k_2 A(t), \quad a_3(\mathbf{X}(t)) = k_3 B(t), \tag{4.5}$$

[1]These are not "rates" but scale factors on reaction event probabilities. A "slow" reaction with low kinetic rate may still occur rapidly, but the probability of this event is low.

and the stoichiometric vectors are

$$\boldsymbol{\nu}_1 = \begin{bmatrix} 1 \\ \\ 0 \end{bmatrix}, \quad \boldsymbol{\nu}_2 = \begin{bmatrix} -1 \\ \\ 1 \end{bmatrix}, \quad \boldsymbol{\nu}_3 = \begin{bmatrix} 0 \\ \\ -1 \end{bmatrix}. \tag{4.6}$$

This mono-molecular chain is interesting since it is a part of general class of biochemical reaction networks that are analytically tractable, though they are only applicable to relatively simple biochemical processes (Jahnke and Huisinga, 2007).

4.2.2.2 Enzyme kinetics

A biologically applicable biochemical reaction network describes the catalytic conversion of a substrate, S, into a product, P, via an enzymatic reaction involving enzyme, E. This is described by Michaelis–Menten enzyme kinetics (Michaelis and Menten, 1913; Rao and Arkin, 2003),

$$\underbrace{E + S \xrightarrow{k_1} C}_{\substack{\text{enzyme and substrate molecules} \\ \text{combine to form a complex}}}, \quad \underbrace{C \xrightarrow{k_2} E + S}_{\text{decay of a complex}}, \quad \underbrace{C \xrightarrow{k_3} E + P}_{\substack{\text{catalytic conversion} \\ \text{of substrate to product}}}, \tag{4.7}$$

with kinetic rate parameters, k_1, k_2, and k_3. This particular enzyme kinetic model is a closed system. Here we have the state vector $\mathbf{X}(t) = [E(t), S(t), C(t), P(t)]^T$, propensity functions

$$a_1(\mathbf{X}(t)) = k_1 E(t) S(t), \quad a_2(\mathbf{X}(t)) = k_2 C(t), \quad a_3(\mathbf{X}(t)) = k_3 C(t), \tag{4.8}$$

and stoichiometric vectors

$$\boldsymbol{\nu}_1 = \begin{bmatrix} -1 \\ \\ -1 \\ \\ 1 \\ \\ 0 \end{bmatrix}, \quad \boldsymbol{\nu}_2 = \begin{bmatrix} 1 \\ \\ 1 \\ \\ -1 \\ \\ 0 \end{bmatrix}, \quad \boldsymbol{\nu}_3 = \begin{bmatrix} 1 \\ \\ 0 \\ \\ -1 \\ \\ 1 \end{bmatrix}. \tag{4.9}$$

Since the first chemical formula involves two reactant molecules, there is significantly less progress that can be made without computational methods.

See `MonoMolecularChain.m` and `MichaelisMenten.m` for example code to generate useful data structures for these biochemical reaction networks. These two biochemical reaction networks have been selected to demonstrate two stereotypical problems. In the first instance, the mono-molecular chain model, the network structure enables progress to be made analytically. The enzyme kinetic model, however, represents a more realistic case in which computational methods are required. The focus on these two representative models is done so that the exposition is clear.

4.3 The forwards problem

Given a biochemical reaction network with known kinetic rate parameters and some initial state vector, $\mathbf{X}(0) = \mathbf{x}_0$, we consider the forwards problem. That is, we wish to predict the future evolution. Since we are dealing with stochastic models, all our methods for dealing with future predictions will involve probabilities, random numbers and uncertainty. We rely on standard code libraries[2] for generating samples from uniform (i.e., all outcomes equally likely), Gaussian (i.e., the bell curve), exponential (i.e., time between events), and Poisson distributions (i.e., number of events over a time interval) and do not discuss algorithms for generating samples from these distributions. This enables the focus of this chapter to be on algorithms specific to biochemical reaction networks.

There are two key aspects to the forwards problem: (i) the simulation of a biochemical reaction network that replicates the random reaction events over time; and (ii) the calculation of average behaviour and variability among all possible sequences of reaction events. Relevant algorithms for dealing with both these aspects are reviewed and demonstrated.

4.3.1 Generation of sample paths

Here, we consider algorithms that deal with the simulation of biochemical reaction network evolution. These algorithms are probabilistic, that is, the output of no two simulations, called

[2]We utilise the Statistics and Machine Learning Toolbox within the MATLAB® environment for generating random samples from any of the standard probability distributions.

sample paths, of the same biochemical reaction network will be identical. The most fundamental stochastic simulation algorithms (SSAs) for sample path generation are based on the work of Gillespie (1977, 2000, 2001), Gibson and Bruck (2000), and Anderson (2007).

4.3.1.1 Exact stochastic simulation algorithms

Exact SSAs generate sample paths, over some interval $t \in [0, T]$, that identically follow the probability laws of the fundamental theory of stochastic chemical kinetics (Gillespie, 1992). Take a sufficiently small time interval, $[t, t + \mathrm{d}t)$, such that the probability of multiple reactions occurring in this interval is zero. In such a case, the reactions are mutually exclusive events. Hence, based on Equations (4.2) and (4.3), the probability of any reaction event occurring in $[t, t + \mathrm{d}t)$ is the sum of the individual event probabilities,

$$\begin{aligned}\mathbb{P}(\text{Any reaction occurs in } [t, t + \mathrm{d}t)) &= \mathbb{P}(\text{Reaction 1 occurs in } [t, t + \mathrm{d}t)) + \cdots \\ &\quad + \mathbb{P}(\text{Reaction } M \text{ occurs in } [t, t + \mathrm{d}t)) \\ &= a_0(\mathbf{X}(t))\mathrm{d}t,\end{aligned}$$

where $a_0(\mathbf{X}(t)) = a_1(\mathbf{X}(t)) + \cdots + a_M(\mathbf{X}(t))$ is the total reaction propensity function. Therefore, if we know that the next reaction occurs at time $s \in [t, t + \mathrm{d}t)$, then we can: (i) randomly select a reaction with probabilities, $a_1(\mathbf{X}(s))/a_0(\mathbf{X}(s)), \ldots, a_M(\mathbf{X}(s))/a_0(\mathbf{X}(s))$; and (ii) update the state vector according to the respective stoichiometric vector, $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_M$.

All that remains for an exact simulation method is to determine the time of the next reaction event. Gillespie (1977) demonstrates that the time interval between reactions, Δt, may be treated as a random variable that is exponentially distributed with rate $a_0(\mathbf{X}(t))$, that is, $\Delta t \sim \text{Exp}(a_0(\mathbf{X}(t)))$. Therefore, we arrive at the most fundamental exact SSA, the *Gillespie direct method*:

1. initialise, time $t = 0$ and state vector $\mathbf{X} = \mathbf{x}_0$;
2. calculate propensities, $a_1(\mathbf{X}), \ldots, a_M(\mathbf{X})$, and $a_0(\mathbf{X}) = a_1(\mathbf{X}) + \cdots + a_M(\mathbf{X})$;
3. generate random sample of the time to next reaction, $\Delta t \sim \text{Exp}(a_0(\mathbf{X}))$;
4. if $t + \Delta t > T$, then terminate the simulation, otherwise, go to step 5;

5. randomly select integer j from the set $\{1, \ldots, M\}$ with
   $\mathbb{P}(j = 1) = a_1(\mathbf{X})/a_0(\mathbf{X}), \ldots, \mathbb{P}(j = M) = a_M(\mathbf{X})/a_0(\mathbf{X})$;

6. update state, $\mathbf{X} = \mathbf{X} + \boldsymbol{\nu}_j$, and time $t = t + \Delta t$, then go to step 2.

An example implementation, `GillespieDirectMethod.m`, and example usage, `DemoGillespie.m` are provided. Figure 4.1 demonstrates sample paths generated by Gillespie direct method for the mono-molecular chain model (Figure 4.1(A)) and the enzyme kinetic model (Figure 4.1(B)).

A mathematically equivalent, but more computationally efficient, exact SSA formulation is derived by Gibson and Bruck (2000). Their method independently tracks the next reaction time of each reaction separately. The next reaction to occur is the one with the smallest next reaction time, therefore no random selection of reaction events is required. It should, however, be noted that the Gillespie direct method may also be improved to yield the *optimised direct method* (Cao et al., 2004) with similar performance benefits. Anderson (2007) further refines the method of Gibson and Bruck by scaling the times of each reaction so that the scaled times between reactions follow unit-rate exponential random variables. This scaling allows the method to be applied to more complex biochemical reaction networks with time-dependent propensity functions, however, the recently proposed Extrande method (Voliotis et al., 2016) is computationally superior. In Anderson's approach, scaled times are tracked for each reaction independently with t_j being the current time at the natural scale of reaction j. This results in the *modified next reaction method*:

1. initialise, global time $t = 0$, state vector $\mathbf{X} = \mathbf{x}_0$ and scaled times
   $t_1 = t_2 = \cdots = t_M = 0$;

2. generate M first reaction times, $s_1, \ldots, s_M \sim \text{Exp}(1)$;

3. calculate propensities, $a_1(\mathbf{X}), \ldots, a_M(\mathbf{X})$;

4. rescale time to next reaction, $\Delta t_j = (s_j - t_j)/a_j(\mathbf{X})$ for $j = 1, 2, \ldots, M$;

5. choose reaction k, such that $\Delta t_k = \min\{\Delta t_1, \ldots, \Delta t_M\}$;

6. if $t + \Delta t_k > T$ terminate simulation, otherwise go to step 7;

7. update rescaled times, $t_j = t_j + a_j(\mathbf{X})\Delta t_k$ for $j = 1, \ldots, M$, state $\mathbf{X} = \mathbf{X} + \boldsymbol{\nu}_k$ and global time $t = t + \Delta t_k$;

8. generate scaled next reaction time for reaction k, $\Delta s_k \sim \text{Exp}(1)$;

9. update next scaled reaction time $s_k = s_k + \Delta s_k$, and go to step 3.

An example implementation, `ModifiedNextReactionMethod.m`, and example usage, `DemoMNRM.m` are provided. Figure 4.1 demonstrates sample paths generated by the modified next reaction method for the mono-molecular chain model (Figure 4.1(C)) and the enzyme kinetic model (Figure 4.1(D)). Note that the sample paths are different to those generated using the Gillespie direct method, despite the random number generators being initialised the same way. However, both represent exact sample paths, that is, sample paths that exactly follow the dynamics of the biochemical reaction network.

While the Gillespie direct method and the more efficient modified next reaction method and optimised direct method represent the most fundamental examples of exact SSAs, other advanced methods are also available to further improve the computational performance for large and complex biochemical reaction networks. Particular techniques include partial-propensity factorisation (Indurkhya and Beal, 2010), rejection-based methods (Thanh et al., 2014; Thanh, 2017) and composition-rejection (Slepoy et al., 2008) methods. We do not discuss these approaches, but we highlight them to indicate that efficient exact SSA method development is still an active area of research.

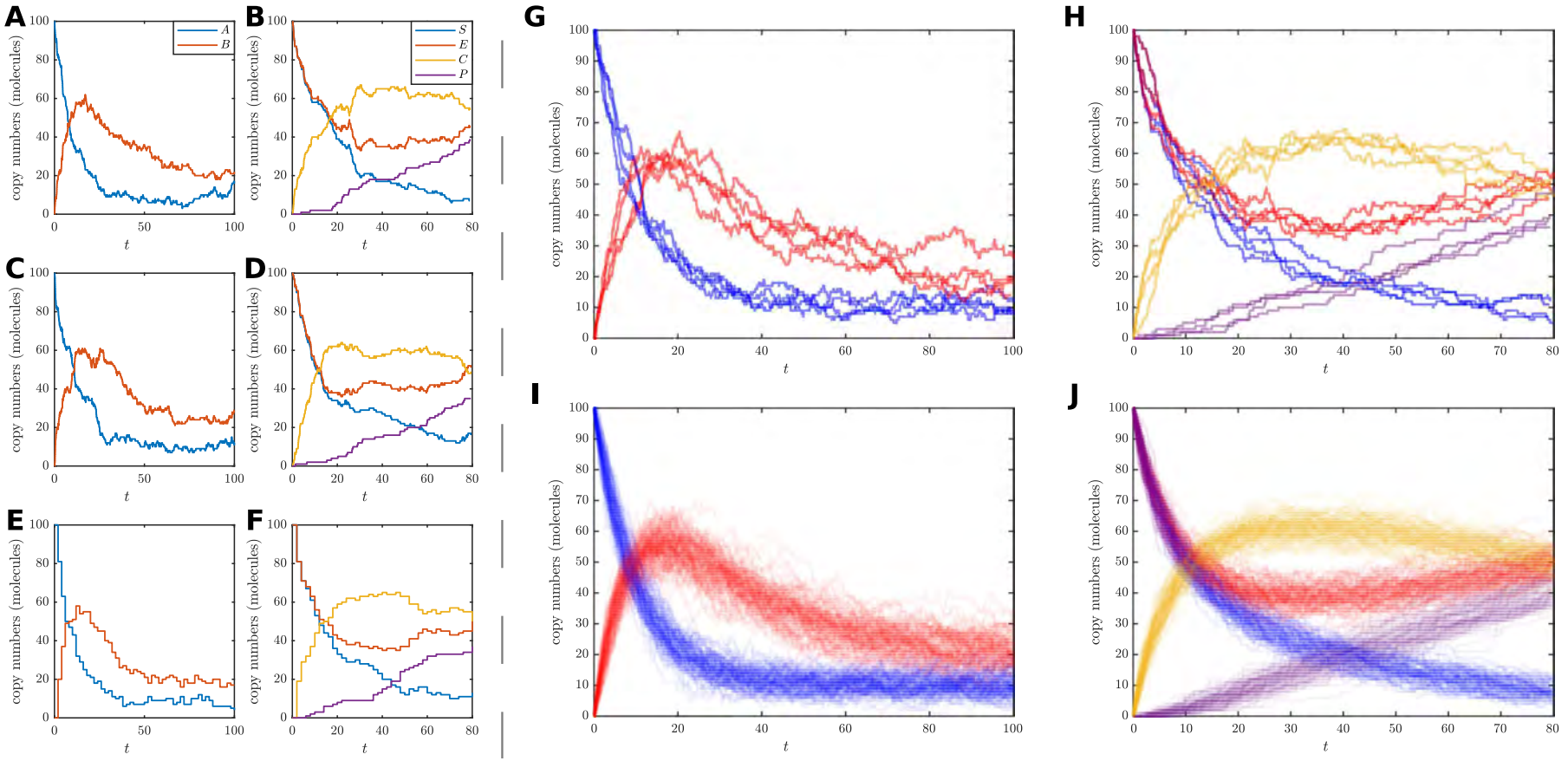

Figure 4.1: Examples of exact sample paths of the mono-molecular chain model using the (A) Gillespie direct method and (C) modified next reaction method; similarly exact sample paths of the enzyme kinetics model using the (B) Gillespie direct method and (D) modified next reaction method. Approximate sample paths may be computed with less computational burden using the tau leaping method with $\tau = 2$, at the expense of accuracy: (E) the mono-molecular chain model and (F) the enzyme kinetics model. Every sample path will be different; as demonstrated by four distinct simulations of (G) the mono-molecular chain model and (H) the enzyme kinetics model. However, trends are revealed when 100 simulations are overlaid to reveal states of higher probability density using (I) the mono-molecular chain model and (J) the enzyme kinetics model. The mono-molecular chain model simulations are configured with parameters $k_1 = 1.0$, $k_2 = 0.1$, $k_3 = 0.05$, and initial state $A(0) = 100$, $B(0) = 0$. The enzyme kinetics model simulations are configured with parameters $k_1 = 0.001$, $k_2 = 0.005$, $k_3 = 0.01$, and initial state $E(0) = 100$, $S(0) = 100$, $C(0) = 0$, $P(0) = 0$.

4.3.1.2 Approximate stochastic simulation algorithms

Despite some computational improvements provided by the modified next reaction method (Anderson, 2007; Gibson and Bruck, 2000), all exact SSAs are computationally intractable for large biochemical populations and with many reactions, since every reaction event is simulated. Several approximate SSAs have been introduced in an attempt to reduce the computational burden while sacrificing accuracy.

The main approximate SSA we consider is also developed by Gillespie (2001) almost 25 years after the development of the Gillespie direct method. The key idea is to evolve the system in discrete time steps of length τ, hold the propensity functions constant over the time interval $[t, t+\tau)$ and count the number of reaction events that occur. The state vector is then updated based on the net effect of all the reaction events. The number of reaction events within the interval can be shown to be a random variable distributed according to a Poisson distribution with mean $a_j(\mathbf{X}(t))\tau$. If Y_j denotes the number of reaction j events in $[t, t+\tau)$ then $Y_j \sim \text{Po}(a_j(\mathbf{X}(t))\tau)$. The result is the *tau leaping method*:

1. initialise, time $t = 0$ and state $\mathbf{Z} = \mathbf{x}_0$;
2. if $t + \tau > T$ then terminate simulation, otherwise continue;
3. calculate propensities, $a_1(\mathbf{Z}), \ldots, a_M(\mathbf{Z})$;
4. generate reaction event counts, $Y_j \sim \text{Po}(a_j(\mathbf{Z})\tau)$ for $j = 1, \ldots, M$;
5. update state, $\mathbf{Z} = \mathbf{Z} + Y_1\boldsymbol{\nu}_1 + \cdots Y_M\boldsymbol{\nu}_M$, and time $t = t + \tau$;
6. go to step 2.

Note that we use the notation $\mathbf{Z}(t)$ to denote an approximation of the true state $\mathbf{X}(t)$. An example implementation, `TauLeapingMethod.m`, and example usage, `DemoTauLeap.m` are provided. Figure 4.1 demonstrates sample paths generated by the tau leaping method for the mono-molecular chain model (Figure 4.1(E)) and the enzyme kinetic model (Figure 4.1(F)). Note that there is a visually obvious difference in the noise patterns of the tau leaping method sample paths and the exact SSA sample paths (Figure 4.1(A)–(D)).

The tau leaping method is the only approximate SSA that we will explicitly discuss as it captures the essence of what approximations try to achieve; trading accuracy for improved performance. Several variations of the tau leaping method have been proposed to improve accuracy, such as adaptive τ selection (Cao et al., 2005a, 2006), implicit schemes (Rathinam et al., 2003) and the replacement of Poisson random variables with binomial random variables (Tian and Burrage, 2004). Hybrid methods that combine exact SSAs and approximations that split reactions into different time-scales are also particularly effective for large scale networks with reactions occurring on very different time-scales (Cao et al., 2005b; E et al., 2005). Other approximate simulation approaches are, for example, based on a continuous state chemical Langevin equation approximation (Cotter and Erban, 2016; Gillespie, 2000; Wilkinson, 2009) and employ numerical schemes for stochastic differential equations (Burrage et al., 2004; Higham, 2001).

4.3.2 Computation of summary statistics

Due to the stochastic nature of biochemical reaction networks, one cannot predict with certainty the future state of the system. Averaging over n sample paths can, however, provide insights into likely future states. Figure 4.1(G),(H) show that there is considerable variation in four independent sample paths, $n = 4$, of the mono-molecular chain and enzyme kinetic models. However, there is still a qualitative similarity between them. This becomes more evident for $n = 100$ sample paths, as in Figure 4.1(I),(J). The natural extension is to consider average behaviour as $n \to \infty$.

4.3.2.1 Using the chemical master equation

From a probability theory perspective, a biochemical reaction network model is a discrete-state, continuous-time Markov process. One key result for discrete-state, continuous-time Markov processes is, given an initial state, $\mathbf{X}(t) = \mathbf{x}_0$, one can describe how the probability distribution

of states evolves. This is given by the chemical master equation (Gillespie, 1992)[3],

$$\frac{\mathrm{d}P(\mathbf{x}, t \mid \mathbf{x}_0)}{\mathrm{d}t} = \underbrace{\sum_{j=1}^{M} a_j(\mathbf{x} - \boldsymbol{\nu}_j) P(\mathbf{x} - \boldsymbol{\nu}_j, t \mid \mathbf{x}_0)}_{\substack{\text{probability increase from events} \\ \text{that cause state change to } \mathbf{x}}} \underbrace{- P(\mathbf{x}, t \mid \mathbf{x}_0) \sum_{j=1}^{M} a_j(\mathbf{x})}_{\substack{\text{probability decrease from events} \\ \text{that cause state change from } \mathbf{x}}}, \tag{4.10}$$

where $P(\mathbf{x}, t \mid \mathbf{x}_0)$ is the probability that $\mathbf{X}(t) = \mathbf{x}$ given $\mathbf{X}(0) = \mathbf{x}_0$. Solving the chemical master equation provides an explicit means of computing the probability of being in any state at any time. Unfortunately, solutions to the chemical master equation are only known for special cases (Gadgil et al., 2005; Jahnke and Huisinga, 2007).

However, the mean and variance of the biochemical reaction network molecule copy numbers can sometimes be derived without solving the chemical master equation. For example, for the mono-molecular chain model (Equation (4.4)), one may use Equation (4.10) to derive the following system of ordinary differential equations (see Section 4.6.1),

$$\frac{\mathrm{d}M_a(t)}{\mathrm{d}t} = k_1 - k_2 M_a(t), \tag{4.11}$$

$$\frac{\mathrm{d}M_b(t)}{\mathrm{d}t} = k_2 M_a(t) - k_3 M_b(t), \tag{4.12}$$

$$\frac{\mathrm{d}V_a(t)}{\mathrm{d}t} = k_1 + k_2 M_a(t) - 2k_2 V_a(t), \tag{4.13}$$

$$\frac{\mathrm{d}V_b(t)}{\mathrm{d}t} = k_2 M_a(t) + k_3 M_b(t) + 2k_2 C_{a,b}(t) - k_3 V_b(t), \tag{4.14}$$

$$\frac{\mathrm{d}C_{a,b}(t)}{\mathrm{d}t} = k_2 V_a(t) - k_2 M_a(t) - (k_2 + k_3) C_{a,b}(t), \tag{4.15}$$

where $M_a(t)$ and $V_a(t)$ ($M_b(t)$ and $V_b(t)$) are the mean and variance of the copy number $A(t)$ ($B(t)$) over all possible sample paths. $C_{a,b}(t)$ is the covariance of between $A(t)$ and $B(t)$. Equations (4.11)–(4.15) are linear ordinary differential equations than can be solved analytically and the solution is plotted in Figure 4.2(A) superimposed with a population of sample paths.

[3] In the theory of Markov processes, this equation is known as the *Kolmogorov forward equation*.

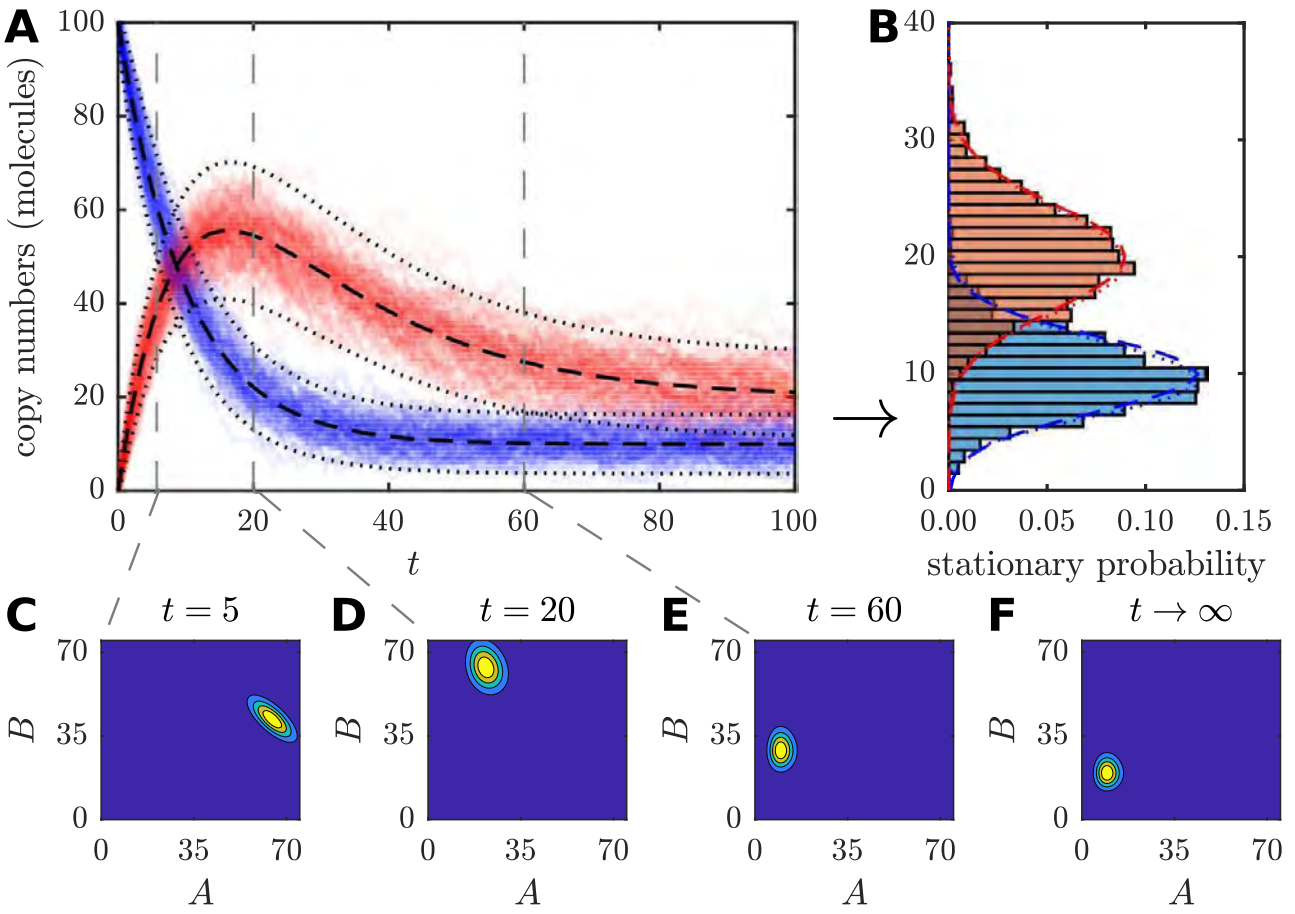

Figure 4.2: (A) The chemical master equation mean (black dashed) ± two standard deviations (black dots) of copy numbers of A (blue) and B (red) chemical species are displayed over simulated sample paths to demonstrate agreement. (B) The stationary distributions of A and B computed using: long running, $T = 1000$, simulated sample paths (blue/red histograms); Gaussian approximation (blue/red dashed) using long time limits of chemical master equation mean and variances; and the long time limit of the full chemical master equation solution (blue/red dots). Transient chemical master equation solution at times (C) $t = 5$, (D) $t = 20$, and (E) $t = 60$. (F) chemical master equation solution stationary distribution. The parameters are $k_1 = 1.0$, $k_2 = 0.1$, $k_3 = 0.05$, and initial state is $A(0) = 100$, $B(0) = 0$.

The long time limit behaviour of a biochemical reaction network can also be determined. For the mono-molecular chain, as $t \to \infty$ we have: $M_a(t) \to k_1/k_2$; $V_a(t) \to k_1/k_2$; $M_b(t) \to k_1/k_3$; $V_b(t) \to k_1/k_3$; and $C_{a,b}(t) \to 0$. We can approximate the long time steady state behaviour, called the *stationary distribution*, of the mono-molecular chain model as two independent Gaussian random variables. That is, as $t \to \infty$, $A(t) \to A_\infty$ with $A_\infty \sim \mathcal{N}(k_1/k_2, k_1/k_2)$. Similarly, $B(t) \to B_\infty$ with $B_\infty \sim \mathcal{N}(k_1/k_3, k_1/k_3)$. This approximation is shown in Figure 4.2(B) against histograms of sample paths generated using large values of T (see example code, `DemoStatDist.m`).

In the case of the mono-molecular chain model, the chemical master equation is analytically tractable (Gadgil et al., 2005; Jahnke and Huisinga, 2007). However, the solution is algebraically complicated and non-trivial to evaluate (see Section 4.6.2). The full chemical master equation solution in Figure 4.2(C)–(E) and the true stationary distribution of two independent Poisson distributions is shown in Figure 4.2(F). The true stationary distribution is also compared with the Gaussian approximation in Figure 4.2(B); the approximation is reasonably accurate, however, we could have also reasonably surmised the true stationary distribution by noting that, for a Poisson distribution, the mean is equal to the variance.

4.3.2.2 Monte Carlo methods

The chemical master equation can yield insight for special cases, however, for practical problems, such as the enzyme kinetic model, the chemical master equation is intractable and numerical methods are required. Here, we consider numerical estimation of the mean state vector at a fixed time, T.

In probability theory, the mean of a distribution is defined via an expectation,

$$\mathbb{E}\left[\mathbf{X}(T)\right] = \sum_{\mathbf{x}\in\Omega} \mathbf{x} P(\mathbf{x}, T \mid \mathbf{x}_0), \tag{4.16}$$

where Ω is the set of all possible states. It is important to note that the methods we describe here are equally valid for a more general expectations of the form $\mathbb{E}\left[f(\mathbf{X}(T))\right]$ where f is some function that satisfies certain conditions (Giles, 2015).

We usually cannot compute Equation (4.16) directly since Ω is typically infinite and the chemical master equation is intractable. However, exact SSAs provide methods for sampling the chemical master equation distribution, $\mathbf{X}(T) \sim P(\mathbf{x}, T \mid \mathbf{x}_0)$. This leads to the Monte Carlo estimator

$$\mathbb{E}\left[\mathbf{X}(T)\right] \approx \hat{\mathbf{X}}(T) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}(T)^{(i)}, \tag{4.17}$$

where $\mathbf{X}(T)^{(1)}, \ldots, \mathbf{X}(T)^{(n)}$ are n independent sample paths of the biochemical reaction network of interest (see example implementation `MonteCarlo.m`).

Unlike Equation (4.16), the Monte Carlo estimates, such as Equation (4.17), are random variables for finite n. This incurs a probabilistic error. A common measure of the accuracy of a Monte Carlo estimator, $\hat{\mu}(T)$, of some expectation, $\mathbb{E}\left[\mu(T)\right]$, is the *mean-square error* that evaluates the average error behaviour and may be decomposed as follows,

$$\underbrace{\mathbb{E}\left[\left(\hat{\mu}(T) - \mathbb{E}\left[\mu(T)\right]\right)^2\right]}_{\text{mean-square error}} = \underbrace{\mathbb{V}\left[\hat{\mu}(T)\right]}_{\text{Estimator variance}} + \left(\underbrace{\mathbb{E}\left[\hat{\mu}(T)\right] - \mathbb{E}\left[\mu(T)\right]}_{\text{Estimator bias}}\right)^2. \tag{4.18}$$

Equation (4.18) highlights that there are two sources of error in a Monte Carlo estimator, the estimator *variance* and *bias*, and much of the discussion that follows deals with how to balance both these error sources in a computationally efficient manner.

Through analysis of the mean-square error of an estimator, the rate at which the mean-square error decays as n increases can be determined. Hence, we can determine how large n needs to be satisfy the condition

$$\sqrt{\mathbb{E}\left[\left(\hat{\mu}(T) - \mathbb{E}\left[\mu(T)\right]\right)^2\right]} \leq ch, \tag{4.19}$$

where c is a positive constant and h is called the *error tolerance*.

Since $\mathbb{E}\left[\hat{\mathbf{X}}(T)\right] = \mathbb{E}\left[\mathbf{X}(T)\right]$, the bias term in Equation (4.18) is zero and we call $\hat{\mathbf{X}}(T)$ an *unbiased* estimator of $\mathbb{E}\left[\mathbf{X}(T)\right]$. For an unbiased estimator, the mean-square error is equal to the estimator variance. Furthermore, $\mathbb{V}\left[\hat{\mathbf{X}}(T)\right] = \mathbb{V}\left[\mathbf{X}(T)\right]/n$, so the estimator variance decreases linearly with n, for sufficiently large n. Therefore, $h \propto 1/\sqrt{n}$. That is, to halve h, one must increase n by a factor of four. This may be prohibitive with exact SSAs, especially for biochemical reaction networks with large variance. In the context of the Monte Carlo estimator using an exact SSA (Equation (4.17)), for n sufficiently large, the central limit theorem (CLT)

states that $\hat{\mathbf{X}}(T) \sim \mathcal{N}\left(\mathbb{E}\left[\mathbf{X}(T)\right], \mathbb{V}\left[\mathbf{X}(T)\right]/n\right)$ (see Wilkinson (2012) for a good discussion on the CLT).

Computational improvements can be achieved by using an approximate SSA, such as the tau leaping method,

$$\mathbb{E}\left[\mathbf{X}(T)\right] \approx \mathbb{E}\left[\mathbf{Z}(T)\right] \approx \hat{\mathbf{Z}}(T) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{Z}(T)^{(i)}, \tag{4.20}$$

where $\mathbf{Z}(T)^{(1)}, \ldots, \mathbf{Z}(T)^{(n)}$ are n independent approximate sample paths of the biochemical reaction network of interest (see the example implementation `MonteCarloTauLeap.m`).

Note that, $\mathbb{E}\left[\hat{\mathbf{Z}}(T)\right] = \mathbb{E}\left[\mathbf{Z}(T)\right]$. Since, $\mathbb{E}\left[\mathbf{Z}(T)\right] \neq \mathbb{E}\left[\mathbf{X}(T)\right]$ in general, we call $\hat{\mathbf{Z}(\mathbf{T})}$ a *biased* estimator. Even in the limit of $n \to \infty$, the bias term in Equation (4.18) may not be zero, which incurs a lower bound on the best achievable error tolerance, h, for fixed τ. However, it has been shown that the bias of the tau leaping method decays linearly with τ (Anderson et al., 2011; Li, 2007). Therefore, to satisfy the error tolerance condition (Equation (4.19)) we not only require $n \propto 1/h^2$, but also $\tau \propto h$. That is, as h decrease the performance improvement of Monte Carlo with the tau leaping method reduces by a factor of τ, because the computational cost of the tau leaping method is proportional to $1/\tau$. In Figure 4.3(A), the decay of the computational advantage in using the tau leaping method for Monte Carlo over the Gillespie direct method is evident, and eventually the tau leaping method will be more computationally burdensome than the Gillespie direct method. By the CLT, for large n, we have $\hat{\mathbf{Z}}(T) \sim \mathcal{N}(\mathbb{E}\left[\mathbf{Z}(T)\right], \mathbb{V}\left[\mathbf{Z}(T)\right]/n)$.

The utility of the tau leaping method for accurate (or exact) Monte Carlo estimation is identified by Anderson and Higham (2012) through extending the idea of multilevel Monte Carlo (MLMC) originally proposed by Giles for stochastic differential equations (Giles, 2008, 2015). Consider a sequence of $L+1$ tau leaping method time steps $\tau_0, \tau_1, \ldots, \tau_L$, with $\tau_\ell < \tau_{\ell-1}$ for $\ell = 1, \ldots, L$. Let $\mathbf{Z}_\ell(T)$ denote the state vector of a tau leaping method approximation using time step τ_ℓ. Assume τ_L is small enough that $\mathbb{E}\left[\mathbf{Z}_L(T)\right]$ is a good approximation of $\mathbb{E}\left[\mathbf{X}(T)\right]$. Note, for large τ_ℓ (small ℓ), sample paths are cheap to generate, but inaccurate; conversely, small τ_ℓ (large ℓ) results in computationally expensive, but accurate sample paths.

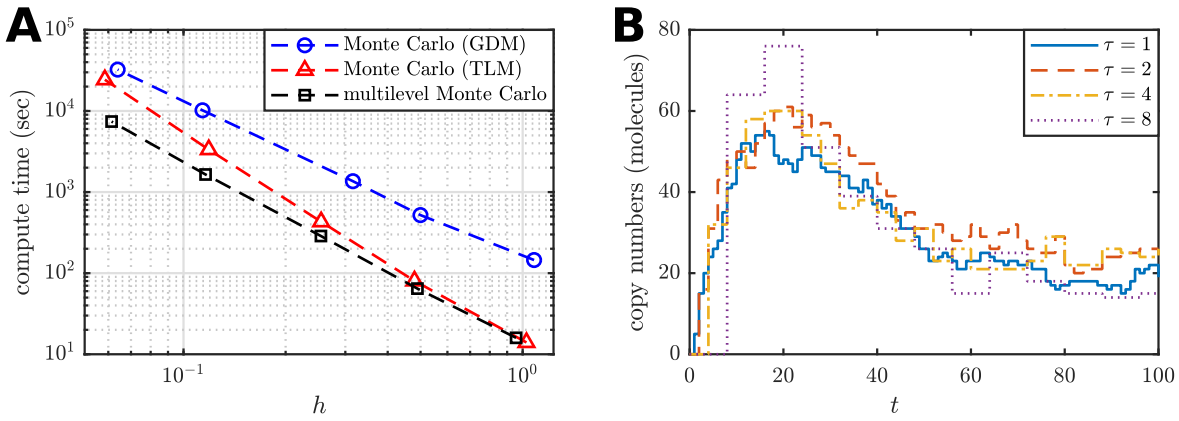

Figure 4.3: (A) Improved performance from MLMC when estimating $\mathbb{E}[B(T)]$ at $T = 100$ using the mono-molecular chain model with parameters $k_1 = 10$, $k_2 = 0.1$, $k_3 = 0.5$, and initial condition $A(0) = 1000$, $B(0) = 0$; the computational advantage of the tau leaping method (red triangles dashed) over the Gillespie direct method (blue circles dashed) for Monte Carlo diminishes as the required error tolerance decreases. The MLMC method (black squares dashed) exploits the correlated tau leaping method to obtain sustained computational efficiency. (B) Demonstration of correlated tau leaping method simulations for nested τ steps; sample paths using a step of $\tau = 2$ (red dashed), $\tau = 4$ (yellow dot-dashed), and $\tau = 8$ (purple dots) are all correlated with the same $\tau = 1$ trajectory (blue solid). Computations are performed using an Intel® Core™ i7-5600U CPU (2.6 GHz).

We can write

$$
\begin{aligned}
\mathbb{E}[\mathbf{X}(T)] &\approx \underbrace{\mathbb{E}[\mathbf{Z}_L(T)]}_{\substack{\text{low bias}\\ \text{approximation}}} \\
&= \underbrace{\mathbb{E}[\mathbf{Z}_{L-1}(T)]}_{\substack{\text{slightly biased}\\ \text{approximation}}} + \underbrace{\mathbb{E}[\mathbf{Z}_L(T) - \mathbf{Z}_{L-1}(T)]}_{\text{bias correction}} \\
&= \underbrace{\mathbb{E}[\mathbf{Z}_{L-2}(T)]}_{\substack{\text{slightly more biased}\\ \text{approximation}}} + \underbrace{\mathbb{E}[\mathbf{Z}_{L-1}(T) - \mathbf{Z}_{L-2}(T)] + \mathbb{E}[\mathbf{Z}_L(T) - \mathbf{Z}_{L-1}(T)]}_{\text{two bias corrections}} \\
&\vdots \\
&= \underbrace{\mathbb{E}[\mathbf{Z}_0(T)]}_{\substack{\text{very biased}\\ \text{approximation}}} + \underbrace{\sum_{\ell=1}^{L} \mathbb{E}[\mathbf{Z}_\ell(T) - \mathbf{Z}_{\ell-1}(T)]}_{L \text{ bias corrections}}.
\end{aligned}
\tag{4.21}
$$

Importantly, the final estimator on the right of Equation (4.21), called the multilevel telescoping summation, is equivalent in bias to $\mathbb{E}[\mathbf{Z}_L(T)]$. At first glance, Equation (4.21) looks to have complicated the computational problem and inevitably decreased performance of the Monte Carlo estimator. The insight of Giles (2008), in the context of stochastic differential equation

models for financial applications, is that the bias correction terms may be computed using Monte Carlo approaches that involve generating highly correlated sample paths in estimation of each of the correction terms, thus reducing the variance in the bias corrections. If the correlation is strong enough, then the variance decays such that few of the most accurate tau leaping method sample paths are required; this can result in significant computational savings.

A contribution of Anderson and Higham (2012) is an efficient method of generating correlated tau leaping method sample path pairs $(\mathbf{Z}_\ell(T), \mathbf{Z}_{\ell-1}(T))$ in the case when $\tau_\ell = \tau_{\ell-1}/\delta$ for some positive integer scale factor δ. The algorithm is based on the property that the sum of two Poisson random variables is also a Poisson random variable. This enables the sample path with $\tau_{\ell-1}$ to be constructed as an approximation to the sample path with τ_ℓ directly. Figure 4.3(B) demonstrates tau leaping method sample paths of $B(t)$ in the mono-molecular chain model with $\tau = 2, 4, 8$ generated directly from a tau leaping method sample path with $\tau = 1$. The algorithm can be thought of as generating multiple approximations of the same exact biochemical reaction network sample path. The algorithm is the *correlated tau leaping method*:

1. initialise time $t = 0$, and states $\mathbf{Z}_\ell$, $\mathbf{Z}_{\ell-1}$ corresponding to sample paths with $\tau = \tau_\ell$ and $\tau = \tau_{\ell-1}$, respectively;

2. if $t + \tau_\ell > T$, then terminate simulation, otherwise, continue;

3. calculate propensities for path $\mathbf{Z}_\ell$, $a_1(\mathbf{Z}_\ell), \ldots, a_M(\mathbf{Z}_\ell)$;

4. if t/τ_ℓ is not an integer multiple of δ, then go to step 6, otherwise continue;

5. calculate propensities for path $\mathbf{Z}_{\ell-1}$, $a_1(\mathbf{Z}_{\ell-1}), \ldots, a_M(\mathbf{Z}_{\ell-1})$, initialise intermediate state $\bar{\mathbf{Z}} = \mathbf{Z}_{\ell-1}$;

6. for each reaction $j = 1, \ldots, M$;

   (a) calculate virtual propensities, $b_{j,1} = \min\{a_j(\mathbf{Z}_\ell), a_j(\mathbf{Z}_{\ell-1})\}$, $b_{j,2} = a_j(\mathbf{Z}_\ell) - b_{j,1}$, and $b_{j,3} = a_j(\mathbf{Z}_{\ell-1}) - b_{j,1}$;

   (b) generate virtual reaction event counts, $Y_{j,1} \sim \text{Po}(b_{j,1}\tau_\ell)$, $Y_{j,2} \sim \text{Po}(b_{j,2}\tau_\ell)$, and $Y_{j,3} \sim \text{Po}(b_{j,3}\tau_\ell)$;

7. set, $\mathbf{Z}_\ell = \mathbf{Z}_\ell + (Y_{1,1} + Y_{1,2})\boldsymbol{\nu}_1 + \cdots + (Y_{M,1} + Y_{M,2})\boldsymbol{\nu}_M$;

8. set, $\bar{\mathbf{Z}} = \bar{\mathbf{Z}} + (Y_{1,1} + Y_{1,3})\boldsymbol{\nu}_1 + \cdots + (Y_{M,1} + Y_{M,3})\boldsymbol{\nu}_M$;

9. if $(t + \tau_\ell)/\tau_\ell$ is an integer multiple of δ, then Set $\mathbf{Z}_{\ell-1} = \bar{\mathbf{Z}}$;

10. update time $t = t + \tau_\ell$, and go to step 2.

See example implementation, `CorTauLeapingMethod.m`, and example usage `DemoCorTauLeap.m`.

Given the correlated tau leaping method, Monte Carlo estimation can be applied to each term in Equation (4.21) to give

$$\hat{\mathbf{Z}}_L(T) = \frac{1}{n_0}\sum_{i=1}^{n_0}\mathbf{Z}_0(T)^{(i)} + \sum_{\ell=1}^{L}\frac{1}{n_\ell}\sum_{i=1}^{n_\ell}\left[\mathbf{Z}_\ell(T)^{(i)} - \mathbf{Z}_{\ell-1}(T)^{(i)}\right], \tag{4.22}$$

where $\mathbf{Z}_0(T)^{(1)}, \ldots, \mathbf{Z}_0(T)^{(n_0)}$ are n_0 independent tau leaping method sample paths with $\tau = \tau_0$, and $(\mathbf{Z}_\ell(T)^{(1)}, \mathbf{Z}_{\ell-1}(T)^{(1)}), \ldots, (\mathbf{Z}_\ell(T)^{(n_\ell)}, \mathbf{Z}_{\ell-1}(T)^{(n_\ell)})$ are n_ℓ paired correlated tau leaping method sample paths with time steps $\tau = \tau_\ell$, $\tau = \tau_{\ell-1}$ and $\tau_{\ell-1} = \delta\tau_\ell$ for each bias correction $\ell = 1, 2, \ldots, L$. Given an error tolerance, h, it is possible to calculate an optimal sequence of sample path numbers $n_0, n_1, \ldots, n_L$ such that the total computation time is optimised (Anderson and Higham, 2012; Lester et al., 2016; Giles, 2008). The results are shown in Figure 4.3(A) for a more computationally challenging parameter set of the mono-molecular chain model. See the example implementation, `MultilevelMonteCarlo.m` and `DemoMonteCarlo.m`, for the full performance comparison.

As formulated here, MLMC results in a biased estimator, though it is significantly more efficient to reduce the bias of this estimator than by direct use of the tau leaping method. If an unbiased estimator is required, then this can be achieved by correlating exact SSA sample paths with approximate SSA sample paths. Anderson and Higham (2012) demonstrate a method for correlating tau leaping method sample paths and modified next reaction method sample paths, and Lester et al. (2016) demonstrate correlating tau leaping method sample paths and Gillespie direct method sample paths. Further refinements such as adaptive and non-nested τ_ℓ steps are also considered by Lester et al. (2015), a multilevel hybrid scheme is developed by Moraes et al. (2016) and Wilson and Baker (2016) use MLMC and maximum entropy methods to generate approximations to the chemical master equation.

4.3.3 Summary of the forwards problem

Significant progress has be made in the study of computational methods for the solution to the forwards problem. As a result, the forwards problem is relatively well understood, particularly for well-mixed systems, such as the biochemical reaction network models we consider in this chapter.

An exact solution to simulation is achieved though the development of exact SSAs. However, if Monte Carlo methods are required to determine expected behaviours, then exact SSAs can be computationally burdensome. While approximate SSAs provide improvements is some cases, highly accurate estimates will still often become burdensome since very small time steps will be required to keep the bias at the same order as the estimator standard deviation. In this respect, MLMC methods provide impressive computational improvements without any loss in accuracy. Such methods have become popular in financial applications (Giles, 2015; Higham, 2015), however, there have been fewer examples in a biological setting.

Beyond the Gillespie direct method, the efficiency of sample path generation has been dramatically improved through the advancements in both exact SSAs and approximate SSAs. While approximate SSAs like the tau leaping method provide computational advantages, they also introduce approximations. Some have noted that the error in these approximations likely to be significantly lower than the modelling error compared with the real biological systems (Wilkinson, 2009). However, there is no general theory or guidelines as to when approximate SSAs are safe to use for applications.

We have only dealt with stochastic models that are well-mixed, that is, spatially homogeneous. The development of robust theory and algorithms accounting for spatial heterogeneity is still an active area of research (Gillespie et al., 2013; Schnoerr et al., 2017; Smith and Grima, 2016). The model of biochemical reaction networks, based on the chemical master equation, can be extended to include spatial dynamics through the *reaction-diffusion master equation* (Isaacson, 2008). However, care must be taken in its application because the kinetics of the reaction-diffusion master equation depend on the spatial discretisation and it not always guaranteed to converge in the continuum limit (Erban and Chapman, 2009; Fange et al., 2010; Gillespie et al., 2014; Isaacson, 2008; Smith and Grima, 2016). We refer the reader to Gillespie et al. (Gillespie et al., 2013) and Isaacson (Isaacson, 2008) for useful discussions on this topic.

State-of-the-art Monte Carlo schemes, such as MLMC methods, have the potential to significantly accelerate the computation of summary statistics for the exploration of the range of biochemical reaction network behaviours. However, these approaches are also known to perform poorly for some biochemical reaction network models (Anderson and Higham, 2012). An open question is related to the characterisation of biochemical reaction network models that will benefit from a MLMC scheme for summary statistic estimation. Furthermore, to our knowledge there has been no application of the MLMC approach to the spatially heterogeneous case. The potential performance gains make MLMC a promising space for future development.

4.4 The inverse problem

When applying stochastic biochemical reaction network models to real applications, one often wants to perform statistical inference to estimate the model parameters. That is, given experimental data, and a biochemical reaction network model, the inverse problem seeks to infer the kinetic rate parameters and quantify the uncertainty in those estimates. Just as with the forwards problem, an enormous volume of literature has been dedicated to the problem of inference in stochastic biochemical reaction network models. Therefore, we cannot cover all computational methods in detail. Rather we focus on a computational Bayesian perspective. For further reading, the monograph by Wilkinson (2012) contains very accessible discussions on inference techniques in a systems biology setting, also the monographs by Gelman et al. (2014) and Sisson et al. (2018) contain a wealth of information on Bayesian methods more generally.

4.4.1 Experimental techniques

Typically, time course data are derived from time-lapse microscopy images and fluorescent reporters (Finkenstädt et al., 2008; Iofalla et al., 2008; Wilkinson, 2009). Advances in microscopy and fluorescent technologies are enabling intracellular processes to be inspected at unprecedented resolutions (Chen et al., 2014; Bajar et al., 2016; Leung and Chou, 2011; Sahl et al., 2017). Despite these advances, the resulting data never provide complete observations since: (i) the number of chemical species that may be observed concurrently is relatively low (Wilkinson,

2009); (ii) two chemical species might be indistinguishable from each other (Golightly and Wilkinson, 2011); and (iii) the relationships between fluorescence levels and actual chemical species copy numbers may not be direct, in particular, the degradation of a protein may be more rapid than that of the fluorescent reporter (Iofalla et al., 2008; Vittadello et al., 2018). That is, inferential methods must be able to deal with uncertainty in observations.

For the purposes of this chapter, we consider time-course data. Specifically, we suppose the data consists of n_t observations of the biochemical reaction network state vector at discrete points in time, $t_1, t_2, \ldots, t_{n_t}$. That is, $\mathbf{Y}_{\text{obs}} = [\mathbf{Y}(t_1), \mathbf{Y}(t_2), \ldots, \mathbf{Y}(t_{n_t})]$, where $\mathbf{Y}(t)$ represents an observation of the state vector sample path $\mathbf{X}(t)$. To model observational uncertainty, it is common to treat observations as subject to additive noise (Finkenstädt et al., 2008; Golightly and Wilkinson, 2011; Schnoerr et al., 2017; Toni et al., 2009), so that

$$\mathbf{Y}(t) = \mathbf{A}\mathbf{X}(t) + \boldsymbol{\xi}, \tag{4.23}$$

where $\mathbf{A}$ is a $K \times N$ matrix and $\boldsymbol{\xi}$ is a $K \times 1$ vector of independent Gaussian random variables. The observation vectors, $\mathbf{Y}(t)$, are $K \times 1$ vectors, with $K \leq N$, reflecting the fact that only a sub-set of chemical species of $\mathbf{X}(t)$ are generally observed, or possibly only a linear combination of chemical species (Golightly and Wilkinson, 2011). The example code, `GenerateObservations.m`, simulates this observation process (Equation (4.23)) given a fully specified biochemical reaction network model.

4.4.1.1 Example data

The computation examples given in this chapter are based on two synthetically generated data sets, corresponding to the biochemical reaction network models given in Section 4.2. This enables the comparison between inference methods and the accuracy of inference.

The data for inference on the mono-molecular chain model (Equation (4.4)) are taken as perfect observations, that is, $K = N$, $\mathbf{A} = \mathbf{I}$ and $\mathbb{P}(\boldsymbol{\xi} = \mathbf{0}) = 1$. A single sample path is generated for the mono-molecular chain model with true parameters, $\boldsymbol{\theta}_{\text{true}} = [1.0, 0.1, 0.05]$, and initial condition $A(0) = 100$ and $B(0) = 0$ using the Gillespie direct method. Observations are made using Equation (4.23) applied at $n_t = 4$ discrete times, $t_1 = 25$, $t_2 = 50$, $t_3 = 75$ and $t_4 = 100$.

The resulting data are given in Section 4.6.3.

The data for inference on the enzyme kinetic model (Equation (4.7)) assumes incomplete and noisy observations. Specifically, only the product is observed, so $K = 1$, $\mathbf{A} = [0, 0, 0, 1]$. Further, we assume that there is some measurement error, $\xi \sim \mathcal{N}(0, 4)$; that is, the error standard deviation is two product molecules. The data is generated using the Gillespie direct method with true parameters $\boldsymbol{\theta}_{\text{true}} = [0.001, 0.005, 0.01]$ and initial condition $E(0) = 100$, $S(0) = 100$, $C(0) = 0$ and $P(0) = 0$. Equation (4.23) is evaluated at $n_t = 5$ discrete times, $t_1 = 0$, $t_2 = 20$, $t_3 = 40$, $t_4 = 60$ and $t_5 = 80$ (including an observation of the initial state), yielding the data in Section 4.6.3.

4.4.2 Bayesian inference

Bayesian methods have been demonstrated to provide a powerful framework for the design of experiments, model selection and parameter estimation, especially in the life sciences (Browning et al., 2017; Ellison, 2004; Liepe et al., 2013; Maclaren et al., 2015, 2017; Vanlier et al., 2014) (see Chapter 2 and 3). Given observations, $\mathbf{Y}_{\text{obs}}$, and a biochemical reaction network model parameterised by the $M \times 1$ real-valued vector of kinetic parameters, $\boldsymbol{\theta} = [k_1, k_2, \ldots, k_M]^T$, the task is to quantify knowledge of the true parameter values in light of the data and prior knowledge. This is expressed mathematically through Bayes' Theorem,

$$p(\boldsymbol{\theta} \mid \mathbf{Y}_{\text{obs}}) = \frac{p(\mathbf{Y}_{\text{obs}} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\mathbf{Y}_{\text{obs}})}. \tag{4.24}$$

The terms in Equation (4.24) are interpreted as follows: $p(\boldsymbol{\theta} \mid \mathbf{Y}_{\text{obs}})$ is the *posterior* distribution, that is, the probability[4] of parameter combinations, $\boldsymbol{\theta}$, given the data, $\mathbf{Y}_{\text{obs}}$; $p(\boldsymbol{\theta})$ is the *prior* distribution, that is, the probability of parameter combinations before taking the data into account; $p(\mathbf{Y}_{\text{obs}} \mid \boldsymbol{\theta})$ is the *likelihood*, that is, the probability of observing the data given a parameter combination; and $p(\mathbf{Y}_{\text{obs}})$ is the *evidence*, that is, the probability of observing the data over all possible parameter combinations. Assumptions about the parameters and the biochemical reaction network model are encoded through the prior and the likelihood,

[4]Since $\boldsymbol{\theta}$ and $\mathbf{Y}_{\text{obs}}$ are real-valued, we are not really dealing with a probability distribution functions, but rather probability *density* functions. We will not continue to make this distinction. However, the main technicality is that it not longer makes sense to say "the probability of $\boldsymbol{\theta} = \boldsymbol{\theta}_0$" but rather only "the probability that $\boldsymbol{\theta}$ is close to $\boldsymbol{\theta}_0$". The probability density function must be integrated over the region "close to $\boldsymbol{\theta}_0$" to yield this probability.

respectively. The evidence acts as a normalisation constant, and ensures the posterior is a true probability distribution[5].

First, consider the special case $\mathbf{Y}(t) = \mathbf{X}(t)$, that is, the biochemical reaction network state can be perfectly observed at time t. In this case, the likelihood is

$$p(\mathbf{Y}_{\text{obs}} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n_t} P\left(\mathbf{Y}(t_i), t_i - t_{i-1} \mid \mathbf{Y}(t_{i-1})\right), \tag{4.25}$$

where the function P is the solution to the chemical master equation (Equation (4.10)) and $t_0 = 0$ (Browning et al., 2019; Wilkinson, 2011). It should be noted that, due to the stochastic nature of $\mathbf{X}(t)$, this perfect observation case is unlikely to recover the true parameters. Regardless of this issue, since the likelihood depends on the solution to the chemical master equation, the exact Bayesian posterior will not be analytically tractable in any practical case. In fact, even for the mono-molecular chain model there are problems since the evidence term is not analytically tractable. The example code, `DemoDirectBayesCME.m`, provides an attempt at such a computation, though this code is not practical for an accurate evaluation. Just as with the forwards problem, we must defer to sampling methods and Monte Carlo.

4.4.3 Sampling methods

For this chapter, we focus on the task of estimating the posterior mean,

$$\mathbb{E}\left[\boldsymbol{\theta} \mid \mathbf{Y}_{\text{obs}}\right] = \int_{\boldsymbol{\Theta}} \boldsymbol{\theta}\, p(\boldsymbol{\theta} \mid \mathbf{Y}_{\text{obs}})\, \mathrm{d}\boldsymbol{\theta}, \tag{4.26}$$

where $\boldsymbol{\Theta}$ is the space of possible parameter combinations. However, the methods presented here are applicable to other quantities of interest. For example, the posterior covariance matrix,

$$\mathbb{C}\left[\boldsymbol{\theta} \mid \mathbf{Y}_{\text{obs}}\right] = \int_{\boldsymbol{\Theta}} \left(\boldsymbol{\theta} - \mathbb{E}\left[\boldsymbol{\theta} \mid \mathbf{Y}_{\text{obs}}\right]\right)\left(\boldsymbol{\theta} - \mathbb{E}\left[\boldsymbol{\theta} \mid \mathbf{Y}_{\text{obs}}\right]\right)^T p(\boldsymbol{\theta} \mid \mathbf{Y}_{\text{obs}})\, \mathrm{d}\boldsymbol{\theta}, \tag{4.27}$$

is of interest as it provides an indicator of uncertainty associated with the inferred parameters. Marginal distributions are extremely useful for visualisation: the marginal posterior distribution

[5] That is, the probability of an event that encompasses all possible outcomes is one.

of the jth kinetic parameter is

$$p(k_j \mid \mathbf{Y}_{\text{obs}}) = \int_{\boldsymbol{\Theta}_j} p(\boldsymbol{\theta} \mid \mathbf{Y}_{\text{obs}}) \prod_{i \neq j} \mathrm{d}k_i, \tag{4.28}$$

where $\boldsymbol{\Theta}_j \subset \boldsymbol{\Theta}$ is the parameter space excluding the jth dimension.

Just as with Monte Carlo for the forwards problem, we can estimate posterior expectations (shown here for Equation (4.26), but the method may be similarly applied to Equation (4.27) and Equation (4.28)) using Monte Carlo,

$$\mathbb{E}\left[\boldsymbol{\theta} \mid \mathbf{Y}_{\text{obs}}\right] \approx \hat{\boldsymbol{\theta}} = \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{\theta}^{(i)}, \tag{4.29}$$

where $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(m)}$ are independent samples from the posterior distribution. In the remainder of this section, we focus on computational schemes for generating such samples. We assume throughout that it is possible to generate samples from the prior distribution.

It is important to note that the sampling algorithms we present are not directly relevant to statistical estimators that are not based on expectations, such as *maximum likelihood estimators* or the *maximum a posteriori*. However, these samplers can be modified through the use of *data cloning* (Lele et al., 2007; Picchini and Anderson, 2017) to approximate these effectively. Estimator variance and confidence intervals may also be estimated using bootstrap methods (Efron, 1979; Rubin, 1981).

4.4.3.1 Exact Bayesian sampling

Assuming the likelihood can be evaluated, that is, the chemical master equation is tractable, a naïve method of generating m samples from the posterior is the *exact rejection sampler*:

1. initialise index $i = 0$,
2. generate a prior sample $\boldsymbol{\theta}^* \sim p(\boldsymbol{\theta})$;
3. calculate acceptance probability $\alpha = p(\mathbf{Y}_{\text{obs}} \mid \boldsymbol{\theta}^*)$;
4. with probability α, accept $\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^*$ and $i = i + 1$;

5. if $i = m$, terminate sampling, otherwise go to step 2;

Unsurprisingly, this approach is almost never viable as the likelihood probabilities are often extremely small. In the code example, `DemoExactBayesRejection.m`, the acceptance probability is never more than 9×10^{-15}.

The most common solution is to construct a type of stochastic process (a Markov chain) in the parameter space to generate m_n steps, $\boldsymbol{\theta}^{(0)}, \ldots, \boldsymbol{\theta}^{(m_n)}$. An essential property of the Markov chain used is that its stationary distribution is the target posterior distribution. This approach is called Markov chain Monte Carlo (MCMC), and a common method is the *Metropolis-Hastings method* (Metropolis et al., 1953; Hastings, 1970):

1. initialise $n = 0$ and select starting point $\boldsymbol{\theta}^{(0)}$;
2. generate a proposal sample, $\boldsymbol{\theta}^* \sim q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(n)})$;
3. calculate acceptance probability $\alpha = \min\left(1, \dfrac{p(\mathbf{Y}_{\text{obs}} \mid \boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^{(n)} \mid \boldsymbol{\theta}^*)}{p(\mathbf{Y}_{\text{obs}} \mid \boldsymbol{\theta}^{(n)})p(\boldsymbol{\theta}^{(n)})q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(n)})}\right)$;
4. with probability α, set $\boldsymbol{\theta}^{(n+1)} = \boldsymbol{\theta}^*$, and with probability $1 - \alpha$, set $\boldsymbol{\theta}^{(n+1)} = \boldsymbol{\theta}^{(n)}$;
5. update time, $n = n + 1$;
6. if $n > m_n$, terminate simulation, otherwise go to step 2;

The acceptance probability is based on the relative likelihood between two parameter configurations, the current configuration $\boldsymbol{\theta}^{(n)}$ and a proposed new configuration $\boldsymbol{\theta}^*$, as generated by the user-defined proposal kernel $q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(n)})$. It is essential to understand that since the samples, $\boldsymbol{\theta}^{(0)}, \ldots, \boldsymbol{\theta}^{(m_n)}$, are produced by a Markov chain we cannot treat the m_n steps as m_n independent posterior samples. Rather, we need to take m_n to be large enough that the m_n steps are effectively equivalent to m independent samples.

One challenge in MCMC sampling is the selection of a proposal kernel such that the size of m_n required for the Markov chain to reach the stationary distribution, called the *mixing time*, is small (Geyer, 1992). If the variance of the proposal is too low, then the acceptance rate is high, but the mixing is poor since only small steps are ever taken. On the other hand, a proposal variance that is too high will almost always make proposals that are rejected, resulting in many

wasted likelihood evaluations before a successful move event. Selecting good proposal kernels is a non-trivial exercise and we refer the reader to Cotter et al. (2013), Green et al. (2015), and Roberts and Rosenthal (2009) for detailed discussions on the wide variety of MCMC samplers including proposal techniques. Other techniques used to reduce correlations in MCMC samples include discarding the first m_b steps, called *burn-in* iterations, or sub-sampling the chain by using only every m_hth step, also called *thinning*. However, in general, the use of thinning decreases the statistical efficiency of the MCMC estimator (Link and Eaton, 2011; MacEachern and Berliner, 1994).

Alternative approaches for exact Bayesian sampling can also be based on *importance sampling* (Glynn and Iglehart, 1989; Wilkinson, 2012). Consider a random variable that cannot be simulated, $X \sim p(x)$, but suppose that it is possible to simulate another random variable, $Y \sim q(y)$. If $X, Y \in \Omega$ and $p(x) = 0$ whenever $q(y) = 0$, then

$$\mathbb{E}[X] = \int_\Omega x p(x)\,\mathrm{d}x = \int_\Omega x \frac{p(x)}{q(x)} q(x)\,\mathrm{d}x \approx \frac{1}{m}\sum_{i=1}^{m} \frac{p(Y^{(i)})}{q(Y^{(i)})} Y^{(i)}, \tag{4.30}$$

where $Y^{(1)}, \ldots, Y^{(m)}$ are independent samples of $q(Y)$. Using Equation (4.30), one can show that if the distributions of the target, $p(x)$, and the proposal, $q(y)$, are similar, then a collection of samples, $Y^{(1)}, \ldots, Y^{(m)}$, can be used to generate m approximate samples from $p(x)$. This is called *importance resampling*:

1. generate samples $Y^{(1)}, \ldots, Y^{(m)} \sim q(y)$;
2. compute weights $w^{(1)} = p(Y^{(1)})/q(Y^{(1)}), \ldots, w^{(m)} = p(Y^{(m)})/q(Y^{(m)})$; for $i = 1, 2, \ldots, m$;
3. generate samples $\{X^{(1)}, \ldots, X^{(m)}\}$ by drawing from $\{Y^{(1)}, \ldots, Y^{(m)}\}$ with replacement using probabilities $\mathbb{P}(X = Y^{(i)}) = w^{(i)} \big/ \sum_{i=1}^{m} w^{(i)}$.

In Bayesian applications the prior is often very different to the posterior. In such a case, importance resampling may be applied using a sequence of intermediate distributions. This *sequential importance resampling* and is one approach from the family of *sequential Monte Carlo* (SMC) (Del Moral et al., 2006) samplers. However, like the Metropolis-Hastings method, these methods also require explicit calculations of the likelihood function in order to compute the weights, thus all of these approaches are infeasible for practical biochemical reaction networks. Therefore, we only present more practical forms of these methods later.

More recently, it has been shown that the MLMC telescoping summation can accelerate the computation of posterior expectations when using MCMC (Dodwell et al., 2015; Efendiev et al., 2015) or SMC (Beskos et al., 2017). The key challenges in these applications is the development of appropriate coupling strategies. We do not cover these technical details in this chapter.

4.4.3.2 Likelihood-free methods

Since exact Bayesian sampling is rarely tractable, due to the intractability of the chemical master equation, alternative, likelihood-free sampling methods are required. Two main classes of likelihood-free methods exist: (i) so-called pseudo-marginal MCMC; and (ii) approximate Bayesian computation (ABC). The focus of this chapter is ABC methods, though we first briefly discuss the pseudo-marginal approach.

The basis of the pseudo-marginal approach is to use a MCMC sampler (e.g., Metropolis-Hastings method), but replace the explicit likelihood evaluation with a likelihood estimate obtained through Monte Carlo simulation of the forwards problem (Andrieu and Roberts, 2009). For example, a direct unbiased Monte Carlo estimator is

$$\begin{aligned} p(\mathbf{Y}_{\text{obs}} \mid \boldsymbol{\theta}) &= \int_{\Omega^{n_t}} \prod_{i=1}^{n_t} p(\mathbf{Y}(t_i) \mid \mathbf{X}(t_i)) P\left(\mathbf{X}(t_i), t_i - t_{i-1} \mid \mathbf{X}(t_{i-1})\right) \mathrm{d}\mathbf{X}(t_i) \\ &\approx \frac{1}{n} \sum_{j=1}^{n} \prod_{i=1}^{n_t} p(\mathbf{Y}(t_i) \mid \mathbf{X}(t_i)^{(j)}), \end{aligned}$$

where $[\mathbf{X}(t_1)^{(j)}, \ldots, \mathbf{X}(t_{n_t})^{(j)}]^T$ for $j = 1, 2, \ldots, n$ are independent sample paths of the biochemical reaction network of interest observed discretely at times $t_1, t_2, \ldots, t_{n_t}$. The most successful class of pseudo-marginal techniques, particle MCMC (Andrieu et al., 2010), apply a SMC sampler to approximate the likelihood and inform the MCMC proposal mechanism. The particle marginal Metropolis-Hastings method is a popular variant (Andrieu et al., 2010; Golightly and Wilkinson, 2011). However, the recent model based proposals variant also holds promise for biochemical reaction networks specifically (Pooley et al., 2015).

A particularly nice feature of pseudo-marginal methods is that they are unbiased[6], that is, the Markov chain will still converge to the exact target posterior distribution. This property

[6]Provided the method of estimating the likelihood is unbiased, or has bias that is independent of $\boldsymbol{\theta}$.

is sometimes called an "exact approximation" (Golightly and Wilkinson, 2011; Wilkinson, 2012). Unfortunately, the Markov chains in these methods typically converge more slowly than their exact counterparts. However, computational improvements have been obtained through application of MLMC (Jasra et al., 2018).

Another popular likelihood-free approach is ABC (Pritchard et al., 1999; Sisson et al., 2018; Tavaré et al., 1997). ABC methods have enabled very complex models to be investigated, that are otherwise intractable (Sunnåker et al., 2013). Furthermore, ABC methods are very intuitive, leading to wide adoption within the scientific community (Sunnåker et al., 2013). Applications are particularly prevalent in the life sciences, especially in evolutionary biology (Beaumont et al., 2002; Pritchard et al., 1999; Ratmann et al., 2012; Tavaré et al., 1997; Wilkinson et al., 2011), cell biology (Johnston et al., 2014; Ross et al., 2017; Vo et al., 2015a), epidemiology (Tanaka et al., 2006), ecology (Csilléry et al., 2010; Stumpf, 2014), and systems biology (Toni et al., 2009; Wilkinson, 2011).

The basis of ABC is a discrepancy metric $\rho(\mathbf{Y}_{\text{obs}}, \mathbf{S}_{\text{obs}})$ that provides a measure of closeness between the data, $\mathbf{Y}_{\text{obs}}$, and simulated data $\mathbf{S}_{\text{obs}}$ generated through stochastic simulation of the biochemical reaction network with simulated measurement error. Thus, acceptance probabilities are replaced with a deterministic acceptance condition, $\rho(\mathbf{Y}_{\text{obs}}, \mathbf{S}_{\text{obs}}) \leq \epsilon$, where ϵ is the discrepancy threshold. This yields an approximation to Bayes' Theorem,

$$\begin{aligned} p(\boldsymbol{\theta} \mid \mathbf{Y}_{\text{obs}}) &\approx p(\boldsymbol{\theta} \mid \rho(\mathbf{Y}_{\text{obs}}, \mathbf{S}_{\text{obs}}) \leq \epsilon) \\ &= \frac{p(\rho(\mathbf{Y}_{\text{obs}}, \mathbf{S}_{\text{obs}}) \leq \epsilon \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\rho(\mathbf{Y}_{\text{obs}}, \mathbf{S}_{\text{obs}}) \leq \epsilon)}. \end{aligned} \tag{4.31}$$

The key insight here is that the ability to produce sample paths of a biochemical reaction network model, that is, the forwards problem, enables an approximate algorithm for inference, that is, the inverse problem, regardless of the intractability or complexity of the likelihood. In fact, a formula for the likelihood need not even be known.

The discrepancy threshold determines the level of approximation, however, under the assumption of model and observation error, Equation (4.31) can be treated as exact (Wilkinson, 2013). As $\epsilon \to 0$ then $p(\boldsymbol{\theta} \mid \rho(\mathbf{Y}_{\text{obs}}, \mathbf{S}_{\text{obs}}) \leq \epsilon) \to p(\boldsymbol{\theta} \mid \mathbf{Y}_{\text{obs}})$ (Barber et al., 2015; Fearnhead and Prangle, 2012). Using data for the mono-molecular chain model, we can demonstrate this convergence, as shown in Figure 4.4. The marginal posterior distributions are shown for

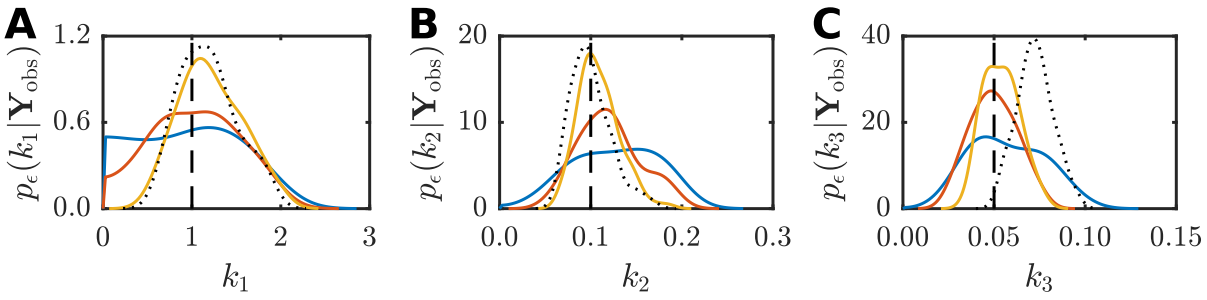

Figure 4.4: Convergence of ABC posterior to the true posterior as $\epsilon \to 0$ for the monomolecular chain inference problem. Marginal posteriors are plotted for $\epsilon = 50$ (blue solid), $\epsilon = 25$ (red solid), $\epsilon = 12.5$ (yellow solid), and $\epsilon = 0$ (black dotted). Here, the $\epsilon = 0$ case corresponds to the exact likelihood using the chemical master equation solution. (A) marginal posteriors for k_1, (B) marginal posteriors for k_2, and (C) marginal posteriors for k_3. The true parameter values (black dashed) are $k_1 = 1.0$, $k_2 = 0.1$ and $k_3 = 0.05$. Note that the exact Bayesian posterior does not recover the true parameter for k_3. The priors used are $k_1 \sim \mathcal{U}(0, 3)$,$k_2 \sim \mathcal{U}(0, 0.3)$, and $k_3 \sim \mathcal{U}(0, 0.15)$.

each parameter, for various values of ϵ and compared with the exact marginal posteriors. The discrepancy metric used is

$$\rho(\mathbf{Y}_{\text{obs}}, \mathbf{S}_{\text{obs}}) = \left[\sum_{i=1}^{n_t} (\mathbf{Y}(t_i) - \mathbf{S}(t_i))^2\right]^{1/2}, \tag{4.32}$$

where $\mathbf{S}(t)$ is simulated data generated using Gillespie direct method and Equation (4.23). The example code, `DemoABCConvergence.m`, is used to generate these marginal distributions.

For any $\epsilon > 0$, ABC methods are biased, just as the tau leaping method is biased for the forwards problem. Therefore, a Monte Carlo estimate of a posterior summary statistic, such as Equation (4.29), needs to take this into bias account. Just like the tau leaping method, the rate of convergence in *mean-square* of ABC based Monte Carlo is degraded because of this bias (Barber et al., 2015; Fearnhead and Prangle, 2012). Furthermore, as the dimensionality of the data increases, small values of ϵ are not computationally feasible. In such cases, the data dimensionality may be reduced by computing lower dimensional summary statistics (Sunnåker et al., 2013), however, these must be *sufficient statistics* in order to ensure the ABC posterior still converges to the correct posterior as $\epsilon \to 0$. We refer the reader to Fearnhead and Prangle (2012) for more detail on this topic.

4.4.3.3 Samplers for approximate Bayesian computation

We now focus on computational methods for generating m samples, $\boldsymbol{\theta}_\epsilon^{(1)}, \ldots, \boldsymbol{\theta}_\epsilon^{(m)}$, from the ABC posterior (Equation (4.31)) with $\epsilon > 0$ and discrepancy metric as given in Equation (4.32). Throughout, we denote $s(\mathbf{S}_{\text{obs}}; \boldsymbol{\theta})$, as the process for generating simulated data given a parameter vector, this process is identical to the processes used to generate our synthetic example data.

In general, the ABC samplers are only computationally viable if the data simulation process is not computationally expensive, in the sense that it is feasible to generate millions of sample paths. However, this is not always realistic, and many extensions exist to standard ABC in an attempt to deal with these more challenging cases. The *Lazy ABC* method (Prangle, 2016) applies an additional probability rule that terminates simulations early if it is likely that $\rho(\mathbf{Y}_{\text{obs}}, \mathbf{S}_{\text{obs}}) > \epsilon$. The *approximate ABC* method (Buzbas and Rosenberg, 2015; Lambert et al., 2018) utilises a small set of data simulations to construct an approximation to the data simulation process.

The most notable early applications of ABC samplers are Beaumont et al. (2002), Pritchard et al. (1999), and Tavaré et al. (1997). The essential development of this early work is that of the *ABC rejection sampler*:

1. initialise index $i = 0$;
2. generate a prior sample $\boldsymbol{\theta}^* \sim p(\boldsymbol{\theta})$;
3. generate simulated data, $\mathbf{S}_{\text{obs}}^* \sim s(\mathbf{S}_{\text{obs}}; \boldsymbol{\theta}^*)$;
4. if $\rho(\mathbf{Y}_{\text{obs}}, \mathbf{S}_{\text{obs}}^*) \leq \epsilon$, accept $\boldsymbol{\theta}_\epsilon^{(i+1)} = \boldsymbol{\theta}^*$ and set $i = i + 1$, otherwise, continue;
5. if $i = m$, terminate sampling, otherwise go to step 2;

There is a clear connection with exact rejection sampler. Note that every accepted sample of the ABC posterior corresponds to at least one simulation of the forwards problem as shown in Figure 4.5. As a result, the computational burden of the inverse problem is significantly higher than the forwards problem, especially for small ϵ. The example code, `ABCRejectionSampler.m`, provides an implementation of the ABC rejection sampler.

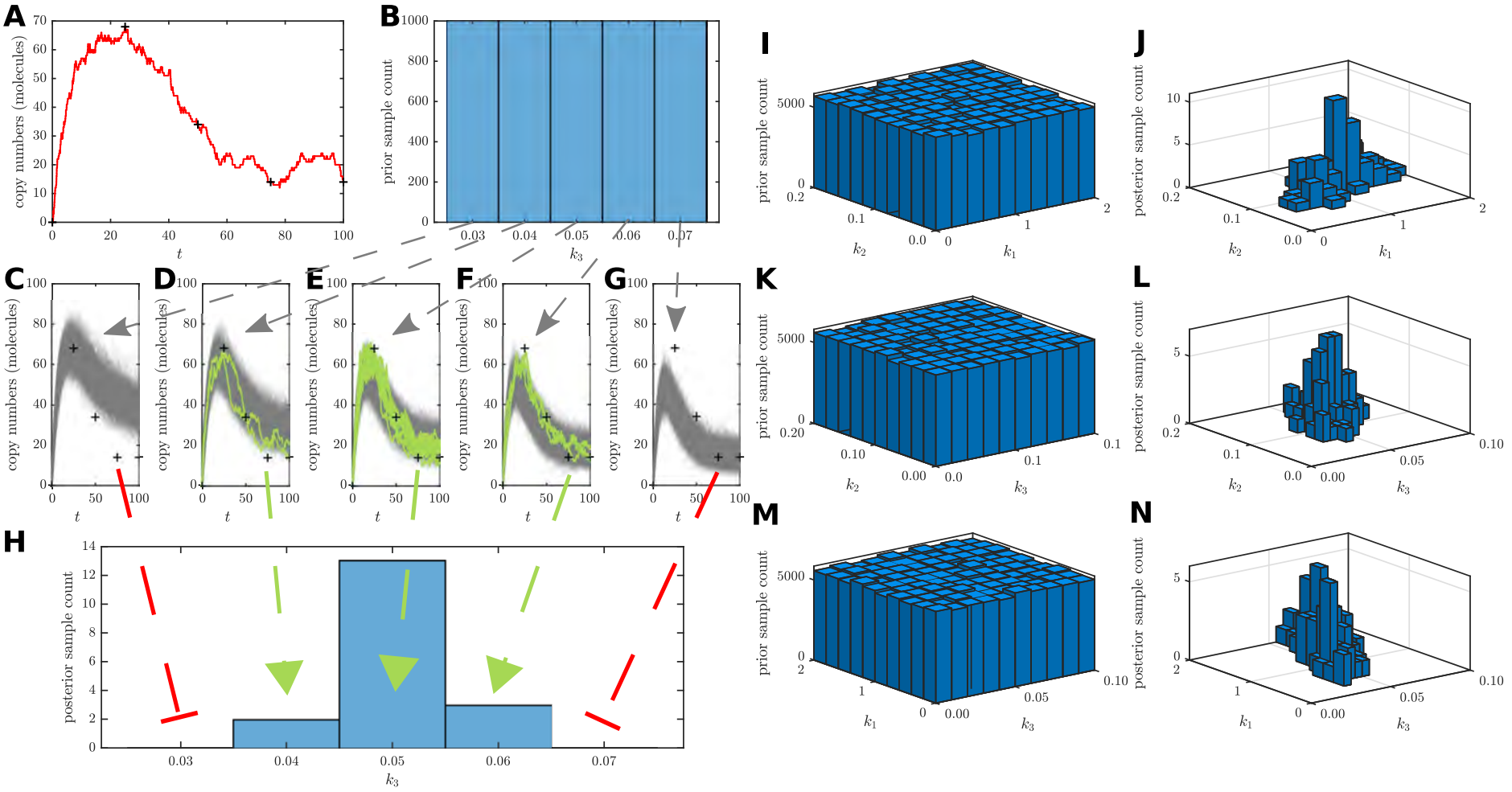

Figure 4.5: Demonstration of the ABC rejection sampler method using the mono-molecular chain model data set. (A) Experimental observations (black crosses) obtained from a true sample path of $B(t)$ in the mono-molecular chain model (red line) with $k_1 = 1.0$, $k_2 = 0.1$ and $k_3 = 0.05$, and initial conditions, $A(0) = 100$ and $B(0) = 0$. (B) Prior for inference on k_3. (C)–(G) Stochastic simulation of many choices of k_3 drawn from the prior, showing accepted (solid green) and rejected (solid gray) sample paths with $\epsilon = 15$ (molecules). (H) ABC posterior for k_3 generated from accepted samples. (I)–(N) Bivariate marginal distributions of the full ABC inference problem on $\boldsymbol{\theta} = \{k_1, k_2, k_3\}$.

Unfortunately, for small ϵ, the computational burden of the ABC rejection sampler may be prohibitive as the acceptance rate is very low (this is especially an issue for biochemical reaction networks with highly variable dynamics). For the example in Figure 4.4, $m = 100$ posterior samples takes approximately one minute for $\epsilon = 25$, but nearly ten hours for $\epsilon = 12.5$. However, for $\epsilon = 12.5$ the marginal ABC posterior for k_3 has not yet converged.

Marjoram et al. (2003) provide a solution via an ABC modification to Metropolis-Hastings method. This results in a Markov chain with the ABC posterior as the stationary distribution. This method is called *ABC Markov chain Monte Carlo* (ABCMCMC):

1. initialise $n = 0$ and select starting point $\boldsymbol{\theta}_\epsilon^{(0)}$;
2. generate a proposal sample, $\boldsymbol{\theta}^* \sim q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_\epsilon^{(n)})$;
3. generate simulated data, $\mathbf{S}_{\text{obs}}^* \sim s(\mathbf{S}_{\text{obs}}; \boldsymbol{\theta}^*)$;
4. if $\rho(\mathbf{Y}_{\text{obs}}, \mathbf{S}_{\text{obs}}^*) > \epsilon$, then set $\boldsymbol{\theta}_\epsilon^{(n+1)} = \boldsymbol{\theta}_\epsilon^{(n)}$and go to step 7, otherwise continue;
5. calculate acceptance probability $\alpha = \min\left(1, \dfrac{p(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}_\epsilon^{(n)} \mid \boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}_\epsilon^{(n)})q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}_\epsilon^{(n)})}\right)$;
6. with probability α, set $\boldsymbol{\theta}_\epsilon^{(n+1)} = \boldsymbol{\theta}^*$, and with probability $1 - \alpha$, set $\boldsymbol{\theta}_\epsilon^{(n+1)} = \boldsymbol{\theta}_\epsilon^{(n)}$;
7. update time, $n = n + 1$;
8. if $n > m_n$, terminate simulation, otherwise go to step 2;

An example implementation, `ABCMCMCSampler.m`, is provided.

Just as with the Metropolis-Hastings method, the efficacy of the ABCMCMC rests upon the non-trivial choice of the proposal kernel, $q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_\epsilon^{(n)})$. The challenge of constructing effective proposal kernels is equally non-trivial for ABCMCMC as for Metropolis-Hastings method. Figure 4.6 highlights different Markov chain behaviours for heuristically chosen proposal kernels based on Gaussian random walks with variances that we alter until the Markov chain seems to be mixing well.

For the mono-molecular chain model (Figure 4.6(A)–(C)), ABCMCMC seems to be performing well with the Markov chain state moving largely in the regions of high posterior probability.

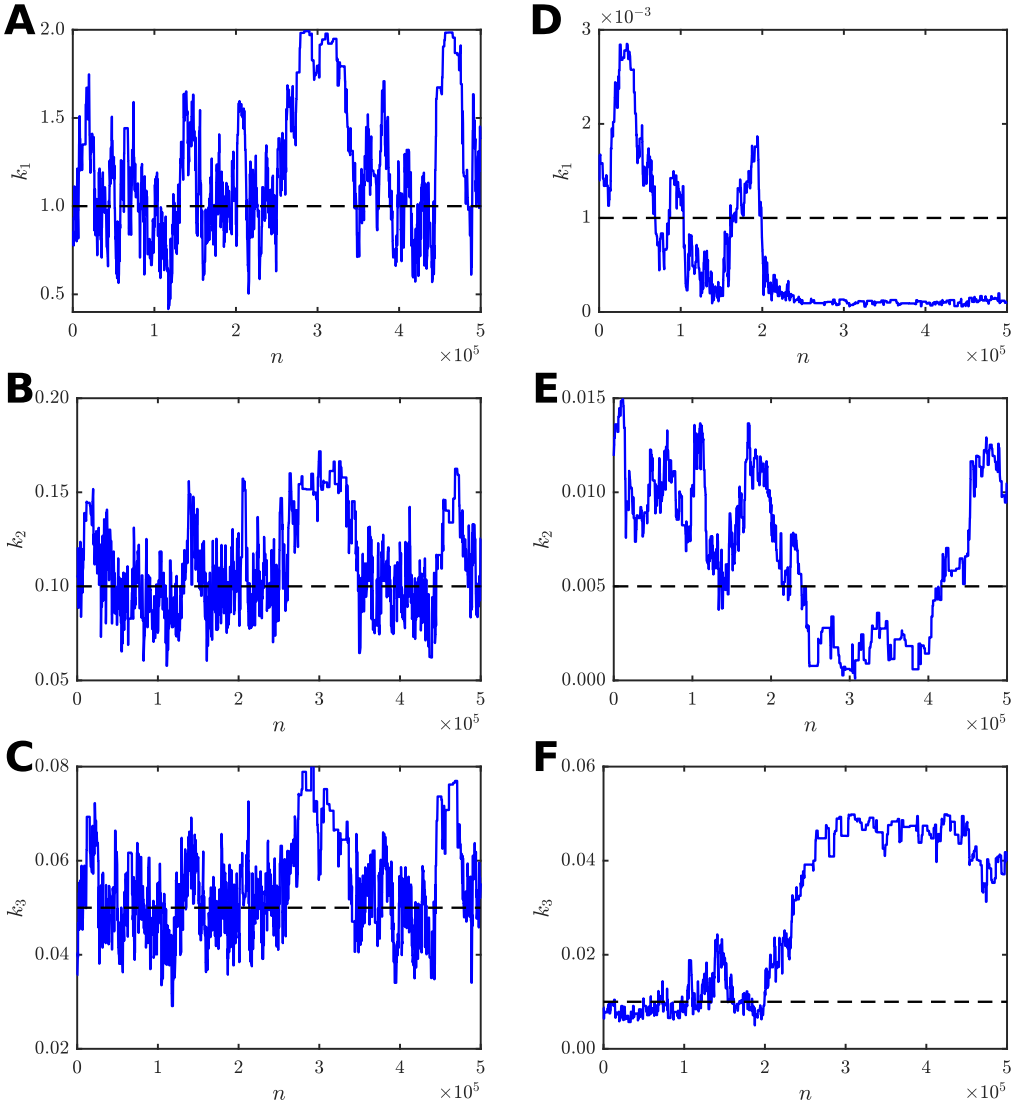

Figure 4.6: $m_n = 500,000$ steps of ABCMCMC for: (A)–(C) the mono-molecular chain model; and (D)–(F) the enzyme kinetics model. True parameter values (dash black) are shown.

However, for the enzyme kinetics model (Figure 4.6(D)–(F)), the Markov chain has taken a long excursion into a region of low posterior probability (see Figure 4.6(F)) where it is a long way from the true parameter value of $k_3 = 0.01$ for many steps. Such an extended path into the low probability region results in significant bias in this region and many many more steps of the algorithm are required to compensate for this (Sisson et al., 2007; Beaumont et al., 2009). Just as with exact MCMC, burn-in samples and thinning may be used to reduce correlations but may not improve the final Monte Carlo estimate.

To avoid the difficulties in ensuring efficient ABCMCMC convergence, Sisson et al. (2007) developed an ABC variant of SMC. Alternative versions of the method are also designed by Beaumont et al. (2009) and Toni et al. (2009). The fundamental idea is to use sequential

importance resampling to propagate m_p samples, called *particles*, through a sequence of $R+1$ ABC posterior distributions defined through a sequence of discrepancy thresholds, $\epsilon_0, \epsilon_1, \ldots, \epsilon_R$ with $\epsilon_r > \epsilon_{r+1}$ for $r = 0, 1, \ldots, R-1$ and $\epsilon_0 = \infty$ (that is, $p(\boldsymbol{\theta}_{\epsilon_0} \mid \mathbf{Y}_{\text{obs}})$ is the prior). This results in *ABC sequential Monte Carlo* (ABCSMC):

1. initialise $r = 0$ and weights $w_r^{(i)} = 1/m_p$ for $i = 1, \ldots, m_p$;
2. generate m_p particles from the prior, $\boldsymbol{\theta}_{\epsilon_r}^{(i)} \sim p(\boldsymbol{\theta})$, for $i = 1, 2, \ldots, m_p$;
3. set index $i = 0$;
4. randomly select integer, j, from the set $\{1, \ldots, m_p\}$ with $\mathbb{P}(j = 1) = w_r^{(1)}, \ldots, \mathbb{P}(j = m_p) = w_r^{(m_p)}$;
5. generate proposal, $\boldsymbol{\theta}^* \sim q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\epsilon_r}^{(j)})$;
6. generate simulated data, $\mathbf{S}_{\text{obs}}^* \sim s(\mathbf{S}_{\text{obs}}; \boldsymbol{\theta}^*)$;
7. if $\rho(\mathbf{Y}_{\text{obs}}, \mathbf{S}_{\text{obs}}^*) > \epsilon_{r+1}$, go to step 4, otherwise continue;
8. set $\boldsymbol{\theta}_{\epsilon_{r+1}}^{(i+1)*} = \boldsymbol{\theta}^*$, $w_{r+1}^{(i+1)*} = p(\boldsymbol{\theta}^*) \Big/ \left[\sum_{j=1}^{m_p} w_r^{(j)} q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}_{\epsilon_r}^{(j)})\right]$ and $i = i + 1$;
9. if $i < m_p$, go to step 4, otherwise continue;
10. set index $i = 0$;
11. randomly select integer, j, from the set $\{1, \ldots, m_p\}$ with $\mathbb{P}(j = 1) = w_{r+1}^{(1)*}, \ldots, \mathbb{P}(j = m_p) = w_{r+1}^{(m_p)*}$;
12. set $\boldsymbol{\theta}_{r+1}^{i+1} = \boldsymbol{\theta}_{r+1}^{(j)*}$ and $i = i + 1$;
13. if $i < m_p$, go to step 11, otherwise continue;
14. set $w_{r+1}^{(i)} = 1/\left[\sum_{j=1}^{m_p} w_{r+1}^{(j)*}\right]$ and $r = r + 1$;
15. if $r < R$, go to step 3, otherwise terminate simulation;

An implementation of the ABCSMC method is provided in `ABCSMCSampler.m`.

The ABCSMC method avoids some issues inherent in ABCMCMC such as strong sample correlation and long excursions into low probability regions. However, it is still the case that

the number of particles, m_p, often needs to be larger than the desired number of independent samples, m, from the ABC posterior with ϵ_R. This is especially true when there is a larger discrepancy between two successive ABC posteriors in the sequence. Although techniques exist to adaptively generate the sequence of acceptance thresholds (Drovandi and Pettitt, 2011). Also note that ABCSMC still requires a proposal kernel to mutate the particles at each step and the efficiency of the method is affected by the choice of proposal kernel. Just as with ABCMCMC, the task of selecting an optimal proposal kernel is non-trivial (Filippi et al., 2013).

The formulation of ABCSMC using a sequence of discrepancy thresholds hints that MLMC ideas could also be applicable. Recently, a variety of MLMC methods for ABC have been proposed (Guha and Tan, 2017; Jasra et al., 2017) (see Chapter 6). All of these approaches are similar in their application of the multilevel telescoping summation to compute expectations with respect to ABC posterior distributions,

$$\mathbb{E}\left[\boldsymbol{\theta}_{\epsilon_L} \mid \mathbf{Y}_{\text{obs}}\right] = \mathbb{E}\left[\boldsymbol{\theta}_{\epsilon_0} \mid \mathbf{Y}_{\text{obs}}\right] + \sum_{\ell=1}^{L} \mathbb{E}\left[\boldsymbol{\theta}_{\epsilon_\ell} - \boldsymbol{\theta}_{\epsilon_{\ell-1}} \mid \mathbf{Y}_{\text{obs}}\right], \tag{4.33}$$

where $\epsilon_0, \epsilon_1, \ldots, \epsilon_L$ is a sequence of discrepancy thresholds with $\epsilon_\ell > \epsilon_{\ell+1}$ for all $\ell = 0, 1, \ldots, L-1$.

Unlike the forwards problem, no exact solution has been found for the generation of correlated ABC posterior pairs, $(\boldsymbol{\theta}_{\epsilon_\ell}, \boldsymbol{\theta}_{\epsilon_{\ell-1}})$, with $\ell = 1, \ldots, L$. Several approaches to this problem have been proposed: Guha and Tan (2017) utilise a sequence of correlated ABCMCMC samplers in a similar manner to Dodwell et al. (2015) and Efendiev et al. (2015); Jasra et al. (2017) apply the MLSMC estimator of Beskos et al. (2017) directly to Equation (4.33); and Chapter 6 develops a MLMC variant of ABC rejection sampler through the introduction of an approximation based on the empirical marginal posterior distributions.

The coupling scheme of Chapter 6 is the simplest to present succinctly. The scheme is based on the inverse transform method for sampling univariate distributions. Given random variable $X \sim p(x)$, the cumulative distribution function is defined as,

$$F(z) = \mathbb{P}(X \leq z) = \int_{-\infty}^{z} p(x)\,\mathrm{d}x.$$

Note that $0 \leq F(z) \leq 1$. If $F(z)$ has an inverse function, $F^{-1}(u)$, and $U^{(1)}, U^{(2)}, \ldots, U^{(m)}$ are

independent samples of $\mathcal{U}(0,1)$, then independent samples from $p(x)$ can be generated using $X^{(1)} = F^{-1}(U^{(1)}), X^{(2)} = F^{-1}(U^{(2)}), \ldots, X^{(m)} = F^{-1}(U^{(m)})$.

Given samples, $\boldsymbol{\theta}_{\epsilon_\ell}^1, \boldsymbol{\theta}_{\epsilon_\ell}^2, \ldots, \boldsymbol{\theta}_{\epsilon_\ell}^{m_\ell}$, an estimate for the posterior marginal cumulative distribution functions can be obtained using

$$F_{\ell,j}(z) \approx \hat{F}_{\ell,j}(z) = \frac{1}{m_\ell} \sum_{i=1}^{m_\ell} \mathbb{1}_z \left(k_{\epsilon_\ell,j}^{(i)} \right),$$

where $k_{\epsilon_\ell,j}^{(i)}$ is the jth component of $\boldsymbol{\theta}_{\epsilon_\ell}^{(i)}$ for all $j = 1, \ldots, M$, and $\mathbb{1}_z(x) = 1$ if $x \leq z$ and $\mathbb{1}_z(x) = 0$ otherwise. Assuming we already have an estimate of the posterior marginal cumulative distribution functions for acceptance threshold $\epsilon_{\ell-1}$, then we apply the transform $k_{\epsilon_{\ell-1},j}^{(i)} = \hat{F}_{\ell-1,j}^{-1}(\hat{F}_{\ell,j}(k_{\epsilon_\ell,j}^{(i)}))$. This results in a coupled marginal pair $(k_{\epsilon_\ell,j}, k_{\epsilon_{\ell-1},j})^{(i)}$ since it is equivalent to using the inverse transform method to generate $k_{\epsilon_\ell,j}^{(i)}$ and $k_{\epsilon_{\ell-1},j}^{(i)}$ using the same uniform sample $U^{(i)} \sim \mathcal{U}(0,1)$. These approximately correlated samples can be obtained using:

1. generate sample from level ℓ ABC posterior $\boldsymbol{\theta}_{\epsilon_\ell} \sim p(\boldsymbol{\theta}_{\epsilon_\ell} \mid \mathbf{Y}_{\text{obs}})$ using the ABC rejection sampler;

2. set $k_{\epsilon_{\ell-1},j} = \hat{F}_{\ell-1,j}^{-1}(\hat{F}_{\ell,j}(k_{\epsilon_\ell,j}))$ for $j = 1, 2, \ldots, M$;

3. set $\boldsymbol{\theta}_{\epsilon_{\ell-1}} = \left[k_{\epsilon_{\ell-1},1}, k_{\epsilon_{\ell-1},2}, \ldots, k_{\epsilon_{\ell-1},M}\right]$;

Thus, the *ABC multilevel Monte Carlo* (ABCMLMC) (Chapter 6) method proceeds though computing the Monte Carlo estimate to Equation (4.33),

$$\hat{\boldsymbol{\theta}}_{\epsilon_\ell} = \frac{1}{m_0} \sum_{i=1}^{m_0} \boldsymbol{\theta}_{\epsilon_0}^{(i)} + \sum_{\ell=1}^{L} \frac{1}{m_\ell} \sum_{i=1}^{m_\ell} \left[\boldsymbol{\theta}_{\epsilon_\ell}^{(i)} - \boldsymbol{\theta}_{\epsilon_{\ell-1}}^{(i)} \right], \tag{4.34}$$

where $\boldsymbol{\theta}_{\epsilon_0}^{(1)}, \ldots, \boldsymbol{\theta}_{\epsilon_0}^{(m_0)}$ are m_0 independent samples using ABC rejection sampler and $(\boldsymbol{\theta}_{\epsilon_\ell}^{(1)}, \boldsymbol{\theta}_{\epsilon_{\ell-1}}^{(1)})$, $\ldots, (\boldsymbol{\theta}_{\epsilon_\ell}^{(m_\ell)}, \boldsymbol{\theta}_{\epsilon_{\ell-1}}^{(m_\ell)})$ are m_ℓ independent sample pairs using the approximately correlated ABC rejection sampler for $\ell = 1, 2, \ldots, L$. It is essential that each bias correction term in Equation (4.34) be estimated in ascending order to ensure the marginal CDF estimates for $\epsilon_{\ell-1}$ are available for use in the coupling scheme for the ℓth bias correction term. The complete algorithm is provided in Section 4.6.5 and the example code is provided in `ABCMLMC.m`. We refer the reader to Chapter 6 for further details.

Provided successive ABC posteriors are similar in *correlation structure*, then the ABCMLMC method accelerates ABC rejection sampler through reduction of the number of m_L samples required for a given error tolerance. However, additional bias is introduced by the approximately correlated ABC rejection sampler since the telescoping summation is not strictly satisfied. In practice, this affects the choice of the acceptance threshold sequence (Chapter 6).

Table 4.1: Comparison of ABC methods for the mono-molecular inference problem. Estimates of the posterior mean are given along with the 95% confidence intervals. Signed relative errors are shown for the confidence interval limits with respect to the true parameter values $k_1 = 1$, $k_2 = 0.1$, and $k_3 = 0.05$. Computations are performed using an Intel® Core™ i7-5600U CPU (2.6 GHz).

Method	Posterior mean estimate	Relative error	Compute Time
ABC rejection sampler	$\hat{k}_1 = 1.1690 \times 10^{0} \pm 7.1130 \times 10^{-2}$	$[9.8\%, 24.0\%]$	$7,268$ (sec)
	$\hat{k}_2 = 1.1011 \times 10^{-1} \pm 4.5105 \times 10^{-3}$	$[5.6\%, 14.6\%]$	
	$\hat{k}_3 = 5.3644 \times 10^{-2} \pm 1.9500 \times 10^{-3}$	$[3.4\%, 11.2\%]$	
ABCMCMC	$\hat{k}_1 = 1.2786 \times 10^{0} \pm 7.8982 \times 10^{-2}$	$[20.0\%, 35.8\%]$	$8,059$ (sec)
	$\hat{k}_2 = 1.1269 \times 10^{-1} \pm 5.2836 \times 10^{-3}$	$[7.4\%, 18.0\%]$	
	$\hat{k}_3 = 5.5644 \times 10^{-2} \pm 2.0561 \times 10^{-3}$	$[7.2\%, 15.4\%]$	
ABCSMC	$\hat{k}_1 = 1.0444 \times 10^{0} \pm 6.4736 \times 10^{-2}$	$[-2.0\%, 10.9\%]$	580 (sec)
	$\hat{k}_2 = 9.8318 \times 10^{-2} \pm 3.2527 \times 10^{-3}$	$[-4.9\%, 1.6\%]$	
	$\hat{k}_3 = 4.8685 \times 10^{-2} \pm 1.8032 \times 10^{-3}$	$[-6.2\%, 1.0\%]$	
ABCMLMC	$\hat{k}_1 = 1.1535 \times 10^{0} \pm 4.8907 \times 10^{-2}$	$[10.5\%, 20.2\%]$	$1,327$ (sec)
	$\hat{k}_2 = 1.0654 \times 10^{-1} \pm 4.8715 \times 10^{-3}$	$[1.7\%, 11.4\%]$	
	$\hat{k}_3 = 5.1265 \times 10^{-2} \pm 2.2210 \times 10^{-3}$	$[-2.0\%, 7.0\%]$	

We provide a comparison of the ABC rejection sampler, ABCMCMC, ABCSMC and ABCMLMC methods as presented here. Posterior means are computed for both the mono-molecular model and enzyme kinetics model inference problems along with the 95% confidence intervals for each estimate[7] (See Supplementary Material). Computation times and parameter estimates are provided in Tables 4.1 and 4.2. Example code, `DemoABCMethodsMonoMol.m` and `DemoABCMethodsMichMent.m`, is provided. Algorithm configurations are supplied in the supplementary material.

Note that no extensive tuning of the ABCMCMC, ABCSMC, or ABCMLMC algorithms is performed, thus these outcomes do not reflect a detailed and optimised implementation of the

[7]We report the 95% confidence interval as an accuracy measure for the Monte Carlo estimate of the true posterior mean. This does not quantify parameter uncertainty that would be achieved through posterior covariances and credible intervals.

Table 4.2: Comparison of ABC methods for the enzyme kinetics model inference problem. Estimates of the posterior mean are given along with the 95% confidence intervals. Signed relative errors are shown for the confidence interval limits with respect to the true parameter values $k_1 = 0.001$, $k_2 = 0.005$, and $k_3 = 0.01$. Computations are performed using an Intel® Core™ i7-5600U CPU (2.6 GHz).

Method	Posterior mean estimate	Relative error	Compute time
ABC rejection sampler	$\hat{k}_1 = 1.0098 \times 10^{-3} \pm 1.7011 \times 10^{-4}$	$[-16.0\%, 18.0\%]$	$2,037$ (sec)
	$\hat{k}_2 = 7.7203 \times 10^{-3} \pm 7.3490 \times 10^{-4}$	$[39.7\%, 69.1\%]$	
	$\hat{k}_3 = 1.5164 \times 10^{-2} \pm 2.1201 \times 10^{-3}$	$[30.4\%, 72.8\%]$	
ABCMCMC	$\hat{k}_1 = 3.0331 \times 10^{-4} \pm 7.5670 \times 10^{-5}$	$[-77.2\%, -62.1\%]$	$2,028$ (sec)
	$\hat{k}_2 = 5.8106 \times 10^{-3} \pm 7.9993 \times 10^{-4}$	$[0.2\%, 32.2\%]$	
	$\hat{k}_3 = 3.3865 \times 10^{-2} \pm 3.0187 \times 10^{-3}$	$[208.5\%, 268.8\%]$	
ABCSMC	$\hat{k}_1 = 9.8972 \times 10^{-4} \pm 1.4630 \times 10^{-4}$	$[-15.7\%, 13.6\%]$	342 (sec)
	$\hat{k}_2 = 9.2680 \times 10^{-3} \pm 9.1997 \times 10^{-4}$	$[67.0\%, 103.8\%]$	
	$\hat{k}_3 = 1.3481 \times 10^{-2} \pm 1.3278 \times 10^{-3}$	$[21.5\%, 48.1\%]$	
ABCMLMC	$\hat{k}_1 = 1.0859 \times 10^{-3} \pm 6.7058 \times 10^{-5}$	$[1.9\%, 15.3\%]$	795 (sec)
	$\hat{k}_2 = 6.8111 \times 10^{-3} \pm 3.2107 \times 10^{-4}$	$[29.9\%, 42.6\%]$	
	$\hat{k}_3 = 1.3663 \times 10^{-2} \pm 1.0729 \times 10^{-3}$	$[25.9\%, 47.4\%]$	

methods. Such a benchmark would be significantly more computationally intensive than the comparison of Monte Carlo methods for the forwards problem (Figure 4.3). The computational statistics literature demonstrates that ABCMCMC (Marjoram et al., 2003), ABCSMC (Beaumont et al., 2009; Sisson et al., 2007) and ABCMLMC (Guha and Tan, 2017; Jasra et al., 2017) (Chapter 6) can be tuned to provide very competitive results on a given inference problem. However, a large number of trial computations are often required to achieve this tuning, or more complex adaptive schemes need to be exploited (Drovandi and Pettitt, 2011; Roberts and Rosenthal, 2009). Instead, our comparisons represent a more practical guide in which computations are kept short and almost no tuning is performed, thus giving a fair comparison for a fixed, short computational budget.

For both inference problems, ABCSMC and ABCMLMC produce more accurate parameter estimates in less computation time than ABC rejection sampler and ABCMCMC. ABCSMC performs better on the mono-molecular chain model and ABCMLMC performs better on the enzyme kinetic model. This is not to say that ABCMCMC cannot be tuned to perform more efficiently, but it does indicate that ABCMCMC is harder to tune without more extensive testing. An advantage of ABCMLMC is that it has fewer components that require tuning, since

the sample numbers may be optimally chosen, and the user need only provide a sequence of acceptance thresholds (Chapter 6). However, the ability to tune the proposal kernel in ABCSMC can be a significant advantage, despite the challenge of determining a good choice for such a proposal (Drovandi and Pettitt, 2011; Filippi et al., 2013).

4.4.4 Summary of the inverse problem

Bayesian approaches for uncertainty quantification and parameter inference have proven to be powerful techniques in the life sciences. However, in the study of intracellular biochemical processes, the intractability of the likelihood causes a significant computational challenge.

A crucial advance towards obtaining numerical solutions in this Bayesian inverse problem setting has been the advent of likelihood-free methods that replace likelihood evaluations with stochastic simulation. Through such methods, especially those based on ABC, a direct connection between the forwards and inverse problems is made explicit. Not only is efficient forwards simulation key, but information obtained through each new sample path must be effectively used. While ABC rejection sampler ignores the latter, advanced methods like ABCMCMC, ABCSMC, and ABCMLMC all provide different mechanisms for incorporating this new information. For a specific inverse problem, however, it will often not be clear which method will perform optimally. In general, significant trial sampling must be performed to get the most out of any of these algorithms.

The application of MLMC methods to ABC-based samplers is a very new area of research. There are many open questions that relate to the appropriateness of various approximations and coupling schemes for a given inference problem. The method of Chapter 6 is conceptually straightforward, however, it is known that additional bias is incurred through the coupling scheme. Alternative coupling approximations that combine MLMC and MCMC (Guha and Tan, 2017) or SMC (Jasra et al., 2017) may be improvements, or a hybrid scheme could be derived through a combination of these methods. Currently, it is not clear which method is the most widely applicable.

4.5 Summary and Outlook

Stochastic models of biochemical reaction networks are routinely used to study various intracellular processes, such as gene regulation. By studying the forwards problem, stochastic models can be explored *in silico* to gain insight into potential hypotheses related to observed phenomena and inform potential experimentation. Models may be calibrated and key kinetic information may be extracted from time course data through the inverse problem. Both the forward and inverse problems are, in practice, reliant upon computation.

Throughout this chapter, we deliberately avoid detailed theoretical derivations in favour of practical computational or algorithmic examples. Since all of our code examples are specifically developed within the user friendly MATLAB® programming environment, our codes do not represent optimal implementations. Certainly, there are software packages available that implement some of these techniques and others that we have not discussed. For example, stochastic simulation is available in `COPASI` (Hoops et al., 2006), `StochKit` (Li et al., 2008; Sanft et al., 2011), `StochPy` (Maarleveld et al., 2013), and `STEPS` (Hepburn et al., 2012), and ABC-based inference is available in `ABC-SysBio` (Liepe et al., 2014), `abctools` (Nunes and Prangle, 2015), and `ABCtoolbox` (Wegmann et al., 2010).

Throughout this chapter, we have focused on two biochemical reaction network models, the mono-molecular chain model for which the chemical master equation can be solved analytically, and the enzyme kinetics model with an intractable chemical master equation. However, the majority of our MATLAB® implementations are completely general and apply to an arbitrary biochemical reaction network model. The biochemical reaction network construction functions, `MonoMolecularChain.m` and `MichaelisMenten.m`, demonstrate how to construct a biochemical reaction network data structure that most of our scripts and functions may directly use, with the only exceptions being those that relate to the analytical solution to the chemical master equation of the mono-molecular chain model. Therefore, our code may be viewed as a useful collection of prototype implementations that may be easily applied in new contexts.

The techniques we present here are powerful tools for the analysis of real biological data and design of experiments. In particular, these methods are highly relevant for the reconstruction of genetic regulatory networks from gene expression microarray data (Cohen et al., 2015;

Gai et al., 2018; Ocone et al., 2015). Cell signalling pathways can also be analysed using mass-spectroscopy proteomic data (Wang et al., 2016; Tian, 2010). Furthermore, Bayesian optimal experimental design using stochastic biochemical reaction networks is key to reliable characterisation of light-induced gene expression (Ruess et al., 2015), a promising technique for external genetic regulation *in vivo* (Jayaraman et al., 2016; Yamada et al., 2018).

This chapter provides researchers in the life sciences with the fundamental concepts and computational tools required to apply stochastic biochemical reaction network models in practice. Through practical demonstration, the current state-of-the-art in simulation and Monte Carlo methods will be more widely and readily applied by the broader life sciences community.

4.6 Supplementary material

4.6.1 Derivation of mono-molecular chain mean and variances using the chemical master equation

In this section, we provide an example on how to derive moments of the chemical master equation (CME) solution without explicit CME evaluation. The presented analysis is specific to the two species mono-molecular chain model as presented in the main text. Our approach is based on the examples from Erban et al. (2007), however, the result is more complex since we deal with a two chemical species, A and B.

For convenience, we restate the model. Here we consider a two species mono-molecular chain,

$$\emptyset \xrightarrow{k_1} A \xrightarrow{k_2} B \xrightarrow{k_3} \emptyset, \tag{4.35}$$

with known kinetic rate parameters k_1, k_2 and k_3. Given the state vector, $\mathbf{X}(t) = [A(t), B(t)]^T$, the respective propensity functions are

$$a_1(\mathbf{X}(t)) = k_1, \quad a_2(\mathbf{X}(t)) = k_2 A(t), \quad a_3(\mathbf{X}(t)) = k_3 B(t). \tag{4.36}$$

The stoichiometric vectors are

$$\nu_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \nu_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \quad \nu_3 = \begin{bmatrix} 0 \\ -1 \end{bmatrix}. \tag{4.37}$$

For $P(\mathbf{x}, t \mid \mathbf{x}_0) = \mathbb{P}(\mathbf{X}(t) = \mathbf{x} \mid \mathbf{X}(0) = \mathbf{x}_0)$, the general form of the CME is

$$\frac{\mathrm{d}P(\mathbf{x}, t \mid \mathbf{x}_0)}{\mathrm{d}t} = \sum_{j=1}^{M} a_j(\mathbf{x} - \nu_j) P(\mathbf{x} - \nu_j, t \mid \mathbf{x}_0) - P(\mathbf{x}, t \mid \mathbf{x}_0) \sum_{j=1}^{M} a_j(\mathbf{x}). \tag{4.38}$$

After substituting the propensity functions (Equation (4.36)) and stoichiometric vectors (Equation (4.37)) into Equation (4.38), we obtain the CME specific to the mono-molecular chain

model (Equation (4.35))

$$\frac{\mathrm{d}P(a,b,t \mid a_0,b_0)}{\mathrm{d}t} = k_1 P(a-1,b,t \mid a_0,b_0) + k_2(a+1)P(a+1,b-1,t \mid a_0,b_0) \\ + k_3(b+1)P(a,b+1,t \mid a_0,b_0) - (k_1 + k_2 a + k_3 b)P(a,b,t \mid a_0,b_0). \tag{4.39}$$

Henceforth, we will denote $p_{a,b}(t)$ as the solution to the mono-molecular CME (Equation (4.39)).

Rather than solve the full CME, we seek a solution to the mean copy number of A at time t,

$$M_a(t) = \sum_{a=0}^{\infty}\sum_{b=0}^{\infty} a p_{a,b}(t), \tag{4.40}$$

the mean copy number of B at time t,

$$M_b(t) = \sum_{a=0}^{\infty}\sum_{b=0}^{\infty} b p_{a,b}(t), \tag{4.41}$$

the variance of A at time t,

$$V_a(t) = \sum_{a=0}^{\infty}\sum_{b=0}^{\infty} \left(a - M_a(t)\right)^2 p_{a,b}(t), \tag{4.42}$$

the variance of B at time t,

$$V_b(t) = \sum_{a=0}^{\infty}\sum_{b=0}^{\infty} \left(b - M_b(t)\right)^2 p_{a,b}(t), \tag{4.43}$$

and the covariance of A and B at time t,

$$C_{a,b}(t) = \sum_{a=0}^{\infty}\sum_{b=0}^{\infty} \left(a - M_a(t)\right)\left(b - M_b(t)\right) p_{a,b}(t). \tag{4.44}$$

We will derive a system of ODEs that describe the evolution of $M_a(t)$,$M_b(t)$, $V_a(t)$, $V_b(t)$, and $C_{a,b}(t)$ without explicitly solving the CME in Equation (4.39). Instead we exploit the linearity of the derivative along with the property,

$$\sum_{a=0}^{\infty}\sum_{b=0}^{\infty} p_{a,b}(t) = 1, \tag{4.45}$$

for all t.

To derive an ODE for $M_a(t)$, we multiply Equation (4.39) by a and sum over all a and b.

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[\sum_{a=0}^{\infty}\sum_{b=0}^{\infty} a p_{a,b}(t)\right] = \sum_{a=1}^{\infty}\sum_{b=0}^{\infty} k_1 a p_{a-1,b}(t) + \sum_{a=0}^{\infty}\sum_{b=1}^{\infty} k_2 a(a+1) p_{a+1,b-1}(t) + \sum_{a=0}^{\infty}\sum_{b=0}^{\infty} k_3 a(b+1) p_{a,b+1}(t) - \sum_{a=0}^{\infty}\sum_{b=0}^{\infty} a(k_1 + k_2 a + k_3 b) p_{a,b}(t).$$

After changing indices ($a-1 \to a$ in the first term, $a+1 \to a$ and $b-1 \to b$ in the second term, and $b+1 \to b$ in the third term), we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[\sum_{a=0}^{\infty}\sum_{b=0}^{\infty} a p_{a,b}(t)\right] = \sum_{a=0}^{\infty}\sum_{b=0}^{\infty} k_1 (a+1) p_{a,b}(t) + \sum_{a=0}^{\infty}\sum_{b=0}^{\infty} k_2 (a-1) a p_{a,b}(t) + \sum_{a=0}^{\infty}\sum_{b=0}^{\infty} k_3 ab p_{a,b}(t) - \sum_{a=0}^{\infty}\sum_{b=0}^{\infty} a(k_1 + k_2 a + k_3 b) p_{a,b}(t).$$

We simplify the right hand side,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[\sum_{a=0}^{\infty}\sum_{b=0}^{\infty} a p_{a,b}(t)\right] = k_1 \sum_{a=0}^{\infty}\sum_{b=0}^{\infty} p_{a,b}(t) - k_2 \sum_{a=0}^{\infty}\sum_{b=0}^{\infty} a p_{a,b}(t),$$

and apply property (4.45) to give

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[\sum_{a=0}^{\infty}\sum_{b=0}^{\infty} a p_{a,b}(t)\right] = k_1 - k_2 \sum_{a=0}^{\infty}\sum_{b=0}^{\infty} a p_{a,b}(t).$$

Using the definition of $M_a(t)$ (Equation (4.40)), we obtain the ODE for the mean of A,

$$\frac{\mathrm{d}M_a(t)}{\mathrm{d}t} = k_1 - k_2 M_a(t). \tag{4.46}$$

Similarly, we derive an ODE for $M_b(t)$ by muliplying Equation (4.39) by b and proceed in the

same manner as we did for $M_a(t)$

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\left[\sum_{a=0}^{\infty}\sum_{b=0}^{\infty} bp_{a,b}(t)\right] &= \sum_{a=1}^{\infty}\sum_{b=0}^{\infty} k_1 bp_{a-1,b}(t) + \sum_{a=0}^{\infty}\sum_{b=1}^{\infty} k_2 b(a+1)p_{a+1,b-1}(t) \\
&\quad + \sum_{a=0}^{\infty}\sum_{b=0}^{\infty} k_3 b(b+1)p_{a,b+1}(t) - \sum_{a=0}^{\infty}\sum_{b=0}^{\infty} b(k_1+k_2a+k_3b)p_{a,b}(t) \\
&= \sum_{a=0}^{\infty}\sum_{b=0}^{\infty} k_1 bp_{a,b}(t) + \sum_{a=0}^{\infty}\sum_{b=0}^{\infty} k_2(b+1)ap_{a,b}(t) \\
&\quad + \sum_{a=0}^{\infty}\sum_{b=0}^{\infty} k_3(b-1)bp_{a,b}(t) - \sum_{a=0}^{\infty}\sum_{b=0}^{\infty} b(k_1+k_2a+k_3b)p_{a,b}(t) \\
&= k_2\sum_{a=0}^{\infty}\sum_{b=0}^{\infty} ap_{a,b}(t) - k_3\sum_{a=0}^{\infty}\sum_{b=0}^{\infty} bp_{a,b}(t).
\end{aligned}$$

Using the definitions of $M_a(t)$ (Equation (4.40)) and $M_b(t)$ (Equation (4.47)) we obtain the ODE for the mean of B,

$$\frac{\mathrm{d}M_b(t)}{\mathrm{d}t} = k_2M_a(t) - k_3M_b(t). \tag{4.47}$$

To derive the ODE for $V_a(t)$, first note that through expanding Equation (4.42) it can be shown that

$$V_a(t) + M_a(t)^2 = \sum_{a=0}^{\infty}\sum_{b=0}^{\infty} a^2p_{a,b}(t). \tag{4.48}$$

Thus, we multiply Equation (4.39) by a^2, sum over all a and b, change indices, and simplify as follows,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\left[\sum_{a=0}^{\infty}\sum_{b=0}^{\infty} a^2p_{a,b}(t)\right] &= \sum_{a=1}^{\infty}\sum_{b=0}^{\infty} k_1a^2p_{a-1,b}(t) + \sum_{a=0}^{\infty}\sum_{b=1}^{\infty} k_2a^2(a+1)p_{a+1,b-1}(t) \\
&\quad + \sum_{a=0}^{\infty}\sum_{b=0}^{\infty} k_3a^2(b+1)p_{a,b+1}(t) - \sum_{a=0}^{\infty}\sum_{b=0}^{\infty} a^2(k_1+k_2a+k_3b)p_{a,b}(t) \\
&= \sum_{a=0}^{\infty}\sum_{b=0}^{\infty} k_1(a+1)^2p_{a,b}(t) + \sum_{a=0}^{\infty}\sum_{b=0}^{\infty} k_2(a-1)^2ap_{a,b}(t) \\
&\quad + \sum_{a=0}^{\infty}\sum_{b=0}^{\infty} k_3a^2bp_{a,b}(t) - \sum_{a=0}^{\infty}\sum_{b=0}^{\infty} a^2(k_1+k_2a+k_3b)p_{a,b}(t) \\
&= k_1\sum_{a=0}^{\infty}\sum_{b=0}^{\infty} p_{a,b}(t) + (2k_1+k_2)\sum_{a=0}^{\infty}\sum_{b=0}^{\infty} ap_{a,b}(t) - 2k_2\sum_{a=0}^{\infty}\sum_{b=0}^{\infty} a^2p_{a,b}(t) \\
&= k_1 + (2k_1+k_2)\sum_{a=0}^{\infty}\sum_{b=0}^{\infty} ap_{a,b}(t) - 2k_2\sum_{a=0}^{\infty}\sum_{b=0}^{\infty} a^2p_{a,b}(t).
\end{aligned}$$

Using the definition of $M_a(t)$ (Equation (4.46)) and property (4.48), we have the ODE,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[V_a(t)+M_a(t)^2\right]=k_1+(2k_1+k_2)M_a(t)-2k_2\left(V_a(t)+M_a(t)^2\right).$$

Apply the chain rule to obtain,

$$\frac{\mathrm{d}V_a(t)}{\mathrm{d}t}=-2M_a(t)\frac{\mathrm{d}M_a(t)}{\mathrm{d}t}+k_1+(2k_1+k_2)M_a(t)-2k_2\left(V_a(t)+M_a(t)^2\right), \tag{4.49}$$

then substitute Equation (4.46) into Equation (4.49) and simplify to arrive at the ODE for the variance of A,

$$\frac{\mathrm{d}V_a(t)}{\mathrm{d}t}=k_1+k_2M_a(t)-2k_2V_a(t). \tag{4.50}$$

Similarly, to derive the ODE for $V_b(t)$ we note that through expanding Equation (4.43) and Equation (4.44) it can be shown that

$$V_b(t)+M_b(t)^2=\sum_{a=0}^{\infty}\sum_{b=0}^{\infty}b^2p_{a,b}(t), \tag{4.51}$$

and

$$C_{a,b}(t)+M_a(t)M_b(t)=\sum_{a=0}^{\infty}\sum_{b=0}^{\infty}abp_{a,b}(t). \tag{4.52}$$

Thus, we multiply Equation (4.39) by b^2, sum over all a and b, change indices, and simplify as follows,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\left[\sum_{a=0}^{\infty}\sum_{b=0}^{\infty}b^2p_{a,b}(t)\right]&=\sum_{a=1}^{\infty}\sum_{b=0}^{\infty}k_1b^2p_{a-1,b}(t)+\sum_{a=0}^{\infty}\sum_{b=1}^{\infty}k_2b^2(a+1)p_{a+1,b-1}(t)\\
&\quad+\sum_{a=0}^{\infty}\sum_{b=0}^{\infty}k_3b^2(b+1)p_{a,b+1}(t)-\sum_{a=0}^{\infty}\sum_{b=0}^{\infty}b^2(k_1+k_2a+k_3b)p_{a,b}(t)\\
&=\sum_{a=0}^{\infty}\sum_{b=0}^{\infty}k_1b^2p_{a,b}(t)+\sum_{a=0}^{\infty}\sum_{b=0}^{\infty}k_2(b+1)^2ap_{a,b}(t)\\
&\quad+\sum_{a=0}^{\infty}\sum_{b=0}^{\infty}k_3(b-1)^2bp_{a,b}(t)-\sum_{a=0}^{\infty}\sum_{b=0}^{\infty}b^2(k_1+k_2a+k_3b)p_{a,b}(t)\\
&=k_2\sum_{a=0}^{\infty}\sum_{b=0}^{\infty}ap_{a,b}(t)+k_3\sum_{a=0}^{\infty}\sum_{b=0}^{\infty}bp_{a,b}(t)\\
&\quad+2k_2\sum_{a=0}^{\infty}\sum_{b=0}^{\infty}abp_{a,b}(t)-2k_3\sum_{a=0}^{\infty}\sum_{b=0}^{\infty}b^2p_{a,b}(t).
\end{aligned}$$

Using the definitions of $M_a(t)$ (Equation (4.40)) and $M_b(t)$ (Equation (4.41)), and properties (4.51) and (4.52) we have the ODE

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[V_b(t) + M_b(t)^2\right] = k_2 M_a(t) + k_3 M_b(t) + 2k_2\left(C_{a,b}(t) + M_a(t)M_b(t)\right) - 2k_3\left(V_b(t) + M_b(t)^2\right).$$

We apply the chain rule

$$\begin{aligned}\frac{\mathrm{d}V_b(t)}{\mathrm{d}t} &= -2M_b(t)\frac{\mathrm{d}M_b(t)}{\mathrm{d}t} + k_2 M_a(t) + k_3 M_b(t) \\ &\quad + 2k_2\left(C_{a,b}(t) + M_a(t)M_b(t)\right) - 2k_3\left(V_b(t) + M_b(t)^2\right),\end{aligned} \tag{4.53}$$

then substitute Equation (4.47) into Equation (4.53) and simplify to obtain the ODE for the variance of B

$$\frac{\mathrm{d}V_b(t)}{\mathrm{d}t} = k_2 M_a(t) + k_3 M_b(t) + 2k_2 C_{a,b} - 2k_3 V_b(t). \tag{4.54}$$

Finally, we derive the ODE for $C_{a,b}(t)$ by multiplying Equation (4.39) by ab, summing over all a and b, changing indices, and simplifying as follows:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left[\sum_{a=0}^{\infty}\sum_{b=0}^{\infty} abp_{a,b}(t)\right] &= \sum_{a=1}^{\infty}\sum_{b=0}^{\infty} k_1 abp_{a-1,b}(t) + \sum_{a=0}^{\infty}\sum_{b=1}^{\infty} k_2 ab(a+1)p_{a+1,b-1}(t) \\ &\quad + \sum_{a=0}^{\infty}\sum_{b=0}^{\infty} k_3 ab(b+1)p_{a,b+1}(t) - \sum_{a=0}^{\infty}\sum_{b=0}^{\infty} ab(k_1 + k_2 a + k_3 b)p_{a,b}(t) \\ &= \sum_{a=0}^{\infty}\sum_{b=0}^{\infty} k_1(a+1)bp_{a,b}(t) + \sum_{a=0}^{\infty}\sum_{b=0}^{\infty} k_2(a-1)(b+1)ap_{a,b}(t) \\ &\quad + \sum_{a=0}^{\infty}\sum_{b=0}^{\infty} k_3 a(b-1)bp_{a,b}(t) - \sum_{a=0}^{\infty}\sum_{b=0}^{\infty} ab(k_1 + k_2 a + k_3 b)p_{a,b}(t) \\ &= k_1\sum_{a=0}^{\infty}\sum_{b=0}^{\infty} bp_{a,b}(t) - k_2\sum_{a=0}^{\infty}\sum_{b=0}^{\infty} bp_{a,b}(t) \\ &\quad - (k_2 + k_3)\sum_{a=0}^{\infty}\sum_{b=0}^{\infty} abp_{a,b}(t) + k_2\sum_{a=0}^{\infty}\sum_{b=0}^{\infty} a^2 p_{a,b}(t).\end{aligned}$$

Using the definition of $M_a(t)$ (Equation (4.40)) and $M_b(t)$ (Equation (4.41)), and properties (4.48)

and (4.52) we obtain the ODE

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[C_{a,b}(t)+M_a(t)M_b(t)\right]=k_1M_a(t)-k_2M_b(t)-(k_2+k_3)\left(C_{a,b}(t)+M_a(t)M_b(t)\right)+k_2\left(V_a(t)+M_a(t)^2\right).$$

Apply the chain rule and product rule

$$\begin{aligned}\frac{\mathrm{d}C_{a,b}(t)}{\mathrm{d}t}=&-M_a(t)\frac{\mathrm{d}M_b(t)}{\mathrm{d}t}-M_b(t)\frac{\mathrm{d}M_a(t)}{\mathrm{d}t}+k_1M_a(t)-k_2M_b(t)\\&-(k_2+k_3)\left(C_{a,b}(t)+M_a(t)M_b(t)\right)+k_2\left(V_a(t)+M_a(t)^2\right),\end{aligned}\tag{4.55}$$

then substitute Equation (4.46) and Equation (4.47) into Equation (4.55) and simplify to obtain the ODE for the covariance of A and B

$$\frac{\mathrm{d}C_{a,b}(t)}{\mathrm{d}t}=k_2V_a(t)-k_2M_a(t)-(k_2+k_3)C_{a,b}(t).\tag{4.56}$$

Therefore, Equations (4.46),(4.47),(4.50),(4.54), and (4.56) form a non-homogeneous linear system of ODEs,

$$\begin{aligned}\frac{\mathrm{d}M_a(t)}{\mathrm{d}t}&=k_1-k_2M_a(t),\\\frac{\mathrm{d}M_b(t)}{\mathrm{d}t}&=k_2M_a(t)-k_3M_b(t),\\\frac{\mathrm{d}V_a(t)}{\mathrm{d}t}&=k_1+k_2M_a(t)-2k_2V_a(t),\\\frac{\mathrm{d}V_b(t)}{\mathrm{d}t}&=k_2M_a(t)+k_3M_b(t)+2k_2C_{a,b}(t)-k_3V_b(t),\\\frac{\mathrm{d}C_{a,b}(t)}{\mathrm{d}t}&=k_2V_a(t)-k_2M_a(t)-(k_2+k_3)C_{a,b}(t).\end{aligned}$$

After solving for the homogeneous solution, a particular solution may be obtained through using the method of undetermined coefficients. Given the initial conditions $A(0)=a_0$ and $B(0)=b_0$

with probability one, the solution, in the case when $k_2 \neq k_3$, is

$$M_a(t) = \frac{k_1}{k_2} + \left(a_0 - \frac{k_1}{k_2}\right) e^{-k_2 t}, \tag{4.57}$$

$$M_b(t) = \frac{k_1}{k_3} + \frac{k_2 a_0 - k_1}{k_3 - k_2} e^{-k_2 t} + \left(b_0 - \frac{k_2 a_0 - k_1}{k_3 - k_2} - \frac{k_1}{k_3}\right) e^{-k_3 t}, \tag{4.58}$$

$$V_a(t) = \frac{k_1}{k_2} + \left(a_0 - \frac{k_1}{k_2}\right) e^{-k_2 t} - a_0 e^{-2k_2 t}, \tag{4.59}$$

$$\begin{aligned} V_b(t) = {} & \frac{k_1}{k_3} + \frac{k_2 a_0 - k_1}{k_3 - k_2} e^{-k_2 t} + \left(b_0 - \frac{k_2 a_0 - k_1}{k_3 - k_2} - \frac{k_1}{k_3}\right) e^{-k_3 t} + \frac{2 a_0 k_2^2}{k_3^2 - k_2^2} e^{-(k_3 + k_2)t} \\ & - \frac{a_0 k_2}{k_3 - k_2} e^{-2k_2 t} + \left[\frac{a_0 k_2}{k_3 - k_2}\left(1 - \frac{2k_2}{k_3 + k_2}\right) - b_0\right] e^{-2k_3 t}, \end{aligned} \tag{4.60}$$

$$C_{a,b}(t) = -\frac{a_0 k_2}{k_3 - k_2} e^{-(k_3 + k_2)t} + \frac{a_0 k_2}{k_3 - k_2} e^{-2k_2 t}. \tag{4.61}$$

Equations (4.57)–(4.61) are the time dependent solutions for the moments and from these solutions we can evaluate the long time limit, $t \to \infty$, to give simple expression for the associated stationary solutions,

$$\lim_{t\to\infty} M_a(t) = \frac{k_1}{k_2}, \quad \lim_{t\to\infty} V_a(t) = \frac{k_1}{k_2}, \quad \lim_{t\to\infty} M_b(t) = \frac{k_1}{k_3}, \quad \lim_{t\to\infty} V_b(t) = \frac{k_1}{k_3}, \quad \lim_{t\to\infty} C_{a,b}(t) = 0.$$

See the example codes `DemoCMEMeanVar.m` and `DemoStationaryDist.m` for the evaluation of this solution.

4.6.2 Evaluation of the mono-molecular chain chemical master equation solution

Jahnke and Huisinga (2007) derive an analytic solution to the CME for a general mono-molecular BCRN. Applying their general solution to the species mono-molecular chain (Equation (4.35)) results in the following general solution to the CME (Equation (4.39)),

$$P(a, b, t \mid a_0, b_0) = \mathcal{P}\left(a, b, \lambda_a(t), \lambda_b(t)\right) * \mathcal{M}\left(a, b, a_0, \alpha_a(t), \alpha_b(t)\right) * \mathcal{M}\left(a, b, b_0, \beta_a(t), \beta_b(t)\right), \tag{4.62}$$

where $*$ is the discrete convolution operation (Jahnke and Huisinga, 2007). $\mathcal{P}\left(a, b, \lambda_a(t), \lambda_b(t)\right)$

is a product Poisson distribution, given by

$$\mathcal{P}\left(a,b,\lambda_a(t),\lambda_b(t)\right)=\begin{cases}\dfrac{\lambda_a(t)^a}{a!}\dfrac{\lambda_b(t)^b}{b!}e^{-(|\lambda_a(t)|+|\lambda_b(t)|)}, & \text{if } a\geq 0, b\geq 0,\\ 0 & \text{otherwise,}\end{cases} \tag{4.63}$$

where the functions $\lambda_a(t)$ and $\lambda_b(t)$ are obtained through the initial value problem (IVP)

$$\frac{\mathrm{d}\lambda_a(t)}{\mathrm{d}t}=k_1-k_2\lambda_a(t),\quad \frac{\mathrm{d}\lambda_b(t)}{\mathrm{d}t}=k_2\lambda_a(t)-k_3\lambda_b(t),\quad t>0, \tag{4.64}$$

with initial conditions $\lambda_a(0)=\lambda_b(0)=0$. $\mathcal{M}\left(a,b,a_0,\alpha_a(t),\alpha_b(t)\right)$ and $\mathcal{M}\left(a,b,b_0,\beta_a(t),\beta_b(t)\right)$ are multinomial distributions, given by

$$\mathcal{M}\left(a,b,a_0,\alpha_a(t),\alpha_b(t)\right)=\begin{cases}a_0!\dfrac{\left(1-|\alpha_a(t)|-|\alpha_b(t)|\right)^{a_0-|a|-|b|}}{(a_0-|a|-|b|)!}\dfrac{\alpha_a(t)^a}{a!}\dfrac{\alpha_b(t)^b}{b!} & \text{if } |a|+|b|\leq a_0,\\ 0 & \text{otherwise,}\end{cases} \tag{4.65}$$

and

$$\mathcal{M}\left(a,b,b_0,\beta_a(t),\beta_b(t)\right)=\begin{cases}b_0!\dfrac{\left(1-|\beta_a(t)|-|\beta_b(t)|\right)^{b_0-|a|-|b|}}{(b_0-|a|-|b|)!}\dfrac{\beta_a(t)^a}{a!}\dfrac{\beta_b(t)^b}{b!} & \text{if } |a|+|b|\leq b_0,\\ 0 & \text{otherwise.}\end{cases} \tag{4.66}$$

The functions $\alpha_a(t)$, $\alpha_b(t)$, $\beta_a(t)$ and $\beta_b(t)$ are obtained through the IVPs

$$\frac{\mathrm{d}\alpha_a(t)}{\mathrm{d}t}=-k_2\alpha_a(t),\quad \frac{\mathrm{d}\alpha_b(t)}{\mathrm{d}t}=k_2\alpha_a(t)-k_3\alpha_b(t),\quad t>0, \tag{4.67}$$

and

$$\frac{\mathrm{d}\beta_a(t)}{\mathrm{d}t}=-k_2\beta_a(t),\quad \frac{\mathrm{d}\beta_b(t)}{\mathrm{d}t}=k_2\beta_a(t)-k_3\beta_b(t),\quad t>0, \tag{4.68}$$

with initial conditions $\alpha_a(0)=1$, $\alpha_b(0)=0$, $\beta_a(0)=0$, and $\beta_b(0)=1$.

Equation (4.62) represents a direct substitution of the two species mono-molecular chain into the general solution by Jahnke and Huisinga (2007). However, direct point-wise evaluation of this solution is not feasible. Specifically, there are two challenges: (i) the two convolutions are taken over an infinite two-dimensional integer lattice; and (ii) the non-zero probabilities

in the product Poisson and Multinomial distribution can be so small that numerical underflow/overflow is almost certain. The first issue can be solved by determining the finite set of lattice sites that contribute to the convolutions, this can be achieved by invoking specific features of Equation (4.35). The second issue requires that we perform calculations using logarithms of probabilities rather than the true probabilities. Extra care must be taken in the convolution summations.

We first simplify the convolution operations to ensure finite computations. Solving the IVPs 4.64,4.67, and 4.68 yields, for $k_2 \neq k_3$,

$$\lambda_a(t) = \frac{k_1}{k_2}\left(1 - e^{-k_2 t}\right), \tag{4.69}$$

$$\lambda_b(t) = \frac{k_1}{k_3} + \frac{k_1}{k_3 - k_2}\left[e^{-k_2 t} + (k_2 - 2k_3)e^{-k_3 t}\right], \tag{4.70}$$

$$\alpha_a(t) = e^{-k_2 t}, \tag{4.71}$$

$$\alpha_b(t) = \frac{k_2}{k_3 - k_2}\left(e^{-k_2 t} - e^{-k_3 t}\right), \tag{4.72}$$

$$\beta_a(t) = 0, \tag{4.73}$$

$$\beta_b(t) = e^{-k_3 t}. \tag{4.74}$$

A key result is that $\beta_a(t)$ is zero for all time (Equation (4.73)). Through substitution of Equation (4.73) into Equation (4.66), we have

$$\mathcal{M}\left(a, b, b_0, 0, \beta_b(t)\right) = \begin{cases} b_0! \dfrac{\left(1 - |\beta_b(t)|\right)^{b_0 - |a| - |b|}}{(b_0 - |a| - |b|)!} \dfrac{0^a}{a!} \dfrac{\beta_b(t)^b}{b!} & \text{if } |a| + |b| \leq b_0, \\ 0 & \text{otherwise.} \end{cases} \tag{4.75}$$

This implies that $\mathcal{M}\left(a, b, b_0, \beta_a(t), \beta_b(t)\right) = 0$ if $a \neq 0$. That is,

$$\mathcal{M}\left(a, b, b_0, 0, \beta_b(t)\right) = \begin{cases} b_0! \dfrac{\left(1 - |\beta_b(t)|\right)^{b_0 - |b|}}{(b_0 - |b|)!} \dfrac{\beta_b(t)^b}{b!} & \text{if } a = 0, \text{ and } |b| \leq b_0, \\ 0 & \text{otherwise.} \end{cases} \tag{4.76}$$

We can now make a significant simplification of the second convolution in Equation (4.62). Let

$\mathcal{M}_a(a,b,t) = \mathcal{M}\left(a,b,a_0,\alpha_a(t),\alpha_b(t)\right)$ and $\mathcal{M}_b(a,b,t) = \mathcal{M}\left(a,b,b_0,0,\beta_b(t)\right)$, we have

$$\begin{aligned}\mathcal{M}_a(a,b,t) * \mathcal{M}_b(a,b,t) &= \sum_{a_w\in\mathbb{N}}\sum_{b_w\in\mathbb{N}} \mathcal{M}_a(a_w,b_w,t)\mathcal{M}_b(a-a_w,b-b_w,t)\\ &= \sum_{b_w\in\mathbb{N}} \mathcal{M}_a(a,b_w,t)\mathcal{M}_b(0,b-b_w,t),\end{aligned}$$

where $\mathbb{N} = \mathbb{Z}^+ \cup \{0\}$. By Equation (4.76), $\mathcal{M}_b(0,b,t) = 0$ if $|b| > b_0$. Furthermore, we have $b \geq 0$ from the definition of the BCRN (Equation (4.35)). It follows that only terms with $b \geq b_w \geq \max(0, b-b_0)$ can contribute to the convolution, that is,

$$\mathcal{M}_a(a,b,t) * \mathcal{M}_b(a,b,t) = \sum_{b_w=\max(0,b-b_0)}^{b} \mathcal{M}_a(a,b_w,t)\mathcal{M}_b(0,b-b_w,t).$$

While this convolution never involves more that b terms, we can apply a further constraint on the upper bound of the index. By Equation (4.65) we have $\mathcal{M}_a(a,b,t) = 0$ if $|a| + |b| > a_0$. Since $a \geq 0$ and $b \geq 0$ from the definition of the BCRN (Equation (4.35). That is, terms with $b_w \geq a_0 - a \geq 0$ will not contribute to the convolution. Therefore, the multinomial convolution term in Equation (4.62) is

$$\mathcal{M}_a(a,b,t) * \mathcal{M}_b(a,b,t) = \sum_{b_w=\max(0,b-b_0)}^{\min(b,\max(0,a_0-a))} \mathcal{M}_a(a,b_w,t)\mathcal{M}_b(0,b-b_w,t). \tag{4.77}$$

Let $\mathcal{P}(a,b,t) = \mathcal{P}\left(a,b,\lambda_a(t),\lambda_b(t)\right)$ and substitute Equation (4.77) into Equation (4.62) to yield

$$\begin{aligned}P(a,b,t \mid a_0,b_0) &= \mathcal{P}\left(a,b,t\right) * \left[\sum_{b_w=\max(0,b-b_0)}^{\min(b,\max(0,a_0-a))} \mathcal{M}_a(a,b_w,t)\mathcal{M}_b(0,b-b_w,t)\right]\\ &= \sum_{a_z\in\mathbb{N}}\sum_{b_z\in\mathbb{N}} \mathcal{P}\left(a-a_z,b-b_z,t\right)\left[\sum_{b_w=\max(0,b_z-b_0)}^{\min(b_z,\max(0,a_0-a_z))} \mathcal{M}_a(a_z,b_w,t)\mathcal{M}_b(0,b_z-b_w,t)\right].\end{aligned}$$

By definition of the product Poisson distribution (Equation (4.63)), $\mathcal{P}\left(a,b,t\right) = 0$ for $a < 0$ or $b < 0$. Hence, only terms with $a \geq a_z$ and $b \geq b_z$ contribute to the convolution. Therefore, we

obtain the following expression for Equation (4.62)

$$P(a,b,t \mid a_0,b_0) = \sum_{a_z=0}^{a}\sum_{b_z=0}^{b} \mathcal{P}\left(a-a_z,b-b_z,t\right) \times \left[\sum_{b_w=\max(0,b_z-b_0)}^{\min(b_z,\max(0,a_0-a_z))} \mathcal{M}_a(a_z,b_w,t)\mathcal{M}_b(0,b_z-b_w,t)\right], \tag{4.78}$$

which requires $\mathcal{O}(ab^2)$ evaluations of either Equation (4.63), Equation (4.65) or Equation (4.76).

Now that we have bounded the number of operations required to evaluate the solution of the CME, we now address the problem of numerical overflow/underflow. There are two possible sources for this type of numerical error. Firstly, the factorials and products of powers involved in the evaluation of Equation (4.63), Equation (4.65), and Equation (4.76) can be very large, causing overflow. Secondly, the probabilities in the convolution terms can be very small, causing underflow.

To avoid these issues we work with logarithms of probabilities. For the non-zero cases of Equations (4.63), (4.65) and (4.76), we have

$$\ln \mathcal{P}\left(a,b,t\right) = a\ln\lambda_a(t) + b\ln\lambda_b(t) - \left(|\lambda_a(t)| + |\lambda_b(t)|\right) - \sum_{a_i=1}^{a}\ln a_i - \sum_{b_i=1}^{b}\ln b_i, \tag{4.79}$$

$$\ln \mathcal{M}_a\left(a,b,t\right) = a\ln\alpha_a(t) + b\ln\alpha_b(t) + (a_0-a-b)\ln\left(1-\alpha_a(t)-\alpha_b(t)\right) + \sum_{a_i=a_0-a-b}^{a_0}\ln a_i - \sum_{a_i=1}^{a}\ln a_i - \sum_{b_i=1}^{b}\ln b_i, \tag{4.80}$$

$$\ln \mathcal{M}_b\left(a,b,t\right) = b\ln\beta_b(t) + (b_0-b)\ln\left(1-\beta_b(t)\right) + \sum_{b_i=b_0-b}^{b_0}\ln b_i - \sum_{b_i=1}^{b}\ln b_i. \tag{4.81}$$

Equations (4.79)–(4.81) enable the computation to proceed with overflow or underflow being significantly less likely. Therefore, we take the logarithm of Equation (4.78) to obtain

$$\ln P(a,b,t \mid a_0,b_0) = \ln\left[\sum_{a_z=0}^{a}\sum_{b_z=0}^{b} e^{\ln\mathcal{P}(a-a_z,b-b_z,t)+\ln\mathcal{F}(a_z,b_z,t)}\right], \tag{4.82}$$

where

$$\ln\mathcal{F}(a_z,b_z,t) = \ln\left[\sum_{b_w=\max(0,b_z-b_0)}^{\min(b_z,\max(0,a_0-a_z))} e^{\ln\mathcal{M}_a(a_z,b_w,t)+\ln\mathcal{M}_b(0,b_z-b_w,t)}\right]. \tag{4.83}$$

Computing the logarithms of summations of exponential functions in Equation (4.82) and Equation (4.83) is still prone to overflow and underflow since the probabilities will be very small in practice. A common solution to numerically stable logarithm of summations of exponential functions is known as the "log-sum-exp trick". This works by noting, for any $x, y \in \mathbb{R}$, that

$$\begin{aligned}\ln\left[e^{x}+e^{y}\right] &= \ln\left[\left(e^{x-\max(x,y)}+e^{y-\max(x,y)}\right)e^{\max(x,y)}\right] \\ &= \ln\left[e^{x-\max(x,y)}+e^{y-\max(x,y)}\right]+\max(x,y).\end{aligned}$$

Thus, computations are re-scaled to the natural scale of the terms in the summation, thus terms that do underflow would not have affected the result significantly. Now, let

$$R(a_z,b_z) = \max_{b_w\in[\max(0,b_z-b_0),\min(b_z,\max(0,a_0-a_z))]} \left\{\ln\mathcal{M}_a(a_z,b_w,t)+\ln\mathcal{M}_b(0,b_z-b_w,t)\right\},$$

and

$$S(a,b) = \max_{[a_z,b_z]\in[0,a]\times[0,b]} \left\{\ln\mathcal{P}\left(a-a_z,b-b_z,t\right)+\ln\mathcal{F}\left(a_z,b_z,t\right)\right\}.$$

Then use $S(a,b)$ and $R(a_z,b_z)$ with the "log-sum-exp trick" to yield a numerically robust form of Equation (4.82). That is,

$$\ln P(a,b,t \mid a_0,b_0) = \ln\left[\sum_{a_z=0}^{a}\sum_{b_z=0}^{b} e^{\ln\mathcal{P}(a-a_z,b-b_z,t)+\ln\mathcal{F}(a_z,b_z,t)-S(a,b)}\right]+S(a,b), \tag{4.84}$$

where

$$\ln\mathcal{F}(a_z,b_z,t) = \ln\left[\sum_{b_w=\max(0,b_z-b_0)}^{\min(b_z,\max(0,a_0-a_z))} e^{\ln\mathcal{M}_a(a_z,b_w,t)+\ln\mathcal{M}_b(0,b_z-b_w,t)-R(a_z,b_z)}\right]+R(a_z,b_z). \tag{4.85}$$

The example code, `CMEsolMonoMol.m` provides a numerical implementation of the CME solution (Equation (4.62)) using Equation (4.84) and Equation (4.85).

4.6.3 Synthetic data

The synthetic data used in the main manuscript and example code is provide in Table 4.3 for the mono-molecular chain model and in Table 4.4 for the enzyme kinetics model.

Table 4.3: Data, $\mathbf{Y}_{\text{obs}}$, used for inference on the mono-molecular chain. Generated using true parameter values $k_1 = 1.0$, $k_2 = 0.1$, and $k_3 = 0.05$ and initial conditions, $A(0) = 100$ and $B(0) = 0$.

	$\mathbf{Y}(t_1)$	$\mathbf{Y}(t_2)$	$\mathbf{Y}(t_3)$	$\mathbf{Y}(t_4)$
t	25	50	75	100
$A(t)$	14	12	17	15
$B(t)$	68	34	14	14

Table 4.4: Data, $\mathbf{Y}_{\text{obs}}$, used for inference on the enzyme kinetic model. Generated using true parameter values $k_1 = 0.001$, $k_2 = 0.005$, and $k_3 = 0.01$ and initial conditions $S(0) = 100$, $E(0) = 100$, $C(0) = 0$ and $P(0) = 0$.

	$\mathbf{Y}(t_1)$	$\mathbf{Y}(t_2)$	$\mathbf{Y}(t_3)$	$\mathbf{Y}(t_4)$	$\mathbf{Y}(t_5)$
t	0	20	40	60	80
$P(t)$	0	5	16	28	39
$P(t) + \xi$	2.04	6.99	14.30	28.71	38.14

4.6.4 Additional ABC results

In the main manuscript only marginal probability densities are used to demonstrate ABC convergence. Here we present plot matrices with the bivariate marginals also.

Since the mono-molecular chain model has three rate parameters, we have three univariate marginals posteriors and three bivariate marginal posteriors. Through application of the ABC with acceptance threshold, ϵ, the equivalent marginals are

$$p(k_1 \mid \rho(\mathbf{Y}_{\text{obs}}, \mathbf{S}_{\text{obs}}) \leq \epsilon) = \iint_{\mathbb{R}^2} p(k_1, k_2, k_3 \mid \rho(\mathbf{Y}_{\text{obs}}, \mathbf{S}_{\text{obs}}) \leq \epsilon)\, \mathrm{d}k_2\, \mathrm{d}k_3, \tag{4.86}$$

$$p(k_2 \mid \rho(\mathbf{Y}_{\text{obs}}, \mathbf{S}_{\text{obs}}) \leq \epsilon) = \iint_{\mathbb{R}^2} p(k_1, k_2, k_3 \mid \rho(\mathbf{Y}_{\text{obs}}, \mathbf{S}_{\text{obs}}) \leq \epsilon)\, \mathrm{d}k_1\, \mathrm{d}k_3, \tag{4.87}$$

$$p(k_3 \mid \rho(\mathbf{Y}_{\text{obs}}, \mathbf{S}_{\text{obs}}) \leq \epsilon) = \iint_{\mathbb{R}^2} p(k_1, k_2, k_3 \mid \rho(\mathbf{Y}_{\text{obs}}, \mathbf{S}_{\text{obs}}) \leq \epsilon)\, \mathrm{d}k_1\, \mathrm{d}k_2, \tag{4.88}$$

$$p(k_1, k_2 \mid \rho(\mathbf{Y}_{\text{obs}}, \mathbf{S}_{\text{obs}}) \leq \epsilon) = \int_{\mathbb{R}} p(k_1, k_2, k_3 \mid \rho(\mathbf{Y}_{\text{obs}}, \mathbf{S}_{\text{obs}}) \leq \epsilon)\, \mathrm{d}k_3, \tag{4.89}$$

$$p(k_1, k_3 \mid \rho(\mathbf{Y}_{\text{obs}}, \mathbf{S}_{\text{obs}}) \leq \epsilon) = \int_{\mathbb{R}} p(k_1, k_2, k_3 \mid \rho(\mathbf{Y}_{\text{obs}}, \mathbf{S}_{\text{obs}}) \leq \epsilon)\, \mathrm{d}k_2, \tag{4.90}$$

$$p(k_2, k_3 \mid \rho(\mathbf{Y}_{\text{obs}}, \mathbf{S}_{\text{obs}}) \leq \epsilon) = \int_{\mathbb{R}} p(k_1, k_2, k_3 \mid \rho(\mathbf{Y}_{\text{obs}}, \mathbf{S}_{\text{obs}}) \leq \epsilon)\, \mathrm{d}k_1. \tag{4.91}$$

The exact univariate and bivariate marginal posteriors are plotted against the ABC posterior for $\epsilon = [50, 25, 12.5, 0]$ (with $\epsilon = 0$ meaning the exact posterior is sampled using the CME-based

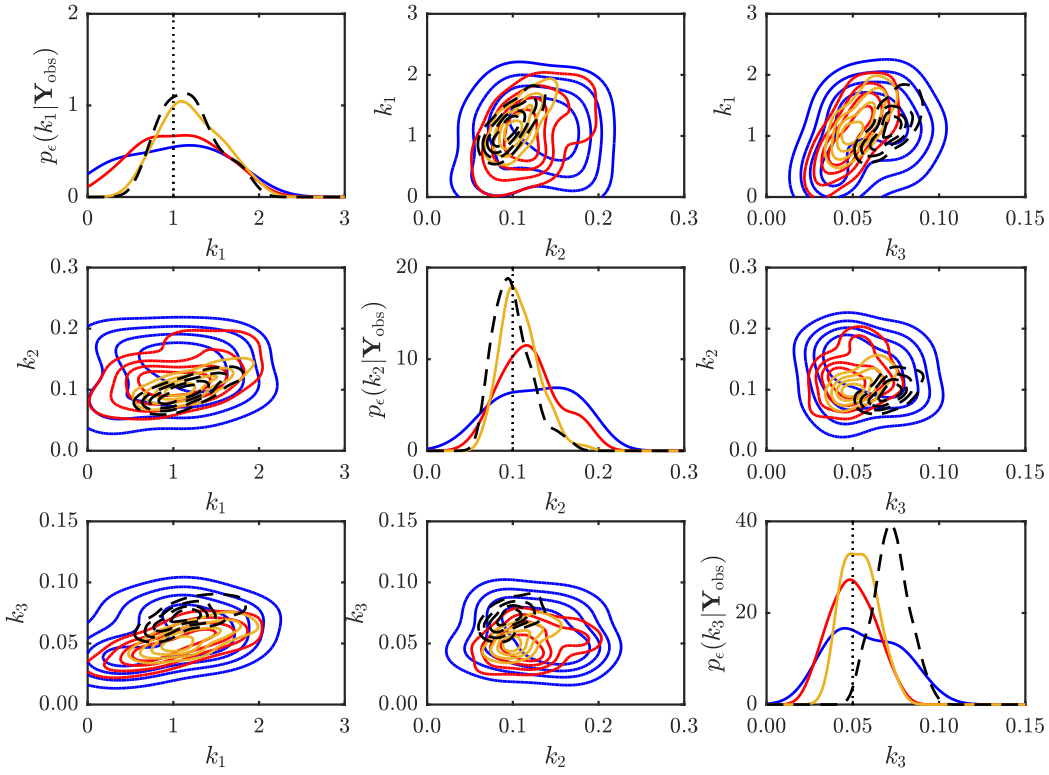

Figure 4.7: Convergence of ABC posterior to the true posterior as $\epsilon \to 0$ for the mono-molecular chain inference problem. Marginal posteriors are plotted for $\epsilon = 50$ (blue solid), $\epsilon = 25$ (red solid), $\epsilon = 12.5$ (yellow solid), and $\epsilon = 0$ (black dashed). Here, the $\epsilon = 0$ case corresponds to the exact likelihoods using the CME solution. Univariate marginals are plotted on the diagonals and bivariate marginals on off diagonal elements. Contour lines in bivariate marginal plots are selected such that six equal probability density intervals are shown. The true parameter values (black dotted) are $k_1 = 1.0$, $k_2 = 0.1$ and $k_3 = 0.05$. Note that the exact Bayesian posterior does not recover the true parameter for k_3.

likelihood). Equations (4.86)–(4.91) are plotted in Figure 4.7.

Reducing ϵ further than 12.5 is prohibitive, even for the mono-molecular chain model. Both Barber et al. (2015) and Fearnhead and Prangle (2012) provide an asymptotic result for the computation time, $\mathcal{C}$, as a function of ϵ, that is, $\mathcal{C} = \mathcal{O}(\epsilon^{-d})$, where d is the dimensionality of the data used in the ABC inference. For the synthetic data we have from Table 4.3, we have $d = n_t N$. Figure 4.8 demonstrates that the computation times we obtain are consistent with this.

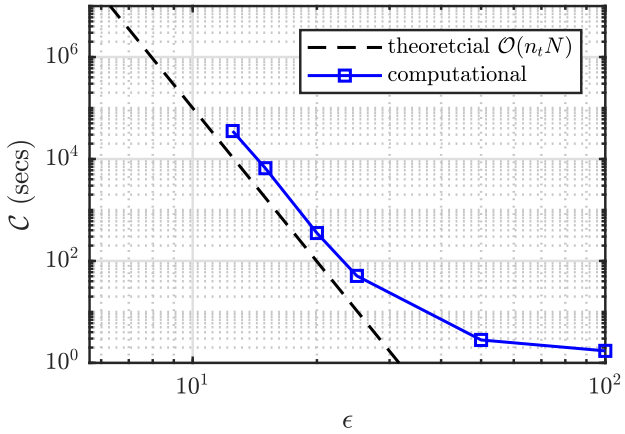

Figure 4.8: Computation time growth of ABC rejection sampling against theoretical result. Computations are performed using an Intel® Core™ i7-5600U CPU (2.6 GHz).

4.6.5 ABC Multilevel Monte Carlo

Here we provide the ABC Multilevel Monte Carlo scheme (ABCMLMC) (see Chapter 6 for more details and the derivation). This particular implementation computes an estimate of the posterior mean. Given a sequence of acceptance thresholds, $\epsilon_0 > \epsilon_1 > \cdots > \epsilon_L = \epsilon$, and a sequence of sample numbers $m_0 > m_1 > \cdots > m_L$ (see Giles (2008) and Chapter 6 for details on optimally computing the sample numbers), ABCMLMC proceeds as follows:

1. initialise $\ell = 0$;
2. set $i = 1$;
3. generate a prior sample $\boldsymbol{\theta}^* \sim p(\boldsymbol{\theta})$;
4. generate simulated data, $\mathbf{S}^*_{\text{obs}} \sim s(\mathbf{S}_{\text{obs}}; \boldsymbol{\theta}^*)$;
5. if $\rho(\mathbf{Y}_{\text{obs}}, \mathbf{S}^*_{\text{obs}}) \leq \epsilon_\ell$, accept $\boldsymbol{\theta}^{(i)}_{\epsilon_\ell} = \boldsymbol{\theta}^*$ and set $i = i + 1$, otherwise continue;
6. if $i \leq m_\ell$, go to step 3, otherwise continue;
7. set $\hat{F}_{\ell,j}(z) = \sum_{i=1}^{m_\ell} \mathbb{1}_z \left(k^{(i)}_{\epsilon_\ell,j} \right) / m_\ell$ for $j = 1, 2, \ldots, M$;
8. if $\ell = 0$, then set $\hat{\boldsymbol{\theta}}_\epsilon = \sum_{i=1}^{m_\ell} \boldsymbol{\theta}^{(i)}_{\epsilon_\ell} / m_\ell$, set $\ell = \ell + 1$, and go to step 2, otherwise continue;
9. set $i = 1$;
10. set $k^{(i)}_{\epsilon_{\ell-1},j} = \hat{F}^{-1}_{\ell-1,j}(\hat{F}_{\ell,j}(k^{(i)}_{\epsilon_\ell,j}))$ for $j = 1, 2, \ldots, M$;

11. set $\boldsymbol{\theta}^{(i)}_{\epsilon_{\ell-1}} = [k^{(i)}_{\epsilon_{\ell-1},1}, k^{(i)}_{\epsilon_{\ell-1},2}, \ldots, k^{(i)}_{\epsilon_{\ell-1},M}]$ and set $i = i + 1$;

12. if $i \leq m_\ell$, then go to step 9, otherwise continue;

13. set $\hat{F}_{\ell,j}(z) = \hat{F}_{\ell-1,j}(z) + \sum_{i=1}^{m_\ell} \left(\mathbb{1}_z \left(k^{(i)}_{\epsilon_\ell,j} \right) - \mathbb{1}_z \left(k^{(i)}_{\epsilon_{\ell-1},j} \right) \right) / m_\ell$;

14. set $\hat{\boldsymbol{\theta}}_\epsilon = \hat{\boldsymbol{\theta}}_\epsilon + \sum_{i=1}^{m_\ell} \left(\boldsymbol{\theta}^{(i)}_{\epsilon_\ell} - \boldsymbol{\theta}^{(i)}_{\epsilon_{\ell-j}} \right) / m_\ell$;

15. if $\ell = L$, then terminate, otherwise set $\ell = \ell + 1$ and go to step 2;

4.6.6 ABC algorithm configurations and additional results

The following algorithm configurations for the ABC rejection sampler, ABCMCMC, ABCSMC and ABCMLMC are used to generate the results in Tables 3 and 4 of the main manuscript. The parameters are also contained in the code examples, `DemoABCMethodsMonoMol.m` and `DemoABCMethodsMichMent.m`.

For the mono-molecular chain model inference problem each algorithm is configured as follows: for the ABC rejection sampler we set $m = 100$ and $\epsilon = 15$; for ABCMCMC we set $m_n = 500,000$, $m_b = 100,000$, $m_h = 10,000$, the proposal kernel is a Gaussian random walk with covariance matrix, $\Sigma = \mathrm{diag}(1 \times 10^{-3}, 1 \times 10^{-5}, 2.5 \times 10^{-5})$, and $\epsilon = 15$; for ABCSMC we use $m_p = 100$, the proposal kernel is a Gaussian random walk with covariance matrix, $\Sigma = \mathrm{diag}(1 \times 10^{-3}, 1 \times 10^{-5}, 2.5 \times 10^{-5})$, and the discrepancy threshold sequence is $\epsilon_1 = 100$ with $\epsilon_{r+1} = \epsilon_r/2$ for $r = 2, 3, \ldots, 5$; for ABCMLMC we use the discrepancy threshold sequence $\epsilon_0 = 100$ with $\epsilon_{\ell+1} = \epsilon_\ell/2$ for $\ell = 1, 2, \ldots, 4$, and the sample number sequence, $m_0 = 800$, with $m_{\ell+1} = M_\ell/2$ for $\ell = 1, 2, \ldots, 4$. For the prior, we assume all parameters are independent of each other and uniformly distributed with $k_1 \sim \mathcal{U}(0, 2)$, $k_2 \sim \mathcal{U}(0, 0.2)$, and $k_3 \sim \mathcal{U}(0, 0.1)$.

Similarly for the enzyme kinetics model inference problem each algorithm is configured as follows: for the ABC rejection sampler we set $m = 100$ and $\epsilon = 2.5$; for ABCMCMC we set $m_n = 500,000$, $m_b = 100,000$, $m_h = 10,000$, the proposal kernel is a Gaussian random walk with covariance matrix, $\Sigma = \mathrm{diag}(2.25 \times 10^{-8}, 5.625 \times 10^{-7}, 6.25 \times 10^{-6})$, and $\epsilon = 2.5$; for ABCSMC we use $m_p = 100$, the proposal kernel is a Gaussian random walk with covariance matrix, $\Sigma = \mathrm{diag}(2.25 \times 10^{-8}, 5.625 \times 10^{-7}, 6.25 \times 10^{-6})$, and the discrepancy

threshold sequence is $\epsilon_1 = 40$ with $\epsilon_{r+1} = \epsilon_r/2$ for $r = 2, 3, \ldots, 5$; for ABCMLMC we use the discrepancy threshold sequence $\epsilon_0 = 40$ with $\epsilon_{\ell+1} = \epsilon_\ell/2$ for $\ell = 1, 2, \ldots, 4$, and the sample number sequence, $m_0 = 800$, with $m_{\ell+1} = M_\ell/2$ for $\ell = 1, 2, \ldots, 4$. For the prior, we assume all parameters are independent of each other and uniformly distributed with $k_1 \sim \mathcal{U}(0, 0.003)$, $k_2 \sim \mathcal{U}(0, 0.015)$, and $k_3 \sim \mathcal{U}(0, 0.05)$.

The resulting marginal posterior distributions are presented in Figure 4.9. ABCSMC and ABCMLMC recover the true parameters effectively and as less computationally intensive. For the enzyme kinetic inference problem, more tuning and samples are required to obtain

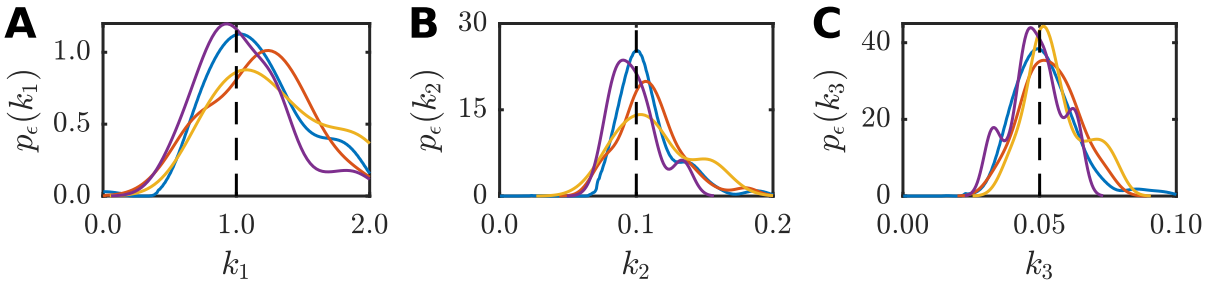

Figure 4.9: Comparison of ABC posteriors generated by the ABC rejection sampler (red solid), ABCMCMC (yellow solid), ABCSMC (purple solid) and ABCMLMC (blue solid) for the mono-molecular chain inference problem. The true parameter values (black dashed) are $k_1 = 1.0$, $k_2 = 0.1$ and $k_3 = 0.05$.

good estimates of the full marginal posterior distributions. Especially, since the ABCMCMC trajectory undergoes a long excursion into the low density tails.

Chapter 5

A Practical Guide to Pseudo-Marginal Methods for Computational Inference in Systems Biology

A paper submitted to *Journal of Theoretical Biology*.

David J. Warne, Ruth E. Baker and Matthew J. Simpson (2019). A practical guide to pseudo-marginal methods for computational inference in systems biology. *ArXiv e-prints*, arXiv:1912.12404 [q-bio.MN]

Abstract For many stochastic models of interest in systems biology, such as those describing biochemical reaction networks, exact quantification of parameter uncertainty through statistical inference is intractable. Likelihood-free computational inference techniques enable parameter inference when the likelihood function for the model is intractable but the generation of many sample paths is feasible through stochastic simulation of the forward problem. The most common likelihood-free method in systems biology is approximate Bayesian computation that accepts parameters that result in low discrepancy between stochastic simulations and measured data. However, it can be difficult to assess how the accuracy of the resulting inferences are affected by the choice of acceptance threshold and discrepancy function. The pseudo-marginal

approach is an alternative likelihood-free inference method that utilises a Monte Carlo estimate of the likelihood function. This approach has several advantages, particularly in the context of noisy, partially observed, time-course data typical in biochemical reaction network studies. Specifically, the pseudo-marginal approach facilitates exact inference and uncertainty quantification, and may be efficiently combined with particle filters for low variance, high-accuracy likelihood estimation. In this review, we provide a practical introduction to the pseudo-marginal approach using inference for biochemical reaction networks as a series of case studies. Implementations of key algorithms and examples are provided using the Julia programming language; a high performance, open source programming language for scientific computing (https://github.com/davidwarne/Warne2019_GuideToPseudoMarginal).

5.1 Introduction

Stochastic models are routinely used in systems biology to facilitate the interpretation and understanding of experimental observations. In particular, stochastic models are often more realistic descriptions, compared with their deterministic counterparts, of many biochemical processes that are naturally affected by extrinsic and intrinsic noise (Kærn et al., 2005; Raj and van Oudenaarden, 2008), such as the biochemical reaction pathways that regulate gene expression (Paulsson et al., 2000; Tian and Burrage, 2006). Such stochastic models enable the exploration of various biochemical network motifs to explain particular phenomena observed through the use of modern, high resolution experimental techniques (Sahl et al., 2017). The validation and comparison of theories against observations can be achieved using statistical inference techniques to quantify the uncertainty in unknown parameters and likelihoods of observations under different models. Recent reviews by Schnoerr et al. (2017) and Chapter 4 highlight the state-of-the-art in computational techniques for simulation of biochemical networks, analysis of the distribution of future states of the biochemical systems, and computational inference from a Bayesian perspective. Both studies point out that, for realistic biochemical reaction networks, the likelihood function is intractable. As a result, likelihood-free computation inference schemes are essential for practical situations.

In Chapter 4, we provide an accessible discussion of a wide range of algorithms for simulation and inference in the context of biochemical systems and provide example implementations for

demonstration purposes. In particular, Chapter 4 highlights the use of approximate Bayesian computation (ABC) (Sisson et al., 2018) for likelihood-free inference of kinetic rate parameters using time-course data. While ABC is a widely applicable and popular likelihood-free approach within the life sciences (Toni et al., 2009), inferences obtained by this method are, as the name implies, approximations, and the accuracy of these approximations are highly dependent upon choices made by the user (Sunnåker et al., 2013).

Time-course data describing temporal variations in particular molecular signals within living cells are often obtained using time-lapse optical microscopy with fluorescent reporters (Figure 5.1(A)) (Bar-Joseph et al., 2012; Locke and Elowitz, 2009; Young et al., 2012). Individual cells are tracked (Figure 5.1(B)) and the luminescence from the reporter is measured over time at discrete intervals (Figure 5.1(C)). These luminescence values are then used to determine concentrations of mRNAs or proteins that may be associated with the expression of a particular gene over time. These data provide information about the dynamics complex gene regulatory networks that can result in stochastic switching (Tian and Burrage, 2006) or oscillatory behaviour (Figure 5.1(C)) (Elowitz and Leibler, 2000; Shimojo et al., 2008).

Common features of gene expression time-course data include sparsity of temporal observations, relatively few concurrent fluorescent reporters, and noisy observations. Therefore, likelihood-free inference methods are essential to deal with statistical inference (Toni et al., 2009). However, complex dynamics observed in real gene regulatory networks, such as stochastic oscillations or bi-stability, can render ABC methods impractical for accurate inferences since most simulations will be rejected, even when the values model parameters are close to the true values.

Pseudo-marginal methods (Andrieu and Roberts, 2009) are an alternative likelihood-free approach that can provide exact inferences under the prescribed model and are significantly less sensitive to user-defined input. Variants of this approach are particularly well suited for Bayesian inference of nonlinear stochastic models using partially observed time-course data (Andrieu et al., 2010). This makes the pseudo-marginal method ideal for the study of biochemical systems, however, to-date, few applications of these approaches are present in the systems biology literature (Golightly and Wilkinson, 2008, 2011). While Chapter 4 briefly discusses the pseudo-marginal approach, no examples or implementations are provided.

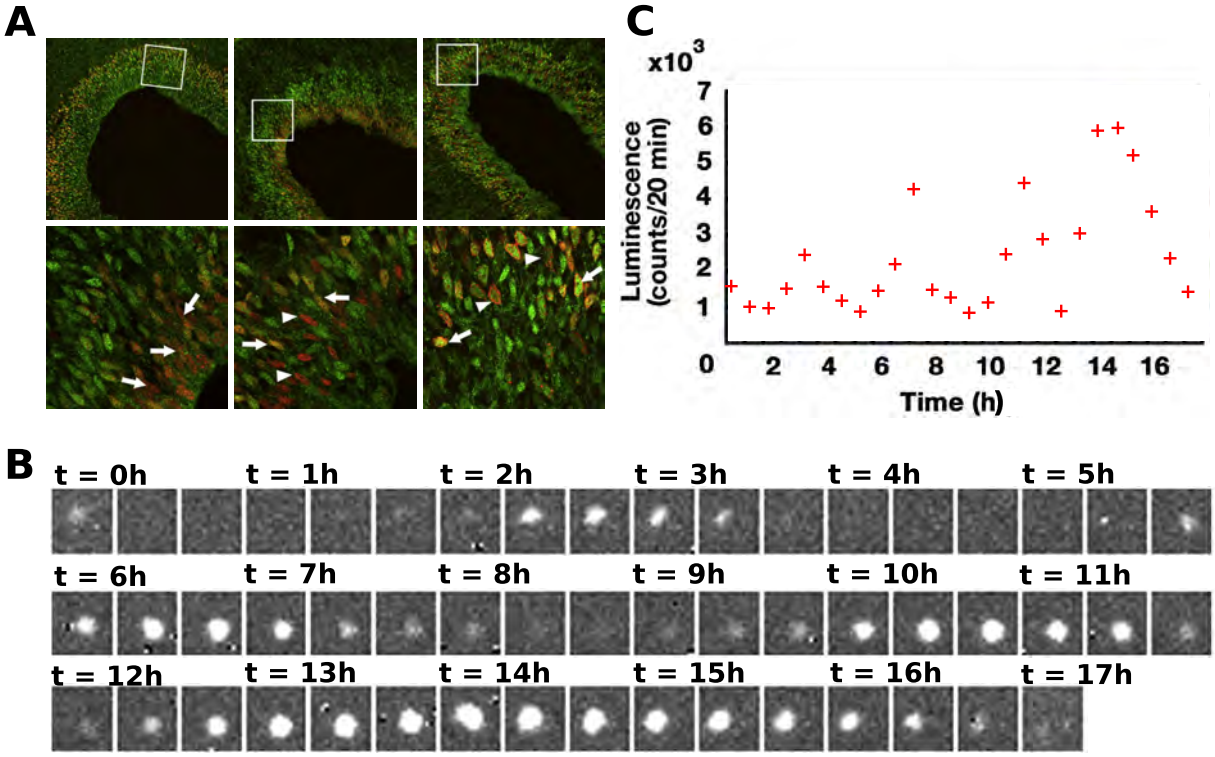

Figure 5.1: Time-course gene expression data. (A) snapshots of time-lapse microscopy using two fluorescent reporters indicating Hes1 gene expression (green) and BrdU incorporation (red) used to indicate cell-cycle phases. (B) Time-series of green fluorescent reporter for a tracked single-cell over 17 hours in 20 minute intervals. (C) the resulting time-series indicating oscillatory Hes1 gene expression. Panels (A),(B), and (C) are modified with permission from Shimojo et al. (2008).

The purpose of this chapter is to complement Chapter 4 and Schnoerr et al. (2017) by providing an accessible, didactic guide to pseudo-marginal methods (Andrieu and Roberts, 2009; Andrieu et al., 2010; Doucet et al., 2015) for the inference of kinetic rate parameters of biochemical reaction network models using the chemical Langevin description. For all of our examples, we provide accessible implementations using the open source, high performance Julia programming language (Besançon et al., 2019; Bezanson et al., 2017)[1].

5.2 Background

In this section, we introduce several concepts that are fundamental to understanding how pseudo-marginal methods work and why they are effective. Firstly, we introduce stochastic biochemical

[1]The Julia code examples and demonstration scripts are available from GitHub https://github.com/davidwarne/Warne2019_GuideToPseudoMarginal

reaction networks and how one might model and simulate these systems using stochastic differential equations (SDEs). The Bayesian inference framework is then described along with the essentials of Markov chain Monte Carlo (MCMC) sampling. Lastly, an analytically tractable inference problem is presented, along with Julia code implementations, in order to solidify the concepts, as they are relied upon in subsequent sections.

5.2.1 Stochastic biochemical reaction networks

A biochemical reaction network consists of N chemical species, $X_1, X_2, \ldots, X_N$ that interact via a network of M reactions,

$$\sum_{i=1}^{N} \nu_{i,j}^{-} X_i \xrightarrow{k_j} \sum_{i=1}^{N} \nu_{i,j}^{+} X_i, \quad j = 1, 2, \ldots, M, \tag{5.1}$$

where $\nu_{i,j}^{-}$ and $\nu_{i,j}^{+}$ are, respectively, the number of reactant and product molecules of species X_i involved in the jth reaction, and k_j is the kinetic rate parameter for the jth reaction. We refer to $\nu_{i,j} = \nu_{i,j}^{+} - \nu_{i,j}^{-}$ as the stoichiometry of species i for reaction j. While spatially extended systems can be considered (Cotter and Erban, 2016; Flegg et al., 2015), we will assume the chemical mixture is spatially uniform for clarity. Under this assumption the law of mass action holds and the probability of the jth reaction occurring in the time interval $[t, t + \mathrm{d}t)$ is $a_j(\mathbf{X}_t)\mathrm{d}t$, where $\mathbf{X}_t = [X_{1,t}, X_{2,t}, \ldots, X_{N,t}]^{\mathrm{T}}$ is an $N \times 1$ state vector consisting of the copy numbers for each species at time t and $a_j(\mathbf{X}_t)$ is the *propensity function* for reaction j (Gillespie, 1977; Kurtz, 1972). Should a reaction j event occur then the state will update by adding the stoichiometric vector $\boldsymbol{\nu}_j = [\nu_{1,j}, \nu_{2,j}, \ldots, \nu_{N,j}]^{\mathrm{T}}$ to the current system state. Example implementations for generating a range of common biochemical reaction network models is provided in `ChemicalReactionNetworkModels.jl`.

In situations where the number of molecules in the system is sufficiently large, the forwards evolution of the biochemical reaction network can be accurately approximated by the *chemical Langevin equation* (Higham, 2008; Gillespie, 2000; Wilkinson, 2009). The chemical Langevin equation is an Itō SDE of the form

$$\mathrm{d}\mathbf{X}_t = \sum_{j=1}^{M} \boldsymbol{\nu}_j a_j(\mathbf{X}_t)\mathrm{d}t + \sum_{j=1}^{M} \boldsymbol{\nu}_j \sqrt{a_j(\mathbf{X}_t)}\mathrm{d}W_t^{(j)}, \tag{5.2}$$

where $\mathbf{X}_t$ takes values in $\mathbb{R}^N$ and $W_t^{(1)}, W_t^{(2)}, \ldots, W_t^{(M)}$ are independent scalar Wiener processes. For a fixed initial condition, $\mathbf{X}_0$, the solution to Equation (5.2), $\{\mathbf{X}_t\}_{0 \leq t}$, can be approximately simulated using numerical methods. In this work, we apply the Euler-Maruyama scheme (Kloeden and Platen, 1999; Maruyama, 1955) which approximates a realisation at $\mathbf{X}_{t+\Delta t}$ given $\mathbf{X}_t$ according to

$$\mathbf{X}_{t+\Delta t} = \mathbf{X}_t + \sum_{j=1}^{M} \boldsymbol{\nu}_j a_j(\mathbf{X}_t)\Delta t + \sum_{j=1}^{M} \boldsymbol{\nu}_j \sqrt{a_j(\mathbf{X}_t)\Delta t}\xi^{(j)},$$

where $\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(M)}$ are independent, identically distributed (i.i.d.) standard normal random variables. It can be shown that the Euler-Maruyama scheme converges with rate $\mathcal{O}(\sqrt{\Delta t})$ to the true path-wise solution (Kloeden and Platen, 1999). While higher-order schemes are possible, tighter restrictions on the SDE form are required. Therefore we restrict ourselves to Euler-Maruyama in this work. For an accessible introduction to numerical methods for SDEs, see Higham (2001), and for a detailed monologue that includes rigorous analysis of convergence rates, see Kloeden and Platen (1999). Example implementations of the Euler-Maruyama scheme for the chemical Langevin equation are provided in `EulerMaruyama.jl` and `ChemicalLangevin.jl`.

5.2.2 Markov chain Monte Carlo for Bayesian inference

In practice, the application of mathematical models to the study of real biochemical networks requires model calibration and parameter inference using experimental data. The data are typically chemical concentrations derived from optical microscopy and fluorescent reporters such as green fluorescent proteins (Finkenstädt et al., 2008; Sahl et al., 2017; Wilkinson, 2011). Let $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ be the vector of unknown model parameters, such as kinetic rate parameters or initial conditions. The task is to quantify the uncertainty in model parameters after taking the experimental data, $\mathcal{D}$, into account. Given a model parameterised by $\boldsymbol{\theta}$ and experimental data, $\mathcal{D}$, uncertainty of the unknown model parameters can be quantified using the Bayesian *posterior probability density*,

$$p(\boldsymbol{\theta} \mid \mathcal{D}) = \frac{\mathcal{L}(\boldsymbol{\theta}; \mathcal{D})p(\boldsymbol{\theta})}{p(\mathcal{D})}, \tag{5.3}$$

where: $p(\boldsymbol{\theta})$ is the *prior probability density* that encodes parameter assumptions; $\mathcal{L}(\boldsymbol{\theta}; \mathcal{D})$ is the *likelihood function* that determines the probability of the data under the assumed model for

fixed $\boldsymbol{\theta}$; and $p(\mathcal{D})$ is the *evidence* that provides a total probability for the data under the assumed model over all possible parameter values.

Parameter uncertainty quantification often involves computing expectations of functionals with respect to the posterior distribution (Equation (5.3)),

$$\mathbb{E}\left[f(\boldsymbol{\theta})\right] = \int_{\boldsymbol{\Theta}} f(\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathcal{D}) \, \mathrm{d}\boldsymbol{\theta},$$

which may be estimated using Monte Carlo integration,

$$\mathbb{E}\left[f(\boldsymbol{\theta})\right] \approx \hat{f}(\boldsymbol{\theta}) = \frac{1}{\mathcal{M}} \sum_{i=1}^{\mathcal{M}} f(\boldsymbol{\theta}^{(i)}),$$

where $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(\mathcal{M})}$ are i.i.d. samples from the posterior distribution, $p(\boldsymbol{\theta} \mid \mathcal{D})$. In particular, the jth marginal posterior probability density, $p(\theta_j \mid \mathcal{D})$ with θ_j the jth dimension of $\boldsymbol{\theta}$, may be estimated using a smoothed kernel density estimate,

$$p(\theta_j \mid \mathcal{D}) \approx \frac{1}{\mathcal{M}h} \sum_{i=1}^{\mathcal{M}} K\left(\frac{\theta_j - \theta_j^{(i)}}{h}\right),$$

where $\theta_j^{(1)}, \theta_j^{(2)}, \ldots, \theta_j^{(\mathcal{M})}$ are the jth dimensions of i.i.d. posterior samples, h is a user prescribed smoothing parameter, and the kernel $K(x)$ is chosen such that $\int_{-\infty}^{\infty} K(x) \, \mathrm{d}x = 1$. Typically, $K(x)$ is a standard Gaussian, and h is chosen using Silverman's rule (Silverman, 1986). In most cases, direct i.i.d. sampling from the posterior distribution is not possible since it is often not from a standard distribution family.

MCMC methods are based on the idea of simulating a discrete time Markov chain, $\{\boldsymbol{\theta}_m\}_{0 \leq m}$, in parameter space, $\boldsymbol{\Theta}$, for which the posterior of interest is its stationary distribution (Green et al., 2015; Roberts and Rosenthal, 2004). A popular MCMC algorithm is the Metropolis-Hastings method (Hastings, 1970; Metropolis et al., 1953) (Algorithm 5.1, an example implementation is provided in `MetropolisHastings.jl`), in which transitions from state $\boldsymbol{\theta}_m$ to a proposed new state $\boldsymbol{\theta}^*$ occur with probability proportional to the relative posterior density between the two locations. The proposals are determined through sampling a proposal kernel distribution that is conditional on $\boldsymbol{\theta}_m$ with density $q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}_m)$. Under some regularity conditions on the proposal density, the resulting Markov chain will converge to the target posterior as its

Algorithm 5.1 The Metropolis-Hastings method for MCMC

1: Given initial condition $\boldsymbol{\theta}_0$ such that $p(\boldsymbol{\theta}_0 \mid \mathcal{D}) > 0$.
2: **for** $m = 1, \ldots, \mathcal{M}$ **do**
3: Sample transition kernel, $\boldsymbol{\theta}^* \sim q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{m-1})$.
4: Calculate acceptance probability

$$\alpha\left(\boldsymbol{\theta}^*, \boldsymbol{\theta}_{m-1}\right) = \min\left(1, \frac{q(\boldsymbol{\theta}_{m-1} \mid \boldsymbol{\theta}^*)p(\boldsymbol{\theta}^* \mid \mathcal{D})}{q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}_{m-1})p(\boldsymbol{\theta}_{m-1} \mid \mathcal{D})}\right).$$

5: Set $\boldsymbol{\theta}_m \leftarrow \boldsymbol{\theta}^*$ with probability $\alpha\left(\boldsymbol{\theta}^*, \boldsymbol{\theta}_{m-1}\right)$, otherwise, set $\boldsymbol{\theta}_m \leftarrow \boldsymbol{\theta}_{m-1}$.
6: **end for**

stationary distribution (Mengersen and Tweedie, 1996). Therefore, computing expectations can be performed with Monte Carlo integration using a sufficiently large *dependent* sequence from the Metropolis-Hastings Markov chain. It is important to note that this is an asymptotic result, and determining when such a sequence is sufficiently large for practical purposes is an active area of research (Cowles and Carlin, 1996; Gelman and Rubin, 1992; Gelman et al., 2014; Vehtari et al., 2019). It is also important to note that the efficiency of MCMC based on Metropolis-Hastings is heavily dependent on the proposal density used (Metropolis et al., 1953). Adaptive schemes may be applied (Roberts and Rosenthal, 2009), however, care must be taken when applying these schemes as the stationary distribution may be altered. In many practical applications, the proposal density and number of iterations is selected heuristically (Hines et al., 2014).

Alternative MCMC algorithms include Gibbs sampling (Geman and Geman, 1984), Hamiltonian Monte Carlo (Duane et al., 1987), and Zig-Zag sampling (Bierkens et al., 2019). In this work, however, we base all discussion and examples on the Metropolis-Hastings method as it is the most natural to extend to challenging inference problems in systems biology (Golightly and Wilkinson, 2011; Marjoram et al., 2003).

5.2.3 A tractable example: the production-degradation model

We demonstrate the application of MCMC to perform exact Bayesian inference using a biochemical reaction network for which an analytic solution to the likelihood can be obtained. This enables us to highlight important MCMC algorithm design considerations before introducing the additional complexity that arises when the likelihood is intractable.

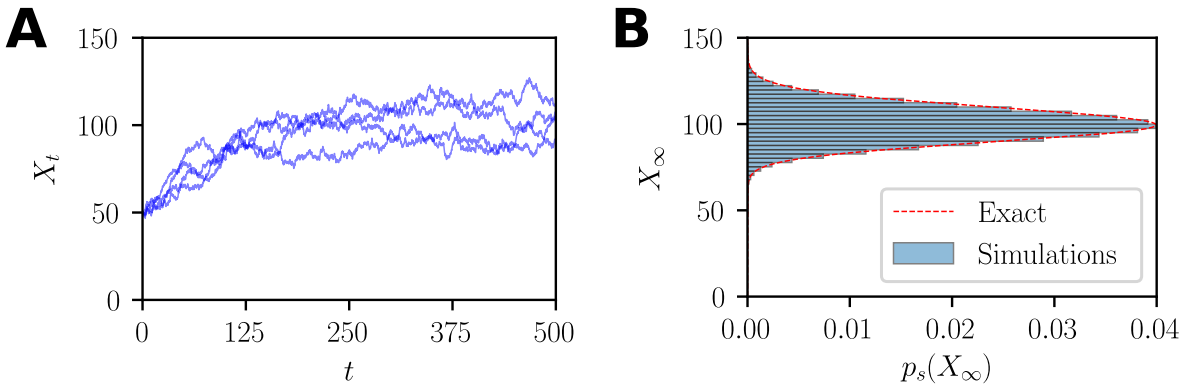

Figure 5.2: (A) Four example realisations of the chemical Langevin SDE for the production-degradation model. (B) The analytic stationary distribution compared with an approximation using simulations. Simulations are performed using the Euler-Maruyama scheme with $\Delta t = 0.1$ and $X_0 = 50$. Kinetic rate parameters are $k_1 = 1.0$ and $k_2 = 0.01$.

Consider a biochemical system consisting a single chemical species, X, involving only production and degradation reactions of the form

$$\underbrace{\emptyset \xrightarrow{k_1} X}_{\substack{\text{external production} \\ \text{of } X \text{ molecules}}} \quad \text{and} \quad \underbrace{X \xrightarrow{k_2} \emptyset}_{\substack{\text{degradation} \\ \text{of } X \text{ molecules}}} . \tag{5.4}$$

Here, $k_1 > 0$ and $k_2 > 0$ are the kinetic parameters for production and degradation, respectively. The propensity functions are given by

$$a_1(X_t) = k_1 \quad \text{and} \quad a_2(X_t) = k_2 X_t,$$

with respective stoichiometries $\nu_1 = 1$ and $\nu_2 = -1$. The chemical Langevin equation for this production-degradation model (Equation (5.4)) is

$$\mathrm{d}X_t = (k_1 - k_2 X_t)\mathrm{d}t + \sqrt{k_1 + k_2 X_t}\mathrm{d}W_t, \tag{5.5}$$

where W_t is a Wiener process. Approximate realisations of the solution process can be generated using the Euler-Maruyama discretisation, as demonstrated in Figure 5.2(A) (see example `DemoProdDeg.jl`). Throughout this work we take time, t, and rate parameters to be dimensionless. However, all results can be re-dimensionalised as appropriate.

Assume that the degradation rate is known, $k_2 = 0.01$. The inference task is to quantify the uncertainty in the production kinetic rate, k_1, using experimental data $\mathcal{D} = \left[Y_{\text{obs}}^{(1)}, Y_{\text{obs}}^{(2)}, \ldots, Y_{\text{obs}}^{(n)}\right]$,

where $Y_{\text{obs}}^{(1)}, Y_{\text{obs}}^{(2)}, \ldots, Y_{\text{obs}}^{(n)}$ are n independent observations of a hypothetical real biochemical production-degradation process that has reached its equilibrium distribution (Section 5.7.1). For simplicity, we also assume these observations are not subject to any observation error, that is, our data is assumed to be exact realisations of the stationary process for the production-degradation model (Equation (5.4)) under the chemical Langevin equation representation (Equation (5.5)).

For inference, we require the Bayesian posterior probability density,

$$p(k_1 \mid \mathcal{D}) \propto \mathcal{L}(k_1; \mathcal{D}) p(k_1), \tag{5.6}$$

where the prior is $p(k_1)$ and the likelihood is

$$\mathcal{L}(k_1; \mathcal{D}) = \prod_{i=1}^{n} p_s\left(Y_{\text{obs}}^{(i)}; k_1\right). \tag{5.7}$$

We prescribe a uniform prior, $k_1 \sim \mathcal{U}(0, 2)$, that contains the true parameter value of $k_1 = 1.0$. In Equation (5.7), $p_s\left(x; k_1\right)$ is the probability density function for the chemical Langevin equation (Equation (5.5)) solution process, $\{X_t\}_{0 \leq t}$, as $t \to \infty$, that is, the stationary process $X_\infty \sim p_s(x; k_1)$. For this particular example, it is possible to obtain an analytical expression for this stationary probability density function. The solution is obtained by formulating the Fokker-Planck equation for the Itō process in Equation (5.5) and solving for the steady state (Section 5.7.1) to yield

$$p_s\left(x; k_1\right) = \frac{\exp\left(-2x + \left(\frac{4k_1}{k_2} - 1\right) \ln\left(k_1 + k_2 x\right)\right)}{\displaystyle\int_0^\infty \exp\left(-2y + \left(\frac{4k_1}{k_2} - 1\right) \ln\left(k_1 + k_2 y\right)\right) \mathrm{d}y}. \tag{5.8}$$

Given a value for k_1, then the denominator can be accurately calculated using quadrature. Figure 5.2(B) overlays this analytical solution against a histogram obtained from the time series of a single very long simulation with end time, $t = 1,000,000$.

Using Equation (5.8), we can now evaluate the likelihood function (Equation (5.7)) point-wise, and hence the posterior density (Equation (5.6)) can be evaluated point-wise up to a normalising constant. Therefore, we can apply the Metropolis-Hastings method for which the acceptance

probability is

$$\alpha\left(\theta^*, \theta\right) = \min\left(1, \frac{q(\theta \mid \theta^*)p(\theta^*)\prod_{i=1}^{n} p_s\left(Y_{\text{obs}}^{(i)}; \theta^*\right)}{q(\theta^* \mid \theta)p(\theta)\prod_{i=1}^{n} p_s\left(Y_{\text{obs}}^{(i)}; \theta\right)}\right), \tag{5.9}$$

where $\theta^* \sim q(\cdot \mid \theta)$ is the proposal mechanism.

The choice of proposal kernel dramatically affects the rate of convergence of the Markov chain. For example, Figure 5.3 demonstrates the Markov chain based on Equation (5.9) using a Gaussian proposal kernel,

$$q(\theta^* \mid \theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\theta^* - \theta)^2}{2\sigma^2}\right), \tag{5.10}$$

for different choices of the standard deviation parameter σ (see example `DemoMH.jl`). In all cases, the initial state of the chain is set in a region of very low posterior density, $\theta_0 = 0.8$, to ensure we can compare transient and stationary behaviour of the Markov chain. For small standard deviation, $\sigma = 0.01$ (Figure 5.3(A)–(B)), the move acceptance rate is high (see Figure 5.3(G)), however, only very small steps are ever taken (Figure 5.3(A)). These small steps lead to an over sampling of the low density region of initialisation before the chain drifts toward the high density region. This over sampling of the tail is still evident after 500 iterations (Figure 5.3)(B)), and almost $10,000$ iterations are required to compensate for this initial transient behaviour. We emphasize that here we refer to transient behaviour of the Metropolis-Hastings Markov chain, and this is not to be confused with any transient behaviour of the underlying model. In Figure 5.3(C)–(D), we show that the use of a larger standard deviation, $\sigma = 0.1$, results in the rejection of significantly more proposals (Figure 5.3(C) and Figure 5.3(G)), however, the larger steps result in rapid convergence to the true target density in almost 500 iterations (Figure 5.3(D)). However, increasing the standard deviation further to $\sigma = 1.0$ (Figure 5.3(E)–(F)), results in proposals that overshoot the high density region frequently and most proposals are rejected (Figure 5.3(G)). Consequently, the chain halts for many iterations and parameter space exploration is inefficient (Figure 5.3). Even after $10,000$ iterations the chain still has not reached stationarity.

In Figure 5.3(B),(D), and (F), the exact posterior density for the production-degradation inference problem, $p(k_1 \mid \mathcal{D})$, is overlaid to demonstrate that all chains are still converging to the same stationary distribution. Having this exact solution also enables us to demonstrate the

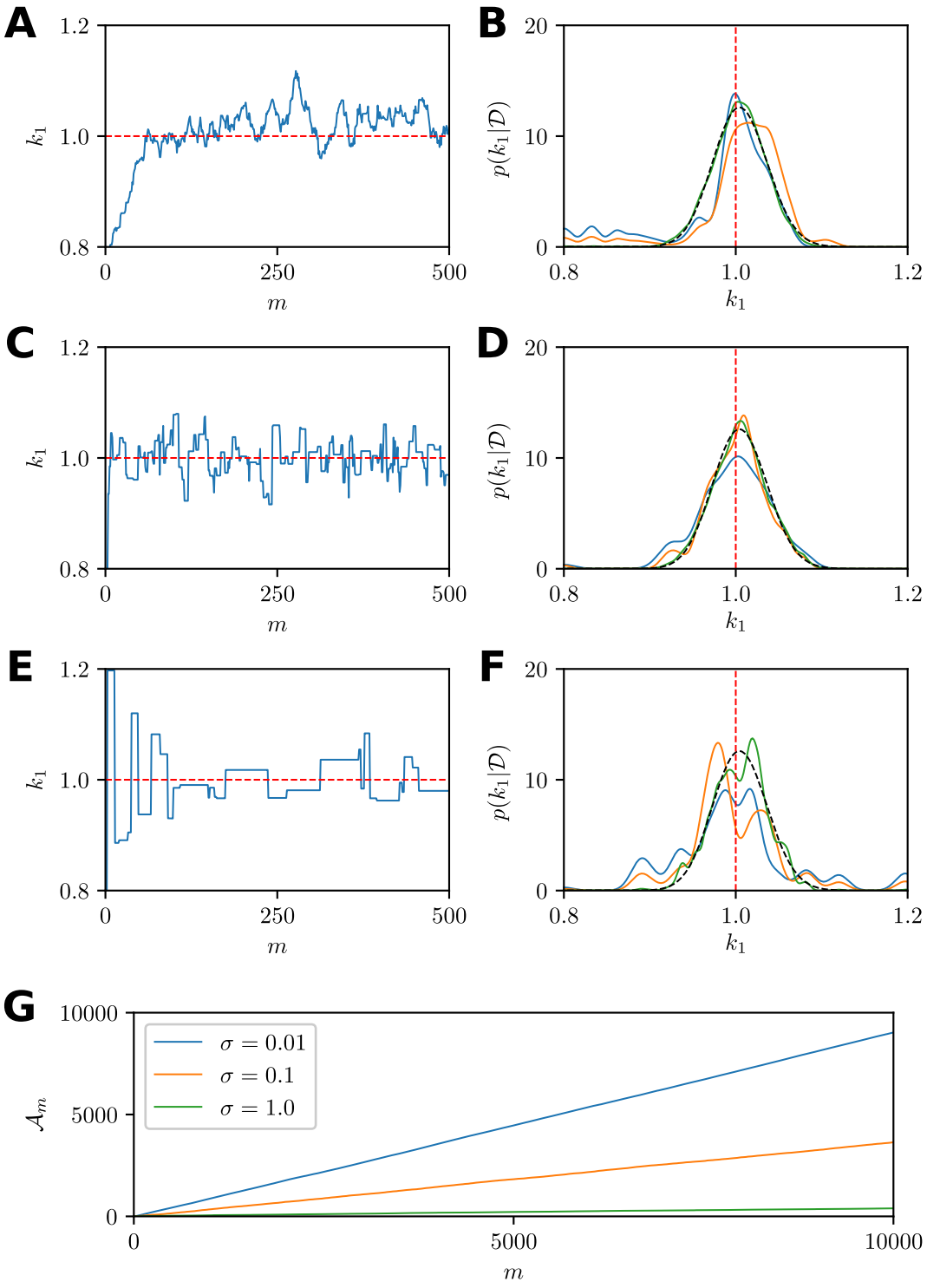

Figure 5.3: The choice of proposal kernel affects the convergence of the Markov chain. (A, C, and E) Trace plots for the first 500 iterations production kinetic parameter k_1, and (B, D, and F) smoothed kernel density estimates for the target Bayesian posteriors are shown for Gaussian proposal kernels with standard deviations of: (A-B) $\sigma = 0.01$; (C-D) $\sigma = 0.1$ and (E-F) $\sigma = 1.0$. Smoothed kernel density estimates for target Bayesian posterior are shown at 250 iterations (solid blue), 500 iterations (solid orange), and 10,000 iterations (solid green) alongside the exact posterior (dashed black). The true production rate of $k_1 = 1.0$ is indicated (dashed red). (G) Comparison of cumulative accepted proposal counts $\mathcal{A}_m$ for different proposal variances, $\sigma = 0.01$ (solid blue), $\sigma = 0.1$ (solid orange) and $\sigma = 1.0$ (solid green).

impact that the choice of proposal kernel has on the efficacy of the Metropolis-Hastings method. In general, the selection of an optimal proposal kernel is an open problem, however, there are techniques that can be applicable in specific cases (Gelman et al., 1996; Roberts and Rosenthal, 2009; Yang and Rodríguez, 2013).

5.3 Likelihood-free MCMC

Nearly all likelihood functions for stochastic biochemical systems of interest are intractable. This renders the standard Metropolis-Hastings method for MCMC sampling (Algorithm 5.1) impossible to implement directly (Sisson et al., 2018; Wilkinson, 2011)(Chapter 4). To deal with this problem, techniques for sampling Bayesian posterior distributions have been developed that avoid the point-wise evaluation of the likelihood. These so-called *likelihood-free* methods fall into two main categories: approximate Bayesian computation; and pseudo-marginal methods.

In this section, we provide a brief description of both approaches in the context of the Metropolis-Hastings method for MCMC sampling. We then demonstrate some important features of these methods in the context of the tractable production-degradation inference problem presented in Section 5.2.3.

5.3.1 Approximate Bayesian computation

ABC is a broad class of Bayesian sampling techniques that are applicable when the likelihood is intractable but simulated data can be generated efficiently for a given parameter vector $\boldsymbol{\theta}$ (Sisson et al., 2018; Sunnåker et al., 2013)(Chapter 4). The fundamental idea is that parameter values that frequently lead to simulated data $\mathcal{D}_s$ that is *similar* to the true observations $\mathcal{D}$ will have higher posterior probability density. In effect, ABC samples from an approximate Bayesian posterior,

$$p(\boldsymbol{\theta} \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon) \propto p(\rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}), \tag{5.11}$$

where the *discrepancy metric*, $\rho(\mathcal{D}, \mathcal{D}_s)$, quantifies how different the two datasets are, the *acceptance threshold*, $\epsilon > 0$, specifies the difference that is considered close, and $\mathcal{D}_s \sim s(\mathcal{D} \mid \boldsymbol{\theta})$

is the data simulation process.

In the context of MCMC sampling, Marjoram et al. (2003), developed a modified Metropolis-Hastings method using the acceptance probability

$$\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}_m) = \begin{cases} \min\left(1, \dfrac{q(\boldsymbol{\theta}_m \mid \boldsymbol{\theta}^*)p(\theta^*)}{q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}_m)p(\theta_m)}\right), & \text{if } \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon, \\ 0, & \text{if } \rho(\mathcal{D}, \mathcal{D}_s) > \epsilon. \end{cases} \tag{5.12}$$

Marjoram et al. (2003) also show that the stationary distribution of the resulting Markov chain is Equation (5.11). Provided that the discrepancy metric, $\rho(\mathcal{D}, \mathcal{D}_s)$, and acceptance threshold, ϵ, are appropriately selected so that $p(\boldsymbol{\theta} \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon) \approx p(\boldsymbol{\theta} \mid \mathcal{D})$, then we can use this Markov chain for inference in the same way that the chain from classical Metropolis-Hastings MCMC (Algorithm 5.1) would be used. An example implementation is provided in `ABCMCMC.jl`.

The choice of $\rho(\mathcal{D}, \mathcal{D}_s)$ and ϵ are critical to both the accuracy of approximate posterior, and the computational cost of the method. Ideally, we require $\rho(\mathcal{D}, \mathcal{D}_s)$ such that we recover the true posterior density in the limit as $\epsilon \to 0$. Using a metric such as the Euclidean distance satisfies this property, however, when the data has high dimensionality it is completely infeasible for small ϵ to accept any parameter proposals. Conversely, metrics based on summary statistics of the data can be used to reduce the data dimensionality so that a smaller ϵ can be used, however, this may not lead to the true posterior as $\epsilon \to 0$. In general, one requires the summary statistics to be sufficient statistics (Fearnhead and Prangle, 2012) and ϵ to be of similar order to the observation error (Toni et al., 2009; Wilkinson, 2013) to obtain accurate posteriors for the purposes of inference.

5.3.2 Pseudo-marginal methods

Pseudo-marginal methods (Andrieu and Roberts, 2009) are an alternative approach to likelihood-free inference with some desirable properties compared with ABC. The pseudo-marginal approach can be used when one has an unbiased Monte Carlo estimator, $\hat{\mathcal{L}}(\theta; \mathcal{D})$, for the point-wise evaluation of the likelihood function $\mathcal{L}(\theta; \mathcal{D})$. This estimator is used directly in place of the true likelihood for the purposes of MCMC.

In the context of Metropolis-Hastings MCMC (Algorithm 5.1), the acceptance probability for

the pseudo-marginal approach is

$$\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}_m) = \min\left(1, \frac{q(\boldsymbol{\theta}_m \mid \boldsymbol{\theta}^*)\hat{\mathcal{L}}(\boldsymbol{\theta}^*; \mathcal{D})p(\theta^*)}{q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}_m)\hat{\mathcal{L}}(\boldsymbol{\theta}_m; \mathcal{D})p(\theta_m)}\right). \tag{5.13}$$

After initial inspection, one would expect the stationary distribution of the resulting Markov chain to be an approximation to the true posterior, just as with the ABC approach using Equation (5.12). Surprisingly, this is not the case; the stationary distribution of the Markov chain using Equation (5.13) is, in fact, the exact posterior distribution. As a result, pseudo-marginal methods have been referred to as *exact approximations* (Golightly and Wilkinson, 2011). For an explanation for why the true posterior is recovered, see Section 5.7.2. For more detail we refer the reader to Andrieu and Roberts (2009), Beaumont (2003), and Golightly and Wilkinson (2011). `PseudoMarginalMetropolisHastings.jl` is provided as an example implementation.

Unlike classical Metropolis Hastings, the acceptance probability, $\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}_m)$ (Equation (5.13)), is still a random variable, given values for $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}_m$. This additional randomness reduces the rate at which the Markov chain approaches stationarity, but the additional noise can be controlled through reducing the variance of the likelihood estimator $\hat{\mathcal{L}}(\boldsymbol{\theta}; \mathcal{D})$. However, reducing the variance necessarily requires higher computation costs since a larger number of Monte Carlo samples will be required for computing $\hat{\mathcal{L}}(\boldsymbol{\theta}; \mathcal{D})$. Research has been undertaken to try to develop methods to choose the number of samples optimally. In particular, Doucet et al. (2015) perform a detailed analysis and find, under some restrictive assumptions, that choosing the number of samples such that $\mathbb{V}\left[\log \hat{\mathcal{L}}(\bar{\boldsymbol{\theta}}; \mathcal{D})\right] \approx 1.2$, where $\bar{\boldsymbol{\theta}}$ is the posterior mean, is the optimal trade-off.

5.3.3 Comparison for an example with a tractable likelihood

We now demonstrate the ABC and pseudo-marginal approaches to MCMC using the tractable production-degradation problem from Section 5.2.3. Specifically, we demonstrate how the ABC acceptance threshold and the pseudo-marginal Monte Carlo estimator variance affect both the rate of convergence and the stationary distribution.

For the ABC case, we can generate simulated data $\mathcal{D}_s = \left[X_T^{(1)}, X_T^{(2)}, \ldots, X_T^{(n)}\right]$ where

$X_T^{(1)}, X_T^{(2)}, \ldots, X_T^{(n)}$ are independent approximate realisations of the production degradation model (Equation (5.5)) using the Euler-Maruyama scheme over the interval $0 \leq t \leq T$ with $\Delta t = 1.0$, $T = 1000.0$, $k_2 = 0.01$, and k_1 is given by the state of the Markov chain θ_m. Given that the data has dimension $n = 10$ (Sections 5.2.3 and 5.7.4), we choose a discrepancy metric that reduces the data dimension for ease of demonstration, that is,

$$\rho(\mathcal{D}, \mathcal{D}_s) = |\hat{\mu}(\mathcal{D}) - \hat{\mu}(\mathcal{D}_s)| + |\hat{\sigma}(\mathcal{D}) - \hat{\sigma}(\mathcal{D}_s)|\,,$$

where $\hat{\mu}(\mathcal{D})$ and $\hat{\sigma}(\mathcal{D})$ are the sample mean and standard deviation of the observations $\mathcal{D}$, and $\hat{\mu}(\mathcal{D}_s)$ and $\hat{\sigma}(\mathcal{D}_s)$ are the sample mean and standard deviation of the simulated data $\mathcal{D}_s$. Figure 5.4(A)–(F) demonstrates the behaviour of ABC MCMC using acceptance thresholds of $\epsilon = 14$ (Figure 5.4(A)–(B)), $\epsilon = 7$ (Figure 5.4(C)–(D)) and $\epsilon = 3.5$ (Figure 5.4(E)–(F)) (see example `DemoABCMCMC.jl`). The Markov chain trajectories shown in Figure 5.4(A), (C), (E), indicate that larger values of ϵ lead to more rapid convergence to stationarity. The converse is true for the accuracy of the stationary distribution as an approximation to the exact posterior, as demonstrated in Figure 5.4(B), (C), (F), with larger values of ϵ leading to a more diffuse, approximate posterior. This highlights a known shortcoming for ABC for the purposes of MCMC sampling; choosing a small ϵ for accuracy will tend to result in a Markov chain that repeatedly gets stuck in the same location (Sisson et al., 2007).

For the pseudo-marginal approach we use a standard smoothed kernel density estimate for the likelihood, that is,

$$\hat{\mathcal{L}}(\boldsymbol{\theta}; \mathcal{D}) = \frac{1}{(Rh)^n} \prod_{i=1}^{n} \sum_{j=1}^{R} K\left(\frac{Y_{\text{obs}}^{(i)} - X_T^{(j)}}{h}\right),$$

where h is the smoothing parameter chosen using Silverman's rule (Silverman, 1986), $K(x)$ is a standard Gaussian smoothing kernel, and $X_T^{(1)}, X_T^{(2)}, \ldots, X_T^{(R)}$ are independent approximate realisations of the production degradation model (Equation (5.5)) using the Euler-Maruyama scheme with identical parameterisation as used for ABC. The variance of the likelihood estimator depends on the number of realisations used in the estimate, R. Figure 5.4(G)–(L) demonstrates the behaviour of the pseudo-marginal approach to MCMC using different realisation numbers of $R = 25$ (Figure 5.4(G)–(H)), $R = 50$ (Figure 5.4(I)–(J)) and $R = 100$ (Figure 5.4(K)–(L)) (see example `DemoPMMH.jl`). As expected, increasing R has the effect of increasing convergence (although not significantly so). More importantly, regardless of the

value R, the same stationary distribution is approached in the limit, that is, the exact posterior distribution.

This highlights a major advantage of pseudo-marginal methods, that is, the stationary distribution is independent of the number of realisations, R; furthermore the stationary distribution is the exact Bayesian posterior distribution. Even with $R = 1$ the method will eventually converge to the exact posterior distribution. This is in stark contrast to ABC methods where the stationary distribution depends on the discrepancy metric and acceptance threshold. Ultimately this means, that user choices only affect the computational performance of pseudo-marginal methods rather than both computational performance and inference accuracy with ABC. This is a clear advantage of the pseudo-marginal approach.

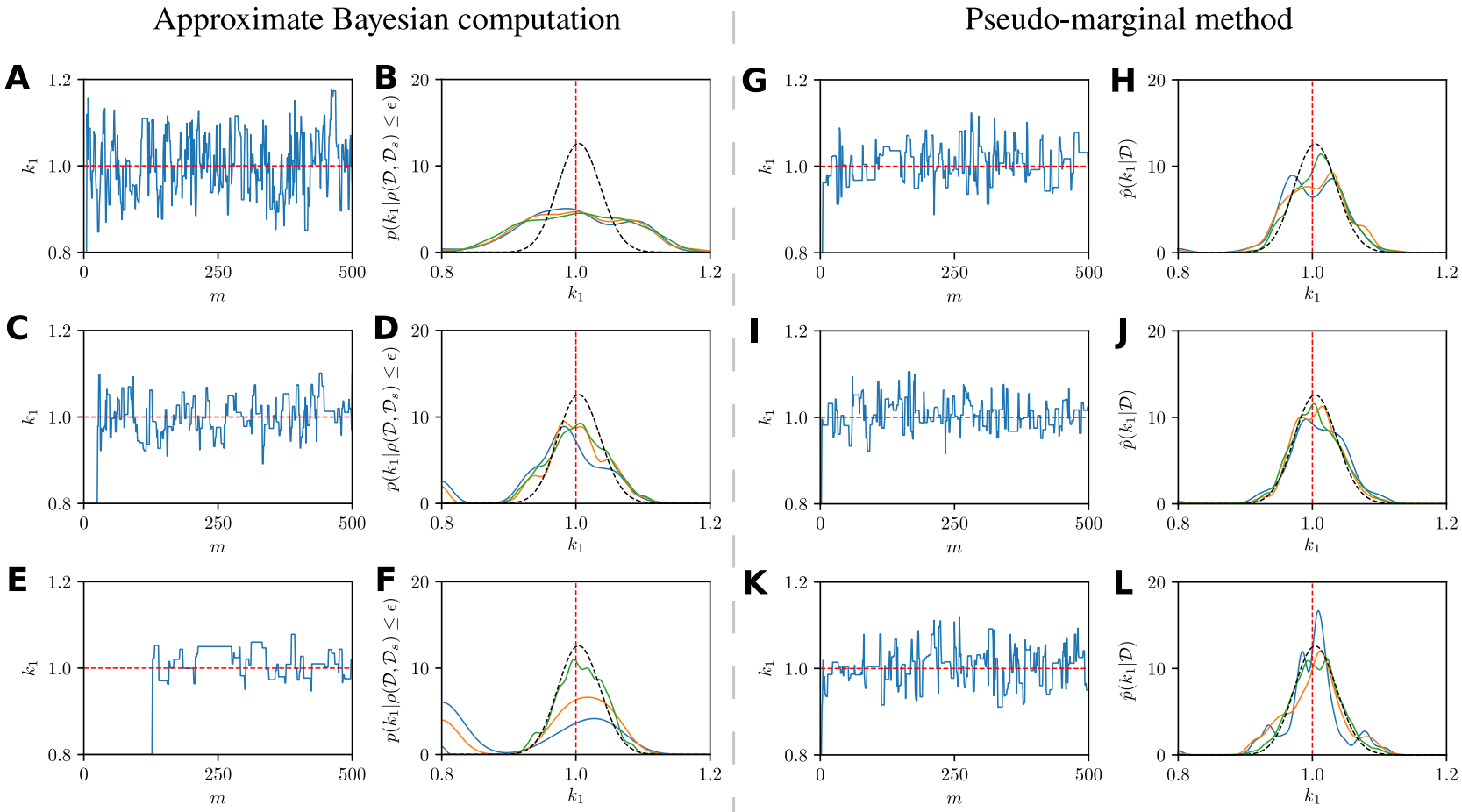

Figure 5.4: Comparison of likelihood-free MCMC algorithms, (A-F) ABC and (G-L) the pseudo-marginal approach, using the tractable production degradation example. ABC trace plots and smoothed kernel density estimates for the target approximate Bayesian posterior are shown for decreasing acceptance thresholds: (A-B) $\epsilon = 14$; (C-D) $\epsilon = 7$ and (E-F) $\epsilon = 3.5$. Pseudo-marginal trace plots and smoothed kernel density estimates for the target approximate Bayesian posterior are shown for increasing sample numbers for likelihood estimation: (G-H) $n = 25$; (I-J) $n = 50$ and (K-L) $n = 100$. Smoothed kernel density estimates for target Bayesian posterior are shown at 250 iterations (solid blue), 500 iterations (solid orange), and 10,000 iterations (solid green) alongside the exact posterior (dashed black). The true production rate of $k_1 = 1.0$ is indicated (dashed red). In all cases, the proposal kernel is Gaussian with variance $\sigma^2 = 0.01$, and the chain is initialised with $\theta_0 = 0.8$.

5.4 Pseudo-marginal methods for biochemical systems

The production-degradation example presented in Section 5.2.3 is useful for highlighting the essential concepts of standard MCMC sampling and likelihood free alternatives. However, this inference problem is very simple compared to practical problems, since real biochemical processes are generally not observed in their stationary state without observation error. Rather, real biochemical process data is often characterised by noisy, time-course data, with few observations in time and only partially observed states (Finkenstädt et al., 2008; Golightly and Wilkinson, 2011)(Chapter 4).

5.4.1 The challenge for time-course data

In the case of time-course data, the observations are samples at discrete points in time, $t_0, t_1, \ldots, t_n$, from a single realisation of a stochastic process, such as a gene regulatory network. The observation process, denoted by $\{\mathbf{Y}_t\}_{0\leq t}$, often has the form, $\mathbf{Y}_t \sim g(\mathbf{Y}_t \mid \mathbf{X}_t)$, where $\{\mathbf{X}_t\}_{0\leq t}$ is the underlying stochastic process, prescribed by the chemical Langevin equation (Equation (5.2)), that governs the biochemical kinetics, and $g(\mathbf{Y}_t \mid \mathbf{X}_t)$ is the observation process. The resulting discrete observations will be $\mathcal{D} = [\mathbf{Y}_{\text{obs}}^{(0)}, \mathbf{Y}_{\text{obs}}^{(1)}, \ldots, \mathbf{Y}_{\text{obs}}^{(n)}]$ with $\mathbf{Y}_{\text{obs}}^{(i)} = \mathbf{Y}_{t_i}$ for $i = 0, 1, \ldots, n$. The likelihood for such observations is

$$\mathcal{L}(\boldsymbol{\theta}; \mathcal{D}) = p(\mathbf{Y}_{\text{obs}}^{(0)}) \prod_{i=1}^{n} p(\mathbf{Y}_{\text{obs}}^{(i)} \mid \mathbf{Y}_{\text{obs}}^{(0)}, \ldots, \mathbf{Y}_{\text{obs}}^{(i-1)}). \tag{5.14}$$

Not only is this likelihood intractable, but a direct Monte Carlo likelihood estimator will be impractical for the pseudo-marginal approach. For example, the following is a direct Monte Carlo estimate for Equation (5.14)

$$\hat{\mathcal{L}}(\boldsymbol{\theta}; \mathcal{D}) = \frac{1}{R} \sum_{j=1}^{R} \prod_{i=1}^{n} g(\mathbf{Y}_{\text{obs}}^{(i)} \mid \mathbf{X}_{t_i}^{(j)}), \tag{5.15}$$

where $[\mathbf{X}_t^{(1)}, \mathbf{X}_t^{(2)}, \ldots, \mathbf{X}_t^{(R)}]$ are R independent realisations of the continuous sample path from the chemical Langevin equation (Equation (5.2)), that are subsequently observed at times $t_0, t_1, \ldots, t_n$. However, a prohibitively large number of sample paths, R, will be required to obtain an acceptable variance in the estimator in Equation (5.15). Consequently, more advanced

approaches to pseudo-marginal are required.

The following observation assists finding an alternative solution,

$$p(\mathbf{Y}_{\text{obs}}^{(i)} \mid \mathbf{Y}_{\text{obs}}^{(0)}, \ldots, \mathbf{Y}_{\text{obs}}^{(i-1)}) = \int_{\mathbb{R}^N} g(\mathbf{Y}_{\text{obs}}^{(i)} \mid \mathbf{X}_{t_i}) p(\mathbf{X}_{t_i} \mid \mathbf{Y}_{\text{obs}}^{(0)}, \ldots, \mathbf{Y}_{\text{obs}}^{(i-1)}) \, \mathrm{d}\mathbf{X}_{t_i}.$$

That is, provided we are able to sample from $p(\mathbf{X}_{t_i} \mid \mathbf{Y}_{\text{obs}}^{(0)}, \ldots, \mathbf{Y}_{\text{obs}}^{(i-1)})$ for any $i = 1, 2, \ldots, n$, then we can use the alternative Monte Carlo estimator

$$\hat{\mathcal{L}}(\boldsymbol{\theta}; \mathcal{D}) = \prod_{i=1}^{n} \frac{1}{R} \sum_{j=1}^{R} g(\mathbf{Y}_{\text{obs}}^{(i)} \mid \mathbf{X}_{t_i}^{(j)}), \tag{5.16}$$

where $[\mathbf{X}_{t_i}^{(1)}, \mathbf{X}_{t_i}^{(2)}, \ldots, \mathbf{X}_{t_i}^{(R)}]$ are R samples from the distribution $p(\mathbf{X}_{t_i} \mid \mathbf{Y}_{\text{obs}}^{(0)}, \ldots, \mathbf{Y}_{\text{obs}}^{(i-1)})$. This estimator will have lower variance because conditioning the samples $[\mathbf{X}_{t_i}^{(1)}, \mathbf{X}_{t_i}^{(2)}, \ldots, \mathbf{X}_{t_i}^{(R)}]$ on all observations taken up to time t_{i-1} automatically removes contributions by trajectories that do not match the observational history. The challenge is in the sampling of $p(\mathbf{X}_{t_i} \mid \mathbf{Y}_{\text{obs}}^{(0)}, \ldots, \mathbf{Y}_{\text{obs}}^{(i-1)})$ and it motivates the use of, so called, *particle filters* (Doucet and Johanson, 2011). We present the mathematical basis for this approach in the next section, along with practical examples that demonstrate how the method works in practice.

5.4.2 Particle MCMC

The bootstrap particle filter (Gordon et al., 1993; Doucet and Johanson, 2011) is a technique based on sequential importance sampling (Del Moral et al., 2006). This enables one to sample from the sequence of distributions $p(\mathbf{X}_{t_1} \mid \mathbf{Y}_{\text{obs}}^{0}), p(\mathbf{X}_{t_2} \mid \mathbf{Y}_{\text{obs}}^{0}, \mathbf{Y}_{\text{obs}}^{1}), \ldots, p(\mathbf{X}_{t_n} \mid \mathbf{Y}_{\text{obs}}^{0}, \ldots, \mathbf{Y}_{\text{obs}}^{n-1})$ and thereby evaluate the lower variance likelihood estimator (Equation (5.16)).

Suppose we have independent samples, called particles,

$$\mathbf{X}_{t_{i-1}}^{(1)}, \mathbf{X}_{t_{i-1}}^{(2)}, \ldots, \mathbf{X}_{t_{i-1}}^{(R)} \sim p(\mathbf{X}_{t_{i-1}} \mid \mathbf{Y}_{\text{obs}}^{0}, \ldots, \mathbf{Y}_{\text{obs}}^{i-1}).$$

Then, using the Euler-Maruyama scheme (or similar), we can simulate each particle forward to time t_i. This results in a new set of independent particles

$$\tilde{\mathbf{X}}_{t_i}^{(1)}, \tilde{\mathbf{X}}_{t_i}^{(2)}, \ldots, \tilde{\mathbf{X}}_{t_i}^{(R)} \sim p(\mathbf{X}_{t_i} \mid \mathbf{Y}_{\text{obs}}^{0}, \ldots, \mathbf{Y}_{\text{obs}}^{i-1}).$$

From these particles, we can evaluate the Monte Carlo estimate for the marginal likelihood at time t_i,

$$\hat{p}(\mathbf{Y}^i_{\text{obs}} \mid \mathbf{Y}^0_{\text{obs}}, \ldots, \mathbf{Y}^{i-1}_{\text{obs}}) = \frac{1}{R}\sum_{k=1}^{R} g(\mathbf{Y}^i_{\text{obs}} \mid \tilde{\mathbf{X}}^{(k)}_{t_i}). \tag{5.17}$$

Provided we can then generate a new set of independent particles,

$$\mathbf{X}^{(1)}_{t_i}, \mathbf{X}^{(2)}_{t_i}, \ldots, \mathbf{X}^{(R)}_{t_i} \sim p(\mathbf{X}_{t_i} \mid \mathbf{Y}^0_{\text{obs}}, \ldots, \mathbf{Y}^i_{\text{obs}}),$$

we can compute Equation (5.17) for all $i = 1, 2, \ldots, n$, and hence compute the likelihood estimator given in Equation (5.16). Progress can be made by noting that, through application of Bayes' Theorem,

$$p(\mathbf{X}_{t_i} \mid \mathbf{Y}^0_{\text{obs}}, \ldots, \mathbf{Y}^i_{\text{obs}}) = \frac{p(\mathbf{Y}^i_{\text{obs}} \mid \mathbf{X}_{t_i}) p(\mathbf{X}_{t_i} \mid \mathbf{Y}^0_{\text{obs}}, \ldots, \mathbf{Y}^{i-1}_{\text{obs}})}{p(\mathbf{Y}^i_{\text{obs}} \mid \mathbf{Y}^0_{\text{obs}}, \ldots, \mathbf{Y}^{i-1}_{\text{obs}})}.$$

Therefore, we can approximate the set of particles $\mathbf{X}^{(1)}_{t_i}, \mathbf{X}^{(2)}_{t_i}, \ldots, \mathbf{X}^{(R)}_{t_i}$ by resampling the particles $\tilde{\mathbf{X}}^{(1)}_{t_i}, \tilde{\mathbf{X}}^{(2)}_{t_i}, \ldots, \tilde{\mathbf{X}}^{(R)}_{t_i}$ with replacement using probabilities,

$$\mathbb{P}(\mathbf{X}_{t_i} = \tilde{\mathbf{X}}^{(k)}_{t_i}) = \frac{g(\mathbf{Y}^i_{\text{obs}} \mid \tilde{\mathbf{X}}^{(k)}_{t_i})}{R\hat{p}(\mathbf{Y}^i_{\text{obs}} \mid \mathbf{Y}^0_{\text{obs}}, \ldots, \mathbf{Y}^{i-1}_{\text{obs}})} = \frac{g(\mathbf{Y}^i_{\text{obs}} \mid \tilde{\mathbf{X}}^{(k)}_{t_i})}{\sum_{j=1}^{R} g(\mathbf{Y}^i_{\text{obs}} \mid \tilde{\mathbf{X}}^{(j)}_{t_i})}.$$

The result is a set of equally weighed particles approximately distributed according to $p(\mathbf{X}_{t_i} \mid \mathbf{Y}^0_{\text{obs}}, \ldots, \mathbf{Y}^i_{\text{obs}})$. This leads to the *bootstrap particle filter* (Algorithm 5.2) (Gordon et al., 1993). An example implementation is provided in `BootstrapParticleFilter.jl`.

Figure 5.5 provides a visual demonstration of this process using a small number of particles, $R = 4$, for ease of visualisation. Figure 5.5(A) shows time-course data with error bars indicating the magnitude of the observation error. This data, with very low time resolution, is typical of many experimental studies, such as the data in Figure 5.1(C). Initially, all four particles are set to the initial data point with equal weighting (Figure 5.5(B)). The particles are then evolved forwards to the next observation time (Figure 5.5(C)), and the weighting is calculated as the probability density that the observation occurred based on each of these particles (Figure 5.5(D)). Note that only one of the four particles contribute significantly to the likelihood after this first forwards step, thus simulating the other three particles forwards

Algorithm 5.2 The bootstrap particle filter for likelihood estimation

1: Initialise $i = 0$ and $\left\{\mathbf{X}_{t_0}^{(k)}\right\}_{k=1}^{R}$ where $\mathbf{X}_{t_0}^{(k)} \sim p(\mathbf{X}_{t_0} \mid \mathbf{Y}_{\text{obs}}^0)$ for $k = 1, 2, \ldots, R$.
2: **for** $i = 1, \ldots, n$ **do**
3: **for** $k = 1, \ldots, R$ **do**
4: Simulate particle forward, $\mathbf{X}_{t_i}^{(k)} \sim s(\mathbf{X}_{t_i}^{(k)} \mid \mathbf{X}_{t_{i-1}}^{(k)})$.
5: Compute weight, $W_i^k \leftarrow g(\mathbf{Y}_{\text{obs}}^i \mid \mathbf{X}_{t_i}^{(k)})$.
6: **end for**
7: Compute marginal likelihood estimate, $\hat{p}(\mathbf{Y}_{\text{obs}}^i \mid \mathbf{Y}_{\text{obs}}^0, \ldots, \mathbf{Y}_{\text{obs}}^{i-1}) \leftarrow \frac{1}{R}\sum_{k=1}^{R} W_i^k$.
8: Resample particles, $\left\{\mathbf{X}_{t_i}^{(k)}\right\}_{k=1}^{R}$, with replacement using probabilities $W_i^k / \left[\sum_{j=1}^{R} W_i^j\right]$ for $k = 1, 2, \ldots, R$.
9: **end for**
10: Compute likelihood estimate, $\hat{\mathcal{L}}(\boldsymbol{\theta}; \mathcal{D}) \leftarrow \prod_{i=1}^{n} \hat{p}(\mathbf{Y}_{\text{obs}}^i \mid \mathbf{Y}_{\text{obs}}^0, \ldots, \mathbf{Y}_{\text{obs}}^{i-1})$.

any further would be a waste of computational effort. As such, we generate four independent continuations of the one highly weighted particle (Figure 5.5(D)) to evolve toward the next observation time (Figure 5.5(E)). The same process is repeated using the second generation of weights (Figure 5.5(F)), in order to perform the last step (Figure 5.5(G)-(H)).

The visualisation in Figure 5.5 highlights the effect of conditioning on past observations. That is, model simulations that lead to a small likelihood in one of the observations are discarded early and new particles are generated by continuing high likelihood simulations. As a result all samples are focused on the higher density region of the likelihood and the variance of the estimator is reduced.

The application of particle filters for likelihood estimators within MCMC schemes is called *Particle MCMC* (Wilkinson, 2012). For the purposes for this article, we directly apply the pseudo-marginal method using Algorithm 5.2 to evaluate the acceptance probability (Equation (5.13)) within the Metropolis-Hastings MCMC scheme (Algorithm 5.1). This turns out to be a special case of the *Particle marginal Metropolis-Hastings sampler* that may be used effectively to sample the joint probability density $p(\boldsymbol{\theta}, \mathbf{X}_{t_0}, \mathbf{X}_{t_1}, \ldots, \mathbf{X}_{t_n} \mid \mathbf{Y}_{\text{obs}}^0, \mathbf{Y}_{\text{obs}}^1, \ldots, \mathbf{Y}_{\text{obs}}^n)$, as demonstrated within a very general framework introduced by (Andrieu et al., 2010).

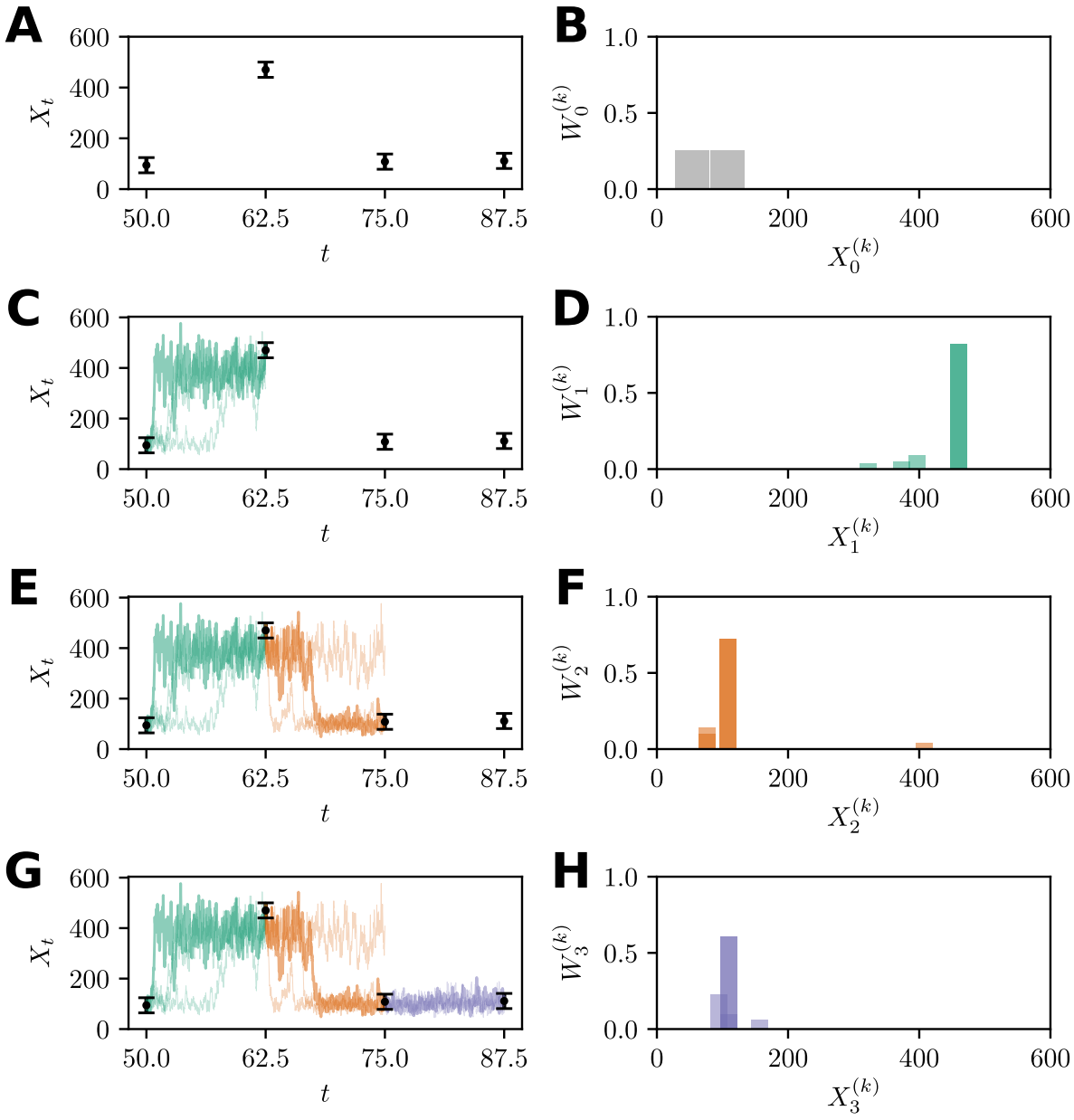

Figure 5.5: Demonstration of the bootstrap particle filter using $R = 4$ particles for demonstration purposes. (A) Observations with error bars indication three standard deviations of the observation noise distribution. (B) Particles are initially weighted equally. (C)–(H) Three iterations of the bootstrap particle filter. (C),(E), and (G) Particle forward trajectories with weights indicated by opacity. (D),(F), and (H) Particle weight distributions computed for resampling.

5.4.3 Practical considerations

There are a number of factors that may affect the performance of particle MCMC sampling in practice. Firstly, an important issue to discuss for sequential importance resampling, such as the bootstrap particle filter, is the problem of particle degeneracy. That is, as the number

of iterations increases, the number of particles with non-zero weights decreases. As a result, the accuracy of approximation to $p(\mathbf{X}_{t_i} \mid \mathbf{Y}^0_{\text{obs}}, \ldots, \mathbf{Y}^i_{\text{obs}})$ degrades as this dependency on a very small particle count introduces bias. While the resampling step reduces the impact of degeneracy, in general a larger number of observations, n, will necessitate a large number of particles, R, for the likelihood estimator (Doucet and Johanson, 2011). The problem of degeneracy becomes even more problematic when the observation error is very small (Golightly and Wilkinson, 2008, 2011), and may require more advanced resampling methods, such as systematic and stratified resampling (Kitagawa, 1996; Carpenter et al., 1999). In this work, we apply a direct multinomial resampling scheme (Doucet and Johanson, 2011).

Just as with the more general pseudo-marginal approach, there is a trade-off between the convergence rate of the Markov chain and the computational cost of each likelihood estimate. While Doucet et al. (2015) provide guidelines for optimally choosing R, these guides may not be feasible to implement since the behaviour of the likelihood estimator about the posterior mean is rarely known (especially since one is often using MCMC in order to compute this quantity).

The performance of particle MCMC methods also depends on the choice of proposal kernel, just as with classical Metropolis-Hastings. When the full inference problem is considered, there are a number of novel proposal schemes (Andrieu et al., 2010; Pooley et al., 2015). A number of asymmetric proposal kernels, such as *preconditioned Crank-Nicholson Langevin proposals* (Cotter et al., 2013), can also be very effective in high dimensional parameter spaces. However, in general, one needs to perform experimentation to elucidate an effective combination of proposal kernel and particle numbers that will converge in an acceptable timeframe.

The question of assessing convergence can be challenging. Typically, the auto-correlation functions (ACF) for each parameter are computed and the potential scale reduction is computed (Geyer, 1992). However, these diagnostics for convergence can be very misleading, especially if the posterior is multimodal. To deal with this, it is common to use multiple chains and assess the within-chain and between-chain variances Gelman et al. (1996, 2014). In this work, we follow the recent recommendations of (Vehtari et al., 2019).

5.5 Examples with intractable likelihoods

As a practical demonstration of the use of particle MCMC, three examples are provided where expressions for the likelihood function are not available.

5.5.1 Example 1: Michaelis-Menten enzyme kinetics

The first example is based on the stochastic variant of the classical model for enzyme kinetics (Michaelis and Menten, 1913; Rao and Arkin, 2003).

5.5.1.1 Model definition

The classical model of Michaelis and Menten (1913) for enzyme kinetics describes the conversion of a chemical substrate S into a product P through the binding of an enzyme E. An enzyme molecule, E, binds to a substrate molecule, S, to form a complex, C, to convert S to P. The stochastic process is describe through three reactions (Rao and Arkin, 2003),

$$\underbrace{E + S \xrightarrow{k_1} C}_{\substack{\text{enzyme and substrate molecules} \\ \text{combine to form a complex}}} \quad , \quad \underbrace{C \xrightarrow{k_2} E + S}_{\text{decay of complex}}, \quad \text{and} \quad \underbrace{C \xrightarrow{k_3} E + P}_{\substack{\text{catalytic conversion} \\ \text{of substrate to product}}} , \tag{5.18}$$

with propensities, $a_1(\mathbf{X}_t) = k_1 E_t S_t$, $a_2(\mathbf{X}_t) = k_2 C_t$, and $a_3(\mathbf{X}_t) = k_3 C_t$, where $\mathbf{X}_t = [E_t, S_t, C_t, P_t]^{\mathrm{T}}$, and stoichiometries

$$\boldsymbol{\nu}_1 = \begin{bmatrix} -1 \\ -1 \\ 1 \\ 0 \end{bmatrix}, \quad \boldsymbol{\nu}_2 = \begin{bmatrix} 1 \\ 1 \\ -1 \\ 0 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\nu}_3 = \begin{bmatrix} 1 \\ 0 \\ -1 \\ 1 \end{bmatrix}.$$

Application of the chemical Langevin approximation (Equation (5.2)) to the Michaelis-Menten model (Equation (5.18)) leads to a coupled system of Itō SDEs

$$\begin{aligned}
\mathrm{d}E_t &= [-k_1 E_t S_t + (k_2 + k_3) C_t]\mathrm{d}t - \sqrt{k_1 E_t S_t}\mathrm{d}W_t^{(1)} + \sqrt{k_2 C_t}\mathrm{d}W_t^{(2)} + \sqrt{k_3 C_t}\mathrm{d}W_t^{(3)}, \\
\mathrm{d}S_t &= (-k_1 E_t S_2 + k_2 C_t)\mathrm{d}t - \sqrt{k_1 E_t S_t}\mathrm{d}W_t^{(1)} + \sqrt{k_2 C_t}\mathrm{d}W_t^{(2)}, \\
\mathrm{d}C_t &= [k_1 E_t S_t - (k_2 + k_3) C_t]\mathrm{d}t + \sqrt{k_1 E_t S_t}\mathrm{d}W_t^{(1)} - \sqrt{k_2 C_t}\mathrm{d}W_t^{(2)} - \sqrt{k_3 C_t}\mathrm{d}W_t^{(3)}, \\
\mathrm{d}P_t &= k_3 C_t \mathrm{d}t + \sqrt{k_3 C_t}\mathrm{d}W_t^{(3)},
\end{aligned} \tag{5.19}$$

where $W_t^{(1)}$, $W_t^{(2)}$ and $W_t^{(3)}$ are independent Wiener processes driving each reaction channel. A typical realisation of the model is provided in Figure 5.6. Note that as $t \to \infty$ the stationary distribution is a product of Dirac distributions, that is, a point mass at $\mathbf{X}_\infty = [E_0 + C_0, 0, 0, S_0 + C_0 + P_0]^\mathrm{T}$ given $\mathbf{X}_0 = [E_0, S_0, C_0, P_0]^\mathrm{T}$. Therefore observations involving the transient behaviour are essential to recover information about the rate parameters.

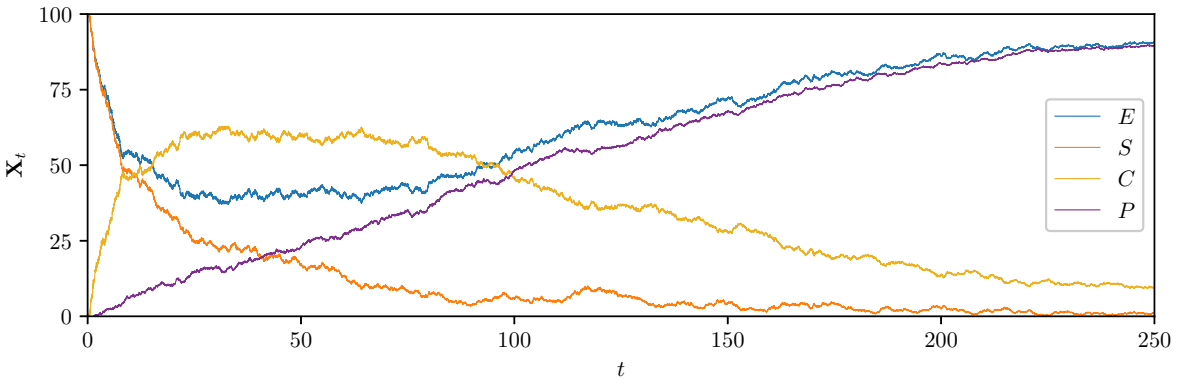

Figure 5.6: Example realisation of the Michaelis-Menten model demonstrating enzyme kinetics. The simulation is produced using the Euler-Maruyama scheme with $\Delta t = 1 \times 10^{-3}$, initial condition $\mathbf{X}_0 = [100, 100, 0, 0]$, and parameters $k_1 = 1 \times 10^{-3}$, $k_2 = 5 \times 10^{-3}$, and $k_3 = 1 \times 10^{-2}$.

While some analytic progress on likelihood approximation can be made using moment closures (Schnoerr et al., 2017), the second order reaction for the production of complexes, C, effectively renders the distribution of the forwards problem analytically intractable.

5.5.1.2 Time-course data and inference problem definition

We generate synthetic data using a single realisation of the Michaelis-Menten chemical Langevin SDE with initial condition $\mathbf{X}_0 = [100, 100, 0, 0]^{\mathrm{T}}$ and kinetic rate parameters $k_1 = 1 \times 10^{-3}$, $k_2 = 5 \times 10^{-3}$ and $k_3 = 1 \times 10^{-2}$. Observations are taken at $n = 20$ uniformly spaced time points $t_1 = 5$, $t_2 = 10, \ldots,$ $t_{20} = 100$. The observation process considers Gaussian noise applied to each chemical species copy number with a standard deviation of $\sigma_{\text{obs}} = 10$, that is, $\mathbf{Y}_{\text{obs}}^{(i)} \sim \mathcal{N}(\mathbf{X}_{t_i}, \sigma_{\text{obs}}^2 \mathbf{I})$ where $\mathbf{I}$ is the 4×4 identity matrix. See Section 5.7.4 for the resulting data table.

We perform inference on all three rate parameters, $\boldsymbol{\theta} = [k_1, k_2, k_3]^{\mathrm{T}}$. We use the particle MCMC approach to sample the Bayesian posterior,

$$p(k_1, k_2, k_3 \mid \mathcal{D}) \propto \mathcal{L}(k_1, k_2, k_3; \mathcal{D}) p(k_1, k_2, k_3),$$

where $p(k_1, k_2, k_3)$ is the joint uniform prior with independent components $k_1 \sim \mathcal{U}(0, 5 \times 10^{-3})$, $k_2 \sim \mathcal{U}(0, 2.5 \times 10^{-2})$ and $k_3 \sim \mathcal{U}(0, 5 \times 10^{-2})$. The likelihood is estimated using the bootstrap particle filter (Algorithm 5.2) with $R = 100$ particles and the Euler-Maruyama method for simulation with $\Delta t = 0.1$.

5.5.1.3 Chain initialisation and proposal tuning

Any application of MCMC requires both a method of initialising the chain and choosing the proposal kernel. To deal with both of these challenges we can apply trial chains.

Four trial chains are simulated for $\mathcal{M} = 8,000$ iterations, each initialised with a random sample from the prior with a non-zero likelihood estimate. The proposal kernel used in all four trial chains is a Gaussian with covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} 5.208 \times 10^{-9} & 0 & 0 \\ 0 & 1.302 \times 10^{-7} & 0 \\ 0 & 0 & 5.208 \times 10^{-7} \end{bmatrix}.$$

The diagonal entries correspond to a proposal density such that one tenth of the prior standard deviation for each parameter is within a single standard deviation of the proposal; such independent proposal kernels are typical choices. However, this proposal is not very efficient, as Figure 5.7(A)–(F) indicates for the first 8,000 iterations of the first chain. However, these chains are not used for inference, just for configuring a new set of more efficient chains.

The tuned proposal kernel is constructed by taking the total covariance matrix using the pooled sample of the four trial chains (total of $32,000$ samples),

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} 2.693 \times 10^{-7} & 9.754 \times 10^{-7} & 1.536 \times 10^{-7} \\ 9.754 \times 10^{-7} & 3.348 \times 10^{-5} & -1.289 \times 10^{-5} \\ 1.536 \times 10^{-7} & -1.289 \times 10^{-5} & 4.677 \times 10^{-5} \end{bmatrix},$$

and applying the optimal scaling rule from Roberts and Rosenthal (2001)

$$\boldsymbol{\Sigma}_{\text{opt}} = \frac{2.38^2}{3}\hat{\boldsymbol{\Sigma}}.$$

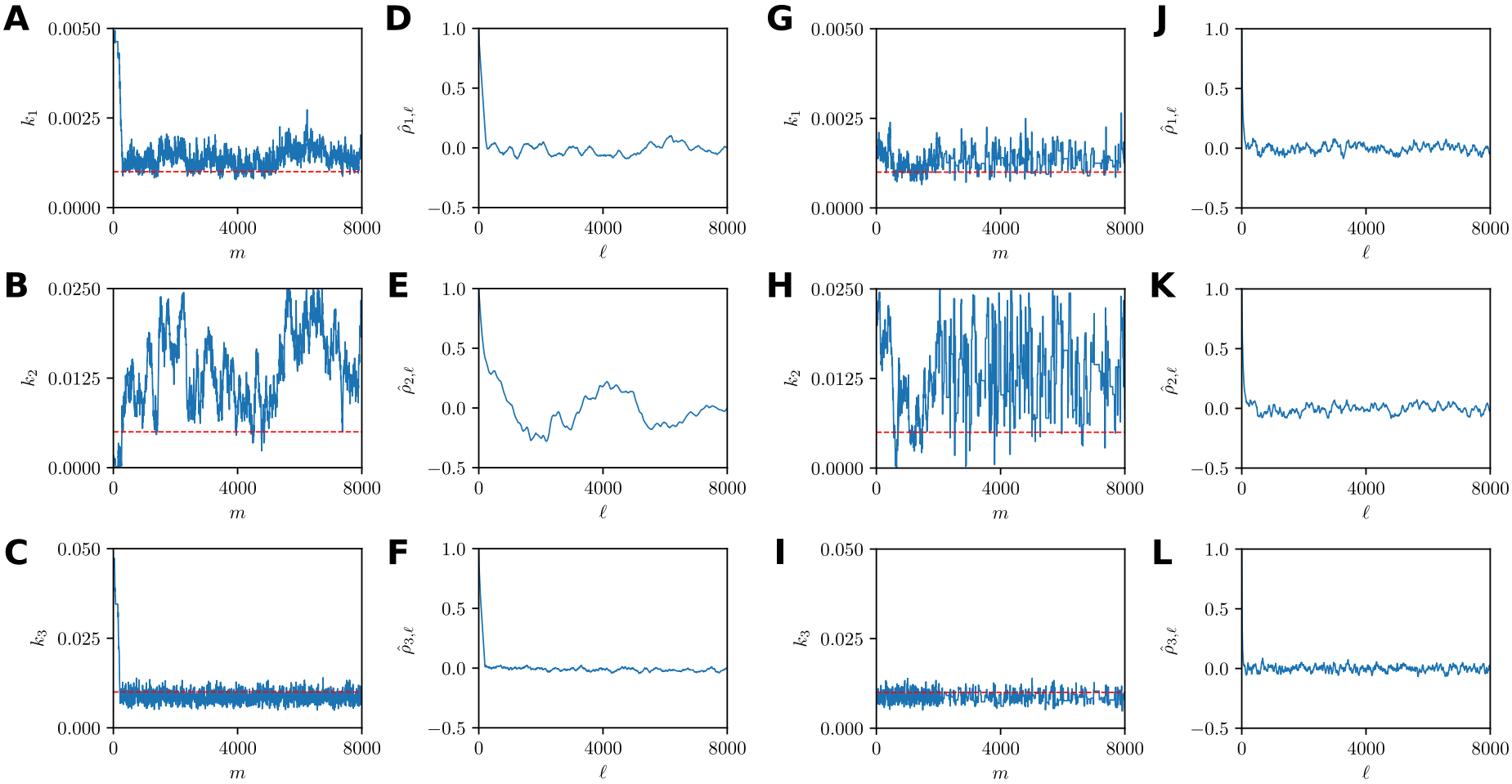

Figure 5.7: Comparison of marginal trace plots and autocorrelation functions (see Section 5.7.3) using (A)-(F) the naïve independent Gaussian proposals and (G)-(L) optimally scaled correlated proposals.

While the optimality of this scaling factor assumes a Gaussian posterior density, this is a useful guide that is widely applied (Roberts and Rosenthal, 2009). The final iteration of the trial chains is then used to initialise four new tuned chains with this optimal proposal covariance. The improvement in convergence behaviour is shown in Figure 5.7(G)–(L).

5.5.1.4 Convergence assessment and parameter estimates

Determining the number of iterations from the tuned chains to ensure valid inference is another practical challenge. Here, we follow the recommendations of Vehtari et al. (2019) and apply the rank normalised $\hat{R}$ statistic along with the multiple chain effective sample size S_{eff} (see Section 5.7.3) using the four tuned Markov chains. Informally, $\hat{R}$ represents the ratio between an estimate of the posterior variance to the average variance of each independent Markov chain, and as $\mathcal{M} \to \infty$ then $\hat{R} \to 1$ (Gelman and Rubin, 1992). The S_{eff} statistic provides a measure of effective number of i.i.d. samples that the Markov chains represent for the purposes of computing an expectation. Larger values of S_{eff} are better, but S_{eff} will typically be much smaller than $\mathcal{M}$.

The results, by parameter, are shown in Table 5.1 after $\mathcal{M} = 15,000$ iterations per chain. Vehtari et al. (2019) recommend that $\hat{R} < 1.01$ and $S_{\text{eff}} > 400$ for each parameter. We conclude that the chains have converged sufficiently for our purposes.

Table 5.1: Convergence diagnostics using four chains each with 15,000 iterations using the optimal proposal with dependent components.

	k_1	k_2	k_3
S_{eff}	986	683	1,909
$\hat{R}$	1.0046	1.0044	1.0023

The resulting inferences are shown in Table 5.2 and Figure 5.8. For all parameters, the true values are within the range of the estimates obtained in Table 5.2. The marginal posterior densities shown in Figure 5.8.

In Figure 5.8, we see that the modes of the marginal posteriors for both k_1 and k_3 are very close to the true values. However, for k_2 the mode overestimates the true parameter. It is important to emphasize, that this is not due to inaccuracy of the pseudo-marginal inference, but is a feature

Table 5.2: Parameter estimates based on estimates of the mean, $\hat{\mu}$, and standard deviation, $\hat{\sigma}$, with respect to the marginal posterior.

	k_1	k_2	k_3
$\boldsymbol{\theta}_{\text{true}}$	1.000×10^{-3}	5.000×10^{-3}	1.000×10^{-2}
$\hat{\mu}$	1.365×10^{-3}	1.381×10^{-2}	8.640×10^{-3}
$\hat{\sigma}$	2.783×10^{-4}	5.441×10^{-3}	1.441×10^{-3}

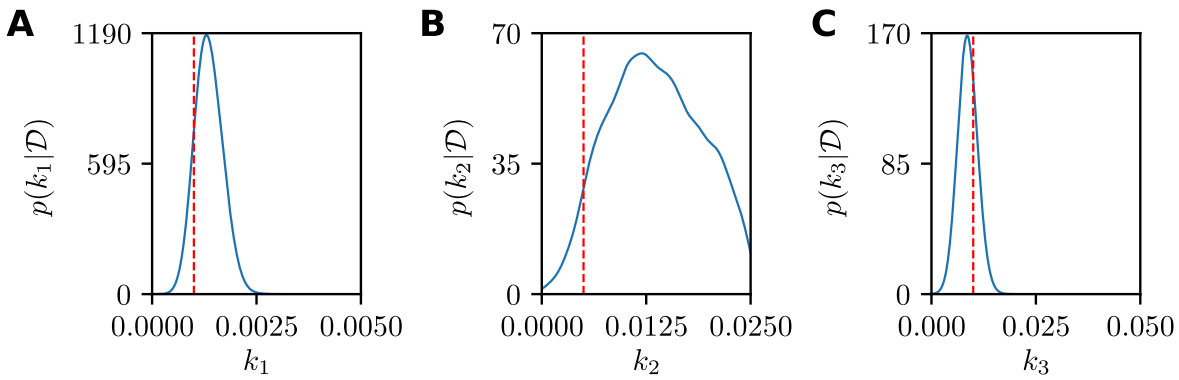

Figure 5.8: Marginal smoothed kernel density estimates for the Michaelis-Menten model using four converged particle MCMC chains. True parameter values are also indicated (red dashed).

of the true posterior density. This result effectively highlights the uncertainty in the k_2 estimate due to partial observations, observation error, and model stochasticity.

The implementation of this inference problem, including data generation, tuning and initialisation steps, convergence assessment, and plotting is given in `DemoMichMentPMCMC.jl`. The rank normalised $\hat{R}$ and S_{eff} statistics are implemented within `Diagnostics.jl`.

5.5.2 Example 2: The Schlögl model

The second example demonstrates the phenomenon of stochastic bi-stability. This leads to a very challenging inference problem that is poorly suited to alternative likelihood free schemes such as ABC.

5.5.2.1 Model definition

This example is a theoretical biochemical network initially studied by Schlögl (1972). This model involves a single chemical species, X_t, that evolves according to four reactions

$$2X \xrightarrow{k_1} 3X, \quad 3X \xrightarrow{k_2} 2X, \quad \emptyset \xrightarrow{k_3} X, \quad \text{and} \quad X \xrightarrow{k_4} \emptyset, \tag{5.20}$$

with propensities $a_1(X_t) = k_1 X_t(X_t - 1)$, $a_2(X_t) = k_2 X_t(X_t - 1)(X_t - 2)$, $a_3(X_t) = k_3$ and $a_4(X_t) = k_4 X_t$ and stoichiometries $\nu_1 = 1$, $\nu_2 = -1$, $\nu_3 = 1$ and $\nu_4 = -1$. The Chemical Langevin Itō SDE is

$$\begin{aligned} \mathrm{d}X_t = &[-k_2 X_t^3 + (k_1 + 3k_2)X_t^2 - (k_1 + 2k_2 + k_4)X_t + k_3]\mathrm{d}t \\ &+ \sqrt{k_2 X_t^3 + (k_1 - 3k_2)X_t^2 + (2k_2 - k_1 + k_4)X_t + k_2}\mathrm{d}W_t, \end{aligned} \tag{5.21}$$

where W_t is a Wiener process. For certain values of the rate parameters the underlying deterministic model has two stable steady states separated by an unstable steady state (Schlögl, 1972; Vellela and Qian, 2009). In the stochastic case, it is possible for the intrinsic noise of the system to drive X_t from around one stable state toward the other; resulting in switching behaviour demonstrated in Figure 5.9 called *stochastic bi-stability*.

The time between switching events is also a random variable, and observations taken from a single realisation will be very difficult to match using simulated data in the ABC setting, therefore acceptance rates will be prohibitively low. On the other hand, particle MCMC is ideally suited to this problem since we condition simulations on the observations, thereby only sampling from realisations that pass closely to the data.

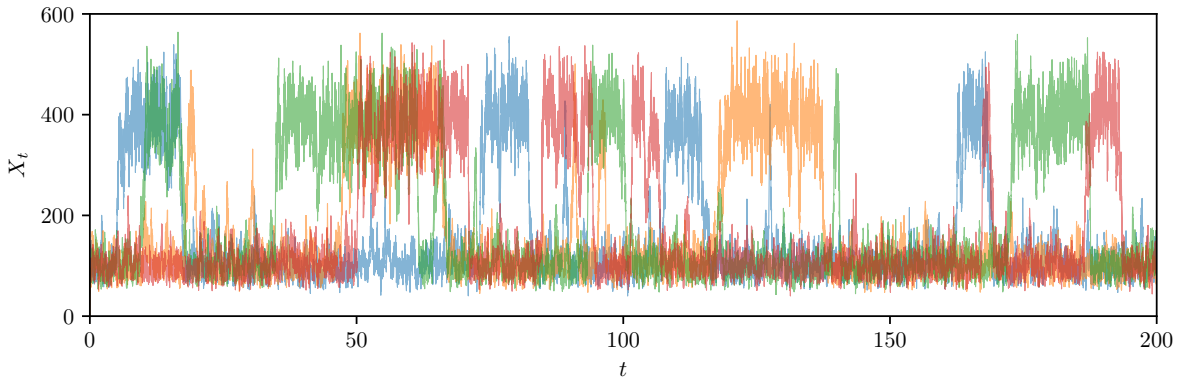

Figure 5.9: Four example realisations of the Schlögl model demonstrating stochastic bistability. The simulations are produced using the Euler-Maruyama scheme with $\Delta t = 10^{-3}$, initial condition $X_0 = 0$, and parameters $k_1 = 1.8 \times 10^{-1}$, $k_2 = 2.5 \times 10^{-4}$, $k_3 = 2.2 \times 10^3$ and $k_4 = 3.75 \times 10^1$.

5.5.2.2 Time-course data and inference problem definition

We generate synthetic data using a single realisation of the Schlögl model chemical Langevin SDE with initial condition $X_0 = 0$ and kinetic rate parameters $k_1 = 1.8 \times 10^{-1}$, $k_2 = 2.5 \times 10^{-4}$, $k_3 = 2.2 \times 10^3$ and $k_4 = 3.75 \times 10^1$. Observations are taken at $n = 16$ uniformly spaced time points $t_1 = 12.5$, $t_2 = 25, \ldots, t_{16} = 200$. The observation process is modelled by Gaussian noise applied to the chemical species copy number with a standard deviation of $\sigma_{\text{obs}} = 10$, that is, $Y_{\text{obs}}^{(i)} \sim \mathcal{N}(X_{t_i}, \sigma_{\text{obs}}^2)$. See Section 5.7.4 for the resulting data table.

We perform inference on all four rate parameters, $\boldsymbol{\theta} = [k_1, k_2, k_3, k_4]^{\mathrm{T}}$. We use the particle MCMC approach to sample for the Bayesian posterior,

$$p(k_1, k_2, k_3, k_4 \mid \mathcal{D}) \propto \mathcal{L}(k_1, k_2, k_{3,4}\,; \mathcal{D}) p(k_1, k_2, k_3, k_4),$$

where $p(k_1, k_2, k_3, k_4)$ is the joint uniform prior with independent components $k_1 \sim \mathcal{U}(0, 5.4 \times 10^{-1})$, $k_2 \sim \mathcal{U}(0, 7.5 \times 10^{-4})$, $k_3 \sim \mathcal{U}(0, 6.6 \times 10^3)$, and $k_4 \sim \mathcal{U}(0, 1.125 \times 10^2)$. The likelihood is estimated using the bootstrap particle filter (Algorithm 5.2) with $R = 100$ particles and Euler-Maruyama for simulation with $\Delta t = 0.1$.

5.5.2.3 Chain initialisation and proposal tuning

To initialise and tune four chains for inference on the four rate parameters of the Schlögl model, we apply the same procedure as described for the Michaelis-Menten inference problem. The only difference is the number of samples applied.

Firstly, four trial chains are simulated for $\mathcal{M} = 20,000$ iterations, each of these chains is initialised with a random sample from the prior with a non-zero likelihood estimate. The proposal kernel used in all four trial chains is Gaussian with covariance

$$\boldsymbol{\Sigma} = \begin{bmatrix} 2.430 \times 10^{-4} & 0.0 & 0.0 & 0.0 \\ 0.0 & 4.688 \times 10^{-10} & 0.0 & 0.0 \\ 0.0 & 0.0 & 3.63 \times 10^{4} & 0.0 \\ 0.0 & 0.0 & 0.0 & 1.055 \times 10^{1} \end{bmatrix}.$$

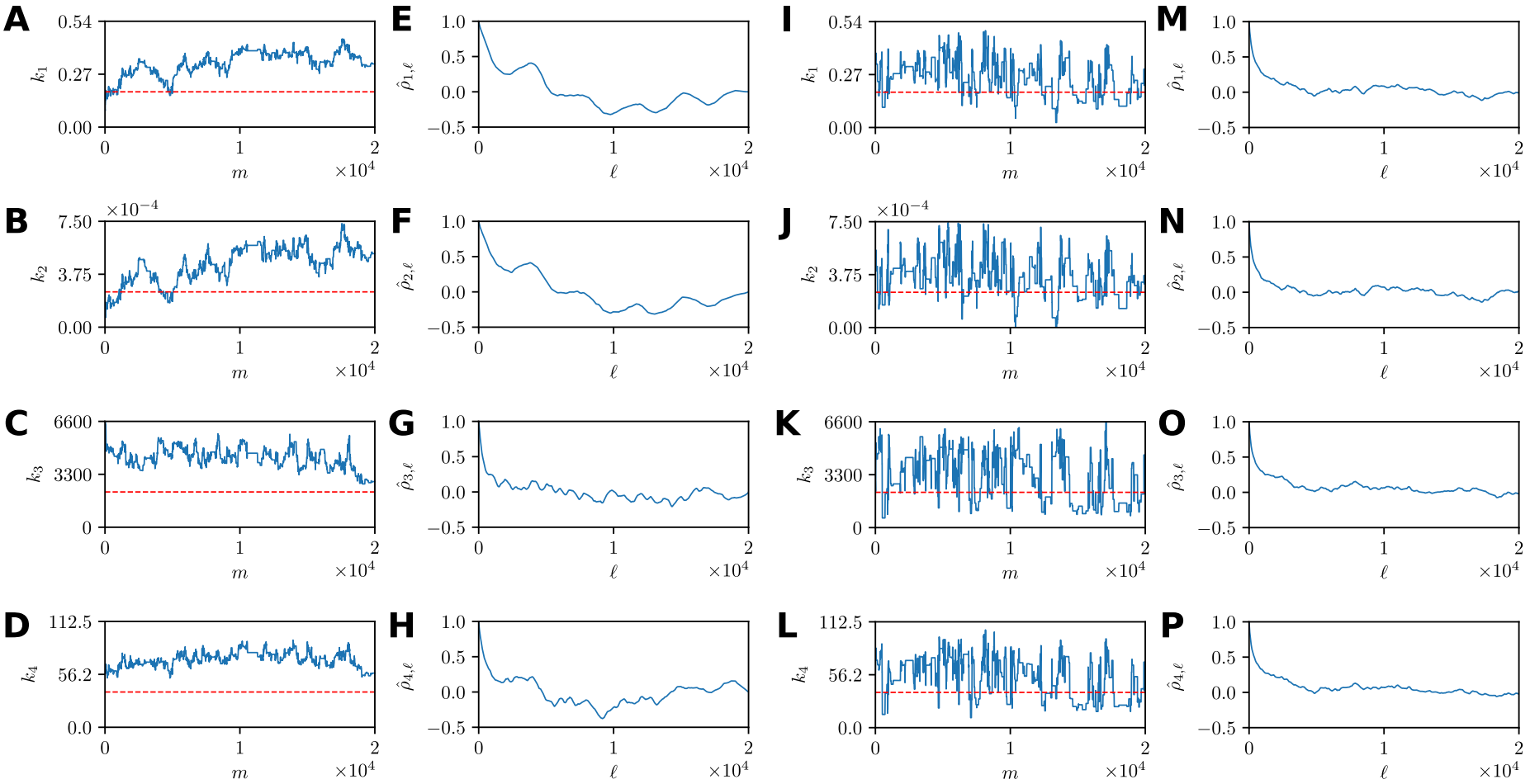

Figure 5.10: Comparison of marginal trace plots and autocorrelation functions (see Section 5.7.3) using (A)–(H) the naïve independent Gaussian proposals and (I)-(P) optimally scaled correlated proposals.

Again, we start with a typical independent proposal kernel with diagonal entries calculated so that one proposal standard deviation in each parameter corresponds to one tenth the prior standard deviation. The tuned proposal kernel is constructed by taking the convariance of the pooled sample of the four trial chains (a total of $80,000$ samples),

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} 4.770 \times 10^{-3} & 6.995 \times 10^{-4} & 4.429 \times 10^{1} & 9.297 \times 10^{-1} \\ 6.994 \times 10^{-6} & 1.511 \times 10^{-8} & 1.640 \times 10^{-3} & 5.929 \times 10^{-4} \\ 4.429 \times 10^{1} & 1.640 \times 10^{-3} & 1.621 \times 10^{6} & 2.061 \times 10^{4} \\ 9.297 \times 10^{-1} & 5.929 \times 10^{-4} & 2.061 \times 10^{4} & 3.199 \times 10^{2} \end{bmatrix},$$

and applying the optimal scaling rule from Roberts and Rosenthal (2001)

$$\boldsymbol{\Sigma}_{\text{opt}} = \frac{2.38^2}{4}\hat{\boldsymbol{\Sigma}}.$$

The final iteration of the trial chains is then used to initialise four new chains using this optimal proposal covariance. Figure 5.10 demonstrates the improvement in convergence behaviour.

5.5.2.4 Convergence assessment and parameter estimates

Convergence diagnostic results, by parameter, are shown in Table 5.3 after $\mathcal{M} = 240,000$ iterations per chain. Again, we ensure that the criteria of $\hat{R} < 1.01$ and $S_{\text{eff}} > 400$ (Vehtari et al., 2019) are satisfied for all parameters. We note that the convergence rate of the MCMC chains is significantly slower than that of the Michaelis-Menten example. In practice, one might consider a fully adaptive proposal scheme for this model to improve convergence rates (Roberts and Rosenthal, 2001, 2009).

Table 5.3: Convergence diagnostics using four chains each with $240,000$ iterations using the optimal proposal with dependent components.

	k_1	k_2	k_3	k_4
S_{eff}	577	656	467	625
$\hat{R}$	1.0054	1.0060	1.0049	1.0039

The resulting inferences are shown in Table 5.4 and Figure 5.11. For all parameters, the true values are within the range of the estimates obtained in Table 5.4. The marginal posterior densities shown in Figure 5.11.

Figure 5.11(A)–(B) demonstrates that both the posterior modes for k_1 and k_2 are very close to the true parameter values, however the uncertainties are asymmetric, indicating a range of possibly appropriate parameter values greater than the true values. The posteriors of k_3 and k_4 are very interesting as they are bi-modal (Figure 5.11(C)–(D)). While the higher density model is closer to the true parameter values, the second lower density mode indicates that an alternative parameter combination in k_3 and k_4 can lead to very similar stochastic bi-stability in the Schlögl model evolution. To observe this posterior bi-modality using ABC methods would be very challenging since ϵ would need to be prohibitively small. Furthermore, most expositions on ABC methods (Sunnåker et al., 2013; Toni et al., 2009)(Chapter 4) do not deal with multimodal posteriors.

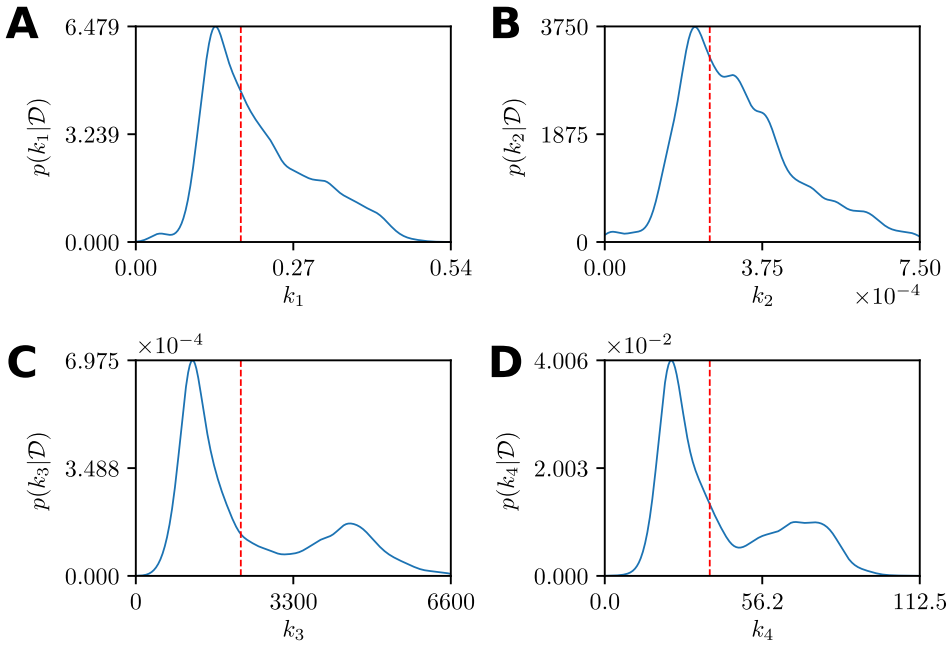

Figure 5.11: Marginal smoothed kernel density estimates for the Schlögl model using four converged particle MCMC chains. True parameter values are also indicated (red dashed).

The apparent bi-modal nature of parameters k_3 and k_4 (Figure 5.11) could be a reason for the increased computational requirements of this model, since all chains must occupy both modes sufficiently to reduce $\hat{R}$ and to increase S_{eff} sufficiently.

Table 5.4: Parameter estimates based on estimates of the mean, $\hat{\mu}$, and standard deviation, $\hat{\sigma}$, with respect to the marginal posterior.

	k_1	k_2	k_3	k_4
$\boldsymbol{\theta}_{\text{true}}$	1.800×10^{-1}	2.500×10^{-4}	2.200×10^{3}	3.750×10^{1}
$\hat{\mu}$	2.118×10^{-1}	3.200×10^{-4}	2.350×10^{3}	4.104×10^{1}
$\hat{\sigma}$	8.988×10^{-2}	1.377×10^{-4}	1.483×10^{3}	2.121×10^{1}

The implementation of this inference problem, including data generation, tuning and initialisation steps, convergence assessment, and plotting is give in `DemoSchloglPMCMC.jl`. The rank normalised $\hat{R}$ and S_{eff} statistics are implemented within `Diagnostics.jl`.

5.5.3 Example 3: The repressilator model

The last example we considered in this work is a gene regulatory network, originally realised synthetically by Elowitz and Leibler (2000), that includes a feedback loop resulting in stochastic oscillatory dynamics in the gene expression. The model is of interest in biological studies (Pokhilko et al., 2012; Potvin-Trottier et al., 2016) and is a challenging benchmark for inference methods (Toni et al., 2009).

5.5.3.1 Model definition

The repressilator consists of three genes where the expression of one gene inhibits the expression of the next gene, forming a feedback loop between the three genes. The regulatory consists of twelve reactions describing the transcription of the three mRNAs, M_1, M_2, and M_3, associated with each gene, G_1, G_2, and G_3, their expression through translation into proteins, P_1, P_2, and P_3, and decay processes for both mRNAs and proteins. For the ith gene we have,

$$\underbrace{G_i \xrightarrow{\alpha_0+\alpha/(1+P_j^n)} M_i}_{\text{mRNA transcription}}, \quad \underbrace{M_i \xrightarrow{\beta} M_i + P_i}_{\text{protein translation}}, \quad \underbrace{P_i \xrightarrow{\beta} \emptyset}_{\text{protein degradation}} \quad , \quad \text{and} \quad \underbrace{M_i \xrightarrow{\gamma} \emptyset}_{\text{mRNA degradation}} \quad , \tag{5.22}$$

where $j = (i + 1 \mod 3) + 1$, α_0 is the leakage transcription rate (transcription rate of maximally inhibited gene), $\alpha + \alpha_0$ is the free transcription rate (uninhibited transcription rate), n is the Hill coefficient that describes the strength of the repressive effect of the inhibitor protein P_j, β is the protein translation and degradation rate, and γ is the mRNA degradation

rate (Elowitz and Leibler, 2000). The gene copy numbers are fixed at $G_i = 1$ for $i = 1, 2, 3$. The resulting chemical Langevin approximation (Equation (5.2)) to the repressilator model (Equation (5.22)) leads to a coupled system of Itō SDEs

$$\begin{aligned}
\mathrm{d}M_{1,t} &= \left(\alpha_0 + \frac{\alpha}{1 + P_{3,t}^n} - \gamma M_{1,t}\right)\mathrm{d}t + \sqrt{\alpha_0 + \frac{\alpha}{1 + P_{3,t}^n}}\mathrm{d}W_t^{(1)} - \sqrt{\gamma M_{1,t}}\mathrm{d}W_t^{(4)}, \\
\mathrm{d}P_{1,t} &= \beta\left(M_{1,t} - P_{1,t}\right)\mathrm{d}t + \sqrt{\beta M_{1,t}}\mathrm{d}W_t^{(2)} - \sqrt{\beta P_{1,t}}\mathrm{d}W_t^{(3)}, \\
\mathrm{d}M_{2,t} &= \left(\alpha_0 + \frac{\alpha}{1 + P_{1,t}^n} - \gamma M_{2,t}\right)\mathrm{d}t + \sqrt{\alpha_0 + \frac{\alpha}{1 + P_{1,t}^n}}\mathrm{d}W_t^{(5)} - \sqrt{\gamma M_{2,t}}\mathrm{d}W_t^{(8)}, \\
\mathrm{d}P_{2,t} &= \beta\left(M_{2,t} - P_{2,t}\right)\mathrm{d}t + \sqrt{\beta M_{2,t}}\mathrm{d}W_t^{(6)} - \sqrt{\beta P_{2,t}}\mathrm{d}W_t^{(7)}, \\
\mathrm{d}M_{3,t} &= \left(\alpha_0 + \frac{\alpha}{1 + P_{2,t}^n} - \gamma M_{3,t}\right)\mathrm{d}t + \sqrt{\alpha_0 + \frac{\alpha}{1 + P_{2,t}^n}}\mathrm{d}W_t^{(9)} - \sqrt{\gamma M_{3,t}}\mathrm{d}W_t^{(12)}, \\
\mathrm{d}P_{3,t} &= \beta\left(M_{3,t} - P_{3,t}\right)\mathrm{d}t + \sqrt{\beta M_{3,t}}\mathrm{d}W_t^{(10)} - \sqrt{\beta P_{3,t}}\mathrm{d}W_t^{(11)},
\end{aligned} \tag{5.23}$$

where $W_t^{(1)}, W_t^{(2)}, \ldots, W_t^{(12)}$ are independent Wiener processes driving each reaction channel. Certain parameter combinations lead to stochastic oscillations in the gene expression levels, that is, the protein copy numbers associated with the expressed gene. Figure 5.12 demonstrates this behaviour in which the expressed gene alternates between G_2, G_1, and G_3 in sequence due to the feedback loop in the gene inhibitor network.

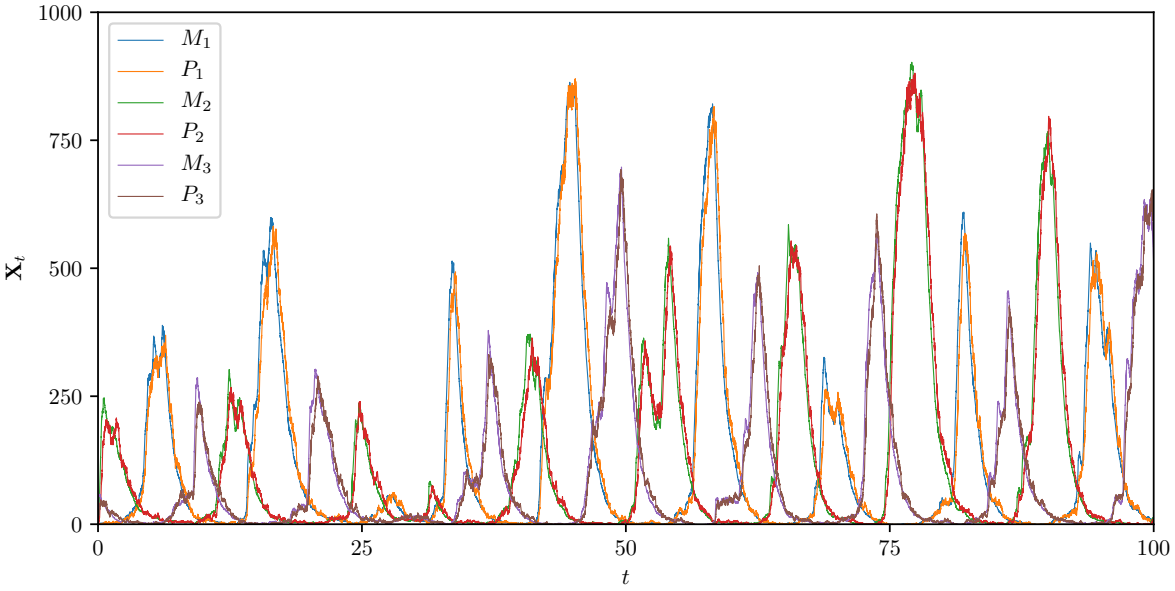

Figure 5.12: Example realisation of the repressilator model demonstrating oscillatory gene expression. The simulation is produced using the Euler-Maruyama scheme with $\Delta t = 1 \times 10^{-3}$, initial condition $\mathbf{X}_0 = [M_{1,0}, P_{1,0}, M_{2,0}, P_{2,0}, M_{3,0}, P_{3,0}]^{\mathrm{T}} = [0, 2, 0, 1, 0, 3]^{\mathrm{T}}$, and parameters $\alpha_0 = 1$, $\alpha = 1000$, $n = 2$, $\beta = 5$, and $\gamma = 1$.

5.5.3.2 Time-course data and inference problem definition

We generate synthetic data using a single realisation of the repressilator chemical Langevin SDE with initial condition $\mathbf{X}_0 = [M_{1,0}, P_{1,0}, M_{2,0}, P_{2,0}, M_{3,0}, P_{3,0}]^{\mathrm{T}} = [0, 2, 0, 1, 0, 3]^{\mathrm{T}}$ and parameters $\alpha_0 = 1$, $\alpha = 1000$, $n = 2$, $\beta = 5$, and $\gamma = 1$. Observations are taken at $n = 20$ uniformly spaced time points $t_1 = 5$, $t_2 = 10$, $\ldots, t_{20} = 100$. Again we consider Gaussian noise applied to each chemical species copy number with a standard deviation of $\sigma_{\text{obs}} = 10$, that is, $\mathbf{Y}_{\text{obs}}^{(i)} \sim \mathcal{N}(\mathbf{X}_{t_i}, \sigma_{\text{obs}}^2 \mathbf{I})$ where $\mathbf{I}$ is the 6×6 identity matrix. See Section 5.7.4 for the resulting data table.

We perform inference on four of the model parameters, $\boldsymbol{\theta} = [\alpha_0, \alpha, n, \beta]^{\mathrm{T}}$, and fix the mRNA degradation rate $\gamma = 1$. We use the particle MCMC approach to sample from the Bayesian posterior,

$$p(\alpha_0, \alpha, n, \beta \mid \mathcal{D}) \propto \mathcal{L}(\alpha_0, \alpha, n, \beta; \mathcal{D}) p(\alpha_0, \alpha, n, \beta),$$

where $p(\alpha_0, \alpha, n, \beta)$ is the joint uniform prior with independent components $\alpha_0 \sim \mathcal{U}(0, 10)$, $\alpha \sim \mathcal{U}(500, 2500)$, $n \sim \mathcal{U}(0, 10)$, and $\beta \sim \mathcal{U}(0, 20)$.

5.5.3.3 Chain initialisation and proposal tuning

The same initialisation and proposal tuning procedure applied to the Michaelis-Menten and Schlögl models is applied here. The resulting tuned proposal kernel convariance is given by

$$\boldsymbol{\Sigma}_{\text{opt}} = \frac{2.38^2}{4} \begin{bmatrix} 43634.288 & 95.328 & 173.584 & 416.368 \\ 95.328 & 0.345 & 1.051 & 1.142 \\ 173.584 & 1.051 & 5.286 & 3.488 \\ 416.368 & 1.142 & 3.488 & 5.939 \end{bmatrix},$$

which is derived through application of the Roberts and Rosenthal (2001) scaling rule to the covariance matrix of the pooled samples from four trial chains, each with $\mathcal{M} = 5,000$ iterations. The trial chains are initialised and constructed in the same way as for the Michaelis-Menten and Schlögl models.

The repressilator model is a good example of when one must be careful to use a large enough number of particles in the bootstrap particle filter. Unlike the Michaelis-Menten and Schlögl models, the repressilator model likelihood estimator is highly variable for low particle numbers. Figure 5.13 demonstrates the effect of the number of particles, R, on the distribution of the logarithm of the likelihood estimator evaluated $Z = \log \hat{\mathcal{L}}(\boldsymbol{\theta}; \mathcal{D})$.

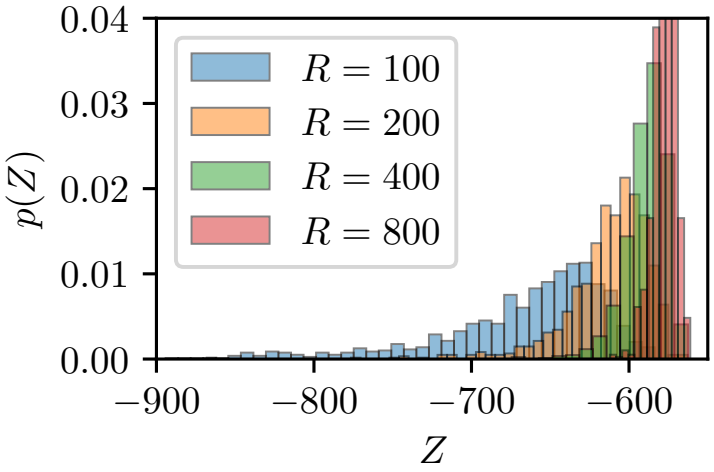

Figure 5.13: Distribution of $1,000$ likelihood estimates around a high density posterior point for different numbers of particles in the bootstrap particle filter. Significant bias is introduced for the lower particle counts.

Note that as R decreases, not only does the variance of the estimator increase, but so does the bias that is seen through the shift in the estimator mode. Here, there is a trade-off, $R = 800$ yields a low variance and is much closer to the optimal criterion of Doucet et al. (2015). However, $R = 400$ has a very similar mode, but slightly higher variance. This motivates the use of $R = 400$ particles to achieve reasonable convergences rates without too much additional computational burden.

5.5.3.4 Convergence assessment and parameter estimates

Convergence diagnostic results, by parameter, are shown in Table 5.5 after $\mathcal{M} = 145,000$ iterations per chain. In this case, the conservative convergence criteria of Vehtari et al. (2019) have not yet been met. We report the results without additional computational effort for the purposes of this chapter, but we emphasise that for a real application more iterations of the MCMC chains should be performed to have confidence in the final inferences. Furthermore,

it is important to note that $\hat{R} < 1.1$ is still a widely used convergence criterion (Gelman and Rubin, 1992; Gelman et al., 2014).

Table 5.5: Convergence diagnostics using four chains each with $145,000$ iterations using the optimal proposal with dependent components.

	α	α_0	β	n
S_{eff}	102	107	152	205
$\hat{R}$	1.0358	1.0455	1.0152	1.0184

The resulting inferences are shown in Figure 5.14 and Table 5.6. For all parameters, the true values are within range of the estimates obtained in Table 5.6. The marginal posterior densities are shown in Figure 5.14.

Table 5.6: Parameter estimates based on estimates of the mean, $\hat{\mu}$, and standard deviation, $\hat{\sigma}$, with respect to the marginal posterior.

	α	α_0	β	n
$\boldsymbol{\theta}_{\text{true}}$	1000.0000	1.0000	5.0000	2.0000
$\hat{\mu}$	871.4327	0.6224	4.7106	1.9960
$\hat{\sigma}$	172.9650	0.4685	1.1671	0.2553

For all parameters, the marginal posterior densities are uni-model, with modes that are close to the true parameter estimates. In particular, the β and n parameters are very accurately retrieved, whereas the marginal posteriors for α and α_0 lead to underestimates. However, these underestimates are consistent with previous results (Toni et al., 2009). A likely cause of this is the temporal sparsity of observations, leading to few observations of the peak gene expression levels (see Figure 5.12 and data in Section 5.7.4), as α and α_0 relate to the transcription rate of the mRNAs. Despite the additional computational complexity associated with the particle filter for this inference problem, the repressilator model provides an insightful example of the efficacy of the pseudo-marginal approach to resolve biological parameters associated with gene regulation using synthetic data that is biological realisable.

The implementation of this inference problem, including data generation, tuning and initialisation steps, convergence assessment, and plotting is give in `DemoRepressilatorPMCMC.jl`. The rank normalised $\hat{R}$ and S_{eff} statistics are implemented within `Diagnostics.jl`.

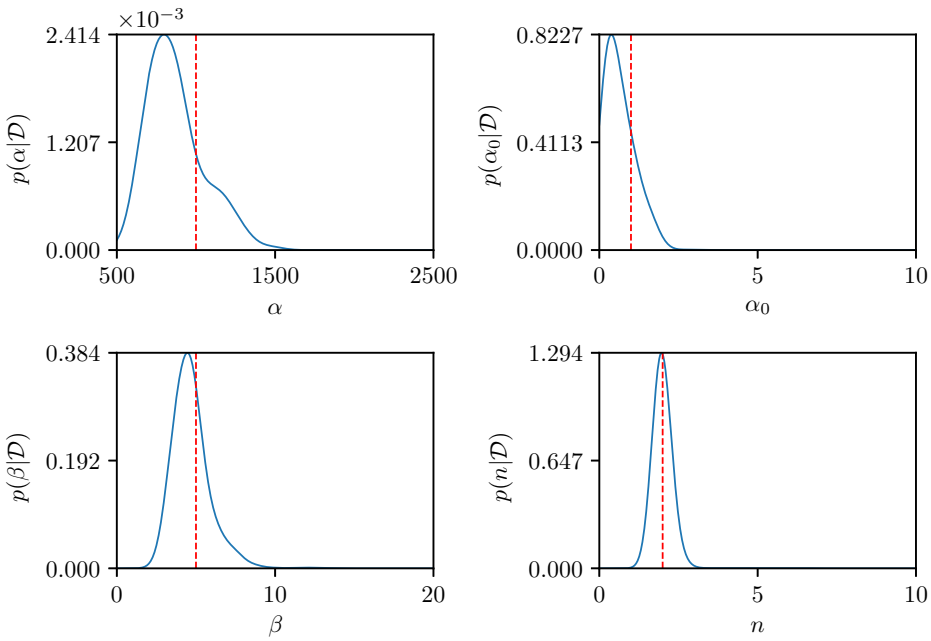

Figure 5.14: Marginal smoothed kernel density estimates for the repressilator model using four particle MCMC chains. True parameter values are also indicated (red dashed).

5.6 Summary

In this work, we provide a practical guide to computational Bayesian inference using the pseudo-marginal approach (Andrieu and Roberts, 2009; Beaumont, 2003; Andrieu et al., 2010). We compare and contrast, using a tractable example, the pseudo-marginal approach with the ABC alternative (Sisson et al., 2018; Sunnåker et al., 2013). Throughout, chemical Langevin SDE descriptions of biochemical reaction networks (Gillespie, 2000; Higham, 2008), of various degrees of complexity, have been employed to demonstrate practical considerations when using these techniques to inference problems with intractable likelihoods.

The ABC approach to likelihood-free inference is widely applicable and used extensively in practical applications (Browning et al., 2018; Johnston et al., 2016; Kursawe et al., 2018; Wilkinson, 2011)(Chapter 3). For some applications, however, it can be difficult to determine *a priori* an appropriate discrepancy metric and acceptance threshold for reliable inference. Furthermore, a sufficiently small threshold for the desired level of accuracy may result in prohibitively low acceptance rates (Sisson et al., 2007). Pseudo-marginal methods do not suffer from these accuracy considerations since they converge to the true posterior target regardless of the variance of the estimator (Golightly and Wilkinson, 2011). As a result, the pseudo-marginal

approach is significantly less sensitive to user-specified algorithm parameters than likelihood-free inference based on ABC.

There are also disadvantages to the pseudo-marginal approach. Firstly, it is not as generally applicable as ABC; the pseudo-marginal method requires an unbiased estimator, whereas ABC only needs a model simulation process. While convergence to the true posterior distribution is not affected by the estimator variance, the rate of convergence is (Andrieu and Roberts, 2009); to obtain optimal likelihood variances, a large number of particles may be required, thus evaluating the likelihood estimate will be very expensive. Alternatively, ABC will only ever use a single simulation per iteration. Furthermore, under the assumption of observation error and model miss-specification, convergence to the true posterior is not always a significant advantage (Wilkinson, 2009; Andrieu et al., 2018) and ABC may be effectively considered exact (Wilkinson, 2013). Lastly, pseudo-marginal methods are not widespread in the systems biology literature and there is a lack of exemplars, despite their suitability for many problems of interest. This chapter is intended to address this by presenting all the steps involved clearly and providing user-friendly implementations in an open access environment (https://github.com/davidwarne/Warne2019_GuideToPseudoMarginal).

For practical illustrative purposes, we focus on the fundamental method of particle marginal Metropolis-Hastings (Andrieu et al., 2010) using the bootstrap particle filter (Gordon et al., 1993) for likelihood estimation. There are many other variants to this classic approach, such as particle Gibbs sampling (Andrieu et al., 2010; Doucet et al., 2015), coupled Markov chains (Dodwell et al., 2015, 2019), and more advanced particle filters (Doucet and Johanson, 2011) and proposal mechanisms (Botha et al., 2019; Cotter et al., 2013). It is also important to note that the pseudo-marginal approach is equally valid for Bayesian sampling strategies based on sequential Monte Carlo (Del Moral et al., 2006; Sisson et al., 2007; Li et al., 2019). Furthermore, advances in stochastic simulation (Schnoerr et al., 2017)(Chapter 4) can also improve the performance of the likelihood estimator, and the application of multilevel Monte Carlo to particle filters can further reduce estimator variance (Jasra et al., 2017, 2018; Gregory et al., 2016).

Likelihood-free methods are essential to modern science, since many mechanistic models of interest have intractable likelihoods. Unlike ABC methods, the pseudo-marginal approach does not affect the stationary distribution for the purposes of MCMC sampling; this is a desirable property. However, one reason for the popularity and success of ABC methods has been

its simplicity to implement. Through this accessible and practical demonstration, along with example open-source codes, the pseudo-marginal approach may become an additional readily available tool for likelihood-free inference within the wider scientific community.

5.7 Supplementary Material

5.7.1 Derivation of stationary distribution for the production-degradation model

Here, we derive the solution to the stationary distribution for the production-degradation model. First, recall that the general form for the chemical Langevin equation is

$$\mathrm{d}\mathbf{X}_t = \sum_{j=1}^{M} \boldsymbol{\nu}_j a_j(\mathbf{X}_t)\mathrm{d}t + \sum_{j=1}^{M} \boldsymbol{\nu}_j \sqrt{a_j(\mathbf{X}_t)}\mathrm{d}W_t^{(j)},$$

where $\mathbf{X}_t$ takes values in $\mathbb{R}^N$, $W_t^{(1)}, W_t^{(2)}, \ldots, W_t^{(M)}$ are independent scalar Wiener processes, $\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \ldots, \boldsymbol{\nu}_M$ are the stoichiometric vectors and $a_1(\mathbf{X}_t), a_2(\mathbf{X}_t), \ldots, a_M(\mathbf{X}_t)$ the propensity functions. Consider the distribution of $\mathbf{X}_t$ over all possible realisations at time t with probability density function $p(\mathbf{x}, t)$. The Fokker-Planck equation describes the forward evolution of this probability density in time. For the general chemical Langevin equation, the Fokker-Planck equation is given by

$$\frac{\partial p(\mathbf{x}, t)}{\partial t} = \frac{1}{2} \sum_{j=1}^{M} \boldsymbol{\nu}_j^{\mathrm{T}} \mathbf{H}\left(a_j(\mathbf{x})p(\mathbf{x}, t)\right) \boldsymbol{\nu}_j - \sum_{j=1}^{M} \nabla\left[a_j(\mathbf{x})p(\mathbf{x}, t)\right] \boldsymbol{\nu}_j, \tag{5.24}$$

where, for reaction j, $\nabla\left[a_j(\mathbf{x})p(\mathbf{x}, t)\right]$ and $\mathbf{H}\left(a_j(\mathbf{x})p(\mathbf{x}, t)\right)$ are, respectively, the gradient vector and Hessian matrix of the product $a_j(\mathbf{x})p(\mathbf{x}, t)$ with respect to the state vector $\mathbf{x}$.

For the production-degradation model, we have a single chemical species X_t, propensity functions

$$a_1(X_t) = k_1 \quad \text{and} \quad a_2(X_t) = k_2 X_t, \tag{5.25}$$

with rate parameters k_1 and k_2, and stoichiometries

$$\nu_1 = 1 \quad \text{and} \quad \nu_2 = -1. \tag{5.26}$$

By substituting Equation (5.25) and Equation (5.26) into Equation (5.24), we obtain the Fokker-Planck equation for the production-degradation model,

$$\frac{\partial p(x, t)}{\partial t} = \frac{\partial^2}{\partial x^2}\left[\frac{k_1 + k_2 x}{2} p(x, t)\right] - \frac{\partial}{\partial x}\left[(k_1 - k_2 x)p(x, t)\right]. \tag{5.27}$$

The stationary distribution of X_t corresponds to the steady state solution of Equation (5.27), that is, $p_s(x) = \lim_{t\to\infty} p(x,t)$. The stationary probability density function, $p_s(x)$, satisfies

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2}\left[\frac{k_1 + k_2 x}{2} p_s(x)\right] - \frac{\mathrm{d}}{\mathrm{d}x}\left[(k_1 - k_2 x) p_s(x)\right] = 0. \tag{5.28}$$

To obtain a solution, integrate Equation (5.28) to obtain

$$\frac{\mathrm{d}p_s(x)}{\mathrm{d}x} + \left(\frac{k_2(1-x) - 2k_1}{k_1 + k_2 x}\right) p_s(x) = C, \tag{5.29}$$

where C is an arbitrary constant. We obtain $C = 0$ by assuming the boundary condition $\lim_{x\to\infty} p_s(x) = 0$. The solution to Equation (5.29) can be obtained using an integrating factor,

$$\begin{aligned} p_s(x) &= \frac{\tilde{C}}{k_1 + k_2 x} \exp\left(2\int_0^x \frac{k_1 - k_2 y}{k_1 + k_2 y}\,\mathrm{d}y\right) \\ &= \tilde{C} \exp\left(-2x + \left(\frac{4k_1}{k_2} - 1\right)\ln\left(k_1 + k_2 x\right)\right), \end{aligned} \tag{5.30}$$

where $\tilde{C}$ is a constant that is obtained by enforcing the condition $\int_{-\infty}^{\infty} p_s(x)\,\mathrm{d}x = 1$, yielding the stationary probability density as given in Equation (5.8).

5.7.2 Pseudo-marginal MCMC as an exact approximation

Here we briefly explain why the stationary distribution of the pseudo-marginal Metropolis-Hastings method is the exact posterior. For more detailed analysis, see Andrieu and Roberts (2009), and Golightly and Wilkinson (2008). Consider the following algebraic manipulations applied to the pseudo-marginal Metropolis-Hastings acceptance probability. We start with

$$\begin{aligned} \alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}_m) &= \frac{q(\boldsymbol{\theta}_m \mid \boldsymbol{\theta}^*)\hat{\mathcal{L}}(\boldsymbol{\theta}^*; \mathcal{D}) p(\boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}_m)\hat{\mathcal{L}}(\boldsymbol{\theta}_m; \mathcal{D}) p(\boldsymbol{\theta}_m)} \\ &= \frac{q(\boldsymbol{\theta}_m \mid \boldsymbol{\theta}^*)\mathcal{L}(\boldsymbol{\theta}^*; \mathcal{D})\left[\dfrac{\hat{\mathcal{L}}(\boldsymbol{\theta}^*; \mathcal{D})}{\mathcal{L}(\boldsymbol{\theta}^*; \mathcal{D})}\right] p(\boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}_m)\mathcal{L}(\boldsymbol{\theta}_m; \mathcal{D})\left[\dfrac{\hat{\mathcal{L}}(\boldsymbol{\theta}_m; \mathcal{D})}{\mathcal{L}(\boldsymbol{\theta}_m; \mathcal{D})}\right] p(\boldsymbol{\theta}_m)}. \end{aligned}$$

Now, define the random variable $Z = \hat{\mathcal{L}}(\mathcal{D};\boldsymbol{\theta})/\mathcal{L}(\mathcal{D};\boldsymbol{\theta})$ with density $p(Z \mid \boldsymbol{\theta})$ that represents a scaled likelihood estimator. We apply this change of variable and perform some straightforward algebra to obtain a new representation for $\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}_m)$ that reveals some interesting structure:

$$\begin{aligned}\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}_m) &= \frac{q(\boldsymbol{\theta}_m \mid \boldsymbol{\theta}^*)\mathcal{L}(\mathcal{D};\boldsymbol{\theta}^*)Z^* p(\boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}_m)\mathcal{L}(\mathcal{D};\boldsymbol{\theta}_m)Zp(\boldsymbol{\theta}_m)} \\ &= \frac{p(Z^* \mid \boldsymbol{\theta}^*)p(Z_m \mid \boldsymbol{\theta}_m)}{p(Z^* \mid \boldsymbol{\theta}^*)p(Z_m \mid \boldsymbol{\theta}_m)} \times \frac{q(\boldsymbol{\theta}_m \mid \boldsymbol{\theta}^*)\mathcal{L}(\mathcal{D};\boldsymbol{\theta}^*)Z^* p(\boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}_m)\mathcal{L}(\mathcal{D};\boldsymbol{\theta}_m)Zp(\boldsymbol{\theta}_m)} \\ &= \frac{p(Z_m \mid \boldsymbol{\theta}_m)q(\boldsymbol{\theta}_m \mid \boldsymbol{\theta}^*)\left[\mathcal{L}(\mathcal{D};\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)Z^* p(Z^* \mid \boldsymbol{\theta}^*)\right]}{p(Z^* \mid \boldsymbol{\theta}^*)q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}_m)\left[\mathcal{L}(\mathcal{D};\boldsymbol{\theta}_m)p(\boldsymbol{\theta}_m)Zp(Z_m \mid \boldsymbol{\theta}_m)\right]}. \end{aligned} \tag{5.31}$$

The expressions outside the brackets can be considered a proposal density, $q(Z^*, \boldsymbol{\theta}^* \mid Z_m, \boldsymbol{\theta}_m) = p(Z^* \mid \boldsymbol{\theta}^*)q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}_m)$, in the product state space $\mathcal{Z} \times \boldsymbol{\Theta}$ where $\mathcal{Z} \subset \mathbb{R}^+$ is the space of values Z can take. By extending the dimension of the Markov chain state by including Z_m, we see that acceptance probability for the original pseudo-marginal Markov chain in $\boldsymbol{\Theta}$, as given in Equation (5.31), can also be considered as an acceptance probability for this new Markov chain in $\mathcal{Z} \times \boldsymbol{\Theta}$ based on exact Metropolis-Hastings MCMC. That is,

$$\alpha((Z^*, \boldsymbol{\theta}^*), (Z_m, \boldsymbol{\theta}_m)) = \frac{q(Z_m, \boldsymbol{\theta}_m \mid Z^*, \boldsymbol{\theta}^*)\left[\mathcal{L}(\mathcal{D};\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)Z^* p(Z^* \mid \boldsymbol{\theta}^*)\right]}{q(Z^*, \boldsymbol{\theta}^* \mid Z_m, \boldsymbol{\theta}_m)\left[\mathcal{L}(\mathcal{D};\boldsymbol{\theta}_m)p(\boldsymbol{\theta}_m)Zp(Z_m \mid \boldsymbol{\theta}_m)\right]}.$$

This Markov chain has the stationary distribution

$$\begin{aligned} p(Z, \boldsymbol{\theta}) &= \mathcal{L}(\mathcal{D};\boldsymbol{\theta})p(\boldsymbol{\theta})Zp(Z \mid \boldsymbol{\theta}) \\ &\propto p(\boldsymbol{\theta} \mid \mathcal{D})Zp(Z \mid \boldsymbol{\theta}). \end{aligned}$$

Integrating out Z we obtain

$$\begin{aligned}\int_{\mathcal{Z}} p(\boldsymbol{\theta} \mid \mathcal{D})Zp(Z \mid \boldsymbol{\theta})\,\mathrm{d}Z &= p(\boldsymbol{\theta} \mid \mathcal{D})\int_{\mathcal{Z}} Zp(Z \mid \boldsymbol{\theta})\,\mathrm{d}Z \\ &= p(\boldsymbol{\theta} \mid \mathcal{D})\mathbb{E}\left[Z \mid \boldsymbol{\theta}\right].\end{aligned}$$

By linearity of expectation we have

$$p(\boldsymbol{\theta} \mid \mathcal{D})\mathbb{E}\left[Z \mid \boldsymbol{\theta}\right] = \frac{p(\boldsymbol{\theta} \mid \mathcal{D})}{\mathcal{L}(\mathcal{D};\boldsymbol{\theta})}\mathbb{E}\left[\hat{\mathcal{L}}(\mathcal{D};\boldsymbol{\theta}) \mid \boldsymbol{\theta}\right].$$

We also have $\mathbb{E}\left[\hat{\mathcal{L}}(\mathcal{D};\boldsymbol{\theta}) \mid \boldsymbol{\theta}\right] = \mathcal{L}(\mathcal{D};\boldsymbol{\theta})$, since the Monte Carlo estimator for the likelihood is unbiased. Therefore,

$$\int_{\mathcal{Z}} p(Z,\boldsymbol{\theta})\,\mathrm{d}Z \propto p(\boldsymbol{\theta} \mid \mathcal{D}).$$

We conclude that the original chain has as its stationary distribution the exact posterior, $p(\boldsymbol{\theta} \mid \mathcal{D})$.

5.7.3 MCMC convergence diagnostics

In the main text we apply the rank normalised $\hat{R}$ statistic and the multiple chain effective sample size measure, S_{eff}, as defined in the recent work by Vehtari et al. (2019) that improves earlier definitions (Gelman and Rubin, 1992; Gelman et al., 2014). Since these diagnostics are relatively recent updates, we present their definitions here (see `Diagnostics.jl` for example implementation).

Consider, $\mathcal{R}$ chains, taking values in $\mathbb{R}^d$, each consisting of an even number of iterations, $\mathcal{M}$. Let $\theta^r_{k,m}$ denote the kth dimension of the mth iteration of the rth chain, then we define the rank normalised transform as

$$z^r_{k,m} = \Phi^{-1}\left(\frac{\eta^r_{k,m} - 1/2}{\mathcal{R}\mathcal{M}}\right),$$

where $\eta^r_{k,m}$ is the rank of $\theta^r_{k,m}$ taken over all $m = 1, 2, \ldots, \mathcal{M}$ and $r = 1, 2, \ldots, \mathcal{R}$, and $\Phi^{-1} : [0,1] \to \mathbb{R}$ is the inverse cumulative distribution function of the standard Normal distribution. Then the rank normalised $\hat{R}_k$ statistic for the k parameter is defined as

$$\hat{R}_k = \sqrt{\frac{V_k}{W_k}},$$

where

$$V_k = \frac{\mathcal{M}-2}{\mathcal{M}} W_k + \frac{2}{\mathcal{M}} B_k,$$

with within-chain variance estimate W_k and between-chain variance estimate B_k. These estimates are given by

$$B_k = \frac{\mathcal{M}}{4\mathcal{R}-2} \sum_{r=1}^{\mathcal{R}} \left(\bar{z}^r_{k,+} - \bar{\bar{z}}_k\right)^2 + \left(\bar{z}^r_{k,-} - \bar{\bar{z}}_k\right)^2 \quad \text{and} \quad W_k = \frac{1}{2\mathcal{R}} \sum_{r=1}^{\mathcal{R}} s^r_{k,+} + s^r_{k,-},$$

where

$$\bar{z}^r_{k,+} = \frac{2}{\mathcal{M}} \sum_{m=\mathcal{M}/2+1}^{\mathcal{M}} z^r_{k,m}, \quad \bar{z}^r_{k,-} = \frac{2}{\mathcal{M}} \sum_{m=1}^{\mathcal{M}/2} z^r_{k,m}, \quad \bar{\bar{z}}_k = \frac{1}{2\mathcal{R}} \sum_{r=1}^{\mathcal{R}} \bar{z}^r_{k,+} + \bar{z}^r_{k,-}$$

$$s^r_{k,+} = \frac{2}{\mathcal{M}-2} \sum_{m=\mathcal{M}/2+1}^{\mathcal{M}} \left(z^r_{k,m} - \bar{z}^r_{k,+}\right)^2 \quad \text{and} \quad s^r_{k,-} = \frac{2}{\mathcal{M}-2} \sum_{m=1}^{\mathcal{M}/2} \left(z^r_{k,m} - \bar{z}^r_{k,-}\right)^2.$$

The multiple chain effective sample size measure is computed according to

$$S_{\text{eff},k} = \frac{\mathcal{R}\mathcal{M}}{\hat{\tau}_k},$$

where

$$\hat{\tau}_k = 1 + 2\sum_{\ell=1}^{L_k} \hat{\rho}_{k,\ell}, \quad \hat{\rho}_{k,\ell} = 1 - \frac{1}{V_k}\left(W_k - \frac{1}{\mathcal{R}}\sum_{r=1}^{\mathcal{R}} \hat{\rho}^r_{k,\ell}\right),$$

and $\hat{\rho}^r_{k,\ell} = \mathbb{C}\left[z^r_{k,m}, z^r_{k,m+\ell}\right] / \mathbb{V}\left[z^r_{k,m}\right]$ is the autocorrelation function for the trace of the kth dimension of the rth chain at lag ℓ. L_k is the largest odd integer such that $\hat{\rho}_{k,\ell+1} + \hat{\rho}_{k,\ell+2} > 0$ for all $\ell = 1, 3, \ldots, L_k - 2$ (Gelman et al., 2014; Vehtari et al., 2019).

To reliably use the chains $\boldsymbol{\theta}^1_m, \boldsymbol{\theta}^2_m, \ldots, \boldsymbol{\theta}^{\mathcal{R}}_m$ for estimation of the posterior mean, Vehtari et al. (2019) recommend that the chains should at least satisfy the conditions $\hat{R}_k < 1.01$ and $S_{\text{eff},k} > 400$ for all $k = 1, 2, \ldots, d$. Of course, this does not guarantee that the chains have converged, but it is a guide that, coupled with trace plots and ACF plots, provide reasonably conservative results.

5.7.4 Observed data

The synthetic data used throughout this chapter and example code is provided in Table 5.7 for the stationary production-degradation model model, Table 5.8 for the Michaelis-Menten model, Table 5.9 for the Schlögl model and Table 5.10 for the repressilator model.

Table 5.7: Data, $\mathcal{D}$, used for inference on the production-degradation model. Generated using parameter values $k_1 = 1.0$ and $k_2 = 0.01$, initial conditions $X_0 = 10$, and final time $t = 1,000,000$.

	$Y^{(1)}_{\text{obs}}$	$Y^{(2)}_{\text{obs}}$	$Y^{(3)}_{\text{obs}}$	$Y^{(4)}_{\text{obs}}$	$Y^{(5)}_{\text{obs}}$	$Y^{(6)}_{\text{obs}}$	$Y^{(7)}_{\text{obs}}$	$Y^{(8)}_{\text{obs}}$	$Y^{(9)}_{\text{obs}}$	$Y^{(10)}_{\text{obs}}$
X_∞	91.68	101.64	88.13	98.88	96.36	119.59	100.62	105.11	105.30	97.00

Table 5.8: Data, $\mathcal{D}$, used for inference on the Michaelis-Menten model. Generated using parameter values $k_1 = 0.001$, $k_2 = 0.05$ and $k_3 = 0.01$, and initial conditions $E_0 = 100$, $S_0 = 100$, $C_0 = 0$ and P_0. The observation error is Gaussian with standard deviation $\sigma = 10$.

	$Y_{\text{obs}}^{(1)}$	$Y_{\text{obs}}^{(2)}$	$Y_{\text{obs}}^{(3)}$	$Y_{\text{obs}}^{(4)}$	$Y_{\text{obs}}^{(5)}$	$Y_{\text{obs}}^{(6)}$	$Y_{\text{obs}}^{(7)}$	$Y_{\text{obs}}^{(8)}$	$Y_{\text{obs}}^{(9)}$	$Y_{\text{obs}}^{(10)}$
t	5	10	15	20	25	30	35	40	45	50
E_t	60.84	47.21	39.53	48.64	28.99	43.53	43.78	73.16	38.40	36.84
S_t	60.77	45.40	46.47	58.84	12.21	48.05	39.03	20.26	0.00	7.73
C_t	42.22	62.48	54.47	59.77	60.34	61.04	57.59	67.03	50.46	64.41
P_t	0.00	0.00	0.00	0.00	21.60	10.04	15.78	20.71	32.32	32.34

	$Y_{\text{obs}}^{(11)}$	$Y_{\text{obs}}^{(12)}$	$Y_{\text{obs}}^{(13)}$	$Y_{\text{obs}}^{(14)}$	$Y_{\text{obs}}^{(15)}$	$Y_{\text{obs}}^{(16)}$	$Y_{\text{obs}}^{(17)}$	$Y_{\text{obs}}^{(18)}$	$Y_{\text{obs}}^{(19)}$	$Y_{\text{obs}}^{(20)}$
t	55	60	65	70	75	80	85	90	95	100
E_t	37.87	37.62	45.81	34.28	49.84	50.68	41.92	42.47	41.36	63.29
S_t	1.13	15.99	17.57	5.06	5.28	0.00	4.07	17.85	19.97	27.57
C_t	64.41	49.31	53.41	62.54	55.42	42.85	43.01	62.41	37.86	38.02
P_t	17.16	36.85	42.20	27.55	41.33	15.40	28.60	29.29	41.10	63.48

Table 5.9: Data, $\mathcal{D}$, used for inference on the Schlögl model. Generated using parameter values $k_1 = 0.18$, $k_2 = 0.00025$, $k_3 = 2200.0$ and $k_4 = 37.5$, and initial condition $X_0 = 0$. The observation error is Gaussian with standard deviation $\sigma = 10$.

	$Y_{\text{obs}}^{(1)}$	$Y_{\text{obs}}^{(2)}$	$Y_{\text{obs}}^{(3)}$	$Y_{\text{obs}}^{(4)}$	$Y_{\text{obs}}^{(5)}$	$Y_{\text{obs}}^{(6)}$	$Y_{\text{obs}}^{(7)}$	$Y_{\text{obs}}^{(8)}$
t	12.5	25	37.5	50	62.5	75	87.5	100
X_t	134.99	95.83	370.91	94.15	470.12	108.17	111.20	59.54

	$Y_{\text{obs}}^{(9)}$	$Y_{\text{obs}}^{(10)}$	$Y_{\text{obs}}^{(11)}$	$Y_{\text{obs}}^{(12)}$	$Y_{\text{obs}}^{(13)}$	$Y_{\text{obs}}^{(14)}$	$Y_{\text{obs}}^{(15)}$	$Y_{\text{obs}}^{(16)}$
t	112.5	125	137.5	150	162.5	175	187.5	200
X_t	99.74	347.01	92.66	377.61	416.85	120.85	361.12	282.14

Table 5.10: Data, $\mathcal{D}$, used for inference on the repressilator model. Generated using parameter values $\alpha = 1000$, $\alpha_0 = 1$, $n = 2$ $\beta = 5$ and $\gamma = 1$, and initial conditions $M_{1,0} = 0$, $P_{1,0} = 2$, $M_{2,0} = 0$, $P_{2,0} = 1$, $M_{3,0} = 0$ and $P_{3,0} = 3$. The observation error is Gaussian with standard deviation $\sigma = 10$.

	$Y_{\text{obs}}^{(1)}$	$Y_{\text{obs}}^{(2)}$	$Y_{\text{obs}}^{(3)}$	$Y_{\text{obs}}^{(4)}$	$Y_{\text{obs}}^{(5)}$	$Y_{\text{obs}}^{(6)}$	$Y_{\text{obs}}^{(7)}$	$Y_{\text{obs}}^{(8)}$	$Y_{\text{obs}}^{(9)}$	$Y_{\text{obs}}^{(10)}$
t	5	10	15	20	25	30	35	40	45	50
$M_{1,t}$	54.09	0.00	303.69	5.46	157.54	0.00	21.27	7.88	0.00	179.26
$P_{1,t}$	45.65	6.54	416.10	17.78	141.51	26.29	4.97	24.36	21.30	264.20
$M_{2,t}$	0.00	498.38	27.63	70.33	44.20	13.44	75.85	0.00	448.41	8.83
$P_{2,t}$	0.00	532.66	2.56	32.25	67.96	0.00	86.08	0.00	562.18	7.73
$M_{3,t}$	20.30	11.97	27.58	213.90	0.00	376.09	6.92	439.23	0.00	323.84
$P_{3,t}$	19.87	46.03	3.61	270.08	7.55	287.98	12.58	413.75	0.00	227.21

	$Y_{\text{obs}}^{(11)}$	$Y_{\text{obs}}^{(12)}$	$Y_{\text{obs}}^{(13)}$	$Y_{\text{obs}}^{(14)}$	$Y_{\text{obs}}^{(15)}$	$Y_{\text{obs}}^{(16)}$	$Y_{\text{obs}}^{(17)}$	$Y_{\text{obs}}^{(18)}$	$Y_{\text{obs}}^{(19)}$	$Y_{\text{obs}}^{(20)}$
t	55	60	65	70	75	80	85	90	95	100
$M_{1,t}$	0.00	277.77	40.45	0.00	222.82	0.00	231.91	58.95	8.22	144.86
$P_{1,t}$	0.13	177.32	24.31	0.00	241.58	19.47	157.78	70.00	6.84	120.19
$M_{2,t}$	66.17	58.86	3.03	608.50	7.38	40.98	61.72	6.91	217.20	0
$P_{2,t}$	46.77	67.67	10.50	579.68	9.09	48.01	95.46	13.10	270.19	13.40
$M_{3,t}$	136.19	4.53	135.78	18.84	8.62	146.78	0.00	404.93	26.80	29.69
$P_{3,t}$	197.56	3.01	136.12	8.00	5.35	174.83	0.00	393.10	22.28	8.00

Chapter 6

Multilevel Rejection Sampling for Approximate Bayesian Computation

A paper published in *Computational Statistics and Data Analysis*.

David J. Warne, Ruth E. Baker and Matthew J. Simpson (2018). Multilevel rejection sampling for approximate Bayesian computation. *Computational Statistics and Data Analysis*, 124:71–86. DOI:10.1016/j.csda.2018.02.009

Abstract Likelihood-free methods, such as approximate Bayesian computation, are powerful tools for practical inference problems with intractable likelihood functions. Markov chain Monte Carlo and sequential Monte Carlo variants of approximate Bayesian computation can be effective techniques for sampling posterior distributions in an approximate Bayesian computation setting. However, without careful consideration of convergence criteria and selection of proposal kernels, such methods can lead to very biased inference or computationally inefficient sampling. In contrast, rejection sampling for approximate Bayesian computation, despite being computationally intensive, results in independent, identically distributed samples from the approximated posterior. An alternative method is proposed for the acceleration of likelihood-free Bayesian inference that applies multilevel Monte Carlo variance reduction techniques directly to rejection sampling. The resulting method retains the accuracy advantages of rejection sampling

while significantly improving the computational efficiency.

6.1 Introduction

Statistical inference is of fundamental importance to all areas of science. Inference enables the testing of theoretical models against observations, and provides a rational means of quantifying uncertainty in existing models. Modern approaches to statistical inference, based on Monte Carlo sampling techniques, provide insight into many complex phenomena (Beaumont et al., 2002; Pooley et al., 2015; Ross et al., 2017; Stumpf, 2014; Sunnåker et al., 2013; Tavaré et al., 1997; Thorne and Stumpf, 2012; Vo et al., 2015a).

Suppose we have: a set of observations, $\mathcal{D}$; a method of determining the likelihood of these observations, $\mathcal{L}(\boldsymbol{\theta}; \mathcal{D})$, under the assumption of some model characterised by parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta}$; and a prior probability density, $p(\boldsymbol{\theta})$. The posterior probability density, $p(\boldsymbol{\theta} \mid \mathcal{D})$, can be computed using Bayes' Theorem,

$$p(\boldsymbol{\theta} \mid \mathcal{D}) = \frac{\mathcal{L}(\boldsymbol{\theta}; \mathcal{D}) p(\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\theta}; \mathcal{D}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}}. \tag{6.1}$$

Explicit expressions for likelihood functions are rarely available (Tavaré et al., 1997; Wilkinson, 2009) (see Chapter 4); motivating the development of likelihood-free methods, such as approximate Bayesian computation (ABC) (Stumpf, 2014; Sunnåker et al., 2013). ABC methods approximate the likelihood through evaluating the discrepancy between data generated by a simulation of the model of interest and the observations, yielding an approximate posterior,

$$p(\boldsymbol{\theta} \mid \rho(\mathcal{D}, \mathcal{D}_s) < \epsilon) \propto \mathbb{P}(\rho(\mathcal{D}, \mathcal{D}_s) < \epsilon \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}). \tag{6.2}$$

Here, $\mathcal{D}_s \sim s(\mathcal{D} \mid \boldsymbol{\theta})$ is data generated by the model simulation process, $s(\mathcal{D} \mid \boldsymbol{\theta})$, ρ is a discrepancy metric, and $\epsilon > 0$ is the acceptance threshold. Due to this approximation, Monte Carlo estimators based on Equation (6.2) are biased (Barber et al., 2015). In spite of this bias, however, ABC methods have proven to be very powerful tools for practical inference applications in many scientific areas, including evolutionary biology (Beaumont et al., 2002; Tavaré et al., 1997; Thorne and Stumpf, 2012), ecology (Stumpf, 2014), cell biology (Ross et al., 2017; Johnston et al., 2014; Vo et al., 2015a) and systems biology (Wilkinson, 2009).

6.1.1 Sampling algorithms for ABC

The most elementary implementation of ABC is ABC rejection sampling (Pritchard et al., 1999; Sunnåker et al., 2013), see Algorithm 6.1. This method generates N independent and identically distributed samples $\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^N$ from the posterior distribution by accepting proposals, $\boldsymbol{\theta}^* \sim p(\boldsymbol{\theta})$, when the data generated by the model simulation process $s(\mathcal{D} \mid \boldsymbol{\theta}^*)$ is within ϵ of the observed data, $\mathcal{D}$, under the discrepancy metric $\rho(\mathcal{D}, \cdot)$. ABC rejection sampling is simple to implement, and samples are independent and identically distributed. Therefore ABC rejection sampling is widely used in many applications (Browning et al., 2018; Grelaud et al., 2009; Navascués et al., 2017; Ross et al., 2017; Vo et al., 2015a). However, ABC rejection sampling can be computationally prohibitive in practice (Barber et al., 2015; Fearnhead and Prangle, 2012). This is especially true when the prior density is highly diffuse compared with the target posterior density (Marin et al., 2012), as most proposals are rejected.

Algorithm 6.1 ABC rejection sampler

1: **for** $i = 1, \ldots, N$ **do**
2: **repeat**
3: Sample prior, $\boldsymbol{\theta}^* \sim p(\boldsymbol{\theta})$.
4: Generate data, $\mathcal{D}_s \sim s(\mathcal{D} \mid \boldsymbol{\theta}^*)$.
5: **until** $\rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon$
6: Set $\boldsymbol{\theta}^i \leftarrow \boldsymbol{\theta}^*$.
7: **end for**

To improve the efficiency of ABC rejection sampling, one can consider a likelihood-free modification of Markov chain Monte Carlo (MCMC) (Beaumont et al., 2002; Marjoram et al., 2003; Tanaka et al., 2006) in which a Markov chain is constructed with a stationary distribution identical to the desired posterior. Given the Markov chain is in state $\boldsymbol{\theta}^i$, a state transition is proposed via a proposal kernel, $q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^i)$.

The Metropolis-Hastings (Hastings, 1970; Metropolis et al., 1953) state transition probability, h, can be modified within an ABC framework to yield

$$h = \begin{cases} \min\left(\dfrac{p(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^i \mid \boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}^i)q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^i)}, 1\right) & \text{if } \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon, \\ 0 & \text{otherwise.} \end{cases}$$

The stationary distribution of such a Markov chain is the desired approximate posterior (Marjoram et al., 2003). Algorithm 6.2 provides a method for computing N_T iterations of this Markov chain.

While MCMC-ABC sampling can be highly efficient, the samples in the sequence, $\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^{N_T}$, are not independent. This can be problematic as it is possible for the Markov chain to take long excursions into regions of low posterior probability. This incurs additional bias that is potentially significant (Sisson et al., 2007). A poor choice of proposal kernel can also have considerable impact upon the efficiency of MCMC-ABC (Green et al., 2015). The question of how to choose the proposal kernel is non-trivial. Typically proposal kernels are determined heuristically. However, automatic and adaptive schemes are available to assist in obtaining near optimal proposals in some cases (Cabras et al., 2015; Roberts and Rosenthal, 2009). Another additional complication is that of determining when the Markov Chain has converged; this is a challenging problem to solve in practice (Roberts and Rosenthal, 2004).

Algorithm 6.2 MCMC-ABC

Given initial sample $\boldsymbol{\theta}^1 \sim p(\boldsymbol{\theta} \mid \rho(\mathcal{D}, \mathcal{D}_s) < \epsilon)$.
for $i = 2, \ldots, N_T$ **do**
    Sample transition kernel, $\boldsymbol{\theta}^* \sim q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{i-1})$.
    Generate data, $\mathcal{D}_s \sim s(\mathcal{D} \mid \boldsymbol{\theta}^*)$.
    if $\rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon$ **then**
        Set $h \leftarrow \min\left(p(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^{i-1} \mid \boldsymbol{\theta}^*)/p(\boldsymbol{\theta}^{i-1})q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{i-1}), 1\right)$.
        Sample uniform distribution, $u \sim \mathcal{U}(0, 1)$.
        if $u \leq h$ **then**
            Set $\boldsymbol{\theta}^i \leftarrow \boldsymbol{\theta}^*$.
        else
            Set $\boldsymbol{\theta}^i \leftarrow \boldsymbol{\theta}^{i-1}$.
        end if
    else
        Set $\boldsymbol{\theta}^i \leftarrow \boldsymbol{\theta}^{i-1}$.
    end if
end for

Sequential Monte Carlo (SMC) sampling was introduced to address these potential inefficiencies (Del Moral et al., 2006) and later extended within an ABC context (Sisson et al., 2007; Drovandi and Pettitt, 2011; Toni et al., 2009). A set of samples, referred to as particles, is evolved through a sequence of ABC posteriors defined through a sequence of T acceptance thresholds, $\epsilon_1, \ldots, \epsilon_T$ (Sisson et al., 2007; Beaumont et al., 2009). At each step, $t \in [0, T]$, the current ABC posterior, $p(\boldsymbol{\theta} \mid \rho(\mathcal{D}, \mathcal{D}_s) < \epsilon_t)$, is approximated by a discrete distribution

constructed from a set of N_P particles $\boldsymbol{\theta}_t^1, \ldots, \boldsymbol{\theta}_t^{N_P}$ with importance weights $W_t^1, \ldots, W_t^{N_P}$ that is, $\mathbb{P}(\boldsymbol{\theta} = \boldsymbol{\theta}_t^i) = W_t^i$ for $i = 1, \ldots, N_P$. The collection is updated to represent the ABC posterior of the next step, $p(\boldsymbol{\theta} \mid \rho(\mathcal{D}, \mathcal{D}_s) < \epsilon_{t+1})$, through application of rejection sampling on particles perturbed by a proposal kernel, $q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{i-1})$. If $p(\boldsymbol{\theta} \mid \rho(\mathcal{D}, \mathcal{D}_s) < \epsilon_t)$ is similar to $p(\boldsymbol{\theta} \mid \rho(\mathcal{D}, \mathcal{D}_s) < \epsilon_{t+1})$, the acceptance rate should be high. The importance weights of the new family of particles are updated using an optimal backwards kernel (Sisson et al., 2007; Beaumont et al., 2009). Algorithm 6.3 outlines the process.

Algorithm 6.3 SMC-ABC

Initialise $\boldsymbol{\theta}_1^i \sim p(\boldsymbol{\theta})$ and $W_1^i = 1/N_P$, for $i = 1, \ldots, N_P$.
for $t = 2, \ldots, T$ **do**
    for $i = 1, \ldots, N_P$ **do**
        repeat
            Set $\boldsymbol{\theta}^* \leftarrow \boldsymbol{\theta}_{t-1}^j$ with probability W_{t-1}^j.
            Sample transition kernel, $\boldsymbol{\theta}^{**} \sim q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^*)$.
            Generate data, $\mathcal{D}_s \sim s(\mathcal{D} \mid \boldsymbol{\theta}^{**})$.
        until $\rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_t$
        Set $\boldsymbol{\theta}_t^i \leftarrow \boldsymbol{\theta}^{**}$.
        Set $W_t^i \leftarrow p(\boldsymbol{\theta}_t^i) / \sum_{j=1}^{N_P} W_{t-1}^j q(\boldsymbol{\theta}_t^i \mid \boldsymbol{\theta}_{t-1}^j)$.
    end for
    Normalise weights so that $\sum_{i=1}^{N_P} W_t^i = 1$.
end for

Through the use of independent weighted particles, SMC-ABC avoids long excursions into the distribution tails that are possible with MCMC-ABC. However, SMC-based methods can be affected by particle degeneracy, and the efficiency of each step is still dependent on the choice of proposal kernel (Green et al., 2015; Filippi et al., 2013). Just as with MCMC-ABC, adaptive schemes are available to guide proposal kernel selection (Beaumont et al., 2009; Del Moral et al., 2012). The choice of the sequence of acceptance thresholds is also important for the efficiency of SMC-ABC. However, there are good solutions to generate these sequences in an adaptive manner (Drovandi and Pettitt, 2011; Silk et al., 2013).

6.1.2 Multilevel Monte Carlo

Multilevel Monte Carlo (MLMC) is a recent development that can significantly reduce the computational burden in the estimation of expectations (Giles, 2008). To demonstrate the basic idea of MLMC, consider computing the expectation of a continuous-time stochastic process X_t

at time T. Let Z_t^τ denote a discrete-time approximation to X_t with time step τ: the expectations of X_T and Z_T^τ are related according to

$$\mathbb{E}\left[X_T\right] = \mathbb{E}\left[Z_T^\tau\right] + \mathbb{E}\left[X_T - Z_T^\tau\right].$$

That is, an estimate of $\mathbb{E}\left[Z_T^\tau\right]$ can be treated as a biased estimate of $\mathbb{E}\left[X_T\right]$. By taking a sequence of time steps $\tau_1 > \cdots > \tau_L$, the indices of which are referred to as *levels*, we can arrive at a telescoping sum,

$$\mathbb{E}\left[Z_T^{\tau_L}\right] = \mathbb{E}\left[Z_T^{\tau_1}\right] + \sum_{\ell=2}^{L} \mathbb{E}\left[Z_T^{\tau_\ell} - Z_T^{\tau_{\ell-1}}\right]. \tag{6.3}$$

Computing this form of the expectation returns the same bias as that returned when computing $\mathbb{E}\left[Z_T^{\tau_L}\right]$ directly. However, Giles (2008) demonstrates that a Monte Carlo estimator for the telescoping sum can be computed more efficiently than directly estimating $\mathbb{E}\left[Z_T^{\tau_L}\right]$ in the context of stochastic differential equations (SDEs). This efficiency comes from exploiting the fact that the bias correction terms, $\mathbb{E}\left[Z_T^{\tau_\ell} - Z_T^{\tau_{\ell-1}}\right]$, measure the expected difference between the estimates on levels ℓ and $\ell - 1$. Therefore, sample paths of $Z_T^{\tau_{\ell-1}}$ need not be independent of sample paths of $Z_T^{\tau_\ell}$. In the case of SDEs, samples may be generated in pairs driven by the same underlying Brownian motion, that is, the pair is coupled. By the strong convergence properties of numerical schemes for SDEs, Giles (2008) shows that this coupling is sufficient to reduce the variance of the Monte Carlo estimator. This reduction in variance is achieved through optimally trading off statistical error and computational cost across all levels. Through this trade-off, an estimator is obtained with the same accuracy in *mean-square* to standard Monte Carlo, but at a reduced computational cost. This saving of computational cost is achieved since fewer samples of the most accurate discretisation are required.

6.1.3 Related work

Recently, several examples of MLMC applications to Bayesian inference problems have appeared in the literature. One of the biggest challenges in the application of MLMC to inverse problems is the introduction of a strong coupling between levels. That is, the construction of a coupling mechanism that reduces the variances of the bias correction terms enough to enable the MLMC estimator to be computed more efficiently than standard Monte Carlo. Dodwell et al. (2015) demonstrate a MLMC scheme for MCMC sampling applicable to high-dimensional

Bayesian inverse problems with closed-form likelihood expressions. The coupling of Dodwell et al. (2015) is based on correlating Markov chains defined on a hierarchy in parameter space. A similar approach is also employed by Efendiev et al. (2015). A multilevel method for ensemble particle filtering is proposed by Gregory et al. (2016) that employs an optimal transport problem to correlate a sequence of particle filters of different levels of accuracy. Due to the computational cost of the transport problem, a local approximation scheme is introduced for multivariate parameters (Gregory et al., 2016). Beskos et al. (2017) look more generally at the case of applying MLMC methods when independent and identically distributed sampling of the distributions on some levels is infeasible, the result is an extension of MLMC in which coupling is replaced with sequential importance sampling, that is, a multilevel variant of SMC (MLSMC).

MLMC has also recently been considered in an ABC context. Guha and Tan (2017) extend the work of Efendiev et al. (2015) by replacing the Metropolis-Hasting acceptance probability in a similar way to the MCMC-ABC method (Marjoram et al., 2003). The MLSMC method (Beskos et al., 2017) is exploited to achieve coupling in an ABC context by Jasra et al. (2019).

6.1.4 Aims and contribution

ABC samplers based on MCMC and SMC are generally more computationally efficient than ABC rejection sampling (Marjoram et al., 2003; Sisson et al., 2007; Toni et al., 2009). However, there are many advantages to using ABC rejection sampling. Specific advantages are: (i) its simple implementation; (ii) it produces truly independent and identically distributed samples, that is, there is no need to re-weight samples; and (iii) there are no algorithm parameters that affect the computational efficiency. In particular, no proposal kernels need be heuristically defined for ABC rejection sampling.

The aim of this work is to design a new algorithm for ABC inference that retains the aforementioned advantages of ABC rejection sampling, while still being efficient and accurate. The aim is not to develop a method that is always superior to MCMC and SMC methods, but rather provide a method that requires little user-defined configuration, and is still computationally reasonable. To this end, we investigate MLMC methods applied directly to ABC rejection sampling.

Various applications of MLMC to ABC inference have been demonstrated very recently in the literature (Guha and Tan, 2017; Jasra et al., 2017), with each implementation contributing different ideas for constructing effective control variates in an ABC context. It is not yet clear when one approach will be superior to another. Thus the question of how best to apply MLMC in an ABC context remains an open problem. Since there are no examples of an application of MLMC methods to the most fundamental of ABC samplers, that is rejection sampling, such an application is a significant contribution to this field.

We describe a new algorithm for ABC rejection sampling, based on MLMC. Our new algorithm, called MLMC-ABC, is as general as standard ABC rejection sampling but is more computationally efficient through the use of variance reduction techniques that employ a novel construction for the coupling problem. We describe the algorithm, its implementation, and validate the method using a tractable Bayesian inverse problem. We also compare the performance of the new method against MCMC-ABC and SMC-ABC using a standard benchmark problem from epidemiology.

Our method benefits from the simplicity of ABC rejection sampling. We require only the discrepancy metric and a sequence of acceptance thresholds to be defined for a given inverse problem. Our approach is also efficient, and achieves comparable or superior performance to MCMC-ABC and SMC-ABC methods, at least for the examples considered in this paper. Therefore, we demonstrate that our algorithm is a promising method that could be extended to design viable alternatives to current state-of-the-art approaches. This work, along with that of Guha and Tan (2017) and Jasra et al. (2017), provides an additional set of computational tools to further enhance the utility of ABC methods in practice.

6.2 Methods

In this section, we demonstrate our application of MLMC ideas to the likelihood-free inference problem given in Equation (6.2). The initial aim is to compute an accurate approximation to the joint posterior cumulative distribution function (CDF) of $\boldsymbol{\theta}$, using as few data generation steps as possible. We define a data generation step to be a simulation of the model of interest given a proposed parameter vector. While the initial presentation is given in the context of estimating the joint CDF, the method naturally extends to expectations of other functions (see

6.5.2). However, the joint CDF is considered first for clarity in the introduction of the MLMC coupling strategy.

We apply MLMC methods to likelihood-free Bayesian inference by reposing the problem of computing the posterior CDF as an expectation calculation. This allows the MLMC telescoping sum idea, as in Equation (6.3), to be applied. In this context, the levels of the MLMC estimator are parameterised by a sequence of L acceptance thresholds $\epsilon_1, \ldots, \epsilon_L$ with $\epsilon_\ell > \epsilon_{\ell+1}$ for all $\ell \in [1, L)$. The efficiency of MLMC relies upon computing the terms of the telescoping sum with low variance. Variance reduction is achieved through exploiting features of the telescoping sum for CDF approximation, and further computational gains are achieved by using properties of nested ABC rejection samplers.

6.2.1 ABC posterior as an expectation

We first reformulate the Bayesian inference problem (Equation (6.1)) as an expectation. To this end, note that, given a k-dimensional parameter space, $\boldsymbol{\Theta}$, and the random parameter vector, $\boldsymbol{\theta}$, if the joint posterior CDF, $F(\mathbf{s}) = \mathbb{P}(\theta_1 \leq s_1, \ldots, \theta_k \leq s_k)$ is differentiable, that is, its probability density function (PDF) exists, then

$$\begin{aligned}\mathbb{P}(\theta_1 \leq s_1, \ldots, \theta_k \leq s_k) &= \int_{-\infty}^{s_1} \cdots \int_{-\infty}^{s_k} p(\boldsymbol{\theta} \mid \mathcal{D}) \, \mathrm{d}\theta_k \ldots \mathrm{d}\theta_1 \\ &= \int_{A_\mathbf{s}} p(\boldsymbol{\theta} \mid \mathcal{D}) \, \mathrm{d}\theta_k \ldots \mathrm{d}\theta_1,\end{aligned}$$

where $A_\mathbf{s} = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \theta_1 \leq s_1, \ldots, \theta_k \leq s_k\}$. This can be expressed as an expectation by noting

$$\begin{aligned}\int_{A_\mathbf{s}} p(\boldsymbol{\theta} \mid \mathcal{D}) \, \mathrm{d}\theta_k \ldots \mathrm{d}\theta_1 &= \int_{\boldsymbol{\Theta}} \mathbb{1}_{A_\mathbf{s}}(\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathcal{D}) \, \mathrm{d}\theta_k \ldots \mathrm{d}\theta_1 \\ &= \mathbb{E}\left[\mathbb{1}_{A_\mathbf{s}}(\boldsymbol{\theta}) \mid \mathcal{D}\right],\end{aligned}$$

where $\mathbb{1}_{A_\mathbf{s}}(\boldsymbol{\theta})$ is the indicator function: $\mathbb{1}_{A_\mathbf{s}}(\boldsymbol{\theta}) = 1$ whenever $\boldsymbol{\theta} \in A_\mathbf{s}$ and $\mathbb{1}_{A_\mathbf{s}}(\boldsymbol{\theta}) = 0$ otherwise. Now, consider the ABC approximation given in Equation (6.2) with discrepancy metric $\rho(\mathcal{D}, \mathcal{D}_s)$ and acceptance threshold ϵ. The ABC posterior CDF, denoted by $F_\epsilon(\mathbf{s})$, will

be

$$\begin{aligned} F_\epsilon(\mathbf{s}) &= \int_{A_\mathbf{s}} p(\boldsymbol{\theta} \mid \rho(\mathcal{D}, \mathcal{D}_s) < \epsilon) \, \mathrm{d}\theta_k \ldots \mathrm{d}\theta_1 \\ &= \mathbb{E}\left[\mathbb{1}_{A_\mathbf{s}}(\boldsymbol{\theta}) \mid \rho(\mathcal{D}, \mathcal{D}_s) < \epsilon\right]. \end{aligned} \tag{6.4}$$

The marginal ABC posterior CDFs, $F_{\epsilon,1}(s), \ldots, F_{\epsilon,k}(s)$, are

$$F_{\epsilon,j}(s) = \lim_{s_{i \neq j} \to \infty} F_\epsilon(\mathbf{s}).$$

6.2.2 Multilevel estimator formulation

We now introduce some notation that will simplify further derivations. We define $\boldsymbol{\theta}_\epsilon$ to be a random vector distributed according to the ABC posterior CDF, $F_\epsilon(\mathbf{s})$, with acceptance threshold, ϵ, as given in Equation (6.4). This provides us with simplification in notation for the ABC posterior PDF, $p(\boldsymbol{\theta}_\epsilon) = p(\boldsymbol{\theta} \mid \rho(\mathcal{D}, \mathcal{D}_s) < \epsilon)$, and the conditional expectation, $\mathbb{E}\left[\mathbb{1}_{A_\mathbf{s}}(\boldsymbol{\theta}_\epsilon)\right] = \mathbb{E}\left[\mathbb{1}_{A_\mathbf{s}}(\boldsymbol{\theta}) \mid \rho(\mathcal{D}, \mathcal{D}_s) < \epsilon\right]$. For any expectation, P, we use $\hat{P}^N$ to denote the Monte Carlo estimate of this expectation using N samples.

The standard Monte Carlo integration approach is to generate N samples $\boldsymbol{\theta}_\epsilon^1, \ldots, \boldsymbol{\theta}_\epsilon^N$ from the ABC posterior, $p(\boldsymbol{\theta}_\epsilon)$, then evaluate the empirical CDF (eCDF),

$$\hat{F}_\epsilon^N(\mathbf{s}) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{A_\mathbf{s}}(\boldsymbol{\theta}_\epsilon^i), \tag{6.5}$$

for $\mathbf{s} \in \mathcal{S}$, where $\mathcal{S}$ is a discretisation of the parameter space $\boldsymbol{\Theta}$. For simplicity, we will consider $\mathcal{S}$ to be a k-dimensional regular lattice. In general, a regular lattice will not be well suited for high dimensional problems. However, since this is the first time that this MLMC approach has been presented, it is most natural to begin the exposition with a regular lattice, and then discuss other more computationally efficient approaches later. More attention is given to other possibilities in Section 6.4.

The eCDF is not, however, the only Monte Carlo approximation to the CDF one may consider. In particular, Giles et al. (2015) demonstrate the application of MLMC to a univariate CDF approximation. We now present a multivariate equivalent of the MLMC CDF of Giles et al.

(2015) in the context of ABC posterior CDF estimation. Consider a sequence of L acceptance thresholds, $\{\epsilon_\ell\}_{\ell=1}^{\ell=L}$, that is strictly decreasing, that is, $\epsilon_\ell > \epsilon_{\ell+1}$. In this work, the problem of constructing optimal sequences is not considered, as the focus is the initial development of the method. More discussion around this problem is given in Section 6.4. Given such a sequence, $\{\epsilon_\ell\}_{\ell=1}^{\ell=L}$, we can represent the CDF (Equation (6.4)) using the telescoping sum

$$F_{\epsilon_L}(\mathbf{s}) = \mathbb{E}\left[\mathbb{1}_{A_\mathbf{s}}(\boldsymbol{\theta}_{\epsilon_L})\right] = \sum_{\ell=1}^{L} Y_\ell(\mathbf{s}), \tag{6.6}$$

where

$$Y_\ell(\mathbf{s}) = \begin{cases} \mathbb{E}\left[\mathbb{1}_{A_\mathbf{s}}(\boldsymbol{\theta}_{\epsilon_1})\right] & \text{if } \ell = 1, \\ \mathbb{E}\left[\mathbb{1}_{A_\mathbf{s}}(\boldsymbol{\theta}_{\epsilon_\ell}) - \mathbb{1}_{A_\mathbf{s}}(\boldsymbol{\theta}_{\epsilon_{\ell-1}})\right] & \text{if } \ell > 1. \end{cases} \tag{6.7}$$

Using our notation, the MLMC estimator for Equation (6.6) and Equation (6.7) is given by

$$\hat{F}_{\epsilon_L}^{N_1,\ldots,N_L}(\mathbf{s}) = \sum_{\ell=1}^{L} \hat{Y}_\ell^{N_\ell}(\mathbf{s}), \tag{6.8}$$

where

$$\hat{Y}_\ell^{N_\ell}(\mathbf{s}) = \begin{cases} \frac{1}{N_1}\sum_{i=1}^{N_1} g_\mathbf{s}(\boldsymbol{\theta}_{\epsilon_1}^i) & \text{if } \ell = 1, \\ \frac{1}{N_\ell}\sum_{i=1}^{N_\ell}\left[g_\mathbf{s}(\boldsymbol{\theta}_{\epsilon_\ell}^i) - g_\mathbf{s}(\boldsymbol{\theta}_{\epsilon_{\ell-1}}^i)\right] & \text{if } \ell > 1, \end{cases} \tag{6.9}$$

and $g_\mathbf{s}(\boldsymbol{\theta})$ is a Lipschitz continuous approximation to the indicator function; this approximation is computed using a tensor product of cubic polynomials,

$$g_\mathbf{s}(\boldsymbol{\theta}) = \prod_{j=1}^{k} \xi\left(\frac{s_j - \theta_j}{\delta_j}\right),$$

where δ_j is the lattice spacing in the jth dimension, and $\xi(x)$ is a piece-wise continuous polynomial,

$$\xi(x) = \begin{cases} 1 & x \leq -1, \\ \frac{5}{8}x^3 - \frac{9}{8}x + \frac{1}{2} & -1 < x < 1, \\ 0 & x \geq 1. \end{cases}$$

This expression is based on the rigorous treatment of smoothing required in the univariate case given by Giles et al. (2015). While other polynomials that satisfy certain conditions are possible (Giles et al., 2015; Reiss, 1981), here we restrict ourselves to this relatively simple

form. Application of the smoothing function improves the quality of the final CDF estimate, just as using smoothing kernels improves the quality of PDF estimators (Silverman, 1986). Such a smoothing is also necessary to avoid convergence issues with MLMC caused by the discontinuity of the indicator function (Avikainen, 2009; Giles et al., 2015).

To compute the $\hat{Y}_1^{N_1}(\mathrm{s})$ term (Equation (6.9)), we generate N_1 samples $\boldsymbol{\theta}_{\epsilon_1}^1, \ldots, \boldsymbol{\theta}_{\epsilon_1}^{N_1}$ from $p(\boldsymbol{\theta}_{\epsilon_1})$; this represents a biased estimate for $F_{\epsilon_L}(\mathrm{s})$. To compensate for this bias, correction terms, $\hat{Y}_\ell^{N_\ell}(\mathrm{s})$, are evaluated for $\ell > 2$, each requiring the generation of N_ℓ samples $\boldsymbol{\theta}_{\epsilon_\ell}^1, \ldots, \boldsymbol{\theta}_{\epsilon_\ell}^{N_\ell}$ from $p(\boldsymbol{\theta}_{\epsilon_\ell})$ and N_ℓ samples $\boldsymbol{\theta}_{\epsilon_{\ell-1}}^1, \ldots, \boldsymbol{\theta}_{\epsilon_{\ell-1}}^{N_\ell}$ from $p(\boldsymbol{\theta}_{\epsilon_{\ell-1}})$, as given in Equation (6.9). It is important to note that the samples, $\boldsymbol{\theta}_{\epsilon_{\ell-1}}^1, \ldots, \boldsymbol{\theta}_{\epsilon_{\ell-1}}^{N_\ell}$, used to compute $\hat{Y}_\ell^{N_\ell}(\mathrm{s})$ are independent of the samples, $\boldsymbol{\theta}_{\epsilon_{\ell-1}}^1, \ldots, \boldsymbol{\theta}_{\epsilon_{\ell-1}}^{N_{\ell-1}}$, used to compute $\hat{Y}_{\ell-1}^{N_{\ell-1}}(\mathrm{s})$.

6.2.3 Variance reduction

The goal is to introduce a coupling between levels that controls the variance of the bias correction terms. With an effective coupling, the result is an estimator with lower variance, hence the number of samples required to obtain an accurate estimate is reduced. Denote v_ℓ as the variance of the estimator $\hat{Y}_\ell^{N_\ell}(\mathrm{s})$. For $\ell \geq 2$ this can be expressed as

$$\begin{aligned} v_\ell &= \mathbb{V}\left[g_{\mathbf{s}}(\boldsymbol{\theta}_{\epsilon_\ell}) - g_{\mathbf{s}}(\boldsymbol{\theta}_{\epsilon_{\ell-1}})\right] \\ &= \mathbb{V}\left[g_{\mathbf{s}}(\boldsymbol{\theta}_{\epsilon_\ell})\right] + \mathbb{V}\left[g_{\mathbf{s}}(\boldsymbol{\theta}_{\epsilon_{\ell-1}})\right] - 2 \cdot \mathbb{C}\left[g_{\mathbf{s}}(\boldsymbol{\theta}_{\epsilon_\ell}), g_{\mathbf{s}}(\boldsymbol{\theta}_{\epsilon_{\ell-1}})\right], \end{aligned}$$

where $\mathbb{V}[\cdot]$ and $\mathbb{C}[\cdot, \cdot]$ denote the variance and covariance, respectively. Introducing a positive correlation between the random variables $\boldsymbol{\theta}_{\epsilon_\ell}$ and $\boldsymbol{\theta}_{\epsilon_{\ell-1}}$ will have the desired effect of reducing the variance of $\hat{Y}_\ell^{N_\ell}(\mathrm{s})$.

In many applications of MLMC, a positive correlation is introduced through driving samplers at both the ℓ and $\ell - 1$ level with the same *randomness*. Properties of Brownian motion or Poisson processes are typically used for the estimation of expectations involving SDEs or Markov processes (Giles, 2008; Anderson and Higham, 2012; Lester et al., 2016). In the context of ABC methods, however, simulation of the quantity of interest is necessarily based on rejection sampling. The reliance on rejection sampling makes a strong coupling, in the true sense of MLMC, a difficult, if not impossible task. Rather, here we introduce a weaker form of coupling

through exploiting the fact that our MLMC estimator is performing the task of computing an estimate of the ABC posterior CDF. We combine this with a property of nested ABC rejection samplers to arrive at an efficient algorithm for computing $\hat{F}_{\epsilon_L}(\mathbf{s})$.

We proceed to establish a correlation between levels as follows. Assume we have computed, for some $\ell < L$, the terms $\hat{Y}_1^{N_1}(\mathbf{s}), \ldots, \hat{Y}_\ell^{N_\ell}(\mathbf{s})$ in Equation (6.8). That is, we have an estimator to the CDF at level ℓ by taking the sum

$$\hat{F}_{\epsilon_\ell}^{N_1,\ldots,N_\ell}(\mathbf{s}) \;=\; \sum_{m=1}^{\ell} \hat{Y}_m^{N_m}(\mathbf{s}),$$

with marginal distributions $\hat{F}_{\epsilon_\ell,j}^{N_1,\ldots,N_\ell}(s_j)$ for $j = 1, \ldots, k$. We can use this to determine a coupling based on matching marginal probabilities when computing $\hat{Y}_{\ell+1}(\mathbf{s})$. After generating $N_{\ell+1}$ samples $\boldsymbol{\theta}_{\epsilon_{\ell+1}}^1, \ldots, \boldsymbol{\theta}_{\epsilon_{\ell+1}}^{N_{\ell+1}}$ from $p(\boldsymbol{\theta}_{\epsilon_{\ell+1}})$, we compute the eCDF, $\hat{F}_{\epsilon_{\ell+1}}^{N_{\ell+1}}(\mathbf{s})$, using Equation (6.5) and obtain the marginal distributions, $\hat{F}_{\epsilon_{\ell+1},j}^{N_{\ell+1}}(s_j)$ for $j = 1, \ldots, k$. We can thus generate $N_{\ell+1}$ coupled pairs $\{\boldsymbol{\theta}_{\epsilon_{\ell+1}}^i, \boldsymbol{\theta}_{\epsilon_\ell}^i\}$ by choosing the $\boldsymbol{\theta}_{\epsilon_\ell}^i$ with the same marginal probabilities as the empirical probability of $\boldsymbol{\theta}_{\epsilon_{\ell+1}}^i$. That is, the jth component of $\boldsymbol{\theta}_{\epsilon_\ell}^i$ is given by

$$\theta_{\epsilon_\ell,j}^i = \hat{G}_{\epsilon_\ell}^{N_1,\ldots,N_\ell}\left(\hat{F}_{\epsilon_{\ell+1},j}^{N_{\ell+1}}(\theta_{\epsilon_{\ell+1},j}^i)\right),$$

where $\theta_{\epsilon_{\ell+1},j}^i$ is the jth component of $\boldsymbol{\theta}_{\epsilon_{\ell+1}}^i$ and $\hat{G}_{\epsilon_\ell,j}^{N_1,\ldots,N_\ell}(s)$ is the inverse of the jth marginal distribution of $\hat{F}_{\epsilon_\ell}^{N_1,\ldots,N_\ell}(\mathbf{s})$. This introduces a positive correlation between the sample pairs, $\{\boldsymbol{\theta}_{\epsilon_{\ell+1}}^i, \boldsymbol{\theta}_{\epsilon_\ell}^i\}$, since an increase in any of the components of $\boldsymbol{\theta}_{\epsilon_{\ell+1}}^i$ will cause an increase in the same component $\boldsymbol{\theta}_{\epsilon_\ell}^i$. This correlation reduces the variance in the bias correction estimator $\hat{Y}_{\epsilon_{\ell+1}}^{N_{\ell+1}}(\mathbf{s})$ computed according to Equation (6.9). We can then update the MLMC CDF to get an improved estimator by using

$$\hat{F}_{\epsilon_{\ell+1}}^{N_1,\ldots,N_{\ell+1}}(\mathbf{s}) = \hat{F}_{\epsilon_\ell}^{N_1,\ldots,N_\ell}(\mathbf{s}) + \hat{Y}_{\epsilon_{\ell+1}}^{N_{\ell+1}}(\mathbf{s}),$$

and apply an adjustment that ensures monotonicity. We continue this process iteratively to obtain $\hat{F}_{\epsilon_L}^{N_1,\ldots,N_L}(\mathbf{s})$.

It must be noted here that this coupling mechanism introduces an approximation for the general inference problem; therefore some additional bias can be introduced. This is made clear when one considers the process in terms of the copula distributions of $\boldsymbol{\theta}_{\epsilon_{\ell+1}}$ and $\boldsymbol{\theta}_{\epsilon_\ell}$. If these copula

distributions are the same, then the coupling is exact and there is no additional bias. The coupling is also exact for the univariate case ($k = 1$). Therefore, under the assumption that the correlation structure does not change significantly between levels, then the bias should be low. In practice, this requirement affects the choice of the acceptance threshold sequence, $\epsilon_1, \ldots, \epsilon_L$; we discuss this in more detail in Section 6.4. In Section 6.3.1, we demonstrate that for sensible choices of this sequence, the introduced bias is small compared with the bias that is inherent in the ABC approximation.

6.2.4 Optimal sample sizes

We now require the sample numbers $N_1, \ldots, N_L$ that are the optimal trade-off between accuracy and efficiency. Denote d_ℓ as the number of data generation steps required during the computation of $\hat{Y}_\ell^{N_\ell}(\mathbf{s})$ and let $c_\ell = d_\ell / N_\ell$ be the average number of data generation steps per accepted ABC posterior sample using acceptance threshold ϵ_ℓ.

Given $v_1 = \mathbb{V}\left[g_{\mathbf{s}}(\boldsymbol{\theta}_{\epsilon_1})\right]$ and $v_\ell = \mathbb{V}\left[g_{\mathbf{s}}(\boldsymbol{\theta}_{\epsilon_\ell}) - g_{\mathbf{s}}(\boldsymbol{\theta}_{\epsilon_{\ell-1}})\right]$, for $\ell > 1$, one can construct the optimal N_ℓ under the constraint $\mathbb{V}\left[\hat{F}_{\epsilon_L}^{N_1,\ldots,N_L}(\mathbf{s})\right] = \mathcal{O}(h^2)$, where h^2 is the target variance of the MLMC CDF estimator. As shown by Giles (2008), using a Lagrange multiplier method, the optimal $N_1, \ldots, N_L$ are given by

$$N_\ell = \mathcal{O}\left(h^{-2}\right) \sqrt{\frac{v_\ell}{c_\ell}} \sum_{m=1}^{L} \sqrt{v_m c_m}, \quad \ell = 1, \ldots, L. \tag{6.10}$$

In practice, the values for $v_1, \ldots, v_L$ and $c_1, \ldots, c_L$ will not have analytic expressions available; rather, we perform a low accuracy trial simulation with all $N_1 = \cdots = N_L = c$, for some comparatively small constant, c, to obtain the relative scaling of variances and data generation requirements.

6.2.5 Improving acceptance rates

A MLMC method based on the estimator in Equation (6.8) and the variance reduction strategy given in Section 6.2.3 would depend on standard ABC rejection sampling (Algorithm 6.1) for the final bias correction term $\hat{Y}_{\epsilon_L}^{N_L}$. For many ABC applications of interest, the computation of

this final term will dominate the computational costs. Therefore, the potential computational gains depend entirely on the size of N_L compared to the number of samples, N, required for the equivalent standard Monte Carlo approach (Equation (6.5)). While this approach is often an improvement over rejection sampling, we can achieve further computational gains by exploiting the iterative computation of the bias correction terms.

Let $\text{supp}(f(x))$ denote the support of a function $f(x)$, and note that, for any $\ell \in [2, L]$, $\text{supp}(p(\boldsymbol{\theta}_{\epsilon_\ell})) \subseteq \text{supp}(p(\boldsymbol{\theta}_{\epsilon_{\ell-1}}))$. This follows from the fact that if, for any $\boldsymbol{\theta}$, $\mathbb{P}(\rho(\mathcal{D}, \mathcal{D}_s) < \epsilon_\ell \mid \boldsymbol{\theta}) > 0$, then $\mathbb{P}(\rho(\mathcal{D}, \mathcal{D}_s) < \epsilon_{\ell-1} \mid \boldsymbol{\theta}) > 0$ since $\epsilon_\ell < \epsilon_{\ell-1}$. That is, any simulated data generated using parameter values taken from outside $\text{supp}(p(\boldsymbol{\theta}_{\epsilon_{\ell-1}}))$ cannot be accepted on level ℓ since $\rho(\mathcal{D}, \mathcal{D}_s) > \epsilon_{\ell-1}$ almost surely. Therefore, we can truncate the prior to the support of $p(\boldsymbol{\theta}_{\epsilon_{\ell-1}})$ when computing $\hat{Y}_\ell^{N_\ell}(\mathbf{s})$, thus increasing the acceptance rate of level ℓ samples. In practice, we need to approximate the support regions through sampling. For simplicity, in this work we restrict sampling of the prior at level ℓ to within the bounding box that contains all the samples generated at level $\ell - 1$. However, more sophisticated approaches could be considered, and may result in further computational improvements.

6.2.6 The algorithm

We now have all the components to construct our MLMC-ABC algorithm. We compute the MLMC CDF estimator (Equation (6.8)) using the coupling technique for the bias correction terms (Section 6.2.3) and prior truncation for improved acceptance rates (Section 6.2.5).

Optimal $N_1, \ldots, N_L$ are estimated as per Equation (6.10) and Section 6.2.4. Once $N_1, \ldots, N_L$ have been estimated, computation of the MLMC-ABC posterior CDF $\hat{F}_{\epsilon_L}(\mathbf{s})$ proceeds according to Algorithm 6.4.

The computational complexity of MLMC-ABC (Algorithm 6.4) is roughly $\mathcal{O}(c_S + N_L(c_L + c_G))$ where c_L is the expected cost of generating a single sample of the ABC posterior with threshold ϵ_L, c_S is the cost of updating the CDF estimate and c_G is the cost of the coupling which involves the marginal CDF inverses. From Algorithm 6.4, we have $c_S = \mathcal{O}(N_1|\mathcal{S}|)$ and from Barber et al. (2015) we have $c_L = \mathcal{O}(\epsilon^{-n})$ where n is the dimensionality of $\mathcal{D}$. The marginal inverse operations require only two steps:

Algorithm 6.4 MLMC-ABC

Initialise $\epsilon_1, \ldots, \epsilon_L$, $N_1, \ldots, N_L$ and prior $p(\boldsymbol{\theta})$.
Set $p(\boldsymbol{\theta}_{\epsilon_0}) \leftarrow p(\boldsymbol{\theta})$.
for $\ell = 1, \ldots, L$ **do**
  for $i = 1, \ldots, N_\ell$ **do**
    repeat
      Sample $\boldsymbol{\theta}^* \sim p(\boldsymbol{\theta})$ restricted to $\mathrm{supp}(p(\boldsymbol{\theta}_{\epsilon_{\ell-1}}))$.
      Generate data, $\mathcal{D}_s \sim s(\mathcal{D} \mid \boldsymbol{\theta}^*)$.
    until $\rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_\ell$.
    Set $\boldsymbol{\theta}^i_{\epsilon_\ell} \leftarrow \boldsymbol{\theta}^*$.
  end for
  for $\mathbf{s} \in \mathcal{S}$ **do**
    Set $\hat{F}^{N_\ell}_{\epsilon_\ell}(\mathbf{s}) \leftarrow \sum_{i=1}^{N_\ell} g_{\mathbf{s}}(\boldsymbol{\theta}^i_{\epsilon_\ell})/N_\ell$.
  end for
  if $\ell > 1$ **then**
    for $i = 1, \ldots, N_\ell$ **do**
      for $j = 1, \ldots, k$ **do**
        Set $\theta^i_{\epsilon_{\ell-1},j} \leftarrow \hat{G}^{N_1,\ldots,N_{\ell-1}}_{\epsilon_{\ell-1},j}\left(\hat{F}^{N_\ell}_{\epsilon_\ell,j}\left(\theta^i_{\epsilon_\ell,j}\right)\right)$.
      end for
    end for
    for $\mathbf{s} \in \mathcal{S}$ **do**
      Set $\hat{Y}^{N_\ell}_{\epsilon_\ell}(\mathbf{s}) \leftarrow \sum_{i=1}^{N_\ell}\left[g_{\mathbf{s}}(\boldsymbol{\theta}^i_{\epsilon_\ell}) - g_{\mathbf{s}}(\boldsymbol{\theta}^i_{\epsilon_{\ell-1}})\right]/N_\ell$.
      Set $\hat{F}^{N_1,\ldots,N_\ell}_{\epsilon_\ell}(\mathbf{s}) \leftarrow \hat{F}^{N_1,\ldots,N_{\ell-1}}_{\epsilon_{\ell-1}}(\mathbf{s}) + \hat{Y}^{N_\ell}_{\epsilon_\ell}(\mathbf{s})$.
    end for
  end if
end for

1. find $\mathbf{s} \in \mathcal{S}$ such that, $s_j \leq \theta^i_{\epsilon_\ell,j} < s_j + \delta_j$. Such an operation is, at most, $\mathcal{O}(\log_k |\mathcal{S}|)$;

2. invert the interpolating cubic spline, which can be done in $\mathcal{O}(1)$.

It follows that $c_G = \mathcal{O}(\log_k |\mathcal{S}|)$. For any practical application, ϵ_L will need to be sufficiently small, that is, $(N_1|\mathcal{S}|/N_L - \log_k |\mathcal{S}|) \ll c_L$, in order for the cost of generating posterior samples at level L to dominate the cost of the marginal CDF inverse operations and the lattice updating. Computational gains over ABC rejection sampling are achieved through decreasing N_L and c_L, via variance reduction and prior truncation.

Our primary focus in Algorithm 6.4 is on using MLMC to estimate the posterior CDF. The coupling mechanism is more readily communicated in this case. However, the MLMC-ABC method is more general and can be used to estimate $\mathbb{E}\left[U(\boldsymbol{\theta}_{\epsilon_\ell})\right]$ where $U(\boldsymbol{\theta}_{\epsilon_\ell})$ is any Lipschitz continuous function. In this more general case, only the marginal CDFs need be accumulated to facilitate the coupling mechanism. For more details see 6.5.2.

6.3 Results

In this section, we provide numerical results to demonstrate the validity, accuracy and performance of our MLMC-ABC method using some common models from epidemiology. In the first instance, we consider a tractable compartmental model, the stochastic SIS (Susceptible-Infected-Susceptible) model (Weiss and Dishon, 1971), to confirm the convergence and accuracy of MLMC-ABC. We then consider the Tuberculosis transmission model introduced by Tanaka et al. (2006) as a benchmark to compare our method with MCMC-ABC (Algorithm 6.2) and SMC-ABC (Algorithm 6.3). While we have chosen to perform our evaluation using an epidemiological model due to the particular prevalence of ABC methods in the life sciences, the techniques outlined in this manuscript are completely general and applicable to many areas of science.

6.3.1 A tractable example

The SIS model is a common model from epidemiology that describes the spread of a disease or infection for which no significant immunity is obtained after recovery; the common cold, for example. The model is given by

$$S + I \xrightarrow{\beta} 2I,$$
$$I \xrightarrow{\gamma} S,$$

with parameters $\boldsymbol{\theta} = \{\beta, \gamma\}$ and hazard functions for infection and recovery given by

$$h_I(S, I) = \beta SI \text{ and } h_R(S, I) = \gamma I,$$

respectively. This process defines a discrete-state continuous-time Markov process with a forward transitional density function that is computationally feasible to evaluate exactly for small populations $N_{pop} = S + I$. That is, for $t > s$, the probability of $S(t) = x$ given $S(s) = y$, with density denoted by $p(x, t \mid y, s; \beta, \gamma)$, has a solution obtained through the xth element of the vector

$$P(y, \beta, \gamma) = \exp(Q(\beta, \gamma)(t - s))\mathbf{y}, \tag{6.11}$$

where the yth element of column vector, $\mathbf{y}$, is one and all other elements zero, $\exp(\cdot)$ denotes the matrix exponential and $Q(\beta, \gamma)$ is the infinitesimal generator matrix of the Markov process; for the SIS model, $Q(\beta, \gamma)$ is a tri-diagonal matrix dependent only on the parameters of the model.

Let $S_{obs}(t)$ be an observation at time t of the number of susceptible individuals in the population. We generate observations using a single realisation of the SIS model with parameters $\beta = 0.003$ and $\gamma = 0.1$, population size $N_{pop} = 101$, and initial conditions $S(0) = 100$ and $I(0) = 1$; Observations are taken at times $t_1 = 4$, $t_2 = 8, \ldots,$ $t_{10} = 40$. Using the analytic solution to the SIS transitional density (using Equation (6.11)), we arrive at the likelihood function

$$\mathcal{L}(\beta, \gamma; \mathcal{D}) = \prod_{i=1}^{10} p(S_{obs}(t_i), t_i \mid S_{obs}(t_{i-1}), t_{i-1}; \beta, \gamma), \tag{6.12}$$

where $S_{obs}(t_0) = s_0$ almost surely. Hence, we can obtain an exact solution to the SIS posterior $p(\beta, \gamma \mid \mathcal{D})$ given the priors $\beta \sim \mathcal{U}(0, 0.06)$ and $\gamma \sim \mathcal{U}(0, 2)$. Given this exact posterior, quadrature can be applied to compute the posterior CDF, given by

$$F(s_1, s_2) = \int_{-\infty}^{s_1} \int_{-\infty}^{s_2} p(\beta, \gamma \mid \mathcal{D}) \, \mathrm{d}\beta \mathrm{d}\gamma.$$

For the ABC approximation we use a discrepancy metric based on the sum of squared errors,

$$\rho(\mathcal{D}, \mathcal{D}_s) = \left[\sum_{i=1}^{10} \left(S_{obs}(t_i) - S^*(t_i) \right)^2 \right]^{1/2},$$

where $S^*(t)$ is a realisation of the model generated using the Gillespie algorithm (Gillespie, 1977) for a given set of parameters, and $\mathcal{D}_s = [S^*(t_1), S^*(t_2), \ldots, S^*(t_{10})]$. An appropriate acceptance threshold sequence for this metric is $\epsilon_1, \ldots, \epsilon_L$, with $\epsilon_\ell = \epsilon_1 m^{1-\ell}$, $m = 2$ and $\epsilon_1 = 75$ (Toni et al., 2009).

We can use the exact likelihood (Equation (6.12)) to evaluate the convergence properties of our MLMC-ABC method. Based on the analysis of ABC rejection sampling by Barber et al. (2015), the rate of decay of the root mean-squared error (RMSE) as the computational cost is increased is slower for ABC than standard Monte Carlo. We find experimentally for ABC rejection sampling, that the decay of RMSE of the SIS model is approximately $\mathcal{O}(C^{-1/4})$,

where C is the expected computational cost of generating N ABC posterior samples. The 95% confidence interval of this experimental rate is $[-0.27, -0.23]$, which is consistent with the expected theoretical result of -0.25 (Barber et al., 2015). This is slower than the expected $\mathcal{O}(C^{-1/2})$ decay typically achieved with standard Monte Carlo. We compare the ABC rejection sampler decay rate with that achieved by our MLMC-ABC methods.

We use the L_∞ norm for the RMSE, that is,

$$\text{RMSE} = \sqrt{\mathbb{E}\left[\left\|F - \hat{F}_{\epsilon_L}\right\|_\infty^2\right]},$$

where F is the exact posterior CDF, evaluated directly from the likelihood (Equation (6.12)), and $\hat{F}_{\epsilon_L}$ is the Monte Carlo estimate of the CDF. A regular lattice that consists of 300×300 points is used for this estimate. The computation required for Gillespie simulations completely dominates the added computation of updating the lattice. The RMSE is computed using 20 independent MLMC-ABC CDF estimations for $L = 1, \ldots, 3$. The sequence of sample numbers, $N_1, \ldots, N_L$, is computed using Equation (6.10) with target variance of $\mathcal{O}(\epsilon^2)$ and 100 trial samples. Figure 6.1 demonstrates the improved convergence rate over the ABC rejection sampler convergence rate. Based on the least-squares fit, the RMSE decay is approximately $\mathcal{O}(C^{-1/3})$. The 95% confidence interval for the rate is $[-0.26, -0.34]$ which is consistent with theory from Giles et al. (2015) for the univariate situation. For smaller target RMSE, this results in an order of magnitude reduction in the computational cost.

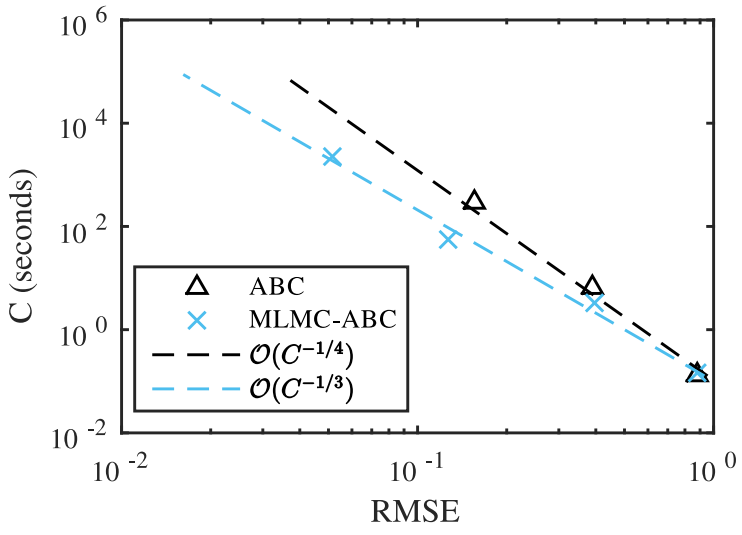

Figure 6.1: RMSE convergence for MLMC-ABC compared with ABC rejection sampling. The RMSE is computed using the exact solution to the posterior density of the SIS model.

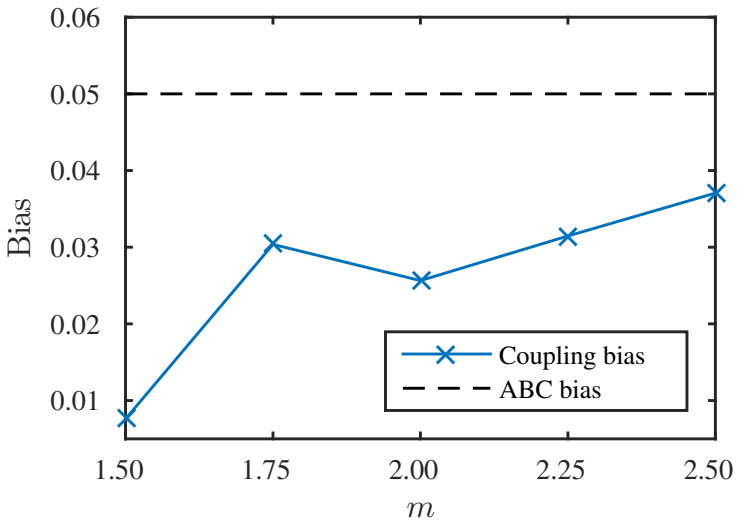

Figure 6.2: Convergence of coupling bias as a function of m.

We also consider the effect of the additional bias introduced through the coupling mechanism as presented in Section 6.2.3. We set the sample numbers at all levels to be 10^4 to ensure the Monte Carlo error is negligible and compare the bias for different values of acceptance threshold scaling factor m. The bias is computed according to the L_∞ norm, that is,

$$\text{Bias} = \mathbb{E}\left[\left\|\hat{F}^c_{\epsilon_L} - \hat{F}^u_{\epsilon_L}\right\|_\infty\right],$$

where $\hat{F}^c_{\epsilon_L}$ denotes the estimator computed according to Algorithm 6.4 and $\hat{F}^u_{\epsilon_L}$ denotes the estimator computed without any coupling. That is, $\hat{F}^u_{\epsilon_L}$ is computed using standard Monte Carlo to evaluate each term in the MLMC telescoping sum (Equation (6.6)). Note that, computationally, $\hat{F}^u_{\epsilon_L}$ will always be inferior to a standard Monte Carlo estimate. Figure 6.2 shows that, not only does the bias decay as m decreases, but also the additional bias is well within the order of bias expected from the ABC approximation. This result, along with Figure 6.1, demonstrates that the reduction in estimator variance can dominate the increase in bias. Thus, compared with standard Monte Carlo, a significantly lower RMSE for the same computational effort is achieved with MLMC using this coupling strategy.

6.3.2 Performance evaluation

We now evaluate the performance of MLMC-ABC using the model developed by Tanaka et al. (2006) in the study of tuberculosis transmission rate parameters using DNA fingerprint data (Small

et al., 1994). This model has been selected due to the availability of published comparative performance evaluations of MCMC-ABC and SMC-ABC (Tanaka et al., 2006; Sisson et al., 2007).

The model proposed by Tanaka et al. (2006) describes the occurrence of tuberculosis infections over time and the mutation of the bacterium responsible, *Myobacterium tuberculosis*. The number of infections, I, at time t is

$$I_t = \sum_{i=1}^{G_t} X_{i,t},$$

where G_t is the number of distinct genotypes and $X_{i,t}$ is the number of infections caused by the ith genotype at time t. For each genotype, new infections occur with rate α, infections terminate with rate δ, and mutation occurs with rate μ; causing an increase in the number of genotypes. This process, as with the SIS model, can be described by a discrete-state continuous-time Markov process. In this case, however, the likelihood is intractable, but the model can still be simulated using the Gillespie algorithm (Gillespie, 1977). After a realisation of the model is completed, either by extinction or when a maximum infection count is reached, a sub-sample of 473 cases is collected and compared against the IS6110 DNA fingerprint data of tuberculosis bacteria samples (Small et al., 1994). The dataset consists of 326 distinct genotypes; the infection cases are clustered according to the genotype responsible for the infection. The collection of clusters can be summarised succinctly as $30^1\,23^1\,15^1\,10^1\,8^1\,5^2\,4^4\,3^{13}\,2^{20}\,1^{282}$, where n^j denotes there are j clusters of size n. The discrepancy metric used is

$$\rho(\mathcal{D}, \mathcal{D}_s) = \frac{1}{n}\left|g(\mathcal{D}) - g(\mathcal{D}_s)\right| + \left|H(\mathcal{D}) - H(\mathcal{D}_s)\right|, \tag{6.13}$$

where n is the size of the population sub-sample (e.g., $n = 473$), $g(\mathcal{D})$ denotes the number of distinct genotypes in the dataset (e.g., $g(\mathcal{D}) = 326$) and the genetic diversity is

$$H(\mathcal{D}) = 1 - \frac{1}{n^2}\sum_{i=1}^{g(\mathcal{D})} n_i(\mathcal{D})^2,$$

where $n_i(\mathcal{D})$ is the cluster size of the ith genotype in the dataset.

We perform likelihood-free inference on the tuberculosis model for the parameters $\boldsymbol{\theta} = \{\alpha, \delta, \mu\}$ with the goal of evaluating the efficiencies of MLMC-ABC, MCMC-ABC and SMC-ABC. We use a target posterior distribution of $p(\boldsymbol{\theta} \mid \rho(\mathcal{D}, \mathcal{D}_s) < \epsilon)$ with $\rho(\mathcal{D}, \mathcal{D}_s)$ as defined in

Equation (6.13) and $\epsilon = 0.0025$. The acceptance threshold sequence, $\epsilon_1, \ldots, \epsilon_{10}$, used for both SMC-ABC and MLMC-ABC is $\epsilon_i = \epsilon_{10} + (\epsilon_{i-1} - \epsilon_{10})/2$ with $\epsilon_1 = 1$ and $\epsilon_{10} = 0.0025$. The improper prior is given by $\alpha \sim \mathcal{U}(0, 5)$, $\delta \sim \mathcal{U}(0, \alpha)$ and $\mu \sim \mathcal{N}(0.198, 0.06735^2)$ (Sisson et al., 2007; Tanaka et al., 2006). For the MCMC-ABC and SMC-ABC algorithms we apply a typical Gaussian proposal kernel,

$$q(\boldsymbol{\theta}^{(i)} \mid \boldsymbol{\theta}^{(i-1)}) = \mathcal{N}\left(\boldsymbol{\theta}^{(i-1)}, \Sigma\right),$$

with covariance matrix

$$\Sigma = \begin{bmatrix} 0.75^2 & 0 & 0 \\ 0 & 0.75^2 & 0 \\ 0 & 0 & 0.03^2 \end{bmatrix}. \tag{6.14}$$

Such a proposal kernel is reasonable to characterise the initial explorations of an ABC posterior as no correlations between parameters are assumed.

While MLMC-ABC does not require a proposal kernel function, some knowledge of the variance of each bias correction term is needed to determine the optimal sample numbers, $N_1, \ldots, N_L$. This is achieved using 100 trial samples of each level. The number of data generation steps is also recorded to compute $N_1, \ldots, N_L$, as per Equation (6.10). The resulting sample numbers are then scaled such that N_L is a user prescribed value.

The efficiency metric we use is the root mean-squared error (RMSE) of the CDF estimate versus the total number of data generation steps, N_s. The mean-squared error (MSE) is taken under the L_∞ norm, that is, $\text{MSE} = \mathbb{E}\left[\|F_{\epsilon_\ell} - \hat{F}_{\epsilon_\ell}\|_\infty^2\right]$ where F_{ϵ_ℓ} is the exact ABC posterior CDF and $\hat{F}_{\epsilon_\ell}$ is the Monte Carlo estimate using a regular lattice with $100 \times 100 \times 100$ points. To compute the RMSE, a high precision solution is computed using 10^6 ABC rejection samples of the target ABC posterior. This is computed over a period of 48 hours using 500 processor cores.

Table 6.1 presents the RMSE for MCMC-ABC, SMC-ABC and MLMC-ABC for different configurations. The RMSE values are computed using 10 independent estimator calculations. The algorithm parameter varied is the number of particles, N_P, for SMC-ABC, the number of

MLMC-ABC			MCMC-ABC			SMC-ABC		
N_L	N_s	RMSE	N_T	N_s	RMSE	N_P	N_s	RMSE
100	102, 246	0.1362	160, 000	163, 800	0.2434	400	347, 281	0.1071
200	255, 443	0.1136	320, 000	323, 841	0.1828	800	701, 100	0.0954
400	577, 312	0.1067	640, 000	643, 930	0.1832	1, 600	1, 385, 790	0.0858
800	1, 040, 140	0.0861	1, 280, 000	1, 283, 590	0.1345	3, 200	2, 792, 510	0.0784

Table 6.1: Comparison of MLMC-ABC against MCMC-ABC and SMC-ABC using a naive proposal kernel.

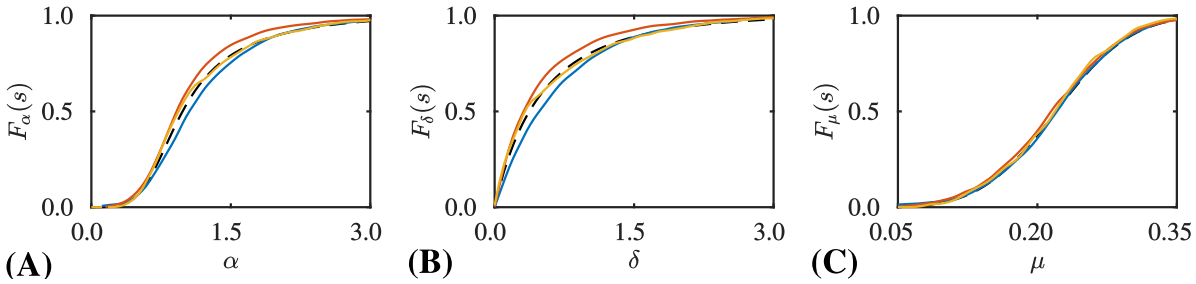

Figure 6.3: Estimated marginal CDFs–(A) α, (B) δ and (C) μ–for the tuberculosis transmission stochastic model. Estimate computed using using MLMC-ABC with $N_L = 800$ (solid yellow), MCMC-ABC over 1.2×10^6 iterations (solid blue), SMC-ABC with $3,200$ particles (solid red) and high precision solution (dashed black).

iterations, N_T, for MCMC-ABC and the level L sample number, N_L, for MLMC-ABC. Using the proposal kernel provided in Equation (6.14), SMC-ABC requires almost 30% more data generation steps than MLMC-ABC to obtain the same RMSE. MLMC-ABC obtains nearly double the accuracy of MCMC-ABC for the same number of data generation steps. Figure 6.3 shows an example of the high precision marginal posterior CDFs, $F_\alpha(s)$, $F_\delta(s)$ and $F_\mu(s)$, compared with the numerical solutions computed using the three methods.

We note that these results represent a typical scenario when solving this problem with a standard choice of proposal densities for MCMC-ABC and SMC-ABC. However, obtaining a good proposal kernel is a difficult open problem, and infeasible to do heuristically for high-dimensional parameter spaces. Therefore, efficient proposal kernels are almost never obtained without significant manual adjustment or additional algorithmic modifications such as adaptive schemes (Beaumont et al., 2009; Del Moral et al., 2012; Roberts and Rosenthal, 2009). Nevertheless, we demonstrate in Table 6.2 the increased efficiency for MCMC-ABC and SMC-ABC when using a highly configured Gaussian proposal kernel with covariance matrix as determined by Tanaka

MLMC-ABC			MCMC-ABC			SMC-ABC		
N_L	N_s	RMSE	N_T	N_s	RMSE	N_P	N_s	RMSE
100	$102,246$	0.1362	$160,000$	$161,604$	0.1718	400	$102,902$	0.1270
200	$255,443$	0.1136	$320,000$	$322,962$	0.1254	800	$198,803$	0.0767
400	$577,312$	0.1067	$640,000$	$642,412$	0.1127	$1,600$	$402,334$	0.0652
800	$1,040,140$	0.0861	$1,280,000$	$1,305,340$	0.0934	$3,200$	$797,893$	0.0560

Table 6.2: Comparison of MLMC-ABC against MCMC-ABC and SMC-ABC using heuristically chosen proposal densities.

et al. (2006)

$$\Sigma = \begin{bmatrix} 0.5^2 & 0.225 & 0 \\ 0.225 & 0.5^2 & 0 \\ 0 & 0 & 0.015^2 \end{bmatrix}. \tag{6.15}$$

We note that Sisson et al. (2007) use a similar proposal kernel, however they do not explicitly state the difference in the covariance matrix Σ; we therefore assume that Equation (6.15) represents a proposal kernel that is heuristically optimal to the target posterior density. We emphasise that it would be incredibly rare to arrive at an optimal proposal kernel without any additional experimentation. Even in this unlikely case, MLMC-ABC is still comparable with MCMC-ABC. However, MLMC-ABC is clearly not as efficient as SMC-ABC for this heuristically optimised scenario. The scenario is intentionally biased toward MCMC-ABC and SMC-ABC, so this result is not unexpected. Future research could consider a comparison of the methods when the extra computational burden of determining the optimal Σ, such as implementation of an adaptive scheme, is taken into account.

6.4 Discussion

Our results indicate that, while SMC-ABC and MCMC-ABC can be heuristically optimised to be highly efficient, an accurate estimate of the parameter posterior can be obtained using the MLMC-ABC method presented here in a relatively automatic fashion. Furthermore, the efficiency of MLMC-ABC is comparable or improved over MCMC-ABC and SMC-ABC, even

in the case when efficient proposal kernels are employed.

The need to estimate the variances of each bias correction term could be considered a limitation of the MLMC-ABC approach. However, we find in practice that these need not be computed to high accuracy and can often be estimated with a relatively small number of samples. There could be examples of Bayesian inference problems where MLMC-ABC is inefficient on account of the variance estimation inaccuracy. We have so far, however, failed to find an example for which 100 samples of each bias correction term is insufficient to obtain a good MLMC-ABC estimator.

There are many modifications one could consider to further improve MLMC-ABC. In this work, we explicitly specify the sequence of acceptance thresholds in advance for both MLMC-ABC and SMC-ABC. As this is the initial presentation of the method, it is appropriate to consider this idealised case. However, it is unclear if an optimal sequence of thresholds for SMC-ABC will also be optimal for MLMC-ABC and vice versa. Furthermore, as mention in Section 6.1, practical applications of SMC-ABC often determine such sequences adaptively (Drovandi and Pettitt, 2011; Silk et al., 2013). A modification of MLMC-ABC to allow for adaptive acceptance thresholds would make MLMC-ABC even more practical as it could be used to minimise the coupling bias. The exact mechanisms to achieve this could be based on similar ideas to adaptive SMC-ABC (Drovandi and Pettitt, 2011). Given the solution at level $\ell - 1$, the next level ℓ could be determined through: (i) sampling a fixed set of *prior* samples; (ii) sorting these samples based on the discrepancy metric; (iii) selecting a discrepancy threshold, ϵ_ℓ, that optimises coupling bias and the variance of the bias correction term. Future work should address these open problems.

Other improvements could focus on the discretisation used for the eCDF calculations. The MLMC-ABC method has no requirement of a regular lattice, and alternative choices would likely enable MLMC-ABC to scale to much higher dimensional parameter spaces. Adaptive grids that refine with each level is one option that could be considered; however, unstructured grids or kernel based methods also have potential.

Improvements to the coupling scheme are also possible avenues for future consideration. The coupling approach we have considered depends only on the computation of the marginal posterior CDFs and assumes nothing about the underlying model. It may be possible to take

advantage of certain model specific features to improve variance reduction. There may also be cases where the rejection sampling scheme is prohibitive for the more accurate acceptance levels. The combination of our coupling scheme with the MLSMC scheme recently proposed by Beskos et al. (2017) and Jasra et al. (2019) is a promising possibility to mitigate these issues. Future work should investigate, compare and contrast the variety of available coupling strategies in the growing body of literature on MLMC for ABC and Bayesian inverse problems.

We have shown, in a practical way, how MLMC techniques can be applied to ABC inference. We also demonstrate that variance reduction strategies, even when applied to simple methods such as rejection sampling, can achieve performance improvements comparable, and in some cases superior, to modern advanced ABC methods based on MCMC and SMC methods. Therefore, the MLMC framework is a promising area for designing improved samplers for complex statistical problems with intractable likelihoods.

6.5 Supplementary material

6.5.1 Derivation of optimal number of samples per level

In this section, we demonstrate how the sequence of sample numbers $N_1, \dots, N_L$ is obtained, as stated in Equation (6.10). First, we assume the variances $v_1 = \mathbb{V}\left[g_{\mathbf{s}}(\boldsymbol{\theta}_{\epsilon_1})\right]$ and $v_\ell = \mathbb{V}\left[g_{\mathbf{s}}(\boldsymbol{\theta}_{\epsilon_\ell}) - g_{\mathbf{s}}(\boldsymbol{\theta}_{\epsilon_{\ell-1}})\right]$, for $\ell > 1$, are known quantities. We also assume that, on average, $d_\ell = c_\ell N_\ell$ data generation steps are required to compute the estimator $\hat{Y}_{\epsilon_\ell}^{N_\ell}(\mathbf{s})$.

Let $C(N_1, \dots, N_L)$ denote the total number of data generation steps to compute $\hat{F}_{\epsilon_L}^{N_1,\dots,N_L}(\mathbf{s})$ given a choice of $N_1, \dots, N_L$. Likewise, let $E(N_1, \dots, N_L)$ denote the Monte Carlo error in the estimator $\hat{F}_{\epsilon_L}^{N_1,\dots,N_L}(\mathbf{s})$.

The sequence $N_1, \dots, N_L$ is considered optimal if $C(N_1, \dots, N_L)$ is minimised subject to the constraint $E(N_1, \dots, N_L) \leq Kh^2$ for some constant $K > 0$ and target Monte Carlo error h^2. To determine optimal $N_1, \dots, N_L$, we consider the Lagrangian

$$\begin{aligned}\mathcal{G}(N_1, \dots, N_L, \lambda) = &C(N_1, \dots, N_L) \\ &+ \lambda\left(E(N_1, \dots, N_L) - Kh^2\right),\end{aligned}$$

and note that solutions to the optimisation problem exist at $\nabla\mathcal{G}(N_1, \dots, N_l, \lambda) = \mathbf{0}$. That is,

$$\nabla\mathcal{G}(N_1, \dots, N_l, \lambda) = \sum_{\ell=1}^{L} \frac{\partial\mathcal{G}}{\partial N_\ell}\mathbf{e}_\ell + \frac{\partial\mathcal{G}}{\partial\lambda}\mathbf{e}_{L+1} = \mathbf{0},$$

$$\sum_{\ell=1}^{L}\left[\frac{\partial C}{\partial N_\ell} + \lambda\frac{\partial E}{\partial N_\ell}\right]\mathbf{e}_\ell + \left[E(N_1, \dots, N_L) - Kh^2\right]\mathbf{e}_{L+1} = \mathbf{0},$$

where $\mathbf{e}_1, \dots, \mathbf{e}_{L+1}$ are the standard orthonormal basis vectors of $(L+1)$-dimensional Euclidean space. We obtain the following system of equations,

$$\frac{\partial C}{\partial N_\ell} = -\lambda\frac{\partial E}{\partial N_\ell},\ \ell = 1, \dots L, \tag{6.16}$$

$$E(N_1, \dots, N_L) = Kh^2. \tag{6.17}$$

First we consider the forms of $C(N_1, \dots, N_L)$ and $E(N_1, \dots, N_L)$. By definition of the MLMC

estimator given in Section 6.2.2, we have

$$E(N_1, \ldots, N_L) = \mathbb{V}\left[\hat{F}_{\epsilon_L}^{N_1,\ldots,N_L}(\mathbf{s})\right] = \mathbb{V}\left[\sum_{\ell=1}^{L} \hat{Y}_\ell^{N_\ell}(\mathbf{s})\right].$$

Furthermore, since $\hat{Y}_1(\mathbf{s}), \ldots, \hat{Y}_L(\mathbf{s})$ are independent,

$$\begin{aligned}
\mathbb{V}\left[\sum_{\ell=1}^{L} \hat{Y}_\ell^{N_\ell}(\mathbf{s})\right] &= \sum_{\ell=1}^{L} \mathbb{V}\left[\hat{Y}_\ell^{N_\ell}(\mathbf{s})\right] \\
&= \mathbb{V}\left[\frac{1}{N_1}\sum_{i=1}^{N_1} g_{\mathbf{s}}(\boldsymbol{\theta}_{\epsilon_1}^i)\right] + \sum_{\ell=2}^{L} \mathbb{V}\left[\frac{1}{N_\ell}\sum_{i=1}^{N_\ell} g_{\mathbf{s}}(\boldsymbol{\theta}_{\epsilon_\ell}^i) - g_{\mathbf{s}}(\boldsymbol{\theta}_{\epsilon_{\ell-1}}^i)\right] \\
&= \sum_{\ell=1}^{L} \frac{1}{N_\ell^2}\sum_{i=1}^{N_\ell} v_\ell.
\end{aligned}$$

That is,

$$E(N_1, \ldots, N_L) = \sum_{\ell=1}^{L} \frac{v_\ell}{N_\ell}. \tag{6.18}$$

The total number of data generation steps is simply

$$C(N_1, \ldots, N_L) = \sum_{\ell=1}^{L} d_\ell = \sum_{\ell=1}^{L} c_\ell N_\ell. \tag{6.19}$$

Substitution of Equation (6.18) and Equation (6.19) into Equation (6.16) yields

$$N_\ell = \sqrt{\lambda \frac{v_\ell}{c_\ell}}. \tag{6.20}$$

Substitution of Equation (6.18) and Equation (6.20) into Equation (6.17) allows us to obtain λ,

$$\sum_{\ell=1}^{L} v_\ell \left(\sqrt{\lambda \frac{v_\ell}{c_\ell}}\right)^{-1} = Kh^2 \Rightarrow \sqrt{\lambda} = \frac{1}{Kh^2}\sum_{\ell=1}^{L} \sqrt{v_\ell c_\ell}. \tag{6.21}$$

Finally, through substitution of Equation (6.21) back into Equation (6.20) we find that the optimal $N_1, \ldots, N_\ell$ is given by,

$$N_\ell = \frac{1}{Kh^2}\sqrt{\frac{v_\ell}{c_\ell}}\sum_{m=1}^{L} \sqrt{v_m c_m}, \tag{6.22}$$

as required.

6.5.2 MLMC for ABC with general Lipschitz functions

Minor modifications of the MLMC-ABC method (Algorithm 6.4) are possible to enable the computation of expectations of the form

$$\mathbb{E}\left[U(\boldsymbol{\theta}) \mid \rho(\mathcal{D}, \mathcal{D}_s) < \epsilon\right] = \int_{\boldsymbol{\Theta}} U(\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \rho(\mathcal{D}, \mathcal{D}_s) < \epsilon) \, \mathrm{d}\theta_k \ldots \mathrm{d}\theta_1,$$

where $U(\boldsymbol{\theta})$ is any Lipschitz continuous function.

Using the same notation as defined in Section 6.2.2, the MLMC telescoping sum may be formed for a given sequence of acceptance thresholds, $\epsilon_1 > \cdots > \epsilon_L$, to compute the expectation $E_{\epsilon_L} = \mathbb{E}\left[U(\boldsymbol{\theta}_{\epsilon_L})\right]$. That is,

$$E_{\epsilon_L} = \sum_{\ell=1}^{L} P_{\epsilon_\ell}, \quad P_{\epsilon_\ell} = \begin{cases} \mathbb{E}\left[U(\boldsymbol{\theta}_{\epsilon_1})\right], & \ell = 1, \\ \mathbb{E}\left[U(\boldsymbol{\theta}_{\epsilon_\ell}) - U(\boldsymbol{\theta}_{\epsilon_{\ell-1}})\right], & \ell > 1. \end{cases} \tag{6.23}$$

From Equation (6.23), it is straightforward to obtain equivalent expressions to Equation (6.8) and Equation (6.9).

The modified MLMC-ABC algorithm proceeds in a very similar fashion to Algorithm 6.4, however, there is no need to hold a complete discretisation of the parameter space, $\boldsymbol{\Theta}$. This is because only the k marginal CDFs, $F_{\epsilon_\ell,1}\left(s_1\right), \ldots, F_{\epsilon_\ell,k}\left(s_k\right)$, are required to form the coupling strategy in Section 6.2.3. Thus, we denote $\mathcal{S}_j$ to be a discretisation of the jth coordinate axis. This significantly reduces the computational burden of the lattice in higher dimensions since $|\mathcal{S}_j| = \mathcal{O}(\log_k |\mathcal{S}|)$. The resulting algorithm is given in Algorithm 6.5.

Algorithm 6.5 Modified MLMC-ABC

Initialise $\epsilon_1, \ldots, \epsilon_L$, $N_1, \ldots, N_L$ and prior $p(\boldsymbol{\theta})$.
Set $p(\boldsymbol{\theta}_{\epsilon_0}) \leftarrow p(\boldsymbol{\theta})$.
for $\ell = 1, \ldots, L$ **do**
  for $i = 1, \ldots, N_\ell$ **do**
    repeat
      Sample $\boldsymbol{\theta}^* \sim p(\boldsymbol{\theta})$ restricted to $\text{supp}(p(\boldsymbol{\theta}_{\epsilon_{\ell-1}}))$.
      Generate data, $\mathcal{D}_s \sim s(\mathcal{D} \mid \boldsymbol{\theta}^*)$.
    until $\rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_\ell$.
    Set $\boldsymbol{\theta}^i_{\epsilon_\ell} \leftarrow \boldsymbol{\theta}^*$.
  end for
  for $j = 1, \ldots, k$ **do**
    for $s \in \mathcal{S}_j$ **do**
      Set $\hat{F}^{N_\ell}_{\epsilon_\ell,j}(s) \leftarrow \sum_{i=1}^{N_\ell} \xi\left((s_j - \theta^i_{\epsilon_\ell,j})/\delta_j\right)/N_\ell$.
    end for
  end for
  if $\ell = 1$ **then**
    $\hat{E}^{N_\ell}_{\epsilon_\ell} \leftarrow \sum_{i=1}^{N_\ell} U(\boldsymbol{\theta}^i_{\epsilon_\ell})/N_\ell$
  else
    for $i = 1, \ldots, N_\ell$ **do**
      for $j = 1, \ldots, k$ **do**
        Set $\theta^i_{\epsilon_{\ell-1},j} \leftarrow \hat{G}^{N_1,\ldots,N_{\ell-1}}_{\epsilon_{\ell-1},j}\left(\hat{F}^{N_\ell}_{\epsilon_\ell,j}\left(\theta^i_{\epsilon_\ell,j}\right)\right)$.
      end for
    end for
    for $j = 1, \ldots, k$ **do**
      for $s \in \mathcal{S}_j$ **do**
        Set $\hat{Y}^{N_\ell}_{\epsilon_\ell,j}(s) \leftarrow \sum_{i=1}^{N_\ell}\left[\xi\left((s_j - \theta^i_{\epsilon_\ell,j})/\delta_j\right) - \xi\left((s_j - \theta^i_{\epsilon_{\ell-1},j})/\delta_j\right)\right]/N_\ell$.
        Set $\hat{F}^{N_1,\ldots,N_\ell}_{\epsilon_\ell,j}(s) \leftarrow \hat{F}^{N_1,\ldots,N_{\ell-1}}_{\epsilon_{\ell-1},j}(s) + \hat{Y}^{N_\ell}_{\epsilon_\ell,j}(s)$.
      end for
    end for
    $\hat{P}^{N_\ell}_{\epsilon_\ell} \leftarrow \sum_{i=1}^{N_\ell}\left[U(\boldsymbol{\theta}^i_{\epsilon_\ell}) - U(\boldsymbol{\theta}^i_{\epsilon_{\ell-1}}))\right]/N_\ell$
    $\hat{E}^{N_1,\ldots,N_\ell}_{\epsilon_\ell} \leftarrow \hat{E}^{N_1,\ldots,N_{\ell-1}}_{\epsilon_{\ell-1}} + \hat{P}^{N_\ell}_{\epsilon_\ell}$.
  end if
end for

Chapter 7

Rapid Bayesian Inference for Expensive Stochastic Models

A paper submitted to *Journal of Computational and Graphical Statistics*.

David J. Warne, Ruth E. Baker and Matthew J. Simpson (2019). Rapid Bayesian inference for expensive stochastic models. *ArXiv Pre-print*, arXiv:1909.06540 [stat.CO]

Abstract Almost all fields of science rely upon statistical inference to estimate unknown parameters in theoretical and computational models. While the performance of modern computer hardware continues to grow, the computational requirements for the simulation of models are growing even faster. This is largely due to the increase in model complexity, often including stochastic dynamics, that is necessary to describe and characterize phenomena observed using modern, high resolution, experimental techniques. Such models are rarely analytically tractable, meaning that extremely large numbers of stochastic simulations are required for parameter inference. In such cases, parameter inference can be practically impossible. In this work, we present new computational Bayesian techniques that accelerate inference for expensive stochastic models by using computationally inexpensive approximations to inform feasible regions in parameter space, and through learning transforms that adjust the biased approximate inferences to closer represent the correct inferences under the expensive stochastic model.

Using topical examples from ecology and cell biology, we demonstrate a speed improvement of an order of magnitude without any loss in accuracy.

7.1 Introduction

Modern experimental techniques allow us to observe the natural world in unprecedented detail and resolution (Chen et al., 2014). Advances in machine learning and artificial intelligence provide many new techniques for pattern recognition and prediction, however, in almost all scientific inquiry there is a need for detailed mathematical models to provide mechanistic insight into the phenomena observed (Baker et al., 2018; Coveney et al., 2016). This is particularly true in the biological and ecological sciences, where detailed stochastic models are routinely applied to develop and validate theory as well as interpret and analyze data (Black and McKane, 2012; Drawert et al., 2017; Wilkinson, 2009).

Two distinct computational challenges arise when stochastic models are considered, they are: (i) the *forwards problem*; and (ii) the *inverse problem*, sometimes called the *backwards problem* Chapter 4. While the computational generation of a single sample path, that is the *forwards problem*, may be feasible, hundreds or thousands or more of such sample paths may be required to gain insight into the range of possible model predictions and to conduct parameter sensitivity analysis (Gunawan et al., 2005; Lester et al., 2017; Marino et al., 2008). The problem is further compounded if the models must be calibrated using experimental data, that is the *inverse problem* of parameter estimation, since millions of sample paths may be necessary.

In many cases, the forwards problem can be sufficiently computationally expensive to render both parameter sensitivity analysis and the inverse problem completely intractable, despite recent advances in computational inference (Sisson et al., 2018). This has prompted recent interest in the use of mathematical approximations to circumvent the computational burden, both in the context of the forwards and inverse problems. For example, linear approximations are applied to the forwards problem of chemical reaction networks with bimolecular and higher-order reactions (Cao and Grima, 2018), and various approximations, including surrogate models (Rynn et al., 2019), emulators (Buzbas and Rosenberg, 2015) and transport maps (Parno and Marzouk, 2018), are applied to inverse problems with expensive forwards models, for example, in the study of climate science (Holden et al., 2018). Furthermore, a number of developments,

such as multilevel Monte Carlo methods (Giles, 2015), have demonstrated that families of approximations can be combined to improve computational performance without sacrificing accuracy.

In recent years, the Bayesian approach to the inverse problem of model calibration and parameter inference has been particularly successful in many fields of science including, astronomy (The Event Horizon Telescope Collaboration et al., 2019), anthropology and archaeology (King et al., 2014; Malaspinas et al., 2016), paleontology and evolution (O'Dea et al., 2016; Pritchard et al., 1999; Tavaré et al., 1997), epidemiology (Liu et al., 2018), biology (Lawson et al., 2018; Guindani et al., 2014; Woods and Barnes, 2016; Vo et al., 2015a), and ecology (Ellison, 2004; Stumpf, 2014). For complex stochastic models, parameterized by $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, computing the likelihood of observing data, $\mathcal{D} \in \mathbb{D}$, is almost always impossible (Browning et al., 2018; Johnston et al., 2014; Vankov et al., 2019). Thus, approximate Bayesian computation (ABC) methods (Sisson et al., 2018) are essential. ABC methods replace likelihood evaluation with an approximation based on stochastic simulations of the proposed model, this is captured directly in *ABC rejection sampling* (Beaumont et al., 2002; Pritchard et al., 1999; Tavaré et al., 1997) (Section 7.2) where samples are generated from an approximate posterior using stochastic simulations of the forwards problem as a replacement for the likelihood.

Unfortunately, ABC rejection sampling can be computationally expensive or even completely prohibitive, especially for high-dimensional parameter spaces, since a very large number of stochastic simulations are required to generate enough samples from the approximate Bayesian posterior distribution (Sisson et al., 2018) (Chapter 5). This is further compounded when the forwards problem is computationally expensive. In contrast, an appropriately chosen approximate model may yield a tractable likelihood that removes the need for ABC methods (Browning et al., 2019) (Chapters 2 and 3). This highlights a key advantage of such approximations because no ABC sampling is required. However, approximations can perform poorly in terms of their predictive capability, and inference based on such models will always be biased, with the extent of the bias dependent on the level of accuracy.

We consider ABC-based inference algorithms for the challenging problem of parameter inference for computationally expensive stochastic models when an appropriate approximation is available to inform the search in parameter space. Under our approach, the approximate model need not be quantitatively accurate in terms of the forwards problem, but must qualitatively

respond to changes in parameter values in a similar way to the stochastic model. In particular, we extend the sequential Monte Carlo ABC sampler (SMC-ABC) of Sisson et al. (2007) (Section 7.2) to exploit the approximate model in two ways: (i) to generate an intermediate proposal distribution, that we call a *preconditioner*, to improve ABC acceptance rates for the stochastic model; and (ii) to construct a biased ABC posterior, then reduce this bias using a *moment-matching* transform. We describe both methods and then present relevant examples from ecology and cell biology. Example calculations demonstrate that our methods generate ABC posteriors with a significant reduction in the number of required expensive stochastic simulations, leading to as much as a tenfold computational speedup.

As a motivating case study for this work, we focus on stochastic models that can replicate many spatiotemporal patterns that naturally arise in biological and ecological systems. Stochastic discrete random walk models (Section 7.3), henceforth called *discrete models*, can accurately characterize the microscale interactions of individual agents, such as animals, plants, microorganisms, and cells (Agnew et al., 2014; Codling et al., 2008; von Hardenberg et al., 2001; Law et al., 2003; Taylor and Hastings, 2005; Vincenot et al., 2016). Mathematical modeling of populations as complex systems of agents can enhance our understanding of real biological and ecological populations with applications in cancer treatment (Böttger et al., 2015), wound healing (Callaghan et al., 2006), wildlife conservation (McLane et al., 2011; DeAngelis and Grimm, 2014), and the management of invasive species (Chkrebtii et al., 2015; Taylor and Hastings, 2005).

For example, the discrete model formulation can replicate many realistic spatiotemporal patterns observed in cell biology. Figure 7.1(A),(B) demonstrates typical microscopy images obtained from *in vitro* cell culture assays; ubiquitous and important experimental techniques used in the study of cell motility, cell proliferation and drug design. Various patterns are observed: prostate cancer cells (PC-3 line) tend to be highly motile, and spread uniformly to invade vacant spaces (Figure 7.1(A)); in contrast breast cancer cells (MBA-MD-231 line) tend be relatively stationary with proliferation events driving the formation of aggregations (Figure 7.1(B)). These phenomena may be captured using a lattice-based discrete model framework by varying the ratio P_p/P_m where $P_p \in [0,1]$ and $P_m \in [0,1]$ are, respectively, the probabilities that an agent attempts to proliferate and attempts to move during a time interval of duration $\tau > 0$. For $P_p/P_m \ll 1$, behavior akin to PC-3 cells is recovered (Figure 7.1(C)–(F)) (Jin et al., 2016a).

Setting $P_p/P_m \gg 1$, as in Figure 7.1(H)–(K), leads to clusters of occupied lattices sites that are similar to the aggregates of MBA-MD-231 cells (Agnew et al., 2014; Simpson et al., 2013).

It is common practice to derive approximate continuum-limit differential equation descriptions of discrete models (Callaghan et al., 2006; Jin et al., 2016a; Simpson et al., 2010) (Section 7.5.2). Such approximations provide a means of performing analysis with significantly reduced computational requirements, since evaluating an exact analytical solution, if available, or otherwise numerically solving a differential equation is typically several orders of magnitude faster than generating a single realization of the discrete model, of which hundreds or thousands may be required for reliable ABC sampling (Browning et al., 2018). However, such approximations are generally only valid within certain parameter regimes, for example here when $P_p/P_m \ll 1$ (Callaghan et al., 2006; Simpson et al., 2010). Consider Figure 7.1(G), the population density growth curve from the continuum-limit logistic growth model is superimposed with stochastic data for four realizations of a discrete model with $P_p/P_m \ll 1$ and $P_p/P_m \gg 1$, under initial conditions simulating a proliferation assay, where each lattice site is randomly populated with constant probability, such that there are no macroscopic gradients present at $t = 0$. The continuum-limit logistic growth model is an excellent match for the $P_p/P_m \ll 1$ case (Figure 7.1(C)–(F)), but severely overestimates the population density when $P_p/P_m \gg 1$ since the mean-field assumptions underpinning the continuum-limit model are violated by the presence of clustering (Figure 7.1(H)–(K)) (Agnew et al., 2014; Simpson et al., 2013).

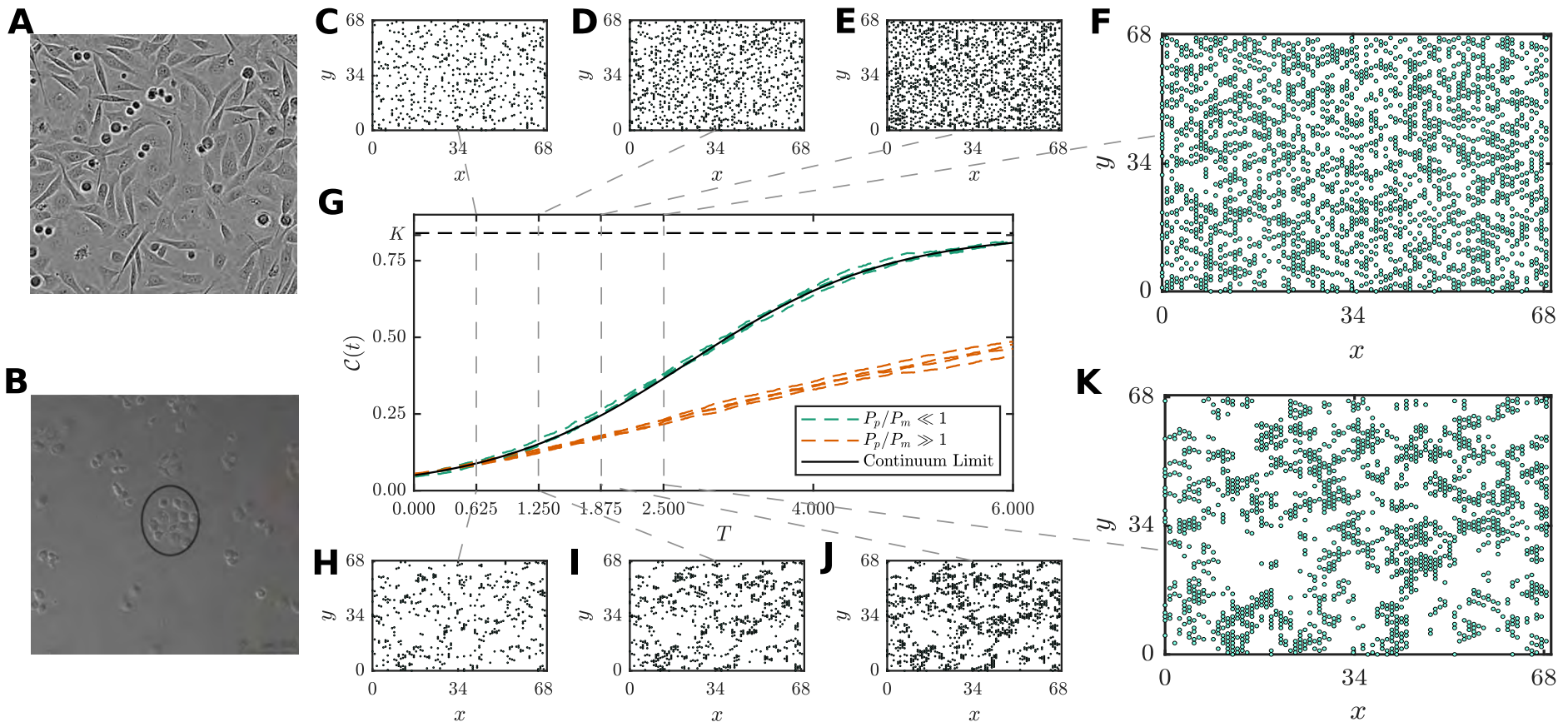

Figure 7.1: Discrete random walk models can replicate observed spatial patterns in cell culture: (A) PC-3 prostate cancer cells (reprinted from Jin et al. (2017) with permission); and (B) MBA-MD-231 breast cancer cells (reprinted from Simpson et al. (2013) with permission). (C)–(F) Discrete simulations with $P_p/P_m \ll 1$ replicate the uniform distribution of (A) PC-3 cells. (H)–(K) Discrete simulations with $P_p/P_m \gg 1$ replicate spatial clustering of (B) MBA-MD-231 cells. (G) Averaged population density profiles $\mathcal{C}(t)$ for the discrete model with highly motile agents, $P_m = 1$ (dashed green), and near stationary agents, $P_m = 5 \times 10^{-4}$ (dashed orange), compared with the logistic growth continuum limit (solid black), time is non-dimensionalised with $T = P_p t/\tau$.

As we demonstrate in Section 7.3, our methods generate accurate ABC posteriors for inference on the discrete problem for a range of biologically relevant parameter regimes, including those where the continuum-limit approximation is poor. In this respect we demonstrate a novel use of approximations that qualitatively respond to changes in parameters in a similar way to the full exact stochastic model.

7.2 Methods

In this section, we present details of two new algorithms for the acceleration of ABC inference for expensive stochastic models when an appropriate approximation is available. First, we present essential background in ABC inference and sequential Monte Carlo (SMC) samplers for ABC (Sisson et al., 2007; Toni et al., 2009). We then describe our extensions to SMC samplers for ABC and provide numerical examples of our approaches using topical examples from ecology and cell biology.

7.2.1 Sequential Monte Carlo for Approximate Bayesian computation

Bayesian analysis techniques are powerful tools for the quantification of uncertainty in parameters, models and predictions (Gelman et al., 2014). Unfortunately, for many stochastic models of practical interest, the likelihood function is intractable. ABC methods replace likelihood evaluation with an approximation based on stochastic simulations of the proposed model, this is captured directly in *ABC rejection sampling* (Pritchard et al., 1999; Tavaré et al., 1997) where $\mathcal{M}$ samples are generated from an approximate posterior, denoted by $p(\boldsymbol{\theta} \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon)$. Here: $\mathcal{D}_s \sim s(\mathcal{D} \mid \boldsymbol{\theta})$ is a data generation process based on simulation of the model; $\rho(\mathcal{D}, \mathcal{D}_s)$ is a discrepancy metric; and ϵ is the discrepancy threshold. The resulting accepted parameter samples are distributed according to $p(\boldsymbol{\theta} \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon) \to p(\boldsymbol{\theta} \mid \mathcal{D})$ as $\epsilon \to 0$.

The average acceptance probability of a proposed parameter sample $\boldsymbol{\theta}^*$ is $\mathcal{O}(\epsilon^d)$ (Fearnhead and Prangle, 2012), where d is the dimensionality of the data space, $\mathbb{D}$. This renders rejection sampling computationally expensive or even completely prohibitive, especially for high-dimensional parameter spaces (Marjoram et al., 2003; Sisson et al., 2007). Summary statistics can reduce the data dimensionality, however, they will often incur information loss (Barnes et

al., 2012; Blum et al., 2013; Fearnhead and Prangle, 2012). However, strategies including regression adjustment and marginal adjustment strategies can improve the accuracy of dimension reductions (Beaumont et al., 2002; Nott et al., 2014).

In the SMC-ABC method, importance resampling is applied to a sequence of R ABC posteriors with discrepancy thresholds $\epsilon_1 > \cdots > \epsilon_R$, with ϵ_R indicating the target ABC posterior. Given $\mathcal{M}$ weighted samples $\{(\boldsymbol{\theta}^i, w^i)\}_{i=1}^{\mathcal{M}}$, called particles, from the prior $p(\boldsymbol{\theta})$, particles are filtered through each ABC posterior using three main steps for each particle $\boldsymbol{\theta}^i$: (i) the particle is perturbed using a proposal kernel density $q_r(\boldsymbol{\theta} \mid \boldsymbol{\theta}^i)$; (ii) an accept/reject step is performed; and (iii) importance weights are updated. Once all particles have been updated and reweighted, resampling of particles is performed to avoid particle degeneracy. For reference, the SMC-ABC algorithm as initially developed by Sisson et al. (2007) and Toni et al. (2009) is given in Section 7.5.1.

Efficient use of SMC-ABC depends critically on the selection of appropriate proposal kernels and threshold sequences. Beaumont et al. (2009) and Filippi et al. (2013) (Section 7.5.1) present methods to determine proposal kernels such that the importance resampling step from ϵ_{r-1} to ϵ_r is efficient, that is, the average acceptance probability is high while still targeting the correct distribution. The efficiency can be evaluated though the Kullback-Leibler divergence (Kullback and Leibler, 1951) of the proposal distribution at ϵ_{r-1} relative to the target distribution at ϵ_r, given by,

$$D_{\mathrm{KL}}(\eta_{r-1}(\cdot); p(\cdot \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_r)) = \int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}_r \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_r) \ln \frac{p(\boldsymbol{\theta}_r \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_r)}{\eta_{r-1}(\boldsymbol{\theta}_r)} \, \mathrm{d}\boldsymbol{\theta}_r,$$

where $\eta_{r-1}(\boldsymbol{\theta}_r)$ is the proposal process

$$\eta_{r-1}(\boldsymbol{\theta}_r) = \sum_{j=1}^{\mathcal{M}} q_r(\boldsymbol{\theta}_r \mid \boldsymbol{\theta}_{r-1}^j) w_{r-1}^j. \tag{7.1}$$

In an adaptive SMC scheme, the proposal kernel, $q_r(\boldsymbol{\theta}^* \mid \boldsymbol{\theta})$, is chosen such that the efficiency is optimal under certain assumptions on the kernel and target distribution families (Section 7.5.1). In this work, we apply such an adaptive scheme (Beaumont et al., 2009; Drovandi and Pettitt, 2011; Filippi et al., 2013) that seeks to select an optimal proposal kernel, $q_r(\boldsymbol{\theta}_r \mid \boldsymbol{\theta}_{r-1})$ based on the weighted samples from the previous iteration $\{(\boldsymbol{\theta}_{r-1}^i, w_{r-1}^i)\}_{i=1}^{\mathcal{M}}$ (Section 7.5.1).

7.2.2 Preconditioning SMC-ABC

Consider a fixed sequence of ABC posteriors for the stochastic model inference problem, $\{p(\boldsymbol{\theta} \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_r)\}_{r=1}^R$. We want to apply SMC-ABC (Section 7.5.1) to efficiently sample from this sequence with adaptive proposal kernels, $\{q_r(\boldsymbol{\theta}^* \mid \boldsymbol{\theta})\}_{r=1}^R$ (Beaumont et al., 2009; Filippi et al., 2013). Our method exploits an approximate model to further improve the average acceptance probability.

7.2.2.1 Algorithm development

Say we have a set of weighted particles that represent the ABC posterior at threshold ϵ_{r-1} using the stochastic model, that is, $\{(\boldsymbol{\theta}_{r-1}^i, w_{r-1}^i)\}_{i=1}^{\mathcal{M}} \approx p(\boldsymbol{\theta}_{r-1} \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_{r-1})$. Now, consider applying the next importance resampling step using an approximate data generation step, $\tilde{\mathcal{D}}_s \sim \tilde{s}(\tilde{\mathcal{D}}_s \mid \boldsymbol{\theta})$, where $\tilde{s}(\tilde{\mathcal{D}}_s \mid \boldsymbol{\theta})$ is the simulation process of an approximate model[1]. Furthermore, assume the computational cost of simulating the approximate model, $\text{Cost}(\tilde{\mathcal{D}}_s)$, is significantly less than the computational cost of the exact model, $\text{Cost}(\mathcal{D}_s)$, that is, $\text{Cost}(\tilde{\mathcal{D}}_s)/\text{Cost}(\mathcal{D}_s) \ll 1$. The result will be a new set of weighted particles that represent the ABC posterior at threshold ϵ_r using this approximate model, denoted by $\{(\tilde{\boldsymbol{\theta}}_r^i, \tilde{w}_r^i)\}_{i=1}^{\mathcal{M}} \approx \tilde{p}(\tilde{\boldsymbol{\theta}}_r \mid \rho(\mathcal{D}, \tilde{\mathcal{D}}_s) \leq \epsilon_r)$. As noted in the examples in Section 7.1, approximate models are not always valid. This implies that $\tilde{p}(\tilde{\boldsymbol{\theta}}_r \mid \rho(\mathcal{D}, \tilde{\mathcal{D}}_s) \leq \epsilon_r)$ is always biased and will not in general converge to $p(\boldsymbol{\theta} \mid \mathcal{D})$ as $\epsilon_r \to 0$. However, since $\text{Cost}(\tilde{\mathcal{D}}_s)/\text{Cost}(\mathcal{D}_s) \ll 1$, it is computationally inexpensive to compute the distribution

$$\tilde{\eta}_r(\boldsymbol{\theta}_r) = \sum_{j=1}^{\mathcal{M}} \tilde{q}_r(\boldsymbol{\theta}_r \mid \tilde{\boldsymbol{\theta}}_r^j)\tilde{w}_r^j, \tag{7.2}$$

by comparison to computing $\eta_{r-1}(\boldsymbol{\theta}_r)$ (Equation (7.1)). In Equation (7.2), the proposal kernel $\tilde{q}_r(\boldsymbol{\theta}_r \mid \tilde{\boldsymbol{\theta}}_r^j)$ is possibly distinct from the $q_r(\boldsymbol{\theta}_r \mid \boldsymbol{\theta}_{r-1}^j)$ used in $\eta_{r-1}(\boldsymbol{\theta}_r)$ (Equation (7.1)). To improve the efficiency of the sampling process we simply require

$$D_{\text{KL}}(\eta_{r-1}(\cdot); p(\cdot \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_r)) > D_{\text{KL}}(\tilde{\eta}_r(\cdot); p(\cdot \mid \rho(\mathcal{D}, \tilde{\mathcal{D}}_s) \leq \epsilon_r)), \tag{7.3}$$

[1]Throughout, the overbar tilde notation, e.g. $\tilde{x}$, is used to refer to the ABC entities related to the approximate model, whereas quantities without the overbar tilde notation, e.g. x, are used to refer fo the ABC entities related to the exact model.

for $\tilde{\eta}_r(\boldsymbol{\theta}_r)$ (Equation (7.2)) to be more efficient as a proposal mechanism compared with $\eta_{r-1}(\boldsymbol{\theta}_r)$ (Equation (7.1)). Provided the condition $\text{Cost}(\tilde{\mathcal{D}}_s)/\text{Cost}(\mathcal{D}_s) \ll 1$ holds, any improvements in sampling efficiency will translate directly into computational performance improvements. That is, it does not matter that $\tilde{p}(\tilde{\boldsymbol{\theta}}_r \mid \rho(\mathcal{D}, \tilde{\mathcal{D}}_s) \leq \epsilon_r)$ is biased, it just needs to be less biased than $p(\boldsymbol{\theta}_{r-1} \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_{r-1})$ and computationally inexpensive.

This idea yields an intuitive new algorithm for SMC-ABC; proceed through the sequential sampling of $\{p(\boldsymbol{\theta} \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_r)\}_{r=1}^{R}$ by applying two resampling steps for each iteration. The first moves the particles from acceptance threshold ϵ_{r-1} to ϵ_r using the computationally inexpensive approximate model, and the second corrects for the bias between $\tilde{p}(\tilde{\boldsymbol{\theta}}_r \mid \rho(\mathcal{D}, \tilde{\mathcal{D}}_s) \leq \epsilon_r)$ and $p(\boldsymbol{\theta}_r \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_r)$ using the expensive stochastic model, but at an improved acceptance rate. Since the intermediate distribution acts on the proposal mechanism to accelerate the

Algorithm 7.1 Preconditioned SMC-ABC

Initialize $\boldsymbol{\theta}_0^i \sim p(\boldsymbol{\theta})$ and $w_0^i = 1/\mathcal{M}$, for $i = 1, \ldots, \mathcal{M}$;
for $r = 1, \ldots, R$ **do**
  for $i = 1, \ldots, \mathcal{M}$ **do**
    repeat
      Set $\boldsymbol{\theta}^* \leftarrow \boldsymbol{\theta}_{r-1}^j$ with probability $w_{r-1}^j / \left[\sum_{k=1}^{\mathcal{M}} w_{r-1}^k\right]$;
      Sample transition kernel, $\tilde{\boldsymbol{\theta}}^{**} \sim q_r(\tilde{\boldsymbol{\theta}} \mid \boldsymbol{\theta}^*)$;
      Generate data, $\tilde{\mathcal{D}}_s \sim \tilde{s}(\mathcal{D} \mid \tilde{\boldsymbol{\theta}}^{**})$;
    until $\rho(\mathcal{D}, \tilde{\mathcal{D}}_s) \leq \epsilon_r$
    Set $\tilde{\boldsymbol{\theta}}_r^i \leftarrow \tilde{\boldsymbol{\theta}}^{**}$;
    Set $\tilde{w}_r^i \leftarrow p(\tilde{\boldsymbol{\theta}}_r^i) / \left[\sum_{j=1}^{\mathcal{M}} w_{r-1}^j q_r(\tilde{\boldsymbol{\theta}}_r^i \mid \boldsymbol{\theta}_{r-1}^j)\right]$;
  end for
  for $i = 1, \ldots, \mathcal{M}$ **do**
    repeat
      Set $\tilde{\boldsymbol{\theta}}^* \leftarrow \tilde{\boldsymbol{\theta}}_r^j$ with probability $\tilde{w}_r^j / \left[\sum_{k=1}^{\mathcal{M}} \tilde{w}_r^k\right]$;
      Sample transition kernel, $\boldsymbol{\theta}^{**} \sim \tilde{q}_r(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}^*)$;
      Generate data, $\mathcal{D}_s \sim s(\mathcal{D} \mid \boldsymbol{\theta}^{**})$;
    until $\rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_r$
    Set $\boldsymbol{\theta}_r^i \leftarrow \boldsymbol{\theta}^{**}$;
    Set $w_r^i \leftarrow p(\boldsymbol{\theta}_r^i) / \left[\sum_{j=1}^{\mathcal{M}} \tilde{w}_r^j \tilde{q}_r(\boldsymbol{\theta}_r^i \mid \tilde{\boldsymbol{\theta}}_r^j)\right]$;
  end for
  Resample weighted particles, $\{(\boldsymbol{\theta}_r^i, w_r^i)\}_{i=1}^{\mathcal{M}}$, with replacement;
  Set $w_r^i \leftarrow 1/\mathcal{M}$ for all $i = 1, \ldots, \mathcal{M}$;
end for

convergence time of SMC-ABC, we denote the sequence $\{\tilde{p}(\tilde{\boldsymbol{\theta}}_r \mid \rho(\mathcal{D}, \tilde{\mathcal{D}}_s) \leq \epsilon_r)\}_{r=1}^{R}$ as the

preconditioner distribution sequence. The algorithm, called *preconditioned SMC-ABC* (PC-SMC-ABC), is given in Algorithm 7.1. We note that similar notions of preconditioning with approximation informed proposals have been applied in the context of Markov chain Monte Carlo samplers (Parno and Marzouk, 2018). However, to the best of our knowledge, our approach represents the first application of preconditioning ideas to SMC-ABC.

One particular advantage of the PC-SMC-ABC method, that demonstrate in the next section, is that it is unbiased. Effectively, one can consider PC-SMC-ABC as standard SMC-ABC method with a specialized proposal mechanism based on the preconditioner distribution. This means that PC-SMC-ABC is completely general, as discussed in Section 7.4, and is independent of the specific stochastic models that we consider here.

7.2.2.2 Analysis of PC-SMC-ABC

Here we demonstrate that the PC-SMC-ABC method is unbiased. Using similar arguments to Del Moral et al. (Del Moral et al., 2006) and Sisson et al. (Sisson et al., 2007), we show that the weighting update scheme can be interpreted as importance sampling on the joint space distribution of particle trajectories and that this joint target density admits the target posterior density as a marginal.

Given the target sequence, $\{p(\boldsymbol{\theta}_r \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_r)\}_{r=1}^R$, and approximate sequence, $\{\tilde{p}(\tilde{\boldsymbol{\theta}}_r \mid \rho(\mathcal{D}, \tilde{\mathcal{D}}_s) \leq \epsilon_r)\}_{r=1}^R$, with prior $p(\boldsymbol{\theta}_0)$, $\epsilon_{r-1} > \epsilon_r$ and proposal kernels $\tilde{q}_r(\boldsymbol{\theta}_r \mid \tilde{\boldsymbol{\theta}}_r)$ and $q_r(\tilde{\boldsymbol{\theta}}_r \mid \boldsymbol{\theta}_{r-1})$, we write the unnormalized weighting update scheme for PC-SMC-ABC. That is,

$$\tilde{w}_r^i = w_{r-1}^i \frac{p(\tilde{\boldsymbol{\theta}}_r^i \mid \rho(\mathcal{D}, \tilde{\mathcal{D}}_s) \leq \epsilon_r) Q_{r-1}(\boldsymbol{\theta}_{r-1}^i \mid \tilde{\boldsymbol{\theta}}_r^i)}{p(\boldsymbol{\theta}_{r-1}^i \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_{r-1}) q_r(\tilde{\boldsymbol{\theta}}_r^i \mid \boldsymbol{\theta}_{r-1}^i)}, \tag{7.4}$$

and

$$w_r^i = \tilde{w}_r^i \frac{p(\boldsymbol{\theta}_r^i \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_r) \tilde{Q}_r(\tilde{\boldsymbol{\theta}}_r^i \mid \boldsymbol{\theta}_r^i)}{p(\tilde{\boldsymbol{\theta}}_r^i \mid \rho(\mathcal{D}, \tilde{\mathcal{D}}_s) \leq \epsilon_r) \tilde{q}_r(\boldsymbol{\theta}_r^i \mid \tilde{\boldsymbol{\theta}}_r^i)}, \tag{7.5}$$

where $Q_{r-1}(\boldsymbol{\theta}_{r-1}^i \mid \tilde{\boldsymbol{\theta}}_r^i)$ and $\tilde{Q}_r(\tilde{\boldsymbol{\theta}}_r^i \mid \boldsymbol{\theta}_r^i)$ are arbitrary backwards kernels. Note that Algorithm 7.1 is not expressed in terms of these arbitrary kernels, but rather we utilize optimal backwards kernels. To proceed we substitute Equation (7.4) into Equation (7.5) and simplify as

follows,

$$
\begin{aligned}
w_r^i &= w_{r-1}^i \frac{p(\boldsymbol{\theta}_r^i \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_r)\tilde{Q}_r(\tilde{\boldsymbol{\theta}}_r^i \mid \boldsymbol{\theta}_r^i)}{p(\tilde{\boldsymbol{\theta}}_r^i \mid \rho(\mathcal{D}, \tilde{\mathcal{D}}_s) \leq \epsilon_r)\tilde{q}_r(\boldsymbol{\theta}_r^i \mid \tilde{\boldsymbol{\theta}}_r^i)} \times \frac{p(\tilde{\boldsymbol{\theta}}_r^i \mid \rho(\mathcal{D}, \tilde{\mathcal{D}}_s) \leq \epsilon_r)Q_{r-1}(\boldsymbol{\theta}_{r-1}^i \mid \tilde{\boldsymbol{\theta}}_r^i)}{p(\boldsymbol{\theta}_{r-1}^i \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_{r-1})q_r(\tilde{\boldsymbol{\theta}}_r^i \mid \boldsymbol{\theta}_{r-1}^i)} \\
&= w_{r-1}^i \frac{p(\boldsymbol{\theta}_r^i \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_r)\tilde{Q}_r(\tilde{\boldsymbol{\theta}}_r^i \mid \boldsymbol{\theta}_r^i)Q_{r-1}(\boldsymbol{\theta}_{r-1}^i \mid \tilde{\boldsymbol{\theta}}_r^i)}{p(\boldsymbol{\theta}_{r-1}^i \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_{r-1})\tilde{q}_r(\boldsymbol{\theta}_r^i \mid \tilde{\boldsymbol{\theta}}_r^i)q(\tilde{\boldsymbol{\theta}}_r^i \mid \boldsymbol{\theta}_{r-1}^i)r} \times \frac{p(\tilde{\boldsymbol{\theta}}_r^i \mid \rho(\mathcal{D}, \tilde{\mathcal{D}}_s) \leq \epsilon_r)}{p(\tilde{\boldsymbol{\theta}}_r^i \mid \rho(\mathcal{D}, \tilde{\mathcal{D}}_s) \leq \epsilon_r)} \\
&= w_{r-1}^i \frac{p(\boldsymbol{\theta}_r^i \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_r)\tilde{Q}_r(\tilde{\boldsymbol{\theta}}_r^i \mid \boldsymbol{\theta}_r^i)Q_{r-1}(\boldsymbol{\theta}_{r-1}^i \mid \tilde{\boldsymbol{\theta}}_r^i)}{p(\boldsymbol{\theta}_{r-1}^i \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_{r-1})\tilde{q}_r(\boldsymbol{\theta}_r^i \mid \tilde{\boldsymbol{\theta}}_r^i)q_r(\tilde{\boldsymbol{\theta}}_r^i \mid \boldsymbol{\theta}_{r-1}^i)}.
\end{aligned} \tag{7.6}
$$

Now, recursively expand the weight update sequence (Equation (7.6)) to obtain the final weight for the ith particle,

$$
\begin{aligned}
w_R^i &= \frac{p(\boldsymbol{\theta}_1^i \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_1)\tilde{Q}_1(\tilde{\boldsymbol{\theta}}_1^i \mid \boldsymbol{\theta}_1^i)Q_0(\boldsymbol{\theta}_0^i \mid \tilde{\boldsymbol{\theta}}_1^i)}{p(\boldsymbol{\theta}_0^i)\tilde{q}_1(\boldsymbol{\theta}_1^i \mid \tilde{\boldsymbol{\theta}}_1^i)q_1(\tilde{\boldsymbol{\theta}}_1^i \mid \boldsymbol{\theta}_0^i)} \\
&\quad \times \prod_{r=2}^{R} \frac{p(\boldsymbol{\theta}_r^i \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_r)\tilde{Q}_r(\tilde{\boldsymbol{\theta}}_r^i \mid \boldsymbol{\theta}_r^i)Q_{r-1}(\boldsymbol{\theta}_{r-1}^i \mid \tilde{\boldsymbol{\theta}}_r^i)}{p(\boldsymbol{\theta}_{r-1}^i \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_{r-1})\tilde{q}_r(\boldsymbol{\theta}_r^i \mid \tilde{\boldsymbol{\theta}}_r^i)q_r(\tilde{\boldsymbol{\theta}}_r^i \mid \boldsymbol{\theta}_{r-1}^i)} \\
&= \frac{p(\boldsymbol{\theta}_R^i \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_R)}{p(\boldsymbol{\theta}_0^i)} \prod_{r=1}^{R} \frac{\tilde{Q}_r(\tilde{\boldsymbol{\theta}}_r^i \mid \boldsymbol{\theta}_r^i)Q_{r-1}(\boldsymbol{\theta}_{r-1}^i \mid \tilde{\boldsymbol{\theta}}_r^i)}{\tilde{q}_r(\boldsymbol{\theta}_r^i \mid \tilde{\boldsymbol{\theta}}_r^i)q_r(\tilde{\boldsymbol{\theta}}_r^i \mid \boldsymbol{\theta}_{r-1}^i)} \\
&= \frac{p(\boldsymbol{\theta}_R^i \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_R)}{p(\boldsymbol{\theta}_0^i)} \frac{\prod_{r=1}^{R} B_{r-1}(\boldsymbol{\theta}_{r-1}^i \mid \boldsymbol{\theta}_r^i)}{\prod_{r=1}^{R} F_r(\boldsymbol{\theta}_r^i \mid \boldsymbol{\theta}_{r-1}^i)},
\end{aligned} \tag{7.7}
$$

where $F_r(\boldsymbol{\theta}_r^i \mid \boldsymbol{\theta}_{r-1}^i) = \tilde{q}_r(\boldsymbol{\theta}_r^i \mid \tilde{\boldsymbol{\theta}}_r^i)q_r(\tilde{\boldsymbol{\theta}}_r^i \mid \boldsymbol{\theta}_{r-1}^i)$ is the composite proposal kernel and $B_{r-1}(\boldsymbol{\theta}_{r-1}^i \mid \boldsymbol{\theta}_r^i) = Q_{r-1}(\boldsymbol{\theta}_{r-1}^i \mid \tilde{\boldsymbol{\theta}}_r^i)\tilde{Q}_r(\tilde{\boldsymbol{\theta}}_r^i \mid \boldsymbol{\theta}_r^i)$ is the composite backward kernel. We observe that Equation (7.7) is equivalent to the weight obtained from direct importance sampling on the joint space of the entire particle trajectory (Del Moral et al., 2006; Sisson et al., 2007), that is,

$$
w_R^i = \frac{\pi_R(\boldsymbol{\theta}_0^i, \boldsymbol{\theta}_1^i, \ldots, \boldsymbol{\theta}_R^i)}{\pi_0(\boldsymbol{\theta}_0^i, \boldsymbol{\theta}_1^i, \ldots, \boldsymbol{\theta}_R^i)}.
$$

Here, the importance distribution, given by

$$
\pi_0(\boldsymbol{\theta}_0^i, \boldsymbol{\theta}_1^i, \ldots, \boldsymbol{\theta}_R^i) = p(\boldsymbol{\theta}_0^i) \prod_{r=1}^{R} F_r(\boldsymbol{\theta}_r^i \mid \boldsymbol{\theta}_{r-1}^i),
$$

is the process of sampling from the prior and performing a sequence of kernel transitions. Finally, we note that the target distribution admits the target ABC posterior as a marginal

density, that is,

$$\int_{\mathbb{R}^R} \pi_R(\boldsymbol{\theta}_0^i, \boldsymbol{\theta}_1^i, \ldots, \boldsymbol{\theta}_R^i)\, \mathrm{d}\boldsymbol{\theta}_0^i \ldots \mathrm{d}\boldsymbol{\theta}_{R-1}^i = p(\boldsymbol{\theta}_R^i \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_R).$$

Therefore, for any function $f(\cdot)$ that satisfies certain regularity conditions,

$$\sum_{i=1}^{\mathcal{M}} f(\boldsymbol{\theta}_R^i) w_R^i \to \int_{\mathbb{R}} f(\boldsymbol{\theta}_R) p(\boldsymbol{\theta}_R \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_R)\, \mathrm{d}\boldsymbol{\theta}_R = \mathbb{E}\left[f(\boldsymbol{\theta}_R)\right],$$

as $\mathcal{M} \to \infty$, that is, the PC-SMC-ABC method is unbiased.

This property of unbiasedness holds even for cases where the approximate model is a poor approximation of the forward dynamics of the model. However, the closer that $\tilde{\eta}_r(\boldsymbol{\theta}_r)$ is to $p(\boldsymbol{\theta}_r \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_r)$ the better the performance improvement will be, as we demonstrate in Section 7.3.

7.2.3 Moment-matching SMC-ABC

The PC-SMC-ABC method is a promising modification to SMC-ABC that can accelerate inference for expensive stochastic models without introducing bias. However, other approaches can be used to obtain further computational improvements. Here, we consider an alternate approach to utilizing approximate models that aims to get the most out of a small sample of expensive stochastic simulations. Unlike PC-SMC-ABC, this method is generally biased, but it has the advantage of yielding a small and fixed computational budget. Specifically, we define a parameter $\alpha \in [0, 1]$, such that $1/\alpha$ is the target computational speedup, for example, $\alpha = 1/10$ should result in approximate 10 times speedup. We apply the SMC-ABC method using $\tilde{\mathcal{M}} = \lfloor (1-\alpha)\mathcal{M} \rfloor$ particles based on the approximate model, and then use $\hat{\mathcal{M}} = \lceil \alpha \mathcal{M} \rceil$ particles based on the stochastic model to construct a hybrid population of $\mathcal{M} = \hat{\mathcal{M}} + \tilde{\mathcal{M}}$ particles that will represent the final inference on the stochastic model. The key idea is that we use the $\hat{\mathcal{M}}$ particles of the expensive stochastic model to inform a transformation on the $\tilde{\mathcal{M}}$ particles of the approximation such that they the emulate particles of expensive stochastic model. Here, $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$ are, respectively, the floor and ceiling functions.

7.2.3.1 Algorithm development

Assume that we have applied SMC-ABC to sequentially sample $\tilde{\mathcal{M}}$ particles through the ABC posteriors from the approximate model, $\{\tilde{p}(\tilde{\boldsymbol{\theta}}_r \mid \rho(\mathcal{D}, \tilde{\mathcal{D}}_s) \leq \epsilon_r)\}_{r=1}^{R}$, with $\epsilon_R = \epsilon$. For the sake of the derivation, say that for all $r \in [1, R]$ we have available the mean vector, $\boldsymbol{\mu}_r$, and the covariance matrix, $\boldsymbol{\Sigma}_r$, of the ABC posterior $p(\boldsymbol{\theta}_r \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_r)$ under the stochastic model. In this case, we use particles $\tilde{\boldsymbol{\theta}}_r^1, \ldots, \tilde{\boldsymbol{\theta}}_r^{\tilde{\mathcal{M}}} \sim \tilde{p}(\tilde{\boldsymbol{\theta}}_r \mid \rho(\mathcal{D}, \tilde{\mathcal{D}}_s) \leq \epsilon_r)$ to emulate particles $\boldsymbol{\theta}_r^1, \ldots, \boldsymbol{\theta}_r^{\hat{\mathcal{M}}} \sim p(\boldsymbol{\theta}_r \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_r)$ by using the moment matching transform (Lei and Bickel, 2011; Sun et al., 2016)

$$\boldsymbol{\theta}_r^i = \mathbf{L}_r \left[\tilde{\mathbf{L}}_r^{-1} \left(\tilde{\boldsymbol{\theta}}_r^i - \tilde{\boldsymbol{\mu}}_r \right) \right] + \boldsymbol{\mu}_r, \quad i = 1, 2, \ldots, \tilde{\mathcal{M}}, \tag{7.8}$$

where $\tilde{\boldsymbol{\mu}}_r$ and $\tilde{\boldsymbol{\Sigma}}_r$ are the empirical mean vector and convariance matrix of particles $\tilde{\boldsymbol{\theta}}_r^1, \ldots, \tilde{\boldsymbol{\theta}}_r^{\tilde{\mathcal{M}}}$, $\mathbf{L}_r$ and $\tilde{\mathbf{L}}_r$ are lower triangular matrices obtained through the Cholesky factorization (Press et al., 1997) of $\boldsymbol{\Sigma}_r$ and $\tilde{\boldsymbol{\Sigma}}_r$, respectively, and $\tilde{\mathbf{L}}_r^{-1}$ is the matrix inverse of $\tilde{\mathbf{L}}_r$. This transform will produce a collection of particles that has a sample mean vector of $\boldsymbol{\mu}_r$ and covariance matrix $\boldsymbol{\Sigma}_r$. That is, the transformed sample matches the ABC posterior under the stochastic model up to the first two moments. In Section 7.3, we demonstrate that matching two moments is sufficient for the problems we investigate here, however, in principle we could extend this matching to higher order moments if required. For discussion on the advantages and disadvantages of matching higher moments, see Section 7.4.

In practice, it would be rare that $\boldsymbol{\mu}_r$ and $\boldsymbol{\Sigma}_r$ are known. If $p(\boldsymbol{\theta}_{r-1} \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_{r-1})$ is available, then we can use importance resampling to obtain $\hat{\mathcal{M}}$ particles, $\boldsymbol{\theta}_r^1, \ldots, \boldsymbol{\theta}_r^{\hat{\mathcal{M}}}$, from $p(\boldsymbol{\theta}_r \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_r)$, that is, we perform a step from SMC-ABC using the expensive stochastic model. We can then use the unbiased estimators

$$\hat{\boldsymbol{\mu}}_r = \frac{1}{\hat{\mathcal{M}}} \sum_{i=1}^{\hat{\mathcal{M}}} \boldsymbol{\theta}_r^i \quad \text{and} \quad \hat{\boldsymbol{\Sigma}}_r = \frac{1}{\hat{\mathcal{M}} - 1} \sum_{i=1}^{\hat{\mathcal{M}}} (\boldsymbol{\theta}_r^i - \hat{\boldsymbol{\mu}}_r)(\boldsymbol{\theta}_r^i - \hat{\boldsymbol{\mu}}_r)^{\mathrm{T}}, \tag{7.9}$$

to obtain estimates of $\boldsymbol{\mu}_r$ and $\boldsymbol{\Sigma}_r$. Substituting Equation (7.9) into Equation (7.8) gives an approximate transform

$$\hat{\boldsymbol{\theta}}_r^i = \hat{\mathbf{L}}_r \left[\tilde{\mathbf{L}}_r^{-1} \left(\tilde{\boldsymbol{\theta}}_r^i - \tilde{\boldsymbol{\mu}}_r \right) \right] + \hat{\boldsymbol{\mu}}_r, \quad i = 1, 2, \ldots, \tilde{\mathcal{M}}, \tag{7.10}$$

where $\hat{\boldsymbol{\Sigma}}_r = \hat{\mathbf{L}}_r \hat{\mathbf{L}}_r^{\mathrm{T}}$. This enables us to construct an estimate of $p(\boldsymbol{\theta}_r \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_r)$ by applying the moment-matching transform (Equation (7.10)) to the particles $\tilde{\boldsymbol{\theta}}_r^1, \ldots, \tilde{\boldsymbol{\theta}}_r^{\tilde{\mathcal{M}}}$ then pooling the transformed particles $\hat{\boldsymbol{\theta}}_r^1, \ldots, \hat{\boldsymbol{\theta}}_r^{\tilde{\mathcal{M}}}$ with the particles $\boldsymbol{\theta}_r^1, \ldots, \boldsymbol{\theta}_r^{\hat{\mathcal{M}}}$ that were used in the estimates $\hat{\boldsymbol{\mu}}_r$ and $\hat{\boldsymbol{\Sigma}}_r$. The goal of the approximate transform application is for the transforms particles $\hat{\boldsymbol{\theta}}_r^1, \ldots, \hat{\boldsymbol{\theta}}_r^{\tilde{\mathcal{M}}}$ to be more accurate in higher moments due, despite only matching the first two moments (see Section 7.3.5 for numerical justification of this property for some specific examples). This results in an approximation of $p(\boldsymbol{\theta}_r \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_r)$ using a set of $\mathcal{M}$ particles $\boldsymbol{\theta}_r^1, \ldots, \boldsymbol{\theta}_r^{\mathcal{M}}$ with $\boldsymbol{\theta}_r^{i+\hat{\mathcal{M}}} = \hat{\boldsymbol{\theta}}_r^i$ where $1 \leq i \leq \tilde{\mathcal{M}}$.

This leads to our *moment-matching* SMC-ABC (MM-SMC-ABC) method. First, SMC-ABC inference is applied using the approximate model with $\tilde{\mathcal{M}}$ particles. Then, given $\mathcal{M}$ samples from the prior, $p(\boldsymbol{\theta})$, we can sequentially approximate
$\{p(\boldsymbol{\theta}_r \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_r)\}_{r=1}^R$. At each iteration the following steps are performed: (i) generate a small number of particles from $p(\boldsymbol{\theta}_r \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_r)$ using importance resampling and stochastic model simulations; (ii) compute $\hat{\boldsymbol{\mu}}_r$ and $\hat{\boldsymbol{\Sigma}}_r = \hat{\mathbf{L}}_r \hat{\mathbf{L}}_r^{\mathrm{T}}$; (iii) apply the transform from Equation (7.10) to the particles at ϵ_r from the approximate model; (iv) pool the resulting particles with the stochastic model samples; and (v) reweight particles and resample. The final MM-SMC-ABC algorithm is provided in Algorithm 7.2.

The performance of this method depends on the choice of α. Note that in Algorithm 7.2, standard SMC-ABC for the expensive stochastic model is recovered as $\alpha \to 1$ (no speedup, inference unbiased), and standard SMC-ABC using the approximate model is recovered as $\alpha \to 0$ (maximum speedup, but inference biased). Therefore we expect there is a choice of $\alpha \in (0, 1)$ that provides an optimal trade-off between computational improvement and accuracy. Clearly, the expected speed improvement is proportional to $1/\alpha$, however, if α is chosen to be too small, then the statistical error in the estimates in Equation (7.9) will be too high. We explore this trade-off in detail in Section 7.3.5 and find that $0.05 \leq \alpha \leq 0.2$ seems to give a reasonable result.

Algorithm 7.2 Moment-matching SMC-ABC

Given $\alpha \in [0,1]$, initialize $\hat{\mathcal{M}} = \lceil \alpha \mathcal{M} \rceil$ and $\tilde{\mathcal{M}} = \lfloor (1-\alpha)\mathcal{M} \rfloor$;
Initialize $\tilde{\boldsymbol{\theta}}_0^i \sim p(\boldsymbol{\theta})$ and $\tilde{w}_0^i = 1/\tilde{\mathcal{M}}$, for $i = 1, \ldots, \tilde{\mathcal{M}}$;
Initialize $\boldsymbol{\theta}_0^i \sim p(\boldsymbol{\theta})$ and $w_0^i = 1/\mathcal{M}$, for $i = 1, \ldots, \mathcal{M}$;
Apply SMC-ABC to generate the sequence of approximate particles $\{(\tilde{\boldsymbol{\theta}}_1^i, \tilde{w}_1^i)\}_{i=1}^{\tilde{\mathcal{M}}}$, $\{(\tilde{\boldsymbol{\theta}}_2^i, \tilde{w}_2^i)\}_{i=1}^{\tilde{\mathcal{M}}}, \ldots, \{(\tilde{\boldsymbol{\theta}}_R^i, \tilde{w}_R^i)\}_{i=1}^{\tilde{\mathcal{M}}}$
for $r = 1, \ldots, R$ **do**
    for $i = 1, \ldots, \hat{\mathcal{M}}$ **do**
        repeat
            Set $\boldsymbol{\theta}^* \leftarrow \boldsymbol{\theta}_{r-1}^j$ with probability $w_r^j / \left[\sum_{k=1}^{\mathcal{M}} w_{r-1}^k\right]$;
            Sample transition kernel, $\boldsymbol{\theta}^{**} \sim q_r(\boldsymbol{\theta} \mid \boldsymbol{\theta}^*)$;
            Generate data, $\mathcal{D}_s \sim s(\mathcal{D} \mid \boldsymbol{\theta}^{**})$;
        until $\rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_r$
        Set $\boldsymbol{\theta}_r^i \leftarrow \boldsymbol{\theta}^{**}$;
        Set $w_r^i \leftarrow p(\boldsymbol{\theta}_r^i) / \left[\sum_{j=1}^{\mathcal{M}} w_{r-1}^j q_r(\boldsymbol{\theta}_r^i \mid \boldsymbol{\theta}_{r-1}^j)\right]$;
    end for
    Estimate means and covariances $\tilde{\boldsymbol{\mu}}_r$, $\tilde{\boldsymbol{\Sigma}}_r$, $\hat{\boldsymbol{\mu}}_r$, and $\hat{\boldsymbol{\Sigma}}_r$;
    Compute Cholesky decompositions $\tilde{\boldsymbol{\Sigma}}_r = \tilde{\mathbf{L}}_r \tilde{\mathbf{L}}_r^{\mathrm{T}}$ and $\hat{\boldsymbol{\Sigma}}_r = \hat{\mathbf{L}}_r \hat{\mathbf{L}}_r^{\mathrm{T}}$
    for $i = 1, \ldots, \tilde{\mathcal{M}}$ **do**
        Set $\boldsymbol{\theta}_r^{i+\hat{\mathcal{M}}} \leftarrow \hat{\mathbf{L}}_r[\tilde{\mathbf{L}}_r^{-1}(\tilde{\boldsymbol{\theta}}_r^i - \tilde{\boldsymbol{\mu}}_r)] + \hat{\boldsymbol{\mu}}_r$ and $w_r^{i+\hat{\mathcal{M}}} \leftarrow \tilde{w}_r^i$;
    end for
    Resample weighted particles $\{(\boldsymbol{\theta}_r^i, w_r^i)\}_{r=1}^{\mathcal{M}}$ with replacement;
    Set $w_r^i \leftarrow 1/\mathcal{M}$ for all $i = 1, \ldots, \mathcal{M}$;
end for

7.3 Results

In this section, we provide numerical examples to demonstrate the accuracy and performance of the PC-SMC-ABC and MM-SMC-ABC methods. For our first example, we consider the analysis of spatially averaged population growth data. The discrete model used in this instance is relevant in the ecological sciences as it describes population growth subject to a weak Allee effect (Taylor and Hastings, 2005). We then analyze data that is typical of *in vitro* cell culture scratch assays in experimental cell biology using a discrete model that leads to the well-studied Fisher-KPP model (Edelstein-Keshet, 2005; Murray, 2002). In both examples, we present the discrete model and its continuum limit, then compute the full Bayesian posterior for the model parameters using the PC-SMC-ABC (Algorithm 7.1) and MM-SMC-ABC (Algorithm 7.2) methods, and compare the results with the SMC-ABC (Section 7.5.1) using either the discrete model or continuum limit alone. We also provide numerical experiments to evaluate the effect of the tuning parameter α on the accuracy and performance of the MM-SMC-ABC method.

It is important to clarify that when we refer to the accuracy of our methods, we refer to their ability to sample from the target ABC posterior under the expensive stochastic model. The evaluation of this accuracy requires sampling from the target ABC posterior under the expensive stochastic model using SMC-ABC. As a result, the target acceptance thresholds are chosen to ensure this is computationally feasible.

7.3.1 Lattice-based stochastic discrete random walk model

The stochastic discrete model we consider is a lattice-based random walk model that is often used to describe populations of motile cells (Jin et al., 2016a). The model involves initially placing a population of N agents of size δ on a lattice, L (Callaghan et al., 2006; Simpson et al., 2010), for example an $I \times J$ hexagonal lattice (Jin et al., 2016a). This hexagonal lattice is defined by a set of indices
$L = \{(i, j) : i \in [0, 1, \ldots, I-1], j \in [0, 1, \ldots, J-1]\}$, and a neighborhood function,

$$\mathcal{N}(i,j) = \begin{cases} \{(i-1,j-1),(i,j-1),(i+1,j-1),(i+1,j),(i,j+1),(i-1,j)\} & \text{if } i \text{ is even,} \\ \{(i-1,j),(i,j-1),(i+1,j),(i+1,j+1),(i,j+1),(i-1,j+1)\} & \text{if } i \text{ is odd.} \end{cases}$$

Lattice indices are mapped to Cartesian coordinates using

$$(x_i, y_j) = \begin{cases} \left(i\dfrac{\sqrt{3}}{2}\delta, j\delta \right) & \text{if } i \text{ is even,} \\ \left(i\dfrac{\sqrt{3}}{2}\delta, \left(j + \dfrac{1}{2} \right)\delta \right) & \text{if } i \text{ is odd.} \end{cases} \tag{7.11}$$

We define an occupancy function such that $C(\ell, t) = 1$ if site ℓ is occupied by an agent at time $t \geq 0$, otherwise $C(\ell, t) = 0$. This means that in our discrete model each lattice site can be occupied by, at most, one agent.

During each discrete time step of duration τ, agents attempt to move with probability $P_m \in [0, 1]$ and attempt to proliferate with probability $P_p \in [0, 1]$. If an agent at site ℓ attempts a motility event, then a neighboring site will be selected randomly with uniform probability. The motility event is aborted if the selected site is occupied, otherwise the agent will move to the selected site (Figure 7.2(A)–(C)). For proliferation events, the local neighborhood average occupancy,

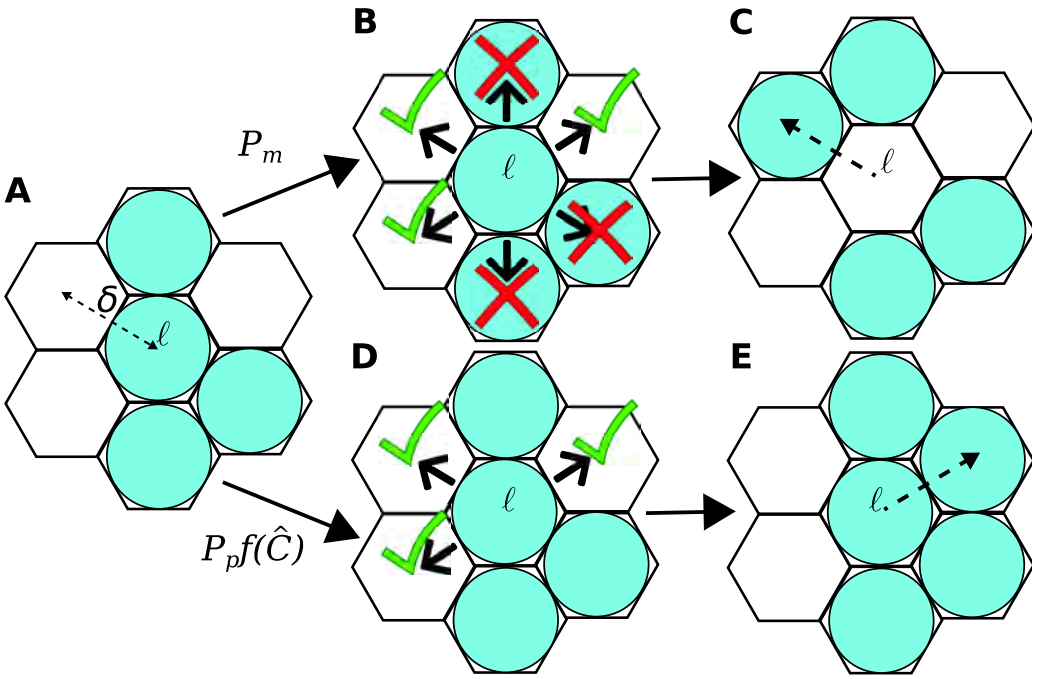

Figure 7.2: Example of movement and proliferation events in a lattice-based random walk model, using a hexagonal lattice with lattice spacing, δ. (A) An example hexagonal lattice neighborhood $\mathcal{N}(\ell)$. An agent at site ℓ attempts a motility event (A)–(C) with probability P_m. (B) Motility events are aborted when the randomly selected neighbor site is occupied. (C) The agent moves to the selected site, if unoccupied. An agent at site ℓ attempts a proliferation event (A),(D)–(E) with probability P_p. (D) Proliferation events are successful with probability $f(\hat{C}(\ell, t))$, resulting in an unoccupied site being selected. (E) The daughter agent is placed at the selected site and the number of agents in the populations is increased by one.

$$\hat{C}(\ell, t) = \frac{1}{6} \sum_{\ell' \in N(\ell)} C(\ell', t),$$

is calculated and a uniform random number $u \sim \mathcal{U}(0, 1)$ is drawn. If $u > f(\hat{C}(\ell, t))$, where $f(\hat{C}(\ell, t)) \in [0, 1]$ is called the *crowding function* (Browning et al., 2017; Jin et al., 2016a), then the proliferation event is aborted due to local crowding effects and contact inhibition. If $u \leq f(\hat{C}(\ell, t))$, then proliferation is successful and a daughter agent is placed at a randomly chosen unoccupied lattice site in $\mathcal{N}(\ell)$ (Figure 7.2(A),(D)–(E)). The evolution of the model is generated through repeating this process though M time steps, $t_1 = \tau$, $t_2 = 2\tau, \ldots,$ $t_M = M\tau$. This approach, based on the work by Jin et al. (2016a), supports a generic proliferation mechanism since $f(\hat{C}(\ell, t))$ is an arbitrary smooth function satisfying $f(0) = 1$ and $f(K) = 0$, where $K > 0$ is the carrying capacity density. However, in the literature there are also examples that include other mechanisms such as cell-cell adhesion (Johnston et al., 2013), directed motility (Binny et al., 2016), and Allee effects (Böttger et al., 2015).

7.3.2 Approximate continuum-limit descriptions

Discrete models do not generally lend themselves to analytical methods, consequently, their application is intrinsically tied to computationally intensive stochastic simulations and Monte Carlo methods (Jin et al., 2016a). As a result, it is common practice to approximate mean behavior using differential equations by invoking mean-field assumptions, that is, to treat the occupancy status of lattice sites as independent (Callaghan et al., 2006; Simpson et al., 2010). The resulting approximate continuum-limit descriptions (Section 7.5.2) are partial differential equations (PDEs) of the form

$$\frac{\partial \mathcal{C}(x,y,t)}{\partial t} = D\nabla^2 \mathcal{C}(x,y,t) + \lambda \mathcal{C}(x,y,t) f(\mathcal{C}(x,y,t)), \tag{7.12}$$

where $\mathcal{C}(x,y,t) = \mathbb{E}\left[C(\ell,t)\right]$, $D = \lim_{\delta\to 0,\tau\to 0} P_m\delta^2/(4\tau)$ is the diffusivity, $\lambda = \lim_{\tau\to 0} P_p/\tau$ is the proliferation rate with $P_p = \mathcal{O}(\tau)$, and $f(\cdot)$ is the crowding function that is related to the proliferation mechanism implemented in the discrete model (Browning et al., 2017; Jin et al., 2016a). For spatially uniform populations there will be no macroscopic spatial gradients on average, that is $\nabla\mathcal{C}(x,y,t) = \mathbf{0}$. Thus, $\mathcal{C}(x,y,t)$ is just a function of time, $\mathcal{C}(t)$, and the continuum limit reduces to an ordinary differential equation (ODE) describing the net population growth,

$$\frac{\mathrm{d}\mathcal{C}(t)}{\mathrm{d}t} = \lambda \mathcal{C}(t) f(\mathcal{C}(t)). \tag{7.13}$$

For many standard discrete models, the crowding function is implicitly $f(\mathcal{C}) = 1-\mathcal{C}$ (Callaghan et al., 2006). That is, the continuum limits in Equation (7.12) and Equation (7.13) yield the Fisher-KPP model (Edelstein-Keshet, 2005; Murray, 2002) and the logistic growth model (Tsoularis and Wallace, 2002) (Chapter 2), respectively. However, non-logistic growth, for example, $f(\mathcal{C}) = (1-\mathcal{C})^n$ for $n > 1$, have also been considered (Jin et al., 2016a; Simpson et al., 2010; Tsoularis and Wallace, 2002).

7.3.3 Temporal example: a weak Allee model

The Allee effect refers to the reduction in growth rate of a population at low densities. This is particularly well studied in ecology where there are many mechanisms that give rise to this phenomenon (Taylor and Hastings, 2005; Johnston et al., 2017). We introduce an Allee effect

into our discrete model by choosing a crowding function of the form

$$f(\hat{C}(\ell,t)) = \left(1 - \frac{\hat{C}(\ell,t)}{K}\right)\left(\frac{A + \hat{C}(\ell,t)}{K}\right),$$

where $\hat{C}(\ell,t) \in [0,1]$ is the local density at the lattice site $\ell \in L$, at time t, $K > 0$ is the carrying capacity density, and A is the Allee parameter which yields a weak Allee effect for $A \geq 0$ (Wang et al., 2019). Note that smaller values of A entail a more pronounced Allee effect with $A < 0$ leading to a strong Allee effect that can lead to species extinction (Wang et al., 2019). For simplicity, we only consider the weak Allee effect here, but our methods are general enough to consider any sufficiently smooth $f(\cdot)$.

Studies in ecology often involve population counts of a particular species over time (Taylor and Hastings, 2005). In the discrete model, the initial occupancy of each lattice site is independent, leading to no macroscopic spatial gradients on average. It is reasonable to summarize simulations of the discrete model at time t by the average occupancy over the entire lattice, $\bar{C}(t) = (1/IJ)\sum_{\ell \in L} C(\ell,t)$. Therefore, the continuum limit for this case is given by (Wang et al., 2019)

$$\frac{\mathrm{d}\mathcal{C}(t)}{\mathrm{d}t} = \lambda \mathcal{C}(t)\left(1 - \frac{\mathcal{C}(t)}{K}\right)\left(\frac{A + \mathcal{C}(t)}{K}\right), \tag{7.14}$$

with $\mathcal{C}(t) = \mathbb{E}\left[\bar{C}(t)\right]$, $\lambda = \lim_{\tau \to 0} P_p/\tau$, and $\mathcal{C}(0) = \mathbb{E}\left[\bar{C}(0)\right]$.

We generate synthetic time-series ecological data using the discrete model, with observations made at times $t_1 = \tau \times 10^3, t_2 = 2\tau \times 10^3, \ldots, t_{10} = \tau \times 10^4$, resulting in data $\mathcal{D} = [C_{\text{obs}}(t_1), C_{\text{obs}}(t_2), \ldots, C_{\text{obs}}(t_{10})]$ with $C_{\text{obs}}(t) = \bar{C}(t)$ where $\bar{C}(t)$ is the average occupancy at time t for a single realization of the discrete model (Section 7.5.3). For this example, we consider an $I \times J$ hexagonal lattice with $I = 80$, $J = 68$, and parameters $P_p = 1/1000$, $P_m = 0$, $\delta = \tau = 1$, $K = 5/6$, and $A = 1/10$. Reflecting boundary conditions are applied at all boundaries and a uniform initial condition is applied, specifically, each site is set to occupied with probability $\mathbb{P}(C(\ell,0) = 1) = 1/4$ for all $\ell \in L$, giving $\mathcal{C}(0) = 1/4$. This combination of parameters is selected since it is known that the continuum limit (Equation (7.14)) will not accurately predict the population growth dynamics of the discrete model in this regime since $P_p/P_m \gg 1$ (Section 7.5.6).

For the inference problem we assume P_m is known, and we seek to compute $p(\boldsymbol{\theta} \mid \mathcal{D})$ under

the discrete model with $\boldsymbol{\theta} = [\lambda, A, K]$ and $\lambda = P_p/\tau$. We utilize uninformative priors, $P_p \sim \mathcal{U}(0, 0.005)$, $K \sim \mathcal{U}(0, 1)$ and $A \sim \mathcal{U}(0, 1)$ with the additional constraint that $A \leq K$, that is A and K are not independent in the prior. The discrepancy metric used is the Euclidean distance. For the discrete model, this is

$$\rho(\mathcal{D}, \mathcal{D}_s) = \left[\sum_{k=1}^{10} \left(C_{\text{obs}}(t_k) - \bar{C}(t_k)\right)^2\right]^{1/2},$$

where $\bar{C}(t_k)$ is the average occupancy at time t_k of a realization of the discrete model given $\boldsymbol{\theta}$. Similarly, for the continuum limit we have

$$\rho(\mathcal{D}, \tilde{\mathcal{D}}_s) = \left[\sum_{j=1}^{10} \left(C_{\text{obs}}(t_k) - \mathcal{C}(t_k)\right)^2\right]^{1/2},$$

where $\mathcal{C}(t_k)$ is the solution to the continuum limit (Equation (7.14)), computed numerically (Fehlberg, 1969; Iserles, 2008) (Section 7.5.4). We compute the posterior using our PC-SMC-ABC and MM-SMC-ABC methods to compare with SMC-ABC under the continuum limit and SMC-ABC under the discrete model. In each instance, $\mathcal{M} = 1000$ particles are used to approach the target threshold $\epsilon = 0.125$ using the sequence $\epsilon_1, \epsilon_2, \ldots, \epsilon_5$ with $\epsilon_r = \epsilon_{r-1}/2$. In the case of MM-SMC-ABC the tuning parameter is $\alpha = 0.1$. The Gaussian proposal kernels, $q_r(\boldsymbol{\theta}_r \mid \boldsymbol{\theta}_{r-1})$ and $\tilde{q}_r(\boldsymbol{\theta}_r \mid \tilde{\boldsymbol{\theta}}_r)$, are selected adaptively (Section 7.5.1).

Figure 7.3 and Table 7.1 present the results. SMC-ABC using the continuum-limit model is a poor approximation for SMC-ABC using the discrete model, especially for the proliferation rate parameter, λ (Figure 7.3(a)), which is expected because $P_m = 0$. However, the posteriors estimated using PC-SMC-ABC are an excellent match to the target posteriors estimated using SMC-ABC with the expensive discrete model, yet the PC-SMC-ABC method requires only half the number of stochastic simulations (Table 7.1). The MM-SMC-ABC method is not quite as accurate as the PC-SMC-ABC method, however, the number of expensive stochastic simulations is reduced by more than a factor of eight (Table 7.1) leading to considerable increase in computational efficiency.

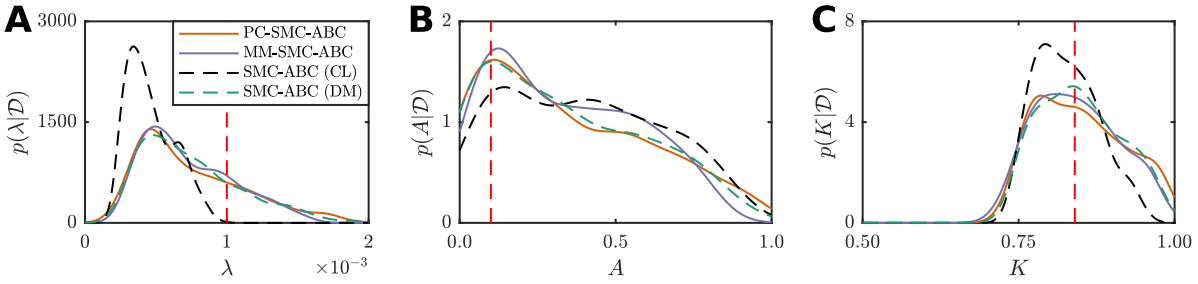

Figure 7.3: Comparison of estimated posterior marginal densities for the weak Allee model. There is a distinct bias in the SMC-ABC density estimate using the continuum limit (CL) (black dashed) compared with SMC-ABC with the discrete model (DM) (green dashed). However, the density estimates computed using the PC-SMC-ABC (orange solid) and MM-SMC-ABC (purple solid) methods match well with a significantly reduced computational overhead.

Table 7.1: Computational performance comparison of the SMC-ABC, PC-SMC-ABC, and MM-SMC-ABC methods for the weak Allee model inference problem. Computations are performed using an Intel® Xeon™ E5-2680v3 CPU (2.5 GHz).

Method	Stochastic samples	Continuum samples	Run time (hours)	Speedup
SMC-ABC	$28,588$	0	47.1	$1\times$
PC-SMC-ABC	$13,799$	$58,752$	21.1	$2\times$
MM-SMC-ABC	$3,342$	$36,908$	5.6	$8\times$

7.3.4 Spatiotemporal example: a scratch assay

We now look to a discrete model commonly used in studies of cell motility and proliferation, and use spatially extended data that is typical of *in vitro* cell culture experiments, specifically scratch assays (Liang et al., 2007).

In this case we use a crowding function of the form $f(\hat{C}(\ell,t)) = 1 - \hat{C}(\ell,t)/K$, where $K > 0$ is the carrying capacity density, since it will lead to a logistic growth source term in Equation (7.12) which characterizes the growth dynamics of many cell types (Simpson et al., 2010) (Chapter 2). The discrete model is initialized such that initial density is independent of y. Therefore, we summarize the discrete simulation by computing the average occupancy for each x coordinate, that is, we average over the y-axis in the hexagonal lattice (Jin et al., 2016a), that is, $\bar{C}(x,t) = (1/J)\sum_{(x,y)\in L} C((x,y),t)$. Thus, one arrives at the Fisher-KPP model (Edelstein-Keshet, 2005; Murray, 2002) for the continuum limit,

$$\frac{\partial \mathcal{C}(x,t)}{\partial t} = D\frac{\partial^2 \mathcal{C}(x,t)}{\partial x^2} + \lambda \mathcal{C}(x,t)\left(1 - \frac{\mathcal{C}(x,t)}{K}\right), \tag{7.15}$$

where $\mathcal{C}(x,t) = \mathbb{E}\left[\bar{C}(x,t)\right]$, $D = \lim_{\delta\to 0,\tau\to 0} P_m\delta^2/(4\tau)$, and $\lambda = \lim_{\tau\to 0} P_p/\tau$.

Just as with the weak Allee model, here we generate synthetic spatiotemporal cell culture data using the discrete model. Observations are made at times $t_1 = 3\tau \times 10^2$, $t_2 = 6\tau \times 10^2, \ldots,$ $t_{10} = 3\tau \times 10^3$, resulting in data

$$\mathcal{D} = \begin{bmatrix} C_{\text{obs}}(x_1,t_1) & C_{\text{obs}}(x_1,t_2) & \ldots & C_{\text{obs}}(x_1,t_{10}) \\ C_{\text{obs}}(x_2,t_1) & C_{\text{obs}}(x_2,t_2) & \ldots & C_{\text{obs}}(x_2,t_{10}) \\ \vdots & \vdots & \ddots & \vdots \\ C_{\text{obs}}(x_I,t_1) & C_{\text{obs}}(x_I,t_2) & \ldots & C_{\text{obs}}(x_I,t_{10}) \end{bmatrix},$$

with $C_{\text{obs}}(x,t) = \bar{C}(x,t)$ where $\bar{C}(x,t)$ is the average occupancy over sites $(x,y_1),(x,y_2),$ $\ldots,(x,y_J)$ at time t for a single realizations of the discrete model. As with the weak Allee model, we consider an $I\times J$ hexagonal lattice with $I = 80$, $J = 68$, and parameters $P_p = 1/1000$, $P_m = 1$, $\delta = \tau = 1$, and $K = 5/6$. We simulate a scratch assay by specifying the center 20 cell columns ($31 \leq i \leq 50$) to be initially unoccupied, and apply a uniform initial condition outside the scratch area such that $\mathbb{E}\left[C(\ell,0)\right] = 1/4$ overall. Reflecting boundary conditions are applied at all boundaries. Note, we have selected a parameter regime with $P_p/P_m \ll 1$ for which the continuum limit is an accurate representation of the discrete model average behavior (Section 7.5.6).

Since we have spatial information for this problem, we assume P_m is also an unknown parameter and perform inference on the discrete model to compute $p(\boldsymbol{\theta} \mid \mathcal{D})$ with $\boldsymbol{\theta} = [\lambda, D, K]$, $\lambda = P_p/\tau$, and $D = P_m\delta^2/4\tau$. We utilize uninformative priors, $P_p \sim \mathcal{U}(0, 0.008)$, $P_m \sim \mathcal{U}(0,1)$, and $K \sim \mathcal{U}(0,1)$. For the discrepancy metric we use the Frobenius norm; for the discrete model, this is

$$\rho(\mathcal{D},\mathcal{D}_s) = \left[\sum_{k=1}^{10}\sum_{i=1}^{I}\left(C_{\text{obs}}(x_i,t_k) - \bar{C}(x_i,t_k)\right)^2\right]^{1/2},$$

where $\bar{C}(x_i,t_k)$ is the average occupancy at site x_i at time t_k of a realization of the discrete

model given parameters $\boldsymbol{\theta}$. Similarly, for the continuum limit we have

$$\rho(\mathcal{D}, \tilde{\mathcal{D}}_s) = \left[\sum_{k=1}^{10}\sum_{i=1}^{I}\left(C_{\text{obs}}(x_i, t_k) - \mathcal{C}(x_i, t_k)\right)^2\right]^{1/2},$$

where $\mathcal{C}(x_i, t_k)$ is the solution to the continuum-limit PDE (Equation (7.15)), computed using a backward-time, centered-space finite difference scheme with fixed-point iteration and adaptive time steps (Simpson et al., 2007b; Sloan and Abbo, 1999) (Section 7.5.4). We estimate the posterior using our PC-SMC-ABC and MM-SMC-ABC methods to compare with SMC-ABC using the continuum limit and SMC-ABC using the discrete model. In each case, $\mathcal{M} = 1000$ particles are used to approach the target threshold $\epsilon = 2$ using the sequence $\epsilon_1, \epsilon_2, \ldots, \epsilon_5$ with $\epsilon_r = \epsilon_{r-1}/2$. In the case of MM-SMC-ABC the tuning parameter is $\alpha = 0.1$. Again, Gaussian proposal kernels, $q_r(\boldsymbol{\theta}_r \mid \boldsymbol{\theta}_{r-1})$ and $\tilde{q}_r(\boldsymbol{\theta}_r \mid \tilde{\boldsymbol{\theta}}_r)$, are selected adaptively (Section 7.5.1).

Results are shown in Figure 7.4 and Table 7.2. Despite the continuum limit being a good approximation of the discrete model average behavior, using solely this continuum limit in the inference problem still leads to bias. Just as with the weak Allee model, both PC-SMC-ABC and MM-SMC-ABC methods produce a more accurate estimate of the SMC-ABC posterior density with the discrete model. Overall, PC-SMC-ABC is unbiased, however, MM-SMC-ABC is still very accurate. The main point for our work is that the PC-SMC-ABC and MM-SMC-ABC methods both produce posteriors that are accurate compared with the expensive stochastic inference problem, whereas the approximate model alone does not. From Table 7.2, both PC-SMC-ABC and MM-SMC-ABC require a reduced number of stochastic simulations of the discrete model compared with direct SMC-ABC. For PC-SMC-ABC, the reduction is almost a factor of four, and for MM-SMC-ABC, the reduction is almost a factor of eleven.

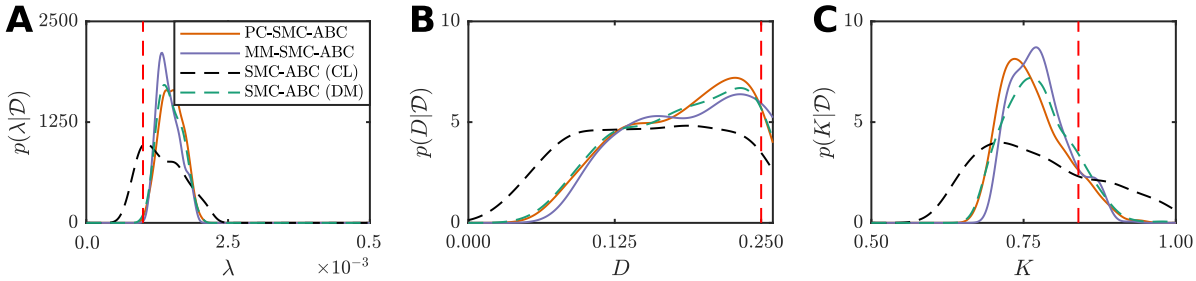

Figure 7.4: Comparison of estimated posterior marginal densities for the scratch assay model. There is a distinct bias in the SMC-ABC density estimate using the continuum limit (CL) (black dashed) compared with SMC-ABC with the discrete model (DM) (green dashed). However, the density estimates computed using the PC-SMC-ABC (orange solid) and MM-SMC-ABC (purple solid) methods match well with a reduced computational overhead.

Table 7.2: Computational performance comparison of the SMC-ABC, PC-SMC-ABC, and MM-SMC-ABC methods, using the scratch assay model inference problem. Computations are performed using an Intel® Xeon™ E5-2680v3 CPU (2.5 GHz).

Method	Stochastic samples	Continuum samples	Run time (hours)	Speedup
SMC-ABC	$46,435$	0	20.6	$1\times$
PC-SMC-ABC	$13,949$	$13,179$	5.6	$4\times$
MM-SMC-ABC	$4,457$	$10,594$	1.9	$11\times$

7.3.5 A guide to selection of α for MM-SMC-ABC

The performance of MM-SMC-ABC is dependent on the tuning parameter $\alpha \in [0, 1]$. Since MM-SMC-ABC will only propagate $\lceil \alpha \mathcal{M} \rceil$ particles based on the expensive stochastic model, α can be considered as a target computational cost reduction factor with $1/\alpha$ being the target speed up factor. However, intuitively there will be a limit as to how small one can choose α before the statistical error incurred from the estimates of $\boldsymbol{\mu}_r$ and $\boldsymbol{\Sigma}_r$ is large enough to render the approximate moment matching transform inaccurate. It is non-trivial to analyze MM-SMC-ABC to obtain a theoretical guideline for choosing α, therefore we perform a computational benchmark to obtain a heuristic.

Here, using different values for α we repeatedly solve the weak Allee model (Section 7.3.3) and the scratch assay model (Section 7.3.4). For both inverse problems we applied MM-SMC-ABC under identical conditions as in Sections 7.3.3 and 7.3.4 with the exception of the tuning parameter α that takes values from the sequence $\{\alpha_k\}_{k=0}^{5}$ with $\alpha_0 = 0.8$ and $\alpha_k = \alpha_{k-1}/2$ for

$k > 0$. For each α_k in the sequence, we consider N independent applications of MM-SMC-ABC. The computational cost for each α_k is denoted by Cost(α_k) and represents the run time in seconds for an application of MM-SMC-ABC with tuning parameter α_k. We also calculate an error metric,

$$\text{Error}(\alpha_k) = \mathcal{E}\left(\boldsymbol{\Theta}_R, \boldsymbol{\Theta}_R(\alpha_k), P\right),$$

where $\boldsymbol{\Theta}_R = \{\boldsymbol{\theta}_R^i\}_{i=1}^{\mathcal{M}}$ is a set of particles from an application of SMC-ABC using the expensive stochastic model, and $\boldsymbol{\Theta}_R(\alpha_k) = \left(\{\boldsymbol{\theta}_R^i\}_{i=1}^{\lceil \alpha_k \mathcal{M} \rceil} \cup \{\bar{\boldsymbol{\theta}}_R^i\}_{i=1}^{\lfloor (1-\alpha_k)\mathcal{M} \rfloor}\right)$ is the pooled exact and approximate transformed particles from the jth application of MM-SMC-ABC. For $P \in \mathbb{N}$, the function $\mathcal{E}(\cdot, \cdot, P)$ is the P-th order empirical moment-matching distance (Liao et al., 2015; Lilliacci and Khammash, 2010; Zechner et al., 2012), given by

$$\mathcal{E}(\mathbf{X}, \mathbf{Y}, P) = \sum_{m=0}^{P} \sum_{\mathbf{b} \in S_m} \frac{1}{|S_m|^m} \left(\frac{\hat{\mu}(\mathbf{X})^{\mathbf{b}} - \hat{\mu}(\mathbf{Y})^{\mathbf{b}}}{\hat{\mu}(\mathbf{X})^{\mathbf{b}}} \right)^2,$$

for two sample sets $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{\mathcal{M}}\}$ and $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{\mathcal{M}}\}$ with $\mathbf{x}_i, \mathbf{y}_i \in \mathbb{R}^n$ for $n \geq 1$, and $S_m = \{\mathbf{b} : \mathbf{b} \in \mathbb{N}^n, \|\mathbf{b}\| = m\}$. For any n-dimensional discrete vector $\mathbf{b} = [b_1, b_2, \ldots, b_n]^{\mathrm{T}} \in \mathbb{N}^n$, then $\hat{\mu}(\mathbf{X})^{\mathbf{b}}$ is the $\mathbf{b}$th empirical raw moment of the sample set $\mathbf{X}$,

$$\hat{\mu}(\mathbf{X})^{\mathbf{b}} = \frac{1}{\mathcal{M}} \sum_{i=1}^{\mathcal{M}} \mathbf{x}_i^{\mathbf{b}},$$

where $\mathbf{x}_i^{\mathbf{b}} = x_{i,1}^{b_1} \times x_{i,2}^{b_2} \times \cdots \times x_{i,n}^{b_n}$. Note that P must be greater than the number of moments that are matched in the approximate transform (Equation (7.10)) to ensure that MM-SMC-ABC is improving the accuracy in higher moments also.

We estimate the average Cost(α_k) and Error(α_k) for each value of α_k for both the weak Allee effect and the scratch assay inverse problems. Figure 7.5 displays the estimates and standard errors given $P = 6$ and $N = 10$, with the value of α_k shown. We emphasize that $P > 2$, that is our error measure compares the first six moments while only two moments are matched. There is clearly a threshold for α such that error becomes highly variable if α is smaller than this threshold. For both the weak Allee effect model (Figure 7.5(A),(C)) and the scratch assay model (Figure 7.5(B),(D)), the optimal choice of α is located between $\alpha_2 = 0.2$ and $\alpha_4 = 0.05$. Therefore, we suggest a heuristic of $\alpha \in [0.1, 0.2]$ to be a reliable choice. If extra performance is needed $\alpha \in [0.05, 0.1)$ may also be acceptable, but if accuracy is of the utmost importance

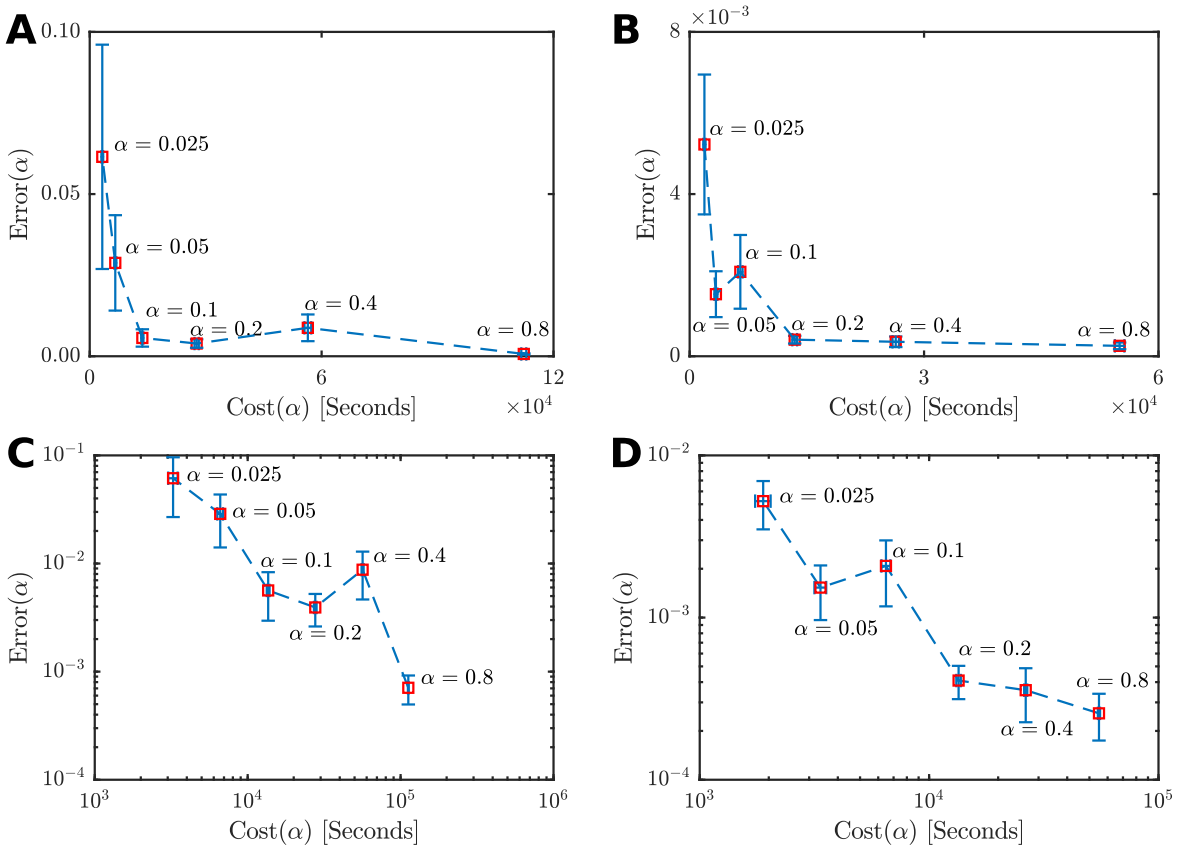

Figure 7.5: Error versus cost plots for different values of the tuning parameter α. Averages and standard errors are shown for $N = 10$ independent applications of the MM-SMC-ABC method to the (A) weak Allee effect model and (B) the scratch assay model. Results in (A)–(B) are shown in (C)–(D) in a log scale for clarity.

then $\alpha \approx 0.2$ seems to be the most robust choice. In general, the choice of optimal α is still an open problem and is likely to be impacted by the specific nature of the relationship between the exact model and the approximate model.

7.3.6 Summary

The two examples presented here demonstrate the efficacy of our two new methods, PC-SMC-ABC and MM-SMC-ABC, for ABC inference with expensive stochastic discrete models. In the first example, the weak Allee model, data were generated using parameters that violate standard continuum-limit assumptions; in the second, the scratch assay model, the Fisher-KPP continuum limit is known to be a good approximation in the parameter regime of the generated data. In both examples, final inferences are biased when the continuum limit is exclusively

relied on in the SMC-ABC sampler. However, the results from our new algorithms, PC-SMC-ABC and MM-SMC-ABC, show significantly more accurate posteriors can be computed at a fraction of the cost of the full SMC-ABC using the discrete model, with speed improvements over an order of magnitude.

As mentioned in Section 7.2.3, the tuning parameter, α, in the MM-SMC-ABC method effectively determines the trade-off between the computational speed of the approximate model and the accuracy of the expensive stochastic model. The values $\alpha = 0$ and $\alpha = 1$ correspond to performing inference exclusively with, respectively, the continuum limit and the stochastic discrete model. Based on numerical experimentation, we find that $\alpha \approx 0.1$ is quite reasonable, however, this conclusion will be dependent on the specific model, the parameter space dimensionality, and the number of particles used for the SMC scheme.

7.4 Discussion

In the life sciences, computationally challenging stochastic discrete models are routinely used to characterize the dynamics of biological populations (Codling et al., 2008; Callaghan et al., 2006; Simpson et al., 2010). In practice, approximations such as the mean-field continuum limit are often derived and used in place of the discrete model for analysis and, more recently, for inference. However, parameter inferences will be biased when the approximate model is solely utilized for inference, even in cases when the approximate model provides an accurate description of the average behavior of the stochastic discrete model.

We provide a new approach to inference for stochastic models that maintains all the methodological benefits of working with discrete mathematical models, while avoiding the computational bottlenecks of relying solely upon repeated expensive stochastic simulations. Our two new algorithms, PC-SMC-ABC and MM-SMC-ABC, utilize samples from the approximate model inference problem in different ways to accelerate SMC-ABC sampling. The PC-SMC-ABC method is unbiased, and we demonstrate computational improvements of up to a factor of almost four are possible. If some bias is acceptable, then MM-SMC-ABC can provide further improvements of approximately a factor of ten. In general, the expected speedup will always be around $1/\alpha$, and $\alpha \approx 0.1$ is reasonable based on our numerical investigations.

There are some assumptions in our approach that could be generalized in future work. First, in PC-SMC-ABC, we assume that the condition in Equation (7.3) holds for all ϵ_r; this is reasonable for the models we consider since we never observe a decrease in performance. However, it may be possible for the bias in the approximate model to be so extreme for some ϵ_r that the condition in Equation (7.3) is violated, leading to a decrease in performance at specific transitions. Acceptance probabilities could be estimated by performing a small set of trial samples from both $\eta_{r-1}(\boldsymbol{\theta}_{r-1})$ and $\tilde{\eta}_r(\boldsymbol{\theta}_r)$ proposal mechanisms, enabling automatic selection of the optimal proposal mechanism. Second, in the moment matching transform proposed in Equation (7.8), we use two moments only as this is sufficient for the problems we consider here with numerical examples demonstrating accuracy in the first six moments. However, our methodology is sufficiently flexible that additional moments can be incorporated if necessary. While including higher moments will improve the accuracy of the moment-matching transform, more samples from the exact model will be required to achieve initial estimates of these moments resulting in eroded performance due to a higher optimal value of α.

While the performance improvements we demonstrate here are significant, it is also possible to obtain improvements of similar order through detailed code optimization techniques applied to standard SMC-ABC. We emphasize, that our schemes would also benefit from such optimizations as advanced vectorization and parallelization to further improve their performance (Lee et al., 2010; Warne et al., 2019d). Our algorithm extensions are also more direct to implement over advanced high performance computing techniques for acceleration of computational schemes.

There are many extensions to our methods that could be considered. We have based our presentation on a form of an SMC-ABC sampler that uses a fixed sequence of thresholds. However, the ideas of using the preconditioning distribution, as in PC-SMC-ABC, and the moment matching transform, as in MM-SMC-ABC, are applicable to SMC schemes that adaptively select thresholds (Drovandi and Pettitt, 2011). Recently, there have been a number of state-of-the-art inference schemes introduced based on multilevel Monte Carlo (MLMC) (Giles, 2015) (Chapter 4). Our new SMC-ABC schemes could exploit MLMC to combine samples from all acceptance thresholds using a coupling scheme and bias correction telescoping summation, such as in the work of Jasra et al. (2019) and Chapter 6. Early accept/rejection schemes, such as those considered by Prangle (2016), Prescott and Baker (2018), and Lester (2018), could

also be introduced for the sampling steps involving the expensive discrete model. Lastly, the PC-SMC-ABC and the MM-SMC-ABC methods could also be applied together and possibly lead to a compounding effect in the performance.

Delayed acceptance schemes (Banterle et al., 2019; Everitt and Rowińska, 2017; Golightly et al., 2015) are also an alternative approach with similar motivations to the methods we propose in this work. However, these approaches can be highly sensitive to false negatives, that is, cases where, for a given $\boldsymbol{\theta}$, the approximate model is rejected but the exact model would have been accepted. Due to this sensitivity to false negatives, the delayed acceptance form of ABC can be biased. Our PC-SMC-ABC approach is not affected by false negatives due to the use of the second set of proposal kernels, $\tilde{q}_r(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})$. This ensures that PC-SMC-ABC is unbiased, which is a distinct advantage over delayed acceptance ABC.

We have demonstrated our methods using a two-dimensional lattice-based discrete random-walk model that leads to mean-field continuum-limit approximations with linear diffusion and a source term of the form $\lambda \mathcal{C} f(\mathcal{C})$. However, our methods are more widely applicable. We could further generalize the model to deal with a more general class of reaction-diffusion continuum limits involving nonlinear diffusion (Witelski, 1995) (Chapter 3) and generalized proliferation mechanisms (Simpson et al., 2013; Tsoularis and Wallace, 2002). Our framework is also relevant to lattice-free discrete models (Codling et al., 2008; Browning et al., 2018) and higher dimensional lattice-based models (Browning et al., 2019); we expect the computational improvements will be even more significant in this case. Many other forms of model combinations are also possible. For example, a sequence of continuum models of increasing complexity could be considered, as in Browning et al. (2019). Alternatively, a sequence of numerical approximations of increasing accuracy could be used for inference using a complex target PDE model (Cotter et al., 2010). Linear mapping approximations of higher order chemical reaction network models, such as in Cao and Grima (2018), could also exploit our approach. Another particularly relevant and very general application in systems biology would be to utilize reaction rate equations, that are deterministic ODEs, as approximations to stochastic chemical kinetics models (Higham, 2008; Wilkinson, 2009).

In this work, novel methods have been presented for exploiting approximate models to accelerate Bayesian inference for expensive stochastic models. We have shown that, even when the approximation leads to biased parameter inferences, it can still inform the proposal mechanisms

for ABC samplers using the stochastic model. The computational improvements are promising and Bayesian analysis of expensive stochastic models will be more computationally feasible as a result.

7.5 Supplementary Material

7.5.1 Proposal efficiency and adaptive proposal kernels

The SMC-ABC algorithm, as presented by Sisson et al. (2007) and Toni et al. (2009) is given in 7.3. Here, the process of sampling at the target threshold, ϵ_r, given the weights of the previous

Algorithm 7.3 SMC-ABC

Initialize $\boldsymbol{\theta}_0^i \sim p(\boldsymbol{\theta})$ and $w_0^i = 1/\mathcal{M}$, for $i = 1, \ldots, \mathcal{M}$;
for $r = 1, \ldots, R$ **do**
  for $i = 1, \ldots, \mathcal{M}$ **do**
    repeat
      Set $\boldsymbol{\theta}^* \leftarrow \boldsymbol{\theta}_{r-1}^j$ with probability $w_{r-1}^j / \left[\sum_{k=1}^{\mathcal{M}} w_{r-1}^k\right]$;
      Sample transition kernel, $\boldsymbol{\theta}^{**} \sim q_r(\boldsymbol{\theta} \mid \boldsymbol{\theta}^*)$;
      Generate data, $\mathcal{D}_s \sim s(\mathcal{D} \mid \boldsymbol{\theta}^{**})$;
    until $\rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_r$
    Set $\boldsymbol{\theta}_r^i \leftarrow \boldsymbol{\theta}^{**}$;
    Set $w_r^i \leftarrow p(\boldsymbol{\theta}_r^i) / \left[\sum_{j=1}^{\mathcal{M}} w_{r-1}^j q_r(\boldsymbol{\theta}_r^i \mid \boldsymbol{\theta}_{r-1}^j)\right]$;
  end for
  Resample weighted particles, $\{(\boldsymbol{\theta}_r^i, w_r^i)\}_{i=1}^{\mathcal{M}}$, with replacement;
  Set $w_r^i \leftarrow 1/\mathcal{M}$ for all $i = 1, \ldots, \mathcal{M}$;
end for

threshold, ϵ_{r-1}, is described by (Filippi et al., 2013)

$$\xi_r\left(\boldsymbol{\theta}_r \mid \mathcal{D}\right) = \frac{1}{a_{r,r-1}} \int_{\boldsymbol{\Theta}} \int_{\mathbb{D}} \mathbb{1}_{B(\mathcal{D},\epsilon_r)}\left(\mathcal{D}_s\right) s(\mathcal{D}_s \mid \boldsymbol{\theta}_r) q_r(\boldsymbol{\theta}_r \mid \boldsymbol{\theta}_{r-1}) w(\boldsymbol{\theta}_{r-1}) \, \mathrm{d}\mathcal{D}_s \mathrm{d}\boldsymbol{\theta}_{r-1}, \quad (7.16)$$

where the data space, $\mathbb{D}$, has dimensionality d, $B(\mathcal{D}, \epsilon_r)$ is a d-dimensional ball centered on the data with radius ϵ_r, and $\mathbb{1}_A(x)$ denotes the indicator function with $\mathbb{1}_A(x) = 1$, if $x \in A$, otherwise $\mathbb{1}_A(x) = 0$. The normalization constant, $a_{r,r-1}$, can be interpreted as the average acceptance probability across all particles. This can be seen by noting that Equation (7.16) can

be reduced to

$$\begin{aligned}\xi_r\left(\boldsymbol{\theta}_r \mid \mathcal{D}\right) &= \frac{1}{a_{r,r-1}} \int_{\mathbb{D}} \mathbb{1}_{B(\mathcal{D},\epsilon_r)}\left(\mathcal{D}_s\right) s(\mathcal{D}_s \mid \boldsymbol{\theta}_r) \left[\int_{\boldsymbol{\Theta}} q_r(\boldsymbol{\theta}_r \mid \boldsymbol{\theta}_{r-1}) w(\boldsymbol{\theta}_{r-1}) \,\mathrm{d}\boldsymbol{\theta}_{r-1}\right] \mathrm{d}\mathcal{D}_s \\ &= \frac{\eta_{r-1}(\boldsymbol{\theta}_r)}{a_{r,r-1}} \int_{\mathbb{D}} \mathbb{1}_{B(\mathcal{D},\epsilon_r)}\left(\mathcal{D}_s\right) s(\mathcal{D}_s \mid \boldsymbol{\theta}_r) \,\mathrm{d}\mathcal{D}_s \\ &= \frac{\eta_{r-1}(\boldsymbol{\theta}_r)\mathbb{E}\left[\mathbb{1}_{B(\mathcal{D},\epsilon_r)}\left(\mathcal{D}_s\right) \mid \boldsymbol{\theta}_r\right]}{a_{r,r-1}} \\ &= \frac{\eta_{r-1}(\boldsymbol{\theta}_r)\mathbb{P}(\mathcal{D}_s \in B(\mathcal{D},\epsilon_r) \mid \boldsymbol{\theta}_r)}{a_{r,r-1}}. \end{aligned} \tag{7.17}$$

Here the distribution $\eta_{r-1}(\boldsymbol{\theta}_r)$ represent the proposal mechanism and $\mathbb{P}(\mathcal{D}_s \in B(\mathcal{D},\epsilon_r) \mid \boldsymbol{\theta}_r)$ is the probability that simulated data is within ϵ_r of the data $\mathcal{D}$ given a parameter value $\boldsymbol{\theta}_r$. Therefore, the normalizing constant is

$$\begin{aligned} a_{r,r-1} &= \int_{\boldsymbol{\Theta}} \eta_{r-1}(\boldsymbol{\theta}_r)\mathbb{P}(\mathcal{D}_s \in B(\mathcal{D},\epsilon_r) \mid \boldsymbol{\theta}_r) \,\mathrm{d}\boldsymbol{\theta}_r \\ &= \mathbb{E}\left[\mathbb{P}(\mathcal{D}_s \in B(\mathcal{D},\epsilon_r) \mid \boldsymbol{\theta}_r)\right], \end{aligned} \tag{7.18}$$

that is, $a_{r,r-1}$ is the average acceptance probability.

From a computational perspective, the goal is to choose $\eta_{r-1}(\boldsymbol{\theta}_r)$ to maximize $a_{r,r-1}$. However, this would not necessarily result in a $\xi_r\left(\boldsymbol{\theta}_r \mid \mathcal{D}\right)$ that is an accurate approximation to the true target ABC posterior $p(\boldsymbol{\theta}_r \mid \rho(\mathcal{D},\mathcal{D}_s) \leq \epsilon_r)$. To achieve this goal, we require $\eta_{r-1}(\boldsymbol{\theta}_r)$ such that the Kullback-Leibler divergence (Kullback and Leibler, 1951), $D_{\mathrm{KL}}\left(\xi_r\left(\cdot \mid \mathcal{D}\right); p(\cdot \mid \rho(\mathcal{D},\mathcal{D}_s) \leq \epsilon_r)\right)$, is minimized.

Beaumont et al. (2009) and Filippi et al. (2013) demonstrate how the latter goal provides insight into how to optimally choose $\eta_{r-1}(\boldsymbol{\theta}_r)$. The key is to note that $D_{\mathrm{KL}}\left(\xi_r\left(\cdot \mid \mathcal{D}\right); p(\cdot \mid \rho(\mathcal{D},\mathcal{D}_s) \leq \epsilon_r)\right)$ can be decomposed as follows,

$$\begin{aligned} D_{\mathrm{KL}}\left(\xi_r\left(\cdot \mid \mathcal{D}\right); p(\cdot \mid \rho(\mathcal{D},\mathcal{D}_s) \leq \epsilon_r)\right) &= \int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}_r \mid \rho(\mathcal{D},\mathcal{D}_s) \leq \epsilon_r) \log_e \left[\frac{p(\boldsymbol{\theta}_r \mid \rho(\mathcal{D},\mathcal{D}_s) \leq \epsilon_r)}{\xi_r(\boldsymbol{\theta}_r \mid \mathcal{D})}\right] \mathrm{d}\boldsymbol{\theta}_r \\ = \int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}_r \mid \rho(\mathcal{D},\mathcal{D}_s) \leq \epsilon_r) \log_e \left[\frac{a_{r,r-1} p(\boldsymbol{\theta}_r \mid \rho(\mathcal{D},\mathcal{D}_s) \leq \epsilon_r)}{\eta_{r-1}(\boldsymbol{\theta}_r) \int_{\mathbb{D}} \mathbb{1}_{B(\mathcal{D},\epsilon_r)}\left(\mathcal{D}_s\right) s(\mathcal{D}_s \mid \boldsymbol{\theta}_r) \,\mathrm{d}\mathcal{D}_s}\right] \mathrm{d}\boldsymbol{\theta}_r \end{aligned}$$

$$
\begin{aligned}
&= \int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}_r \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_r) \log_e \left[\frac{p(\boldsymbol{\theta}_r \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_r)}{\eta_{r-1}(\boldsymbol{\theta}_r)} \right] \mathrm{d}\boldsymbol{\theta}_r \\
&\quad + \int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}_r \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_r) \log_e \left[a_{r,r-1} \right] \mathrm{d}\boldsymbol{\theta}_r \\
&\quad - \int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}_r \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_r) \log_e \left[\int_{\mathbb{D}} \mathbb{1}_{B(\mathcal{D}, \epsilon_r)}(\mathcal{D}_s)\, s(\mathcal{D}_s \mid \boldsymbol{\theta}_r)\, \mathrm{d}\mathcal{D}_s \right] \mathrm{d}\boldsymbol{\theta}_r \\
&= D_{\mathrm{KL}}\left(\eta_{r-1}(\cdot)\,; p(\cdot \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_r)\right) - \mathbb{E}\left[\log_e \left(\int_{\mathbb{D}} \mathbb{1}_{B(\mathcal{D}, \epsilon_r)}(\mathcal{D}_s)\, s(\mathcal{D}_s \mid \boldsymbol{\theta}_r)\, \mathrm{d}\mathcal{D}_s \right) \right] + \log_e a_{r,r-1} \\
&= D_{\mathrm{KL}}\left(\eta_{r-1}(\cdot)\,; p(\cdot \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_r)\right) - E(\boldsymbol{\theta}_r) + \log_e a_{r,r-1}, \qquad (7.19)
\end{aligned}
$$

where $E(\boldsymbol{\theta}_r) = \mathbb{E}\left[\log_e \left(\int_{\mathbb{D}} \mathbb{1}_{B(\mathcal{D}, \epsilon_r)}(\mathcal{D}_s)\, s(\mathcal{D}_s \mid \boldsymbol{\theta}_r)\, \mathrm{d}\mathcal{D}_s\right)\right]$ is independent of $\eta_{r-1}(\boldsymbol{\theta}_r)$. By rearranging Equation (7.19), we obtain

$$
D_{\mathrm{KL}}\left(\eta_{r-1}(\cdot)\,; p(\cdot \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_r)\right) = D_{\mathrm{KL}}\left(\xi_r(\cdot \mid \mathcal{D})\,; p(\cdot \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_r)\right) + E(\boldsymbol{\theta}_r) - \log_e a_{r,r-1}.
$$

That is, minimizing $D_{\mathrm{KL}}\left(\eta_{r-1}(\cdot)\,; p(\cdot \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_r)\right)$ is equivalent to minimizing $D_{\mathrm{KL}}\left(\xi_r(\cdot \mid \mathcal{D})\,; p(\cdot \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_r)\right)$ and maximizing $a_{r,r-1}$ simultaneously. Therefore, any proposal mechanism that is closer, in the Kullback-Liebler sense, to $p(\cdot \mid \rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon_r)$ is more efficient.

We apply the optimal adaptive scheme of Beaumont et al. (2009) and Filippi et al. (2013) for multivariate Gaussian proposals. That is, we set

$$
q_r(\boldsymbol{\theta}_r \mid \boldsymbol{\theta}_{r-1}) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} \exp\left(-(\boldsymbol{\theta}_r - \boldsymbol{\theta}_{r-1})^{\mathrm{T}} \Sigma^{-1} (\boldsymbol{\theta}_r - \boldsymbol{\theta}_{r-1})/2\right),
$$

where n is the dimensionality of parameter space, $\boldsymbol{\Theta}$, and

$$
\Sigma = \frac{2}{\mathcal{M}-1} \sum_{i=1}^{\mathcal{M}} (\boldsymbol{\theta}_{r-1}^i - \mu_{r-1})(\boldsymbol{\theta}_{r-1}^i - \mu_{r-1})^{\mathrm{T}} \quad \text{with} \quad \mu_{r-1} = \frac{1}{\mathcal{M}} \sum_{i=1}^{\mathcal{M}} \boldsymbol{\theta}_{r-1}^i.
$$

In the PC-SMC-ABC method, we find that independent Gaussian proposals are more effective for the second stage proposal step $\tilde{q}_r(\boldsymbol{\theta}_r^i \mid \tilde{\boldsymbol{\theta}}_r^i)$ since the covariance of the approximate particles $\{\tilde{\boldsymbol{\theta}}_r^i\}_{i=1}^{\mathcal{M}}$ is not a good candidate for the exact target.

7.5.2 Derivation of approximate continuum-limit description

Here we derive the approximate continuum-limit description for our hexagonal lattice based discrete random walk model with generalized crowding function $f : [0,1] \to [-1,1]$. We follow the method of Simpson et al. (2010) and Jin et al. (2016a).

Our main modification is dealing with a potentially negative crowding function. To deal with this case, we define two auxiliary functions,

$$f^+(C) = \begin{cases} f(C) & \text{if } f(C) \geq 0, \\ 0 & \text{otherwise,} \end{cases}, \quad f^-(C) = \begin{cases} |f(C)| & \text{if } f(C) < 0, \\ 0 & \text{otherwise.} \end{cases} \tag{7.20}$$

Note that $f(C) = f^+(C) - f^-(C)$, as this is important later.

We assume a mean-field, that is, for any two lattice sites, $(x_1, y_1), (x_2, y_2) \in \mathbb{R}^2$, then their occupancy probabilities are independent, which results in the property, $\mathbb{E}\left[C((x_1,y_1),t)C((x_2,y_2),t)\right] = \mathbb{E}\left[C((x_1,y_1),t)\right]\mathbb{E}\left[C((x_2,y_2),t)\right]$. Using this property, we denote $\mathcal{C}(x,y,t) = \mathbb{E}\left[C((x,y),t)\right]$ and write the conservation of probability equation that describes the change in occupancy probability of a site over a single time step,

$$\begin{aligned} \Delta\mathcal{C}(x,y,t) = & P_m\left(1-\mathcal{C}(x,y,t)\right)\hat{\mathcal{C}}(x,y,t) - P_m\mathcal{C}(x,y,t)\left(1-\hat{\mathcal{C}}(x,y,t)\right) \\ & + \frac{P_p}{6}\left(1-\mathcal{C}(x,y,t)\right)\left(\sum_{(x',y')\in N(x,y)} \mathcal{C}(x',y',t)\frac{f^+(\hat{\mathcal{C}}(x',y',t)}{1-\hat{\mathcal{C}}(x',y',t)}\right) \\ & - P_p\mathcal{C}(x,y,t)f^-(\hat{\mathcal{C}}(x,y,t), \end{aligned} \tag{7.21}$$

where $\Delta\mathcal{C}(x,y,t) = \mathcal{C}(x,y,t+\tau) - \mathcal{C}(x,y,t)$,

$$\hat{\mathcal{C}}(x,y,t) = \frac{1}{6}\sum_{(x',y')\in N(x,y)} \mathcal{C}(x',y',t), \tag{7.22}$$

and

$$\hat{\mathcal{C}}(x',y',t) = \frac{1}{6}\sum_{(x'',y'')\in N(x',y')} \mathcal{C}(x'',y'',t). \tag{7.23}$$

The conservation of probability equation (Equation 7.21) deserves some interpretation. The first term is the probability that site (x,y) is unoccupied at time t and becomes occupied at

time $t+\tau$ due to a successful motility event that moves an agent from an occupied neighboring site into (x, y). The second term is the probability that site (x, y) is occupied at time t and becomes unoccupied at time $t+\tau$ due to a successful motility event that moves the agent away to a neighboring site. The third term is the probability that site (x, y) is unoccupied at time t and becomes occupied at time $t+\tau$ due to a successful proliferation event from an occupied neighboring site. The final term is the probability that the site (x, y) is occupied at time t and becomes unoccupied at time $t+\tau$ due to $f(\hat{C}((x, y), t)) < 0$.

For a hexagonal lattice site, (x, y), we have six immediate neighboring lattice sites,
$(x_1, y_1) = (x, y+\delta)$, $(x_2, y_2) = (x, y-\delta)$, $(x_3, y_3) = (x+\delta\sqrt{3}/2, y+\delta/2)$,
$(x_4, y_4) = (x-\delta\sqrt{3}/2, y+\delta/2)$, $(x_5, y_5) = (x+\delta\sqrt{3}/2, y-\delta/2)$, and
$(x_6, y_6) = (x-\delta\sqrt{3}/2, y-\delta/2)$. The positions of these neighbors are shown in the schematic in Figure 7.6.

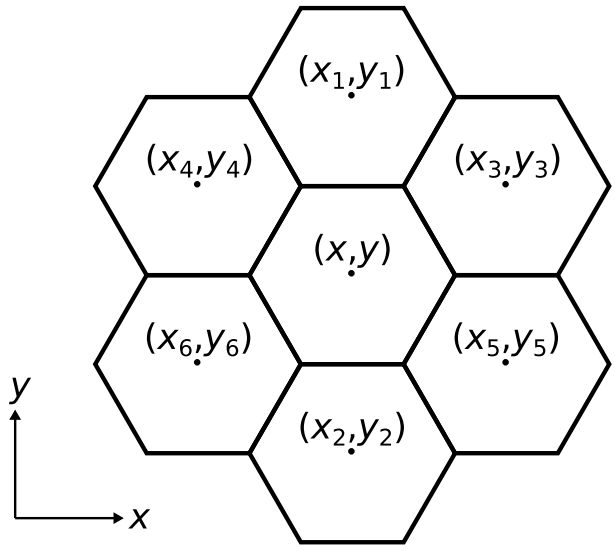

Figure 7.6: Schematic of the local neighborhood for the hexagonal lattice. Neighbor coordinates are $(x_1, y_1) = (x, y+\delta)$, $(x_2, y_2) = (x, y-\delta)$, $(x_3, y_3) = (x+\delta\sqrt{3}/2, y+\delta/2)$, $(x_4, y_4) = (x-\delta\sqrt{3}/2, y+\delta/2)$, $(x_5, y_5) = (x+\delta\sqrt{3}/2, y-\delta/2)$, and $(x_6, y_6) = (x-\delta\sqrt{3}/2, y-\delta/2)$.

We obtain an expression for $\mathcal{C}(x, y, t)$ at each of the six neighboring lattice points by associating $\mathcal{C}(x, y, t)$ with a continuous function and performing a Taylor expansion about the point (x, y).

The result is,

$$
\begin{aligned}
\mathcal{C}(x_1,y_1,t) =& \mathcal{C}(x,y,t) + \delta\frac{\partial\mathcal{C}(x,y,t)}{\partial y} + \frac{\delta^2}{2}\frac{\partial^2\mathcal{C}(x,y,t)}{\partial y^2} + \mathcal{O}(\delta^3),\\
\mathcal{C}(x_2,y_2,t) =& \mathcal{C}(x,y,t) - \delta\frac{\partial\mathcal{C}(x,y,t)}{\partial y} + \frac{\delta^2}{2}\frac{\partial^2\mathcal{C}(x,y,t)}{\partial y^2} + \mathcal{O}(\delta^3),\\
\mathcal{C}(x_3,y_3,t) =& \mathcal{C}(x,y,t) + \frac{\delta\sqrt{3}}{2}\frac{\partial\mathcal{C}(x,y,t)}{\partial x} + \frac{\delta}{2}\frac{\partial\mathcal{C}(x,y,t)}{\partial y}\\
&+ \frac{\delta^2}{2}\left[\frac{3}{4}\frac{\partial^2\mathcal{C}(x,y,t)}{\partial x^2} + \frac{\sqrt{3}}{2}\frac{\partial^2\mathcal{C}(x,y,t)}{\partial x\partial y} + \frac{1}{4}\frac{\partial^2\mathcal{C}(x,y,t)}{\partial y^2}\right] + \mathcal{O}(\delta^3),\\
\mathcal{C}(x_4,y_4,t) =& \mathcal{C}(x,y,t) - \frac{\delta\sqrt{3}}{2}\frac{\partial\mathcal{C}(x,y,t)}{\partial x} + \frac{\delta}{2}\frac{\partial\mathcal{C}(x,y,t)}{\partial y}\\
&+ \frac{\delta^2}{2}\left[\frac{3}{4}\frac{\partial^2\mathcal{C}(x,y,t)}{\partial x^2} - \frac{\sqrt{3}}{2}\frac{\partial^2\mathcal{C}(x,y,t)}{\partial x\partial y} + \frac{1}{4}\frac{\partial^2\mathcal{C}(x,y,t)}{\partial y^2}\right] + \mathcal{O}(\delta^3),\\
\mathcal{C}(x_5,y_5,t) =& \mathcal{C}(x,y,t) + \frac{\delta\sqrt{3}}{2}\frac{\partial\mathcal{C}(x,y,t)}{\partial x} - \frac{\delta}{2}\frac{\partial\mathcal{C}(x,y,t)}{\partial y}\\
&+ \frac{\delta^2}{2}\left[\frac{3}{4}\frac{\partial^2\mathcal{C}(x,y,t)}{\partial x^2} + \frac{\sqrt{3}}{2}\frac{\partial^2\mathcal{C}(x,y,t)}{\partial x\partial y} + \frac{1}{4}\frac{\partial^2\mathcal{C}(x,y,t)}{\partial y^2}\right] + \mathcal{O}(\delta^3),\\
\mathcal{C}(x_6,y_6,t) =& \mathcal{C}(x,y,t) - \frac{\delta\sqrt{3}}{2}\frac{\partial\mathcal{C}(x,y,t)}{\partial x} - \frac{\delta}{2}\frac{\partial\mathcal{C}(x,y,t)}{\partial y}\\
&+ \frac{\delta^2}{2}\left[\frac{3}{4}\frac{\partial^2\mathcal{C}(x,y,t)}{\partial x^2} + \frac{\sqrt{3}}{2}\frac{\partial^2\mathcal{C}(x,y,t)}{\partial x\partial y} + \frac{1}{4}\frac{\partial^2\mathcal{C}(x,y,t)}{\partial y^2}\right] + \mathcal{O}(\delta^3).
\end{aligned}
\tag{7.24}
$$

Therefore, we obtain an expression for $\hat{\mathcal{C}}(x,y,t)$,

$$
\begin{aligned}
\hat{\mathcal{C}}(x,y,t) &= \frac{1}{6}\sum_{(x',y')\in N(x,y)} \mathcal{C}(x',y',t)\\
&= \frac{1}{6}\sum_{k=1}^{6}\mathcal{C}(x_k,y_k,t)\\
&= \mathcal{C}(x,y,t) + \frac{\delta^2}{4}\left[\frac{\partial^2\mathcal{C}(x,y,t)}{\partial x^2} + \frac{\partial^2\mathcal{C}(x,y,t)}{\partial y^2}\right] + \mathcal{O}(\delta^3).
\end{aligned}
\tag{7.25}
$$

This expression is required for the first, second, and fourth terms in the conservation equation (Equation (7.21)).

To deal with the third term in Equation (7.21) we require an expression for

$g^+(\hat{\mathcal{C}}(x_k, y_k, t)) = f^+(\hat{\mathcal{C}}(x_k, y_k, t))/(1 - \hat{\mathcal{C}}(x_k, y_k, t))$, $k = 1, 2, \ldots, 6$. By combining Equation (7.24) and Equation (7.25) we obtain

$$
\begin{aligned}
\hat{\mathcal{C}}(x_1, y_1, t) =& \mathcal{C}(x, y, t) + \delta \frac{\partial \mathcal{C}(x, y, t)}{\partial y} + \frac{\delta^2}{4}\left[\frac{\partial^2 \mathcal{C}(x, y, t)}{\partial x^2} + 3\frac{\partial^2 \mathcal{C}(x, y, t)}{\partial y^2}\right] + \mathcal{O}(\delta^3) \\
\hat{\mathcal{C}}(x_2, y_2, t) =& \mathcal{C}(x, y, t) + \delta \frac{\partial \mathcal{C}(x, y, t)}{\partial y} + \frac{\delta^2}{4}\left[\frac{\partial^2 \mathcal{C}(x, y, t)}{\partial x^2} + 3\frac{\partial^2 \mathcal{C}(x, y, t)}{\partial y^2}\right] + \mathcal{O}(\delta^3) \\
\hat{\mathcal{C}}(x_3, y_3, t) =& \mathcal{C}(x, y, t) + \frac{\delta\sqrt{3}}{2}\frac{\partial \mathcal{C}(x, y, t)}{\partial x} + \frac{\delta}{2}\frac{\partial \mathcal{C}(x, y, t)}{\partial y} \\
&+ \frac{\delta^2}{4}\left[\frac{5}{2}\frac{\partial^2 \mathcal{C}(x, y, t)}{\partial x^2} + \sqrt{3}\frac{\partial^2 \mathcal{C}(x, y, t)}{\partial x^2} y + \frac{3}{2}\frac{\partial^2 \mathcal{C}(x, y, t)}{\partial y^2}\right] + \mathcal{O}(\delta^3) \\
\hat{\mathcal{C}}(x_4, y_4, t) =& \mathcal{C}(x, y, t) - \frac{\delta\sqrt{3}}{2}\frac{\partial \mathcal{C}(x, y, t)}{\partial x} + \frac{\delta}{2}\frac{\partial \mathcal{C}(x, y, t)}{\partial y} \\
&+ \frac{\delta^2}{4}\left[\frac{5}{2}\frac{\partial^2 \mathcal{C}(x, y, t)}{\partial x^2} - \sqrt{3}\frac{\partial^2 \mathcal{C}(x, y, t)}{\partial x^2} y + \frac{3}{2}\frac{\partial^2 \mathcal{C}(x, y, t)}{\partial y^2}\right] + \mathcal{O}(\delta^3), \\
\hat{\mathcal{C}}(x_5, y_5, t) =& \mathcal{C}(x, y, t) + \frac{\delta\sqrt{3}}{2}\frac{\partial \mathcal{C}(x, y, t)}{\partial x} - \frac{\delta}{2}\frac{\partial \mathcal{C}(x, y, t)}{\partial y} \\
&+ \frac{\delta^2}{4}\left[\frac{5}{2}\frac{\partial^2 \mathcal{C}(x, y, t)}{\partial x^2} - \sqrt{3}\frac{\partial^2 \mathcal{C}(x, y, t)}{\partial x^2} y + \frac{3}{2}\frac{\partial^2 \mathcal{C}(x, y, t)}{\partial y^2}\right] + \mathcal{O}(\delta^3), \\
\hat{\mathcal{C}}(x_3, y_3, t) =& \mathcal{C}(x, y, t) - \frac{\delta\sqrt{3}}{2}\frac{\partial \mathcal{C}(x, y, t)}{\partial x} - \frac{\delta}{2}\frac{\partial \mathcal{C}(x, y, t)}{\partial y} \\
&+ \frac{\delta^2}{4}\left[\frac{5}{2}\frac{\partial^2 \mathcal{C}(x, y, t)}{\partial x^2} + \sqrt{3}\frac{\partial^2 \mathcal{C}(x, y, t)}{\partial x^2} y + \frac{3}{2}\frac{\partial^2 \mathcal{C}(x, y, t)}{\partial y^2}\right] + \mathcal{O}(\delta^3).
\end{aligned}
\tag{7.26}
$$

Each expression in Equation (7.26) is of the form $\hat{\mathcal{C}}(x_k, y_k, t) = \mathcal{C}(x, y, t) + \bar{\mathcal{C}}_k$, where $\bar{\mathcal{C}}_k = \mathcal{O}(\delta)$. This allows us to consider the Taylor expansion of $g^+(\hat{\mathcal{C}}(x_k, y_k, t))$ about the density $\mathcal{C}(x, y, t)$, that is,

$$
g^+(\mathcal{C} + \bar{\mathcal{C}}_k) = g^+(\mathcal{C}) + \bar{\mathcal{C}}_k \frac{\mathrm{d}g^+(\mathcal{C})}{\mathrm{d}\mathcal{C}} + \frac{\bar{\mathcal{C}}_k^2}{2}\frac{\mathrm{d}^2 g^+(\mathcal{C})}{\mathrm{d}\mathcal{C}^2} + \mathcal{O}(\delta^3). \tag{7.27}
$$

Using Equation (7.27), we can obtain an expression for the summation within the third term of the conservation equation (Equation (7.21)), that is,

$$
\begin{aligned}
\sum_{(x', y') \in N(x, y)} \mathcal{C}(x', y', t)\frac{f^+(\hat{\mathcal{C}}(x', y', t)}{1 - \hat{\mathcal{C}}(x', y', t)} &= \sum_{k=1}^{6} \mathcal{C}(x_k, y_k, t) g^+(\mathcal{C}(x, y, t) + \bar{\mathcal{C}}_k) \\
&= g^+(\mathcal{C}(x, y, t)) \sum_{k=1}^{6} \mathcal{C}(x_k, y_k, t) + \frac{\mathrm{d}g^+(\mathcal{C})}{\mathrm{d}\mathcal{C}} \sum_{k=1}^{6} \bar{\mathcal{C}}_k \mathcal{C}(x_k, y_k, t) \\
&\quad + \frac{\mathrm{d}^2 g^+(\mathcal{C})}{\mathrm{d}\mathcal{C}^2} \sum_{k=1}^{6} \frac{\bar{\mathcal{C}}_k^2}{2} \mathcal{C}(x_k, y_k, t) + \mathcal{O}(\delta^3).
\end{aligned}
\tag{7.28}
$$

After substitution of Equation (7.26) into Equation (7.28) and some tedious algebra we arrive at

$$\begin{aligned}\sum_{(x',y')\in N(x,y)} \mathcal{C}(x',y',t)\frac{f^+(\hat{\mathcal{C}}(x',y',t)}{1-\hat{\mathcal{C}}(x',y',t)} =& 6\mathcal{C}(x,y,t)g^+(\mathcal{C}(x,y,t)) \\ &+ \frac{3\delta^2}{2}g^+(\mathcal{C}(x,y,t))\left[\frac{\partial^2\mathcal{C}(x,y,t)}{\partial x^2}+\frac{\partial^2\mathcal{C}(x,y,t)}{\partial y^2}\right] \\ &+ 3\delta^2\mathcal{C}(x,y,t)\frac{\mathrm{d}g^+(\mathcal{C})}{\mathrm{d}\mathcal{C}}\left[\frac{\partial^2\mathcal{C}(x,y,t)}{\partial x^2}+\frac{\partial^2\mathcal{C}(x,y,t)}{\partial y^2}\right] \\ &+ 3\delta^2\frac{\mathrm{d}g^+(\mathcal{C})}{\mathrm{d}\mathcal{C}}\left[\left(\frac{\partial\mathcal{C}(x,y,t)}{\partial x}\right)^2+\left(\frac{\partial\mathcal{C}(x,y,t)}{\partial y}\right)^2\right] \\ &+ \frac{3\delta^2}{2}\mathcal{C}(x,y,t)\frac{\mathrm{d}^2g^+(\mathcal{C})}{\mathrm{d}\mathcal{C}^2}\left[\left(\frac{\partial\mathcal{C}(x,y,t)}{\partial x}\right)^2+\left(\frac{\partial\mathcal{C}(x,y,t)}{\partial y}\right)^2\right] \\ &+ \mathcal{O}(\delta^3). \end{aligned} \tag{7.29}$$

Finally, we substitute Equation (7.25) and Equation (7.29) into Equation (7.21) to obtain

$$\begin{aligned}\Delta\mathcal{C}(x,y,t) =& P_m\left(1-\mathcal{C}(x,y,t)\right)\left[\mathcal{C}(x,y,t)+\frac{\delta^2}{4}\nabla^2\mathcal{C}(x,y,t)\right] \\ &- P_m\mathcal{C}(x,y,t)\left[1-\mathcal{C}(x,y,t)-\frac{\delta^2}{4}\nabla^2\mathcal{C}(x,y,t)\right] \\ &+ \frac{P_p}{6}\left(1-\mathcal{C}(x,y,t)\right)\Bigg[6\mathcal{C}(x,y,t)g^+(\mathcal{C}(x,y,t))+\frac{3\delta^2}{2}g^+(\mathcal{C}(x,y,t))\nabla^2\mathcal{C}(x,y,t) \\ &+ 3\delta^2\mathcal{C}(x,y,t)\frac{\mathrm{d}g^+(\mathcal{C})}{\mathrm{d}\mathcal{C}}\nabla^2\mathcal{C}(x,y,t)+3\delta^2\frac{\mathrm{d}g^+(\mathcal{C})}{\mathrm{d}\mathcal{C}}\left(\nabla\mathcal{C}(x,y,t)\cdot\nabla\mathcal{C}(x,y,t)\right) \\ &+ \frac{3\delta^2}{2}\mathcal{C}(x,y,t)\frac{\mathrm{d}^2g^+(\mathcal{C})}{\mathrm{d}\mathcal{C}^2}\left(\nabla\mathcal{C}(x,y,t)\cdot\nabla\mathcal{C}(x,y,t)\right)\Bigg] \\ &- P_p\mathcal{C}(x,y,t)f^-(\mathcal{C}(x,y,t)+\mathcal{O}(\delta^3). \end{aligned} \tag{7.30}$$

Here we have used the notation $\nabla^2\mathcal{C} = \frac{\partial^2\mathcal{C}}{\partial x^2}+\frac{\partial^2\mathcal{C}}{\partial y^2}$, and $\nabla\mathcal{C}\cdot\nabla\mathcal{C} = (\frac{\partial\mathcal{C}}{\partial x})^2+(\frac{\partial\mathcal{C}}{\partial y})^2$. After rearranging Equation (7.30), we obtain

$$\begin{aligned}\Delta\mathcal{C}(x,y,t) =& \frac{P_m\delta^2}{4}\nabla^2\mathcal{C}(x,y,t)+P_p\mathcal{C}(x,y,t)\left[\left(1-\mathcal{C}(x,y,t)\right)g^+(\mathcal{C}(x,y,t))-f^-(\mathcal{C}(x,y,t))\right] \\ &+ P_p\delta^2H(\mathcal{C}(x,y,t))+\mathcal{O}(\delta^3) \end{aligned} \tag{7.31}$$

where

$$H(\mathcal{C}(x,y,t)) = \frac{1}{4}g^+(\mathcal{C}(x,y,t))\nabla^2\mathcal{C}(x,y,t) + \frac{1}{2}\mathcal{C}(x,y,t)\frac{\mathrm{d}g^+(\mathcal{C})}{\mathrm{d}\mathcal{C}}\nabla^2\mathcal{C}(x,y,t) + \frac{1}{2}\frac{\mathrm{d}g^+(\mathcal{C})}{\mathrm{d}\mathcal{C}}\left(\nabla\mathcal{C}(x,y,t)\cdot\nabla\mathcal{C}(x,y,t)\right) + \frac{1}{4}\mathcal{C}(x,y,t)\frac{\mathrm{d}^2g^+(\mathcal{C})}{\mathrm{d}\mathcal{C}^2}\left(\nabla\mathcal{C}(x,y,t)\cdot\nabla\mathcal{C}(x,y,t)\right).$$

Note that $f^+(\mathcal{C}(x,y,t)) = g^+(\mathcal{C}(x,y,t))(1-\mathcal{C}(x,y,t))$ and, by Equation (7.20), that $f(\mathcal{C}(x,y,t)) = f^+(\mathcal{C}(x,y,t)) - f^-(\mathcal{C}(x,y,t))$. Then Equation (7.31) becomes

$$\Delta\mathcal{C}(x,y,t) = \frac{P_m\delta^2}{4}\nabla^2\mathcal{C}(x,y,t) + P_p\mathcal{C}(x,y,t)f(\mathcal{C}(x,y,t)) + P_p\delta^2H(\mathcal{C}(x,y,t)) + \mathcal{O}(\delta^3).$$

After dividing by the time interval, τ, and choosing $\delta^2 = \mathcal{O}(\tau)$ and $P_p = \mathcal{O}(\tau)$, then we have the following limits,

$$\lim_{\tau\to 0}\frac{P_m\delta^2}{4\tau} = D, \quad \lim_{\tau\to 0}\frac{P_p}{\tau} = \lambda, \quad \lim_{\tau\to 0}\frac{\Delta\mathcal{C}(x,y,t)}{\tau} = \frac{\partial\mathcal{C}(x,y,t)}{\partial t}, \quad \lim_{\tau\to 0}\frac{P_p\delta^2}{\tau} = 0.$$

Therefore, we arrive at the continuum-limit approximation

$$\frac{\partial\mathcal{C}(x,y,t)}{\partial t} = D\nabla^2\mathcal{C}(x,y,t) + \lambda\mathcal{C}(x,y,t)f(\mathcal{C}(x,y,t)).$$

7.5.3 Stochastic simulation of the discrete model

In the work of Jin et al. (2016a), they consider crowding functions with $f : [0,1] \to [0,1]$. This is not quite sufficient to enable a crowding function with a carrying capacity $K < 1$ since motility events will cause the carrying capacity to be violated. Therefore, we take $f : [0,1] \to [-1,1]$. If $f(\hat{C}(\ell_s)) < 0$ then the site ℓ_s is removed from the set of occupied sites at time t, denoted by $\mathcal{L}(t)$, with probability $P_p|f(\hat{C}(\ell_s))|$. Given these definitions, the lattice-based random walk proceeds according to Algorithm 7.4.

Algorithm 7.4 Lattice-based random walk model

Initialize $\mathcal{L}(0) \subset L$ with $|\mathcal{L}(0)| = N$ and $t \leftarrow 0$;
while $t < T$ **do**
  $N \leftarrow |\mathcal{L}(t)|$;
  for $i = [1, 2, \ldots, N]$ **do**
    Choose $\ell_s \in \mathcal{L}(t)$ uniformly at random with probability $1/N$;
    Choose $\ell_m \in \mathcal{N}(\ell_s)$ uniformly at random with probability $1/|\mathcal{N}(\ell_s)|$;
    if $C(\ell_m, t) = 0$ **then**
      Generate $u \sim \mathcal{U}(0, 1)$;
      if $u \leq P_m$ **then**
        Set $\mathcal{L}(t) \leftarrow \{\ell_m\} \cup \mathcal{L}(t) \setminus \{\ell_s\}$;
      end if
    end if
  end for
  for $i = [1, 2, \ldots, N]$ **do**
    Choose $\ell_s \in \mathcal{L}(t)$ uniformly at random with probability $1/N$;
    Set $\hat{C}(\ell_s) \leftarrow [1/|\mathcal{N}(\ell_s)|] \sum_{\ell'_s \in \mathcal{N}(\ell_s)} C(\ell'_s, t)$;
    Generate $u \sim \mathcal{U}(0, 1)$;
    if $u \leq P_p |f(\hat{C}(\ell_s))|$ **then**
      if $f(\hat{C}(\ell_s)) \geq 0$ **then**
        Choose $\ell_p \in \{\ell'_s \in \mathcal{N}(\ell_s) : C(\ell'_s, t) = 0\}$ uniformly at random with probability $1 - \hat{C}(\ell_s)$;
        Set $\mathcal{L}(t) \leftarrow \{\ell_p\} \cup \mathcal{L}(t)$;
      else
        Set $\mathcal{L}(t) \leftarrow \mathcal{L}(t) \setminus \{\ell_s\}$;
      end if
    end if
  end for
  $t \leftarrow t + \tau$
end while

7.5.4 Numerical solutions of continuum-limit differential equations

In our case, the approximate model is a deterministic continuum equation that is either a nonlinear ODE or PDE. In both cases we apply numerical schemes that automatically adapt the time step size to control the truncation error.

7.5.4.1 ODE numerical solutions

The Runge-Kutta-Fehlberg fourth-order-fifth-order method (RKF45) (Fehlberg, 1969) is a numerical scheme in the Runge-Kutta family of methods to approximate the solution of a nonlinear

ODE of the form,

$$\frac{\mathrm{d}\mathcal{C}(t)}{\mathrm{d}t} = h(t, \mathcal{C}(t)), \quad 0 < t,$$

where $h(t, \mathcal{C}(t))$ is a function that satisfies certain regularity conditions and the initial condition, $\mathcal{C}(0)$.

Given the approximate solution, $c_i \approx \mathcal{C}(t_i)$, an embedded pair of Runge-Kutta methods, specifically a fourth and fifth order pair, are used to advanced the solution to $c_{i+1} \approx \mathcal{C}(t_i + \Delta t) + \mathcal{O}(\Delta t^4)$ and estimate the truncation error that can be used to adaptively adjust Δt to ensure the error is always within some specified tolerance τ.

The fourth and fifth order estimates are, respectively,

$$c_{i+1} = c_i + \Delta t \left(\frac{25}{216} k_1 + \frac{1,408}{2,565} k_3 + \frac{2,197}{4,104} k_4 - \frac{1}{5} k_5 \right), \tag{7.32}$$

and

$$c^*_{i+1} = c_i + \Delta t \left(\frac{16}{135} k_1 + \frac{6,656}{12,825} k_3 + \frac{28,561}{56,430} k_4 - \frac{9}{50} k_5 + \frac{2}{55} k_6 \right), \tag{7.33}$$

where

$$\begin{aligned}
k_1 &= h(t_i, c_i), \\
k_2 &= h\left(t_i + \frac{\Delta t}{4}, c_i + \frac{\Delta t}{4} k_1 \right), \\
k_3 &= h\left(t_i + \frac{3\Delta t}{8}, c_i + \Delta t \left[\frac{3}{32} k_1 + \frac{9}{32} k_2 \right] \right), \\
k_4 &= h\left(t_i + \frac{12\Delta t}{13}, c_i + \Delta t \left[\frac{1,932}{2,197} k_1 - \frac{7,200}{2,197} k_2 + \frac{7,296}{2,197} k_3 \right] \right), \\
k_5 &= h\left(t_i + \Delta t, c_i + \Delta t \left[\frac{439}{216} k_1 - 8 k_2 + \frac{3,680}{513} k_3 - \frac{845}{4,104} k_4 \right] \right), \\
k_6 &= h\left(t_i + \frac{\Delta t}{2}, c_i + \Delta t \left[-\frac{8}{27} k_1 + 2 k_2 - \frac{3,544}{2,565} k_3 + \frac{1,859}{4,104} k_4 - \frac{11}{40} k_5 \right] \right).
\end{aligned}$$

Note that the truncation error can be estimated by $\epsilon = |c_{i+1} - c^*_{i+1}|$. After each evaluation of Equation (7.32) and Equation (7.33), a new step size is determined by $\Delta t \leftarrow s \Delta t$, where $s = [\tau/(2\epsilon)]^{1/4}$. If $\epsilon \leq \tau$, then the solution is accepted and the new Δt is used for the next iteration. Alternatively, if $\epsilon > \tau$ then the solution is rejected and a new attempt is made using the new time step. This process is repeated until the solution advances to some desired time point T. For more details and analysis of this method, see Iserles (2008).

7.5.4.2 PDE numerical solutions

In this work, we consider numerical solutions to PDEs of the form

$$\frac{\partial \mathcal{C}(x,t)}{\partial t} = D\frac{\partial^2 \mathcal{C}(x,t)}{\partial x^2} + \lambda \mathcal{C}(x,t) f(\mathcal{C}(x,t)), \quad 0 < t, \quad 0 < x < L, \tag{7.34}$$

with initial conditions

$$\mathcal{C}(x,0) = c(x), \quad t = 0,$$

and Neumann boundary conditions

$$\frac{\partial \mathcal{C}(x,t)}{\partial x} = 0, \quad x = 0 \text{ and } x = L.$$

Given the approximate solution, $c_i^j \approx \mathcal{C}(x_i, t_j)$ for $i = 1, 2, \ldots N$, then we apply a first order backward Euler discretization in time and first order central differences in space to yield

$$\begin{aligned} \frac{c_2^{j+1} - c_1^{j+1}}{\Delta x} &= 0, \\ \frac{c_i^{j+1} - c_i^j}{\Delta t} &= D\frac{c_{i+1}^{j+1} - 2c_i^{j+1} + c_{i-1}^{j+1}}{\Delta x^2} + \lambda c_i^{j+1} f(c_i^{j+1}), \quad i = 2, 3, \ldots, N-1, \\ \frac{c_N^{j+1} - c_{N-1}^{j+1}}{\Delta x} &= 0, \end{aligned} \tag{7.35}$$

where $c_{i\pm1}^{j\pm1} \approx \mathcal{C}(x_i \pm \Delta x, t_j \pm \Delta t)$. The solution is stepped forward in time using fixed point iterations that are initialized by a first order forward Euler estimate.

The truncation error is estimated by

$$\epsilon = \frac{\Delta t}{2} \max_{1 \le i \le N} \left| \left(\frac{\mathrm{d}c}{\mathrm{d}t}\right)_i^{j+1} - \left(\frac{\mathrm{d}c}{\mathrm{d}t}\right)_i^{j} \right|, \text{ with } \left(\frac{\mathrm{d}c}{\mathrm{d}t}\right)_i^{j+1} \approx \frac{c_i^{j+1} - c_i^j}{\Delta t}.$$

After solving the nonlinear system (Equation (7.35)) using fixed point iteration, a new step size is determined by $\Delta t \leftarrow s\Delta t$, where $s = 0.9\sqrt{\tau/\epsilon}$ and τ is the truncation error tolerance (Simpson et al., 2007b; Sloan and Abbo, 1999). If $\epsilon \le \tau$, then the solution is accepted and the new Δt is used for the next time step. Alternatively, if $\epsilon > \tau$ then the solution is rejected and a new attempt is made using the new time step. This process is repeated until the solution advances to some desired time point T.

7.5.5 Additional results

We provide the multivariate posteriors for the results provided in the main text. In both Figure 7.7 and Figure 7.8 the contours of the posterior bivariate marginals generated, at a significant reduction in computational cost, both the PC-SMC-ABC and MM-SMC-ABC methods align very well with contours of the posterior bivariate marginals SMC-ABC using the expensive discrete model alone, however, the is a clear bias in the contours posterior bivariate marginals generated with SMC-ABC using the continuum-limit approximation alone.

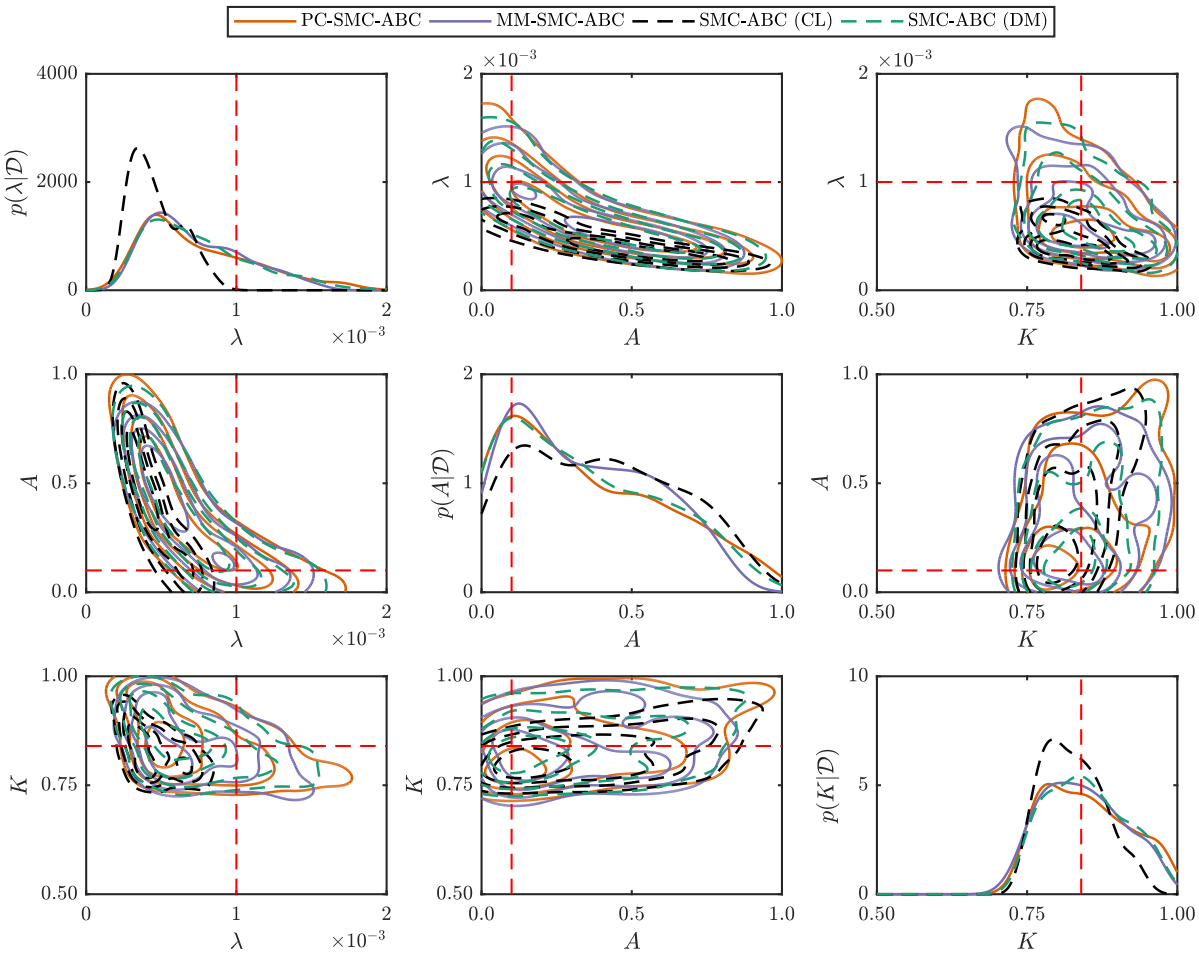

Figure 7.7: Comparison of estimated posterior marginal densities for the weak Allee model. There is a distinct bias in the SMC-ABC density estimate using the continuum limit (CL) (black dashed) compared with the SMC-ABC method with the discrete model (DM) (green dashed). However, the density estimates computed using the PC-SMC-ABC (orange solid) and MM-SMC-ABC (purple solid) methods match well with a reduced computational overhead.

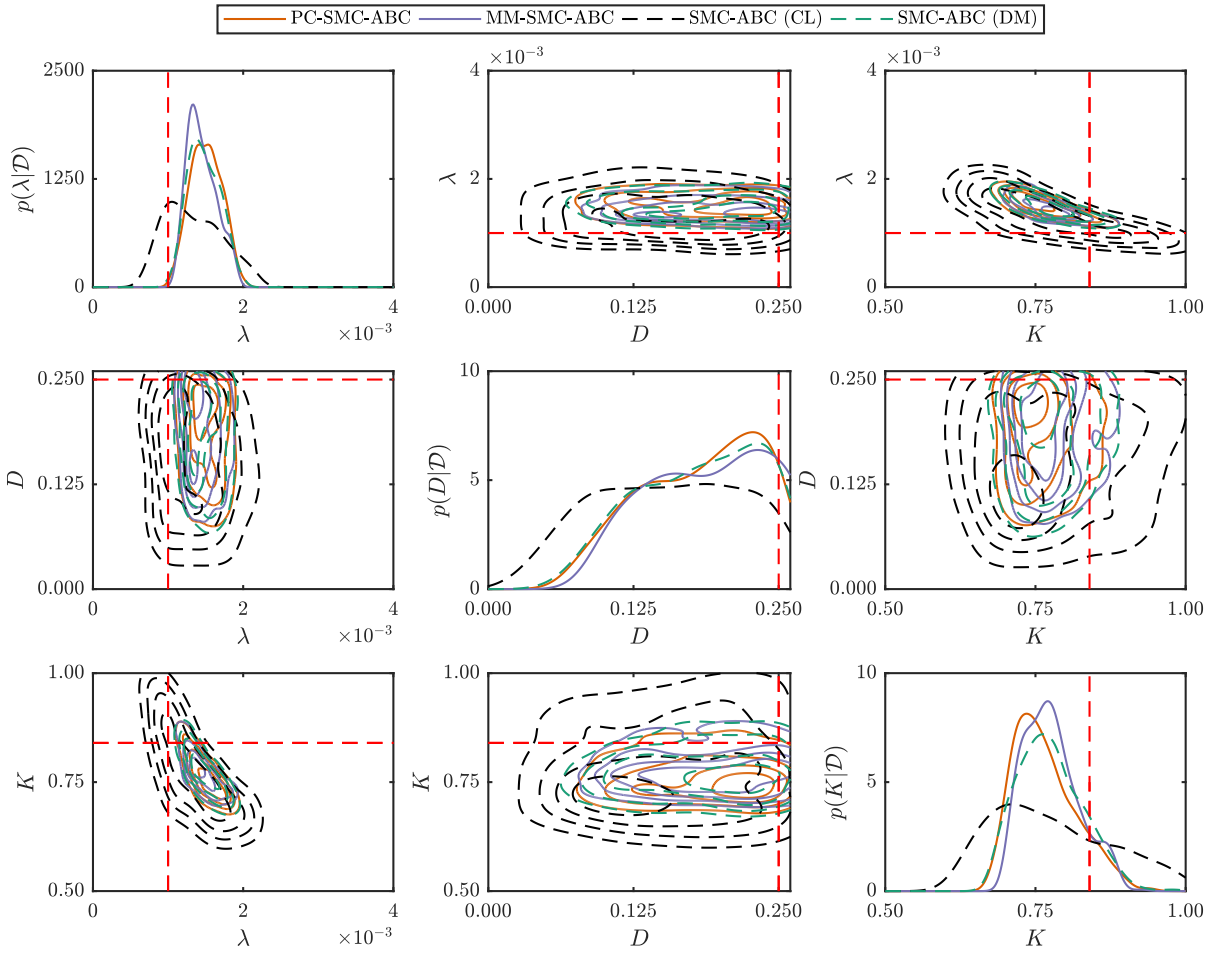

Figure 7.8: Comparison of estimated posterior marginal densities for the scratch assay model. There is a distinct bias in the SMC-ABC density estimate using the continuum limit (CL) (black dashed) compared with the SMC-ABC method with the discrete model (DM) (green dashed). However, the density estimates computed using the PC-SMC-ABC (orange solid) and MM-SMC-ABC (purple solid) methods match well with a reduced computational overhead.

7.5.6 Effect of motility rate on continuum-limit approximation

Here we demonstrate the ability (or lack thereof) of the continuum-limit approximation to capture the average behavior of the discrete model for the two examples considered in the main manuscript: the weak Allee model; and the scratch assay model. Figure 7.9 plots the solutions to the continuum-limit differential equation against realizations of the discrete model for both motile, $P_m = 1$, and non-motile, $P_m = 0$, agents.

In the case of the weak Allee model (Figure 7.9(A)), the continuum limit does not match the average behavior in either case, though it does perform better when $P_m = 1$ than when $P_m = 0$. The remaining discrepancy when $P_m = 1$ is can be related to the ratio P_p/P_m and to the effect of the neighborhood radius, r (see Jin et al. (2016a) for further explanation). Decreasing P_p/P_m and increasing r will correct this discrepancy. The scratch assay continuum limit captures the average behavior of the discrete model very well when $P_m = 1$ (Figure 7.9(B)), but does not capture the scratch closing behavior of the discrete model when $P_m = 0$ (Figure 7.9(C)).

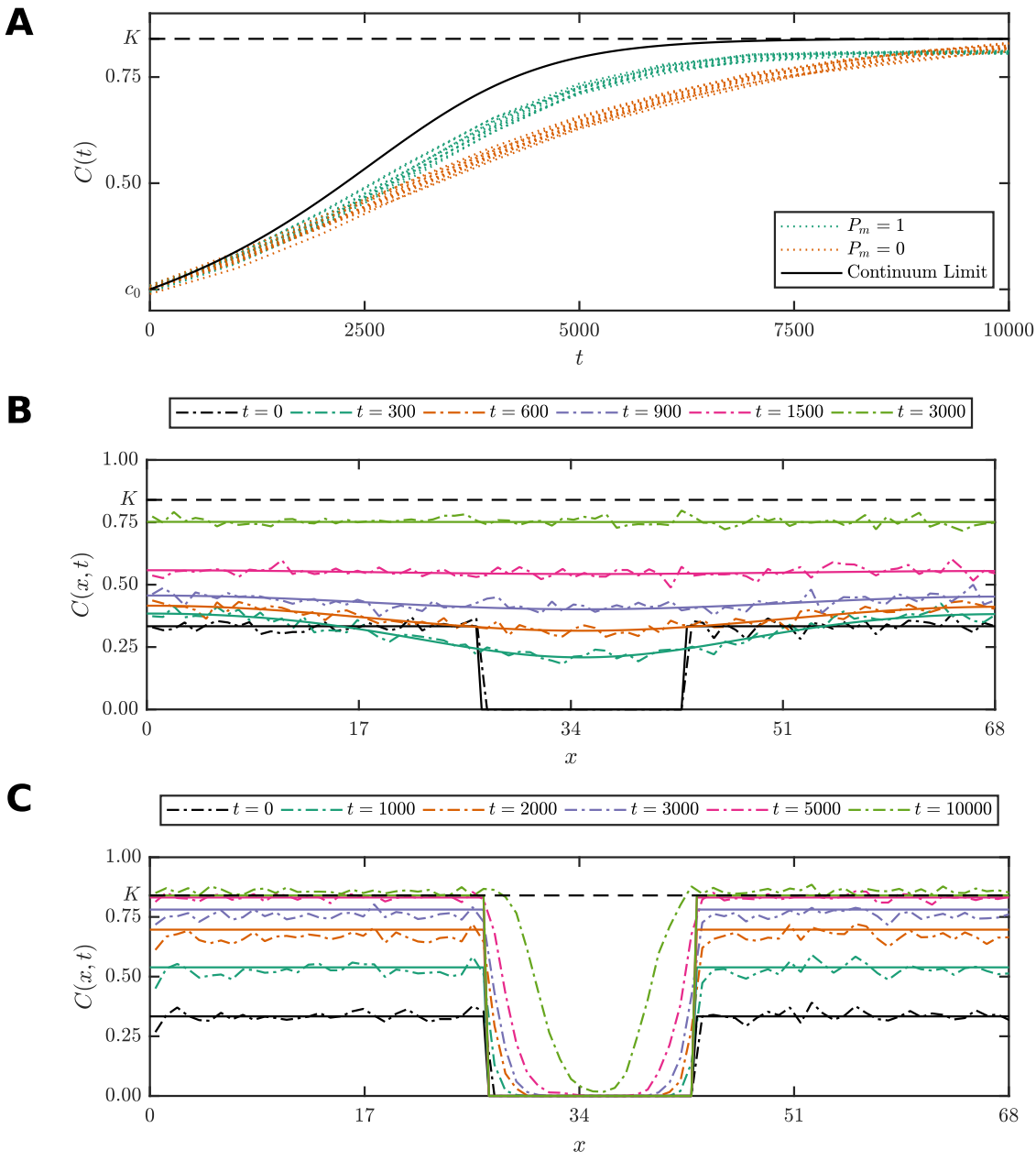

Figure 7.9: Continuum-limit approximations for the (A) weak Allee model and (B)-(C) the scratch assay model plotted against stochastic simulations of the discrete model. (A) For the weak Allee model, the continuum limit (black solid) does not capture the average behavior of the discrete model for motile agents, $P_m = 1$, (green dotted) or non-motile agents, $P_m = 0$, (orange dotted). (B)-(C) For the scratch assay model, the continuum limit (solid) is a very good match of the average behavior of the (B) discrete model (dot-dashed) for motile agents, $P_m = 1$, but a very poor approximation of the average behavior of the (C) discrete model (dot-dashed) for non-motile agents, $P_m = 0$, especially in the scratch region. Parameters used in the simulations are $P_p = 1/1000$, $\lambda = P_p/\tau$ $D = P_m\delta^2/4\tau$, $K = 5/6$, and $A = 1/10$. The stochastic simulations are performed on an $I \times J$ hexagonal lattice with $I = 80$, $J = 68$, $\tau = 1$ and $\delta = 1$.

7.5.7 Synthetic data

The full synthetic data used in the main manuscript for computational examples is provided in Table 7.3 for the Allee effect model and in Table 7.4 for the scratch assay model.

Table 7.3: Synthetic data used for the weak Allee effect model example.

t	0	$1,000$	$2,000$	$3,000$	$4,000$	$5,000$	$6,000$	$7,000$	$8,000$	$9,000$	$10,000$
$\bar{C}(t)$	0.27	0.33	0.41	0.50	0.57	0.63	0.70	0.74	0.77	0.80	0.82

Table 7.4: Synthetic data used for the cell culture assay model example.

	$\bar{C}(x,t)$	t_0	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}
i	x_i	0	300	600	900	1200	1500	1800	2100	2400	2700	3000
0	0.43	0.26	0.43	0.44	0.50	0.57	0.51	0.65	0.68	0.71	0.76	0.79
1	1.30	0.25	0.43	0.51	0.46	0.37	0.47	0.63	0.69	0.74	0.78	0.76
2	2.17	0.37	0.40	0.38	0.43	0.57	0.53	0.66	0.74	0.69	0.75	0.82
3	3.03	0.28	0.43	0.40	0.41	0.47	0.46	0.65	0.75	0.72	0.65	0.72
4	3.90	0.32	0.37	0.47	0.46	0.43	0.53	0.62	0.66	0.71	0.75	0.74
5	4.76	0.32	0.35	0.44	0.47	0.53	0.54	0.72	0.69	0.66	0.66	0.81
6	5.63	0.37	0.35	0.40	0.44	0.50	0.63	0.57	0.71	0.66	0.66	0.76
7	6.50	0.37	0.41	0.50	0.44	0.50	0.57	0.72	0.69	0.75	0.74	0.81
8	7.36	0.26	0.38	0.49	0.47	0.53	0.62	0.60	0.65	0.62	0.79	0.66
9	8.23	0.37	0.35	0.47	0.51	0.56	0.43	0.56	0.66	0.69	0.72	0.76
10	9.09	0.35	0.38	0.41	0.50	0.40	0.57	0.60	0.59	0.71	0.84	0.78
11	9.96	0.37	0.35	0.46	0.41	0.49	0.63	0.54	0.59	0.72	0.78	0.71
12	10.83	0.32	0.31	0.40	0.43	0.50	0.57	0.57	0.63	0.81	0.71	0.81
13	11.69	0.28	0.43	0.32	0.41	0.53	0.57	0.49	0.72	0.71	0.75	0.81
14	12.56	0.35	0.47	0.31	0.47	0.51	0.63	0.54	0.60	0.74	0.81	0.74
15	13.42	0.26	0.37	0.40	0.51	0.53	0.59	0.62	0.59	0.72	0.78	0.66
16	14.29	0.35	0.24	0.24	0.49	0.56	0.57	0.69	0.71	0.62	0.76	0.72
17	15.16	0.34	0.25	0.41	0.43	0.47	0.69	0.68	0.63	0.65	0.68	0.74
18	16.02	0.34	0.50	0.34	0.44	0.51	0.47	0.57	0.65	0.74	0.71	0.78
19	16.89	0.50	0.32	0.41	0.35	0.46	0.49	0.54	0.71	0.71	0.66	0.79
20	17.75	0.29	0.31	0.38	0.49	0.54	0.49	0.53	0.65	0.63	0.66	0.76
21	18.62	0.46	0.38	0.44	0.44	0.46	0.50	0.54	0.63	0.69	0.71	0.56
22	19.49	0.38	0.29	0.32	0.41	0.51	0.53	0.62	0.66	0.71	0.74	0.76
23	20.35	0.41	0.35	0.40	0.37	0.47	0.53	0.65	0.57	0.85	0.76	0.75
24	21.22	0.32	0.34	0.37	0.37	0.40	0.66	0.69	0.59	0.65	0.75	0.76
25	22.08	0.21	0.29	0.35	0.41	0.49	0.65	0.59	0.63	0.69	0.68	0.75
26	22.95	0.34	0.21	0.32	0.38	0.40	0.51	0.68	0.65	0.62	0.71	0.79
27	23.82	0.29	0.24	0.22	0.41	0.46	0.49	0.71	0.59	0.68	0.75	0.71
28	24.68	0.41	0.34	0.29	0.44	0.49	0.59	0.60	0.65	0.71	0.62	0.71
29	25.55	0.29	0.25	0.32	0.38	0.40	0.57	0.54	0.76	0.56	0.74	0.82
30	26.41	0.26	0.22	0.46	0.29	0.56	0.60	0.65	0.65	0.68	0.81	0.69
31	27.28	0.00	0.25	0.41	0.43	0.46	0.56	0.69	0.60	0.76	0.79	0.76
32	28.15	0.00	0.18	0.29	0.44	0.51	0.47	0.49	0.54	0.66	0.71	0.75
33	29.01	0.00	0.21	0.35	0.43	0.54	0.51	0.63	0.68	0.62	0.66	0.72
34	29.88	0.00	0.22	0.24	0.40	0.49	0.56	0.63	0.63	0.74	0.72	0.79
35	30.74	0.00	0.19	0.37	0.43	0.44	0.56	0.57	0.71	0.57	0.85	0.76
36	31.61	0.00	0.21	0.40	0.34	0.50	0.59	0.53	0.65	0.78	0.71	0.74
37	32.48	0.00	0.21	0.40	0.34	0.54	0.50	0.63	0.75	0.79	0.75	0.72
38	33.34	0.00	0.21	0.34	0.47	0.38	0.60	0.65	0.62	0.78	0.72	0.75
39	34.21	0.00	0.15	0.37	0.47	0.51	0.63	0.54	0.71	0.65	0.78	0.82
40	35.07	0.00	0.21	0.34	0.40	0.40	0.46	0.63	0.69	0.66	0.71	0.79
41	35.94	0.00	0.19	0.26	0.46	0.53	0.51	0.65	0.68	0.65	0.76	0.79
42	36.81	0.00	0.26	0.32	0.29	0.53	0.57	0.68	0.66	0.68	0.63	0.79
43	37.67	0.00	0.16	0.28	0.50	0.44	0.60	0.60	0.54	0.74	0.72	0.74
44	38.54	0.00	0.21	0.31	0.49	0.59	0.50	0.66	0.74	0.72	0.69	0.72
45	39.40	0.00	0.22	0.31	0.37	0.57	0.53	0.62	0.62	0.78	0.72	0.76
46	40.27	0.00	0.24	0.32	0.41	0.49	0.59	0.65	0.72	0.68	0.72	0.82
47	41.14	0.00	0.22	0.37	0.51	0.46	0.46	0.56	0.62	0.69	0.72	0.76
48	42.00	0.00	0.28	0.24	0.44	0.56	0.59	0.68	0.63	0.60	0.68	0.74
49	42.87	0.00	0.25	0.35	0.37	0.47	0.51	0.60	0.71	0.71	0.71	0.75
50	43.73	0.32	0.21	0.41	0.37	0.51	0.66	0.46	0.63	0.59	0.75	0.76
51	44.60	0.32	0.21	0.26	0.38	0.57	0.56	0.56	0.74	0.66	0.78	0.76
52	45.47	0.47	0.21	0.26	0.49	0.37	0.57	0.76	0.59	0.74	0.71	0.78
53	46.33	0.40	0.26	0.38	0.46	0.50	0.51	0.59	0.53	0.62	0.74	0.66
54	47.20	0.40	0.25	0.31	0.46	0.51	0.56	0.66	0.71	0.78	0.63	0.75
55	48.06	0.34	0.29	0.34	0.51	0.51	0.49	0.53	0.60	0.76	0.76	0.71
56	48.93	0.25	0.31	0.37	0.44	0.38	0.51	0.63	0.62	0.66	0.75	0.79
57	49.80	0.35	0.24	0.44	0.38	0.43	0.51	0.63	0.76	0.71	0.72	0.76
58	50.66	0.26	0.31	0.29	0.43	0.50	0.65	0.59	0.66	0.68	0.68	0.76
59	51.53	0.31	0.31	0.32	0.41	0.51	0.63	0.68	0.51	0.69	0.74	0.76
60	52.39	0.29	0.40	0.37	0.41	0.49	0.56	0.66	0.63	0.71	0.87	0.75
61	53.26	0.41	0.37	0.28	0.35	0.51	0.54	0.71	0.65	0.68	0.69	0.75
62	54.13	0.19	0.24	0.32	0.37	0.51	0.69	0.65	0.68	0.65	0.76	0.68
63	54.99	0.31	0.38	0.49	0.37	0.44	0.37	0.65	0.68	0.63	0.71	0.81
64	55.86	0.37	0.29	0.40	0.43	0.49	0.62	0.65	0.68	0.66	0.71	0.75
65	56.72	0.26	0.37	0.37	0.43	0.44	0.56	0.51	0.60	0.59	0.84	0.78
66	57.59	0.29	0.26	0.49	0.46	0.41	0.57	0.53	0.60	0.82	0.63	0.71
67	58.46	0.35	0.28	0.50	0.40	0.56	0.53	0.63	0.74	0.66	0.78	0.81
68	59.32	0.29	0.31	0.40	0.44	0.50	0.62	0.57	0.63	0.71	0.65	0.71
69	60.19	0.40	0.41	0.32	0.44	0.49	0.63	0.66	0.71	0.71	0.68	0.76
70	61.05	0.26	0.44	0.38	0.43	0.47	0.62	0.62	0.66	0.68	0.74	0.78
71	61.92	0.40	0.29	0.35	0.50	0.51	0.57	0.40	0.62	0.60	0.68	0.78
72	62.79	0.26	0.41	0.41	0.34	0.47	0.66	0.60	0.60	0.72	0.76	0.72
73	63.65	0.40	0.50	0.38	0.43	0.56	0.59	0.53	0.69	0.76	0.74	0.72
74	64.52	0.35	0.29	0.43	0.62	0.43	0.53	0.59	0.66	0.63	0.76	0.68
75	65.38	0.34	0.43	0.47	0.47	0.51	0.65	0.54	0.62	0.72	0.76	0.69
76	66.25	0.28	0.47	0.40	0.53	0.53	0.54	0.72	0.65	0.68	0.68	0.74
77	67.12	0.35	0.37	0.37	0.56	0.54	0.49	0.62	0.74	0.75	0.78	0.82
78	67.98	0.37	0.47	0.38	0.41	0.63	0.62	0.59	0.59	0.81	0.76	0.75
79	68.85	0.28	0.35	0.40	0.37	0.51	0.50	0.71	0.71	0.72	0.66	0.74

Chapter 8

Conclusion and Future Work

This chapter summarises and contextualise the contributions of the work in this thesis. We highlight the novel aspects of the work and discuss the relevance and significance of these contributions in the context of the current literature. We present new research opportunities that arise from the results of this research along with potential future work. The thesis concludes with final remarks.

8.1 Summary and contribution

This thesis explores the application of Bayesian methods for computational inference for the advancement of the biological sciences. In Chapters 2 and 3, we demonstrate the utility of the Bayesian approach as a powerful tool to inform the design of *in vitro* cell culture, and the selection of mathematical models that are optimal trade-offs between model complexity and fitness. The reviews in Chapters 4 and 5 discuss a range of challenges related to applications requiring expensive stochastic models, and provide practical guidelines to implement and tune state-of-the-art numerical methods for simulation and inference. Chapters 6 and 7 develop new computation schemes that significantly improve the computational efficiency of likelihood-free inference methods for applications involving expensive stochastic models.

Chapter 2 presents a new rigorous analysis of proliferation assay experimental protocols with

the goal of inferring the carrying capacity density, K, with minimal uncertainty. We demonstrate theoretical lower bounds on the uncertainty in K under the scenario of fixed observation numbers, Gaussian observation error, and unbounded experimental duration. This lower bound is independent of stochastic fluctuations in the initial condition. However, for realistic experiments with finite time, the duration of the experiment decreases the uncertainty more significantly than the number of observations, with optimal duration close to the inflection point of the logistic growth model. Furthermore, if experimental duration is fixed, then using multiple identically prepared replicates is more effective in reducing uncertainty in parameter estimates than multiple observations of the same assay at different times. Proliferation assays are a standard tool for the study of collective cell behaviour, however, many assay durations are used in practice with little or no justification; for example, Huang et al. (2017) use 4 hours, Chen et al. (2017) use 24 hours, and Jin et al. (2017) use 48 hours. Practitioners now have clear, evidence-based, guidelines for the duration and configuration of proliferation assays.

In Chapter 3, we apply Bayesian parameter inferences and information criteria to explore the efficacy of the Bayesian approach over the maximum likelihood approach. A topical case study, involving reaction–diffusion models for the analysis of scratch assay data, demonstrates that maximum likelihood estimates favor overly complex models (Box, 1976; Gelman et al., 2014; Johnson and Omland, 2004; Stoica and Selen, 2004). We advise caution when including additional mechanisms in a model that cannot be validated using the data. The work of this chapter provides evidence that the popular Fisher-KPP model (Fisher, 1937; Kolmogorov et al., 1937; Maini et al., 2004a,b; Johnston et al., 2015) is not appropriate for the description of PC-3 cells as suggested by (Jin et al., 2016b). Analysis also determines that the improved fitness of the Generalised Porous Fisher model (Cai et al., 2007; Jin et al., 2016b; King and McCabe, 2003; Sherratt and Murray, 1990; Simpson et al., 2011; Witelski, 1995) is not justified over the simpler Porous Fisher model (Gurney and Nisbet, 1975; Sengers et al., 2007). This is important and significant since reaction–diffusion models are routinely applied to compare hypotheses for various biological processes, such as wound healing (Bianchi et al., 2016; Flegg et al., 2009; Jin et al., 2016b), without proper consideration of effects of overparameterisation. Furthermore, many studies have applied Fisher-KPP, or Porous Fisher without clear justification for selecting one over the other (Sengers et al., 2007; Sherratt and Murray, 1990; Simpson et al., 2011), and this study provides a framework for determining this in a structured way. Lastly, to the best of our knowledge, this is the first time various information criteria, such as AIC (Akaike, 1974),

BIC (Schwarz, 1978) and DIC (Spiegelhalter et al., 2002), have been used in the mathematical biology literature related to collective cell behaviour.

Chapter 4 is a comprehensive and accessible review on stochastic simulation and computational inference algorithms for biochemical reaction networks (Erban et al., 2007; Schnoerr et al., 2017; Wilkinson, 2009). This review highlights many key algorithmic developments that have lead to the current state-of-the-art in stochastic simulation (Anderson, 2007; Gibson and Bruck, 2000; Gillespie, 1977, 2000; Voliotis et al., 2016), advanced Monte Carlo methods (Giles, 2008; Higham, 2015), and likelihood-free Bayesian inference methods (Marjoram et al., 2003; Pritchard et al., 1999; Sisson et al., 2007, 2018; Tavaré et al., 1997). The significance of this work is the drawing of connections between the forwards and inverse problems for biochemical reaction networks, and the accessible demonstration of advanced multilevel Monte Carlo methods for the computation of expectations in the context of the forwards problem (Anderson and Higham, 2012; Lester et al., 2015; Wilson and Baker, 2016) and the inverse problem (Guha and Tan, 2017; Jasra et al., 2019) (Chapter 6). The accessible review of these aspects of stochastic modelling in biology will enable researches in the life sciences to apply these techniques in practice. Furthermore, this review provides practical, user friendly example implementations of all presented algorithms using the MATLAB® programming language.

Chapter 5 introduces the applied mathematical biology community to the pseudo-marginal approach (Andrieu and Roberts, 2009) for likelihood-free Bayesian inference. While the approximate Bayesian computation approach is growing in popularity in the biological sciences (Browning et al., 2018; Lambert et al., 2018; Parker et al., 2018; Ross et al., 2017), the pseudo-marginal approach has an advantageous property of targeting the exact Bayesian posterior distribution that would be obtained with the unavailable exact likelihood (Andrieu and Roberts, 2009). The contribution in this chapter is the practical demonstration of several well-known concepts from the computational statistics literature. In particular, this work presents to the mathematical biology community, for the first time, important concepts in Bayesian statistics such as tuning of various Markov chain Monte Carlo samplers (Gelman et al., 1996; Geyer, 1992; Mengersen and Tweedie, 1996), unbiased likelihood estimators (Beaumont, 2003; Andrieu and Roberts, 2009), particle filters (Gordon et al., 1993; Doucet and Johanson, 2011), particle Markov chain Monte Carlo (Doucet et al., 2015; Golightly and Wilkinson, 2008, 2011), and convergence diagnostics (Gelman and Rubin, 1992; Gelman et al., 2014; Vehtari et al., 2019). The key

examples of the stochastically bistable Schlögl model (Schlögl, 1972; Vellela and Qian, 2009) and the oscillatory repressilator gene regulatory network model (Elowitz and Leibler, 2000) are very challenging to solve using approximate Bayesian computation (Toni et al., 2009), however, the particle Markov chain Monte Carlo method can solve these inference problems efficiently. Furthermore, this Chapter provides practical example implementations of all presented algorithms using the high performance, open source Julia programming language (Bezanson et al., 2017).

Chapter 6 represents the first presentation of multilevel methods to approximate Bayesian computation rejection sampling. Prior to this work, first posted on ArXiv in 2017 and published in 2018 (Warne et al., 2018) (Chapter 6), there were no examples of multilevel methods applied in an approximate Bayesian setting. The publication of this work is timely as there is a growing interest in multilevel and multifidelity approaches to computational inference (Dodwell et al., 2015, 2019; Efendiev et al., 2015; Gregory et al., 2016; Guha and Tan, 2017; Jasra et al., 2019; Lester, 2018; Prescott and Baker, 2018; Sisson et al., 2018), and there are many challenges that are unresolved. In particular, the most significant challenge is the development of coupling schemes to reduce the variance of the bias correction terms in the multilevel telescoping sum. While Jasra et al. (2019) consider our reliance on rejection sampling to be a limitation, they also state that our coupling scheme is a novel contribution to the field (Jasra et al., 2019). Other approaches to coupling problem include sequential importance sampling (Jasra et al., 2019) and a hierarchy of correlated Markov chains (Guha and Tan, 2017). Determining the optimal approach for a given inference problem is a new and unresolved problem. Another significant feature of the work of Chapter 6 is its generality; we make no assumptions about the underlying forwards problem and our method can be applied in any situation where standard approximate Bayesian schemes are employed (Pritchard et al., 1999; Marjoram et al., 2003; Sisson et al., 2007).

In Chapter 7, we develop new approximate Bayesian computation schemes for inference problems involving expensive stochastic models when a low cost approximate model with similar qualitative behaviour is available. This work contributes to an idea that is growing in popularity, that of using approximations for acceleration whilst attempting to maintain accurate inferences with respect to the exact stochastic model (Banterle et al., 2019; Everitt and Rowińska, 2017; Lester, 2018; Prescott and Baker, 2018). A successful methodology of this form is a significant

contribution to science, since the resolution of modern experimental techniques (Chen et al., 2014) is driving the need for more detailed computational models to analyse and characterise observed phenomena (Baker et al., 2018; Black and McKane, 2012; Coveney et al., 2016; Drawert et al., 2017). Our contribution includes an unbiased sequential Monte Carlo method that is up to four times more computationally efficient than modern adaptive schemes, and a method based on a moment-matching transform that is in excess of ten times more computationally efficient than modern adaptive schemes, but incurs a bias that can be minimised with tuning. These methods are also widely applicable to many classes of models and approximations used in systems biology (Cao and Grima, 2018), cell biology (Browning et al., 2019), climate science (Holden et al., 2018), astronomy (The Event Horizon Telescope Collaboration et al., 2019), and engineering (Cotter et al., 2010).

8.2 Future work

The developments and findings of this thesis identify a number of new research questions and avenues for further investigation. We present future work that is specific to particular chapters along with new ideas emerging from the thesis as a whole.

The chapters of Part I apply specific forms of continuum models for applications of mathematical modelling to investigate collective cell behaviour. In particular, we assume the logistic growth model for cell proliferation in both Chapters 2 and 3. However, our approach is more general than these specific models. For example, the analysis in Chapters 2 and 3 is applicable to more general proliferation functions of the form

$$S(\mathbf{x}, t) = \lambda C(\mathbf{x}, t) f(C(\mathbf{x}, t)),$$

where λ is the proliferation rate, $C(\mathbf{x}, t)$ is the cell density at position $\mathbf{x}$ and time t, and $f(C(\mathbf{x}, t))$ is a crowding function as defined in Browning et al. (2017); Jin et al. (2016a); Simpson et al. (2010) and Chapter 7 with $f(0) = 1$ and $f(K) = 0$. In this more general case, some modifications to the analysis in Chapter 2 are necessary if the resulting growth model does not yield an analytical solution, however, many of the results can still be acquired numerically.

The recommendations of Chapter 2 also assume that the proliferation rate, λ, is known precisely.

In reality, there will be some uncertainty associated with this estimate. Furthermore, Chapter 3 identifies that a trade-off may exist between the optimal estimation of λ and K. Can an *in vitro* cell culture assay be designed to optimally minimise uncertainty in both λ and K together? This is an interesting and relevant question that requires more detailed analysis since the posterior distribution will be bivariate and potentially more complex.

Chapter 3 highlights that, while complex models may provide insight into possible mechanisms that drive certain biological processes, often it is challenging to validate these models, particularly when the commonly used maximum likelihood estimator favors complexity and overparameterisation. The analysis in Chapter 3 provides a framework for determining when additional complexity is justified from the data. An interesting extension is to consider, from an optimal experimental design perspective, the experimental protocol necessary to reliably distinguish the complex model from the simpler model. Investigating this problem requires extensions to the methods used in Chapter 2 and more sophisticated ideas from experiment design literature could provide solutions (Browning et al., 2017; Drovandi and Pettitt, 2013; Liepe et al., 2013; Ryan et al., 2016; Silk et al., 2014).

Part II of this thesis also reveals a number of unresolved computational challenges in likelihood-free Bayesian inference, and new avenues for future developments. Particularly in the area of combining multiple models (Chapter 7) (Prescott and Baker, 2018) or multiple inference approximation levels (Chapter 6) (Lester, 2018; Jasra et al., 2019) to reduce computational costs while maintaining accuracy. This line of research is very much in its infancy, but the potential to revolutionise computational inference is significant.

The reviews of the current literature from Chapters 4 and 5 highlight a number of recommendations for optimal algorithm design choices for many standard and state-of-the-art inference methods. For example, the are many guidelines for optimal selection of proposal kernels (Beaumont et al., 2009; Filippi et al., 2013) and target distribution sequences (Drovandi and Pettitt, 2011; Silk et al., 2013) in sequential Monte Carlo sampling, and for proposal kernels (Gelman et al., 1996; Roberts and Rosenthal, 2001, 2004) and convergence diagnostics (Gelman and Rubin, 1992; Geyer, 1992; Vehtari et al., 2019) in Markov chain Monte Carlo sampling. However, similar ideas have not yet been explored in the multilevel Monte Carlo context (Chapter 6), or when approximate models are applied (Chapter 7). Chapter 6 applies acceptance threshold sequences reported in the literature for sequential Monte Carlo benchmark models (Sisson

et al., 2007; Tanaka et al., 2006; Toni et al., 2009), however, these may be far from optimal for multilevel Monte Carlo. Furthermore, Chapter 7 applies the adaptive proposal schemes of Beaumont et al. (2009) and Filippi et al. (2013), but alternative asymmetric proposals may be more optimal (Cotter et al., 2013).

In the multilevel Monte Carlo literature, there is an emphasis on proving the multilevel estimator converges in mean-square at a faster asymptotic rate than direct Monte Carlo. As noted by Jasra et al. (2019), the novel coupling scheme in Chapter 6 is non-trivial to analyse. Determining the conditions in which our multilevel estimator will be superior to direct Monte Carlo is an open problem. There is some theory in the asymptotic behaviour of approximate Bayesian computation (Beskos et al., 2017; Fearnhead and Prangle, 2012), however, the usual multilevel condition to ensure the appropriate strong convergence error rate is unclear for our scheme. Discovering these key criteria will enable our multilevel scheme to be confidently applied in practice.

The multifidelity framework of Prescott and Baker (2018) is based on rejection sampling, however, Chapter 7 presents new algorithms based on sequential Monte Carlo sampling. There are connections between the methods of Prescott and Baker (2018) and our approaches, and a promising area of future research is the application of the multifidelity weighting scheme within the sequential Monte Carlo framework from Chapter 7. Similar ideas are also being explored in the context of delayed acceptance methods (Banterle et al., 2019; Everitt and Rowińska, 2017; Golightly et al., 2015), however, the multifidelity approach is more general than delayed acceptance. While extending multifidelity methods to be applicable to sequential Monte Carlo sampling presents challenges, the potential computational benefit of such a scheme is significant.

8.3 Final remarks

Mathematical modelling is an essential part of science, and increasingly so in the life sciences. In practical applications, experimental design, model selection, model simulation, model calibration, and parameter inference all require a variety of design elements to be prescribed by the practitioner to ensure results and conclusions are reliable. Many of these decisions need to balance trade-offs between accuracy and computational costs. The work in this thesis

explores a range of these trade-offs, from the design of experiments and model selection to the tuning of computational methods for simulation and inference. The context of the work is biological applications with a focus on models and algorithms commonly used in cell biology, biochemistry and systems biology. However, the methods of this thesis are widely applicable since the tools, techniques, and algorithms developed here do not rely on features specific to models in biology. The thesis outcomes provide new insights into acceleration of computational inference methods and open new avenues for future developments in data analysis for the study of complex biological processes.

Bibliography

Abercrombie, M., 1970. Contact inhibition in tissue culture. *In Vitro* 6:128–142. DOI:10.1007/BF02616114

Abercrombie, M., 1979. Contact inhibition and malignancy. *Nature* 281:259–262. DOI:10.1038/281259a0

Abkowitz, J.L., Catlin, S.N., Guttorp, P., 1996. Evidence that hematopoiesis may be a stochastic process *in vivo*. *Nature Medicine* 2:190–197. DOI:10.1038/nm0296-190

Agnew, D.J.G., Green, J.E.F., Brown, T.M., Simpson, M.J., Binder, B.J., 2014. Distinguishing between mechanisms of cell aggregation using pair-correlation functions. *Journal of Theoretical Biology* 352:16–23. DOI:10.1016/j.jtbi.2014.02.033

Akaike, H., 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19:716–723. DOI:10.1109/TAC.1974.1100705

Anderson, D.F., 2007. A modified next reaction method for simulating chemical systems with time dependent propensities and delays. *The Journal of Chemical Physics* 127:214107. DOI:10.1063/1.2799998

Anderson, D.F., Ganguly, A., Kurtz, T.G., 2011. Error analysis of tau-leap simulation methods. *Annals of Applied Probability* 21:2226–2262. DOI:10.1214/10-AAP756

Anderson, D.F., Higham, D.J., 2012. Multilevel Monte Carlo for continuous time Markov chains, with applications in biochemical kinetics. *Multiscale Modeling and Simulation* 10:146–179. DOI:10.1137/110840546

Andrieu, C., Doucet, A., Holenstein, R., 2010. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 72:269–342. DOI:10.1111/j.1467-9868.2009.00736.x

Andrieu, C., Lee, A., Vihola, M., 2018. Theoretical and methodological aspects of MCMC computations with noisy likelihoods. In *Handbook of Approximate Bayesian Computation*, Sisson, S.A., Fan, Y., and Beaumont, M.A., (Eds.), 1st edn. Chapman & Hall/CRC Press.

Andrieu, C., Roberts, G.O., 2009. The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics* 37:697–725, 2009. DOI:10.1214/07-AOS574

Ang, A., Tang, W.H., 2007. *Probability Concepts in Engineering*. Wiley, New York.

Arkin, A., Ross, J., McAdams, H.H., 1998. Stochastic kinetic analysis of developmental pathway bifurcation in phage l-infected *Escherichia Coli* cells. *Genetics* 149:1633–1648,.

Armstrong, N.J., Painter, K.J., Sherratt, J.A., 2009. Adding adhesion to a chemical signaling model for somite formation. *Bulletin of Mathematical Biology* 71:1–24. DOI:10.1007/s11538-008-9350-1

Avikainen, R., 2009. On irregular functionals of SDEs and the Euler scheme. *Finance and Stochastics* 13:381–401. DOI:10.1007/s00780-009-0099-7

Bajar, B.T., Wang, E.S., Zhang, S., Lin, M.Z., Chu, J., 2016. A guide to fluorescent protein FRET pairs. *Sensors* 16:1488, 2016. DOI:10.3390/s16091488

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* 14:20170660. DOI:10.1098/rsbl.2017.0660

Banterle, M., Grazian, C., Lee, A., Robert, C.P., 2019. Accelerating Metropolis-Hastings algorithms by delayed acceptance. *Foundations of Data Science* 1:103–128. DOI:10.3934/fods.2019005

Barber, S., Voss, J., Webster, M., 2015. The rate of convergence for approximate Bayesian computation. *Electronic Journal of Statistics* 9:80–105. DOI:10.1214/15-EJS988

Barenblatt, G.I., 2003. *Scaling*. Cambridge University Press, Cambridge, UK.

Bar-Joseph, Z., Gitter, A., Simon, I., 2012. Studying and modelling dynamic biological processes using time-series gene expression data. *Nature Reviews Genetics* 13:552–562. DOI:10.1038/nrg3244

Barnes, C.P., Filippi, S., Stumpf, M.P.H., Thorne, T., 2012. Considerate approaches to constructing summary statistics for ABC model selection. *Statistics and Computing* 22:1181–1197. DOI:10.1007/s11222-012-9335-7

Beaumont, K.A., Mohana-Kumaran, N., Haass, N.K., 2014. Modeling melanoma *in vitro* and *in vivo*. *Healthcare (Basel)* 2:27–46. DOI:10.3390/healthcare2010027

Beaumont, M.A., Zhang, W., Balding, D.J., 2002. Approximate Bayesian computation in population genetics. *Genetics* 162:2025–2035.

Beaumont, M.A., 2003. Estimation of population growth or decline in genetically monitored populations. *Genetics* 164:1139–1160.

Beaumont, M.A., Cornuet, J.M., Marin, J.M., Robert, C.P., 2009. Adaptive approximate Bayesian computation. *Biometrika* 96:983–990.

Berger, J., 2006. The case for objective Bayesian analysis. *Bayesian Analysis* 1:385–402. DOI:10.1214/06-BA115

Besançon, M., Anthoff, D., Arslan, A., Byrne, S., Lin, D., Papamarkou, T., Pearson, J., 2019. Distributions.jl: Definition and Modeling of Probability Distributions in the JuliaStats Ecosystem. *arXiv e-prints*, `arXiv:1907.08611 [stat.CO]`

Beskos, A., Jasra, A., Law, K., Tempone, R., Zhou, Y., 2017. Multilevel sequential Monte Carlo samplers. *Stochastic Processes and their Applications* 127:1417–1440. DOI:10.1016/j.spa.2016.08.004

Bezanson, J., Edelman, A., Karpinski, S., Shah, V.B., 2017. Julia: A fresh approach to numerical computing. *SIAM Review* 59:65–98. DOI:10.1137/141000671

Bianchi, A., Painter, K.J., Sherratt, J.A., 2016. Spatio-temporal models of lymphangiogenesis in wound healing. *Bulletin of Mathematical Biology* 78:1904–1941. DOI:10.1007/s11538-016-0205-x

Bierkens, J., Fearnhead, P., Roberts, G., 2019. The Zig-Zag process and super-efficient sampling for Bayesian analysis of big data *The Annals of Statistics* 47:1288-1320.DOI:10.1214/18-AOS1715

Binny, R.N., Haridas, P., James, A., Law, R., Simpson, M.J., Plank, M.J., 2016. Spatial structure arising from neighbour-dependent bias in collective cell movement. *PeerJ* 4:e1689. DOI:10.7717/peerj.1689

Black, A.J., McKane, A.J., 2012. Stochastic formulation of ecological models and their applications. *Trends in Ecology & Evolution* 27:337–345. DOI:10.1016/j.tree.2012.01.014

Blum, M.G.B., 2010. Approximate Bayesian computation: A nonparametric perspective. *Journal of the American Statistical Association* 105:1178–1187. DOI:10.1198/jasa.2010.tm09448

Blum, M.G.B., Nunes, M.A., Prangle, D., Sisson, S.A., 2013. A comparative review of dimension reduction methods in approximate Bayesian computation. *Statistical Science* 28:189–208. DOI:10.1214/12-STS406

Box, G.E.P., 1976. Science and statistics. *Journal of the American Statistical Association* 71:791–799. DOI:10.1080/01621459.1976.10480949

Bressloff, P.C., 2017. Stochastic switching in biology: from genotype to phenotype. *Journal of Physics A: Mathematical and Theoretical* 50:133001. DOI:10.1088/1751-8121/aa5db4

Browning, A.P., McCue, S.W., Simpson, M.J., 2017. A Bayesian computational approach to explore the optimal duration of a cell proliferation assay. *Bulletin of Mathematical Biology* 79:1888–1906. DOI:10.1007/s11538-017-0311-4

Browning, A.P., Haridas, P., Simpson, M.J., 2019. A Bayesian sequential learning framework to parameterise continuum models of melanoma invasion into human skin. *Bulletin of Mathematical Biology* 81:676-698. DOI:10.1007/s11538-018-0532-1.

Browning, A.P., McCue, S.W., Binny, R.N., Plank, M.J., Shah, E.T., Simpson, M.J., 2018. Inferring parameters for a lattice-free model of cell migration and proliferation using experimental data. *Journal of Theoretical Biology* 437:251–260. DOI:10.1016/j.jtbi.2017.10.032

Böttger, K., Hatzikirou, H., Voss-Böhme, A., Cavalcanti-Adam, E.A., Herrero, M.A., Deutsch, A., 2015. An emerging Allee effect is critical for tumor initiation and persistence. *PLOS Computational Biology* 11:e1004366. DOI:10.1371/journal.pcbi.1004366

Botha, I., Kohn, R., Drovandi, C., 2019. Particle methods for stochastic differential equation mixed effects models. *arXiv e-prints*, `arXiv:1907.11017 [stat.CO]`

Burrage, K., Burrage, P.M., Tian, T., 2004. Numerical methods for strong solutions of stochastic differential equations: an overview. *Proceedings of the Royal Society of London Series A: Mathematical, Physical and Engineering Sciences* 460:373–402. DOI:10.1098/rspa.2003.1247

Buzbas, E.O., Rosenberg, N.A., 2015. AABC: Approximate approximate Bayesian computation for inference in population-genetic models. *Theoretical Population Biology* 99:31–42. DOI:10.1016/j.tpb.2014.09.002

Cabras, S., Castellanos Neuda, M.E., Ruli, E., 2015. Approximate Bayesian computation by modelling summary statistics in a quasi-likelihood framework. *Bayesian Analysis* 10:411–439. DOI:10.1214/14-BA921

Cai, A.Q., Landman, K.A., Hughes, B.D., 2007. Multi-scale modeling of a wound-healing cell migration assay. *Journal of Theoretical Biology* 245:576–594. DOI:10.1016/j.jtbi.2006.10.024

Callaghan, T., Khain, E., Sander, L.M., Ziff, R.M., 2006. A stochastic model for wound healing. *Journal of Statistical Physics* 122:909–924. DOI:10.1007/s10955-006-9022-1

Cao, Z., Grima, R., 2018. Linear mapping approximations of gene regulatory networks with stochastic dynamics. *Nature Communications* 9:3305. DOI:10.1038/s41467-018-05822-0

Cao, Y., Li, H., Petzold, L.R., 2004. Efficient formulation of the stochastic simulation algorithm for chemically reacting systems. *The Journal of Chemical Physics* 121:4059–4067. DOI:10.1063/1.1778376

Cao, Y., Gillespie, D.T., Petzold, L.R., 2005a. Avoiding negative populations in explicit Poisson tau-leaping. *The Journal of Chemical Physics* 123:054104. DOI:10.1063/1.1992473

Cao, Y., Gillespie, D.T., Petzold, L.R., 2005b. The slow-scale stochastic simulation algorithm. *The Journal of Chemical Physics* 122:014116. DOI:10.1063/1.1824902

Cao, Y., Gillespie, D.T., Petzold, L.R., 2006. Efficient step size selection for the tau-leaping simulation method. *The Journal of Chemical Physics* 124:044109. DOI:10.1063/1.2159468

Carpenter, J., Clifford, P., Fearnhead, P., 1999. Improved particle filter form nonlinear problems. *IEEE Pproceedings – Radar, Sonar and Navigation* 146:2–7. DOI:10.1049/ip-rsn:19990255

Chen, B.-C., Legant, W.R., Wang, K., Shao, L., Milkie, D.E., Davidson, M.W., Janetopoulos, C., Wu, X.S., Hammer, J.A., Liu, Z., English, B.P., Mimori-Kiyosue, Y., Romero, D.P., Ritter, A.T., Lippincott-Schwartz, J., Fritz-Laylin, L., Mullins, R.D., Mitchell, D.M., Bembenek, J.N., Reymann, A.-C., Böhme, R., Grill, S.W., Wang, J.T., Seydoux, G., Tulu, U.S., Kiekart, D.P., Betzig, E., 2014. Lattice light sheet microscopy: Imaging molecules to embryos at high spatiotemporal resolution. *Science* 346:1257998. DOI:10.1126/science.1257998

Chen, C.L.P., Zhang, C.-Y., 2014. Data-intensive applications, challenges, techniques and technologies: a survey on big data. *Information Sciences* 275:314–347. DOI:10.1016/j.ins.2014.01.015

Chen, K., Liu, Q., Tsang, L.L., Ye, Q., Chan, H.C., Sun, Y., Jiang, X., 2017. Human MSCs promotes colorectal cancer epithelial-mesenchymal transition and progression via CCL5/-catenin/Slug pathway. *Cell Death and Disease* 8:e2819. DOI:10.1038/cddis.2017.138

Chkrebtii, O.A., Cameron, E.K., Campbell, D.A., Bayne, E.M., 2015. Transdimensional approximate Bayesian computation for inference on invasive species models with latent variables of unknown dimension. *Computational Statistics & Data Analysis* 86:97–110. DOI:10.1016/j.csda.2015.01.002

Clyde, M., George, E.I., 2004. Model uncertainty. *Statistical Science* 19:81–94. DOI:10.1214/088342304000000035

Codling, E.A., Plank, M.J., Benhamou, S., 2008. Random walk models in biology. *Journal of The Royal Society Interface* 5:813–834. DOI:10.1098/rsif.2008.0014

Cohen, Y., Galiano, G., 2013. Evolutionary distributions and competition by way of reaction-diffusion and by way of convolution. *Bulletin of Mathematical Biology* 75:2305–2323. DOI:10.1007/s11538-013-9890-x

Cohen, M., Kicheva, A., Ribeiro, A., Blassberg, R., Page, K.M., Barnes, C.P., Briscoe, J., 2015. Ptch1 and Gli regulate Shh signalling dynamics via multiple mechanisms. *Nature Communications* 6:6709. DOI:10.1038/ncomms7709

Consonni, G., Fouskakis, D., Liseo, B., Ntzoufras, I. 2018. Prior distributions for objective Bayesian analysis. *Bayesian Analysis* 13:627–679. DOI:10.1214/18-BA1103

Cotter, S.L., Dashti, M., Stuart, A.M., 2010. Approximation of Bayesian inverse problems for PDEs. *SIAM Journal on Numerical Analysis* 48:322–345. DOI:10.1137/090770734

Cotter, S.L., Roberts, G.O., Stuart, A.M., White, D., 2013. MCMC methods for functions: modifying old algorithms to make them faster. *Statistical Science* 28:424–446. DOI:10.1214/13-STS421

Cotter, S.L., Erban, R., 2016. Error analysis of diffusion approximation methods for multiscale systems in reaction kinetics. *SIAM Journal on Scientific Computing* 38:B144–B163. DOI:10.1137/14100052X

Coveney, P.V., Dougherty, E.R., Highfield, R.R., 2016. Big data need big theory too. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374:20160153. DOI:10.1098/rsta.2016.0153

Cowles, M.K., Carlin, B.P., 1996. Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association* 91:883–904. DOI:10.2307/2291683

Crank, J., 1975. *The Mathematics of Diffusion*. Oxford University Press, Oxford, UK.

Csilléry, K., Blum, M.G.B., Gaggiotti, O.E., François, O., 2010. Approximate Bayesian computation (ABC) in practice. *Trends in Ecology & Evolution* 25:410–418. DOI:10.1016/j.tree.2010.04.001

DeAngelis, D.L., Grimm, V., 2014. Individual-based models in ecology after four decades. *F1000Prime Reports* 6:39. DOI:10.12703/P6-39

Delarue, M., Montel, F., Vignjevic, D., Prost, J., Joanny, J.-F., Cappello, G., 2014. Compressive stress inhibits proliferation in tumor spheroids through a volume limitation. *Biophysical Journal* 107:1821–1828. DOI:0.1016/j.bpj.2014.08.031

Del Moral, P., Doucet, A., Jasra, A., 2006. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68:411–436. DOI:10.1111/j.1467-9868.2006.00553.x

Del Moral, P., Doucet, A., Jasra, A., 2012. An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Statistics and Computing* 22:1009–1020. DOI:10.1007/s11222-011-9271-y

Dodwell, T.J., Ketelsen, C., Scheichl, R., Teckentrup, A.L., 2015. A hierarchical multilevel Markov chain Monte Carlo algorithm with applications to uncertainty quantification in subsurface flow. *SIAM/ASA Journal on Uncertainty Quantification* 3:1075–1108. DOI:10.1137/130915005

Dodwell, T.J., Ketelsen, C., Scheichl, R., Teckentrup, A.L., 2019. Multilevel Markov chain Monte Carlo. *SIAM Review* 61:509–545.DOI:10.1137/19M126966X

Doucet, A., Johansen, A., 2011. A tutorial on particle filtering and smoothing: fifteen years later. *The Oxford Handbook of Nonlinear Filtering*, Oxford University Press, New York, 656-704.

Doucet, A., Pitt, M.K., Deligiannidis, G., Kohn, R., 2015. Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika* 102:295–313

Drawert, B., Griesemer, M., Petzold, L.R., Briggs, C.J., 2017. Using stochastic epidemiological models to evaluate conservation strategies for endangered amphibians. *Journal of the Royal Society Interface* 14:20170480. DOI:10.1098/rsif.2017.0480

Drovandi, C.C., Pettitt, A.N., 2011. Estimation of parameters for macroparasite population evolution using approximate Bayesian computation. *Biometrics* 67:225–233. DOI:10.1111/j.1541-0420.2010.01410.x

Drovandi, C.C., Pettitt, A.N., 2013. Bayesian experimental design for models with intractable likelihoods. *Biometrics* 69:937–948. DOI:10.1111/biom.12081

Duane, S., Kennedy, A.D., Pendleton, B.J., Roweth, D., 1987. Hybrid Monte Carlo. *Physics Letters B* 195:216222. DOI:10.1016/0370-2693(87)91197-X

E, W., Liu, D., Vanden-Eijnden, E., 2005. Nested stochastic simulation algorithm for chemical kinetic systems with disparate rates. *The Journal of Chemical Physics* 123:194107. DOI:10.1063/1.2109987

Edelstein-Keshet, L., 2005. *Mathematical Models in Biology*, 6th edn. SIAM, Philadelphia.

Efendiev, Y., Jin, B., Michael, P., Tan, X., 2015. Multilevel Markov chain Monte Carlo method for high-contrast single-phase flow problems. *Communications in Computational Physics* 17:259–286. DOI:10.4208/cicp.021013.260614a

Efron, B., 1979. Bootstrap methods: another look at the jackknife. *The Annals of Statistics* 7:1–26. DOI:10.1214/aos/1176344552

Efron, B., 1986. Why isn't everyone a Bayesian? *The American Statistician* 40:1–5. DOI:10.1080/00031305.1986.10475342

Ellison, A.M., 2004. Bayesian inference in ecology. *Ecology Letters* 7:509–520. DOI:10.1111/j.1461-0248.2004.00603.x

Eldar, A., Elowitz, M.B., 2010. Functional roles for noise in genetic circuits. *Nature* 467:167–173. DOI:10.1038/nature09326

Elowitz, M.B., Leibler, S., 2000. A synthetic oscillatory network of transcriptional regulators. *Nature* 403:335–338. DOI:10.1038/35002125

Elowitz, M.B., Levine, A.J., Siggia, E.D., Swain, P.S., 2002. Stochastic gene expression in a single cell. *Science* 297:1183–1186. DOI:10.1126/science.1070919

Enderling, H., Chaplain, M.A.J., 2014. Mathematical modeling of tumor growth and treatment. *Current Pharmaceutical Design* 20:4934–4940. DOI:10.2174/1381612819666131125150434

Erban, R., Chapman, S.J., Maini, P.K., 2007. A practical guide to stochastic simulation of reaction-diffusion processes. *ArXiv e-prints*, `arXiv:0704.1908 [q-bio.SC]`

Erban, R., Chapman, S.J., 2009. Stochastic modelling of reactiondiffusion processes: algorithms for bimolecular reactions. *Physical Biology* 6:046001. DOI:10.1088/1478-3975/6/4/046001

Everitt, R.G., Rowińska, P.A., 2017. Delayed acceptance ABC-SMC. *ArXiv e-prints*, arXiv:1708.02230 [stat.CO]

Fange, D., Berg, O.G., Sjöberg, P., Elf, J., 2010. Stochastic reaction-diffusion kinetics in the microscopic limit. *Proceedings of the National Academy of Sciences* 107:19820–19825. DOI:10.1073/pnas.1006565107

Fearnhead, P., Prangle, D., 2012. Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society Series B (Statistical Methodology)* 74:419–474. DOI:10.1111/j.1467-9868.2011.01010.x

Fedoroff, N., Fontana, W., 2002. Small numbers of big molecules. *Science* 297:1129–1131. DOI:10.1126/science.1075988

Fehlberg, E., 1969. Low-order classical Runge-Kutta formulas with step size control and their application to some heat transfer problems. *NASA Technical Report* R-315.

Feinberg, A.P., 2014. Epigenetic stochasticity, nuclear structure and cancer: the implications for medicine. *Journal of Internal Medicine* 276:5–11. DOI:10.1111/joim.12224

Filippi, S., Barnes, C.P., Cornebise, J., Stumpf, M.P.H., 2013. On optimality of kernels for approximate Bayesian computation using sequential Monte Carlo. *Statistical Applications in Genetics and Molecular Biology* 12:87–107. DOI:10.1515/sagmb-2012-0069

Finkenstädt, B., Heron, E.A., Komorowski, M., Edwards, K., Tang, S., Harper, C.V., Julian, 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* 24:2901–2907. DOI:10.1093/bioinformatics/btn562

Fisher, R.A., 1937. The wave of advance of advantageous genes. *Annals of Eugenics* 7:355–369. DOI:10.1111/j.1469-1809.1937.tb02153.x

Flegg, J.A., Byrne, H.M., McElwain, D.L.S., 2010. Mathematical model of hyperbaric oxygen therapy applied to chronic diabetic wounds. *Bulletin of Mathematical Biology* 72:1867–1891. DOI:10.1007/s11538-010-9514-7

Flegg, M.B., Hellander, S., Erban, R., 2015. Convergence of methods for coupling of microscopic and mesoscopic reactiondiffusion simulations. *Journal of Computational Physics* 289:1–17. DOI:10.1016/j.jcp.2015.01.030

Flegg, J.A., McElwain, D.L.S., Byrne, H.M., Turner, I.W., 2009. A three species model to simulate application of hyperbaric oxygen therapy to chronic wounds. *PLOS Computational Biology* 5:e1000451. DOI:10.1371/journal.pcbi.1000451

Fortelius, M., Geritz, S., Gyllenberg, M., Toivonen, J., 2015. Adaptive dynamics on an environmental gradient that changes over a geological time-scale. *Journal of Theoretical Biology* 376:91–104. DOI:10.1016/j.jtbi.2015.03.036

Gadgil, C., Lee, C.H., Othmer, H.G., 2005. A stochastic analysis of first-order reaction networks. *Bulletin of Mathematical Biology* 67:901–946. DOI:10.1016/j.bulm.2004.09.009

Gao, N.P., Ud-Dean, S.M.M, Gandrillon, O., Gunawan, R., 2018. SINCERITIES: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinformatics* 34:258–266. DOI:10.1093/bioinformatics/btx575

Gelman, A., 2008a. Objections to Bayesian statistics. *Bayesian Analysis* 3:445–450. DOI:10.1214/08-BA318

Gelman, A., 2008b. Rejoinder. *Bayesian Analysis* 3:467–478. DOI:10.1214/08-BA318REJ

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2014. *Bayesian Data Analysis*, 3rd edn. Chapman & Hall/CRC.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2004. *Bayesian Data Analysis*, 2nd edn. Chapman & Hall/CRC.

Gelman, A., Roberts, G.O., Gilks, W.R., 1996. Efficient Metropolis jumping rules. *Bayesian Statistics* 5:599–607.

Gelman, A., Rubin, D.B., 1992. Inference from iterative simulation using multiple sequences. *Statistical Sciences* 7:457–472.

Geman, S., Geman, D., 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6:721741. DOI:10.1109/TPAMI.1984.4767596

Gerlee, P., 2013. The model muddle: In search of tumor growth laws. *Cancer Research* 73:2407–2411. DOI:10.1158/0008-5472.CAN-12-4355

Geyer, C.J., 1992. Practical Markov Chain Monte Carlo. *Statistical Science* 7:473–483. DOI:10.1214/ss/1177011137

Gibson, M.A., Bruck, J., 2000. Efficient exact stochastic simulation of chemical systems with many species and many channels. *The Journal of Physical Chemistry* 104:1876–1889. DOI:10.1021/jp993732q

Giles, M.B., 2008. Multilevel Monte Carlo path simulation. *Operations Research* 56:607–617. DOI:10.1287/opre.1070.0496

Giles, M.B., 2015. Multilevel Monte Carlo methods. *Acta Numerica* 24:259–328. DOI:10.1017/S096249291500001X

Giles, M.B., Nagapetyan, T., Ritter, K., 2015. Multilevel Monte Carlo approximation of cumulative distribution function and probability densities. *SIAM/ASA Journal on Uncertainty Quantification* 3:267–295. DOI:10.1137/140960086

Gillespie, D.T., 1977. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry* 81:2340–2361. DOI:10.1021/j100540a008

Gillespie, D.T., 1992. A rigorous derivation of the chemical master equation. *Physica A* 188:404–425. DOI:10.1016/0378-4371(92)90283-V

Gillespie, D.T., 2000. The chemical Langevin equation. *The Journal of Chemical Physics* 113:297–306. DOI:10.1063/1.481811

Gillespie, D.T., 2001. Approximate accelerated simulation of chemically reacting systems. *The Journal of Chemical Physics* 115:1716–1733. DOI:10.1063/1.1378322

Gillespie, D.T., Hellander, A., Petzold, L.R., 2013. Perspective: Stochastic algorithms for chemical kinetics. *The Journal of Chemical Physics* 138:170901. DOI:10.1063/1.4801941

Gillespie, D.T., Petzold, L.R., Seitaridou, E., 2014. Validity conditions for stochastic chemical kinetics in diffusion-limited system. *The Journal of Chemical Physics* 140:054111. DOI:10.1063/1.4863990

Glynn, P.W., Iglehart, D.L., 1989. Importance sampling for stochastic simulations. *Management Science* 35:1367–1392. DOI:10.1287/mnsc.35.11.1367

Golightly, A., Henderson, D.A., Sherlock, C., 2015. Delayed acceptance particle MCMC for exact inference in stochastic kinetic models. *Statistics and Computing* 25:1039–1055. DOI:10.1007/s11222-014-9469-x

Golightly, A., Wilkinson, D.J., 2008. Bayesian inference for nonlinear multivariate diffusion models observed with error. *Computational Statistics & Data Analysis* 52:1674–1693. DOI:10.1016/j.csda.2007.05.019

Golightly, A., Wilkinson, D.J., 2011. Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo. *Interface Focus* 1:807–820. DOI:10.1098/rsfs.2011.0047

Gordon, N.J., Salmond, D.J, and Smith, A.F.M., 1993. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F - Radar and Signal Processing* 140:107113. DOI:10.1049/ip-f-2.1993.0015

Green, P.J., Łatuszyński, K., Pereyra, M., Robert, C.P., 2015. Bayesian computation: a summary of the current state, and samples backwards and forwards. *Statistics and Computing* 25:835–862. DOI:10.1007/s11222-015-9574-5

Gregory, A., Cotter, C.J., Reich, S., 2016. Multilevel ensemble transform particle filtering. *SIAM Journal on Scientific Computing* 38:A1317–A1338. DOI:10.1137/15M1038232

Grelaud, A., Marin, J.-M., Robert, C.P., Rodolphe, F., Tallay, J.-F., 2009. ABC likelihood-free methods for model choice in Gibbs random fields. *Bayesian Analysis* 4:317–335. DOI:10.1214/09-BA412

Guha, N., Tan, X., 2017. Multilevel approximate Bayesian approaches for flows in highly heterogeneous porous media and their applications. *Journal of Computational and Applied Mathematics* 317:700–717. DOI:10.1016/j.cam.2016.10.008

Guindani, M., Sepúlveda, N., Paulino, C.D., Müller, P., 2014. A Bayesian semiparametric approach for the differential analysis of sequence counts data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 63:385–404. DOI:10.1111/rssc.12041

Gunawan, R., Cao, Y., Petzold, L., Doyle III, F.J., 2005. Sensitivity analysis of discrete stochastic systems. *Biophysical Journal* 88:2530–2540. DOI:10.1529/biophysj.104.053405

Gupta, P., Fillmore, C.M., Jiang, G., Shapira, S.D., Tao, K., Kuperwasser, C., Lander, E.S., 2011. Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells. *Cell* 146:633–644. DOI:10.1016/j.cell.2011.07.026

Gurney, W., Nisbet, R., 1975. The regulation of inhomogeneous populations. *Journal of Theoretical Biology* 52:441–457. DOI:10.1016/0022-5193(75)90011-9

Haridas, P., McGovern, J.A., McElwain, D.L.S., Simpson, M.J., 2017. Quantitative comparison of the spreading and invasion of radial growth phase and metastatic melanoma cells in a three-dimensional human skin equivalent model. *PeerJ* 5:e3754. DOI:10.7717/peerj.3754

Harris, S., 2004. Fisher equation with density-dependent diffusion: Special solutions. *Journal of Physics A: Mathematical and General* 37:6267. DOI:10.1088/0305-4470/37/24/005

Hass, N.K., Beaumont, K.A., Hill, D.S., Anfosso, A., Mrass, P., Munoz, M.A., Kinjyo, I., Weninger, W., 2014. Realtime cell cycle imaging during melanoma growth, invasion, and drug response. *Pigment Cell and Melanoma Research* 27:764–776. DOI:10.1111/pcmr.12274

Hastings, W.K., 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57:97–109.

Hepburn, I., Chen, W., Wils, S., De Schutter, E., 2012. STEPS: efficient simulation of stochastic reaction–diffusion models in realistic morphologies. *BMC Systems Biology* 6:36. DOI:10.1186/1752-0509-6-36

Higham, D.J., 2001. An algorithmic introduction to numerical simulation of stochastic differential equations. *SIAM Review* 43:525–546. DOI:10.1137/S0036144500378302

Higham, D.J., 2008. Modeling and simulating chemical reactions. *SIAM Review* 50:347–368. DOI:10.1137/060666457

Higham, D.J., 2015. An introduction to multilevel Monte Carlo for option valuation. *International Journal of Computer Mathematics* 82:2347–2360. DOI:10.1080/00207160.2015.1077236

Hines, K.E., Middendorf, T.R., Aldrich, R.W., 2014. Determination of parameter identifiability in nonlinear biophysical models: A Bayesian approach. *The Journal of General Physiology* 143:401–416. DOI:10.1085/jgp.201311116

Holden, P.B., Edwards, N.R., Ridgwell, A., Wilkinson, R.D., Fraedrich, K., Lunkeit, F., Pollit, H., Mercure, J.F., Salas, P., Lam, A., Knobloch, F., Chewpreecha, U., Viñuales, J.E., 2018. Climatecarbon cycle uncertainties and the Paris Agreement. *Nature Climate Change* 8:609–613. DOI:10.1038/s41558-018-0197-7

Hoops, S., Sahle, S., Gauges, R., Lee, C., Pahle, J., Simus, N., Singhal, M., Xu, L., Mendes, P., Kummer, U., 2006. COPASI–a COmplex PAthway SImulator. *Bioinformatics* 22:3067–3074. DOI:10.1093/bioinformatics/btl485

Huang, F.-T., Chen, W.-F., Gu, Z.-Q., Zhuang, Y.-Y., Li, C.-Q., Wang, L.-Y., Peng, J.-F., Zhu, Z., Luo, X., Li, Y.-H., Yao, H.-R., Zhang, S.-N., 2017. The novel long intergenic noncoding RNA UCC promotes colorectal cancer progression by sponging miR-143. *Cell Death and Disease* 8:e2778. DOI:10.1038/cddis.2017.191

Iafolla, M.A.J., Mazumder, M., Sardana, V., Velauthapillai, T., Pannu, K., McMillen, D.R., 2008. Dark proteins: Effect of inclusion body formation on quantification of protein expression. *Proteins: Structure, Function, and Bioinformatics* 72:1233–1242. DOI:10.1002/prot.22024

Indurkhya, S., Beal, J., 2010. Reaction factoring and bipartite update graphs accelerate the Gillespie algorithm for large-scale biochemical systems. *PLOS One* 5:e8125. DOI:10.1371/journal.pone.0008125

Isaacson, S.A., 2008. Relationship between the reaction–diffusion master equation and particle tracking models. *Journal of Physics A: Mathematical and Theoretical* 41:065003. DOI:10.1088/1751-8113/41/6/065003

Iserles, A., 2008. *A First Course in the Numerical Analysis of Differential Equations*, 2nd edn. Cambridge University Press.

Jackson, P.R., Juliano, J., Hawkins-Daarud, A., Rockne, R.C., Swanson, K.R., 2015. Patient-specific mathematical neuro-oncology: Using a simple proliferation and invasion

tumor model to inform clinical practice. *Bulletin of Mathematical Biology* 77:846–856. DOI:10.1007/s11538-015-0067-7

Jahnke, T., Huisinga, W., 2007. Solving the chemical master equation for monomolecular reaction systems analytically. *Journal of Mathematical Biology* 54:1–26. DOI:10.1007/s00285-006-0034-x

Jasra, A., Jo, S., Nott, D., Shoemaker, C., Tempone, R., 2019. Multilevel Monte Carlo in approximate Bayesian computation. *Stochastic Analysis and Applications*, DOI:10.1080/07362994.2019.1566006

Jasra, A., Kamatani, K., Law, K., Zhou, Y., 2017. Multilevel particle filters. *SIAM Journal on Numerical Analysis* 55:3068–3096. DOI:10.1137/17M1111553

Jasra, A., Kamatani, K., Law, K., Zhou, Y., 2018. Bayesian static parameter estimation for partially observed diffusions via multilevel Monte Carlo *SIAM Journal on Scientific Computing* 40:A887–A902. DOI:10.1137/17M1112595

Jayaraman, P., Devarajan, K., Chua, T.K., Zhang, H., Gunawan, E., Poh, C.L., 2016. Blue light-mediated transcriptional activation and repression of gene expression in bacteria. *Nucleic Acids Research* 44:6994–7005. DOI:10.1093/nar/gkw548

Jin, W., Penington, C.J., McCue, S.W., Simpson, M.J. 2016a. Stochastic simulation tools and continuum models for describing two-dimensional collective cell spreading with universal growth functions. *Physical Biology* 13:056003. DOI:10.1088/1478-3975/13/5/056003

Jin, W., Shah, E.T., Penington, C.J., McCue, S.W., Chopin, L.K., Simpson, M.J., 2016b. Reproducibility of scratch assays is affected by the initial degree of confluence: Experiments, modelling and model selection. *Journal of Theoretical Biology* 390:136–145. DOI:10.1016/j.jtbi.2015.10.040

Jin, W., Shah, E.T., Penington, C.J., McCue, S.W., Maini, P.K., Simpson, M.J., 2017. Logistic proliferation of cells in scratch assays is delayed. *Bulletin of Mathematical Biology* 79:1028–1050. DOI:10.1007/s11538-017-0267-4

Johnson, J.B., Omland, K.S., 2004. Model selection in ecology and evolution. *Trends in Ecology & Evolution* 19:101–108. DOI:10.1016/j.tree.2003.10.013

Johnston, S.T., Baker, R.E., McElwain, D.L.S., Simpson, M.J. 2017. Co-operation, competition and crowding: a discrete framework linking Allee kinetics, nonlinear diffusion, shocks and sharp-fronted travelling waves. *Scientific Reports* 7:42134. DOI:10.1038/srep42134

Johnston, S.T., Ross, J.V., Binder, B.J., McElwain, D.L.S., Haridas, P., Simpson, M.J., 2016. Quantifying the effect of experimental design choices for in vitro scratch assays. *Journal of Theoretical Biology* 400:19–31. DOI:10.1016/j.jtbi.2016.04.012

Johnston, S.T., Shah, E.T., Chopin, L.K., McElwain, D.L.S., Simpson, M.J., 2015. Estimating cell diffusivity and cell proliferation rate by interpreting IncuCyte ZOOM™ assay data using the Fisher–Kolmogorov model. *BMC Systems Biology* 9:38. DOI:10.1186/s12918-015-0182-y

Johnston, S.T., Simpson, M.J., McElwain, D.L.S., Binder, B.J., Ross, J.V., 2014 Interpreting scratch assays using pair density dynamics and approximate Bayesian computation. *Open Biology* 4:140097. DOI:10.1098/rsob.140097

Johnston, S.T., Simpson, M.J., Plank, M.J., 2013. Lattice-free descriptions of collective motion with crowding and adhesion. *Physical Review E* 88:062720. DOI:10.1103/PhysRevE.88.062720

Kærn, M., Elston, T.C., Blake, W.J., Collins, J.J., 2005. Stochasticity in gene expression: from theories to phenotypes. *Nature Reviews Genetics* 9:451–464. DOI:10.1038/nrg1615

Kass, R.E., Wasserman, L., 1996. The selection of prior distributions by formal rules. *Journal of the American Statistical Association* 91:1343–1370. DOI:10.2307/2291752

Keren, L., van Dijk, D., Weingarten-Gabbay, S., David, D., Jona, G., Weinberger, A., Milo, R., Segal, E., 2015. Noise in gene expression is coupled to growth rate. *Genome Research* 25:1893–1902. DOI:10.1101/gr.191635.115

King, T.E., Fortes, G.G., Balaresque, P., Thomas, M.G., Balding, D., Delser, P.M., Neumann, R., Parson, W., Knapp, M., Walsh, S., Tonasso, L., Holt, J., Kayser, M., Appleby, J., Forster, P., Ekserdjian, D., Hofreiter, M., Schürer, K., 2014. Identification of the remains of King Richard III. *Nature Communications* 5:5631. DOI:10.1038/ncomms6631

King, J.R., McCabe, P.M., 2003. On the Fisher–KPP equation with fast nonlinear diffusion. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* 459:2529–2546. DOI:10.1098/rspa.2003.1134

Kitagawa, G., 1996. Monte Carlo filter and smoother for non-Gaussian nonlinear stats space models. *Journal of Computational and Graphical Statistics* 5:1–15. DOI:10.2307/1390750

Kloeden, P.E., Platen, E., 1999. *Numerical Solution of Stochastic Differential Equations*, 3rd edn. Springer, New York.

Kolmogorov, A., Petrovsky, N., Piscounov, S., 1937. Etude de l'quations de la Diffusion avec Croissance de la Quantite de Matiere et Son Application a un Probleme Biologique. *Moscow University Mathematics Bulletin* 1:1–25.

Kramer, N., Walzl, A., Unger, C., Rosner, M., Krupitza, G., Hengstschläger, M., Dolznig, H., 2013. In vitro cell migration and invasion assays. *Mutation Research/Reviews in Mutation Research*, 752:10–24. DOI:10.1016/j.mrrev.2012.08.001

Kullback, S., Leibler, R.A., 1951. On information and sufficiency. *The Annals of Mathematical Statistics* 22:79–86. DOI:10.1214/aoms/1177729694

Kursawe, J., Baker, R.E., Fletcher, A.G., 2018. Approximate Bayesian computation reveals the importance of repeated measurements for parameterising cell-based models of growing tissues. *Journal of Theoretical Biology* 443:66–81. DOI:10.1016/j.jtbi.2018.01.020

Kurtz, T.G., 1972. The relationship between stochastic and deterministic models for chemical reactions. *The Journal of Chemical Physics* 57:2976–2978. DOI:10.1063/1.1678692

Lambert, B., 2018. *A Student's Guide to Bayesian Statistics*, 1st edn. Sage Publications.

Lambert, B., MacLean, A.L., Fletcher, A.G., Combes, A.N., Little, M.H., Byrne, H.M., 2018. Bayesian inference of agent-based models: A tool for studying kidney branching morphogenesis. *Journal of Mathematical Biology* 76:1673–1697. DOI:10.1007/s00285-018-1208-z

Landman, K.A., Cai, A.Q., Hughes, B.D., 2007. Travelling waves of attached and detached cells in a wound-healing cell migration assay. *Bulletin of Mathematical Biology* 69:2119–2138. DOI:10.1007/s11538-007-9206-0

Law, R., Murrell, D.J., Dieckmann, U., 2003. Population growth in space and time: spatial logistic equations. *Ecology* 84:252–262. DOI:10.1890/0012-9658(2003)084[0252:PGISAT]2.0.CO;2

Lawson, B.A.J., Drovandi, C.C., Cusimano, N., Burrage, P., Rodriguez, B., Burrage, K., 2018. Unlocking data sets by calibrating populations of models to data density: A study in atrial electrophysiology. *Science Advances* 4:e1701676. DOI:10.1126/sciadv.1701676

Lee, A., Yau, C., Giles, M.B., Doucet, A., Holmes, C.C., 2010. On the utility of graphics cards to perform massively parallel simulation of advanced Monte Carlo methods. *Journal of Computational and Graphical Statistics* 19:769–789. DOI:10.1198/jcgs.2010.10039

Lei, J., Bickel, P., 2011. A moment matching ensemble filter for nonlinear non-Gaussian data assimilation. *Monthly Weather Review* 139:3964–3973. DOI:10.1175/2011MWR3553.1

Lele, S.R., Dennis, B., Lutscher, F., 2007. Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. *Ecology Letters* 10:551–563. DOI:10.1111/j.1461-0248.2007.01047.x

Lester, C., 2018. Multi-level approximate Bayesian computation. *ArXiv e-prints*, `arXiv:1811.08866 [q-bio.QM]`

Lester, C., Yates, C.A., Baker, R.E., 2017. Efficient parameter sensitivity computation for spatially extended reaction networks. *The Journal of Chemical Physics* 146:044106. DOI:10.1063/1.4973219

Lester, C., Baker, R.E., Giles, M.B., Yates, C.A., 2016. Extending the multi-level method for the simulation of stochastic biological systems. *Bulletin of Mathematical Biology* 78:1640–1677. DOI:10.1007/s11538-016-0178-9

Lester, C., Yates, C.A., Giles, M.B., Baker, R.E., 2015. An adaptive multi-level simulation algorithm for stochastic biological systems. *The Journal of Chemical Physics* 142:024113. DOI:10.1063/1.4904980

Leung, B.O., Chou, K.C., 2011. Review of super-resolution fluorescent microscopy for biology *Applied Spectroscopy* 65:967–980. DOI:10.1366/11-06398

Levine, E.M., Becker, Y., Boone, C.W., Eagle, H., 1965. Contact inhibition, macromolecular synthesis, and polyribosomes in cultured human diploid fibroblasts. *Proceedings of the National Academy of Sciences* 53:350–356. DOI:0.1073/pnas.53.2.350

Li, T., 2007. Analysis of explicit tau-leaping schemes for simulating chemically reacting systems. *Multiscale Modeling and Simulation* 6:417–436. DOI:10.1137/06066792X

Li, H., Cao, Y., Petzold, L.R., Gillespie, D.T., 2008. Algorithms and software for stochastic simulation of biochemical reacting systems. *Biotechnology Progress* 24:56–61. DOI:10.1021/bp070255h

Li, D., Clements, A., Drovandi, C., 2019. Efficient Bayesian estimation for GARCH-type models via sequential Monte Carlo. *arXiv e-prints*, `arXiv:1906.03828 [stat.Ap]`

Liang, C.C., Park, A., Guan, J.L., 2007. In vitro scratch assay: A convenient and inexpensive method for analysis of cell migration in vitro. *Nature Protocols* 2:329–333. DOI:10.1038/nprot.2007.30

Liao, S., Vejchodský, T., Erban, R., 2015. Tensor methods for parameter estimation and bifurcation analysis of stochastic reaction networks. *Journal of the Royal Society Interface* 12:20150233. DOI:10.1098/rsif.2015.0233

Liepe, J., Filippi, S., Komorowski, M., Stumpf, M.P.H., 2013. Maximizing the information content of experiments in systems biology. *PLOS Computational Biology* 9:e1002888. DOI:10.1371/journal.pcbi.1002888

Liepe, J., Kirk, P., Filippi, S., Toni, T., Barnes, C.P., Stumpf, M.P.H., 2014. A framework for parameter estimation and model selection from experimental data in systems biology using approximate Bayesian computation. *Nature Protocols* 9:439–456. DOI:10.1038/nprot.2014.025

Lillacci, G., Khammash, M., 2010. Parameter estimation and model selection in computational biology. *PLOS Computational Biology* 6:e1000696. DOI:10.1371/journal.pcbi.1000696

Link, W.A., Eaton, M.J., 2011. On thinning of chains in MCMC. *Methods in Ecology and Evolution* 3:112–115. DOI:10.1111/j.2041-210X.2011.00131.x

Liu, Q.-H., Ajelli, M., Aleta, A., Merler, S., Moreno, Y., Vespignani, A., 2018. Measurability of the epidemic reproduction number in data-driven contact networks. *The Proceedings of the National Academy of Sciences* 115:12680–12685. DOI:10.1073/pnas.1811115115

Liu, L., Sun, B., Pedersen, J.N., Yong, K.M.A., Getzenberg, R.H., Stone, H.A., Austin, R.H., 2011. Probing the invasiveness of prostate cancer cells in a 3D microfabricated landscape. *Proceedings of the National Academy of Sciences* 108:6853–6856. DOI:10.1073/pnas.1102808108

Locke, J.C.W., Elowitz, M.B., 2009. Using movies to analyse gene circuit dynamics in single cells. *Nature Reviews Microbiology* 7:383–392. DOI:10.1038/nrmicro2056

Maarleveld, T.R., Olivier, B.G., Bruggeman, F.J., 2013. StochPy: A comprehensive, user-friendly tool for simulating stochastic biological processes. *PLOS One* 8:e79345. DOI:10.1371/journal.pone.0079345

MacEachern, S.N., Berliner, L.M., 1994. Subsampling the Gibbs Sampler. *The American Statistician* 48:188–190. DOI:10.1080/00031305.1994.10476054

Maclaren, O.J., Byrne, H.M., Fletcher, A.G., Maini, P.K., 2015. Models, measurement and inference in epithelial tissue dynamics. *Mathematical Biosciences and Engineering* 12:1321–1340. DOI:10.3934/mbe.2015.12.1321

Maclaren, O.J., Parker, A., Pin, C., Carding, S.R., Watson, A.J., Fletcher, A.G., Byrne, H.M., Maini, P.K., 2017. A hierarchical Bayesian model for understanding the spatiotemporal dynamics of the intestinal epithelium. *PLOS Computational Biology* 13:e1005688. DOI:10.1371/journal.pcbi.1005688

Maini, P., McElwain, D.L.S., Leavesley, D.I., 2004a. Travelling waves in a wound healing assay. *Applied Mathematics Letters* 17:575–580. DOI:10.1016/S0893-9659(04)90128-0

Maini, P., McElwain, D.L.S., Leavesley, D.I., 2004b. Traveling wave model to interpret a wound-healing cell migration assay for human peritoneal mesothelial cells. *Tissue Engineering* 10:475–482. DOI:10.1089/107632704323061834

Malaspinas, A.-S., Westaway, M.C., Muller, C., Sousa, V.C., Lao, O., Alves, I., Bergström, A., Athanasiadis, G., Cheng, J.Y., Crawford, J.E., et al., 2016. A genomic history of Aboriginal Australia. *Nature* 538:207–214. DOI:10.1038/nature18299

Marchant, B.P., Norbury, J., Sherratt, J.A., 2001. Travelling wave solutions to a haptotaxis-dominated model of malignant invasion. *Nonlinearity* 14:1653–1671. DOI:10.1088/0951-7715/14/6/313

Marchetti, L., Priami, C., Thanh, V.H., 2016. HRSSA Efficient hybrid stochastic simulation for spatially homogeneous biochemical reaction network. *Journal of Computational Physics* 317:301–317. DOI:10.1016/j.jcp.2016.04.056

Marino, S., Hogue, I.B., Ray, C.J., Kirschner, D.E., 2008. A methodology for performing global uncertainty and sensitivity analysis in systems biology. *Journal of Theoretical Biology* 254:178–196. DOI:10.1016/j.jtbi.2008.04.011

Marjoram, P., Molitor, J., Plagnol, V., Tavaré, S., 2003. Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences* 100:15324–15328. DOI:10.1073/pnas.0306899100

Marin, J.-M., Pudlo, P., Robert, C.P., Ryder, R.J., 2012. Approximate Bayesian computational methods. *Statistics and Computing* 22:1167–1180. DOI:10.1007/s11222-011-9288-2

Maruyama, G., 1955. Continuous Markov processes and stochastic equations. *Rendiconti del Circolo Matematico di Palermo* 4:48–90. DOI:10.1007/BF02846028

Matsiaka, O.M., Baker, R.E., Shah, E.T., Simpson, M.J., 2019. Mechanistic and experimental models of cell migration reveal the importance of intercellular interactions in cell invasion. *Biomedical Physics and Engineering Express* 5:045009. DOI:10.1088/2057-1976/ab1b01

McAdams, H.H., Arkin, A., 1997. Stochastic mechanisms in gene expression. *Proceedings of the National Academy of Sciences* 94:814–819. DOI:10.1073/pnas.94.3.814

McLane, A.J., Semeniuk, C., McDermid, G.J., Marceau, D.J., 2011. The role of agent-based models in wildlife ecology and management. *Ecology Modelling* 222:1544–1556. DOI:10.1016/j.ecolmodel.2011.01.020

Mengersen, K.L., Tweedie, R.L., 1996. Rates of convergence of the Hastings and Metropolis algorithms. *The Annals of Statistics* 24:101–121.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E., 1953. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics* 21:1087–1092. DOI:10.1063/1.1699114

Michaelis, L., Menten, M.L., 1913. Die kinetik der invertinwirkung. *Biochem Z* 49:333–369.

Moraes, A., Tempone, R., Vilanova, P., 2016. A multilevel adaptive reaction-splitting simulation method for stochastic reaction networks. *SIAM Journal on Scientific Computing* 38:A2091–A2117. DOI:10.1137/140972081

Moss, F., Pei, X., 1995. Neurons in parallel. *Nature* 376:211–212. DOI:10.1038/376211a0

Murray, J.D., 2002. *Mathematical Biology: I. An Introduction*. Springer, New York.

Nardini, J.T., Chapnick, D.A., Liu, X., Bortz, D.M., 2016. Modeling keratinocyte wound healing dynamics: cell-cell adhesion promotes sustained collective migration. *Journal of Theoretical Biology* 400:103–117. DOI:10.1016/j.jtbi.2016.04.015

Navascués, M., Leblois, R., Burgarella, C., 2017. Demographic inference through approximate-Bayesian-computation skyline plots. *PeerJ* 5:e3530. DOI:10.7717/peerj.3530

Nott, D.J., Fan, Y., Marshall, L., Sisson, S.A., 2014. Approximate Bayesian computation and Bayes' linear analysis: toward high-dimensional ABC. *Journal of Computational and Graphical Statistics* 23:65–86. DOI:10.1080/10618600.2012.751874

Nunes, M.A., Prangle, D., 2015. abctools: An R package for tuning approximate Bayesian computation analysis. *The R Journal* 7:189–205.

Ocone, A., Haghverdi, L., Mueller, N.S., Theis, F.J., 2015. Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data. *Bioinformatics* 31:i89–i96. DOI:10.1093/bioinformatics/btv257

O'Dea, A., Lessios, H.A., Coates, A.G., Eytan, R.I., Restrepo-Moreno, S.A., Cione, A.L., Collins, L.S., de Queiroz, A., Farris, D.W., Norris, R.D., et al., 2016. Formation of the Isthmus of Panama. *Science Advances* 2:e1600883. DOI:10.1126/sciadv.1600883

Parker, A., Simpson, M.J., Baker, R.E., 2018. The impact of experimental design choices on parameter inference for models of growing cell colonies. *Royal Society Open Science* 5:180384. DOI:10.1098/rsos.180384

Parno, M.D., Marzouk, Y.M., 2018. Transport map accelerated Markov chain Monte Carlo. *SIAM/ASA Journal on Uncertainty Quantification* 6:645–682. DOI:10.1137/17M1134640

Paulsson, J., Berg, O.G., Ehrenberg, M., 2000. Stochastic focusing: Fluctuation-enhanced sensitivity of intracellular regulation *Proceedings of the National Academy of Sciences* 97:7148–7153. DOI:10.1073/pnas.110057697

Picchini, U., Anderson, R., 2017. Approximate maximum likelihood estimation using data-cloning ABC. *Computational Statistics & Data Analysis* 105:166–183. DOI:10.1016/j.csda.2016.08.006

Pokhilko, A., Fernández, A.P., Edwards, K.D., Southern, M.M., Halliday, K.J., Millar, A.J., 2012. The clock gene circuit in Arabidopsis includes a repressilator with additional feedback loops. *Molecular Systems Biology* 8:574. DOI:10.1038/msb.2012.6

Pooley, C.M., Bishop, S.C., Marion, G., 2015. Using model-based proposals for fast parameter inference on discrete state space, continuous-time Markov processes. *Journal of the Royal Society Interface* 12:20150225. DOI:10.1098/rsif.2015.0225

Pooley, C.M., Marion, G., 2018. Bayesian model evidence as a practical alternative to deviance information criterion. *Royal Society Open Science* 5:171519. DOI:10.1098/rsos.171519

Potvin-Trottier, L., Lord, N.D., Vinnicombe, G., Paulsson, J., 2016. Synchronous long-term oscillations in a synthetic gene circuit. *Nature* 538:514–517. DOI:10.1038/nature19841

Prangle, D., 2016. Lazy ABC. *Statistics and Computing* 26:171–185. DOI:10.1007/s11222-014-9544-3

Prescott, T.P., Baker, R.E., 2018. Multifidelity approximate Bayesian computation. *ArXiv e-prints*, `arXiv:1811.09550 [stat.CO]`

Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., 1997. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press.

Pritchard, J.K, Seielstad, M.T., Perez-Lezaun, A., Feldman, M.W., 1999. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution* 16:1791–1798. DOI:10.1093/oxfordjournals.molbev.a026091

Puliafito, A., Hufnagel, L., Neveu, P., Streichan, S., Sigal, A., Fygenson, D.K., Shraiman, B.I., 2012. Collective and single cell behavior in epithelial contact inhibition. *Proceedings of the National Academy of Sciences* 109:739–744. DOI:10.1073/pnas.1007809109.

Raj, A., van Oudenaarden, A., 2008. Nature, nurture, or chance: stochastic gene expression and its Consequences. *Cell* 135:216–226. DOI:10.1016/j.cell.2008.09.050

Rao, C.V., Arkin, A.P., 2003. Stochastic chemical kinetics and the quasi-steady-state assumption: application to the Gillespie algorithm. *The Journal of Chemical Physics* 118:4999–5010. DOI:10.1063/1.1545446

Rathinam, M., Petzold, L.R., Cao, Y., Gillespie, D.T., 2003. Stiffness in stochastic chemically reacting systems: The implicit tau-leaping method. *The Journal of Chemical Physics* 119:12784–12794. DOI:10.1063/1.1627296

Ratmann, O., Donker, G., Meijer, A., Fraser, C., Koelle, K., 2012. Phylodynamic inference and model assessment with approximate Bayesian computation: Influenza as a case study. *PLOS Computational Biology* 8:e1002835. DOI:10.1371/journal.pcbi.1002835

Reiss, R.-D., 1981. Nonparametric estimation of smooth distribution functions. *Scandinavian Journal of Statistics* 8:116–119.

Roberts, G.O., Rosenthal, J.S., 2001. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science* 16:351–367. DOI:10.1214/ss/1015346320

Roberts, G.O., Rosenthal, J.S., 2004. General state space Markov chains and MCMC algorithms. *Probability Surveys* 1:20–71. DOI:10.1214/154957804100000024

Roberts, G.O., Rosenthal, J.S., 2009. Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics* 18:349–367. DOI:10.1198/jcgs.2009.06134

Ross, R.J.H., Baker, R.E., Parker, A., Ford, M.J., Mort, R.L., Yates, C.A., 2017. Using approximate Bayesian computation to quantify cell-cell adhesion parameters in cell migratory process. *npj Systems Biology and Applications* 3:9. DOI:10.1038/s41540-017-0010-7

Rubin, D.B., 1981. The Bayesian bootstrap. *The Annals of Statistics* 9:130–134. DOI:10.1214/aos/1176345338

Ruess, J., Parise, F., Milias-Argeitis, A., Khammash, M., Lygeros, J., 2015. Iterative experiment design guides the characterization of a light-inducible gene expression circuit. *Proceedings of the National Academy of Sciences* 112:8148–8153. DOI:10.1073/pnas.1423947112

Ryan, E.G., Drovandi, C.C., McGree, J.M., Pettitt, A.N., 2016. A review of modern computational algorithms for Bayesian optimal design. *International Statistical Review* 84:128–154. DOI:10.1111/insr.12107

Rynn, J.A.J., Cotter, S.L., Powell, C.E., Wright, L., 2019. Surrogate accelerated Bayesian inversion for the determination of the thermal diffusivity of a material. *Metrologia* 59:015018. DOI:10.1088/1681-7575/aaf984

Sarapata, E.A., de Pillis, L.G. 2014. A comparison and catalog of intrinsic tumor growth models. *Bulletin of Mathematical Biology* 76:2010–2024. DOI:10.1007/s11538-014-9986-y

Sahl, S.J., Hell, S.W., Jakobs, S., 2017. Fluorescence nanoscopy in cell biology. *Nature Reviews Molecular Cell Biology* 18:685–701. DOI:10.1038/nrm.2017.71

Sanft, K.R., Wu, S., Roh, M., Fu, J., Lim, R.K., Petzold, L.R., 2011. StochKit2: software for discrete stochastic simulation of biochemical systems with events. *Bioinformatics* 27:2457–2458. DOI:10.1093/bioinformatics/btr401

Satija, R., Shalek, A.K., 2014. Heterogeneity in immune responses: from populations to single cells. *Trends in Immunology* 35:219–229. DOI:10.1016/j.it.2014.03.004

Savla, U., Olson, L.E., Waters, C.M., 2004. Mathematical modeling of airway epithelial wound closure during cyclic mechanical strain. *Journal of Applied Physiology* 96:566–574. DOI:10.1152/japplphysiol.00510.2003

Schlögl, F., 1972. Chemical reaction models for non-equilibrium phase transitions. *Z. Physik* 253:147–161.

Schnoerr, D., Sanguinetti, G., Grima, R., 2017. Approximation and inference methods for stochastic biochemical kineticsa tutorial review. *Journal of Physics A: Mathematical and Theoretical* 50:093001. DOI:10.1088/1751-8121/aa54d9

Schwarz, G., 1978. Estimating the dimension of a model. *Annals of Statistics* 6:461–464. DOI:10.1214/aos/1176344136

Scott, J.G., Basanta, D., Anderson, A.R.A., Gerlee, P., 2013. A mathematical model of tumor self-seeding reveals secondary metastatic deposits as drivers of primary tumour growth. *Journal of the Royal Society Interface* 10:20130011. DOI:10.1098/rsif.2013.0011

Sengers, B.G., Please, C.P., Oreffo, R.O., 2007. Experimental characterization and computational modelling of two–dimensional cell spreading for skeletal regeneration. *Journal of the Royal Society Interface* 4:1107–1117. DOI:10.1098/rsif.2007.0233

Shannon, C.E., 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27:379–423. DOI:10.1002/j.1538-7305.1948.tb01338.x

Sherratt, J.A., 2015. Using wavelength and slope to infer the historical origin of semiarid vegetation bands. *The Proceedings of the National Academy of Sciences* 112:4202–4207. DOI:10.1073/pnas.1420171112

Sherratt, J.A., 2016. When does colonisation of a semi-arid hillslope generate vegetation patterns? *Journal of Mathematical Biology* 73:199–226. DOI:10.1007/s00285-015-0942-8

Sherratt, J.A., Murray, J.D., 1990. Models of epidermal wound healing. *Proceedings of the Royal Society of London. Series B, Biological Sciences* 241:29–36. DOI:10.1098/rspb.1990.0061

Shimojo, H., Ohtsuka, T., Kageyama, R., 2008. Oscillations in notch signaling regulate maintenance of neural progenitors. *Neuron* 58:52–64. DOI:10.1016/j.neuron.2008.02.014

Silk, D., Filippi, S., Stumpf, M.P.H., 2013. Optimizing threshold-schedules for sequential approximate Bayesian computation: applications to molecular systems. *Statistical Applications in Genetics and Molecular Biology* 12:603–618. DOI:10.1515/sagmb-2012-004

Silk, D., Kirk, P.D.W., Barnes, C.P., Toni, T., Stumpf, M.P.H., 2014. Model selection in systems biology depends on experimental design. *PLOS Computational Biology* 10:e1003650. DOI:10.1371/journal.pcbi.1003650

Silverman, B.W., 1986. Density estimation for statistics and data analysis. *Monographs on Statistics and Applied Probability*. Boca Ranton: Chapman & Hall/CRC.

Simpson, M.J., Baker, R.E., McCue, S.W., 2011. Models of collective cell spreading with variable cell aspect ratio: A motivation for degenerate diffusion models. *Physical Review E* 83:021901. DOI:10.1103/PhysRevE.83.021901

Simpson, M.J., Landman, K.A., Bhaganagarapu, K., 2007a. Coalescence of interacting cell populations. *Journal of Theoretical Biology* 247:525–543. DOI:10.1016/j.jtbi.2007.02.020

Simpson, M.J., Landman, K.A., Hughes, B.D., Newgreen, D.F., 2006. Looking inside an invasion wave of cells using continuum models: Proliferation is the key. *Journal of Theoretical Biology* 243:343–360. DOI:10.1016/j.jtbi.2006.06.021

Simpson, M.J., Landman, K.A., Hughes, B.D., 2010. Cell invasion with proliferation mechanisms motivated by time-lapse data. *Physica A: Statistical Mechanics and its Applications* 389:3779–3790. DOI:10.1016/j.physa.2010.05.020

Simpson, M.J., Binder, B.J., Haridas, P., Wood, B.K., Treloar,K.K., McElwain, D.L.S., Baker, R.E., 2013. Experimental and modelling investigation of monolayer development with clustering. *Bulletin of Mathematical Biology* 75:871–889. DOI:10.1007/s11538-013-9839-0

Simpson, M.J., Zhang, D.C., Mariani, M., Landman, K.A., Newgreen, D.F., 2007b. Cell proliferation drives neural crest cell invasion of the intestine. *Developmental Biology* 302:553–568. DOI:10.1016/j.ydbio.2006.10.017

Sisson, S.A., Fan, Y., Beaumont, M., 2018. *Handbook of approximate Bayesian computation*, 1st edn. Chapman & Hall/CRC.

Sisson, S.A., Fan, Y., Tanaka, M.M., 2007. Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences* 104:1760–1765. DOI:10.1073/pnas.0607208104

Skellam, J.G., 1951. Random dispersal in theoretical populations. *Biometrika* 38:196–218. DOI:10.2307/2332328

Slepoy, A., Thompson, A.P., Plimpton, S.J., 2008. A constant-time kinetic Monte Carlo algorithm for simulation of large biochemical reaction networks. *The Journal of Chemical Physics* 128:205101. DOI:10.1063/1.2919546

Slezak, F.D., Surez, C., Cecchi, G.A., Marshall, G., Stolovitzky, G., 2010. When the optimal is not the best: Parameter estimation in complex biological models. *PLOS One* 5:e13283. DOI:10.1371/journal.pone.0013283

Sloan, S.W., Abbo, A.J., 1999. Biot consolidation analysis with automatic time stepping and error control Part 1: theory and implementation. *International Journal for Numerical and Analytical Methods in Geomechanics* 23:467–492. DOI:10.1002/(SICI)1096-9853(199905)23:6¡467::AID-NAG949¿3.0.CO;2-R

Small, P.M., Hopewell, P.C., Singh, S.P., Paz, A., Parsonnet, J., Ruston, D.C., Schecter, G.F., Daley, C.L., Schoolnik, G.K., 1994. The epidemiology of tuberculosis in San Francisco – a population-based study using conventional and molecular methods. *New England Journal of Medicine* 330:1703–1709. DOI:10.1056/NEJM199406163302402

Smith, S., Grima, R., 2016. Breakdown of the reaction-diffusion master equation with nonelementary rates. *Physical Review E* 93:052135. DOI:10.1103/PhysRevE.93.052135

Smith, S., Grima, R., 2018. Single-cell variability in multicellular organisms. *Nature Communications* 9:345. DOI:10.1038/s41467-017-02710-x

Soltani, M., Vargas-Garcia, C.A., Antunes, D., Singh, A., 2016. Intercellular variability in protein levels from stochastic expression and noisy cell cycle processes. *PLOS Computational Biology* 12:e1004972. DOI:10.1371/journal.pcbi.1004972

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64:583–639. DOI:10.1111/1467-9868.00353

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A., (2014) The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76:485–493. DOI:10.1111/rssb.12062

Stoica, P., Selen, Y., 2004. Model-order selection: A review of information criterion rules. *IEEE Signal Processing Magazine* 21:36–47. DOI:10.1109/MSP.2004.1311138

Stumpf, M.P.H., 2014. Approximate Bayesian inference for complex ecosystems. *F1000Prime Reports* 6:60. DOI:10.12703/P6-60

Sun, B., Fen, J., Saenko, K., 2016. Return of frustratingly easy domain adaptation. *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence* AAAI Press, 2058–2065.

Sunnåker, M., Busetto, A.G., Numminen, E., Corander, J., Foll, M., Dessimoz, C., 2013. Approximate Bayesian computation. *PLOS Computational Biology* 9:e1002803. DOI:10.1371/journal.pcbi.1002803

Swanson, K.R., Alvord, E.C., Murray, J.D., 2002. Virtual brain tumours (gliomas) enhance the reality of medical imaging and highlight inadequacies of current therapy. *British Journal of Cancer* 86:14–18. DOI:10.1038/sj.bjc.6600021

Swanson, K.R., Bridge, C., Murray, J.D., Alvord, E.C., 2003. Virtual and real brain tumors: Using mathematical modeling to quantify glioma growth and invasion. *Journal of the Neurological Sciences* 216:1–10. DOI:10.1016/j.jns.2003.06.001

Tanaka, M.M., Francis, A.R., Luciani, F., Sisson, S.A., 2006. Using approximate Bayesian computation to estimate tuberculosis transmission parameter from genotype data. *Genetics* 173:1511–1520. DOI:10.1534/genetics.106.055574

Taniguchi, Y., Choi, P.J., Li, G.-W., Chen, H., Babu, M., Heam, J., Emili, A., Xie, X.S., 2010. Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science* 329:533–538. DOI:10.1126/science.1188308

Tavaré, S., Balding, D.J., Griffiths, R.C., Donnelly, P., 1997. Inferring coalescence times from DNA sequence data. *Genetics* 145:505–518.

Taylor, C.M., Hastings, A., 2005. Allee effects in biological invasions. *Ecology Letters* 8:895–908. DOI:10.1111/j.1461-0248.2005.00787.x

Thanh, V.H., 2017. Stochastic simulation of biochemical reactions with partial-propensity and rejection-based approaches. *Mathematical Biosciences* 292:67–75. DOI:10.1016/j.mbs.2017.08.001

Thanh, V.H., Priami, C., Zunino, R., 2014. Efficient rejection-based simulation of biochemical reactions with stochastic noise and delays. *The Journal of Chemical Physics* 141:134116. DOI:10.1063/1.4896985

Thattai, M., van Oudenaarden, A., 2001. Intrinsic noise in gene regulatory networks. *Proceedings of the National Academy of Sciences* 98:8614–8619. DOI:10.1073/pnas.151588598

The Event Horizon Telescope Collaboration, Akiyama, K., Alberdi, A., Alef, W., Asada, K., Azulay, R., Baczko, A.-K., Ball, D., Baloković, M., Barrett, J., et al., 2019. First M87 Event Horizon Telescope results. VI. The shadow and mass of the central black hole. *The Astrophysical Journal Letters* 875:L6. DOI:10.3847/2041-8213/ab1141

Thorne, T., Stumpf, M.P.H., 2012. Graph spectral analysis of protein interaction network evolution. *Journal of The Royal Society Interface* 9:2653–2666. DOI:10.1098/rsif.2012.0220

Tian, T., Burrage, K., 2004. Binomial leap methods for simulating stochastic chemical kinetics. *The Journal of Chemical Physics* 121:10356–10364. DOI:10.1063/1.1810475

Tian, T., Burrage, K., 2006. Stochastic models for regulatory networks of the genetic toggle switch. *Proceedings of the National Academy of Sciences* 103:8372–8377. DOI:10.1073/pnas.0507818103

Tian, T., 2010. Stochastic models for inferring genetic regulation from microarray gene expression data. *Biosystems* 99:192–200. DOI:10.1016/j.biosystems.2009.11.002

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., Stumpf, M.P.H., 2009. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface* 6:187–202. DOI:10.1098/rsif.2008.0172

Treloar, K.K., Simpson, M.J., Haridas, P., Manton, K.J., Leavesley, D.I., McElwain, D.L.S., Baker, R.E., 2013. Multiple types of data are required to identify the mechanisms influencing the spatial expansion of melanoma cell colonies. *BMC Systems Biology* 7:137. DOI:10.1186/1752-0509-7-137

Treloar, K.K., Simpson, M.J., McElwain, D.L.S., Baker, R.E., 2014. Are *in vitro* estimates of cell diffusivity and cell proliferation rate sensitive to assay geometry? *Journal of Theoretical Biology* 356:71–84. DOI:10.1016/j.jtbi.2014.04.026

Tsoularis, A., Wallace, J., 2002. Analysis of logistic growth models. *Mathematical Biosciences* 179:21–55. DOI:10.1016/S0025-5564(02)00096-2

Turner, T.E., Schnell, S., Burrage, K., 2004. Stochastic approaches for modelling *in vivo* reactions. *Computational Biology and Chemistry* 28:165–178. DOI:10.1016/j.compbiolchem.2004.05.001

Vankov, E.R., Guindani, M., Ensor, K.B. 2019. Filtering and estimation for a class of stochastic volatility models with intractable likelihoods. *Bayesian Analysis* 14:29–52. DOI:10.1214/18-BA1099

Vanlier, J., Tiemann, C.A., Hilbers, P.A.J., van Riel, N.A.W., 2012. A Bayesian approach to targeted experiment design. *Bioinformatics* 28:1136–1142. DOI:10.1093/bioinformatics/bts092

Vanlier, J., Tiemann, C.A., Hilbers, P.A.J., van Riel, N.A.W., 2014. Optimal experiment design for model selection in biochemical networks. *BMC Systems Biology* 8:20. DOI:10.1186/1752-0509-8-20

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C., 2019. Rank-normalization, folding, and localization: an improved $\hat{R}$ for assessing convergence of MCMC. *arXiv e-print*, `arXiv:1903.08008 [stat.CO]`

Vellela, M., Qian, H., 2009. Stochastic dynamics and non-equilibrium thermodynamics of a bistable chemical system: the Schlögl model revisited. *Journal of the Royal Society Interface* 6:925–940. DOI:10.1098/rsif.2008.0476

Verhulst, P.F., 1838. Notice sur la loi que la populations suit dans son accroissement. *correspondence mathematique et physique* X:113–121.

Vincenot, C.E., Carteni, F., Mazzoleni, S., Rietkerk, M., Giannino, F., 2016. Spatial self-organization of vegetation subject to climatic stressinsights from a system dynamicsindividual-based hybrid model. *Frontiers in Plant Science* 7:636. DOI:10.3389/fpls.2016.0063

Vittadello, S.T., McCue, S.W., Gunasingh, G., Haass, N.K., Simpson, M.J., 2018. Mathematical models for cell migration with real-time cell cycle dynamics. *Biophysical Journal* 114:1241–1253. DOI:10.1016/j.bpj.2017.12.041

Vo, B.N., Drovandi, C.C., Pettit, A.N., Simpson, M.J., 2015a. Quantifying uncertainty in parameter estimates for stochastic models of collective cell spreading using approximate Bayesian computation. *Mathematical Biosciences* 263:133–142. DOI:10.1016/j.mbs.2015.02.010

Vo, B.N., Drovandi, C.C., Pettit, A.N., Pettet,G.J., 2015b. Melanoma cell colony expansion parameters revealed by approximate Bayesian computation. *PLOS Computational Biology* 11:e1004635. DOI:10.1371/journal.pcbi.1004635

Voliotis, M., Thomas, P., Grima, R., Bowsher, C.G., 2016. Stochastic simulation of biomolecular networks in dynamic environments. *PLOS Computational Biology* 12:e1004923. DOI:10.1371/journal.pcbi.1004923

von Hardenberg, J., Meron, E., Shachak, M., Zarmi, Y., 2001. Diversity of vegetation patterns and desertification. *Physical Review Letters* 87:198101. DOI:10.1103/PhysRevLett.87.198101

Wang, Y., Shi, J., Wang, J.J., 2019. Persistence and extinction of population in reactiondiffusionadvection model with strong Allee effect growth. *Journal of Mathematical Biology* 78:2093–2140. DOI:10.1007/s00285-019-01334-7

Wang, J., Wu, Q., Hu, X.T., Tian, T., 2016. An integrated approach to infer dynamic protein-gene interactions a case study of the human P53 protein. *Methods* 110:3–13. DOI:10.1016/j.ymeth.2016.08.001

Warne, D.J., Baker, R.E., Simpson M.J., 2017. Optimal quantification of contact inhibition in cell populations. *Biophysical Journal* 113:1920–1924. DOI:10.1016/j.bpj.2017.09.016

Warne, D.J., Baker, R.E., Simpson, M.J., 2018. Multilevel rejection sampling for approximate Bayesian computation. *Computational Statistics & Data Analysis* 124:71–86. DOI:10.1016/j.csda.2018.02.009

Warne, D.J., Baker, R.E., Simpson, M.J., 2019a. Simulation and inference algorithms for stochastic biochemical reaction networks: form basic concepts to state-of-the-art. *Journal of the Royal Society Interface* 16:20180943. DOI:10.1098/rsif.2018.0943

Warne, D.J., Baker, R.E., Simpson, M.J., 2019b. Using experimental data and information criteria to guide model selection for reaction–diffusion problems in mathematical biology. *Bulletin of Mathematical Biology* 81:1760–1804. DOI:10.1007/s11538-019-00589-x

Warne, D.J., Baker, R.E., Simpson, M.J., 2019c. A practical guide to pseudo-marginal methods for computational inference in systems biology. *ArXiv e-prints* `arXiv:1912.12404 [q-bio.MN]`.

Warne, D.J., Sisson, S.A., Drovandi, C. 2019d. Acceleration of expensive computations in Bayesian statistics using vector operations. *ArXiv e-prints* `arXiv:1902:09046 [stat.CO]`

Wegmann, D., Leuenberger, C., Neuenschwander, S., Excoffier, L., 2010. ABCtoolbox: a versatile toolkit for approximate Bayesian computations. *BMC Bioinformatics* 11:116. DOI:10.1186/1471-2105-11-116

Weiss, G.H., Dishon, M., 1971. On the asymptotic behavior of the stochastic and deterministic models of an epidemic. *Mathematical Biosciences* 11:261–265. DOI:10.1016/0025-5564(71)90087-3

Wilson, D., Baker, R.E., 2016. Multi-level methods and approximating distribution functions. *AIP Advances* 6:075020. DOI:10.1063/1.4960118

Wilkinson, D.J., 2009. Stochastic modelling for quantitative description of heterogeneous biological systems. *Nature Reviews Genetics* 10:122–133. DOI:10.1038/nrg2509

Wilkinson, D.J., 2011. Parameter inference for stochastic kinetic models of bacterial gene regulation: A Bayesian approach to systems biology. *Bayesian Statistics 9: Proceedings of the Ninth Valencia International Meeting*, Oxford University Press, 679–706. DOI:10.1093/acprof:oso/9780199694587.003.0023

Wilkinson, D.J., 2012. *Stochastic Modelling for Systems Biology*, 2nd edn. CRC Press, 2nd edition.

Wilkinson, R.D., 2013. Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. *Statistical Applications in Genetics and Molecular Biology* 12:129–141. DOI:10.1515/sagmb-2013-0010

Wilkinson, R.D., Steiper, M.E., Soligo, C., Martin, R.D., Yang, Z., Tavaré, S., 2011. Dating primate divergences through an integrated analysis of palaeontological and molecular data. *Systematic Biology* 60:16–31. DOI:10.1093/sysbio/syq054

Witelski, T.P., 1995. Merging traveling waves for the porous-Fisher's equation. *Applied Mathematics Letters* 8:57–62. DOI:10.1016/0893-9659(95)00047-T

Woods, M.L., Barnes, C.P., 2016. Mechanistic modelling and Bayesian inference elucidates the variable dynamics of double-strand break repair. *PLOS Computational Biology* 12:e1005131. DOI:10.1371/journal.pcbi.1005131

Yamada, M., Suzuki, Y., Nagasaki, S.C., Okuno, H., Imayoshi, I., 2018. Light control of the Tet gene expression system in mammalian cells. *Cell Reports* 25:487–500. DOI:10.1016/j.celrep.2018.09.026

Yang, Y., 2005. Can the strengths of AIC and BIC be shared? a conflict between model identification and regression estimation. *Biometrika* 92:937–950. DOI:10.2307/20441246

Yang, Z., Rodríguez, C.E., 2013. Searching for efficient Markov chain Monte Carlo proposal kernels. *Proceedings of the National Academy of Sciences* 110:19307–19312. DOI:10.1073/pnas.1311790110

Young, J.W., Locke, J.C.W., Altinok, A., Rosenfeld, N., Bacarian, T., Swain, P.S., Mjolsness, E., Elowitz, M.B., 2012. Measuring single-cell gene expression dynamics in bacteria using fluorescence time-lapse microscopy. *Nature Protocols* 7:80–88. DOI:10.1038/nprot.2011.432

Zechner, C., Ruess, J., Krenn, P., Pelet, S., Peter, M., Lygeros, J., Koeppl, H., 2012. Moment-based inference predicts bimodality in transient gene expression. *Proceedings of the National Academy of Sciences* 109:8340–8345, 2012. DOI:10.1073/pnas.1200161109