MONASH University

Exact and approximate methods of sampling from posterior distributions of parameters in non-linear ordinary differential equations

by

Amani Alahmadi
School of Mathematics
Faculty of Sciences
Monash University

A dissertation submitted in partial fulfilment
of the requirements for the degree of
Doctor of Philosophy

Tuesday 21st July, 2020

To my late daughter Leen.
My beloved husband Fahad, and my son Yazan,
Without whom none of my success would be possible.

Acknowledgements

First and foremost I would like to thank my supervisor Associate Prof. Jonathan M.Keith for the time he spent in supporting me through the course of my doctoral studies and for the efforts he put into sharpening my skills. I am grateful to him for all his encouragement, patience and insightful suggestions. I am also indebted to my associate supervisor Associate Prof. Jennifer Flegg for her enthusiasm, motivation and her valuable contribution for supervising me with her invaluable knowledge.

I would thank all the co-authors, Associate Prof. Christopher Drovandi, whose experience in the field of Approximate Bayesian computation has been immense and extremely valuable, and Davis Cochrane for his help and suggestions.

I want to thank my PhD panel members: Associate Prof. Tianhai Tian, Associate Prof. Tim Garoni and Dr. Tiangang Cui for their valuable comments and feedback. I am grateful to my sponsor, Shaqra University, for their generous sponsorship, in funding my studies, and their unlimited support for my family and me.

I wish to thank my friends and colleagues Farhana Sadia and Yuqin Ke for their support and the many fruitful discussions that we shared during my research. I would also acknowledge the support that I got from Monash University and the administration staff in the faculty of science for all their efforts, especially, John

Chan for his continuous support and coordinating role throughout my PhD.

Finally, I want to thank my husband, Fahad Alrehaili for his unconditional love, support, encouragement, understanding and optimism that allowed me to finish this journey. A special thank you goes to my son Yazan Alrehaili who was just one year old when I started my PhD, he made these past few years much more enjoyable and inspirable. I would like to extend my sincere gratitude for the generous support that I got endlessly from my parents and my siblings.

Abstract

Ordinary differential equations (ODEs) are an essential tool for describing physical and biological processes. ODE models contain parameters that represent quantities of biological and physical importance, such as kinetic rates and initial concentrations. These parameters are often hard to measure experimentally and have uncertainty in their values. Statistical methods are often used to perform parameter inference, from noisy observations. However, current standard methods suffer from some limitations and can produce inefficient results. The aim of this thesis is to conduct an investigation of some of the limitations of these methods and provide novel approaches that improve the efficiency of parameter estimation in the context of ODE parameters.

Bayesian inference, which is the focus of this thesis, provides an essential methodology for parameter inference in ODE models. A common Bayesian approach, Markov chain Monte Carlo (MCMC), uses pseudo-random parameter samples from a posterior distribution to estimate relevant integrals. However, a poorly designed MCMC method can exhibit slow mixing and poor convergence. These problems are exacerbated when the parameter space is of high dimension, parameters suffer from identifiability issues, or the posterior distribution is multi-modal. Approximate Bayesian Computation (ABC) is a likelihood-free Bayesian inference methodology that provides an approximation to the posterior when the likelihood is intractable or too computationally intensive to evaluate. This thesis makes

three main contributions to computational Bayesian methods for ODEs. Namely, we highlight major problems due to model misspecification in current widely used ABC methods for ODE parameter estimation, develop a novel ABC scheme based on sequential Monte Carlo (SMC) to address such problems, and propose an alternative method that explores the parameter space more efficiently than MCMC using Quasi Monte Carlo (QMC) methods.

The first contribution of this thesis is related to the current practice of using ABC for parameter inference in ODEs. This thesis conducted a study to demonstrate problems caused by neglecting the simulation of random errors and illustrates for both simple and complex epidemiological models that this can produce serious errors in the estimated posterior distributions. The second contribution is to propose a modification to the current SMC ABC method. A summary statistic was used to aid the construction of acceptance criteria that allow correct approximation of the posterior distribution. Including estimation of the error terms in the inference process improves significantly the efficiency of the estimated posterior distribution as well as the accuracy of predictions. Finally, a new method was proposed in this thesis, which allows the exploration of the parameter space conducted using low discrepancy point sets instead of random points as in MCMC. This method improves estimation of the posterior and outperforms MCMC when the posterior density has multiple modes. Using predetermined point sets, this method allows the algorithm to be implemented in a parallel computing environment, and reduces the computational cost significantly. In this thesis, I highlight the usefulness of the proposed methods for a variety of nonlinear ODEs with noisy observations.

Declaration

This thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

Amani Alahmadi

Thesis including published works declaration

This thesis includes one original paper published in a peer reviewed journal, two submitted articles and one submitted joint authorship article. The core theme of the thesis is to address some of the limitations of current Bayesian approaches to parameter inference for ODEs and to develop new statistical methods for this purpose. The ideas, development and writing up of all the papers in the thesis were the principal responsibility of myself, the student, working within the School of Mathematics at Monash University under the supervision of Associate Prof. Jonathan M. Keith and Associate Prof.Jennifer Flegg.

In the case of Chapters 3, 4, 5 and Appendix B, my contribution to the work involved the following:

Thesis Chapter	**Publication Title**	**Status** *(published, in press, accepted or returned for revision)*	**Nature and % of student contribution**	**Co-author name(s) Nature and % of Co-author's contribution***	**Co-author(s), Monash student Y/N***
3	*A comparison of approximate versus exact techniques for bayesian parameter inference in nonlinear ordinary differential equation models*	*Published*	*60 % produced the final codes, performed the analysis, interpreted results and wrote the manuscript*	*Jennifer Flegg conceived of the presented idea, supervised the findings of this work and revised the manuscript, 15%*	*No*
				Davis Cochrane participated in developing the initial draft of the code and the paper, 5%	Yes
				Christopher Drovandi provided guidance on the ABC implementation and revised the manuscript, 5%.	No
				Jonathan Keith conceived of the presented idea, supervised the findings of this work and revised the manuscript, 15%.	No

4	Estimating error parameters in dynamical systems models using Approximate Bayesian Computation	Submitted	80 % Concept, proposed the method, performed the analysis, interpreted the results and wrote the code and the manuscript	Jennifer Flegg supervised, input into manuscript and proofread, 10%.	*No*
				Jonathan Keith supervised, input into manuscript and proofread, 10 %.	*No*
5	Low Discrepancy Sequences for Bayesian Estimation in Ordinary Differential Equations	Submitted	80 % Concept, proposed the method, performed the analysis, interpreted the results and wrote the code and the manuscript	Jennifer Flegg supervised, input into manuscript and proofread, 10%.	*No*
				Jonathan Keith supervised, input into manuscript and proofread, 10 %.	*No*
Append ix B	Influencing public health policy with data-informed mathematical models of infectious diseases: Recent developments and new challenges	Submitted	5.55%	Andrew Black, Deborah Cromer, Pavithra Jayasundara, Jonathan Keith, James McCaw, Rob Moss, Freya Shearer, Sai Thein Than Tun, James Walker, Lisa White, Jason Whyte, Ada Yan and Alexander Zarebski, 5.55%	No
				Sarah Belet, 5.55%	Yes
				Jennifer Flegg, Thomas House and Joshua Ross, 5.55% each. the corresponding Authors	No

I have not renumbered sections of submitted or published papers in order to generate a consistent presentation within the thesis.

Student signature: **Date:15/04/2020**

The undersigned hereby certify that the above declaration correctly reflects the nature and extent of the student's and co-authors' contributions to this work. In instances where I am not the responsible author I have consulted with the responsible author to agree on the respective contributions of the authors.

Main Supervisor signature: **Date:15/04/2020**

Contents

List of Figures

List of Tables

List of Abbreviations

ABC	Approximate Bayesian computation
CDF	Cumulative Distribution Function
CRs	Credible Intervals
HMC	Hamiltonian Monte Carlo
I	Infected
KL	Kullback-Liebler
LDS	low discrepancy sequences
MAD	Median Absolute Deviation
MAE	Mean Absolute Error
MC	Monte Carlo
MCMC	Markov chain Monte Carlo
ODEs	Ordinary Differential Equations

P	Population
PDEs	Partial Differential Equations
PPD	Posterior Predictive Distribution
QMC	Quasi Monte Carlo
R	Recovered
RQMC	Randomized QMC
RWMH	Random Walk Metropolis Hastings
S	Susceptible
SIR	Susceptible-Infected-Recovered
SMC	sequential Monte Carlo

Chapter 1

Introduction

1.1 Motivation

Ordinary Differential Equations (ODEs) are a popular mathematical tool for describing physical and biological processes in the real world. ODEs describe the differential structure of dynamical variables, often in terms of some unknown parameters that must be estimated using noisy observations. The task of estimating the unknown parameters is considered a significant statistical problem and is sometimes referred to as the inverse problem for ODEs.

Bayesian approaches to such estimation problems are popular, not only because they are able to quantify inherent uncertainty, but also because they have been found to deliver excellent results for complex non-linear ODEs (see Vaart, Prangle, and Sibly (2018); Prangle et al. (2017); Ghosh, Dasmahapatra, and Maharatna (2017); Calderhead, Girolami, and Lawrence (2008); Dondelinger, Rogers, and Husmeier (2013); Wang and Barber (2014); Toni, Welch, Strelkowa, Ipsen, and Stumpf (2009) for examples). Typically, Bayesian estimation involves sampling parameters from a posterior distribution using a computational method. However,

some of these approaches involve integration over a subset of parameters to obtain a marginal likelihood. Commonly, for non-linear ODE models, these integrals are either not tractable or they are computationally expensive. Identifying the limitations of current approaches and development of new statistical methods are required to improve parameter inference and future predictions in the context of ODEs.

The parameter space can be explored using pseudo-random sampling, which is a deterministic process that generates numbers which appear to be distributed according to a given distribution, as in Monte Carlo integration. Markov chain Monte Carlo (MCMC) algorithms were first developed by Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller (1953) and are considered one of the most important tools to carry out statistical analyses in a Bayesian context. This class of technique creates a Markov chain, that has the posterior distribution as its limiting distribution. In complex applications, the likelihood function can be intractable or computationally very expensive to evaluate. Thus, alternative likelihood-free methods have been developed, such as Approximate Bayesian Computation (ABC), which first appeared in Tavaré, Balding, Griffiths, and Donnelly (1997) and Pritchard, Seielstad, Perez-Lezaun, and Feldman (1999). The standard ABC method involves comparing a summary statistic from the simulated data and the observed data using a discrepancy function. To improve the low acceptance rate in the basic ABC algorithm, population-based methods such as the sequential Monte Carlo (SMC) algorithm were proposed in Sisson, Fan, and Tanaka (2007), based on the SMC methodology developed by Del Moral, Doucet, and Jasra (2006). An application of SMC ABC to ODE models was described in Toni et al. (2009). However, some of the current ABC approaches fail to adequately model the error terms associated with observed data because the acceptance probability depends on a discrepancy function that does not take into consideration the error term. One objective of this thesis is to address this limitation.

In this thesis, to overcome the limitations of existing ABC approaches for parame-

ter inference within ODEs, I developed a novel modification to the standard SMC ABC. This new approach combines the benefits of using a likelihood-free method with a framework that is capable of accurately quantifying the uncertainty and improving the efficiency of the estimation. This method takes observation error into account when performing parameter inference using the SMC ABC method.

Moreover, in the case where the likelihood is tractable when performing parameter inference in ODEs, a poorly designed MCMC method can suffer from slow mixing and poor convergence. This problem is particularly acute when the parameter space is of high dimension or parameters are unidentifiable. Furthermore, many MCMC approaches explore the parameter space by making local moves (Chumbley, Friston, Fearn, & Kiebel, 2007). Consequently, the sampler may get stuck in some high-density region and miss exploring an area with low density, especially in the case where the posterior distribution is multi-modal. This thesis develops efficient and inexpensive methods that explore the parameter space from a global point of view. This second contribution of this thesis uses low discrepancy point sets to construct Quasi Monte Carlo (QMC) methods to explore the parameter space. The results show that using QMC point sets can outperform MCMC in terms of computational time and accuracy of estimation, especially, when the target posterior is multi-modal. This approach is simple and easy to apply. Additionally, statistical software is under development for this thesis and it will be publicly available when it is finished. This software is likely to benefit practitioners working in a wide range of fields.

1.2 Research Aim and Objectives

The overall aim of this thesis is to address some of the limitations of current Bayesian approaches to parameter inference for ODEs and to develop new statistical methods for this purpose. This overall aim can be divided into the following

objectives:

1. **Objective 1**
   Clarify the problems arising from model misspecification in ABC methods for ODE parameter estimation. Addressing this issue is very important to ensure that practitioners of such methods are extracting valid information from observed data and applying ABC methods appropriately.

2. **Objective 2**
   Develop an ABC SMC algorithm to overcome the current misspecification issue outlined in Objective 1 above. The approach I propose uses an improved acceptance criteria that appropriately accounts for observational error and produces comparable accuracy to MCMC.

3. **Objective 3**
   Propose a new approach that explores the parameter space in a more even fashion. In this approach, instead of using random sampling as in MCMC, I use predetermined QMC point sets. This significantly reduces the computational cost compared to MCMC, particularly when the posterior is multimodal.

4. **Objective 4**
   Evaluate the proposed methods by applying them to real world research problems in epidemiological modelling (one simple and one more complex ODE system) and ecological modelling, using simulated data and actual observations.

1.3 Thesis Structure and Contributions

This thesis comprises published and submitted papers; thus, the main chapters that contribute to the thesis aim are published or submitted. The candidate is

the corresponding author on all the articles included in the main body of the thesis. The thesis is organised in six chapters, including the current one. The main contributions of Chapters 2-6 are outlined as follows:

- Chapter 2, **Monte Carlo Methods for Bayesian Statistics**
  In this chapter, I provide a brief introduction to Bayesian statistics for inferring parameters of an ODE system and then provide a self-contained overview of the fundamentals of MCMC, QMC and ABC methods that are relevant to the contributions made in this thesis. There is some overlap between this chapter and the main chapters of this thesis (Chapters 3 - 5). These later chapters contain published or submitted journal articles.

- Chapter 3, **A Comparison of Approximate Versus Exact Techniques for Bayesian Parameter Inference in Non-linear ODE Models**
  In this chapter, we highlight, clarify and explore several limitations of existing ABC approaches for ODE models (Objective 1). Specifically, we argue that some common ABC approaches fail to sufficiently quantify the uncertainty in parameter values, which significantly reduces the accuracy of posterior estimation and posterior predictive credible intervals. We demonstrate this problem by applying ABC approaches to two ODE epidemiological models with simulated data and one with real data concerning malaria transmission in Afghanistan.

- Chapter 4, **Estimating error parameters in dynamical systems models using Approximate Bayesian Computation**
  In this chapter, a new SMC ABC method is proposed to overcome the limitation previously identified (Objective 2). In this method, we propose a summary statistic that facilitates accurate estimation of the noise associated with observations. The new summary statistic incorporates knowledge about the error term into the choice of the tolerance sequence, thus producing more efficient acceptance criteria. The advantages of this method are

illustrated when applying the proposed method to several challenging test problems.

- Chapter 5, **Low Discrepancy Sequences for Bayesian Estimation in Ordinary Differential Equations**
  In this chapter, a method based on using predetermined QMC point sets is proposed. This method involves two main contributions. First, the new method explores the parameter space of non-linear ODEs more evenly and efficiently using QMC point sets. Second, I propose a new and easy method to visualise posterior marginals using cumulative summation. Our results show that QMC outperforms MCMC in terms of computational cost, ease of implementation and the accuracy of posterior estimation, especially when the posterior distribution has a multi-modal surface (Objective 3). Applications of the method to a Lotka-Volterra model and an epidemiological compartmental model with real data illustrate the advantages of the proposed QMC-based method.

- Chapter 6, **Summary and main contributions**
  In this chapter, I summarise the three major contributions of this research and discuss potential areas for future work.

1.4 Publications

1. The first contribution addressing the limitation of using ABC methods within ODE models has been published in the following article:

   Alahmadi, A. A., Flegg, J. A., Cochrane, D. G., Drovandi, C. C. & Keith, J. M. (2020). A comparison of approximate versus exact techniques for Bayesian parameter inference in nonlinear ordinary differential equation models. *Royal Society Open Science*, 7(3), 191315.

2. The second contribution proposing a modification to the SMC ABC method to overcome the current limitations outlined in the first publication has been submitted in the following article:

   Alahmadi, A. A., Flegg, J. A. & Keith, J. M. Estimating error parameters in dynamical systems models using Approximate Bayesian Computation. *Computational Statistics*. Submitted 1/3/2020.

3. The third contribution proposing an approach that explores the parameter space using QMC point sets has been submitted in the following article:

   Alahmadi, A. A., Flegg, J. A. Keith, J. M. Low Discrepancy Sequences for Bayesian Estimation in Ordinary Differential Equations. *Statistics and Computing*. Submitted 31/3/2020.

4. Joint work that reviews recent progress towards solving the challenges of influencing public health policy with data-informed mathematical models of infectious diseases was submitted in the following article:

   Alahmadi, A. A., Belet, S., Black, A., Cromer, D., Flegg, J. A., House, T., Jayasundara, P., Keith, J. M., McCaw, J. M., Moss, R., Ross, J., Shearer, F. M., Tun, S. T. T., Walker, J., White, J., Whyte, J. M., Yan, W. C. Zarebski, A. E. Influencing public health policy with data-informed mathematical models of infectious diseases: Recent developments and new challenges. *Epidemics*. Submitted 24/11/2019.

   Details of this work appear in Appendix B. This work was completed during my candidature but is not considered a direct contribution of this thesis.

Chapter 2

Monte Carlo Methods for Bayesian Statistics

2.1 Introduction

Many exact and approximate Bayesian methods have been developed to sample from a posterior distribution. In this chapter, I give a brief introduction to the Bayesian paradigm as it applies to inferring parameters of ODE systems. Although there are numerous other computational Bayesian methods, this chapter only provides a review of the fundamentals of the MCMC, QMC and ABC methods that are used throughout this thesis.

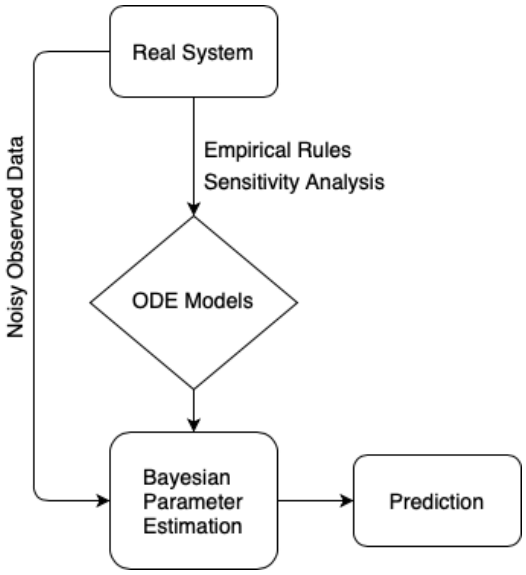

Figure 2.1: Summary of steps involved in making predictions from a model of a real system. In the Bayesian framework, parameter estimation depends on the chosen ODE model and on the noisy observations of the real system. The accuracy of such estimation affects future predictions.

2.2 Bayesian Parameter Estimation in ODE Systems

ODEs represent one of the most widely used techniques for modeling continuous-time, nonlinear, dynamical systems. Typically, there are two prime challenges for researchers to deal with in this context: selecting the most appropriate model and estimating the parameters of the selected model (see Figure 2.1). Solving these problems is generally worthwhile since it leads to an enhanced understanding of the real system and quantifies the associated uncertainty can improve the accuracy of the predictions about future observations. Figure 2.1 illustrates these challenges. In this thesis, the focus is on methods for estimating parameters under a Bayesian framework.

Consider a Q-dimensional dynamical system for the state variable vector, $\mathbf{x}(t)$, described by the system of ODEs:

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}, \boldsymbol{\theta}, t), \tag{2.1}$$

where $\mathbf{x}$ is a $Q \times 1$ vector of the dependent variables and $\mathbf{f}$ is a $Q \times 1$ vector-valued Lipschitz continuous function with respect to $\mathbf{x}$ (this condition is necessary to guarantee the existence and uniqueness of the ODEs solution (Walter, 1998)). The parameter vector, $\boldsymbol{\theta}$, is an $M \times 1$ vector of model parameters, t is the independent variable (often time) and $\dot{\mathbf{x}}$ represents the derivative of $\mathbf{x}$ with respect to the independent variable. Given the dynamical system in Equation (2.1), along with values for the parameter vector $\boldsymbol{\theta}$, and the initial condition $\mathbf{x}_0$, the solution to the system can be approximated numerically, for example, using Euler's method or Runge-Kutta methods (Atkinson, Han, & Stewart, 2011). There are many Runge-Kutta approximation methods; however, the fourth order Runge-Kutta method is the one used in this thesis.

I denote an experimental observation at time t_k by the $Q \times 1$ vector $\mathbf{y}_k$. Experimental observations are taken at K time points, with times stored in a $K \times 1$ vector $\mathbf{t} = (t_1, t_2, ..., t_K)^T$ and observations stored in the $Q \times K$ matrix $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_K)$. These observations are usually associated with some unknown noise process, characterised by one or more variance parameters, say $\boldsymbol{\sigma^2} = (\sigma_1^2, \ldots, \sigma_Q^2)$. The (approximate) solution for the dependent variables at time t_k, given $\boldsymbol{\theta}$ and $\mathbf{x}_0$, is denoted by the $Q \times 1$ vector $\hat{\mathbf{x}}(t_k; \boldsymbol{\theta}, \mathbf{x}_0)$. The solution for the dependent variables at times $\mathbf{t}$ is stored in the $Q \times K$ matrix $\hat{\mathbf{x}}(\mathbf{t}; \boldsymbol{\theta}, \mathbf{x}_0) = (\hat{\mathbf{x}}(t_1; \boldsymbol{\theta}, \mathbf{x}_0), \hat{\mathbf{x}}(t_2; \boldsymbol{\theta}, \mathbf{x}_0), ..., \hat{\mathbf{x}}(t_K; \boldsymbol{\theta}, \mathbf{x}_0))$. In a Bayesian setting, the posterior distribution for $\boldsymbol{\theta}$ and $\boldsymbol{\sigma}^2$ given $\mathbf{y}$ is:

$$p(\boldsymbol{\theta}, \boldsymbol{\sigma}^2, \boldsymbol{x_0}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\sigma}^2, \boldsymbol{x_0}) p(\boldsymbol{\theta}) p(\boldsymbol{\sigma}^2) p(\boldsymbol{x_0}), \tag{2.2}$$

where $p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\sigma}^2, \boldsymbol{x_0})$ is the likelihood, and $p(\boldsymbol{\theta})$, $p(\boldsymbol{\sigma}^2)$ and $p(\boldsymbol{x_0})$ are independent

priors for $\boldsymbol{\theta}$, $\boldsymbol{\sigma}^2$ and $\boldsymbol{x_0}$, respectively.

2.2.1 Noisy Observations

Throughout the thesis I assume that each observation, $\mathbf{y}_k$ for $k = 1, \ldots, K$, has an associated additive noise process, $\boldsymbol{\delta}_k$, such that

$$\mathbf{y}_k = \hat{\mathbf{x}}(t_k; \boldsymbol{\theta}, \mathbf{x}_0) + \boldsymbol{\delta}_k, \tag{2.3}$$

where $\boldsymbol{\delta}_k$ is a $Q \times 1$ vector and $\hat{\mathbf{x}}(t_k; \boldsymbol{\theta}, \mathbf{x}_0)$ is the solution for the dependent variables at time t_k, given $\boldsymbol{\theta}$ and $\mathbf{x}_0$. Under a Gaussian error model and assuming the $\boldsymbol{\delta}_k$ are independent of each other, $\mathbf{y}_k$ follows a multivariate normal distribution:

$$\mathbf{y}_k \sim \text{MVN}(\hat{\mathbf{x}}(t_k; \boldsymbol{\theta}, \mathbf{x}_0), \boldsymbol{\Sigma}(\boldsymbol{\sigma}^2)) \tag{2.4}$$

where $\boldsymbol{\Sigma}(\boldsymbol{\sigma}^2)$ is a diagonal matrix with diagonal elements $\boldsymbol{\sigma}^2 = (\sigma_1^2, \sigma_2^2, ..., \sigma_Q^2)^T$ associated with the Q dependent variables. Hence, the likelihood function is given by

$$\mathcal{L}(\mathbf{y}|\hat{\mathbf{x}}(\mathbf{t}; \boldsymbol{\theta}, \mathbf{x}_0), \boldsymbol{\sigma}^2) = \prod_{k=1}^{K} \text{MVN}(\mathbf{Y}_k; \hat{\mathbf{x}}(t_k; \boldsymbol{\theta}, \mathbf{x}_0), \boldsymbol{\Sigma}(\boldsymbol{\sigma}^2)) \tag{2.5}$$

and the posterior density is

$$p(\boldsymbol{\theta}, \boldsymbol{\sigma}^2, \mathbf{x}_0|\mathbf{y}) \propto p(\boldsymbol{\theta})p(\boldsymbol{\sigma}^2)p(\mathbf{x}_0) \prod_{k=1}^{K} \text{MVN}(\mathbf{Y}_k; \hat{\mathbf{x}}(t_k; \boldsymbol{\theta}, \mathbf{x}_0), \boldsymbol{\Sigma}(\boldsymbol{\sigma}^2)). \tag{2.6}$$

Although we can assume any error model depending on the specific applications under the study, assuming a Gaussian error model is a common choice and standard in many relevant applications (see Toni et al. (2009); Silk, Filippi, and Stumpf (2013); Vaart et al. (2018)). One reason is that a Gaussian distribution has many

convenient analytic properties. Another reason is that noise in the data typically does not come from just one source; it may include contributions from observational measurement error, model misspecification or a combination of both (Vaart et al., 2018). Each of these independent sources of noise can influence the data and, while these sources of noise may not be normally distributed in general, their overall effect can often be modelled as a Gaussian random variable according to the central limit theorem.

Bayesian computational techniques are generally used to sample from the posterior distribution of $\{\boldsymbol{\theta}, \mathbf{x}_0, \boldsymbol{\sigma}^2\}$ in the case of MCMC or an approximation to the posterior in the case of ABC methods. In the following section, I provide a brief review of MCMC, QMC and ABC approaches.

2.3 Markov Chain Monte Carlo (MCMC)

MCMC is a general method for drawing samples from a distribution when this distribution is not available in closed form, or is otherwise not amenable to more direct sampling methods. In the case of inferring parameters in ODEs, Equation (2.6) may fail to have a closed form for at least one of two main reasons: i) the normalising constant of proportionality in Bayes' theorem is an integral that cannot be evaluated analytically or ii) the solution $\hat{\mathbf{x}}(t_k; \boldsymbol{\theta}, \mathbf{x}_0)$ of the ODE model is not available in closed form. The MCMC approach combines the Monte Carlo strategy of using a large number of random samples from a target distribution to estimate integrals and marginal posterior distributions with a Markov chain strategy of using each random sample to generate the next random sample based on a transition kernel (Van Ravenzwaaij, Cassey, & Brown, 2018). Generally, MCMC samples are correlated and equally weighted.

In Bayesian analysis, there are many popular MCMC methods. Throughout this thesis, I have used the Metropolis-Hastings (MH) algorithm or in some cases

Hamiltonian Monte Carlo (HMC). These are each discussed in turn below.

2.3.1 Metropolis-Hastings (MH) Algorithm

The MH algorithm generates a sequence of random samples from an arbitrary target distribution. It is often applicable even when the dimension of the parameter space is large (Metropolis et al., 1953; Hastings, 1970). The MH algorithm constructs a Markov chain for which the stationary and limiting distribution is the target posterior distribution. After running the chain for a sufficient amount of time, say (K iterations), and discarding the *burn-in* samples ($\boldsymbol{\theta^0}, \boldsymbol{\theta^1}, \ldots, \boldsymbol{\theta^K}$), subsequent samples from the chain can be considered draws from the posterior distribution, where transition probabilities from $\boldsymbol{\theta^t}$ to $\boldsymbol{\theta^{t+1}}$ depend only on the position $\boldsymbol{\theta^t}$ rather than any past positions ($\boldsymbol{\theta^0}, \boldsymbol{\theta^1}, \ldots, \boldsymbol{\theta^{t-1}}$). Under certain conditions, an MCMC chain converges to a unique stationary distribution, in the limit as the number of iterations goes to infinity. This is true for any starting position $\boldsymbol{\theta^0}$, even if $\boldsymbol{\theta^0}$ comes from a low-probability region. The initial burn-in samples are usually discarded because they are highly dependent on the starting value $\boldsymbol{\theta^0}$ (Gelman et al., 2014). An implementation of the Metropolis-Hastings algorithm is given in Algorithm 2.3.1.

Algorithm 2.1. The Metropolis-Hastings Algorithm (Metropolis et al., 1953; Hastings, 1970)

1: Initialise $\boldsymbol{\theta}_0$.
2: **for** $t = 1$ *to* T **do**
3: Propose $\boldsymbol{\theta}^*$ from a proposal distribution $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}_{t-1})$.
4: Calculate $\alpha(\boldsymbol{\theta}_{t-1} \rightarrow \boldsymbol{\theta}^*) = \min\left(1, \frac{p(\mathbf{y}|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}_{t-1}|\boldsymbol{\theta}^*)}{p(\mathbf{y}|\boldsymbol{\theta}_{t-1})p(\boldsymbol{\theta}_{t-1})q(\boldsymbol{\theta}^*|\boldsymbol{\theta}_{t-1})}\right)$.
5: Set $\boldsymbol{\theta}_t = \boldsymbol{\theta}^*$ with probability α, else set $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1}$.
6: **end for**

In Algorithm 2.3.1 the choice of proposal distribution $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}_{t-1})$ can greatly impact the performance of the sampler (Roberts, Gelman, & Gilks, 1997; Van Ravenzwaaij et al., 2018). For example, $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}_{t-1})$ can be chosen to be

$\text{MVN}(\boldsymbol{\theta}^*|\boldsymbol{\theta}_{t-1}, \boldsymbol{\Sigma})$-the multivariate normal distribution with mean vector $\boldsymbol{\theta}_{t-1}$ and covariance matrix $\boldsymbol{\Sigma}$. However, if the component variances of the normal proposal distribution are very large, many proposals may fall in the low density region of the target distribution, which results in a high rejection rate. On the other hand, if the component variances are small, this may increase the time for an MCMC chain to converge to the target distribution or increase the chance of getting stuck in a local mode of the distribution (Van Ravenzwaaij et al., 2018). In practice, there is an adaptive approach that can tune the proposals in Algorithm 2.3.1 to maintain an acceptance rate between 0.3 and 0.5; more detail can be found in Gelman and Rubin (1992).

2.3.2 Hamiltonian Monte Carlo (HMC)

The HMC method adapts an idea from statistical mechanics that allows an MCMC chain to explore the parameter space more efficiently. This is achieved by introducing an auxiliary momentum variable $\boldsymbol{\phi}$ for the parameter vector $\boldsymbol{\theta}$, in which both have the same dimension, say M. Then, draws from a joint density

$$p(\boldsymbol{\phi}, \boldsymbol{\theta}) = p(\boldsymbol{\phi}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{2.7}$$

are sampled using a new Metropolis algorithm, where the proposal distribution for $\boldsymbol{\theta}$ is influenced by $\boldsymbol{\phi}$. In most applications of HMC, $\boldsymbol{\phi}$ is considered to have a multivariate normal distribution with zero mean and covariance matrix $\mathbf{C}$, which is commonly a diagonal matrix:

$$\boldsymbol{\phi} \sim \text{MVN}(\mathbf{0}, \mathbf{C}). \tag{2.8}$$

Thus, each ϕ_m is independent of the others and has a normal distribution,

$N(0, \mathbf{C}_{mm})$ for $m = 1, \ldots, M$. The joint density in Equation (2.7) can be used to define a fictitious Hamiltonian system (Hoffman & Gelman, 2014) as follows:

$$H(\boldsymbol{\phi}, \boldsymbol{\theta}) = -\log\, p(\boldsymbol{\phi}, \boldsymbol{\theta}) = -\log\, p(\boldsymbol{\phi}|\boldsymbol{\theta}) - \log\, p(\boldsymbol{\theta}) \tag{2.9}$$

where $\boldsymbol{\theta}$ denotes a particle's position in M-dimensional parameter space, $\boldsymbol{\phi}$ denotes the associated momentum of that particle in Mth-dimensional space, $-\log\, p(\boldsymbol{\phi}|\boldsymbol{\theta})$ is the kinetic energy and $-\log\, p(\boldsymbol{\theta})$ is the negative potential energy (Hoffman & Gelman, 2014). The simulation over time of the Hamiltonian dynamics can be performed using a leapfrog integrator, which is a numerical integration algorithm; for more details, I refer the interested reader to Brooks, Gelman, Jones, and Meng (2011). In this thesis I have used the software Stan (Stan Development Team, 2019) to implement the HMC algorithm. The basic steps of this implementation are as follows (Gelman et al., 2014):

1. Specify the initial value of the parameter vector $\boldsymbol{\theta}_0$.
2. Repeat steps 3-5 for $(t = 1, \ldots, T)$ times:
3. A new momentum vector, $\boldsymbol{\phi}_{t-1}$, is sampled from a multivariate normal distribution.
4. Apply leapfrog updates to the position and momentum variables $\boldsymbol{\theta}_{t-1}$ and $\boldsymbol{\phi}_{t-1}$ with discretization time d and number of steps L. This will generate a new proposal for the position-momentum pair $\boldsymbol{\theta}^*$ and $\boldsymbol{\phi}^*$.
5. Apply a Metropolis acceptance step with acceptance probability

$$\alpha = \min(1, \exp(H(\boldsymbol{\phi}_{t-1}, \boldsymbol{\theta}_{t-1}) - H(\boldsymbol{\phi}^*, \boldsymbol{\theta}^*))$$

to decide whether to accept or reject updating the new state to obtain $(\boldsymbol{\theta}^t, \boldsymbol{\phi}^t)$.

2.3.3 Limitations of Standard MCMC

Standard MCMC algorithms, such as the random walk Metropolis-Hastings (RWMH) sampler, explore the parameter space by making local moves using, for example, a normal proposal density (Chumbley et al., 2007). RWMH can converge slowly to the target posterior density when the number of parameters is large (Sengupta, Friston, & Penny, 2016; Feng & Li, 2015). In addition, unidentifiable parameters in nonlinear models may also cause slow convergence and mixing of MCMC algorithms (Kim & Li, 2012). Such poor mixing can greatly increase the computational burden of inference due to the need to explicitly solve the ODEs numerically for each proposal of the parameter vector of interest (Wang & Barber, 2014). Multimodality in the posterior distribution is often the reason MCMC methods have poor convergence or even fail to converge (Neal, 1993, 2012; Celeux, Hurn, & Robert, 2000; Neal, 2001; Rudoy & Wolfe, 2006; Sminchisescu & Welling, 2007; Craiu, Rosenthal, & Yang, 2009). When the posterior distribution is multi-modal, MCMC methods may fail to traverse low probability regions between modes (Lan, Streets, & Shahbaba, 2014). HMC also suffers from this problem because the probability of sampling a sufficiently large momentum to explore this low density region is very small (Levy, Hoffman, & Sohl-Dickstein, 2017). Although HMC is a more efficient and sophisticated method than MCMC, when dealing with discrete parameters, HMC dose not work. As a result, HMC requires calculating the gradient of the posterior distribution with respect to the parameters, which is undefined in the case of discrete parameters (Kruschke, 2014). Regardless of the various methods that have been proposed to tackle these problems, especially in the context of inverse problems for ODEs, there is still a need to develop new computational methods for Bayesian inference to address the problems of mixing and poor convergence in a broader context.

2.4 Quasi Monte Carlo (QMC)

In this section, I give a brief overview of QMC and randomised QMC (RQMC) methods. First, I discuss QMC integration and address some features of low-discrepancy point sets. Second, I describe the construction of the point sets that I have used in this thesis.

For an integrable function $f : [0,1)^M \to \mathbb{R}$, the basic Monte Carlo (MC) approximation to the integral over the $[0,1)^M$ hypercube is:

$$I = \int_{[0,1)^M} f(x)dx \qquad \text{and} \qquad \hat{I}_N = \frac{1}{N}\sum_{i=1}^{N} f(x_i) \tag{2.10}$$

where the points $x_i \sim \mathcal{U}([0,1)^M)$ and $\mathcal{U}(x)$ are the uniform distribution. MC methods depend on the random samples, and the error of this approximation is $\mathcal{O}(N^{-\frac{1}{2}})$ (Caflisch, 1998). This rate of convergence can be improved using QMC, which involves generating a deterministic sequence (called a low-discrepancy sequence) $x_1, \ldots, x_N$ such that the corresponding summation in Equation (2.10) converges faster than the Monte Carlo estimate. Specifically, the integration error when using QMC is bounded above by $\mathcal{O}(N^{-1}(\log N)^M)$. The enhanced rate of convergence is due to the fact that QMC sequences are distributed more evenly in the hypercube than a uniform random sequence (Dick, Kuo, & Sloan, 2013).

2.4.1 Discrepancy

To measure the evenness of the spread of the points, one can use the star discrepancy, which is the distance between the empirical distribution on x_i and the continuous uniform distribution on $[0,1)^M$ calculated by the Kolmogorov-Smirnoff statistic. The star discrepancy generalises the Kolmogorov-Smirnov distance, and

to define it we first consider

$$\Delta(a) = \text{Vol}([0,a]) - \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{x_i \in [0,a]\}$$

to be the local discrepancy function at a point $a \in [0,1)^M$ where $\text{Vol}(A)$ is the M-dimensional volume of the measurable set A, $[0,a]$ is the M-dimensional hyperrectangle that has a corner at the origin and an opposing corner at $a = (a_1, \ldots, a_M)$ and $\mathbb{1}\{x_i \in [0,a]\}$ is the indicator function. Thus the volume of $[0,a]$ is $\prod_{j=1}^{M} a_j$. Then, the star discrepancy of a point set $\mathcal{P}_N = \{x_1, \ldots, x_N\} \subset [0,1)^M$ is given by

$$D^*(\mathcal{P}_N) = \sup_{a \in [0,1)^M} \left| \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{x_i \in [0,a]\} - \prod_{j=1}^{M} a_j \right|. \tag{2.11}$$

When $D^* \to 0$, the sample mean $\hat{I}_N$ approaches the theoretical mean I given by the integral in Equation (2.10), by a deterministic version of the law of large numbers that applies for QMC (Owen & Tribble, 2005). The importance of the star discrepancy arises from the Koksma-Hlawka inequality (Hickernell, 2014), which states that for QMC integration, if the function f has variation $V(f)$ in the sense of Hardy and Krause (Hardy, 1905), then

$$|\hat{I}_N - I| \leq D^*(\mathcal{P}_N) V(f). \tag{2.12}$$

From this it follows that the integration error of QMC is bounded above by $\mathcal{O}(N^{-1}(\log N)^M)$. Many examples show that QMC integration outperforms MC integration even for a small number of points, N (Owen & Tribble, 2005; Buchholz & Chopin, 2019).

2.4.2 Low-Discrepancy Sequences

QMC sequences can be widely classified into two groups: digital nets and sequences, and lattice rules. In all numerical examples in this thesis, I have generated points using a Sobol sequence, which is a kind of digital net. One reason for choosing to work with Sobol sequence is the availability of software (an R package called randtoolbox) to generate such point sets.

The digital nets presented in this section are known as (t, M)-nets in base b, where t, M and b are integer parameters and M corresponds to the dimension of the space for a vector of points $\boldsymbol{x}$. For $b \geqslant 2$, an *elementary interval* in base b is a subinterval of $[0,1)^M$ of the form

$$E = \prod_{j=1}^{M} \left[\frac{c_j}{b^{k_j}}, \frac{c_{j+1}}{b^{k_j}} \right)$$

where k_j and c_j are some integers satisfying $k_j \geqslant 0$ and $0 \leqslant c_j < b^{k_j}$.

Definition 2.4.1 Let $m \geqslant t \geqslant 0$ be integers. The sequence $x_1, \ldots, x_{b^m} \in [0,1)^M$ is a (t, m, M)-net in base b if every elementary interval in base b of volume b^{t-m} contains exactly b^t points of the sequence.

Definition 2.4.2 For $t \geqslant 0$, the infinite sequence $x_1, x_2, \cdots \in [0,1)^M$ is a (t, M)-sequence in base b if for all $k \geqslant 0$ and $m \geqslant t$ the sequence $x_{kb^m, \ldots, (k+1)b^m}$ is a (t, m, M)-net in base b.
The concept of a (t, M)-sequence in base b is essential in constructing Sobol sequences and it was shown by Niederreiter (1992) that all (t, M)-sequences are low-discrepancy sequences.

Sobol Sequences

Sobol (1976) was the first to introduce the construction of (t, M)-sequences, which are now known as Sobol sequences when the base $b = 2$. One way to generate a Sobol sequence is to consider the primitive polynomials $p_1, \ldots, p_M \in \mathbb{F}_2$, where $\mathbb{F}_2 = \{0, 1\}$, are ordered according to degree, and let

$$p_j(x) = x^q + a_{1,j}x^{q-1} + \cdots + a_{q-1}x + 1$$

for $j = 1, \ldots, M$. Then, one chooses odd natural numbers $1 \leqslant m_{1,j}, \ldots, m_{q,j}$ such that $m_{k,j} < 2^k$ for $1 \leqslant k \leqslant q$, but for all $k > q$ define $m_{k,j}$ recursively by

$$m_{k,j} = 2a_{1,j}m_{k-1,j} \oplus \cdots \oplus 2^{q-1}a_{q-1}m_{k-q+1,j} \oplus m_{k-q,j}.$$

where $\oplus$ is the bit-by-bit operator. Then define the direction numbers as

$$v_{k,j} = \frac{m_{k,j}}{2^k}.$$

After that, for $n \in \mathbb{N}_0$, where $\mathbb{N}_0 = \{0, 1, 2, \ldots, \infty\}$, with base 2 expansion $n = n_0 + 2n_1 + \cdots + 2^{r-1}n_{r-1}$ we identify

$$x_{n,j} = n_0v_{1,j} \oplus n_1v_{2,j} \oplus \cdots \oplus n_{r-1}v_{r,j} \quad \textit{and} \quad x_n = (x_{n,1}, \ldots, x_{n,s}).$$

Then a Sobol sequence is the sequence of points $(x_n)_{n\in\mathbb{N}_0}$ (Leobacher & Pillichshammer, 2014). As noted above, these points are defined over a unit hypercube, but one can use a linear transformation to define them over a hypercube with different bounds.

2.4.3 Randomization

Using QMC sequences is inconvenient for constructing estimators because the point construction is deterministic. However, randomness can be reintroduced through scrambling and other related randomization methods (Chi & Mascagni, 2007; Vandewoestyne, Chi, Mascagni, & Cools, 2007). Then, these sequences are known as randomized QMC (RQMC) sequences, and the estimator $\hat{I}_N = \frac{1}{N}\sum_{i=1}^{N} f(x_i)$ is an unbiased estimate of the integral in Equation (2.10) (Buchholz & Chopin, 2019; Wenzel, Buchholz, & Mandt, 2018). To obtain an RQMC sequence, we use a method called scrambled nets, which was introduced by Owen et al. (1997) and then modified in Owen et al. (2008). In this method, for $b \geqslant 2$, let $A_j = a_{j1}b^{-1} + a_{j2}b^{-2} + \ldots$, where $a_{j1}, a_{j2}, \cdots \in \{0, 1, \ldots, b-1\}$ and $j = 1, \ldots, n$. Then, for each value of a_{j1}, there is a permutation $\pi_{a_{j1}}$ of $\{0, 1, \ldots, b-1\}$ that can define a second digit as $\pi_{a_{j1}}(a_{j2})$. This process can be continued to define more digits as $\pi_{a_{j1},a_{j2}}(a_{j3})$, $\pi_{a_{j1},a_{j2},a_{j3}}(a_{j4}), \ldots$. Each permutation is considered to be uniformly distributed, and they are mutually independent (Owen, 1998). Next, let $X = (X_1, X_2, \ldots, X_n)$ be a (t, m, M)-net and define that $X_j = \sum_{k=1}^{\infty} x_{jk}b^k$, for $j = 1, \ldots, n$. Then, the scrambled X is obtained by taking x_{jk} to be the permutations of the digits $a_{j1}, a_{j2}, \ldots$ as follows:

$$x_{jk} = \pi_{a_{j1},a_{j2},\ldots,a_{j(k-1)}}(a_{jk}).$$

2.5 Approximate Bayesian Computation (ABC)

ABC approaches give alternative parameter estimation methods when the likelihood is difficult or impossible to evaluate or too computationally costly. ABC methods were first introduced by Pritchard et al. (1999) in the form of the ABC rejection sampler. This approach involves replacing the calculation of the likeli-

hood with a comparison between the observed data, $\mathbf{y}$, and simulated data, $\mathbf{z}$. The simplest ABC algorithm involves first sampling a parameter, $\boldsymbol{\theta}^*$, from the prior, $p(\boldsymbol{\theta})$, and then using a generative model to simulate a data set, $\mathbf{z} \sim f(\mathbf{z}|\boldsymbol{\theta}^*)$. Proposals of the parameter $\boldsymbol{\theta}^*$ are accepted if the distance between the observation and the simulated data sets falls below a chosen threshold, ϵ, where distance is quantified using a discrepancy function $\rho(\mathbf{z}, \mathbf{y})$. This procedure is repeated until a desired number of samples is accepted. A comprehensive reference for ABC methods can be found in Sisson, Fan, and Beaumont (2018). The ABC rejection algorithm is detailed in Algorithm 2.2.

Algorithm 2.2. ABC Rejection Algorithm, pre-specification of ϵ

1: **while** number of accepted $\boldsymbol{\theta}^* < N$ **do**
2: Draw $\boldsymbol{\theta}^* \sim p(\boldsymbol{\theta})$.
3: Simulate $\mathbf{z}^*$ from model given $\boldsymbol{\theta}^*$.
4: **if** $\rho(\mathbf{z}^*, \mathbf{y}) \leq \epsilon$ **then**
5: Accept $\boldsymbol{\theta}^*$.
6: **end if**
7: **end while**

Alternatively, ϵ could be specified after drawing T samples from the prior as demonstrated in Algorithm 2.3. A constant, $0 < \alpha < 1$, defined to represent the percentage of draws that are to be accepted, is used to set ϵ. Although this scheme provides flexibility in terms of trading accuracy for speed, it can require large storage requirements if T is large.

Algorithm 2.3. ABC Rejection Algorithm, post-determination of ϵ

1: **for** $t = 1$ *to* T **do**
2: Draw $\boldsymbol{\theta}_t \sim p(\boldsymbol{\theta})$ and simulate $\mathbf{z}_t$ from model given $\boldsymbol{\theta}_t$.
3: Compute discrepancy function $\rho_t = \rho(\mathbf{z}_t, \mathbf{y})$.
4: **end for**
5: Sort $\{\boldsymbol{\theta}_t, \rho_t\}_{t=1}^T$ into ascending order, based on ρ.
6: Keep $N = \alpha T$ of $\boldsymbol{\theta}_t$ with the lowest discrepancy, hence defining ϵ.

ABC targets an approximate posterior (Frazier, Martin, Robert, & Rousseau,

2018):

$$p_\epsilon(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y}) \propto \mathbb{1}(\rho(\mathbf{z}, \mathbf{y}) \leq \epsilon) p(\boldsymbol{\theta}) f(\mathbf{z}|\boldsymbol{\theta}), \tag{2.13}$$

where $\mathbb{1}$ is an indicator function that takes the value one if its logical argument is true and zero otherwise and $f(\mathbf{z}|\boldsymbol{\theta})$ is the model that generates new simulations. The accuracy of ABC approaches depends on choosing a suitable discrepancy function $\rho(\mathbf{z}, \mathbf{y})$ and an appropriate tolerance ϵ (Marjoram, Molitor, Plagnol, & Tavare, 2003). In practice, the discrepancy function typically compares sets of summary statistics $s(\cdot)$ for the observed and simulated datasets.

2.5.1 MCMC ABC

ABC rejection sampling can suffer from extremely low acceptance rates when the prior and posterior distributions are quite different (Marjoram et al., 2003). To overcome this deficiency, a more efficient ABC technique based on MCMC was developed (Marjoram et al., 2003). The implementation of an early rejection step (Picchini & Forman, 2014) can improve the efficiency of the method since data is only simulated under the model when necessary, as shown in Algorithm 2.4. MCMC ABC is motivated by a desire to keep proposals for $\boldsymbol{\theta}$ within non-negligible posterior regions. MCMC ABC aims to combat the heavy storage requirements of the ABC rejection sampler, while allowing efficient exploration of the parameter space. The approach strongly depends on the selection of the proposal distribution q, prior distribution and ϵ.

Algorithm 2.4. MCMC ABC with Early Rejection (Picchini & Forman, 2014)

1: Obtain $\boldsymbol{\theta}_0$ and $\mathbf{z_0}$ using ABC rejection sampling.
2: **for** $t = 1$ *to* T **do**
3: Draw $\boldsymbol{\theta}^* \sim q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{t-1})$.
4: Compute $r = \frac{p(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}_{t-1}|\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}_{t-1})q(\boldsymbol{\theta}^*|\boldsymbol{\theta}_{t-1})}$.
5: **if** $\mathcal{U}(0,1) < r$ **then**
6: Simulate $\mathbf{z}^*$ from model given $\boldsymbol{\theta}^*$.
7: **if** $\mathbb{1}\{\rho(\mathbf{z}^*, \mathbf{y}) \leq \epsilon\}$ **then**
8: Set $\boldsymbol{\theta}_t = \boldsymbol{\theta}^*$ else, set $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1}$.
9: **end if**
10: **else**
11: $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1}$.
12: **end if**
13: **end for**

MCMC ABC tolerance

The selection of tolerance level plays an important role in the efficiency of MCMC ABC, and choosing an appropriate ϵ is problematic. In this thesis, in order to select an appropriate tolerance level, I adopted the method of Vaart et al. (2018), which is to solve the ODE model with different proposals of the parameters from the priors, find all the distances between these solutions and the observed data and then choose the one that minimises this distance. Next, I used the best-fitting solution to estimate the value of ϵ, and used it as a guide to manually adjust the value of the tolerance to maintain an appropriate acceptance rate. This strategy addresses one of the main drawbacks of MCMC ABC, which is that choosing a very small ϵ increases the rejection rate, which in turn causes poor mixing of the Markov chain (M. A. Beaumont, 2010).

2.5.2 SMC ABC

In order to improve the low acceptance rate in the basic ABC algorithm, an SMC ABC algorithm was proposed in Sisson et al. (2007), based on the SMC sampler methodology developed by Del Moral et al. (2006). This algorithm requires identifying a sequence of tolerance values, $\epsilon_t > \epsilon_{t+1}$, where $t = 1, \ldots, T$ is the identification number of the population, instead of using a single fixed tolerance value as in MCMC ABC. Here T represents the number of population in the SMC ABC, not the number of draws as in previous ABC approaches. The algorithm is a type of particle filter. Particles are sampled from the prior distribution in the first population and then by using an acceptance criteria, these particles are filtered throughout a series of intermediate distributions. This process can be continued until the resulting final population is considered an approximate sample of the posterior distribution (see Figure 2.2). An implementation of the SMC ABC algorithm is given in Algorithm 2.5.

The efficiency of the SMC ABC algorithm depends not only on the model complexity and the amount of data available, but also on the choice of the decreasing sequence of ϵ_t (the tolerances), and the choice of perturbation kernel K_t, according to Filippi, Barnes, Cornebise, and Stumpf (2013).

Tolerance

There are various ways to construct the decreasing sequence of ϵ_t, called the tolerance sequence, either manually or adaptively as proposed in Drovandi and Pettitt (2011) and Del Moral, Doucet, and Jasra (2012). In the adaptive method, the value of ϵ_1 is chosen to be large to avoid having a low acceptance rate. Then, each ϵ_t, for $t = 2, ..., T$, is selected to be the α-th quantile of the discrepancies between the observed data and the simulated data that was generated in the

Algorithm 2.5. SMC ABC Algorithm (Del Moral et al., 2006; Sisson et al., 2007; Toni et al., 2009)

1: Initialise $\epsilon_t \geq 0$ for $t = 1, ..., T$ where $\epsilon_t > \epsilon_{t+1} > 0$.
2: **for** $t = 0$ *to* T **do**
3: **for** $i = 1$ *to* N **do**
4: **if** $t = 0$ **then**
5: Sample $\boldsymbol{\theta}^{**}$ from $p(\boldsymbol{\theta})$.
6: **else**
7: Sample $\boldsymbol{\theta}^{*}$ from the previous population $\boldsymbol{\theta}_{t-1}^{(i)}$ with normalised weights $w_{t-1}^{(i)}$ and use a perturbation kernel K_t to sample $\boldsymbol{\theta}^{**} \sim K_t(\cdot|\boldsymbol{\theta}^{*})$.
8: **end if**
9: **if** $p(\boldsymbol{\theta}^{**}) = 0$ **then**
10: Go to line 4.
11: **else**
12: Simulate $\mathbf{z}^{*}$ from model given $\boldsymbol{\theta}^{**}$.
13: **end if**
14: **if** $\rho(\mathbf{z}^{*}, \mathbf{y}) \geq \epsilon_t$ **then**
15: Go to line 4.
16: **else**
17: Set $\boldsymbol{\theta}_t^{(i)} = \boldsymbol{\theta}^{**}$ and calculate the weight for the particle $\boldsymbol{\theta}_t^{(i)}$:

$$w_t^{(i)} = \begin{cases} 1, & \text{if } t = 0, \\ \frac{p(\boldsymbol{\theta}_t^{(i)})}{\sum_{j=1}^{N} w_{t-1}^{(j)} K_t(\boldsymbol{\theta}_t^{(i)}|\boldsymbol{\theta}_{t-1}^{(j)})} & \text{if } t > 0. \end{cases}$$

18: **end if**
19: **end for**
20: set ϵ_{t+1} to be α-quantile of saved distances vector
21: Normalise the weights.
22: **end for**
23: Return particles $\boldsymbol{\theta}_T^{(i)}$.

$(t-1)^{th}$ population, as described in Algorithm 2.5, where $0 \leq \alpha \leq 1$. Throughout this thesis, I have used the latter method of selecting the sequence of tolerance thresholds with $\alpha = 0.1$, and I have stopped the algorithm when it reached a final ϵ_t that achieved the desired final agreement between simulated and real data (Liepe et al., 2014).

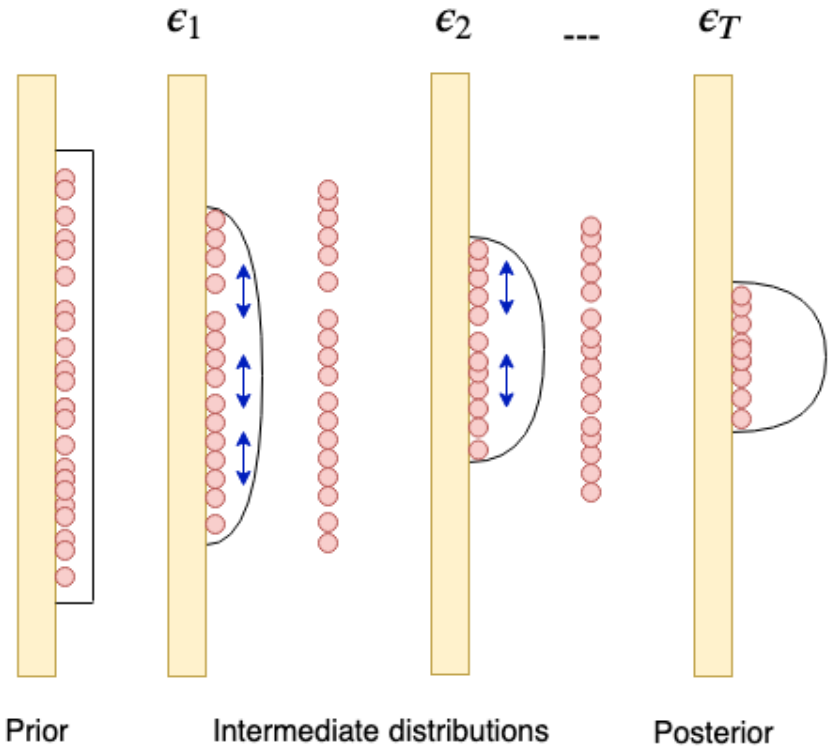

Figure 2.2: Schematic representation of SMC ABC algorithm shows that the probability of samples taken first from the prior is updated throughout the process via filtering steps. The black curves represent the intermediate distribution and the final population represents an approximation of the posterior distribution (Toni et al., 2009).

Stopping criteria

In practice, when implementing Algorithm 2.5, stopping criteria are required to terminate the process. One simple way is to stop the algorithm when it reaches a specified ϵ_T close to zero or achieves a target acceptance rate, which is calculated in each iteration by finding the ratio of the number of accepted particles to the total number of simulations needed (Abdessalem, Dervilis, Wagg, & Worden, 2018). Another termination criterion is to use a specified number of total simulations as a tuning parameter (Prangle et al., 2017).

Perturbation kernel

The choice of perturbation kernel affects the acceptance rate in SMC ABC and the time consumed by the algorithm as explained in Filippi et al. (2013). Perturbation kernels can be divided into two classes: component-wise perturbation kernels and multivariate perturbation kernels. For component-wise perturbation kernels, one can use a uniform distribution or a univariate Gaussian distribution to perturb the particle $\boldsymbol{\theta}^*$ sampled from the previous population $\{\boldsymbol{\theta}_{t-1}^{(i)}\}_{i=1}^{N}$. The standard deviation of the kernel can be fixed in advance for each population, but more recently, practitioners have been adaptively choosing the width of the kernel. For a detailed discussion, the reader is referred to M. Beaumont, Cornuet, Marin, and Rober (2009); Didelot, Everitt, Johansen, Lawson, et al. (2011); Filippi et al. (2013).

2.5.3 Limitations of SMC ABC

ABC methods have been extensively developed, since their first appearance, in terms of their theoretical foundations and practical methodology. For example, Barber, Voss, Webster, et al. (2015) provided a theoretical foundation for understanding the convergence properties of the ABC method in general and Li and Fearnhead (2018) and Frazier et al. (2018) investigated the asymptotic behaviour of the posterior distribution of ABC and its expectation as the number of observations increases.

According to Frazier et al. (2018), the main and most important problem of the ABC approach is that the approximated ABC posterior is exact only when the tolerance $\epsilon = 0$, which in the case of ABC applied to ODEs is, for many existing implementations, not practical or reasonable for reasons discussed in detail in Chapter 3. However, appropriate handling of the error term in the ABC algorithm may overcome this limitation, and work inspired by Wilkinson (2013) in Vaart et al. (2018) argues that the acceptance of the proposed parameters should be with respect to the error term rather than with respect to some tolerance level.

The SMC ABC algorithm applied to the inverse problem in non-linear ODEs was presented in Toni and Stumpf (2009). Three different suggestions for implementing the SMC ABC algorithm appeared in the thesis by Toni (2010) as below, where in the notation of Toni (2010), σ is the standard deviation of additive, normally distributed noise and $\boldsymbol{\theta}$ is the parameter vector that needs to be inferred:

1. No noise added in the simulation step, $\boldsymbol{\theta}$ is unknown parameter. This framework has been introduced in Chapter 3 and used throughout this thesis.

2. Noise added in the simulation step, σ is considered known and $\boldsymbol{\theta}$ unknown.

3. Noise added in the simulation step, both σ and $\boldsymbol{\theta}$ are unknown.

However, using the three approaches above leads to misspecified inference results as follows:

1. The implementation of SMC ABC with the first suggestion above without adding an error term leads to the data generation model being misspecified. This occurs because the simulated data is deterministic in the sense that it is the unique solution to a system of ODEs. Consequently, the underlying likelihood distribution is a point mass concentrated at the solution of the system as I outline in more detail in Chapter 3. Thus, this approach fails to correctly characterise the uncertainty represented by the posterior distribution.

2. The second suggestion is to implement the SMC ABC approach assuming that the noise variance σ is known, as exemplified by Silk et al. (2013). In this method, a noisy simulation $\mathbf{y}_{sim}$ is generated by solving the ODEs and then adding noise sampled from an assumed known error model. When this assumed error model is a good approximation to the actual noisy observation process, the final approximate ABC posterior may be a good approximation. However, in real world problems, assuming a known noise variance is unrealistic.

3. In the third suggestion above, the inference of the error term σ is carried out by proposing values and accepting or rejecting them, exactly as for the other parameters being estimated. Such approaches encounter two problems. First, the residual vector has high dimension, which can dramatically reduce the acceptance probability of proposed values. In addition, it is difficult to define the distance function $\rho(\mathbf{y}_k, \mathbf{y}_{\boldsymbol{sim}})$ in such a way as to accurately reflect the noise associated with the data. Appropriately accounting for knowledge about the error terms when choosing ϵ_t remains a major challenge, which we address in Chapter 4.

Regardless of the limitations of SMC ABC, the algorithm presented in Toni and Stumpf (2009) gives a useful alternative to exact Bayesian methods for inference in ODE. This likelihood-free method is simple to apply, and the sampler produces uncorrelated samples. Furthermore, the SMC ABC algorithm does not suffer from the poor mixing problems that cause issues for MCMC and ABC based on MCMC.

Chapter 3

A comparison of approximate versus exact techniques for Bayesian parameter inference in nonlinear ordinary differential equation models

Preamble

The purpose of this chapter is to investigate the ABC method that is commonly used to estimate parameters in ODE models. We satisfy Objective 1 of this thesis by demonstrating that several popular ABC approaches fail to adequately model the error associated with observations that have been described by ODEs. In the current implementation of ABC approaches, the acceptance probability depends on the choice of the discrepancy function and the tolerance without any consideration of the error term. We observe that the so-called posterior distributions derived from such methods do not accurately reflect the epistemic uncertainties in parameter values. Moreover, our findings confirm that these methods provide minimal computational advantages over exact Bayesian methods. Applications to two ODE epidemiological models with simulated data and the other with real data contributes to Objective 4 of this thesis.

ROYAL SOCIETY OPEN SCIENCE

royalsocietypublishing.org/journal/rsos

Research

Cite this article: Alahmadi AA, Flegg JA, Cochrane DG, Drovandi CC, Keith JM. 2020 A comparison of approximate versus exact techniques for Bayesian parameter inference in nonlinear ordinary differential equation models. *R. Soc. open sci.* **7**: 191315.
http://dx.doi.org/10.1098/rsos.191315

Received: 8 August 2019
Accepted: 27 January 2020

Subject Category:
Mathematics

Subject Areas:
statistics/health and disease and epidemiology/differential equations

Keywords:
Bayesian inference, Markov chain Monte Carlo, approximate Bayesian computation, compartmental models, ordinary differential equations, epidemiology models

Author for correspondence:
Amani A. Alahmadi
e-mail: amani.alahmadi@monash.edu

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.4880280.

THE ROYAL SOCIETY PUBLISHING

A comparison of approximate versus exact techniques for Bayesian parameter inference in nonlinear ordinary differential equation models

Amani A. Alahmadi[1,4], Jennifer A. Flegg[2], Davis G. Cochrane[1], Christopher C. Drovandi[3] and Jonathan M. Keith[1]

[1]School of Mathematics, Monash University, Clayton, Victoria, Australia
[2]School of Mathematics and Statistics, University of Melbourne, Parkville, Victoria, Australia
[3]School of Mathematical Sciences, Queensland University of Technology, Brisbane, Queensland, Australia
[4]College of Science and Humanities, Shaqra University, Shaqra, Saudi Arabia

AAA, 0000-0001-5359-1340; JAF, 0000-0002-8809-726X; CCD, 0000-0001-9222-8763; JMK, 0000-0002-9675-3976

The behaviour of many processes in science and engineering can be accurately described by dynamical system models consisting of a set of ordinary differential equations (ODEs). Often these models have several unknown parameters that are difficult to estimate from experimental data, in which case Bayesian inference can be a useful tool. In principle, exact Bayesian inference using Markov chain Monte Carlo (MCMC) techniques is possible; however, in practice, such methods may suffer from slow convergence and poor mixing. To address this problem, several approaches based on approximate Bayesian computation (ABC) have been introduced, including Markov chain Monte Carlo ABC (MCMC ABC) and sequential Monte Carlo ABC (SMC ABC). While the system of ODEs describes the underlying process that generates the data, the observed measurements invariably include errors. In this paper, we argue that several popular ABC approaches fail to adequately model these errors because the acceptance probability depends on the choice of the discrepancy function and the tolerance without any consideration of the error term. We observe that the so-called posterior distributions derived from such methods do not accurately reflect the epistemic uncertainties in parameter values. Moreover, we demonstrate that these methods provide minimal computational advantages

over exact Bayesian methods when applied to two ODE epidemiological models with simulated data and one with real data concerning malaria transmission in Afghanistan.

1. Introduction

Models of dynamical systems consisting of sets of ordinary differential equations (ODEs) are an essential tool to describe many processes in science and engineering. ODE models contain parameters such as kinetic rates and initial concentrations. However, these parameters often cannot be measured directly by experiments, or there is inherent uncertainty in the parameter values. As such, these parameter values need to be estimated using statistical techniques such as maximum likelihood estimation or Bayesian inference. In the last decade much research has focused on estimating the unknown parameters of ODE systems under a Bayesian framework. One reason is that the Bayesian approach provides appropriate quantification of the uncertainty of parameters (and hence model predictions) through the posterior distribution.

Markov chain Monte Carlo (MCMC) techniques were first developed by Metropolis *et al.* [1]. This class of technique creates a Markov chain which has the posterior distribution as its limiting distribution. The state of the chain after a number of steps is used as a sample from the posterior distribution and the quality of this sample improves as the number of steps gets larger. The original algorithm proposed by Metropolis was generalized by Hastings [2] to give the Metropolis–Hastings algorithm.

Exact Bayesian inference techniques have grown steadily more sophisticated over time, increasing the efficiency and complexity of sampling schemes. The modern Bayesian toolbox now includes schemes such as sequential Monte Carlo (SMC) [3], the Metropolis adjusted Langevin algorithm (MALA) [4] and hybrid (Hamiltonian) Monte-Carlo (HMC) [5,6]. These schemes improve the Metropolis–Hastings algorithm, enabling efficient sampling from high dimensional, strongly correlated posterior distributions.

However, there are many models that possess a computationally intractable likelihood function, ruling out exact Bayesian methods. This has led to the development of approximate Bayesian computation (ABC). The ABC methodology first appeared as the ABC rejection algorithm [7] which avoids calculation of the likelihood function. The theory was generalized and substantiated by Beaumont *et al*. [8]. To obtain samples more efficiently, a MCMC approach to ABC was formulated by Marjoram *et al.* [9].

In the context of dynamical systems, both approximate and exact Bayesian techniques involve numerical solution of the set of ODEs for each proposed set of parameters in order to evaluate how well the numerical solution matches the observed data. A desire to avoid the computational costs associated with numerical solution of the ODEs has led to the development of Gaussian Process (GP) models [10–12] for ODE parameter inference.

Dass *et al.* [13] proposed a two-step method to approximate the posterior distribution of unknown parameters in an ODE model. In the first step, data are generated from the ODE using a numerical method and then the second step uses the Laplace approximation to marginalize the posterior for each parameter. This method gives a fast approach compared to a full Bayesian computational scheme.

ABC methods based on SMC have been proposed [14] and many authors have developed approaches to improve the performance of the SMC ABC algorithm (see for example Beaumont *et al*. [15]). An SMC ABC approach was developed by Toni *et al.* [16], with application to dynamical systems. Their algorithm is theoretically sound, but we question the validity of the Bayesian posteriors they produce when they apply ABC to several examples involving ODE models. The authors apply ABC where they take the observed data as synthetically generated, where the ODE model is solved at an assumed true parameter value and measurement error added. However, when 'simulating' data in their ABC procedure, the ODE model is solved only, without generating measurement error. In this paper, we show that such an approach generates parameter distributions that are sensitive to the ABC tolerance, and will eventually converge onto a point mass if the tolerance is continually reduced. Thus this approach fails to correctly characterize the uncertainty as a Bayesian approach would aim to do.

In order to 'correctly' apply ABC to ODE models, one must simulate from the assumed measurement error model after solving the ODE. However, we also show in this paper that an exact Bayesian approach is more computationally efficient than this 'correct' ABC implementation, questioning the need for considering ABC in the first place when attempting to estimate the posterior distribution for ODE models.

Given that the Toni *et al.* [16] paper is highly cited, we are concerned that other researchers might follow their ABC approach for calibrating ODE models. For example, Gupta *et al.* [17] compared the performance of MCMC, parallel tempering (PT) and SMC ABC (using the ABC Sys-Bio package) in

estimating parameters in ODE models. The authors analysed simulated data with measurement error added, taken from the ABC-SysBio package. Then when applying MCMC and PT to infer model parameters, they assumed 'a likelihood function with 1% Gaussian error'. However, when using SMC ABC, no noise was added to the simulated data. Consequently, not only does this make the comparison invalid, but also the resulting approximate posterior distribution produced by SMC ABC does not represent the uncertainty around the parameter values. In another example, Silk *et al.* [18] present applications to molecular dynamical systems in which they 'have focused on the sequential ABC algorithm proposed by Toni *et al.* [16]'. Silk *et al.* [18] mention that they simulate the model 'subject to some small added zero-mean Gaussian noise with covariance $0.01I$' so they have clearly used 'noise added in the simulation step, σ is considered known' (option 2 from Toni's thesis [19, p. 154]). However, for many real problems, this is not practical and we fear that ABC users might revert to the option of simulating without noise. Two other examples of assuming known noise are in da Costa *et al.* [20] and Costa *et al.* [21]. The authors assumed the uncertainties to be 'additive, uncorrelated, Gaussian, with zero mean and' a known standard deviation, as they stated on pages 2801 and 1295, respectively. Moreover, Barnes *et al.* [22] presented an implementation of ABC SMC for ODEs (section 4.1 of their paper) and used the SysBio package. We show in our paper that using this package with ODE models can give an incorrect approximation to the posterior when not considering estimation of the noise. There is no explanation in Barnes *et al.* [22] regarding the authors' assumption about the noise. The same issue appeared in Toni & Stumpf [23] and Sun *et al.* [24]: they applied SMC ABC for an ODE model, but there are no details regarding the authors' assumptions about the noise. Understanding the overall noise (uncertainty) associated with the unknown parameter values when conducting parameter estimation using ODE models is important, especially when we aim to use these ODE models to inform real-world applications.

The remainder of this paper is organized as follows. In §2, we introduce a simple method of exact Bayesian inference and two methods of approximate Bayesian computation (MCMC ABC and SMC ABC), complemented with a discussion on the approximation to a point mass that results from SMC ABC and MCMC ABC. Application of MCMC, MCMC ABC and SMC ABC to two ODE epidemiological models with simulated data and one with real epidemiological data are presented in §3. Section 4 presents further discussion, comparison of the presented methods and our conclusions.

2. Bayesian techniques for ODE parameter inference

Bayesian techniques such as Markov chain Monte Carlo (MCMC) methodologies are sampling-based methods that involves sampling the posterior density

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}), \tag{2.1}$$

or an approximation to equation (2.1) in the case of approximate Bayesian computation (ABC) approaches, to calculate the desired density, where $\mathbf{y} = (y_1, \ldots, y_n)$ is the observed data, $\boldsymbol{\theta}$ are the unknown parameters, $p(\boldsymbol{\theta} \mid \mathbf{y})$ is the posterior distribution, $p(\mathbf{y} \mid \boldsymbol{\theta})$ is the likelihood and $p(\boldsymbol{\theta})$ is the prior. In this section we discuss application of these Bayesian frameworks in the context of inferring parameters for ODE models.

MCMC techniques as developed by Metropolis *et al.* [1] and Hastings [2] can be used to sample from the posterior distribution in equation (2.1). The Metropolis–Hastings algorithm constructs a Markov chain for which the stationary and limiting distribution is the posterior distribution. After running the chain for a sufficient amount of time,[1] samples from the chain can be considered draws from the posterior distribution. An implementation of the Metropolis–Hastings algorithm is given in appendix A. However, MCMC methods require the computation of the likelihood function, $p(\mathbf{y} \mid \boldsymbol{\theta})$, in equation (2.1). As a result, ABC methods were developed to sample from an approximation to the posterior in cases for which the likelihood is intractable or too computationally costly to compute. Instead of calculating the likelihood as before, a distance between the observed data, $\mathbf{y}$, and simulated data, $\mathbf{z}$, is calculated and for sufficiently small distance the parameter proposals are accepted. For more explanation see appendix A. ABC targets an approximate posterior [26]:

$$p_\epsilon(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y}) \propto \mathbb{1}(\rho(\mathbf{z}, \mathbf{y}) \leq \epsilon)p(\boldsymbol{\theta})f(\mathbf{z}|\boldsymbol{\theta}), \tag{2.2}$$

[1]Sufficient time in the context of MCMC can be taken to mean that the chain is close to convergence. In practice this is often assessed by checking that multiple chains produce a Gelman–Rubin diagnostic less than 1.05 [25].

where $\mathbb{1}$ is an indicator function that takes the value one if its logical argument is true and zero otherwise and $f(\mathbf{z} \mid \boldsymbol{\theta})$ is the model that generates simulations $\mathbf{z}$ giving $\boldsymbol{\theta}$. The accuracy of ABC approaches depends on choosing a suitable discrepancy function $\rho(\mathbf{z}, \mathbf{y})$ and an appropriate tolerance $\boldsymbol{\epsilon}$ [9]. In practice, the discrepancy function typically compares sets of summary statistics $s(\cdot)$ for the observed and simulated datasets. ABC rejection sampling is very simple to implement, though it can suffer from extremely low acceptance rates when the prior distribution is dissimilar to the posterior distribution [9]. To counteract this deficiency, a more efficient ABC technique based on MCMC was developed [9]. For more details see appendix A. Furthermore, in order to improve the low acceptance rate in the basic ABC algorithm, an SMC ABC algorithm was proposed in Sisson *et al.* [14], based on the SMC sampler methodology developed by Del Moral *et al.* [27]. The SMC ABC algorithm converges to the approximate posterior distribution through a number of intermediate distributions with a distance threshold that is sequentially decreased, see appendix A. The efficiency of the SMC ABC algorithm depends not only on the model complexity and the amount of data available, but also on the choice of the decreasing sequence of $\boldsymbol{\epsilon}_t$ (the tolerances), and the choice of perturbation kernel K_t, according to Filippi *et al.* [28]. There are various ways to construct the decreasing sequence of $\boldsymbol{\epsilon}_t$, either manually or adaptively as proposed in Drovandi & Pettitt [29] and Del Moral *et al.* [30]. In the adaptive method, the value of $\boldsymbol{\epsilon}_t$ is chosen to be the αth quantile of the discrepancies between the observed data and the simulated data that was generated in the $(t-1)$th population (see appendix A), where $0 \leq \alpha \leq 1$. In this paper, we used the latter method of selecting the sequence of tolerance thresholds and we stopped the algorithm when we reached a final $\boldsymbol{\epsilon}_t$ that setting the desired final agreement between simulated and real data Liepe *et al.* [31].

The choice of perturbation kernel affects the acceptance rate in SMC ABC and the time consumed by the algorithm as explained in Filippi *et al.* [28]. Perturbation kernels can be divided into two classes: component-wise perturbation kernels and multivariate perturbation kernels. For component-wise perturbation kernels, one can use a uniform distribution or a univariate Gaussian distribution to perturb the particle $\boldsymbol{\theta}^*$ sampled from the previous population $\{\boldsymbol{\theta}_{t-1}^{(i)}\}_{i=1}^{N}$. The standard deviation of the kernel can be fixed in advance for each population, but more recently practitioners are adaptively choosing the width of the kernel (Beaumont *et al.* [15], Didelot *et al.* [32], Filippi *et al.* [28]).

If the model parameters are correlated, a component-wise perturbation kernel can fail to capture the structure of the true posterior, leading to a low acceptance rate. To overcome this problem, a multivariate normal distribution with a covariance matrix $\boldsymbol{\Sigma}^{(t)}$ that depends on the covariance of the previous populations can be used to perturb the particles [28]:

$$\boldsymbol{\Sigma}^{(t)} = \sum_{i=1}^{N} \sum_{k=1}^{N_0} w_{t-1}^{(i)} \hat{w}^{(k)} (\hat{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}_{t-1}^{(i)})(\hat{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}_{t-1}^{(i)})^T, \tag{2.3}$$

where $\{\hat{\boldsymbol{\theta}}^{(k)}\}_{1 \leq k \leq N_0}$ are the particles from the previous populations for which the corresponding simulated data $\boldsymbol{z}^{(k)}$ satisfy $\rho(\boldsymbol{z}^{(k)}, \boldsymbol{y}) < \boldsymbol{\epsilon}_t$ (remembering $\boldsymbol{\epsilon}_t < \boldsymbol{\epsilon}_{t-1}$) and $\hat{w}^{(k)}$ are the associated weights.

To further improve the performance of SMC ABC, we adopted a method proposed in Prangle [33] to adaptively update the discrepancy function, $\rho(\mathbf{z}, \mathbf{y})$. We used a weighted Euclidean distance function:

$$\rho(\mathbf{z}, \mathbf{y}) = \sum_{j=1}^{n} \left(\frac{z_j - y_j}{\zeta_j} \right)^2, \tag{2.4}$$

where y_j is the jth observation, z_j is the jth simulated observation in the simulated data $\boldsymbol{z} = (z_1, \ldots, z_n)$ and ζ_j is a tunable scaling factor that allows the contribution to the discrepancy function of the jth coordinate to be normalized. The reason for normalizing the coordinates is to prevent any of them dominating the acceptance decision in the algorithm. In non-adaptive methods, the values of ζ_j are determined in advance and fixed. Fixing ζ_j from the first iteration in SMC ABC will not guarantee that the jth coordinate will be normalized in later iterations because in SMC ABC after the first round we are not sampling from the prior so the scale to normalize needs to be adapted.

To adapt the values of ζ_j in each iteration, Prangle [33] proposed calculating the median absolute deviation (MAD) of the jth coordinate of the simulated data vectors from the previous iteration (including those rejected). The value of the next $\boldsymbol{\epsilon}_t$ is also determined using these distances; for more details see algorithm 4 in [33]. Note that Prangle [33] defined the discrepancy function in terms of summary statistics for $\boldsymbol{z}$ and $\boldsymbol{y}$, as is usual in ABC. Here we have used the coordinates of $\boldsymbol{z}$ and $\boldsymbol{y}$ directly, following the approach of Toni *et al.* [16] for inference of ODE model parameters.

Vaart *et al*. [34] proposed an ABC method called *error-calibrated ABC* that implements a general methodology introduced by Wilkinson [35]. In their method, they incorporated the estimation of the noise into the ABC technique by identifying an ABC acceptance probability in which the noise is assumed to be normally distributed and independent. This results in improved estimates of the parameter values and their uncertainty. An implementation of the error-calibrated ABC algorithm is given in algorithm 1.

Algorithm 1. Error-calibrated ABC algorithm, Vaart *et al.* [34].

1: Repeat n times:.
   (a) Draw $\theta^* \sim p(\theta)$.
   (b) Simulate $\mathbf{z}^*$ from model given θ^*.
2: Find $\hat{\mathbf{z}}$, the simulated value that minimizes $\rho(\mathbf{z}, \mathbf{y})$.
3: For each data type k, calculate $\hat{\lambda}_k$, the standard deviation of all corresponding $\hat{\mathbf{z}}_n - \mathbf{y}$.
4: Accept $(\theta^*, \mathbf{z}^*)$ with probability $p_{\chi^2_{\mathbf{l}}}(\mathbf{s})\mathbf{s}^{1-\frac{1}{2}}/\mathbf{c}$, where $\mathbf{s} = \sum_{j=1}^{\mathbf{l}} \left(\frac{\mathbf{z}_j^* - \mathbf{y}_\mathbf{j}}{\hat{\lambda}_k} \right)^2$ and $\mathbf{c}$ is equal to the maximum acceptance probability across all runs.

2.1. Bayesian inference for parameters in ODE models

Consider a Q-dimensional dynamical system for the state variable vector, $\mathbf{x}(t)$, described by the system of ODEs:

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}, \boldsymbol{\theta}, t), \tag{2.5}$$

where $\mathbf{x}$ is a $Q \times 1$ vector of the dependent variables, $\mathbf{f}$ is a $Q \times 1$ vector-valued Lipschitz continuous function with respect to $\mathbf{x}$, $\boldsymbol{\theta}$ is an $M \times 1$ vector of model parameters, t is the independent variable (often time) and $\dot{\mathbf{x}}$ represents the derivative of $\mathbf{x}$ with respect to the independent variable. Given the dynamical system in equation (2.5), along with values for the parameter vector, $\boldsymbol{\theta}$, and the initial condition, $\mathbf{x}_0$, the solution to the system can be approximated numerically.

We denote an experimental observation at time t_k by the $Q \times 1$ vector $\mathbf{y}_k$. Experimental observations are taken at K time points; the times are stored in a $K \times 1$ vector $\mathbf{t} = (t_1, t_2, \ldots, t_K)^T$ and the observations are stored in the $Q \times K$ matrix $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_K)$. These observations are usually associated with some unknown noise process, characterized by one or more variance parameters, say $\boldsymbol{\sigma}^2$. The (approximate) solution for the dependent variables at time t_k, given $\boldsymbol{\theta}$ and $\mathbf{x}_0$, is denoted by the $Q \times 1$ vector $\hat{\mathbf{x}}(t_k; \boldsymbol{\theta}, \mathbf{x}_0)$. The solution for the dependent variables at times $\mathbf{t}$ is stored in the $Q \times K$ matrix $\hat{\mathbf{x}}(\mathbf{t}; \boldsymbol{\theta}, \mathbf{x}_0) = (\hat{\mathbf{x}}(t_1; \boldsymbol{\theta}, \mathbf{x}_0), \hat{\mathbf{x}}(t_2; \boldsymbol{\theta}, \mathbf{x}_0), \ldots, \hat{\mathbf{x}}(t_K; \boldsymbol{\theta}, \mathbf{x}_0))$. In a Bayesian setting, the posterior distribution for $\boldsymbol{\theta}$ and $\boldsymbol{\sigma}^2$ given $\mathbf{y}$ is:

$$p(\boldsymbol{\theta}, \boldsymbol{\sigma}^2 | \mathbf{y}) \propto p(\mathbf{y} | \boldsymbol{\theta}, \boldsymbol{\sigma}^2) p(\boldsymbol{\theta}) p(\boldsymbol{\sigma}^2), \tag{2.6}$$

where $p(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\sigma}^2)$ is the likelihood, $p(\boldsymbol{\theta})$ and $p(\boldsymbol{\sigma}^2)$ are independent priors for $\boldsymbol{\theta}$ and $\boldsymbol{\sigma}^2$ respectively.

2.1.1. Observation model

In this paper, we assume that each observation, $\mathbf{y}_k$ for $k = 1, \ldots, K$, has an associated additive noise process, $\boldsymbol{\delta}_k$, such that

$$\mathbf{y}_k = \hat{\mathbf{x}}(t_k; \boldsymbol{\theta}, \mathbf{x}_0) + \boldsymbol{\delta}_k, \tag{2.7}$$

where $\boldsymbol{\delta}_k$ is a $Q \times 1$ vector and $\hat{\mathbf{x}}(t_k; \boldsymbol{\theta}, \mathbf{x}_0)$ is the solution for the dependent variables at time t_k, given $\boldsymbol{\theta}$ and $\mathbf{x}_0$. Under a Gaussian error model (we assumed Gaussian model for simplicity and illustration purposes but we can assume any kind of error model), and assuming the $\boldsymbol{\delta}_k$ are independent of each other, $\mathbf{y}_k$ follows a multivariate normal distribution:

$$\mathbf{y}_k \sim \text{MVN}(\hat{\mathbf{x}}(t_k; \boldsymbol{\theta}, \mathbf{x}_0), \boldsymbol{\Sigma}(\boldsymbol{\sigma}^2)), \tag{2.8}$$

where $\boldsymbol{\Sigma}(\boldsymbol{\sigma}^2)$ is a diagonal matrix with diagonal elements $\boldsymbol{\sigma}^2 = (\sigma_1^2, \sigma_2^2, \ldots, \sigma_Q^2)^T$ associated with the Q dependent variables. Hence, the likelihood function is given by

$$\mathcal{L}(\mathbf{y}|\hat{\mathbf{x}}(\mathbf{t}; \boldsymbol{\theta}, \mathbf{x}_0), \boldsymbol{\sigma}^2) = \prod_{k=1}^{K} \mathrm{MVN}(\mathbf{Y}_k; \hat{\mathbf{x}}(t_k; \boldsymbol{\theta}, \mathbf{x}_0), \boldsymbol{\Sigma}(\boldsymbol{\sigma}^2)) \tag{2.9}$$

and the posterior density is

$$p(\boldsymbol{\theta}, \boldsymbol{\sigma}^2, \mathbf{x}_0|\mathbf{y}) \propto p(\boldsymbol{\theta})p(\boldsymbol{\sigma}^2)p(\mathbf{x}_0) \prod_{k=1}^{K} \mathrm{MVN}(\mathbf{Y}_k; \hat{\mathbf{x}}(t_k; \boldsymbol{\theta}, \mathbf{x}_0), \boldsymbol{\Sigma}(\boldsymbol{\sigma}^2)). \tag{2.10}$$

The Bayesian techniques discussed in appendix A can be used to sample from the posterior distribution of $\{\boldsymbol{\theta}, \mathbf{x}_0, \boldsymbol{\sigma}^2\}$ in the case of MCMC and an approximation to the posterior in the case of ABC methods.

2.2. Model misspecification in ABC methods for ODE models

Exact Bayesian methods, such as MCMC, generate samples directly from $p(\boldsymbol{\theta}, \boldsymbol{\sigma}^2 \,|\, \mathbf{y})$ (at least in the limiting sense), rather than from the approximate posterior $p_\epsilon(\boldsymbol{\theta}, \boldsymbol{z} \,|\, \boldsymbol{y})$ shown in equation (2.2). This applies in general, but in particular for the case of parameter inference in ODE models, as long as an observational model has been defined, leading to a posterior distribution such as in equation (2.10). Evaluating this expression requires solving a system of ODEs to obtain $\hat{\mathbf{x}}(t_k; \boldsymbol{\theta}, \mathbf{x}_0)$ for a specific collection of parameters. It is then straightforward, at least for this simple noise model, to evaluate the likelihood and the posterior density up to a normalizing constant.

One of the major motivations for using likelihood-free methods, such as ABC, is that they are applicable even when evaluating the likelihood is difficult or impossible. That motivation is not present here, since solving the system of ODEs for each proposed parameter vector $\boldsymbol{\theta}$ is the main computational burden involved in evaluating equation (2.10), and this is still necessary for ABC, at least using the method proposed by Toni *et al.* [16]. Their method does still have a computational advantage, in that it avoids evaluating the density of the noise model, which may be prohibitive for certain models. However, where the simple independent Gaussian noise model of equation (2.2) is appropriate, or some other simple noise model applies, the contribution of these density evaluations to the overall computational burden will be negligible.

The method of Toni *et al.* [16] actually avoids even simulating draws from the noise model. This is made clear in the following text which appears in the thesis by Toni [19, p. 154], which is the basis of the work in Toni *et al.* [16]: 'We explore the differences between three different inference approaches:

1. No noise added in the simulation step, θ is the unknown parameter. This framework has been introduced in Chapter 3 and used throughout this thesis.
2. Noise added in the simulation step, σ is considered known and θ unknown.
3. Noise added in the simulation step, both σ and θ are unknown.'

Since option 1 (the approach taken throughout the thesis and associated papers) avoids adding noise in the simulation step, the method is applicable regardless of the noise model. However, this generality comes at a cost and, as we explain in this section, results in an approximate distribution of the form equation (2.2) that contains no information about parameter uncertainty.

Toni *et al.* [16] adapted the Sisson *et al.* [14] SMC ABC algorithm and used it to infer parameters in ODE models. However, the method they devised differs in two crucial aspects from standard practice in implementing ABC. The first is that they do not simulate data vectors $\mathbf{z}^*$ from the same model they assume for the data, which is of the form shown in equation (2.9). Instead, they generate $\mathbf{z}^*$ by merely solving the underlying system of ODEs for each proposed value of the parameter vector $\boldsymbol{\theta}^{**}$. The simulated data $\mathbf{z}^*$ is thus a deterministic function of $\boldsymbol{\theta}^{**}$, without any added noise, and in effect the underlying likelihood distribution model used in the resulting ABC algorithm is a point mass concentrated at the solution of the system of ODEs. In this sense, the data generation model used by Toni *et al.* [16] is misspecified.

A second departure from standard ABC practice is that the discrepancy function used by Toni *et al.* [16] directly computes a distance between the simulated and observed data, originally using Euclidean distance

$$\rho(\mathbf{z}, \mathbf{y}) = \sum_{j=1}^{n} (z_j - y_j)^2, \tag{2.11}$$

but they also experimented with alternative metrics. In this paper, we experiment with the more general adaptively weighted distance function in equation (2.4). All of these discrepancy functions have in

common that they can only take a zero value when the simulated data $\mathbf{z}^*$ exactly corresponds to the observed data $\boldsymbol{y}$. In contrast, practical ABC methods more commonly use a discrepancy function based on the distance between vectors of summary statistics $s(\boldsymbol{z}^*)$ and $s(\boldsymbol{y})$, which have much lower dimension than the simulated and observed data vectors $\boldsymbol{z}^*$ and $\boldsymbol{y}$. In that case, the discrepancy function is zero whenever $s(\boldsymbol{z}^*) = s(\boldsymbol{y})$, which can occur even if $\boldsymbol{z}^* \neq \boldsymbol{y}$. A practical reason for basing the discrepancy on a vector of summary statistics is that this places weaker constraints on the acceptability of proposed pairs $(\boldsymbol{\theta}^{**}, \boldsymbol{z}^*)$. If the summary statistics are sufficient for $\boldsymbol{\theta}$, nothing is lost by using $s(\boldsymbol{y})$ instead of $\boldsymbol{y}$, but more usually the summary statistics capture much, but not all, of the information $\boldsymbol{y}$ can reveal about $\boldsymbol{\theta}$.

As a result of these two departures from standard ABC practice, it will not in general be possible for the discrepancy $\rho(\mathbf{z}, \mathbf{y})$ in equation (2.11) to be arbitrarily small. The problem is that since the generative data model does not include a noise term, there may be no parameter vector for which the solution to the system of ODEs exactly corresponds to the observations $\boldsymbol{y}$, and hence there is some minimum allowable discrepancy $\epsilon_0 > 0$. Consequently, $\rho(\mathbf{z}, \mathbf{y})$ will always be greater than 0, which is considered a misspecification in ABC estimation, according to Frazier *et al*. [36]. Under ideal conditions, for the function $\boldsymbol{\theta} \mapsto \rho(z(\boldsymbol{\theta}), \boldsymbol{y})$ there is a unique $\boldsymbol{\theta}_0$ such that $\rho(z(\boldsymbol{\theta}_0), \boldsymbol{y}) = \epsilon_0 > 0$, where $z(\boldsymbol{\theta})$ is the unique solution to the system of ODEs with parameter vector $\boldsymbol{\theta}$. Therefore, as $\epsilon \to \epsilon_0$ from above, the approximate posterior $p_\epsilon(\boldsymbol{\theta}, z \mid \boldsymbol{y})$ approaches a Dirac delta function at the point $(\boldsymbol{\theta}_0, z(\boldsymbol{\theta}_0))$.

It follows that the approximate 'posterior' $p_\epsilon(\boldsymbol{\theta}, z \mid \boldsymbol{y})$ targeted by the method of Toni *et al*. [16] contains no information about the posterior variance of parameters. A practical demonstration of this is provided in the results below, in which small to moderate changes in the noise model used to simulate the observations resulted in no change in the posterior variance estimated by ABC methods. On the other hand, changing the noise model used to simulate observations did affect the location of the posterior and the final ϵ that guaranteed a good acceptance rate.

However, our results presented below demonstrate that the *shapes* of the contours of distributions of the form of equation (2.2) for $\epsilon > \epsilon_0$ may resemble those of the true posterior, and we propose that it may be possible to find some $\epsilon > \epsilon_0$ for which p_ϵ approximates the true posterior. Finding a good way to do this is left for future work.

3. Test problems

The ABC and MCMC techniques described in appendix A were compared against each other when conducting parameter inference for one epidemiological compartmental model. The Bayesian parameter inference software developed in this paper was validated using the method of posterior quantiles [37] on a computationally inexpensive model described in §3.1, before being implemented on a more demanding nonlinear system of ODEs describing malaria transmission in §3.2.

3.1. Test problem 1—susceptible–infected–recovered model

Susceptible–infected–recovered (SIR) models categorize hosts into one of three different compartments at time t. Individuals are considered susceptible (S), if they are able to be infected by the pathogen, infected (I) if currently infected with the pathogen or recovered (R) if they have successfully cleared the pathogen. The flow of individuals between compartments in the SIR model is visualized in figure 1.

SIR models and their variants, in both deterministic and stochastic forms, are among the most fundamental epidemiological models and have found use describing diseases as diverse as influenza, herpes and malaria [38]. In this test problem we use the SIR model to represent the fraction of the total population (P) in each category as follows:

$$s(t) = \frac{S(t)}{P}, \quad i(t) = \frac{I(t)}{P} \quad \text{and} \quad r(t) = \frac{R(t)}{P},$$

where $S(t)$, $I(t)$ and $R(t)$ are the numbers of susceptible, infected and recovered individuals in the population at time t (weeks). The deterministic, constant population, SIR model without demographics can be described mathematically as:

$$\left.\begin{aligned} \frac{\mathrm{d}s}{\mathrm{d}t} &= -\beta i(t)s(t), \\ \frac{\mathrm{d}i}{\mathrm{d}t} &= \beta i(t)s(t) - \gamma i(t) \\ \text{and} \quad \frac{\mathrm{d}r}{\mathrm{d}t} &= \gamma i(t), \end{aligned}\right\} \tag{3.1}$$

where β is the infection rate and γ is the recovery rate.

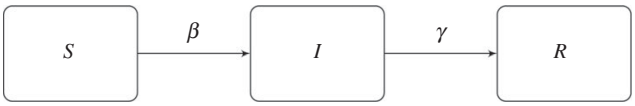

Figure 1. The transition of individuals between susceptible, infected and recovered states in an epidemiological compartmental model.

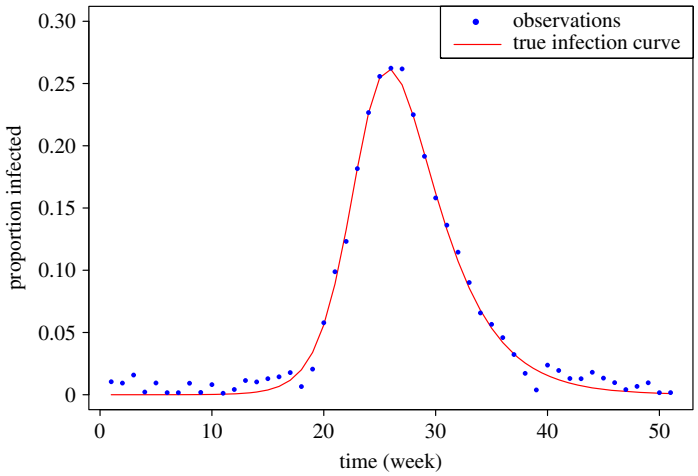

Figure 2. Test data showing the proportion of population infected over time, obtained from equation (3.1) with $\beta = 0.9$, $\gamma = 0.333$, $S(0) = 1 - 1.27 \times 10^{-6}$, $I(0) = 1.27 \times 10^{-6}$ and $R(0) = 0$. Red line shows the continuous infection curve while the blue points are the observations to be used to infer the model parameters.

When conducting parameter inference for this system, the observable data is $\mathbf{y} = (y_1, \ldots, y_n)$, where $y_k = i(t_k)$ is the proportion of the population infected at time t_k, for $k = 1, \ldots, n$. The parameters of interest are $\boldsymbol{\theta} = \{\beta, \gamma\}$.

3.1.1. Simulation results

A test dataset was generated by solving the system of equations (3.1) in the interval [0, 50] using a fourth order Runge–Kutta method and storing the solution at weekly intervals (figure 2), using true model parameters $\boldsymbol{\theta} = (\beta = 0.9, \gamma = \frac{1}{3})^T$. To generate observations $\mathbf{y}$, normal noise $\mathcal{N}(0, \sigma^2 = 0.0001)$ was added to the solution. For ABC approaches, a discrepancy function $\rho(\mathbf{z}, \mathbf{y})$ was used to compare infected proportions in the dataset $\mathbf{y}$ with a solution to the equations $\mathbf{z} = (z_1, \ldots, z_n)$ for proposed parameters as follows:

$$\rho(\mathbf{z}, \mathbf{y}) = \frac{1}{n} \sum_{i=1}^{n} (z_i - y_i)^2, \tag{3.2}$$

where n is the number of observed data points. The priors for β, γ and σ^2 were taken to be vague:

$$\beta \sim \mathcal{U}(0, 2), \quad \gamma \sim \mathcal{U}(0, 2) \quad \text{and} \quad \sigma^2 \sim \mathcal{IG}(1, 1) \tag{3.3}$$

where $\mathcal{U}(\cdot, \cdot)$ is the uniform distribution and $\mathcal{IG}(\cdot, \cdot)$ is the inverse-gamma distribution. For MCMC approach, normal proposal distributions were used with adaptive approach tuning parameters in the algorithm to maintain an acceptance ratio between 0.3 and 0.5 [25].

Given the observations $\mathbf{y}$, the parameter vector $\boldsymbol{\theta} = \{\beta, \gamma\}$ was estimated using MCMC and SMC ABC and the results from these methods were compared. The noise σ^2 was additionally estimated when using MCMC. As discussed in §2.2, the distributions derived using the ABC approaches are not an approximation to the true posterior of the ODE model parameters since the noise is not estimated; we therefore cannot use the standard deviation of the distributions from ABC approaches as a measure of performance. Instead, we compared the CPU times, the number of iterations and the mean absolute

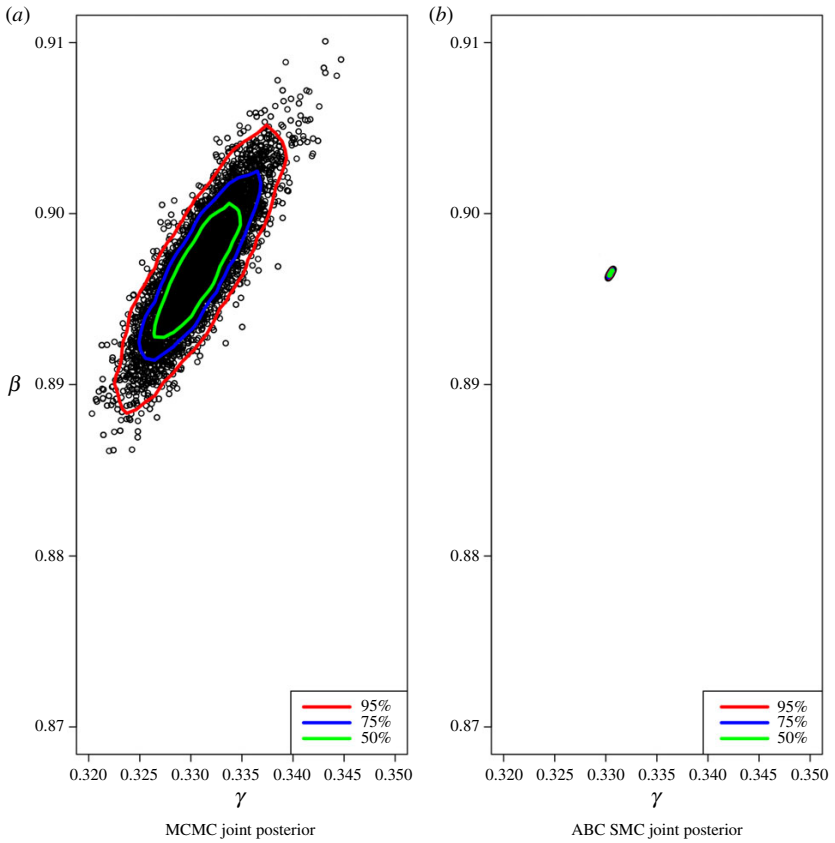

Figure 3. Scatter plot of sample draws for γ and β using MCMC (*a*) and SMC ABC (*b*). The contour lines contain the stated proportions of sample draws from the joint posterior and are produced using the R function 'HPDregionplot'.

Table 1. The number of iterations, computational time (min) and mean absolute error for parameter inference in the SIR model.

	iterations	CPU time	MAE (β)	MAE (γ)
MCMC	12 401	6.58 min	0.0038	0.0034
SMC ABC	141 408	11.29 min	0.0035	0.0029

errors (MAE), although, MAE may favour over-concentrated posterior approximations. The formula used for MAE is MAE $= \sum_{i=1}^{n} (|\theta_i - \theta_{\text{true}}|)/n$, and the θ_i are posterior samples for θ, for each method. We applied the MCMC approach using appendix A, algorithm 2. Table 1 shows that MCMC chain converged after 12 401 steps to reach convergence and this took approximately 6.58 min.

We next applied SMC ABC as outlined in appendix A, algorithm 6 using our own implementation in R. In SMC ABC code we used $T = 11$ populations, each with 1000 particles, used component-wise uniform kernels that adapted their width from the previous particle distributions [28] and used uniform priors 2 units wide and centred at zero for both parameters. The tolerance sequence was selected adaptively such that in population i the new threshold ϵ_i was the 25th percentile of the distances in the previous iteration, $t-1$ (as explained in §2). The algorithms terminated when we reached a challenge tolerance of $\epsilon = 0.067056$ that had been chosen by finding the distance between the true ODE solution and the generated observations $\mathbf{y}$.

Comparing SMC ABC with MCMC, we found that SMC ABC consumed run times longer than MCMC with 11.26 min. In addition, It can be seen in figure 3*b* that the estimated joint posterior resulting from the 11th population of the SMC ABC method has the smaller variance compared with MCMC method. Table 2 shows that all the methods have achieved good point estimation for both parameters β and γ.

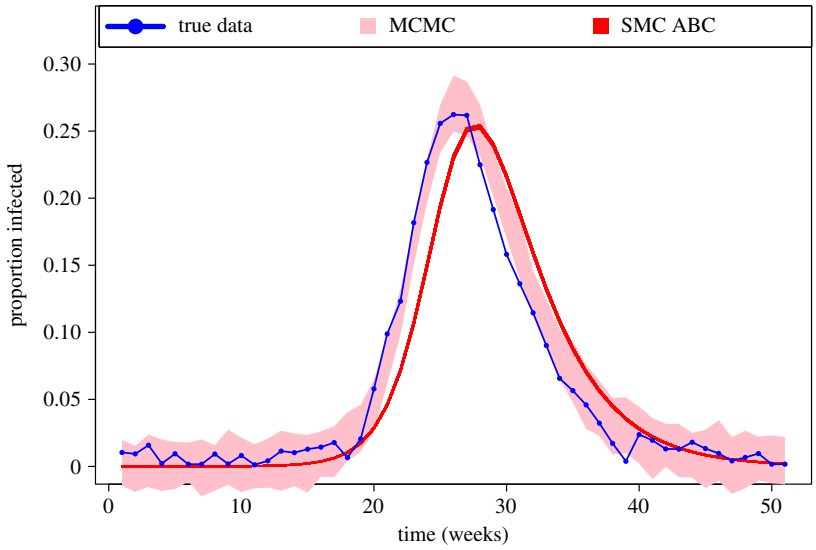

Figure 4. Plot of the posterior predictive credible intervals estimated using MCMC and SMC ABC fitted with the weekly infection cases. The blue dots represent the noisy data. The shaded areas are created using the posterior predictive samples. It can be seen that the result derived from MCMC covered most of the data points while the ABC derived result produces unrealistically narrow credible intervals.

Table 2. True values of the parameters β and γ and their estimated values (each estimate is the median of the sampled values) using MCMC and SMC ABC for SIR model example.

parameter	true value	MCMC	SMC ABC
β	0.9	0.8968	0.8964
γ	0.3333	0.3308	0.3304

We also compared credible intervals for the solution to the system of ODEs, using MCMC and ABC. To do this, we followed the procedure in Gelman & Rubin [25] to simulate the posterior predictive distribution (PPD) for a future observation, $\mathbf{y}_{\mathbf{rep}}$. This produces a PPD that has variance that depends on the posterior uncertainty (and hence the observational noise). This was not a problem within MCMC because we take samples from the posterior and solve the model and then simulate the noise, which has been estimated within MCMC. However, with SMC ABC method, we take the samples from the posterior and solve the model without simulating the noise. So, as the tolerance gets smaller and smaller the predictive intervals will also get smaller, and do not have the correct coverage of the observed data. Figure 4 shows that the credible intervals obtained using SMC ABC are much narrower than those obtained using MCMC, highlighting that the variation within the sample does not contain useful information about the inferred uncertainty of the estimates. Note that this problem is not reduced by the tunable elements of the algorithm; for example, we tried several different perturbation kernels proposed by Filippi *et al.* [28], such as component-wise perturbation kernels that adaptively choose width based on the previous population and multivariate normal perturbation kernels that are sometimes useful when parameters are highly correlated. The resulting credible intervals were not affected significantly by the choice of perturbation kernel (comparison results not shown).

In addition, we used the proposed method in Prangle [33] within our R code, involving an adaptive distance function to improve the performance of the SMC ABC method. In this algorithm, the scale parameters ζ_j in the distance function (equation (2.4)) are updated in each iteration (calculated using MAD) and are used to choose the value of the next ϵ_t. In principle, it might be possible to use a variant of this technique to choose ϵ_t so that the resulting sample reflects the shape and spread of the true posterior distribution. However, it is not clear what number of rounds of adaptation would produce such an ϵ_t. In other words, it is not clear how and when to terminate the SMC ABC algorithm. In the literature, there

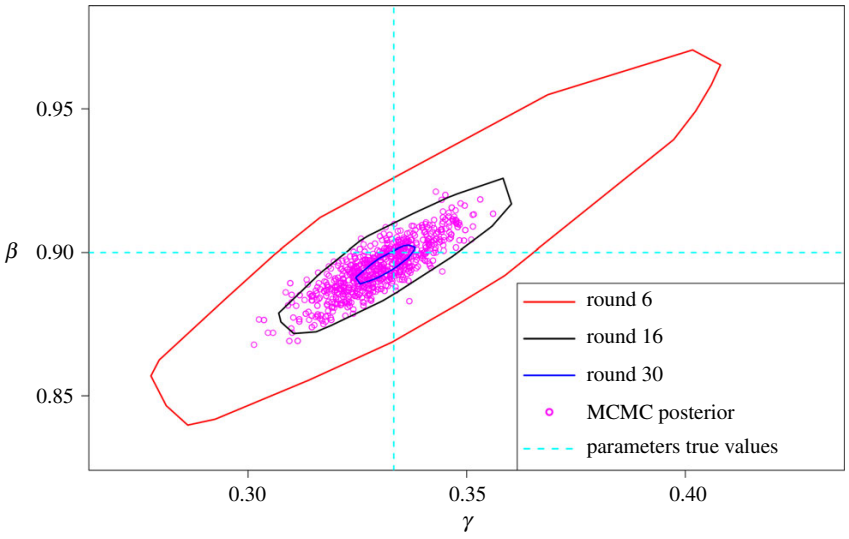

Figure 5. An illustration of SMC ABC for the SIR model using algorithm 4 in Prangle [33]. The red curve represents the posterior estimate resulting from the sixth round and is clearly overdispersed (by comparison to the sample obtained using MCMC—purple scatter points). The black curve, obtained after 16 rounds of SMC ABC, is the closest approximation to the posterior obtained using SMC ABC. The blue curve is the posterior estimate obtained after 30 rounds of SMC ABC, and has clearly shrunk too much around the true parameter values (dashed light blue lines).

are several methods one can use to terminate, such as when the algorithm reaches a certain value of ϵ or a target acceptance rate, or one can use a specified number of a total simulations as a tuning parameter and the algorithm terminates when a further simulation is required [33]. One option that is available when using an adaptive kernel width is to stop the algorithm when the width of the kernel becomes negligible or when $1-(\epsilon_{t+1}/\epsilon_t)$ falls below some threshold. However, we found that none of these methods terminate the algorithm in such a way as to produce the correct shape and spread of the posterior distribution. Figure 5 illustrates that if we run the algorithm for 30 rounds the estimated posterior will shrink towards a point estimate of the parameters. If we run the algorithm for six rounds, the resulting estimate is not a good representation of the true posterior: it is too wide and hence misleading. An estimated posterior distribution similar to what we obtain using MCMC can be somewhat artificially generated if we run the SMC ABC algorithm for 16 rounds, but at present the algorithm lacks an independent way of identifying this. As far as we are aware, there are as yet no clear guidelines to follow to determine how many rounds of SMC ABC are needed when dealing with an ODE model to guarantee that a good approximation to the true posterior of the parameters has been achieved.

3.1.2. The impact of the noise on the inference

To demonstrate the impact of the noise on the parameter inference we applied MCMC and SMC ABC to observations $\mathbf{y}$ generated using different values of σ^2, specifically $\sigma^2 \in \{0.0001, 0.0005, 0.001\}$. We plot the resulting posterior marginals for both parameters in figure 6. The variance of the estimated posterior derived from MCMC increases as the value of σ^2 increases, indicating that the variance of the posterior is affected by the amount of noise, as expected. On the other hand, although the noise has been increased, the variances of the estimated posteriors derived from SMC ABC are almost the same for small to moderate amounts of noise. However, when the noise parameter is increased further, the location of the estimated posterior is changed. This illustrates what has been discussed in §2.2 that these posteriors do not provide valid information about the uncertainty in the parameter estimates. As a result, conducting parameter estimation for ODE models using this ABC framework is not recommended.

3.1.3. Including the error term in the ABC algorithm

Including the error term in the ABC algorithm may overcome this limitation and work inspired by Wilkinson [35] in Vaart *et al.* [34], as has been explained in §2, argued that the acceptance of the

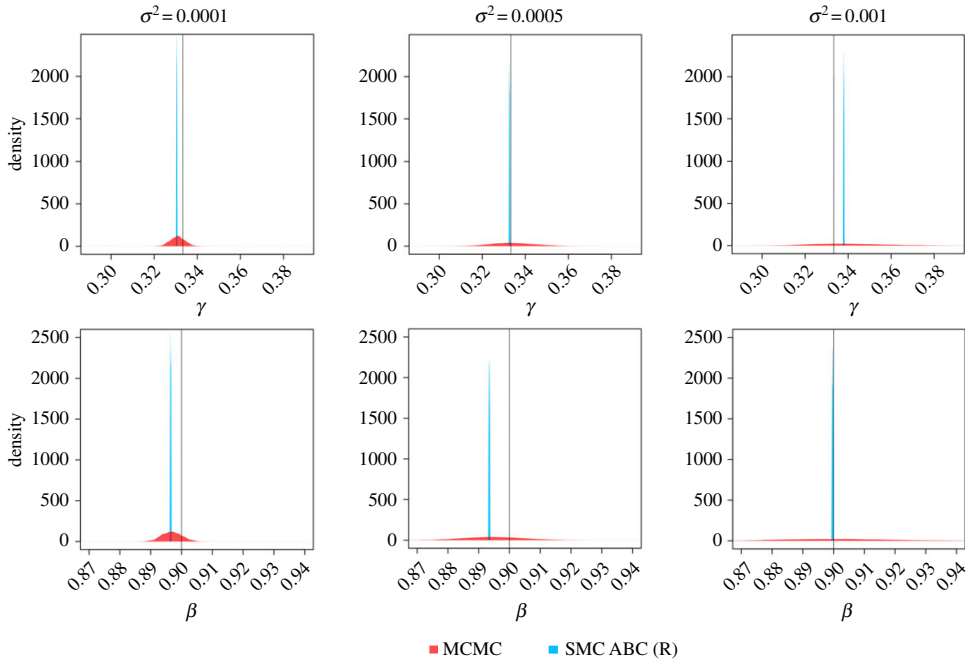

Figure 6. Plots of the estimated posterior marginal densities for parameter β and γ obtained using MCMC (red) and SMC ABC (blue) with different amounts of noise. The black solid line represents the true value of the parameters. It is clear that the variance of the posteriors derived from MCMC is affected by increasing the noise, but this is not the case for posteriors obtained using SMC ABC.

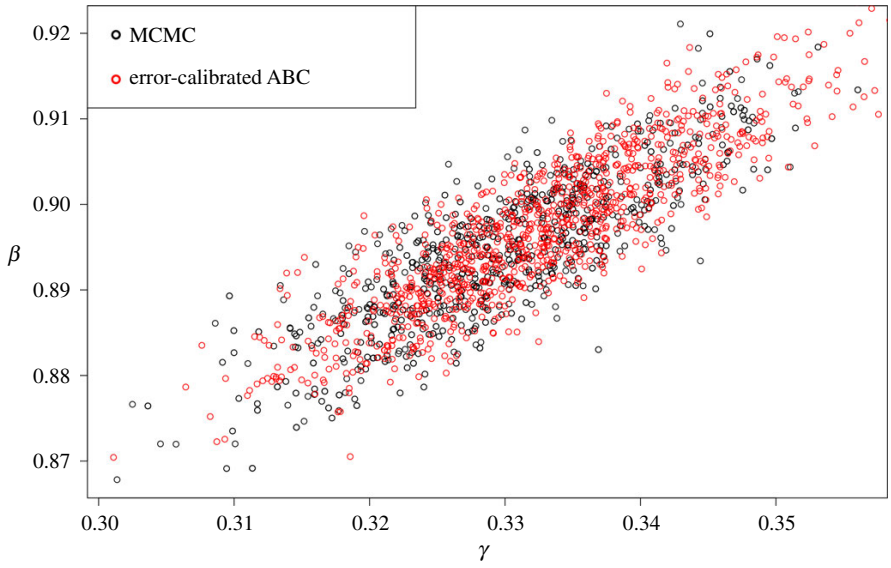

Figure 7. Scatter plot of posterior distribution sample draws for β and γ from MCMC (black) and error-calibrated ABC (red) for the SIR model in SIR model example.

proposed parameters should be with respect to the error term rather than with respect to some tolerance level. In their method they assumed that the error term follows a normal distribution. This method is promising and can capture similar posterior shapes compared to the one derived from MCMC, as figure 7 shows. A significant drawback that we found for this approach is that the acceptance rate is very

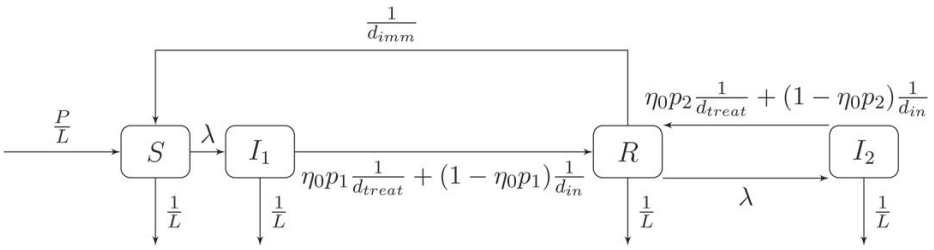

Figure 8. The flow of individuals between susceptible, infected and recovered states in the model of White *et al.* [39].

low and a large number of simulations are needed. This leads to a longer computational time, which is prohibitive for the case of ODE model parameter inference. The total CPU time when we applied the Vaart *et al.* [34] algorithm to the first test problem to have 1000 samples is 29 h which was derived from 2×10^6 simulations and this time is much larger than when we used the MCMC method for this example (5.25 min).

3.2. Example 2—nonlinear ODE model of malaria transmission

Work by White *et al.* [39], acknowledging the lack of reliable data in some countries where malaria control or elimination is particularly desirable, showed the utility of a compartmental mathematical model in predicting effects of various elimination strategies compared to the more complex models of Gu *et al.* [40] and Maire *et al.* [41]. The model describes population dynamics using four population compartments in the transmission of malaria:

$S(t)$: Uninfected and non-immune.
$I_1(t)$: Infected with no prior immunity.
$R(t)$: Uninfected with immunity.
$I_2(t)$: Infected with prior immunity.

The model comprises four ODEs that govern the temporal evolution of the population compartments. The model is illustrated in figure 8 and can be described mathematically by the following equations:

$$\left.\begin{aligned}
\frac{\mathrm{d}S}{\mathrm{d}t} &= \frac{P}{L} - \left(\lambda + \frac{1}{L}\right)S + \frac{1}{d_{\mathrm{imm}}}R,\\
\frac{\mathrm{d}I_1}{\mathrm{d}t} &= \lambda S - \left(\frac{\eta_0 p_1}{d_{\mathrm{treat}}} + \frac{1-\eta_0 p_1}{d_{\mathrm{in}}} + \frac{1}{L}\right)I_1,\\
\frac{\mathrm{d}I_2}{\mathrm{d}t} &= \lambda R - \left(\frac{\eta_0 p_2}{d_{\mathrm{treat}}} + \frac{1-\eta_0 p_2}{d_{\mathrm{in}}} + \frac{1}{L}\right)I_2\\
\text{and}\quad \frac{\mathrm{d}R}{\mathrm{d}t} &= \left(\frac{\eta_0 p_1}{d_{\mathrm{treat}}} + \frac{1-\eta_0 p_1}{d_{\mathrm{in}}}\right)I_1 + \left(\frac{\eta_0 p_2}{d_{\mathrm{treat}}} + \frac{1-\eta_0 p_2}{d_{\mathrm{in}}}\right)I_2 - \left(\lambda + \frac{1}{d_{\mathrm{imm}}} + \frac{1}{L}\right)R.
\end{aligned}\right\} \tag{3.4}$$

Here λ is the force of infection and is given by

$$\lambda = R_0\left(\frac{1}{L} + \frac{1}{d_{\mathrm{treat}}}\right)\frac{(I_1 + I_2)}{P},$$

where R_0, the average number of secondary infections arising from a single infected individual in a susceptible population, is expressed as a function of time to incorporate the seasonal forcing associated with malaria transmission and is of the form, $R_0(t) = A\cos 2\pi(t-\phi) + r_0$. The model is parametrized in terms of a number of constants as described in table 3.

The observed data, $\mathbf{y(t)}$, is taken to be the number of observable clinical infections, $C(t)$ as follows:

$$\mathbf{y(t)} = \{C_1(t_1), C_2(t_2), \ldots, C_{n-1}(t_{n-1}), C_n(t_n)\} \tag{3.5}$$

and

$$C_n(t_n) = p_1\mathbf{I_1}(t_n) + p_2\mathbf{I_2}(t_n). \tag{3.6}$$

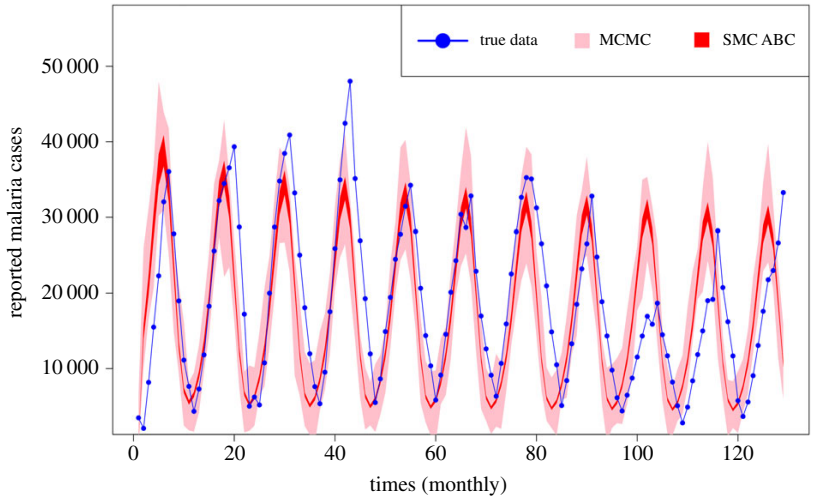

Figure 9. Plot of the posterior predictive credible intervals from MCMC and SMC ABC fitted with the monthly Afghanistan data. The blue dots represent the data. The shaded areas are created by the posterior predictive samples. The result derived from MCMC covered most of the data points while the ABC methods were unable to cover most of the data.

Table 3. The parameter values used in simulation of the White *et al.* [39] model.

parameter		value	source
P	(people)	29 203 486	Worldometers [42]
L	(years)	66.67	Maude *et al.* [43]
d_{imm}	(years)	0.93	Aguas *et al.* [44]
d_{in}	(years)	0.11	assumed
d_{treat0}	(weeks)	2	Maude *et al.* [43]
p_1		0.87	Aguas *et al.* [44]
p_2		0.08	Aguas *et al.* [44]
A		0.67	assumed
r_0		1.23	assumed
ϕ		3/12	assumed
η_0		0.11	assumed

For our purposes here, the parameter vector of interest is $\boldsymbol{\theta} = (\eta_0, d_{\text{in}})$ and in the case of using MCMC, the parameter vector of interest is $\boldsymbol{\theta} = (\eta_0, d_{\text{in}}, \sigma^2)$, where η_0 is the percentage of individuals with clinical infection that receive treatment, d_{in} is the average duration of an untreated sensitive infection and σ^2 is the noise associated with the data which we assumed to be normally distributed.

3.2.1. Application: malaria in Afghanistan

Afghanistan is a landlocked country located between South Asia and Central Asia. Despite the fact that most of the country is desert, there is significant rainfall and snowfall [45], which provides a fertile environment for mosquito-borne diseases such as malaria. We use monthly data from cases registered nationwide across all regions of Afghanistan in the period from January 2005 to September 2015 from Anwar *et al.* [46] as shown in figure 9.

In the ODE system in equation (3.4), I_1 and I_2 represent the number of infected individuals with no prior immunity and prior immunity, respectively. However, in the case of the data from Afghanistan, each data point represents the total number of malaria cases that arrived at hospitals in the month. In order to calculate

Table 4. The estimated values (the median of the posterior) of the parameters η_0 and d_{in} from MCMC, MCMC ABC and SMC ABC for the Afghanistan data.

parameter	MCMC	MCMC ABC	SMC ABC
η_0	0.0525	0.04685	0.0459
d_{in}	0.23035	0.2483	0.2453

Table 5. The number of iterations and computational time (min) for parameter inference in the malaria model, applied to the Afghanistan data.

	iterations	CPU time
MCMC	5838	45.13 min
MCMC ABC	4050	39.99 min
SMC ABC	143 031	522.34 min

the cumulative number of cases over time we added an extra ODE which takes the form:

$$\frac{\mathrm{d}W}{\mathrm{d}t} = \lambda S \eta_0 p_1 + \lambda R \eta_0 p_2, \tag{3.7}$$

where $W(t)$ is the cumulative number of observed (that is, treated) cases. To compute the number of new cases in each month we subtract the cumulative cases from consecutive months.

The values of the model parameters used are shown in table 3 and the initial conditions are given by the equilibrium solution of the system in equation (3.4) with the addition of $W(t=0)=0$ for the new ODE. As with the first test problem, for the ABC approaches we used the discrepancy function in equation (3.2) to compare the clinical infections given in the dataset $\mathbf{y}$ with a simulated solution $\mathbf{x}$. The priors for η_0, d_{in} and σ^2 were taken as follows:

$$\left.\begin{aligned} p(\eta_0) &= \mathcal{B}(1,1) \in [0,1], \\ p(d_{\text{in}}) &= \mathcal{GA}(1,1) \in [0,\infty) \\ \text{and} \quad p(\sigma^2) &= \mathcal{IG}(1,1) \in [0,\infty). \end{aligned}\right\} \tag{3.8}$$

A logistic transformation was used to transform $\eta_0 \in [0,1]$ while a log transform was applied to $d_{\text{in}} \in [0,\infty)$ and $\sigma^2 \in [0,\infty)$ so that each transformed parameter had support over the real line. This step was used to improve the acceptance rate of the proposals.

In MCMC method, zero mean normal proposal distributions were used with standard deviations equal to (0.007, 0.07, 0.1) for the parameters (η, d_{in}, σ^2) respectively. The same proposal distributions were used with MCMC ABC, but with standard deviations equal to (0.1, 0.1) for the parameters (η, d_{in}).

All of the methods have achieved convergence to similar values for both parameters under investigation (η_0, d_{in}), as shown in table 4. SMC ABC consumed significantly longer CPU time compared with the other methods as shown in table 5. Also, SMC ABC needed 143 031 model simulations to get 500 accepted values, while MCMC and MCMC ABC needed just 5838 and 4050 iterations, respectively to converge (tables 4 and 5).

Since a real dataset has been used here, the true parameter values are unknown. As a consequence, applying MCMC ABC was difficult because this lack of information makes the choice of an appropriate ϵ problematic. In this paper, in order to select an appropriate tolerance level we adopted the method of Vaart *et al.* [34], which is to solve the ODE model with different proposals of the parameters from the priors, find all the distances using 3.2 between these solutions and the true data, and then choose the one that minimizes this distance. We then used the best fitting solution to estimate the value of the MCMC ABC tolerance which was $\epsilon = 116230.8$. Then, we applied SMC ABC for six populations with an adaptively chosen sequence of tolerance $\epsilon = (244616.4, 244616.4, 244616.4, 176677.1, 116966.8, 100042.7)$.

The estimated joint posteriors of η_0 and d_{in} can be seen in figure 10. All have the same shape and similar position, but the variances are very different. Figure 9 shows that the posterior predictive distribution from MCMC covers most of the data points; however, the predictive intervals from ABC methods (here, showing only the SMC ABC result from the last population) are very tight and poorly

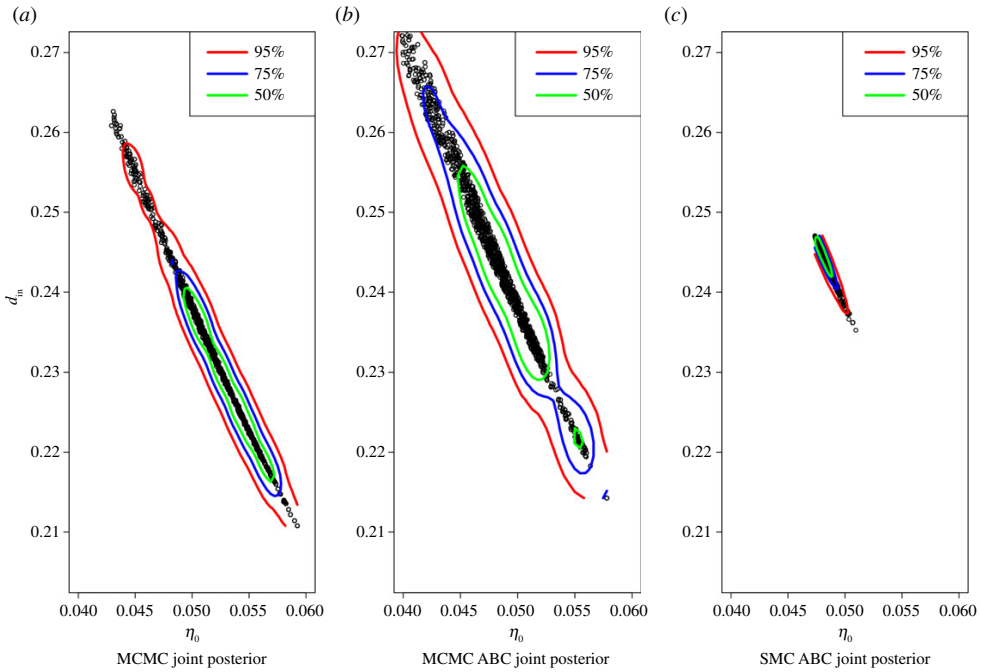

Figure 10. Scatter plot of posterior distribution sample draws for η_0 and d_{in} obtained using MCMC (*a*), ABC MCMC (*b*) and SMC ABC (*c*) for model applied to Afghanistan data. The contour lines contain the stated proportion of sample draws and they were produced using the R function 'HPDregionplot'.

cover the data points. The reason for this is, as the SMC ABC algorithm converges towards the point estimate, without any consideration to the noise in the data, the tolerance gets smaller and then most of the accepted parameter values are from a tight region around a point estimate (figure 10). Thus, the predictive intervals from ABC methods do not enable appropriate coverage of the data, consistent with the discussion in §2.2.

4. Discussion

Our investigation into exact and approximate methods for inferring parameters in ODE dynamical systems under a Bayesian framework highlights some limitations of current methods. The main problem we identified is that the observations of the system are often noisy, so when we infer the parameters for such a system it is inappropriate to not simulate the noise process. Estimation of the noise parameter is standard using exact Bayesian inference (MCMC), but not with the current practice with ABC-based approaches when applying to a system of ODEs that we investigated here. The general idea when applying the ABC-based methods considered here to a dynamical system is to compare the noisy observations to solutions generated using the ODEs (which is a deterministic model that does not take into consideration the noise in the observations) for each set of parameters proposed. The parameters are accepted based on some tolerance ϵ that also does not depend on the noise term.

To illustrate this limitation, we compared the popular methods MCMC, MCMC ABC and SMC ABC for estimating model parameters in ODEs. We can see in the second example presented in this paper that the computational time consumed by MCMC ABC is shorter compared to the other methods (MCMC and SMC ABC). However, when dealing with a deterministic model, the estimated posteriors derived from current ABC methods do not provide useful estimates of the true posteriors. In particular, they do not contain appropriate information about the uncertainty of the parameter values. Being able to naturally quantify uncertainty in posterior distributions is one of the main advantages of Bayesian statistical inference over other approaches, given that the output of Bayesian inference is a probability distribution rather than a point estimate. Here, we have shown that the ABC methods are not able to capture the distribution, but instead converge to point estimates of the best parameter values. To

demonstrate the effect of not including the noise term in the estimation we compared the results in the first example for different noise values. It is clear that the distribution estimated using the ABC methods is not affected by the noise values (for low to moderate amounts of noise), which illustrates that the posterior variances estimated using ABC do not depend on the noise in the data. As a result, the posteriors do not represent the whole probability distribution of the parameters under investigation.

In order to improve SMC ABC we tried an adaptive SMC ABC on the first test problem by using several different perturbation kernels that are proposed in Filippi *et al.* [28]; component-wise perturbation kernels and multivariate normal perturbation kernels that adapt their width from the previous population in the algorithm steps. Adapting the width of the kernels may affect the accuracy of the estimations, but still does not capture the true posterior shape when comparing with the posterior obtained using an MCMC method.

We also tried another adaptive SMC ABC method in this paper; an adaptive distance function as in §2, proposed in Prangle [33]. This method did make it possible to obtain a distribution that more closely approximates the posterior using SMC ABC, but the problem with this algorithm is that there is no existing criteria to identify an appropriate iteration at which to terminate.

Including the error term in the ABC algorithm may improve the ABC posterior as we have seen when we applied Vaart *et al.* [34] method in §2; however, the long computational times required by this approach are considered as a remarkable drawback in the case of ODE model parameter inference.

The comparison conducted in this paper demonstrates that using exact Bayesian inference (MCMC) for ODE parameter estimation is a practical alternative (Gelman *et al.* [47]), despite the difficulty involved in calculation of the likelihood. We found that MCMC gave accurate estimation of the parameter values and the resulting posterior gave appropriate information about the uncertainty of the parameters. Furthermore, the variance of the MCMC posterior distribution changed as the noise in the data changed, as one would expect. The same was not true for the ABC methods considered. The time consumed by the MCMC algorithms was slightly larger than MCMC ABC; however, since the resulting posteriors were more appropriate, the extra effort to calculate the likelihood is deemed worthwhile. In addition, choosing an appropriate ϵ when applying MCMC ABC is difficult, especially when working with real data. With simulated data, it is possible to find an appropriate ϵ from the distance between the true solution of the ODE model and the noisy data, but this is not possible in a real application where the true solution is unknown. In this case, in order to determine an acceptable tolerance level, we adapt the work of Vaart *et al.* [34] to find the best fit solution and then find an appropriate ϵ. We found that among all the methods, applying SMC ABC is the easiest to implement, but consumes the most computational time. Moreover, as we have observed, this method produces inappropriately shaped posterior distributions.

The first example presented in this paper involves a likelihood function that is easy to compute, so using a likelihood-based approach such as MCMC or an importance sampling method like SMC is certainly to be preferred over likelihood-free methods (such as ABC). Most of the computational cost of MCMC and SMC ABC method is consumed in solving the ODE models several times to compute the likelihood for MCMC or to do the simulations for SMC ABC. However, in the second example we found that more effort is needed to construct the likelihood functions when applying MCMC. In addition, when we chose an uninformative prior for the parameters, the SMC ABC algorithm located the appropriate region of the parameters space easily, while it was more difficult to choose appropriate initial parameters to achieve rapid convergence with MCMC. We would currently recommend users of ABC methods be careful when using it with ODEs, unless a sensible choice of error model and summary statistics can be made. Deciding what are sensible choices for the ABC algorithm is still difficult and an important topic of current and future work.

Data accessibility. All the code and the data required to reproduce the results presented in this paper is available as electronic supplementary material and appendix.

Authors' contributions. A.A.A. produced the final codes of the study, performed the analysis, interpreted results and wrote the paper. D.G.C. participated in developing the initial draft of the code and the paper. C.C.D. provided guidance on the ABC implementation and revised the manuscript. J.A.F. and J.M.K. conceived of the presented idea, supervised the findings of this work and revised the manuscript.

Competing interests. We declare we have no competing interests.

Funding. The authors are grateful to the Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers for their support of this project no. CE140100049.

Acknowledgements. We thank Prof. Richard Wilkinson, the editors and the six anonymous reviewers, whose comments helped us improve our work.

References

1. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. 1953 Equations of state by fast computing machines. *J. Chem. Phys.* **21**, 1087–1092. (doi:10.1063/1.1699114)
2. Hastings W. 1970 Monte Carlo sampling methods using Markov chain and their applications. *Biometrika* **57**, 97–109. (doi:10.1093/biomet/57.1.97)
3. Chopin N. 2002 A sequential particle filter method for static models. *Biometrika* **89**, 539–551. (doi:10.1093/biomet/89.3.539)
4. Roberts GO, Stramer O. 2002 Langevin diffusions and Metropolis-Hastings algorithms. *Methodol. Comput. Appl. Probab.* **4**, 337–357. (doi:10.1023/A:1023562417138)
5. Duane S, Kennedy A, Pendleton BJ, Roweth D. 1987 Hybrid Monte Carlo. *Phys. Lett. B* **195**, 216–222. (doi:10.1016/0370-2693(87)91197-X)
6. Girolami M, Calderhead B. 2011 Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. R. Stat. Soc. Series B Stat. Methodol.* **73**, 123–214. (doi:10.1111/j.1467-9868.2010.00765.x)
7. Pritchard J, Seielstad M, Perez-Lezaun A, Feldman M. 1999 Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol. Biol. Evol.* **16**, 1791–1798. (doi:10.1093/oxfordjournals.molbev.a026091)
8. Beaumont M, Zhang W, Balding D. 2002 Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025–2035.
9. Marjoram P, Molitor J, Plagnol V, Tavare S. 2003 Markov chain Monte Carlo without likelihoods. *Proc. Natl Acad. Sci. USA* **100**, 15 324–15 328. (doi:10.1073/pnas.0306899100)
10. Calderhead B, Girolami M, Lawrence ND. 2008 Accelerating Bayesian inference over nonlinear differential equations with Gaussian processes. In *Advances in Neural Information Processing Systems 21 (NIPS 2008)* (eds D Koller, D Schuurmans, Y Bengio, L Bottou). Neural Information Processing Systems.
11. Dondelinger F, Filippone M, Rogers S, Husmeier D. 2013 ODE parameter inference using adaptive gradient matching with Gaussian processes. In *Proc. of the 16th Int. Conf. on Artificial Intelligence and Statistics (AISTATS 2013), Scottsdale, AZ, 29 April–1 May*, pp. 216–228.
12. Wang Y, Barber D. 2014 Gaussian processes for Bayesian parameter estimation in ordinary differential equations. *Journal of Machine Learning Research - Workshop and Conference Proceedings (ICML)*, **32**, 1485–1493.
13. Dass SC, Lee J, Lee K, Park J. 2017 Laplace based approximate posterior inference for differential equation models. *Stat. Comput.* **27**, 679–698. (doi:10.1007/s11222-016-9647-0)
14. Sisson S, Fan Y, Tanaka M. 2007 Sequential Monte Carlo without likelihoods. *Proc. Natl Acad. Sci. USA* **104**, 1760–1765. (doi:10.1073/pnas.0607208104)
15. Beaumont M, Cornuet J, Marin J, Rober C. 2009 Adaptive approximate Bayesian computation. *Biometrika* **96**, 983–990. (doi:10.1093/biomet/asp052)
16. Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf MP. 2009 Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface* **6**, 187–202. (doi:10.1098/rsif.2008.0172)
17. Gupta S, Hainsworth L, Hogg J, Lee R, Faeder J. 2018 Evaluation of parallel tempering to accelerate Bayesian parameter estimation in systems biology. In *Proc. Euromicro Int. Conf. on Parallel, Distributed and Network-based Processing*, pp. 690–697. IEEE. See https://doi.org/10.1109/PDP2018.2018.00114.
18. Silk D, Filippi S, Stumpf MP. 2013 Optimizing threshold-schedules for sequential approximate Bayesian computation: applications to molecular systems. *Stat. Appl. Genet. Mol. Biol.* **12**, 603–618. (doi:10.1515/sagmb-2012-0043)
19. Toni T. 2010 Approximate Bayesian computation for parameter inference and model selection in systems biology. PhD thesis, Imperial College London.
20. da Costa JMJ, Orlande HRB, da Silva WB. 2018 Model selection and parameter estimation in tumor growth models using approximate Bayesian computation-ABC. *Comput. Appl. Math.* **37**, 2795–2815. (doi:10.1007/s40314-017-0479-0)
21. Costa JM, Orlande HR, Lione VO, Lima AG, Cardoso TC, Varon LA. 2018 Simultaneous model selection and model calibration for the proliferation of tumor and normal cells during in vitro chemotherapy experiments. *J. Comput. Biol.* **25**, 1285–1300. (doi:10.1089/cmb.2017.0130)
22. Barnes CP, Silk D, Stumpf MP. 2011 Bayesian design strategies for synthetic biology. *Interface Focus* **1**, 895–908. (doi:10.1098/rsfs.2011.0056)
23. Toni T, Stumpf MP. 2009 Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics* **26**, 104–110. (doi:10.1093/bioinformatics/btp619)
24. Sun L, Lee C, Hoeting JA. 2015 Parameter inference and model selection in deterministic and stochastic dynamical models via approximate Bayesian computation: modeling a wildlife epidemic. *Environmetrics* **26**, 451–462. (doi:10.1002/env.2353)
25. Gelman A, Rubin D. 1992 Inference from iterative simulation using multiple sequences. *Stat. Sci.* **7**, 457–472. (doi:10.1214/ss/1177011136)
26. Frazier DT, Martin GM, Robert CP, Rousseau J. 2016 Asymptotic properties of approximate Bayesian computation. (http://arxiv.org/abs/1607.06903)
27. Del Moral P, Doucet A, Jasra A. 2006 Sequential Monte Carlo samplers. *J. R. Stat. Soc. Series B Stat. Methodol.* **68**, 411–436. (doi:10.1111/j.1467-9868.2006.00553.x)
28. Filippi S, Barnes CP, Cornebise J, Stumpf MP. 2013 On optimality of kernels for approximate Bayesian computation using sequential Monte Carlo. *Stat. Appl. Genet. Mol. Biol.* **12**, 87–107. (doi:10.1515/sagmb-2012-0069)
29. Drovandi CC, Pettitt AN. 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)
30. Del Moral P, Doucet A, Jasra A. 2012 An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Stat. Comput.* **22**, 1009–1020. (doi:10.1007/s11222-011-9271-y)
31. Liepe J, Kirk P, Filippi S, Toni T, Barnes CP, Stumpf MP. 2014 A framework for parameter estimation and model selection from experimental data in systems biology using approximate Bayesian computation. *Nat. Protoc.* **9**, 439. (doi:10.1038/nprot.2014.025)
32. Didelot X *et al.* 2011 Likelihood-free estimation of model evidence. *Bayesian Anal.* **6**, 49–76. (doi:10.1214/11-BA602)
33. Prangle D. 2017 Adapting the ABC distance function. *Bayesian Anal.* **12**, 289–309. (doi:10.1214/16-BA1002)
34. Vaart E, Prangle D, Sibly RM. 2018 Taking error into account when fitting models using approximate Bayesian computation. *Ecol. Appl.* **28**, 267–274. (doi:10.1002/eap.1656)
35. Wilkinson RD. 2013 Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. *Stat. Appl. Genet. Mol. Biol.* **12**, 129–141. (doi:10.1515/sagmb-2013-0010)
36. Frazier DT, Robert CP, Rousseau J. 2017 Model misspecification in ABC: consequences and diagnostics. (http://arxiv.org/abs/1708.01974)
37. Cook SR, Gelman A, Rubin DB. 2006 Validation of software for Bayesian models using posterior quantiles. *J. Comput. Graph. Stat.* **15**, 675–692. (doi:10.1198/106186006X136976)
38. Hethcote HW. 2000 The mathematics of infectious diseases. *SIAM Rev.* **42**, 599–653. (doi:10.1137/S0036144500371907)
39. White LJ, Maude RJ, Pongtavornpinyo W, Saralamba S, Aguas R, Van Effelterre T, Day NP, White NJ. 2009 The role of simple mathematical model in malaria elimination strategy design. *Malar. J.* **8**, 1–10. (doi:10.1186/1475-2875-8-1)
40. Gu W, Killeen GF, Mbogo CM, Regens JL, Githure JI, Beier JC. 2003 An individual based model of *Plasmodium falciparum* malaria transmission on the coast of Kenya. *Trans. R. Soc. Trop. Med. Hyg.* **97**, 43–50. (doi:10.1016/S0035-9203(03)90018-6)
41. Maire N, Tediosi F, Ross A, Smith T. 2006 Predictions of the epidemiological impact of introducing a pre-erythocytic vaccine into the expanded program on immunization in sub-Saharan Africa. *Am. J. Trop. Med. Hyg.* **75**, 111–118. (doi:10.4269/ajtmh.2006.75.111)
42. Worldometers. 2018 Afghanistan population. See https://www.worldometers.info.
43. Maude RJ, Pontavornpinyo W, Saralamba S, Aguas R, Yeung S, Dondorp AM, Day NP, White NJ, White LJ. 2009 The last man standing is the most resistant: eliminating artemisinin-resistant malaria in Cambodia. *Malar. J.* **8**, 31. (doi:10.1186/1475-2875-8-31)

44. Aguas R, White L, Snow R, Gomes M. 2008 Prospects for malaria eradication in sub-Saharan Africa. *PLoS ONE* **3**, e1767. (doi:10.1371/journal.pone.0001767)
45. Brookfield M. 1998 The evolution of the great river systems of southern Asia during the Cenozoic India-Asia collision: rivers draining southwards. *Geomorphology* **22**, 285–312. (doi:10.1016/S0169-555X(97)00082-2)
46. Anwar MY, Lewnard JA, Parikh S, Pitzer VE. 2016 Time series analysis of malaria in Afghanistan: using ARIMA models to predict future trends in incidence. *Malar. J.* **15**, 566. (doi:10.1186/s12936-016-1602-1)
47. Gelman A, Bois F, Jiang J. 1996 Physiological pharmacokinetic analysis using population modeling and informative prior distributions. *J. Am. Stat. Assoc.* **91**, 1400–1412. (doi:10.1080/01621459.1996.10476708)
48. Roberts G, Gelman A, Gilks W. 1997 Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.* **7**, 110–120. (doi:10.1214/aoap/1034625254)
49. Picchini U, Forman J. 2014 Accelerating inference for diffusions observed with measurement error and large sample sizes using approximate Bayesian computation. *J. Stat. Comput. Simul.* **86**, 195–213. (doi:10.1080/00949655.2014.1002101)

Chapter 4

Estimating error parameters in dynamical systems models using Approximate Bayesian Computation

Preamble

The purpose of this chapter is to develop a new Modified SMC ABC method to overcome the problem of estimating error parameters identified in Chapter 3. Thus, this chapter addresses Objective 2. This is achieved by proposing a summary statistic that quantifies the noise associated with the observation. The benefit of the Modified SMC ABC is that the new summary statistic facilitates the use of knowledge about the error term to guide the choice of the tolerance sequence which improves the acceptance criteria. The advantages of this method are illustrated by applying the proposed method to several challenging examples.

4.1 Abstract

Approximate Bayesian Computation (ABC) is a popular tool for estimating the parameters of dynamical systems models, and in particular non-linear differential equation models. It is a Monte Carlo method designed specifically for models in which the likelihood is computationally intractable or expensive, but for which data is relatively easy to simulate. One variant of ABC, known as Sequential Monte Carlo ABC (SMC ABC), shows promise as an efficient methodology for parameter estimation, but some current implementations fail to accurately estimate the posterior distribution of noise variance when applied to Ordinary Differential Equation (ODE) models.

Here we present a modified SMC ABC algorithm and propose a new summary statistic that facilitates accurate estimation of noise variance in ODE models. These innovations also result in improved posterior predictive intervals. We apply the proposed method to two ODE epidemiological models, and demonstrate that it outperforms standard SMC ABC in terms of accuracy, and compares favourably with a Markov chain Monte Carlo (MCMC) method in terms of both accuracy and overall computational effort.

4.2 Introduction

Ordinary Differential Equations (ODEs) are a popular mathematical tool for describing physical and biological processes in the real world. These ODEs are often characterised by some unknown parameters that must be estimated using noisy observations $\mathbf{y}_k$. Bayesian approaches to such estimation problems are popular and produce excellent results even for complex non-linear ODEs. Typically, this kind of estimation involves sampling parameters from a posterior distribution using a computational method. However, some of these approaches involve in-

tegration over a subset of parameters to obtain a marginal likelihood, and these integrals may be intractable or computationally expensive. This is often the case for non-linear ODE models.

Approximate Bayesian Computation (ABC) is a likelihood-free method that has been extensively studied since its first appearance in Tavaré et al. (1997) and Pritchard et al. (1999). The method is by now well developed in terms of both its theoretical foundations and practical implementation. Recent theoretical advances include Barber et al. (2015), which analysed the convergence properties of the ABC method in general, and the related papers Li and Fearnhead (2018) and Frazier et al. (2018), which investigate the asymptotic behaviour of the approximate posterior distribution on which ABC is based, as the number of observations increases.

In the ABC approach, data is simulated under an assumed model, with the parameters of the model proposed via some mechanism. The several varieties of ABC differ in the proposal mechanism. The simulated data is then compared to the actual observed data, and the difference between simulated and observed data is quantified, ideally using a metric to evaluate the distance between two vectors of summary statistics. If the difference between the simulated and observed data is below some threshold ϵ, the parameters are accepted, meaning that they are considered to have been sampled from an approximation to the posterior distribution over parameter space, conditional on the data. Typically, the resulting ABC approximation to the posterior distribution is exact only when $\epsilon = 0$.

In widely used and highly cited applications of ABC to ODEs, such as, Toni et al. (2009); Liepe et al. (2010); Filippi et al. (2013); Prangle et al. (2017), a numerical solution to a system of ODEs is directly compared to the observed data, without taking into account the noise associated with experimental data. In such cases, ϵ cannot be set arbitrarily small, as the noisy observations differ from any exact solution to the system, and the ABC approximation to the posterior distribution

typically shrinks to a point mass as ϵ approaches its minimum (Frazier, Robert, & Rousseau, 2017; Alahmadi, Flegg, Cochrane, Drovandi, & Keith, 2020). It is therefore important to include the error term in the ABC algorithm in some way, although how this should be done is unclear. Vaart et al. (2018) argued, in work inspired by Wilkinson (2013), that acceptance of the proposed parameters should be based on an estimate of the error term rather than based on some tolerance level. In this paper, we present a new approach implementing this idea.

The method of Toni and Stumpf (2009) applies a variant of ABC based on sequential Monte Carlo (SMC) to solve inverse problems in non-linear ODEs involving both parameter estimation and model selection. This SMC ABC approach potentially provides a useful alternative to exact Bayesian methods for differential equations. It is a likelihood-free method, meaning that at no stage does the likelihood have to be evaluated. It is also simple to apply and produces uncorrelated samples. Furthermore, the SMC ABC algorithm does not suffer from "poor mixing" problems, in which the sampler gets stuck in low probability areas for extended periods of time, as sometimes happens in Markov chain Monte Carlo (MCMC) and in variants of ABC based on MCMC.

Regardless of the advantages of SMC ABC, its applications to ODE parameter inference typically suffer from at least one of three crucial defects. The first of these defects applies to several approaches that simulate data without noise, as noted above. In these approaches, the data generation model is mis-specified. The simulated data is deterministic in the sense that it is the unique solution to a system of ODEs. As a result, the underlying likelihood distribution is a point mass concentrated at the solution of the system. Thus, this approach fails to correctly characterise the uncertainty represented by the posterior distribution. Our earlier work in Alahmadi et al. (2020) discusses this problem in detail.

The second defect applies to SMC ABC approaches that assume the noise variance $\boldsymbol{\sigma^2}$ is known, such as Silk et al. (2013). In this approach, a noisy simulation $\mathbf{y}_{sim}$ is

generated by solving the ODEs and then adding noise drawn from a known error model. If the error model is a good approximation to the actual noisy observation process, the resulting approximate ABC posterior may be a good approximation. However, this approach is unrealistic in real world problems, because the variance associated with the residuals is typically unknown, and must be estimated.

The third defect applies to approaches in which the inference of the error term $\boldsymbol{\sigma^2}$ is conducted by proposing values and accepting or rejecting them, exactly as for the other parameters being estimated, as suggested in Toni (2010). Such approaches encounter two problems. The residual vector has high dimension, which can dramatically reduce the acceptance probability of proposed values. For example, if the algorithm proposes a value of $\boldsymbol{\sigma^2}$ that is close to the true $\boldsymbol{\sigma^2}$ value, but the simulated deterministic data are located far from the observations, this will result in rejecting $\boldsymbol{\sigma^2}$. In addition, it is difficult to define the distance function $\rho(\mathbf{y}_k, \mathbf{y}_{\boldsymbol{sim}})$ in such a way as to accurately reflect the noise associated with the data. We show in this paper that such methods tend to underestimate $\boldsymbol{\sigma^2}$, which in turn reduces the accuracy of estimates of the other parameters.

The aim of this paper is to modify the standard SMC ABC, in particular by controlling the threshold schedule used in Algorithm 4.1, which is crucially important in determining which parameter values to accept and which to reject. We achieve this by finding a summary statistic that can represent the noise associated with the observation $\mathbf{y}_k$ and add a new condition to standard SMC ABC that allows it to more accurately estimate the true marginal posterior distribution of $\boldsymbol{\sigma^2}$. The novelty of our proposed approach lies in that, instead of accepting proposals of $\boldsymbol{\sigma^2}$ with respect to a tolerance level on distance, we accept them based on the ratio between the variance of the differences between the noisy *observations* and the deterministic solution of the ODEs and the variance of the differences between the noisy *simulations* and the deterministic solution, using an F-test.

The rest of the paper is organised as follows. In Section 5.3 we lay out a Bayesian

inference framework for ODEs and the resulting posterior distrbution. Then, in Section 5.4 we introduce the standard SMC ABC algorithm for parameter estimation in differential equations. In Section 5.5 we introduce the modified SMC ABC algorithm and subsequently in Section 5.6 we apply the proposed method to two ODE epidemiological models and compare the results with MCMC and standard SMC ABC. Discussions and summary of our innovations are presented in Section 4.7.

4.3 Bayesian inference

We consider a continuous time Q-dimensional dynamical system described by a set of Q ODEs

$$\dot{\mathbf{x}} = \frac{d\mathbf{x}}{dt} = f(\mathbf{x}, \boldsymbol{\theta}), \tag{4.1}$$

where $\mathbf{x} = \mathbf{x}(\mathbf{t}) = (x_1(t), ..., x_Q(t))$ is a vector of Q variables all dependent on a continuous variable t (often representing time), f is a Q-dimensional smooth function, and $\boldsymbol{\theta} = (\theta_1, ..., \theta_M)$ is a vector of M unknown parameters. The data is discretized, in the sense that $\mathbf{x}$ is evaluated only at a finite number of observation times $t_k \in \{t_1, \ldots, t_K\}$. Given values for the parameter vector $\boldsymbol{\theta}$ and an initial condition $\mathbf{x_0}$, a solution $\hat{\mathbf{x}}(t, \boldsymbol{\theta}, \mathbf{x_0})$ to the system (5.1) can be approximated numerically. Then, by adding an independent $\boldsymbol{\delta_k}$ error term to this solution, Q-dimensional observations $\mathbf{y_k}$ can be constructed for each t_k, such that

$$\mathbf{y_k} = \hat{\mathbf{x}}(t_k; \boldsymbol{\theta}, \mathbf{x_0}) + \boldsymbol{\delta_k}. \tag{4.2}$$

This error term is assumed to arise from a Gaussian error model, $\boldsymbol{\delta_k} \sim MVN(\mathbf{0}, \Sigma(\boldsymbol{\sigma^2}))$, where $\Sigma(\boldsymbol{\sigma^2})$ is a diagonal matrix with diagonal elements $\boldsymbol{\sigma^2} = (\sigma_1^2, \ldots, \sigma_Q^2)$. Accordingly, the observations $\mathbf{y_k}$ follow a multivariate nor-

mal distribution:

$$\mathbf{y_k} \sim MVN(\hat{\mathbf{x}}(t, \boldsymbol{\theta}, \mathbf{x_0}), \Sigma(\boldsymbol{\sigma^2})). \tag{4.3}$$

4.3.1 Posterior computation

The full joint posterior distribution of the parameters $\boldsymbol{\theta}$, $\mathbf{x_0}$ and $\boldsymbol{\sigma^2}$ given the noisy observations $\mathbf{y_k}$ has the form

$$p(\boldsymbol{\theta}, \mathbf{x_0}, \boldsymbol{\sigma^2}|\mathbf{y_k}) \propto p(\boldsymbol{\theta})p(\mathbf{x_0})p(\boldsymbol{\sigma^2})\mathcal{L}(\mathbf{y_k}|\hat{\mathbf{x}}(t, \boldsymbol{\theta}, \mathbf{x_0}), \boldsymbol{\sigma^2}), \tag{4.4}$$

where the likelihood is

$$\mathcal{L}(\mathbf{y_k}|\hat{\mathbf{x}}(t, \boldsymbol{\theta}, \mathbf{x_0}), \boldsymbol{\sigma^2}) = \prod_{k=1}^{K} MVN(\mathbf{y_k}; \hat{\mathbf{x}}(t, \boldsymbol{\theta}, \mathbf{x_0}), \Sigma(\boldsymbol{\sigma^2})) \tag{4.5}$$

and $p(\boldsymbol{\theta})$, $p(\mathbf{x_0})$ and $p(\boldsymbol{\sigma^2})$ are the prior distributions over $\boldsymbol{\theta}$, $\mathbf{x_0}$ and $\boldsymbol{\sigma^2}$ respectively. Here we have assumed that the priors are independent. In the Bayesian framework, the aim is to infer this joint posterior from the noisy observations over the parameters $\boldsymbol{\theta}$, $\mathbf{x_0}$ and $\boldsymbol{\sigma^2}$.

4.4 SMC ABC

In order to improve the low acceptance rate in the basic ABC algorithm, an SMC ABC algorithm was proposed in Sisson et al. (2007), based on the SMC sampler methodology developed by Del Moral et al. (2006). Toni et al. (2009) applied an SMC ABC to estimate parameters in ODE models.

The efficiency of the SMC ABC algorithm depends not only on the model com-

Algorithm 4.1. SMC ABC Algorithm (Del Moral et al., 2006; Sisson et al., 2007)

1: Initialise $\epsilon_0 > 0$ and $0 < \alpha < 1$.
2: **for** $t = 0$ *to* T **do**
3: **for** $i = 1$ *to* N **do**
4: **if** $t = 0$ **then**
5: Sample $\boldsymbol{\theta}^{**}$ from $p(\boldsymbol{\theta})$.
6: **else**
7: Sample $\boldsymbol{\theta}^{*}$ from the previous population $\boldsymbol{\theta}_{t-1}^{(i)}$ with normalised weights $w_{t-1}^{(i)}$ and use a perturbation kernel K_t to sample $\boldsymbol{\theta}^{**} \sim K_t(\cdot|\boldsymbol{\theta}^{*})$.
8: **end if**
9: **if** $p(\boldsymbol{\theta}^{**}) = 0$ **then**
10: Go to line 4.
11: **else**
12: Simulate $\mathbf{z}^{*}$ from model given $\boldsymbol{\theta}^{**}$.
13: **if** $\rho(\mathbf{z}^{*}, \mathbf{y}) \geq \epsilon_t$ **then**
14: Go to line 4.
15: **else**
16: Set $\boldsymbol{\theta}_t^{(i)} = \boldsymbol{\theta}^{**}$ and calculate the weight for particle $\boldsymbol{\theta}_t^{(i)}$:

$$w_t^{(i)} = \begin{cases} 1, & \text{if } t = 0, \\ \frac{p(\boldsymbol{\theta}_t^{(i)})}{\sum_{j=1}^{N} w_{t-1}^{(j)} K_t(\boldsymbol{\theta}_t^{(i)}|\boldsymbol{\theta}_{t-1}^{(j)})} & \text{if } t > 0. \end{cases}$$

17: **end if**
18: **end if**
19:
20: **end for**
21: Set ϵ_{t+1} to be α-quantile of saved distances vector
22: Normalise the weights.
23: **end for**
24: Return particles $\boldsymbol{\theta}_T^{(i)}$.

plexity and the amount of data available, but also on the choice of the decreasing sequence of ϵ_t (the tolerances), and the choice of perturbation kernel K_t, according to Filippi et al. (2013).

There are various ways to construct the decreasing sequence of ϵ_t, either manually or adaptively as proposed in Drovandi and Pettitt (2011); Del Moral et al. (2012); Liepe et al. (2014). In the adaptive method, the value of ϵ_1 is chosen to be large to avoid having a low acceptance rate. Then, for $t = 2, ..., T$, the tolerance ϵ_t is selected to be the α-th quantile of the discrepancies between the observed data and the simulated data that was generated in the $(t-1)^{th}$ population, where $0 \leq \alpha \leq 1$. The resulting method is summarised in Algorithm 4.1. Throughout this paper, we used this method of selecting the sequence of tolerance thresholds with $\alpha = 0.1$ and we stopped the Algorithm when we reached a final ϵ_t that achieved the desired agreement between simulated and real data. The choice of perturbation kernel K_t affects the acceptance rate in SMC ABC and the time consumed by the algorithm as explained in Filippi et al. (2013). Perturbation kernels can be divided into two classes: component-wise perturbation kernels and multivariate perturbation kernels. For component-wise perturbation kernels, one can use a uniform distribution or a univariate Gaussian distribution to perturb a particle $\boldsymbol{\theta}^*$ sampled from the previous population $\{\boldsymbol{\theta}_{t-1}^{(i)}\}_{i=1}^{N}$, where N is the number of particles in population $t-1$. The standard deviation of the kernel can be fixed in advance for each population, but more recently practitioners are adaptively choosing the width of the kernel. For a detailed discussion, the reader is referred to M. Beaumont et al. (2009); Didelot et al. (2011); Filippi et al. (2013).

In practice when implementing Algorithm 4.1, a stopping criterion is required to terminate the process. One simple approach is to stop the algorithm when it reaches a specified ϵ_T close to zero or target acceptance rate that is calculated in each iteration by finding the ratio of the number of accepted particles to the total number of simulations produced (Abdessalem et al., 2018). An alternative

stopping criterion is to use a specified number of total simulations as a tuning parameter (Prangle et al., 2017).

4.5 Modified SMC ABC Algorithm

This paper concentrates on improving the estimation procedure when using SMC ABC to estimate ODE parameters. We propose a novel method to estimate the noise within the SMC ABC approach when the scale $\boldsymbol{\sigma^2}$ is unknown and needs to be inferred. Using this new method we accurately estimate the true joint posterior distribution of the parameters including the noise. The new method is a modification to the standard SMC ABC algorithm summarised in Algorithm 4.1. In particular, we have added a new condition to improve the handling of the proposed noise parameters.

In our modified SMC ABC approach, we use a new summary statistic that represents the noise associated with the observation $\mathbf{y_k}$. In Equation (5.2), the noisy observations are the sum of the deterministic solution to the ODE model with parameter $\boldsymbol{\theta}$ (which we want to infer) and some additive normal noise $\boldsymbol{\delta}_k \sim MVN(\mathbf{0}, \Sigma(\sigma^2))$. Given $\mathbf{y_k}$, we aim to approximately sample the posterior distribution for $\boldsymbol{\theta}$ and $\boldsymbol{\sigma^2}$. Therefore, given proposals of the parameter values $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\sigma}}^{\mathbf{2}}$ from their prior distributions we can generate noisy simulated data $\mathbf{y}_{sim}$ as:

$$\mathbf{y}_{sim} = \mathbf{z}(t; \hat{\boldsymbol{\theta}}, \mathbf{x}_0) + \boldsymbol{\delta}_{sim}, \tag{4.6}$$

where $\mathbf{z}(t; \hat{\boldsymbol{\theta}}, \mathbf{x}_0)$ is the deterministic simulation generated by the ODE model and $\boldsymbol{\delta}_{sim}$ is the additive normal noise generated by using a proposal value of $\hat{\boldsymbol{\sigma}}^{\mathbf{2}}$. In the standard SMC ABC method, the proposed $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\sigma}}^{\mathbf{2}}$ are accepted if $\rho(\mathbf{y}_k, \mathbf{y}_{\boldsymbol{sim}}) < \epsilon$, for a chosen $\epsilon > 0$. However, our modified SMC ABC algorithm

Algorithm 4.2. Modified SMC ABC Algorithm

1: Initialise ϵ_0¿0 and $0 < \alpha < 1$.
2: **for** $t = 0$ *to* T **do**
3: **for** $i = 1$ *to* N **do**
4: **if** $t = 0$ **then**
5: Sample $\boldsymbol{\theta}^{**}$ and $\boldsymbol{\sigma^2}^{**}$ from $p(\boldsymbol{\theta})$ and $p(\boldsymbol{\sigma^2})$, respectively.
6: **else**
7: Sample $\boldsymbol{\theta}^{*}$ and $\boldsymbol{\sigma^2}^{*}$ from the previous populations $\boldsymbol{\theta}_{t-1}^{(i)}$ and $\boldsymbol{\sigma^2}_{t-1}^{(i)}$, respectively, with their corresponding normalised weights $w_{t-1}^{(i)}$ and use a perturbation kernel K_t to sample $\boldsymbol{\theta}^{**} \sim K_t(\cdot|\boldsymbol{\theta}^{*})$ and $\boldsymbol{\sigma^2}^{**} \sim K_t(\cdot|\boldsymbol{\sigma^2}^{*})$.
8: **end if**
9: **if** $p(\boldsymbol{\theta}^{**}) \times p(\boldsymbol{\sigma^2}^{**}) = 0$ **then**
10: Go to line 4.
11: **else**
12: Simulate $\mathbf{z}^{*}$ from model given $\boldsymbol{\theta}^{**}$, then generate noisy simulation $\mathbf{y_{sim}}$ by adding normal noise $\boldsymbol{\delta_{sim}} \sim MVN(\mathbf{0}, \Sigma(\boldsymbol{\sigma^2}^{**}))$ to the simulation $\mathbf{z}^{*}$.
13: Find $\boldsymbol{\delta_{diff}}$ by calculating the difference between $\mathbf{z}^{*}$ and $\mathbf{y_k}$, then find the sample variance $\boldsymbol{\sigma^2}_{diff}$ of $\boldsymbol{\delta_{diff}}$.
14: Calculate F-score by finding the ratio between $\boldsymbol{\sigma^2}^{**}$ and $\boldsymbol{\sigma^2}_{diff}$ and find the associated p-value.
15: **if** p-value ≥ 0.05 **then**
16: **if** $\rho(\mathbf{y_{sim}}, \mathbf{y_k}) < \epsilon_t$ **then**
17: Accept $\boldsymbol{\sigma^2}^{**}$ and Set $\boldsymbol{\theta}_t^{(i)} = \boldsymbol{\theta}^{**}$ and calculate the weight for the particle $\boldsymbol{\theta}_t^{(i)}$ and $\boldsymbol{\sigma^2}^{**}$:
$$w_t^{(i)} = \begin{cases} 1, & \text{if } t = 0, \\ \frac{p(\boldsymbol{\theta}_t^{(i)})}{\sum_{j=1}^{N} w_{t-1}^{(j)} K_t(\boldsymbol{\theta}_t^{(i)}|\boldsymbol{\theta}_{t-1}^{(j)})} & \text{if } t > 0. \end{cases}$$
18: **else**
19: Go to line 4.
20: **end if**
21: **end if**
22: **end if**
23: **end for**
24: Set ϵ_{t+1} to be α-quantile of saved distances vector
25: Normalise the weights.
26: **end for**
27: Return particles $\boldsymbol{\theta}_T^{(i)}$ and all accepted $\boldsymbol{\sigma^2}^{**}$.

decides whether to accept or reject the proposed $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\sigma}}^{\mathbf{2}}$ using a new summary statistic that compares the distribution of the differences between the observations and $\mathbf{z}(t; \hat{\boldsymbol{\theta}}, \mathbf{x}_0)$,

$$\boldsymbol{\delta}_{diff} = \mathbf{y}_k - \mathbf{z}(t; \hat{\boldsymbol{\theta}}, \mathbf{x}_0), \tag{4.7}$$

to the distribution of $\boldsymbol{\delta}_{sim}$. If the two distributions are similar, then the proposed value $\hat{\boldsymbol{\sigma}}^{\mathbf{2}}$ has generated a similar pattern of noise to the true value associated with the observations. In the proposed approach, we use an F-test as a way to accept or reject the proposed $\hat{\boldsymbol{\sigma}}^{\mathbf{2}}$. One reason to choose an F-test is our assumption that the additive noise is independent and normally distributed, in which case the variance ratio of $\boldsymbol{\delta}_{sim}$ and $\boldsymbol{\delta}_{diff}$ has a Fisher–Snedecor F-distribution (Forbes, Evans, Hastings, & Peacock, 2011). It is important to mention that in all the examples used in this paper we calculate the sample variance of $\boldsymbol{\delta}_{sim}$ and $\boldsymbol{\delta}_{diff}$ under the assumption that the mean of the additive noise is known and equal to the zero vector.

The new acceptance criteria is based on the associated p-value: the Algorithm rejects proposals of the parameters and $\hat{\boldsymbol{\sigma}}^{\mathbf{2}}$ when the p-value is less than or equal to 0.05 and $\rho(\mathbf{y}_k, \mathbf{y}_{\boldsymbol{sim}}) < \epsilon$. By repeating the process thousands of times with different parameters and $\hat{\boldsymbol{\sigma}}^{\mathbf{2}}$ values for t populations, we generate samples that approximate the joint posterior distribution of the parameters, including the noise parameter. Algorithm 4.2 describes the overall procedure, which we refer to as Modified SMC ABC.

4.6 Illustrations of the Algorithm

To test this new Modified SMC ABC, we applied it to two ODE epidemiological models with simulated data. In each case, we compared its results to MCMC

and the standard SMC ABC method and highlighted the benefits of using our proposed method.

4.6.1 Test Problem 1 — SIR Model

The SIR model is an ODE system that describes the spread of an infectious disease in a large population. The SIR model assumes the population consists of three types of individual: susceptible, infected and recovered. The number of susceptible individuals is S, the number of infected individuals is I and the number of recovered or removed individuals is R, where $P = S + I + R$ is the total population. In this test problem, we use the SIR model to represent the fraction of the total population (P) in each category as follows:

$$s(t) = \frac{S(t)}{P}, i(t) = \frac{I(t)}{P} \text{ and } r(t) = \frac{R(t)}{P}.$$

The SIR system, with constant population, is given by

$$\begin{aligned} \frac{ds}{dt} &= -\beta s(t) I(t) \\ \frac{di}{dt} &= (\beta s(t) - \gamma) i(t) \\ \frac{dr}{dt} &= \gamma i(t) \end{aligned} \tag{4.8}$$

where $\boldsymbol{\beta} > 0$ is the disease transmission rate and $\boldsymbol{\gamma} > 0$ is the recovery rate.

Simulation Results

We generated a test data set by solving the system Equation (4.8) over the time interval $[0, 50]$ (weeks) using a 4th order Runge-Kutta method and storing the

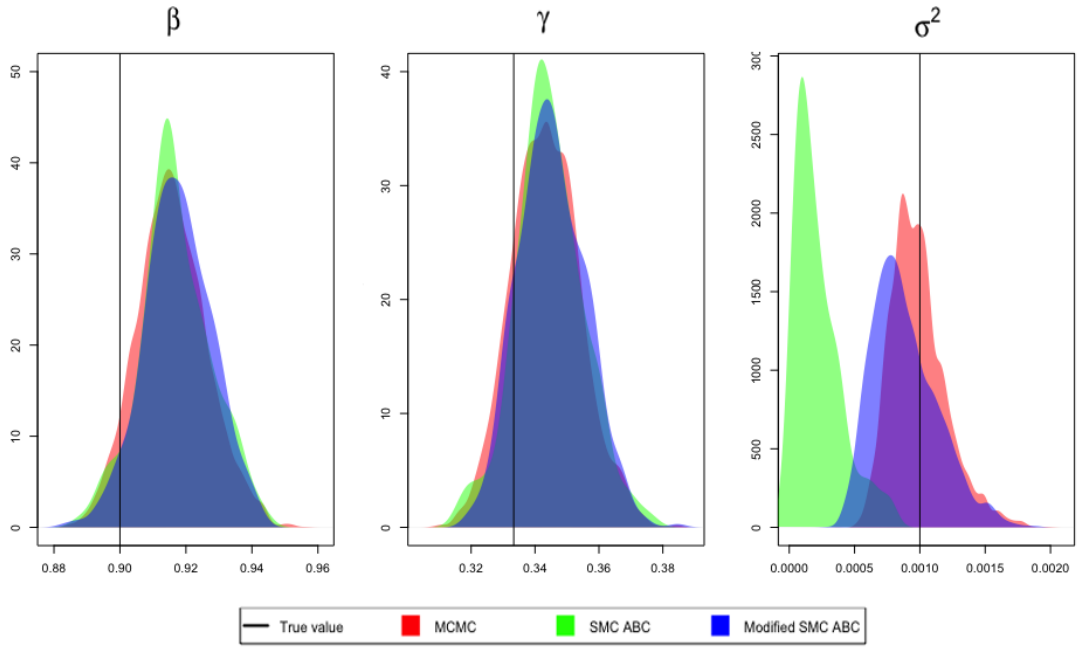

Figure 4.1: Estimated marginal posterior distributions for β, γ and σ^2 produced using MCMC (red) SMC ABC (green) and Modified SMC ABC (blue).

solution at weekly intervals, using true model parameters $\boldsymbol{\theta} = (\beta = 0.9, \gamma = \frac{1}{3})^T$. The $K = 50$ observations $\mathbf{y_k}$, $(k = 1, \ldots, K)$, were constructed by adding normal noise $\mathcal{N}(0, \sigma^2 = 0.001)$ to the proportion of the population infected at time t_k. For ABC approaches, we used a discrepancy function $\rho(\mathbf{y}_{sim}, \mathbf{y_k})$ to compare infected proportions in the data set $\mathbf{y_k}$ with a noisy simulation $\mathbf{y}_{sim} = (y^1_{sim}, \ldots, y^K_{sim})$ for proposed parameters as follows:

$$\rho(\mathbf{y}_{sim}, \mathbf{y_k}) = \frac{1}{K} \sum_{k=1}^{K} (y^k_{sim} - y_k)^2 \tag{4.9}$$

where n is the number of observed data points. The priors for β, γ and σ^2 were

taken to be:

$$\begin{aligned} \beta &\sim \mathcal{U}(0,2) \\ \gamma &\sim \mathcal{U}(0,2) \\ \sigma^2 &\sim \mathcal{U}(0,1) \end{aligned} \tag{4.10}$$

where $\mathcal{U}(\cdot,\cdot)$ is the Uniform distribution.

Given the observations $\mathbf{y_k}$, the parameter vector $\boldsymbol{\theta} = \{\beta, \gamma, \sigma^2\}$ was estimated using MCMC, standard SMC ABC and Modified SMC ABC. When applying the MCMC approach, normal proposal distributions were used with an adaptive approach that tuned parameters in the algorithm to maintain an appropriate acceptance ratio between 0.3 and 0.5, as suggested by Gelman and Rubin (1992).

Then, we performed the standard SMC ABC for $T = 5$ populations. Following Line 5 in Algorithm 4.1, we sampled β^*, γ^* and σ^{2*} from their priors and then we used σ^{2*} to add normal noise $N(0, \sigma^{2*})$ to the simulation generated by solving Equation (4.8), using the samples (β^*, γ^*) as values for the ODE parameters. Consequently, we obtain a noisy simulation $\mathbf{y}_{sim}$ instead of a deterministic simulation. Then we compute the distance between $\mathbf{y}_{sim}$ and $\mathbf{y_k}$ using a discrepancy function $\rho(\mathbf{y}_{sim}, \mathbf{y_k})$, as in Equation (4.9). At Line 13 of Algorithm 4.1 we accept the samples $(\beta^*, \gamma^*, \sigma^{2*})$ that make $\rho(\mathbf{y}_{sim}, \mathbf{y_k}) \leq \epsilon_t$. The tolerances were chosen adaptively as discussed in Section 5.4 to be $\epsilon = (0.927, 0.500, 0.339, 0.277, 0.250)$.

We applied the Modified SMC ABC, also for $T = 5$. Samples of $\beta^*, \gamma^*$and σ^{2*} were accepted when they passed the conditions on Lines 15 and 16 in Algorithm 4.2. The adaptive tolerance sequence for modified SMC ABC was $\epsilon = (0.923, 0.498, 0.339, 0.282, 0.255)$.

Results obtained using the standard SMC ABC and the Modified SMC ABC were compared with results derived from an MCMC approach: see Figure 4.1. The improved handling of the noise in the modified SMC ABC approach has improved the accuracy of the approximation to the marginal posterior for β and γ and pro-

		True value	MCMC	Modified SMC ABC	SMC ABC
Median	β	0.9	0.893	0.897	0.895
	γ	0.333	0.327	0.331	0.330
	σ^2	0.001	0.0011	0.0011	0.00016
95% CI	β		$(0.87, 0.92)$	$(0.87, 0.93)$	$(0.87, 0.92)$
	γ		$(0.31, 0.35)$	$(0.30, 0.36)$	$(0.30, 0.36)$
	σ^2		$(0.0008, 0.0018)$	$(0.0007, 0.0019)$	$(6 \times 10^{-}6, 9 \times 10^{-}4)$
CPU time			1.69 mins	126.55 mins	106.70 mins
Iterations			8759	626040	532979

Table 4.1: Posterior median and 95% credible intervals for all parameters in the SIR model, number of iterations and computational times for MCMC, Modified SMC ABC and SMC ABC.

duced results similar to those derived from MCMC. All the methods should ideally sample approximately the same posterior distribution; however, the approximate posterior distribution of σ^2 derived from the standard SMC ABC is noticeably different from those obtained using MCMC or Modified SMC ABC and the high density region does not even contain the true σ^2. This poor approximation also affected the posterior predictive distribution for the standard SMC ABC, as shown in Figure 4.2. The posterior predictive distribution derived from standard SMC ABC is narrower than the others and does not cover the appropriate number of observations. By contrast, the posterior predictive plots obtained using MCMC and Modified SMC ABC covers almost all the observations.

Table 4.1 shows that the ODE parameters were estimated accurately and all methods have produced similar credible intervals. However, the standard SMC ABC failed to approximate the true value of σ^2, while the values estimated by MCMC and Modified SMC ABC are accurate. Both SMC ABC and the Modified SMC ABC consumed more time (106.7, 126.55, respectively), compered to MCMC (1.69), because of the need to perform a large number of iterations in the SMC ABC process.

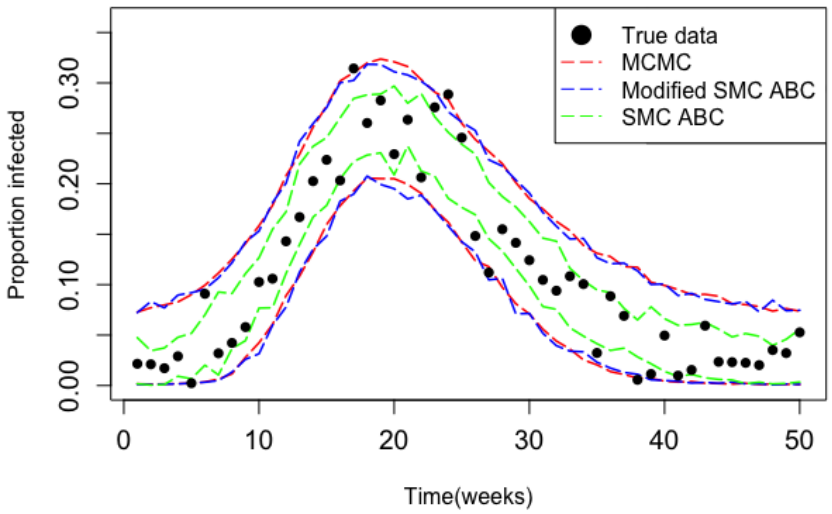

Figure 4.2: Posterior predictive 95% credible intervals estimated using MCMC, SMC ABC and Modified SMC ABC plotted with the weekly number of infected cases. The black dots represent the noisy data. The shaded areas are created using posterior predictive samples. It can be seen that the result derived from MCMC and Modified SMC ABC cover most of the data points while the SMC ABC result produces unrealistically narrow credible intervals.

4.6.2 Test Problem 2 — Non-Linear ODE Model of Malaria Transmission

Work by White, Maude, Pongtavornpinyo, et al. (2009), acknowledging the lack of reliable data in some countries where malaria control or elimination is particularly desirable, showed the utility of a compartmental model in predicting the effects of various elimination strategies compared to more complex models of Gu, Killeen, Mbogo, et al. (2003) and Maire, Tediosi, Ross, and Smith (2006). The model describes population dynamics using four population compartments in the transmission of malaria:

$S(t)$: Uninfected and non-immune,
$I_1(t)$: Infected with no-prior immunity,
$R(t)$: Uninfected with immunity,
$I_2(t)$: Infected with prior immunity.

The model comprises 4 ODEs that govern the temporal evolution of the population compartments. The model can be described mathematically by the following equations:

$$\begin{aligned}
\frac{dS}{dt} &= \frac{P}{L} - \left(\lambda + \frac{1}{L}\right) S + \frac{1}{d_{imm}} R \\
\frac{dI_1}{dt} &= \lambda S - \left(\frac{\eta_0 p_1}{d_{treat}} + \frac{1-\eta_0 p_1}{d_{in}} + \frac{1}{L}\right) I_1 \\
\frac{dI_2}{dt} &= \lambda R - \left(\frac{\eta_0 p_2}{d_{treat}} + \frac{1-\eta_0 p_2}{d_{in}} + \frac{1}{L}\right) I_2 \\
\frac{dR}{dt} &= \left(\frac{\eta_0 p_1}{d_{treat}} + \frac{1-\eta_0 p_1}{d_{in}}\right) I_1 + \left(\frac{\eta_0 p_2}{d_{treat}}\right. \\
&\qquad \left. + \frac{1-\eta_0 p_2}{d_{in}} I_2 - \left(\lambda + \frac{1}{d_{imm}} + \frac{1}{L}\right) R\right.
\end{aligned} \tag{4.11}$$

Here λ is the force of infection and is given by

$$\lambda = R_0 \left(\frac{1}{L} + \frac{1}{d_{treat}}\right) \frac{(I_1 + I_2)}{P}.$$

The model is parameterised in terms of a number of constants as described in Table A.1.

The simulated data, $\mathbf{y}$, is taken to be the number of observable clinical infections

$$C_k \quad = \quad p_1 I_1(t_k) + p_2 I_2(t_k). \tag{4.12}$$

To construct the data $\mathbf{y_k}$, we added normal noise, $N(0, 10^9)$, to the clinical infections in Equation (4.12). Here, the parameter vector of interest is $\boldsymbol{\theta} = (\eta_0, d_{in}, \sigma^2)$, where η_0 is the percentage of individuals with clinical infection that receive treatment, d_{in} is the average duration of an untreated infection and σ^2 is the noise associated with the data.

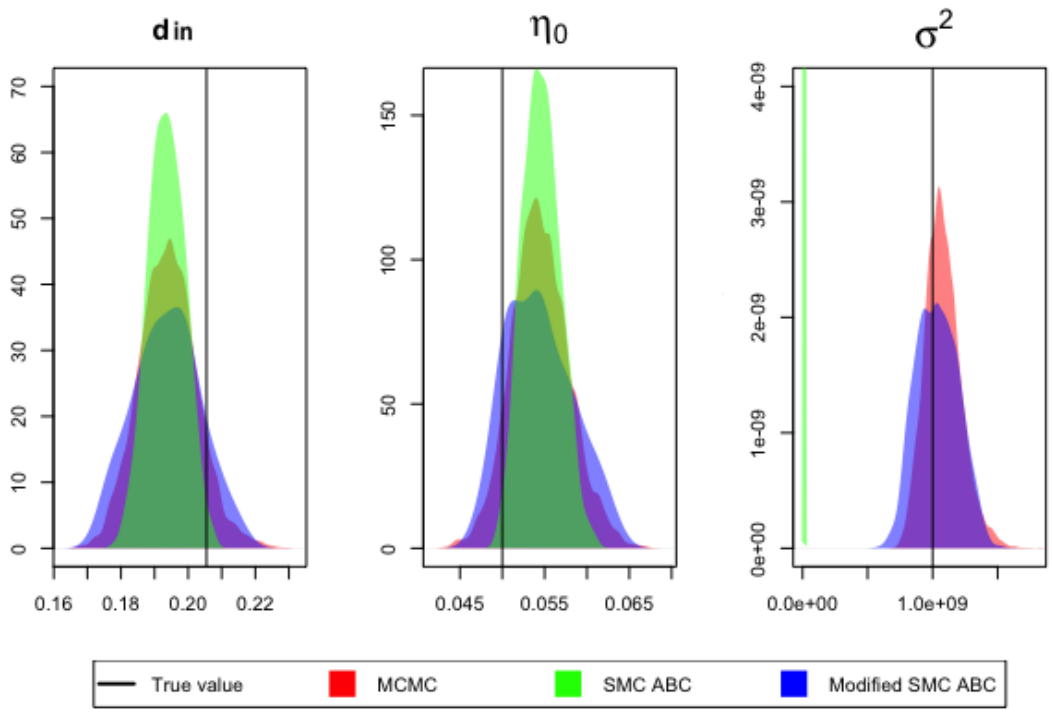

Figure 4.3: Estimated marginal posterior distributions for η_0, d_{in} and σ^2 produced using MCMC (red) SMC ABC (green) and modified SMC ABC (blue).

Simulation Results

A data set of 121 simulated points was generated by solving Equations (4.11) in the interval $[0 - 10]$ years using a 4th order Runge-Kutta method. The values of the model parameters used are shown in Table A.1 and the initial conditions are

given by the equilibrium solution of the system in Equation (4.11).

parameter		*value*	*source*
P	[People]	3.2×10^6	Assumed
L	[Years]	66.67	Maude, Pongtavornpinyo, et al. (2009)
d_{imm}	[Years]	0.93	Aguas, White, Snow, and Gomes (2008)
d_{in}	[Years]	0.2	Assumed
d_{treat0}	[Weeks]	2	Maude et al. (2009)
p_1		0.87	Aguas et al. (2008)
p_2		0.08	Aguas et al. (2008)
η_0		0.05	Assumed

Table 4.2: The parameter values used in simulation of the White et al. (2009) model.

As with the first test problem, for the ABC approaches we used the discrepancy function in Equation 4.9 to compare the clinical infections given in the data set $\mathbf{y_k}$ with the noisy simulations $\mathbf{y_{sim}}$. The priors for η_0, d_{in} and σ^2 were taken as follows:

$$\begin{aligned} p(\eta_0) &= \mathcal{B}(1,1) \in [0,1], \\ p(d_{in}) &= \mathcal{GA}(1,1) \in [0,\infty), \\ p(\sigma^2) &= \mathcal{IG}(1,1) \in [0,\infty). \end{aligned} \tag{4.13}$$

For the MCMC approach, a logistic transformation was used to transform $\eta_0 \in [0,1]$ while a log transform was applied to $d_{in} \in [0,\infty)$ and $\sigma^2 \in [0,\infty)$, so that each transformed parameter had support over the real line.

We first applied the MCMC method to estimate $\boldsymbol{\theta} = (\eta_0, d_{in}, \sigma^2)$. We ran the sampler until the chain converged, which needed 32349 iterations and approximately 68.25 minutes. The convergence was assessed using MCMC convergence diagnostics described in Gelman et al. (2014). The median of the sampled values for each parameter is shown in Table 4.3.

We applied standard SMC ABC as described in Algorithm 4.1 for $T = 4$ with adaptive tolerances, $\epsilon = (2106083.6, 432384.4, 372201.2, 356149.6)$. In order to get

		True value	MCMC	Modified SMC ABC	SMC ABC
Median	η_0	0.05	0.054	0.055	0.055
	d_{in}	0.205	0.193	0.192	0.193
	σ^2	10^9	10^9	9×10^8	5×10^7
95% CI	η_0		$(0.048, 0.062)$	$(0.048, 0.068)$	$(0.048, 0.062)$
	d_{in}		$(0.177, 0.212)$	$(0.167, 0.222)$	$(0.178, 0.208)$
	σ^2		$(83 \times 10^7, 13 \times 10^8)$	$(67 \times 10^7, 12 \times 10^8)$	$(18 \times 10^5, 25 \times 10^7)$
CPU time			68.25 mins	300.71 mins	261.43 mins
Iterations			32349	155759	143003

Table 4.3: Posterior median and 95% credible intervals (CIs) for parameters in the malaria model, number of iterations and computation times for MCMC, SMC ABC and Modified SMC ABC.

1000 accepted values we needed 143003 simulations, which took 261.43 minutes to produce.

Finally, we applied our Modified SMC ABC for $T = 4$ populations and generated 1000 accepted particles. This required 300.71 minutes and 155759 simulations. The algorithm adapted the values of the tolerances as $\epsilon = (2106083.6, 472757.5, 419652.1, 393456.2)$.

As in the first example, MCMC and Modified SMC ABC produced good approximations to all parameters. However, standard SMC ABC failed to estimate the σ^2 value accurately, as shown in Table 4.3 and Figure 4.3. In this example, the effect of misspecifing the noise model is clearer, given that the noise associated with the observation is large and thus the approximate posterior predictive distribution generated by standard SMC ABC is unrealistically narrow compared to results obtained using MCMC and Modified SMC ABC, as shown in Figure 4.4.

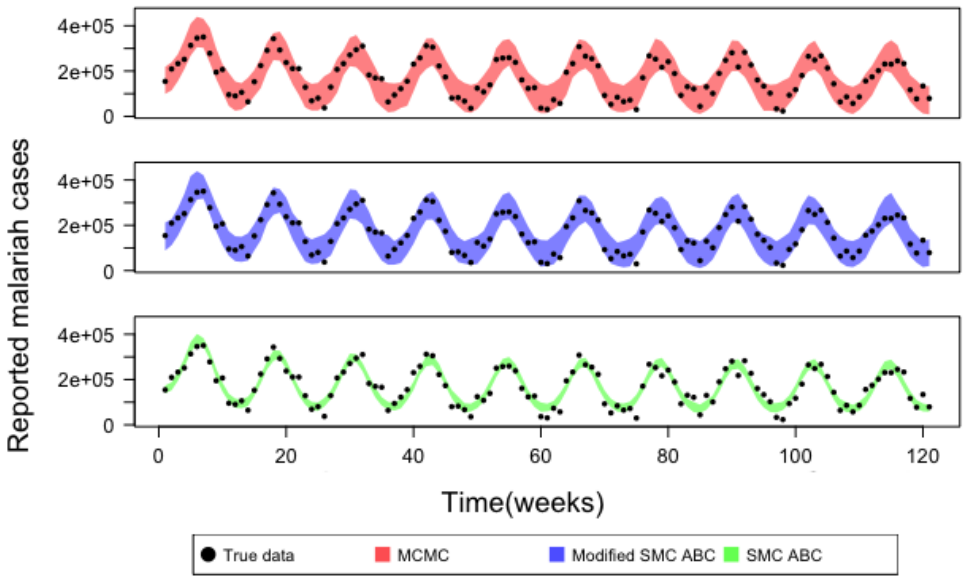

Figure 4.4: Plot of the posterior predictive 95% credible intervals estimated using MCMC, SMC ABC and Modified SMC ABC fitted with the weekly infection cases. The black dots represent the noisy data. The shaded areas are created using the posterior predictive samples. It can be seen that the result derived from MCMC and modified SMC ABC covered most of the data points while the SMC ABC derived result produces unrealistically narrow credible intervals.

4.7 Discussion

In this paper we have proposed a method to significantly improve the approximation of the posterior distribution of the ODE parameters and especially the noise parameter when using an SMC ABC approach.

We achieve this by introducing a new summary statistic that guides the acceptance of the noise parameter proposed when applying the SMC ABC algorithm. This new summary statistic increases the probability that only samples from the true posterior distribution of the noise can be accepted. The key idea behind our method lies in deciding whether the distribution of the differences between the

noisy observations and the deterministic solution of the ODEs is likely to be similar to the distribution of the differences between the noisy simulations and the deterministic solution, using an F-test. For the two examples presented in this paper, the Modified SMC ABC has improved posterior accuracy for both the ODE parameters and the noise parameter, and gives results comparable to MCMC. The posterior predictive checks show that both methods - modified SMC ABC and MCMC - cover an appropriate number of observations. Interestingly, the presence of the noise in the standard SMC ABC approach does allow the standard SMC ABC algorithm to achieve an approximation to the parameter posterior distribution similar to that generated using MCMC. However, it fails to give an accurate estimate of the posterior distribution of the noise, which significantly affects the posterior predictive distribution, especially when the noise is large as in Test Problem 2. The new SMC ABC approach that we have presented in this paper is promising for researchers who would prefer to use ABC methods for parameter inference in ODE models.

Chapter 5

Low Discrepancy Sequences for Bayesian Estimation in Ordinary Differential Equations

Preamble

The purpose of this chapter is to develop a new method for exploring the parameter space of non-linear ODEs more efficiently using Quasi Monte Carlo (QMC) point sets. Our results show that the QMC method out-performs Markov chain Monte Carlo (MCMC) in terms of computational time and accuracy of estimation, particularly when the posterior distribution is multi-modal. Furthermore, it is easy to implement relative to MCMC. In this way, we have satisfied Objective 3. The proposed method is tested on low-dimensional parameter spaces, 3-5 parameters, in a Lotka-Volterra model and an epidemiological compartmental model which contributes to Objective 4.

5.1 Abstract

Ordinary differential equations (ODEs) are a popular mathematical tool for modelling dynamic processes. Typically, ODEs are associated with parameters that need to be estimated, often based on noisy data. Bayesian approaches to such estimation problems often involve sampling parameters from a posterior distribution using a computational method such as Markov chain Monte Carlo (MCMC). However, a poorly designed MCMC method can suffer from slow mixing and poor convergence. This problem is exacerbated when the parameter space is of high dimension, parameters are unidentifiable or the posterior distribution is multimodal. In this paper, we propose a new method that has two contributions. First, we explore the parameter space of non-linear ODEs more efficiently using Quasi Monte Carlo (QMC) point sets. Second, we propose a new and easy method to visualise posterior marginals using cumulative summation. Our results show that the QMC method outperforms MCMC in terms of computational time, accuracy of estimation (particularly when the posterior distribution is multi-modal) and ease of implementation relative to MCMC. The proposed method is tested on low-dimensional parameter spaces (3-5 parameters) in a Lotka-Volterra model and an epidemiological compartmental model.

5.2 Introduction

In many fields, modelling by ordinary differential equations (ODEs) is an important method for building a better understanding of a physical system and for accurately predicting its future behaviour. Some parameters of differential equations can be directly measured; many others, however, must be inferred from data. Thus, unknown parameter values need to be estimated via statistical approaches. This inverse problem in ODEs is computationally challenging and time-consuming

because many current approaches involve explicitly solving the ODEs many times for different sets of candidate parameters. This problem is a feature of the method of maximum likelihood (a non-Bayesian approach) and of Bayesian approaches involving Markov chain Monte Carlo (MCMC) methods (such as Gibbs sampling, Metropolis-Hastings and Hamiltonian MCMC).

A common approach to estimating ODE parameters is to use least squares to minimise the differences between the observations and the numerical solution of the ODEs Macdonald, Higham, and Husmeier (2015). On the other hand, some approaches avoid solving the ODEs numerically to reduce computational time, using a Gaussian process to model the trajectory of state variables. In the next section we review recent methodologies for parameter inference in ODE models.

5.2.1 Statistical parameter estimation in ODE models

Varah (1982) suggested a two-stage method. In the first stage, least squares regression is used to fit the given data using cubic spline functions with fixed knots. The resulting model is then differentiated with respect to time. In the next step, the unknown parameters are estimated by minimising the difference between the ODE model and the estimated derivative of the solution. This technique has been developed by Ramsay (2006) and Qi and Zhao (2010) to iterate between smoothing the data and estimating the ODE parameters under investigation. Ramsay (1996) proposed a method known as principal differential analysis (PDA) for estimation of dynamic parameters. This technique works by fitting discrete data using a spline model; then the estimated values are substituted into an ODE, and the differential equation parameters are estimated with a simple least squares procedure.

Ramsay, Hooker, Campbell, and Cao (2007) proposed a generalised profiling approach to estimate ODE parameters. The computational burden is considerably

lower than other methods as it does not involve the computational cost of finding a numerical ODE solution. Qi and Zhao (2010) explored the statistical properties of this approach, such as consistency and asymptotic normality. Campbell and Steele (2012) proposed a new population MCMC method for posterior estimation of a parameter vector using the parallel tempering algorithm and a generalized profiling approach, calling this the *smooth functional tempering* approach. Peng, Li, Yang, and Wang (2009) established a novel approach based on integration theory which used the idea that the solution to the ODE system is constrained to a particular function with parameters to be estimated.

Calderhead, Girolami, and Lawrence (2009) fitted a Gaussian Process (GP) with hyperparameters to the data before estimating the parameters of the ODEs. This method is based on the product of experts approach and marginalisation over the state derivatives. The advantage of this method is that all parameters can be estimated from the data; however, the estimation accuracy of the GP hyperparameters relies only on the data without feedback from the ODE system Wang and Barber (2014). To improve this method, Dondelinger et al. (2013) introduced a bidirectional link between the ODE and GP parameters. Wang and Barber (2014) improved these two approaches, establishing a simpler generative model (GP-ODE) that connects the state derivatives to the system's observations directly using a GP. Currently, there are issues related to the efficiency of the existing GP-ODE approach. To simplify the method, the authors of the GP-ODE approach have been equating the elimination of the state variable with its marginalisation. As a consequence, their method suffers from identifiability problems, especially when data are systematically missing, as Macdonald et al. (2015) pointed out.

Bhaumik, Ghosal, et al. (2015) proposed a two-step Bayesian method of parameter estimation. In this approach, the posterior distribution of parameters is approximated using a restricted random series based on B-spline basis functions. The parameters are approximated by minimising the distance between the non-

parametrically estimated derivative and the derivative proposed by the ODE. This method has been extended to a higher order ODE model Bhaumik and Ghosal (2015).

Dass, Lee, Lee, and Park (2016) proposed a two-step method to approximate the posterior distribution of the unknown parameters. In the first step, data are generated from the ODE using a numerical method and then the second step uses the Laplace approximation to marginalise the posterior for each parameter. This method gives a fast approximate approach compared to a full Bayesian computational scheme.

In most current methods the hyperparameter space is explored using either a pseudo-random set of points or a deterministic grid of points. The problem with using pseudo-random points is that the points are not evenly distributed throughout the parameter space and tend to form clusters and gaps. To improve the approximation accuracy, a large number of random points is needed. Each additional point increases the computational time required to solve the ODEs numerically. On the other hand, point sets generated based on a grid suffer from three critical problems. The first problem is that the choice of grid matrix can be a complex task Joshi and Wilson (2011). For example, if we use the approach of Rue, Martino, and Chopin (2009) to choose a grid set, we have to find the posterior mode and then find a support interval around it, but in real world problems usually the mode of the posterior is unknown. It is true that the posterior mode can be identified using numerical methods; however, in some applications, these methods can increase the computational burden, particularly when the number of parameters is large. The second problem is that even if we can determine the support interval, evaluating the marginal posterior on the grid points can fail to capture the shape of the true posterior distribution when the distribution is multi-modal or skewed Joshi, Brown, and Joe (2016). The third issue is that the number of grid points increases exponentially with the number of parameters that need to be estimated.

Standard MCMC algorithms such as the random walk Metropolis-Hastings (RWMH) sampler explore the parameter space by making local moves Chumbley et al. (2007). However, RWMH can converge slowly to the target posterior density when the number of parameters is large Sengupta et al. (2016); Feng and Li (2015). In addition, unidentifiable parameters in nonlinear models may also cause slow convergence and mixing of MCMC algorithms Kim and Li (2012). Such poor mixing can greatly increase the computational burden, due to the need to explicitly solve the ODE numerically for each proposal of the parameter of interest Wang and Barber (2014). Multimodality in the posterior distribution is often the reason MCMC methods have poor convergence or even fail to converge Neal (1993, 2012); Celeux et al. (2000); Neal (2001); Rudoy and Wolfe (2006); Sminchisescu and Welling (2007); Craiu et al. (2009). When the posterior distribution is multi-modal, MCMC methods may fail to traverse low probability regions between modes Lan et al. (2014).

For most practitioners aiming to solve inverse problem in ODEs using Bayesian computation, methods based on MCMC appear to be the default. However, in this paper we argue that QMC is a more appropriate default method for such problems and can outperform MCMC in many aspects.

QMC methods have previously been used in solving the forward problem in differential equation models. For example, Coulibaly and Lécot (1999); Lécot and Koudiraty (2001) generalized the Runge-Kutta Monte Carlo algorithm to apply to ODEs and in addition they used the Runge-Kutta (quasi) Monte Carlo method. In the context of partial differential equations (PDEs), Dick, Gantner, Le Gia, and Schwab (2019) used higher order QMC methods to improve convergence rates regardless of the dimension of the parameter space, and performed a mathematical analysis of the resulting algorithm, but, instead of addressing the problem of locating the high density region in the parameter space, the authors assumed it as pre-specified. However, work by Schillings and Schwab (2016) used Quasi-Newton methods to identify computationally the high concentration of the posterior, then,

applied dimension-adaptive Smolyak quadratures on PDEs. A survey of QMC methods in computational statistics is provided in Shaw (1988).

The aim of this paper is to explore the use of the QMC approach to improve Bayesian analysis of inverse problems in ODEs. We view this as a numerical integration problem. Several numerical integration or approximation methods, such as those in Smith, Skene, Shaw, Naylor, and Dransfield (1985); Tierney and Kadane (1986); Smith, Skene, Shaw, and Naylor (1987), require careful choices of likelihood and prior in order for the evaluation of such integrals to be possible Smith and Gelfand (1992). Here we propose an easily implemented two-stage approach that first identifies an interval of the parameter space containing the highest posterior density interval. Our method is in contrast with Schillings and Schwab (2016), in which the uncertainty around the marginal posterior distributions is first identified using predetermined points generated using low discrepancy sequences (LDS) and then numerical integration is applied over that interval. The advantage of our proposed method is that we need fewer points to accurately approximate the posterior than MCMC methods. Moreover, using a pre-determined point set allows parallel evaluation of the posterior. Consequently, we significantly reduce the time spent iteratively solving the ODE numerically compared to MCMC or grid-based approaches. The accuracy achieved using LDS is comparable to that obtained using either random points or grid points. Furthermore, we propose to use the cumulative summation of the normalised posterior generated using QMC points in Stage 2 to visualize the marginal posteriors. This produces remarkably similar posterior approximations to those that result from an MCMC approach.

The rest of the paper is organised as follows. In Section 5.3 a brief introduction to Bayesian inference and Quasi Monte Carlo are provided. In Section 5.4 the proposed methodology and the posterior computation are described in further detail. In Section 5.5 we examine the quality of the proposed method using simulated data sets in the Lotka-Volterra model and real data within an epidemiological

compartmental model. Discussions are presented in Section 5.6.

5.3 Mathematical background

5.3.1 Bayesian inference

We consider a continuous time Q-dimensional dynamical system described by the equation

$$\dot{x} = \frac{dx}{dt} = f(x, \theta) \tag{5.1}$$

where $x = x(t) = (x_1(t), ..., x_Q(t))$ is a vector containing Q variables all dependent on a continuous variable t (often representing time), f is a Q-dimensional smooth function, and $\theta = (\theta_1, ..., \theta_M)$ is a vector of M unknown parameters. The system is discretized, in the sense that x is evaluated only at a finite number of observation times $t_k \in \{t_1, \ldots, t_K\}$. An approximate solution $\hat{x}(t, \theta, x_0)$ to the system (5.1) can be calculated numerically given values for the parameter vector θ and an initial condition x_0. Then Q-dimensional observations $\mathbf{y}_k$ for each t_k can be obtained from this solution by adding an independent δ_k noise, such that

$$\mathbf{y}_k = \hat{x}(t, \theta, x_0) + \delta_k. \tag{5.2}$$

Under a Gaussian error model, $\delta_k \sim MVN(0, \Sigma)$, where Σ is a diagonal matrix with diagonal elements $\sigma^2 = (\sigma_1^2, \ldots, \sigma_Q^2)$. Thus $\mathbf{y}_k$ follows a multivariate normal distribution:

$$\mathbf{y}_k \sim MVN(\hat{x}(t, \theta, x_0), \Sigma). \tag{5.3}$$

In a Bayesian setting, given these noisy observations, the aim is to infer a joint posterior distribution over the parameters θ, x_0 and σ^2 if the latter is unknown,

such as

$$p(\theta, x_0, \sigma^2|\mathbf{y}_k) \propto p(\theta)p(x_0)p(\sigma^2)\mathcal{L}(\mathbf{y}_k|\hat{x}(t,\theta,x_0),\sigma^2) \tag{5.4}$$

where the likelihood is

$$\mathcal{L}(\mathbf{y}_k|\hat{x}(t,\theta,x_0),\sigma^2) = \prod_{k=1}^{K} MVN(\mathbf{y}_k;\hat{x}(t,\theta,x_0),\Sigma)) \tag{5.5}$$

and $p(\theta)$, $p(x_0)$ and $p(\sigma^2)$ are the prior distributions over θ, x_0 and σ^2 respectively.

Samples from the posterior can be drawn using Bayesian techniques such as MCMC.

5.3.2 Quasi-Monte Carlo

In this section we give a brief overview of quasi-Monte Carlo (QMC) and randomized QMC (RQMC) methods. For a full exposition, see Lemieux (2009); Leobacher and Pillichshammer (2014).

For an integrable function $f : [0,1)^M \to \mathbb{R}$, the basic Monte Carlo (MC) approximation to the integral over the $[0,1)^M$ hypercube is:

$$I = \int_{[0,1)^M} f(x)dx \qquad \text{and} \qquad \hat{I}_N = \frac{1}{N}\sum_{i=1}^{N} f(x_i) \tag{5.6}$$

where the points $x_i \sim \mathcal{U}([0,1)^M)$. The error of this approximation using MC is $\mathcal{O}(N^{-\frac{1}{2}})$ Caflisch (1998). This rate of convergence can be improved using QMC, which involves generating a deterministic sequence (called a low-discrepancy sequence) $x_1, \ldots, x_N$ such that the sum in (5.6) converges faster than the Monte

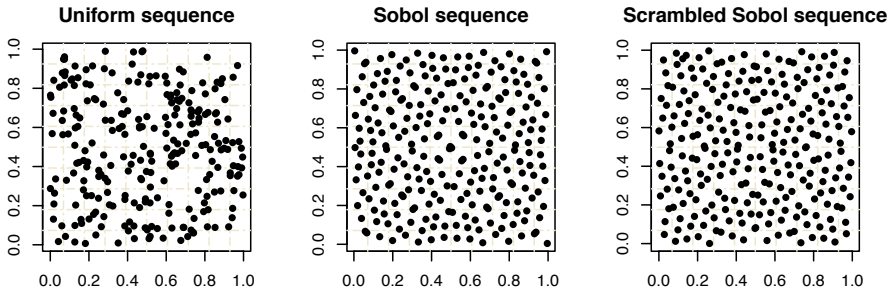

Figure 5.1: Left: uniformly random Monte Carlo (MC) points in the unit square. Centre: Sobol quasi-Monte Carlo (QMC) points. Right: Scrambled or randomized Sobol (RQMC) points. Each sequence contains 256 points generated on $[0,1]^2$. Note that QMC and RQMC points are distributed more evenly than a uniform random sequence.

Carlo estimate, as a result of which the QMC method has recently become popular. Specifically, the integration error when using QMC is bounded above by $\mathcal{O}(N^{-1}(\log N)^M)$. The enhanced rate of convergence is due to the fact that QMC sequences are distributed more evenly in the hypercube than a uniform random sequence Dick et al. (2013). See Figure 5.1 for a comparison of MC and QMC sequences on a unit square.

One way to quantify the evenness of the spread of the points is using star discrepancy, which is the distance between the discrete uniform distribution on x_i and the continuous uniform distribution on $[0,1)^M$ under a suitable metric. The star discrepancy generalises the Kolmogorov-Smirnov distance, and to define it we first consider $\Delta(a) = Vol([0,a]) - \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{x_i \in [0,a]\}$ to be the local discrepancy function at a point $a \in [0,1)^M$ where $Vol(A)$ is the M-dimensional volume of the measurable set A, $[0,a]$ is the M-dimensional hyperrectangle that has a corner at the origin and an opposing corner at $a = (a_1, \ldots, a_M)$ and $\mathbb{1}\{x_i \in [0,a]\}$ is the indicator function. Thus the volume of $[0,a]$ is $\prod_{j=1}^{M} a_j$. Then, the star discrepancy

of a point set $\mathcal{P}_N = \{x_1, \ldots, x_N\} \subset [0,1)^M$ is given by

$$D^*(\mathcal{P}_N) = \sup_{a \in [0,1)^M} \left| \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{x_i \in [0,a]\} - \prod_{j=1}^{M} a_j \right|. \tag{5.7}$$

When $D^* \to 0$, the sample mean $\hat{I}_N$ approaches the theoretical mean I given by the integral in (5.6), by a deterministic version of the law of large numbers that applies for QMC Owen and Tribble (2005). The importance of the star discrepancy arises from the Koksma-Hlawka inequality Hickernell (2014), which states that for QMC integration, if the function f has variation $V(f)$ in the sense of Hardy and Krause Hardy (1905), then

$$|\hat{I}_N - I| \leq D^*(\mathcal{P}_N) V(f). \tag{5.8}$$

From this it follows that the integration error of QMC is bounded above by $\mathcal{O}(N^{-1}(\log N)^M)$. Many examples show that QMC integration outperforms MC integration even for small N Owen and Tribble (2005); Buchholz and Chopin (2019).

Using QMC sequences is inconvenient for constructing estimators because the point construction is deterministic, however, one can reintroduce randomness and preserve the same structure of the sequence by using randomized QMC (RQMC) sequences. Then, the estimator $\hat{I}_N = \frac{1}{N} \sum_{i=1}^{N} f(x_i)$ is an unbiased estimate of the integral in (5.6) Buchholz and Chopin (2019); Wenzel et al. (2018). To obtain an RQMC sequence we used a method called scrambled nets which was introduced by Owen et al. (1997) and then modified in Owen et al. (2008).

QMC sequences can be widely classified into two groups: lattice rules, and digital nets and sequences. In all numerical examples in this paper we have generated points using a Sobol sequence, which is a kind of digital net. We generated RQMC

Sobol points using an R package called randtoolbox. As noted above, QMC points are defined over a unit hypercube, but one can use a linear transformation to define them over a hypercube with different bounds. In the following section we demonstrate how to use RQMC sequences to approximate the posterior distribution and visualise posterior marginals.

5.4 Methodology

5.4.1 Approximation to the posterior using deterministic point sets

As mentinoned in Section 5.2.1, MCMC methods may suffer from slow convergence when the number of parameters is large, if one or more parameters are unidentifiable, or if the posterior distribution has a multi-modal topology. Thus it may be necessary to run an MCMC chain for a long time to guarantee that the chain has reached stationarity, incurring a high computational burden. This problem is particularly acute when inferring the parameters of ODEs, due to the need to solve the ODE system for many proposed parameters.

In this paper, inspired by the work of Brown (2019) in which they explored some standard posterior distributions using deterministic point sets generated from a low discrepancy sequence, we propose a two stage LDS-based method for Bayesian inference to sample from a posterior distribution of the parameters of a non-linear ODE system. In the first stage, the algorithm locates the posterior high density region instead of assuming that the posterior mode is known. Then, in the second stage we use the resulting normalised posterior as weights to approximate the marginals instead of fitting a least-squares polynomial as in Brown (2019). We compared our results in all examples with standard MCMC methods.

5.4.2 Posterior approximation

The full joint posterior of $\theta = (\theta_1, \ldots, \theta_M)$, x_0 and $\sigma^2 = (\sigma_1^2, \ldots, \sigma_Q^2)$ given multivariate normal observations $Y_k \sim \mathcal{N}(\hat{x}_k, \sigma^2 I_Q)$, where $k = 1, \ldots, K$, Y_k and $\hat{x}_k$ have dimension Q, σ^2 is the noise associated with the data types and I_Q is the $Q \times Q$ identity matrix, is:

$$p(\theta, x_0, \sigma^2|Y) = \frac{\mathcal{L}(Y|\theta, x_0, \sigma^2)p(\theta_1, \theta_2, \ldots, \theta_M)p(x_0)p(\sigma^2)}{p(Y)}. \tag{5.9}$$

Here $\mathcal{L}(Y|\theta, x_0, \sigma^2)$ is the likelihood of the observations $Y = (Y_1, \ldots, Y_K)$ given θ, x_0 and σ^2, $p(\theta_1, \theta_2, \ldots, \theta_M)$ is any joint prior distribution on the unknown parameter vector θ, $p(x_0)$ is the prior distribution of the state initial values and $p(\sigma^2)$ is the prior distribution on σ^2. In this paper, to simplify the notation, we consider from now on that σ^2 and x_0 are a part of the parameter vector θ to be estimated. The denominator $p(Y)$ may be regarded as the normalising constant for the posterior distribution.

Computing the denominator of (5.9) may not be possible analytically, and is not necessary when using MCMC. However, we show how $p(Y)$ can be used to normalise the posterior and use this normalization as weights to approximate the posterior marginals for each parameter.

In this paper we view the approximation of the unnormalised joint posterior $\hat{p}(\theta|Y) = \mathcal{L}(Y|\theta)p(\theta)$ as a numerical integration problem. That is, we calculate the M dimensional integral:

$$\int_{\prod_{m=1}^{M}[a_m, b_m)} \hat{p}(\theta|Y)d\theta \approx \int_{\prod_{m=1}^{M}[a_m, b_m)} \mathcal{L}(Y|\theta)p(\theta)d\theta, \tag{5.10}$$

where $a_m = (a_1, \ldots, a_M) \in \mathbb{R}^M, b_m = (b_1, \ldots, b_M) \in \mathbb{R}^M$ are left and right limits respectively of the parameter intervals for each component of the θ vector. The integral in (5.10) is also hard to obtain explicitly. Hence, we use the following approximation for each data type Q:

$$\int_{\prod_{m=1}^{M}[a_m,b_m)} \hat{p}(\theta|Y)d\theta \approx \frac{\prod_{m=1}^{M}(b_m - a_m)}{N} \sum_{s=1}^{N} \hat{p}(\theta^s|Y) \tag{5.11}$$

where $\theta^s \in \mathcal{P}_N, s = 1, \ldots, N$, is the sth approximated value of each θ in the set of points $\mathcal{P}_N = \{(\theta_1^1, \ldots, \theta_1^N), \ldots, (\theta_M^1, \ldots, \theta_M^N)\}$. The term $\prod_{m=1}^{M}(b_m - a_m)$ is necessary since we transformed the parameter space from the unit hypercube to $[a, b)^M$. There are many approaches to determine $\mathcal{P}_N$. In the method proposed here we used a Sobol sequence to generate a low discrepancy set of points in $[a, b) \subset \mathbb{R}^M$. To compute the approximation (5.11), we evaluate the unnormalised joint posterior:

$$\hat{p}(\theta^s|Y) \propto \mathcal{L}(Y|\theta^s)p(\theta^s) \tag{5.12}$$

on all $\theta^s \in \mathcal{P}_N$ and normalise by dividing by the sum of these values.

If there is some prior information about the parameter values, one can generate a low discrepancy point set and then use, for example, a linear transformation to map the point set from the M dimensional unit hypercube $[0, 1)^M$ to a suitable hypercube $[a_m, b_m)$. However, in many cases we lack such prior information. In order to overcome this problem and identify an appropriate bounded region in the parameter space, we run the algorithm in two stages. First, we choose values of a_m and b_m to explore the parameter space conservatively and locate a region with high posterior density. For instance, in Example 5.5.1 we use the maximum posterior value resulting from stage 1 to approximate the posterior mode and then the new parameters intervals limits, $[a_m, b_m]$, are determined to be around

this mode, for example, twice the posterior mode value. However, when the posterior has multi-modal shape as in Example 5.5.2 we cannot use this approach, so instead we choose parameter values (using predetermined thresholds) from the resulting posterior marginals in Stage 1 for each parameter to be the new bounds of the interval $[a_m, b_m]$. Second, we reapply the algorithm to a smaller hypercube containing the high-density region. As the examples below demonstrate, this second stage substantially improves the accuracy of estimation at the cost of doubling the number of times the ODE system must be solved numerically.

The manner in which QMC point sets are distributed in parameter space allows this approach to explore the parameter space from a global point of view, whereas MCMC methods explore the space in a local and sequential manner. This is particularly advantageous when the posterior distribution is multi-modal, as we demonstrate in Example 5.5.2.

The proposed approach is summarised in the following procedure.

Algorithm 5.1. QMC Algorithm

Stage one

1: Identify an initial interval $\mathcal{I}_m$ for each paramter containing the high density region of the marginal prior distribution for each parameter $m = 1, \ldots, M$.
2: Generate a low discrepancy set of points $\mathcal{P}_{\mathcal{N}} \subset \mathcal{I}_m = [I_1 \times \ldots \times I_M]$.
3: **for** each $\theta^s \in \mathcal{P}_N$ **do**
4: Solve the ODE using numerical integration.
5: Using the solution of the ODE, the observations Y and the priors, evaluate the joint posterior $\hat{p}(\theta^s, |Y)$ in (5.10).
6: **end for**

Stage two

7: Construct a smaller hypercube $\mathcal{I}'_m$ containing those points obtained by thresholding (at level $d = 0.999$) cumulative summations of the posterior density evaluated in Step 5.
8: Repeat Steps 2-5 for this smaller hypercube.

5.4.3 The estimation of the marginal posterior

The resulting joint posterior vector produced from Stage 2 in Algorithm 5.1 contains the evaluation of the posterior at the proposed parameter values. Some of these values represent high density regions and some are negligible or low. In this thesis, we propose using the cumulative summations of the normalised posterior evaluations vector to identify the high density region of the posterior in the following manner. Sort the normalized posterior vector into descending order of density and then calculate the cumulative summations and store them in a vector denoted *cumsum*. Since the posterior vector is normalized, the cumulative summations vector *cumsum* has an upper bound 1, so the cumulative sum gradually increases until it comes close to this upper bound. The low posterior values have a very small effect on the increasing sum. Therefore, elements of the posterior vector corresponding to cumulative sums greater than some threshold $d \in [0, 1]$ are discarded. In all numerical examples presented in this paper, we found that using $d = 0.999$ gives satisfactory results. To visualize the marginal posterior distribution, we plot the weighted density of these points (weighted by the corresponding posterior value) as in Figures 5.3, 5.8, 5.9 and 5.10.

5.5 Comparison with existing methods

This section compares the performance of our method with that of Stan: a state-of-the-art Hamiltonian Monte Carlo (HMC) sampler. As test cases, we have used the Lotka-Volterra model and an epidemiological compartmental model.

5.5.1 Lotka-Volterra

The Lotka-Volterra model is a well-known model of ecological systems comprised of two non-linear differential equations describing the interaction between two species, a predator species W and a prey species S:

$$\begin{aligned} \frac{dS}{dt} &= S(\alpha - \beta W) \\ \frac{dW}{dt} &= -W(\gamma - \delta S) \end{aligned} \tag{5.13}$$

where $\theta = [\alpha, \beta, \gamma, \delta]$ are the parameters of interest. We first generated data by solving the system (5.13) numerically over the time interval $[0, 2]$ with step size 0.1, $\theta = [\alpha = 2, \beta = 1, \gamma = 4, \delta = 1]$ and initial values of the state variable $x(0) = [S(0) = 5, W(0) = 3]$. Then, to generate an observation vector Y, normal noise $\mathcal{N}(0, \sigma^2 = 0.2)$ was added to the solution, as shown in Figure 5.2.

Weakly informative priors were used for all the parameters. We chose a Gamma prior distribution, $\mathcal{G}(4, 0.5)$, for each of α, β, γ and δ, and an inverse Gamma, $\mathcal{IG}(1, 1)$, prior for σ^2.

For comparison to our method, the posterior distribution was sampled using Stan Stan Development Team (2019). Five chains were run with 65536 iterations each, of which the first 5000 were discarded as burn-in. To check for adequate convergence and mixing we reviewed the $\hat{R}$ statistics and the effective sample sizes for each chain Gelman et al. (2013). The $\hat{R}$ statistics were all close to 1, which is consistent with all chains having converged. The effective sample size estimates for each parameter were $(75370, 73197, 71248, 74350, 142843)$ for the parameters $[\alpha, \beta, \gamma, \delta, \sigma]$, respectively.

QMC analysis was performed using Algorithm 5.1. A total of 2^{16} RQMC points distributed over the interval $[0, 10]$ for each parameter were used in Stage 1 of

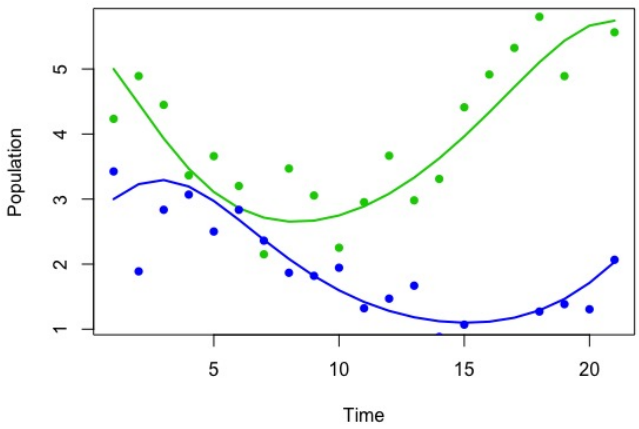

Figure 5.2: Simulated trajectories of predator (green) and prey (blue) populations, obtained as solutions of the Lotka-Volterra model (5.13) with $\alpha = 2, \beta = 1, \gamma = 4, \delta = 1, S(0) = 5$ and $W(0) = 3$. Curves show the continuous population trajectories, while scatter points represent simulated observations to be used to infer the model parameters.

the algorithm to locate the high density region of the posterior distribution. In Stage 2 we used the resulting modes as a guide to generate 2^{18} RQMC points in the high density region.

One great feature of RQMC points is that once determined the function evaluations can be performed in parallel to speed up computation. For example, in this test case we used three cores to perform parallel posterior evaluations. In contrast, MCMC methods are necessarily serial. One can run multiple MCMC chains in parallel, as we have done here, but each chain has to converge separately.

In this example, the QMC method implemented on three cores ran in 19.56 minutes, compared to 22.56 minutes per chain for Stan. Figure 5.3 shows the marginal posteriors estimated using MCMC and QMC. Both methods have similar outputs, however, the QMC method produces more accurate estimates of the parameters (γ, δ), as shown in Table 5.1. Table 5.1 also demonstrates that Algorithm 5.1

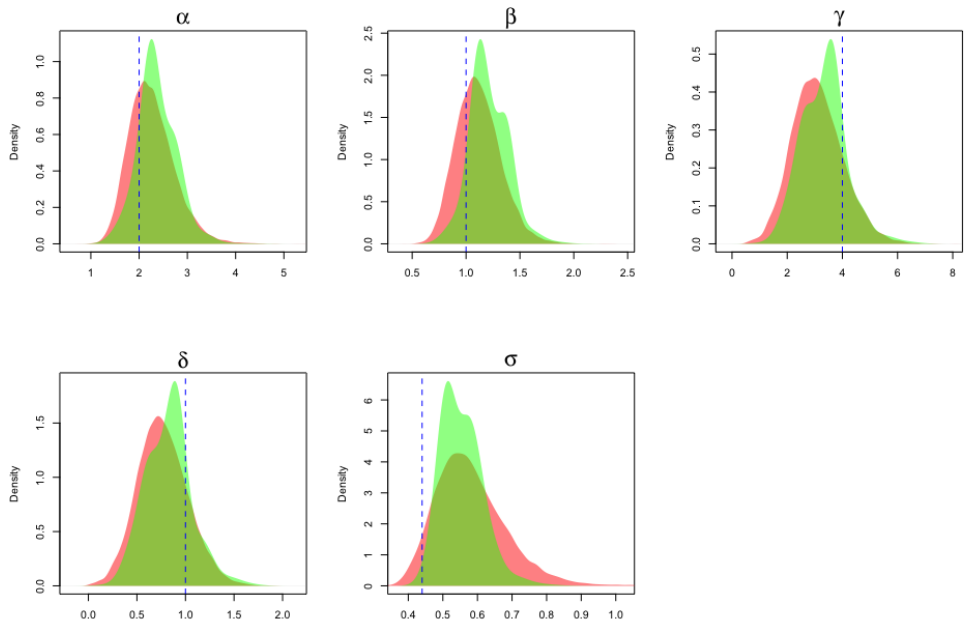

Figure 5.3: Marginal posteriors for α, β, γ, δ and σ from model in (5.13). Stan (red) and QMC (green) produced similar approximations to the posterior distribution. The dashed lines represent the true values of each parameter.

produces wider credible intervals than Stan for all parameters. We suggest that the more systematic and global approach to covering the parameter space used by QMC does a better job of exploring the tails of the distribution efficiently. MCMC methods like Stan make sequential local moves and consequently expend more computational effort exploring high density regions, resulting in relatively poor characterisation of the posterior tails. In other respects, Stan and QMC produced similar results. Figures 5.4 and 5.5 show that both methods produce similar pairwise marginal posterior distributions and similar posterior predictive results, respectively.

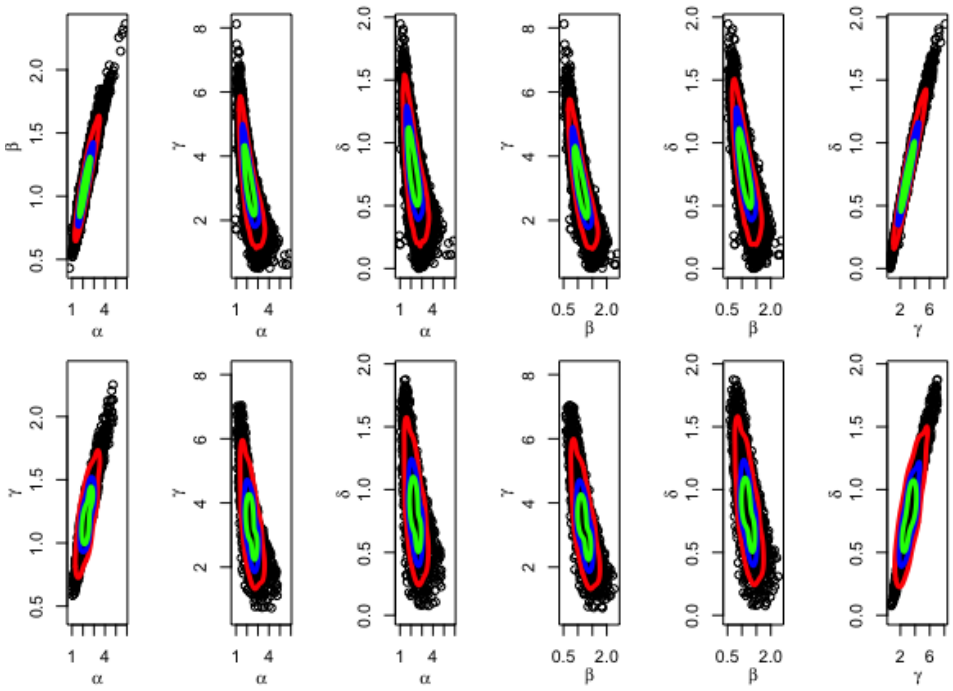

Figure 5.4: Scatterplots and contours of the bivariate posterior marginal distributions for each pair of parameters in the Lotka-Volterra model using Stan (first row) and QMC (second row). The red, blue and green contours contain respectively the proportions $0.95, 0.75$ and 0.5 of samples drawn from the joint posterior distribution. Contours were produced using the R function "HPDregionplot".

		MCMC	QMC
Mode	α	2.105	2.286
	β	1.029	1.139
	γ	3.377	3.610
	δ	0.861	0.906
	σ^2	0.542	0.514
95% CI	α	$(1.467, 3.069)$	$(1.021, 4.390)$
	β	$(0.7442, 1.444)$	$(0.591, 2.0261)$
	γ	$(2.117, 5.398)$	$(1.197, 6.579)$
	δ	$(0.502, 1.406)$	$(0.198, 1.707)$
	σ^2	$(0.423, 0.827)$	$(0.401, 0.802)$
Computing time		22.56 mins per chain	19.56 mins
Iterations		65536×5	327680
Software		R and C++	R

Table 5.1: Posterior modes and credible intervals (CIs) for all parameters in the Lotka-Volterra model, number of iterations and computation times for Stan and QMC.

5.5.2 SIR model

The SIR model is an ODE system that describes the spread of an infectious disease in a large population. The SIR model assumes the population consists of three types of individuals: the number of susceptibles S, the number of infecteds I and the number of recovered or removed individuals R, and $N = S + I + R$ is the total population. The ODE system is given by

$$\begin{aligned} \frac{dS}{dt} &= -\beta S(t) I(t) \\ \frac{dI}{dt} &= (\beta S(t) - \gamma) I(t) \\ \frac{dR}{dt} &= \gamma I(t) \end{aligned} \tag{5.14}$$

where $\beta > 0$ is the the disease transmission rate and $\gamma > 0$ is the recovery rate.

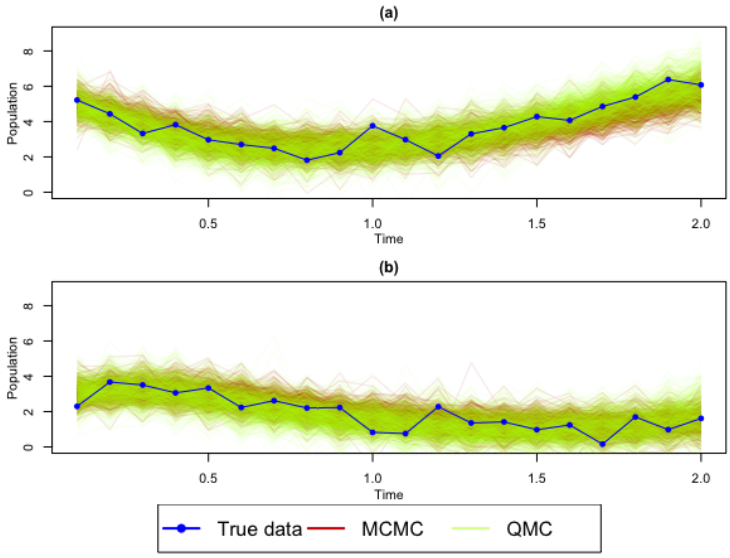

Figure 5.5: Posterior predictive simulations based on parameters estimated by Stan (red) and QMC (green) given the noisy data (blue dots). Panel (a) shows posterior predictive simulations for the prey species S and (b) shows the same for the predator species W.

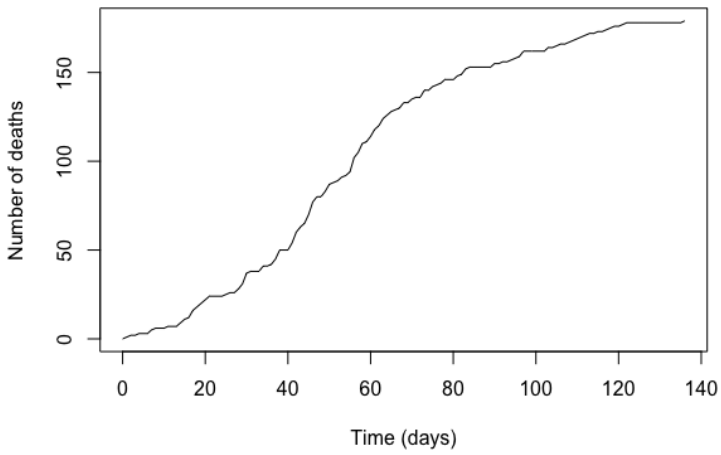

Figure 5.6: Accumulated number of deaths during the second black plague outbreak in the village of Eyam, UK during the 136 days during the period from June 19, 1666 to November 1, 1666 with only 83 surviving villagers.

We use cumulative daily deaths, y_t, recorded during the second black plague outbreak in the village of Eyam, UK in the period from June 19, 1666 to November 1, 1666 Massad, Coutinho, Burattini, and Lopez (2004). Figure 5.6 shows the accumulated number of deaths at times $t \in \{t_1, ...t_n\}$ corresponding to $n = 136$ consecutive days. Since the village decided to seal themselves off from other surrounding villages to avoid spreading the disease, the total population is fixed at $N = 261$ with only 83 surviving villagers. The number of removed individuals, $R(t)$, is equivalent to the number of deaths up to time t since there was no recovery from the disease under the model Campbell and Lele (2014); Golchi and Campbell (2016). The initial number of removed individuals is $R(0) = 0$ and thus $S(0) = N - I(0)$, but the initial number of infected individuals $I(0)$ is unknown. We therefore consider $I(0)$ a parameter to be estimated, so that $\theta = (\beta, \gamma, I(0))$.

The only observed data is $Y = (y_1, \ldots, y_n)$, the cumulative deaths on each day. However, the number of infected individuals at the end of the epidemic is 0, and as there was only one death on the final day of the epidemic as appeared in the Eyam data, the number of infected individuals at time $n - 1$ is 1 Campbell and Lele (2014); Jonoska Stojkova (2017). Hence, following the approach of Jonoska Stojkova (2017), there are two additional data points $x_{n-1} = 1$ and $x_n = 0$ representing the number of infected individuals at times $n - 1$ and n respectively.

Following Jonoska Stojkova (2017), we modelled each y_t as having a Binomial distribution with mean $R_{(\beta,\gamma,I(0))}(t)$ obtained by solving the system in (5.14). Likewise, we modelled x_{n-1} and x_n as each having a Binomial distribution with mean $I_{(\beta,\gamma,I(0))}(t))$ obtained by solving the system in (5.14). Thus, the likelihood

takes the form:

$$p(Y|\beta,\gamma,I(0)) = \prod_{i=1}^{n} \text{Binomial}\left(y_i|N, \frac{R_{(\beta,\gamma,I(0))}(t_i)}{N}\right) \times \prod_{i=n-1}^{n} \text{Binomial}\left(x_i|N, \frac{I_{(\beta,\gamma,I(0))}(t_i)}{N}\right). \tag{5.15}$$

This model has a significant defect in that the likelihood assigns positive probability to the set of trajectories Y in which $y_t > y_{t-1}$ for some t. A better model without this defect could certainly be devised, but this published model is adequate for our present purposes.

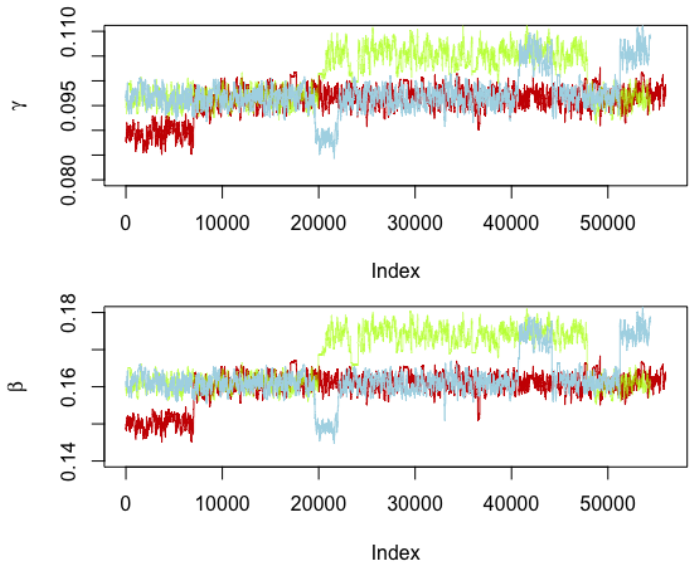

Figure 5.7: Trajectories of MCMC samples for γ and β in the SIR model for three different MCMC chains.

Again following Jonoska Stojkova (2017), we set vague priors for β and γ to be $\mathcal{G}(1,1)$ and the prior distribution for $I(0)$ to be Binomial$(N, 5/N)$. We also experimented with a Binomial$(N, 7/N)$ prior for $I(0)$, noting that Raggett (1982) estimated $I(0)$ to be 7. These two prior distributions for $I(0)$ produced

indistinguishable output; the results presented below are for an expected $I(0)$ value of 5).

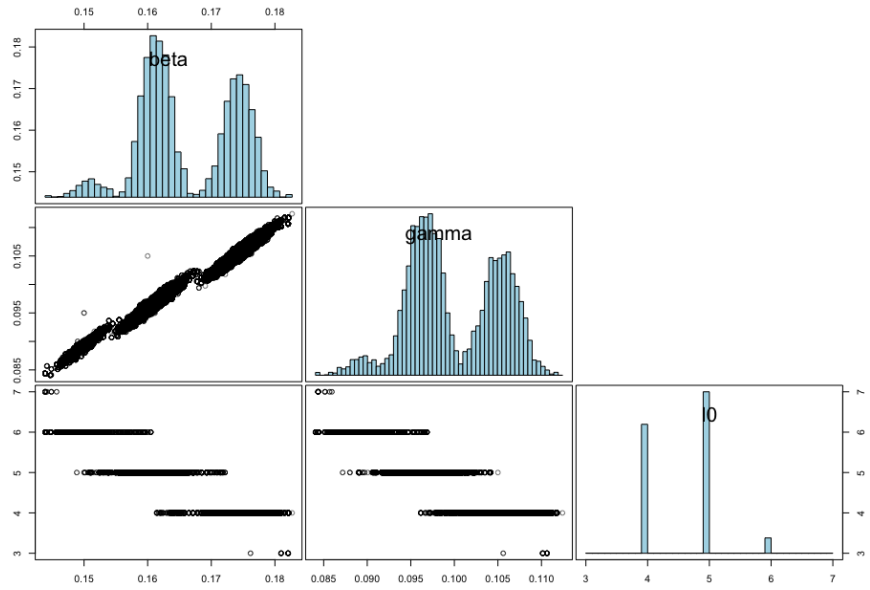

Figure 5.8: Marginal (diagonal) and joint (off-diagonal) posterior distributions of γ, β and $I(0)$ in the SIR model using MCMC.

For this example we cannot compare to Stan because the parameter $I(0)$ is discrete and Stan does not currently support discrete parameters. Instead we coded a simple Metropolis-Hastings MCMC algorithm in R to sample $\theta = [\beta, \gamma, I(0)]$.

Multimodality of the posterior distribution is often a cause of poor mixing when using MCMC, but for this model we were able to obtain useful results in feasible run times (although there were some modes not sampled as discussed below). Three chains were run from different starting points for both β and γ equal to $(0.05, 0.1, 0.15)$ and for $I(0) = (3, 5, 7)$. For the discrete parameter $I(0)$ we set the proposal to be a normal distribution with standard deviation equal to one. For β and γ we chose normal proposals that randomly choose the standard deviation from $[0.001, 0.01, 0.1]$ in each iteration. This variability of scale in the proposal

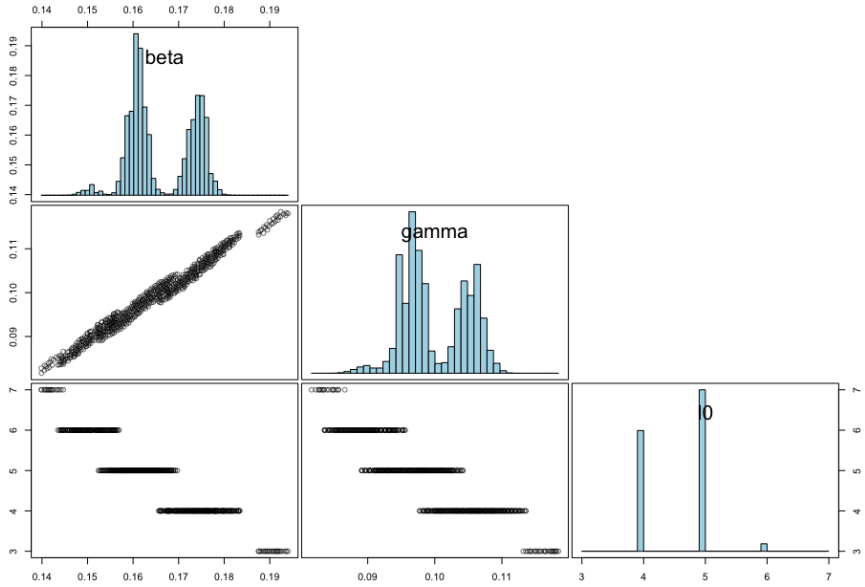

Figure 5.9: Marginal (diagonal) and joint (off-diagonal) posterior distributions of γ, β and $I(0)$ in the SIR model using QMC.

distribution may help the MCMC sampler to traverse between modes. Figure 5.7 shows the trajectories of β and γ for the three chains, illustrating how different chains have explored the posterior results. One chain has found three modes, but two other chains failed to discover a third mode. Taking all the samples from all the chains we can see three distinct modes of the posterior distribution, as shown in Figure 5.8. This indicates all three chains have not converged, and need to be run for a longer time, ideally until all modes have been sampled by all chains. However, we chose to stop the chain when it reached the same number of posterior evaluations as used in QMC for the sake of fair comparison.

The QMC method in Algorithm 5.1 was also used to estimate the parameters $\theta = [\beta, \gamma, I(0)]$. A total of 2^{16} RQMC points distributed on the hypercube $[0,1]^3$ were used in Stage 1 of the Algorithm. For the continuous parameters $[\beta, \gamma]$, we did not apply a transformation. For the discrete parameter $I(0)$ we did not use RQMC points - instead we assigned equal numbers of points to the values

1 to 7. Using the maximum and minimum values identified in Stage 1 for each parameter, a new set of RQMC points were generated in the high density region of the posterior distribution, to use in Stage 2 of Algorithm 5.1.

The QMC method produced sensible marginal distributions for each parameter and was more efficient than MCMC in terms of the computational time and exploration of the modes of the posterior distribution. For fair comparison, we run the same number of posterior evaluations for both methods. Whereas the QMC approach required 7.63 minutes, involving 2^{16} posterior density evaluations in Stage 1 and 2^{17} evaluations in Stage 2, the MCMC consumed nearly 10 minutes per chain for an average of 65536 iterations.

		MCMC	QMC
Modes	γ	–	0.083
		0.089	0.089
		0.096	0.096
		0.105	0.105
		–	0.117
	β	–	0.141
		0.151	0.150
		0.161	0.161
		0.175	0.175
		–	0.191

Table 5.2: Posterior modes for all parameters in the SIR model for MCMC (3 modes identified) and QMC (5 modes identified).

Figure 5.8 presents the results obtained using the MCMC sampler, in particular showing that only three initial state values $I(0) = \{4, 5, 6\}$ were sampled by any of the chains. Figure 5.9 presents the results obtained using RQMC, showing that 5 different initial state values $I(0) = \{3, 4, 5, 6, 7\}$ were identified as having non-neglible marginal posterior probabilities as shown in Figure 5.10. The two plots in Figure 5.10 show the corresponding marginal distributions of the two continuous parameters γ and β, which have multiple modes corresponding to

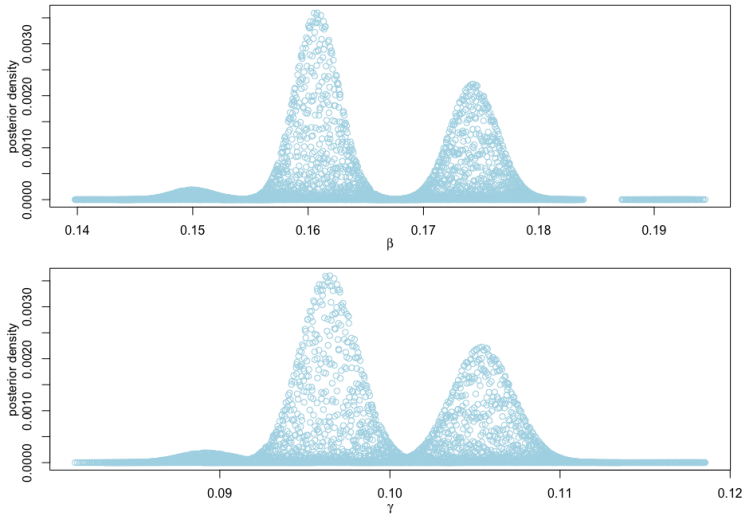

Figure 5.10: Samples from the marginal posterior distribution of β (top) and γ (bottom) from the SIR model example, showing five modes identified using QMC.

different numbers of initially infected individuals.

The MCMC algorithm is clearly sub-optimal, and needs to be run for longer to sample the low probability modes, since the three chains sampled only the three modes with higher density and neglected the two with low density as shown in Table 5.2 and Figure 5.10. Nevertheless, both methods appear to be adequate for prediction purposes. Figure 5.11 shows posterior predictive results for 1000 simulations using posterior samples from MCMC (red) and QMC (green). Both methods produced prediction plots that covered the data points, in spite of the fact that the MCMC sampler failed to sample all posterior modes.

5.6 Discussion

In this paper, we proposed a method using point sets constructed using LDS to estimate marginal distributions of parameters in low-dimensional systems of ODEs. Moreover, we proposed a new method to visualise the marginal posterior using

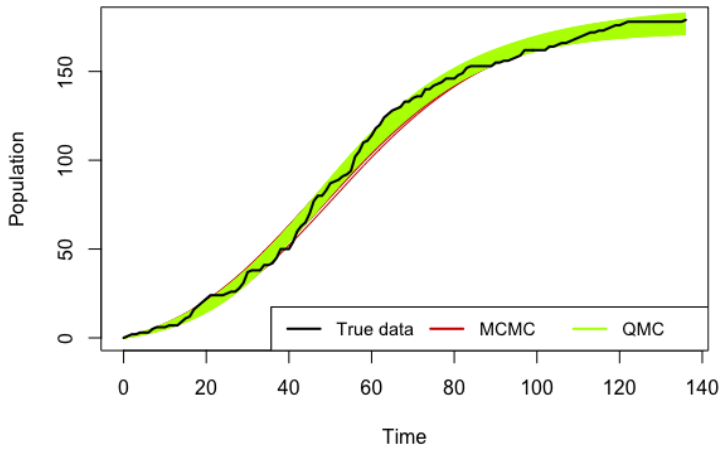

Figure 5.11: Posterior predictive distribution (90% credible interval) estimated using MCMC (red) and QMC (green) using the noisy data (black line).

cumulative summation. There are five important advantages of our approach. First, QMC points provide a direct global picture of the posterior distribution, whereas MCMC relies on local exploration to sequentially explore the parameter space. This advantage is particularly important when the posterior density has multiple modes. Second, given that QMC points are predetermined, posterior evaluations can be done independently in parallel. Third, Algorithm 5.1 does not suffer from the slow mixing problems to which MCMC methods are prone. Fourth, QMC methods are generic, easy to understand and implement, whereas efficient proposal distributions for MCMC algorithms can be problem-specific and difficult to design. Fifth, the cumulative summation procedure assesses the contribution of each positive posterior value resulting from a pair of proposed parameters and provides a quick and easy method to determine whether this contribution is negligible (when proposed parameters fall in a low posterior probability region) or significant (when these proposed parameters are located in a high density region). This provides a useful way to visualise the marginal posteriors.

In the illustrative examples presented here, the QMC method outperformed MCMC in both computational time and accuracy of estimation. In particular, the second example illustrates the advantages of the QMC method when the posterior distribution has multiple modes.

Here we have applied QMC to parameter spaces of up to five dimensions. In higher dimensions, QMC becomes infeasible, and other methods may become preferable. Practical guidelines for determining whether QMC is preferable for a specific problem are not currently available. Such guidelines would be of use to practitioners and are an important topic for future work.

Chapter 6

Summary and main contribution

The first section of this Chapter summarises how I achieved the research objectives and contributed to fulfilling the overall aim of addressing some of the current Bayesian approach limitations when used within ODEs and developed statistical methods that overcome these limitations. Brief descriptions of the four research objectives are provided. The second part of this Chapter summarises the limitations of this work and possible future directions.

6.1 Summary

The overall aim of this thesis is to investigate the crucial limitations of the current Bayesian approach when applied to ODE models and then to develop new methods to conquer such problems. For convenience, I have summarised the four research objectives as follows:

1. Demonstrate the problem of misspecification in ABC methods for ODE parameter estimation.
2. Develop a solution for SMC ABC to overcome the current misspecification

issue.

3. Propose a new approach that explores the parameter space more efficiently, significantly reduces computational cost, and outperforms MCMC in the case of a multi-modal posterior.

4. Evaluate the proposed methods for high impact, real world, research problems.

Objective 1 was achieved in Chapter 3 of this thesis. Deep investigations were conducted on various ABC algorithms that could be applied to estimate parameters in ODEs. These investigations highlighted the danger of neglecting the simulation of random errors, which are commonly associated with observations, and demonstrated that doing so can produce serious errors in the results. Applications to simple and complex ODE models illustrated that approximation to the Bayesian posterior became severely over-concentrated when ABC methods were applied without consideration of the noise in the observations.

A modification to the current SMC ABC method was developed in Chapter 4, satisfying Objective 2. A summary statistic was proposed to aid the construction of acceptance criteria that correctly approximate the error posterior. Moreover, including the error term in the estimation affected the process of choosing the tolerance sequence, which consequently improves the efficiency of estimation in terms of producing true parameter posteriors and improves the quality of the posterior predictive intervals.

A new method was proposed in Chapter 5 that allows the efficient exploration of the parameter space, conducted using evenly distributed point sets instead of random points as in MCMC, which satisfies Objective 3. This method enables one to sample from the parameter space from a global point of view, which improves

the parameter estimation, particularly when the posterior density has multiple modes. Given the fact that these point sets are predetermined, this allows the algorithm to work independently in parallel, which reduces the computational cost significantly. Furthermore, this method does not suffer from poor mixing, which is a common problem for MCMC. In addition, we proposed a simple method that approximates the marginal posterior using the cumulative summation of the normalized posterior.

Various ODEs were used throughout the thesis, ranging from a simple SIR model to more complex models such as the malaria model developed by White et al. (2009), which appeared in Chapter 3 with real data and in Chapter 4 with simulated data. The method proposed in Chapter 4 improved remarkably the estimation efficiency compared with using the standard SMC ABC and produced results that are similar to the results derived from MCMC.

An SIR model was used in Chapters 3, 4 and 5; however, in Chapter 5 true data were used and a multi-modal posterior was approximated. Results from applying the method proposed in Chapter 4 to the SIR model demonstrates how the approximation of the posterior improved in terms of the accurate estimation of posterior credible intervals. Moreover, the results obtained by applying the method proposed in Chapter 5 to the SIR model outperformed the standard MCMC. Estimation of the parameters conducted in this thesis was in low-dimensional parameter spaces: three parameters for the malaria and SIR models and five parameters for the Lotka-Volterra model. Applying the proposed methods to different ODE models improves the accessibility of such methods and illustrated the capabilities of the new approaches, which fulfils Objective 4.

6.2 Future work

In this section, I will outline some of the future research directions for investigating the usage of an adaptive QMC method to approximate the posterior distribution using a multivariate normal distribution. Then, I will outline the current progress in the development of an R package implementing the method proposed in Chapter 5, for which the candidate will be the corresponding author. Moreover, I will illustrate some of the possible future directions for extending the Modified SMC ABC.

6.2.1 Adaptive QMC method

Using the QMC method improves the computational efficiency of parameter inference for ODEs. The good scaling properties of the QMC approach create the possibility of using QMC point sets with a higher parameter dimension. These properties, such as the promising results that we have seen in Chapter 5 when the posterior distribution has multiple modes and the ability to run the estimator in parallel, would in the case of ODEs reduce the computational burden.

We have seen in Chapter 5 that QMC methods can be used to approximate the integral of a function $f : [0,1)^M \to \mathbb{R}$, defined on the M-dimensional unit cube as in Equation (5.6). QMC methods can be used to compute expected values of functions on $\mathbb{R}^M$. For example, suppose $X = (X_1, \ldots, X_M)$ is a random vector comprised of independent random variables, each with a standard normal distribution, and consider $h : \mathbb{R}^M \to \mathbb{R}$. Then

$$E[h(X)] = \int_{\mathbb{R}^M} h(X)\phi(X)dX \tag{6.1}$$

where ϕ is the standard normal density. Let $X = \Phi^{-1}(t)$, where $t = t_1, \ldots, t_M$ and Φ is the standard normal cumulative distribution function. Then, using the transformation

$$dX = \frac{1}{\Phi'(\Phi^{-1}(t))}dt$$
$$\phi(X) = \phi(\Phi^{-1}(t)),$$

the integral in Equation (6.1) has the form

$$\begin{aligned} E[h(X)] &= \int_{[0,1)^M} h(\Phi^{-1}(t_1), \ldots, \Phi^{-1}(t_M))\phi(\Phi^{-1}(t))\frac{1}{\Phi'(\Phi^{-1}(t))}dt \\ &= \int_{[0,1)^M} h(\Phi^{-1}(t_1), \ldots, \Phi^{-1}(t_M)))dt \approx \frac{1}{n}\sum_{i=1}^{n} f(t_i) \end{aligned} \tag{6.2}$$

and

$$f(t_1, \ldots, t_M) = h(\Phi^{-1}(t_1), \ldots, \Phi^{-1}(t_M)).$$

Here we have used the change of variables $t_i = \Phi(x_i)$ for $i = 1, \ldots, M$.

Now suppose that Y is a random vector on $\mathbb{R}^M$ with a multivariate normal distribution having mean vector μ and positive definite covariance matrix $\Sigma = LL^T$ where L is a lower triangular matrix (that is, a Cholesky factor). Then $X = L^{-1}(Y - \mu)$ has independent standard normal components. Consider $g : \mathbb{R}^d \to \mathbb{R}$, then $E[g(Y)] = E[g(\mu + LX)]$ can be approximated as above with $h(X) = g(\mu + LX)$.

Next consider a random vector Z having a density that is a weighted average of multivariate normal densities. That is, the density of Z is $\sum_{j=1}^{d} \pi_j N(z; \mu_j, \Sigma_j)$, where $\pi_1, \ldots, \pi_d \in [0, 1]$, $\sum_{j=1}^{d} \pi_j = 1$ and $N(z; \mu_j, \Sigma_j)$ is the multivariate normal

density with mean μ_j and positive definite covariance matrix Σ_j for $j = 1, \ldots, d$. Given $n \in \mathbb{N}$, define $n_j = [\pi_j n]$ (that is, the nearest integer to $\pi_j n$) and construct Sobol sequences $t_{j1}, \ldots, t_{jn_j}$ for $j = 1, \ldots, d$. Then

$$E[g(Z)] = \sum_{j=1}^{d} \pi_j E[g(Y_j)]$$

where $Y_j \sim N(\mu_j, \Sigma_j)$. Thus

$$E[g(Z)] \approx \sum_{j=1}^{d} \pi_j \frac{1}{n_j} \sum_{i=1}^{n_j} f_j(t_{ji}) \approx \frac{1}{n} \sum_{j=1}^{d} \sum_{i=1}^{n_j} f_j(t_{ji}) \tag{6.3}$$

where $f_j(t_1, \ldots, t_M) = h_j(\Phi^{-1}(t_1), \ldots, \Phi^{-1}(t_M))$ and $h_j(X) = g(\mu_j + L_j X)$ with $L_j L_j^T$ being the Cholesky factorisation of Σ_j.

Then, consider a random vector W with a continuous density p over $\mathbb{R}^M$. We aim to approximate the density of W using a weighted average of d multivariate normal densities. Specifically, we want to find $q(w) = \sum_{j=1}^{d} \pi_j N(w; \mu_j, \Sigma_j)$ to minimise the Kullback-Liebler (KL) divergence:

$$D(q, p) = -\int_{\mathbb{R}^M} q(w) \log\left(\frac{p(w)}{q(w)}\right) dw = -E\left[\log\left(\frac{p(W)}{q(W)}\right)\right].$$

Then we propose to approximate $D(p, q)$ as above, so that

$$D(q, p) \approx \frac{1}{n} \sum_{j=1}^{d} \sum_{i=1}^{n_j} f_j(t_{ji}) \tag{6.4}$$

where the values of n_j and t_{ji} and the functions f_j are determined as above for given values of (π_j, μ_j, Σ_j) for $j = 1, \ldots, d$ with $g(w) = -\log(p(w)/q(w))$. The problem is, to determine n_j, t_{ji} and f_j we need to know (π_j, μ_j, Σ_j) for $j = 1, \ldots, d$, but we want to find values of these parameters that will minimise $D(q, p)$. We therefore suggest an iterative algorithm:

Algorithm 6.1. Adaptive QMC

1: Initialize N=0, π_j, μ_j and Σ_j, for $j = 1, ..., m$.
2: Compute lower triangular Cholesky factor L for each Σ_j.
3: Repeat N times.
4: Determine n_j, t_{ji} and f_j for $j = 1, \ldots, m$ to use in 6.3.
5: Given n_j, t_{ji} and f_j, find (π_j, μ_j, Σ_j) for $j = 1, \ldots, m$ to minimise the approximation to Kullback-Liebler divergence given by 6.4.

Then, once q has been determined, we can use importance sampling to evaluate expectations of the form

$$E[k(W)] = \int_{\mathbb{R}^M} k(w)\frac{p(w)}{q(w)}q(w)dw \approx \frac{1}{n}\sum_{j=1}^{d}\sum_{i=1}^{n_j} f_j(t_{ji})$$

where we have redefined f_j using $g(w) = k(w)(p(w)/q(w))$ for an arbitrary function k on $\mathbb{R}^M$.

In future work, this algorithm should be investigated using empirical studies and explorations of its theoretical properties. For example, a key theoretical question is under what conditions Algorithm 6.1 converges. Another important theoretical question is whether the approach finds the minimizer of KL divergence as $n \to \infty$.

6.2.2 qmcposterior: An R package for the estimation of parameters in ODEs:

'qmcposterior' is an R package providing a toolkit to estimate parameter values in ODEs, using the QMC method I have developed during the course of my candidature. It is 50% complete, and it will be released to CRAN when it is finished. This package has functions that evaluate the posterior distribution value over QMC point sets. Then, by using cumulative summation of the joint posterior

vector, marginal posteriors can be visualized. Given that QMC point sets are predetermined, all computations can be performed independently over multiple cores, which reduces the computational time significantly.

6.2.3 Modified SMC ABC

One of the main objectives of this thesis is to improve the SMC ABC method. The Modified SMC ABC method introduced in Chapter 4 has been shown to produce more efficient results than the standard SMC ABC, but still there are some limitations that need to be addressed.

To increase the efficiency of the estimation when using SMC ABC, a large number of samples have to be drawn, which makes this method computationally expensive due to the need to solve the ODEs with each parameter sample for a number of populations. Although the modification to SMC ABC proposed in this thesis improves the approximated posterior distribution for both the parameters within ODEs and the noise parameter, the method is still time-consuming compared with MCMC. Ghosh et al. (2017) proposed a modification to SMC ABC that sped up the estimation process; however, there are some challenges that need to be addressed such as the choice of the prior distribution. Choosing narrow priors can lead to population degeneracy and choosing wider priors can cause a low acceptance rate.

In Chapter 4, experimental evidence has illustrated how Modified SMC ABC could give more accurate results than the standard SMC ABC. It would be of interest to investigate the theoretical properties of the proposed method, for example, to study the form that the new ABC posterior approximation would have after using the proposed modification. Moreover, conducting a mathematical argument that generalized this finding would make the new acceptance criterion for updating the noise parameter stronger. Conducting such an investigation is left for further

research.

6.3 Concluding Remarks

This thesis has addressed some of the limitations of current methods that are used to estimate parameters of ODEs and developed novel and efficient methods in this context. These methods facilitate inference by tackling some of the challenges that appear when conducting estimation for ODEs, such as computational time, multimodality and quantification of the uncertainty in parameter values.

Appendix A

Appendix Chapter 3

A.1 Software Validation

This section summarizes the method that has been used to validate the software in Chapter 3. To validate our software for the MCMC approach we used the method of Cook, Gelman, and Rubin (2006). This method is based on the idea that posterior quantiles corresponding to true parameter values should be uniformly distributed in the interval $[0, 1]$. For example, the 95th percentile of the marginal posterior distribution for a given parameter should contain the true parameter value with 0.95 probability, and similarly for any other quantile. In this paper, we implemented Cook's simulation-based validation method as follows: First, we perform a simulation for $N_{rep} = 100$ replications to validate the MCMC software. For each replication, we generated a sample of $L = 5000$ draws from the posterior distribution of the parameters $\boldsymbol{\theta}$. Then we calculated the following statistic for each posterior quantile q_i for each $\boldsymbol{\theta_i}$:

$$X_\theta^2 = \sum_{i=1}^{N_{rep}} (\Phi^{-1}(q_i))^2, \tag{A.1}$$

where $\Phi^{-1}(q_i)$ is the inverse of the standard normal Cumulative Distribution Function (CDF) evaluated at q_i. For correct software, the statistic in Equation (A.1) should follow a χ^2 distribution with N_{rep} degrees of freedom. To determine whether the posterior quantiles deviate from uniformity, we calculated the associated p_θ values for each X^2_θ and then we transformed the p_θ values into z_θ statistics, where $z_\theta = \Phi^{-1}(p_\theta)$, and plotted the absolute values of the z_θ statistics. Any extreme value of z_θ (i.e. $|z_\theta| > 2$) suggests an error in the software (Cook et al., 2006). Figure A.1 shows the absolute values of z_θ statistics from this simulation for each parameter in the first test problem. All the z_θ statistics are less than 2 which provides evidence that the software has been written correctly.

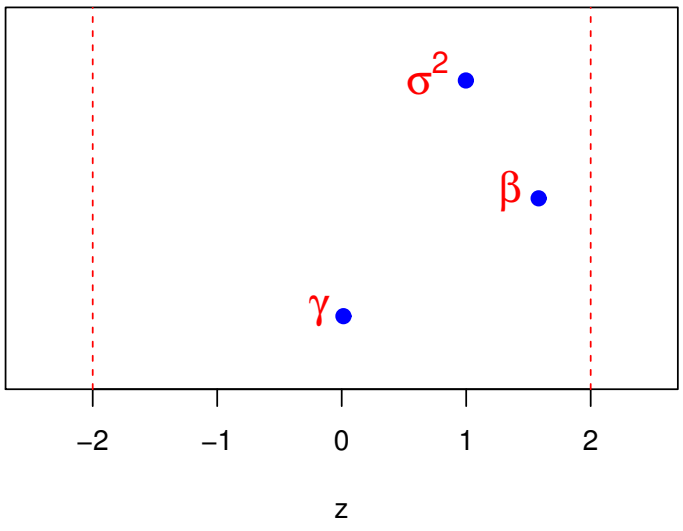

Figure A.1: The absolute values of the z_θ statistics for the validation of the MCMC algorithm implementation for the first test problem.

A.2 Non-Linear ODE Model of Malaria Transmission

This section shows the results of applying MCMC, MCMC ABC and SMC ABC on the malaria model that appeared in Chapter 3 with simulated data.

A.2.1 Simulation Results

A data set of 360 simulated data points was generated by solving Equation (4.11) in the interval $[0, 30\,(\text{years})]$ using a 4th order Runge-Kutta method. The values of the model parameters used are shown in Table A.1 and the initial conditions are given by the equilibrium solution of the system in Equation (4.11). Observations,

parameter		*value*	*source*
P	[People]	3.2×10^6	Assumed
L	[Years]	66.67	(Maude et al., 2009)
d_{imm}	[Years]	0.93	(Aguas et al., 2008)
d_{in}	[Years]	0.2	Assumed
d_{treat0}	[Weeks]	2	(Maude et al., 2009)
p_1		0.87	(Aguas et al., 2008)
p_2		0.08	(Aguas et al., 2008)
η_0		0.05	Assumed

Table A.1: The parameter values used in simulation of the White et al. (2009) model.

$\mathbf{y}$, were constructed by adding normal noise $\mathcal{N}(0, 1 \times 10^9)$ to the simulated data points. As with the first test problem, for the ABC approaches we used the discrepancy function in Equation (4.9) to compare the clinical infections given in the data set $\mathbf{y}$ with a simulated solution $\mathbf{x}$. The priors for η_0, d_{in} and σ^2 were

taken as follows:

$$\begin{aligned} p(\eta_0) &= \mathcal{B}(1,1) \in [0,1], \\ p(d_{in}) &= \mathcal{GA}(1,1) \in [0,\infty), \\ p(\sigma^2) &= \mathcal{IG}(1,1) \in [0,\infty). \end{aligned} \tag{A.2}$$

A logistic transformation was used to transform $\eta_0 \in [0,1]$ while a log transform was applied to $d_{in} \in [0,\infty)$ and $\sigma^2 \in [0,\infty)$ so that each transformed parameter had support over the real line.

	Iterations	Time
MCMC	6183	26.91 mins
MCMC ABC	4264	18.87 mins
SMC ABC	108020	316.2 mins

Table A.2: The number of iterations and computational time (mins) for parameter inference in the malaria simulation model.

We first applied the MCMC method to estimate $\boldsymbol{\theta} = (\eta_0, d_{in}, \sigma^2)$. We ran the sampler for 6183 iterations to get 1000 accepted values, which took approximately 26 minutes. Then we applied MCMC ABC as described in Algorithm 2.4. In order to get 1000 accepted values we only needed to generate 4264 simulations. The result was similar to that for the MCMC approach, with slightly less computational time.

Parameter	True value	MCMC	MCMC ABC	SMC ABC
η_0	0.05	0.0504	0.0503	0.0506
d_{in}	0.2055	0.2040	0.2041	0.2032

Table A.3: True values of the parameters η_0 and d_{in} with their estimated values (median) from MCMC, MCMC ABC and SMC ABC.

In applying SMC ABC we took perturbation kernels for both parameters to be uniform ($K_t = U(-0.005, 0.005)$) and to achieve good convergence for all the pa-

rameters we ran the algorithm for $T = 10$ populations, where each one has 1000 particles and the Algorithm adapted the values of the tolerance as ϵ =(2189977.9, 1553143.8, 1096850.5, 686344.1, 417953.8, 285161.4, 236941.6, 211657.7, 202114.8, 199320.0). When applying SMC ABC, although the parameter medians are comparable to those achieved using the other methods, the number of data generation steps required was considerably large and as a result the computational time was also large (Table A.2).

It can be seen in Table A.3 that both parameters achieved good estimation (The median of the sampled values) to their true values with all three approaches. Figure A.2 shows that the joint posterior estimated using MCMC ABC is the least dispersed. Figure A.3 shows that the posterior predictive intervals for the ABC methods fail to cover the observed data, while the MCMC method produces posterior predictive intervals that cover the majority of the data. This finding is consistent with the first example.

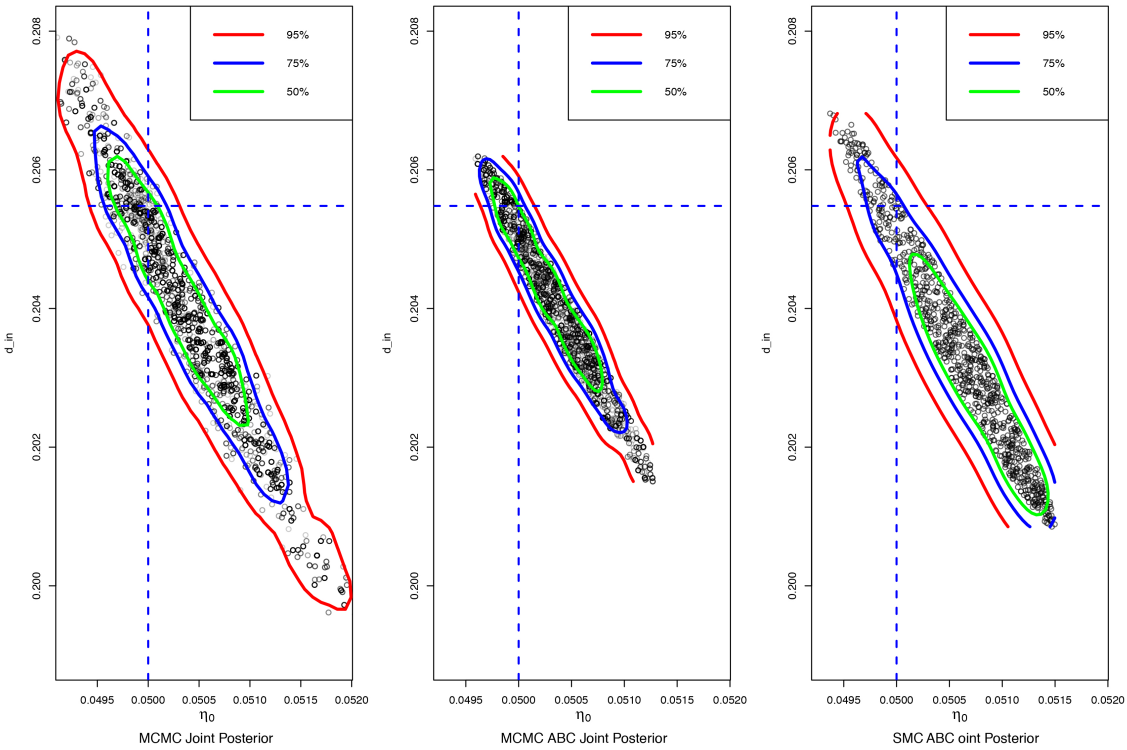

Figure A.2: Scatter plot of posterior distribution sample draws for η_0 and d_{in} from MCMC (left), ABC MCMC (middle) and SMC ABC (right). The contour lines contain the stated proportion of sample draws and they are produced using the R function "HPDregionplot".

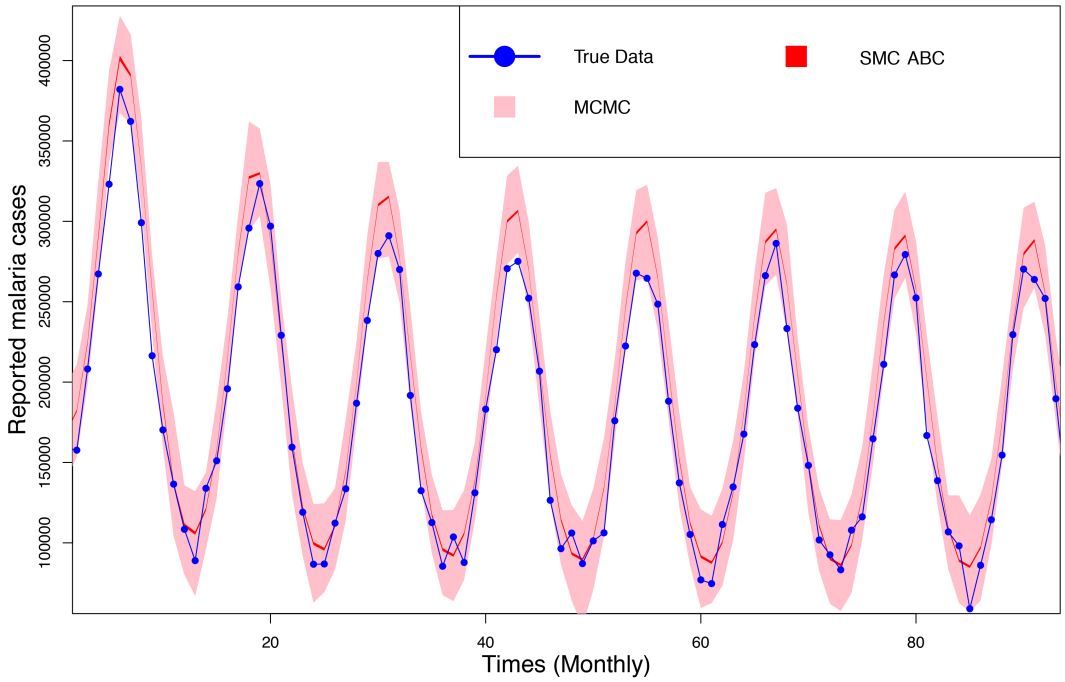

Figure A.3: Plot of the posterior predictive credible intervals from MCMC and SMC ABC fitted with the monthly malaria cases. The blue dots represent the noisy data. The shaded areas are created by the posterior predictive samples. The result derived from MCMC covered most of the data points while the ABC methods were unable to.

Appendix B

Influencing public health policy with data-informed mathematical models of infectious diseases: Recent developments and new challenges

Preamble

This appendix contains a joint paper that discusses the recent progress in some of the challenges that appear when dealing with parameter inference within dynamic disease models. Addressing these challenges will help in enhancing models' usefulness in prediction and policy.

Influencing public health policy with data-informed mathematical models of infectious diseases: Recent developments and new challenges

Amani Alahmadi[1], Sarah Belet[1], Andrew Black[2], Deborah Cromer[3], Jennifer A. Flegg[4,‡], Thomas House[5,6,‡], Pavithra Jayasundara[7], Jonathan M. Keith[1], James M. McCaw[4,8,‡], Rob Moss[8], Joshua V. Ross[2,‡], Freya M. Shearer[8], Sai Thein Than Tun[9], James Walker[2], Lisa White[9], Jason M. Whyte[10,11], Ada W. C. Yan[12], Alexander E. Zarebski[13]

‡Corresponding Authors: jennifer.flegg@unimelb.edu.au; thomas.house@manchester.ac.uk; jamesm@unimelb.edu.au; joshua.ross@adelaide.edu.au

Abstract

Modern data and computational resources, coupled with algorithmic and theoretical advances to exploit these, allow disease dynamic models to be parameterised with increasing detail and accuracy. While this enhances models' usefulness in prediction and policy, major challenges remain. In particular, lack of identifiability of a model's parameters may limit the usefulness of the model. While lack of parameter identifiability may be resolved through incorporation into an inference procedure of prior knowledge, formulating such knowledge is often difficult. Furthermore, there are practical challenges associated with acquiring data of sufficient quantity and quality. Here, we discuss recent progress on these issues.

[1] School of Mathematics and Statistics, Faculty of Science, Monash University, Melbourne, Australia.
[2] School of Mathematical Sciences, University of Adelaide, Adelaide, Australia.
[3] School of Mathematics and Statistics, UNSW Sydney, Sydney, Australia and Kirby Institute for Infection and Immunity, UNSW Sydney, Sydney, Australia.
[4] School of Mathematics and Statistics, University of Melbourne, Melbourne, Australia.
[5] School of Mathematics, University of Manchester, Manchester, UK.
[6] IBM Research, Hartree Centre, Sci-Tech Daresbury, Warrington, UK.
[7] School of Public Health and Community Medicine, UNSW Sydney, Sydney, Australia.
[8] Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, Faculty of Medicine, Dentistry and Health Sciences, University of Melbourne, Melbourne, Australia.
[9] Big Data Institute, Nuffield Department of Medicine, University of Oxford, UK.
[10] Centre of Excellence for Biosecurity Risk Analysis, School of BioSciences, University of Melbourne, Melbourne, Australia.
[11] Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS), School of Mathematics and Statistics, University of Melbourne, Melbourne, Australia.
[12] MRC Centre for Global Infectious Disease Analysis, Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, London, UK.
[13] Department of Zoology, The University of Oxford, Oxford, UK.

1. Introduction

Despite progress on many fronts, infectious diseases remain a key threat to human health worldwide [HEA15]. From 1-12 July 2019, we participated in the workshop "Influencing public health policy with data-informed mathematical models of infectious diseases" at the MATRIX institute in Victoria, Australia [MAT19]. Much of the discussion and scientific work at this event concerned the challenges identified five years ago following the Infectious Disease Dynamics 2013 programme at the Isaac Newton Institute, particularly those related to the integration of multiple datasets [DeA15]. In this paper, we return to several of the challenges identified by [DeA15] and consider both recent progress and perspectives for this rapidly developing field.

One key challenge relates to the structure of the underlying assumed mechanistic model and the observed data; in particular, whether the model parameters can be estimated given the model and the observations, and whether we can obtain analytical insights into "parameter identifiability", a property of a model that must be satisfied for precise parameter inference to be possible. If parameter identifiability is an issue, then can/should the model be reparametrized, and how should this be done? A natural question that arises here is whether we can measure something else in the process (included in the model yet or not) that can help resolve the issue, or build in existing (that is, prior) knowledge in a structured manner. For identifiable models, particularly if these are very complicated, there is a question of what we can reasonably do with current inference methods, for example Markov chain Monte Carlo (MCMC) or Maximum Likelihood estimation.

In this paper, we focus mainly on a Bayesian approach to parameter inference, which has been commonly adopted in the infectious disease modelling field since the work of O'Neill and Roberts [ONE99]. This approach has proliferated for various reasons. One is the difficulty in interpreting an epidemic in terms of frequentist statistical theory, as a small sample from a larger population, making the Bayesian approach to parameter estimation more philosophically natural [McK03]. Another advantage of Bayesian methods is their ability to accommodate incomplete observations of the epidemic [ONE99], by treating missing data as latent variables. Bayesian methods facilitate data assimilation and uncertainty quantification in a natural and unified framework [DeA15]. Finally, computational methods are rapidly advancing in this space (see Section 5) making application of those methods to real-world data sets increasingly feasible.

In Bayesian inference, parameters θ are considered as random variables and the aim of the inference is to estimate their distribution. The posterior distribution, $p(\theta|y)$, is derived from the likelihood, $p(y|\theta)$, which comes from a probability model for the observed data y, and the prior, $p(\theta)$, which encodes knowledge available before the current data was observed:

$$p(y) \propto p(y|\theta)p(\theta)$$

The essence of Bayesian inference is to update what you believe about the parameters through the observation of data. This then poses another major challenge: how do we use prior knowledge in a consistent and convenient way within models of infectious disease? Within a Bayesian setting, it is natural to specify priors on the model parameters themselves. However, experts typically cannot easily quantify their beliefs about the parameters directly; rather they will have knowledge of (and so be able to construct a prior on) observables

associated with the underlying process, such as the expected peak prevalence or the duration of the outbreak.

Statistical inference based on data generated by a single type of observation process is routine, but challenges remain when performing inference based on multiple observation processes, and/or different types of data from a variety of sources. Another challenge relates to issues around data and specifically how multiple types of data, drawn from alternative sources, can be included in the modelling framework. For example, when modelling a nascent epidemic, we might have access to case notification data, special studies (such as First Few Hundred studies), and phylogenetic data. Statistical models that integrate multiple data sources are beginning to gain traction in infectious disease modelling [DEM18, CAM19].

There is a need across all of the above issues to develop computational algorithms that can help end users to automate some of these processes. With increased volumes of and access to data (from multiple sources), algorithms need to be efficient and make use of recent advancements in computational hardware. However, there is still an important place for expert human input to gain mechanistic model insight (rather than relying, say, on machine learning techniques only, which also have their place, but are not a focus of this paper). This insight will be enhanced by addressing each of the challenges we have focused on, including considerations of parameter identifiability, the construction of priors around process observables, and the integration of data from multiple sources.

Finally, the field of mathematical epidemiology is intimately tied to the life sciences, epidemiology, and the practice of public health itself. To make an impact, that is to contribute to policy with the purpose of reducing the burden of disease and saving lives, mathematical epidemiologists need to consider the context in which their work exists. Towards the end of the article, we provide some commentary on this broader context, and how it may influence our practice.

2. Identifiability

The mere ability to fit a model to data does not guarantee that the model's parameters can be uniquely determined; parameters may not be identifiable. As an example, consider the well-known deterministic SIR model:

$$\begin{aligned} \frac{dS}{dt} &= -\frac{\beta SI}{N}, \\ \frac{dI}{dt} &= \frac{\beta SI}{N} - \gamma I, \\ \frac{dR}{dt} &= \gamma I. \end{aligned} \qquad \ldots(1)$$

At the start of the epidemic, we can make the approximation that susceptibles are not depleted ($S(0) \approx N$), which results in an approximation described by:

$$I(t) \approx \exp((\beta - \gamma)t). \qquad \ldots(2)$$

By inspection we see that the function $I(t)$ does not change if β and γ are changed, provided the difference $r = \beta - \gamma$ remains the same. Consequently, even if $I(t)$ is perfectly observed, only the value of r can be inferred, not the values of β and γ themselves. Given observation of prevalence then, β and γ are considered unidentifiable, while r is considered to be identifiable. For non-trivial models, identifiability analysis – using analytic and/or simulation-based techniques – is required to determine which model parameters are identifiable.

There are two main aspects to identifiability. "Structural identifiability" [BEL70, WHY13] concerns whether different parameter vectors produce different probability distributions of observed data. Structural identifiability ensures true parameter values can be inferred under idealised conditions: that the model is an exact representation of the system under study, and that the observations uniquely determine the probability distribution of the data. (This latter condition is only possible given an infinite number of observations.) Structural identifiability is thus a property of the model, not the specific data observed. For example, a continuous-time *deterministic* compartmental model can only be structurally identifiable if different parameter vectors generate distinct output trajectories, regardless of how precisely those trajectories are determined by a specific data set. A continuous-time *stochastic* compartmental model is structurally identifiable if different parameter vectors result in different probability distributions for the observables.

On the other hand, "practical identifiability" [GOD83] concerns whether or not parameter values can be uniquely, precisely and accurately determined for realistic measurement frequencies, quantity and quality, and in light of discrepancies between the model and the real-world process under study. Practical identifiability is thus less rigorously defined, and dependent on the specific data observed. For example, when applying likelihood-based inference methods to a continuous-time deterministic compartmental model with stochasticity in the observation process only, the model is practically identifiable if one set of parameter values maximises the likelihood, given measurements for a realistic number of measurement time points (and replicates where relevant).

Structural identifiability is typically assessed using analytic or numerical methods, whereas practical identifiability analysis is often assessed by undertaking a simulation/re-estimation study. In such a study, one selects a particular parameter vector, uses it to simulate data from the model subject to noise, conducts parameter inference, and then investigates the features of parameter estimates to determine if estimates adequately approximate assumed values.

Within structural identifiability, a distinction is made between global identifiability, whereby a unique parameter vector in the whole parameter space can be determined, versus local identifiability, where a unique parameter set can be determined in the neighbourhood of the true parameter set. Within practical identifiability, there is the distinction between *a priori* identifiability – where the results of identifiability analysis apply to all realistic data sets – and identifiability given a particular data set.

It may be taken as given that unidentifiability of some number of model parameters will greatly reduce the insight that can be drawn from fitting a model to data, and this is indeed often the case. However, the situation is more nuanced. One must consider whether the desired scientific insight rests upon the values of parameters themselves ("are the parameter values of intrinsic interest?"); or lies in use of the model to make predictions; or in testing competing mechanistic hypotheses. For these latter situations, non-identifiability is not

necessarily as significant an issue. We now discuss two types of challenges associated with identifiability. First, what challenges lie in determining *a priori* which parameters may be identifiable given a data collection process? And second, what challenges are faced when interpreting the results of fitting an unidentifiable model to data?

Challenges in determining identifiability

Identifiability analysis has not yet seen widespread uptake in biological system modelling, as described by Nguyen *et al*. [NGU16]. The authors noted that (Page 2): " … the booming works on mathematical models in biological and medical research over the last years have been accompanied with a disproportionately low amount of assessments on parameter validity …". They drew this conclusion by interrogating publication records in PubMed Central from 1990 to 2015. The authors noted that publications featuring models composed of ordinary differential equations were much more common than those which also listed keywords relating to some form of identifiability. See [NGU16, Figure 1] and its associated supplementary text for details.

Part of the challenge in increasing the uptake of identifiability analysis is raising awareness of the many pitfalls that can occur in the absence of such analysis, and thus the necessity of performing such an analysis. We note that some authors in the field have recognised the value of analytical or numerical methods of scrutinising models, and have advocated for their inclusion in modelling practice (see [BOI15] for a relatively recent example). Many studies use simulation/re-estimation methods as described above to determine whether model parameters can be estimated, but do not explicitly describe this as identifiability analysis. The clear labelling of identifiability analyses as such, and references to established methods, would raise awareness of the benefits of identifiability analysis and change community perceptions both regarding the need to conduct such analyses, and the ease of doing so.

However, even where researchers are aware of the importance of identifiability analysis, several practical problems can arise. First, ideally, a model should be practically identifiable *a priori*, but this is difficult to establish. Instead, most methods assess practical identifiability for a given (simulated or actual) data set or a given set of 'true' parameter values [RAU09, YAN19]. On the other hand, there is a proliferation of methods to assess structural identifiability (as reviewed by [CHI11]), but structural identifiability is necessary but not sufficient for practical identifiability. A challenge arises then in either improving methods for *a priori* practical identifiability analysis and/or making them more accessible, or developing methods to combine the results of structural and practical identifiability analyses for selected data sets.

Identifiability analysis presents a major technical barrier, especially to the non-specialist. The barrier is perceived to be particularly high for structural identifiability analysis, as it involves manipulation of model equations rather than simulation and parameter estimation, the latter of which are more readily accessible skills, already used in fitting models to data. A challenge for the field is to promote the use of automated tools for structural identifiability analysis. Many tools have already been developed in the context of systems biology [BEL07, CHI11, MES14, KAR12], but awareness of their utility in the epidemiological community remains low, and these are often implemented in proprietary software, limiting their accessibility.

Difficulties of structural identifiability analysis have encouraged the alternative of practical identifiability analysis using numerical methods, although as discussed above, structural and practical identifiability analyses serve slightly different purposes and are not strictly interchangeable. The simulation and estimation processes for practical identifiability analysis are relatively straightforward (although may be time-consuming), but automated tools would still lower the barrier for their application. More importantly, decisions on how to conduct such a study and interpret results are not straightforward, and there may remain issues with convincing journals, editors and reviewers that simulation/re-estimation studies are worthy of publication in and of themselves and indeed required before data analyses are undertaken and reported.

Returning to the actual conduct of practical identifiability, a number of questions arise. For example, what parameter values should be used to simulate data? How many parameter vectors should be used to provide confidence that results can be considered general? When interpreting parameter confidence intervals or posterior distributions, how narrow do they have to be for us to claim that a parameter is identifiable? Proposing robust numerical methods to provide insight into undesirable model features, and guide their remediation, remains a challenge for the community.

A review of structural identifiability analysis methods and their suitability for different model structures in infectious disease modelling would help guide practitioners in their choice of methods, as has been conducted for systems biology models [CHI11]. Although the suitability of analysis methods is obviously independent of the physical interpretation of model equations, a similar review for common model structures in infectious disease modelling would enable easier comparison between reviewed models and a particular model of interest.

Challenges in interpreting the results of fitting an unidentifiable model to data

If a model is unidentifiable, obtaining meaningful results from fitting to data and interpreting these results can be extremely challenging. Appropriate interpretation begins with an acknowledgment that the adequacy of the inference depends on how the model is to be used.

If the aim of fitting the model to data is to determine values of parameters that are of intrinsic interest, then one can examine whether it is possible (or likely) that changes to the model or to planned data collection will remedy any lack of identifiability. For example, does holding some parameters constant (or imposing strong priors on some parameters in the Bayesian context) or acquiring additional data result in an identifiable model? Reparameterisation is unlikely to help in this situation, as the reparameterised model may be identifiable, but the original unidentifiable parameters of interest will no longer appear in the model. However, information on identifiable parameters can guide reparameterization, enabling one to make stronger claims about the values of the new parameters, the biological interpretation of which can then be investigated. One advantage of testing a model for global structural identifiability is that it can reveal the parameter combinations ("observational parameters" [JAC85]) which can be determined uniquely under the idealised conditions of the test. Used with a method appropriate for the model class (e.g. one employing the notion of "structural equivalence" for linear state-space models [VAJ84]), knowledge of these combinations can guide the reparameterization of the model into one that is globally identifiable.

On the other hand, if the aim of parameter inference is to make model predictions – either for the unperturbed system or in the context of an intervention – we may seek to propagate parameter uncertainty through our model so as to produce a range of predictions, allowing us to quantify prediction uncertainty. It is possible that although parameters are not individually identifiable, parameter sets consistent with observations make similar (or identical) predictions, or that the quantitative behaviour of certain subsets or functions of parameters are well determined despite lack of parameter-level identifiability (e.g. [YAN19]). However, if interventions act by changing the values of unidentifiable parameters or unidentifiable combinations thereof, predictions are unlikely to be consistent. For example, consider the SIR model and suppose that there exists an intervention which halves β for all time after the intervention is applied to System (2). Figure 1 shows that alternative values of $(\beta/2, \gamma)$ lead to a wide range of predictions for I over this post-intervention time period. As such, uncertainty over the true values of β and γ creates doubt over the benefit of the proposed intervention. Additional data collection and/or fixing model parameters may be required, as discussed above. Reparameterisation is unlikely to be helpful in this context, as the new parameters will be identifiable but not linked directly to the intervention (for example, reparameterising system (2) as $I(t) = I(0)\exp(rt)$, removing β). An exception is if new parameters are subject to stronger priors than the original parameters, enabling more precise inference of the values of the original parameters.

In summary, addressing the challenges associated with identifiability will provide the disease modelling community with a systematic means of comparing models and evaluating their usefulness. In turn, we expect this to enable progress on the discipline's fundamental challenges, in using models to direct resources towards ensuring better health outcomes.

3. Incorporating prior knowledge

The ability to identify parameters, or at least to have distributions on parameters which capture our full knowledge of the disease process, is dependent upon use of prior knowledge. When analysing a (new) data set within the Bayesian framework, we must specify a prior distribution on the parameters of the model. This provides a natural way to incorporate existing information (obtained from the literature, past experience etc.) about plausible values for the parameters. The prior distribution also offers a way to incorporate information about observable quantities (i.e., properties of the system that can be measured and expressed as a function of model parameters). This requires a clear distinction between knowledge of the real-world system (and our understanding of it) and knowledge of the model we are using to represent it [GEL17, CRA97].

The Bernstein-Von-Mises theorem tells us the likelihood will (asymptotically) come to dominate over the prior as the amount of data increases provided relatively mild technical conditions are met [KLE12]. However, one should not use this as an excuse to neglect the choice of prior distribution. In cases where data is limited, there is a risk of a (potentially incorrect) prior dominating the analysis; the result of the analysis will not reflect the data (and desired distribution over parameters of interest) but rather the (potentially incorrect) prior information.

Furthermore, when analysing infectious disease data, it is unusual to have large amounts of high-quality data, and hence it is very attractive to supplement our data with prior knowledge. A weakly informative prior can assist with some of the statistical identifiability issues discussed in the section above on Identifiability, but if the prior distribution is poorly

specified (even if a non-informative prior is used) it can lead to misleading results [GEL17]. A desirable solution is to select a prior distribution that concentrates prior probability on plausible parameter values but does not dominate the posterior distribution, i.e. it allows the information in the data to determine the outcome of the inference. But doing so may be difficult.

Mechanistic models of infectious disease dynamics are often expressed in terms of parameters far removed from the aspects of the process that are observed. It is arguably rare for scientists to have good knowledge of model parameters; rather, they have, or could construct, an informed view of various system observables. To make explicit the difference between model parameters and observables, consider a mathematical model of influenza infection within a host. *A priori*, one may not know plausible orders of magnitude for the rate parameters (of the mathematical model), but likely they will know that an influenza infection resolves in days, rather than hours or months.

Given that scientists often have a better grasp on the value of observable quantities rather than the model parameters themselves, it makes more sense to articulate our prior belief in terms of these observable quantities. But of course, we then have a task to translate between the two. This challenge is not universal across models. Consider logistic growth describing the size of a population through time, *N(t)*, with the following differential equation:

$$\frac{dN}{dt} = rN\left(1 - \frac{N}{K}\right).$$

It is simple to parameterise this model in terms of the growth rate, *r*, and the carrying capacity, *K*, both of which are natural quantities to observe. Consequently, if one were interested in developing a prior distribution for this model, one would only need to specify the distribution of plausible values to observe for these quantities. Considering the alternative equation

$$\frac{dN}{dt} = rN - cN^2,$$

the observable consequences of different values of the parameter *c* are less obvious, and hence it is unclear what plausible values may be. This situation is common in mathematical epidemiology: the observable quantities are themselves a function of the solution to the model, and most disease dynamic models do not have any closed analytic form for these quantities. Consequently, it is infeasible to develop a clearly interpretable parameterisation – in terms of writing down a prior based on past observations – of these models.

The problem extends beyond just the parameterisation used. It is tempting to think that it is sufficient to use published estimates of individual, model-inferred parameters to construct a prior distribution. However, even when this is possible, issues will still arise. Reported parameter estimates are model dependent. The quintessential example being use of an SIR or an SEIR model, whereby inference may lead to the same estimates of the basic reproduction number but different estimates of the rates of infection and recovery. Understanding the marginal distributions of parameters is insufficient to construct a plausible prior distribution when correlations determine model behaviour.

If we are serious about incorporating prior knowledge into future analyses, then it seems sensible that we should turn to fields where this has already successfully been implemented.

Priors informed by expert opinion (obtained via expert elicitation) or the results of previous studies have proved popular in the field of ecological where they have found diverse applications [HEM17, LOW09]. However, the use of expert elicitation to inform priors has become much more popular among ecologists over the past decade [DRE13]. Using expert opinion in conjunction with Bayesian inference is an attractive option for researchers in ecology, where the types of data collected are likely to involve a high degree of uncertainty [KUH10, MAR05], or are difficult or expensive to collect [OHA98, MAR12].

Martin et al. [MAR05] solicited expert opinion for a study regarding the effect of livestock grazing on various Australian birds. In cases where experts agreed with each other, the resulting credible sets for parameters are typically tightened, this can be an effective and cost-effective way to improve estimates. This improvement was particularly noticeable in situations where existing data was weak but the precision, or agreement, between expert information was high. Just as important is the case of incorporating expert knowledge where there was noticeable disagreement between experts; in this situation, results did not differ significantly from analysis where expert opinion was not used.

Expert opinion can also improve confidence in parameter estimates obtained from analysing these data, although care should be taken to ensure the elicitation of this information is carried out correctly [MAR05,MOR13]. There is a need to be rigorous in the selection of experts and the execution of the elicitation [DRE13]. Kadane and Wolfson [KAD98] and Chaloner et al. [CHA83] have developed methodology which abstracts much of the mathematical detail, simplifying the elicitation process. However, these methodologies only cover certain types of models, and are limited in scope and the degree to which they scale with the complexity of the model being considered. We believe it would be beneficial for the modelling community to extend this work, and develop statistical methodology that enables “automatic elicitation” for a wider range of models.

While peer-reviewed literature provides a rich source of 'prior knowledge' and would typically form the basis for determination of the prior, caution should be exercised due to potential systematic effects in publication practice that can make it difficult to reliably source and account for all (published) primary sources [REI11]. In that paper, Reich et al. demonstrated that, due to publication and referencing practices, so called 'medical facts' can become enshrined as truth in the absence of strong and sufficient empirical evidence. This strongly suggests that when using the peer-reviewed literature to establish a prior, one must proceed carefully, being sure to establish a process to identify relevant primary source literature and avoid the pitfalls identified by Reich et al.

Given the importance of the prior distribution and that there may be substantial amounts of knowledge about observable quantities of the process being modelled, how might one go about using this knowledge to specify a prior distribution? An idealised work-flow to utilise prior knowledge may consist of the following steps: (1) determine a set of relevant observable quantities which characterise the system and for which there is some quantitative understanding; (2) construct a prior distribution on the parameters of the model which reflects this understanding; and (3) carry out the remainder of the inference process as normal. There are three aspects to this work-flow which are challenging: (1) determining appropriate observable quantities; (2) representing this information in the prior distribution of parameters for an arbitrary model; and (3) devising a way to do this which is not prohibitively computationally expensive. We now expand on these three challenges.

(1) *Which observable quantities?* The first aspect is the most specific to the particular application, in that it requires some knowledge of the system being modelled to know what observable quantities of the process characterise it. When multiple observable quantities are being used, strong correlations between the quantities leads to redundancy; ideally independent observable quantities would be used. Above, the time required to resolve an influenza infection was given as an example of an observed quantity. This quantity may be expressed in terms of the solution to a particular model, even if it is not one of the parameters of the model. For example, if one were modelling the number of people hospitalised with influenza during an epidemic, one might use the total number of patients hospitalised as an observable quantity. Historical records of hospitalisation could be used to estimate, a distribution for this for previous epidemics. In the SIR model, the final size is determined by the initial condition and basic reproduction number, hence information about the final size can constrain the prior distribution for these parameters [MIL12].

(2) *How to represent prior knowledge?* The second aspect involves the process of taking a representation of the uncertainty in, potentially several, observable quantities, and translating this into a prior distribution over the parameters of a model. This is a non-trivial task, even when the distributions on the observable quantities are self-consistent (and it is easy to construct examples where this is not the case). Moreover, when there are correlations between observable quantities this can further complicate matters if one is not content to assume a joint distribution with independent components.

(3) *Can choosing a prior distribution be made easier?* Finally, the third aspect involves finding a way to efficiently choose a prior distribution. As discussed above, extensive research from the field of psychology suggests that eliciting information from experts in a defensible way is labour-intensive. However, there are alternatives to obtaining information from domain experts. With the rise of “big data” there will be an increasing amount of data to be mined to inform prior distributions. The use of additional data sources brings its own challenges, e.g., accounting for correlations between the data sets as discussed further in the subsequent section on Data Challenges.

While solutions to these problems would improve our ability to carry out inference, there is another equally important conceptual contribution from this work-flow. Decoupling prior knowledge of observable quantities and prior distributions from specific mathematical models allows us to create prior distributions which can be shared between models and capture the same information, independent of how that information translates into a distribution on the parameters of that particular model. This itself is a powerful idea, as it increases the portability of parameter estimates as they are no longer attached to the particular model with which they were obtained. For example, suppose models X and Y share observable quantities but do not have the same set of parameters; it is possible to use parameter estimates obtained with X to construct an equivalent prior for Y, where "equivalent" refers to having the same distribution on observable quantities. Moreover, since such observable quantities will often involve a combination of the model parameters, even if the prior distributions of the observable quantities are independent, the prior distribution they enforce on the parameters may have a rich correlation structure.

Spending considerable effort on choosing a prior distribution can seem indulgent and, provided there is sufficient data, it will often have only a small effect on the results. Debate over the choice of prior has inspired many developments in Bayesian statistics, and there are cases where it has a strong effect on the inferences drawn [MOSS19]. In this section we have

described some open problems regarding the practical application of informative prior distributions while attempting to motivate why, despite these difficulties, they can still be very beneficial. In the words of Judea Pearl, "It is plain silly to ignore what we know" [PE01], however this comes with the caveat that it relies on this knowledge being appropriate and accurate.

4. Challenges posed by data

As discussed by De Angelis et al. [DeA15], precise and accurate model inference relies on the volume and quality of data, with quality encompassing both variability and bias. If the observation process is well-characterised, parameter estimation can still be performed, although the precision of these estimates will be limited by the quality of the data. Combining data from multiple sources can increase the precision and accuracy of estimates but presents its own challenges. In this section we will first discuss challenges in characterising the observation process for a single epidemiological data set, before moving on to those for combining multiple data sets.

Challenges within a single epidemiological data set

When data measuring one quantity is used as a proxy for another quantity, the relationships involved should be well-characterised to reduce bias in inferred quantities. For example, the proportion of individuals reporting influenza-like illness (ILI) in a weekly community survey can be used as a proxy for influenza prevalence in the general population [ADL14, CAR10]. Here, ILI is used as a proxy for influenza virus infection; however, influenza infection prevalence may be overestimated due to ILI caused by non-influenza pathogens, or underestimated due to asymptomatic infections. Moreover, prevalence in the survey population is used as a proxy for prevalence in the general population, but demographics of the survey population may not reflect that of the general population. To reduce bias during inference, the relationship between reported ILI and influenza virus infection, and between the survey and general population, should be explicit in the observation process. For example, the former can be achieved by specifying a reporting probability conditional on influenza infection, and a background observation probability due to illnesses other than influenza that may vary over time.

The biases which are likely to affect inference, and are thus important to model, are likely to differ by both situation and data type. For example, when performing inference during outbreak scenarios, if lags in data collection are either ignored or mis-specified, inference may be poorly affected [AZM14, MOSS19]. However, even for the same data type (such as incidence data), such lags may not drastically affect inference in endemic scenarios. Reporting rates are also more likely to vary during the course of an outbreak, as indicated by Flutracking data [CAR10]. Many studies calculating time-varying effective reproduction numbers (e.g. [ROS15]) are not robust to time-dependent reporting rates, so if these methods are used, time-dependent reporting rates should be included in the observation process.

Where available, data sources covering the same timeframe as the primary data should be used to inform biases. For example, community survey data can be used to infer changes in healthcare seeking behaviours and testing practices over the course of an epidemic, which can then be used to improve epidemic forecasts using a different data source [PEP17, MOSS19a]. In other situations, sensible observation models and/or parameter values can be obtained using historical data. For example, observation noise can be estimated for previous

epidemics, and resampled from when proposing parameters to fit to data from a new epidemic [ERT18]. Where the relationship between historical and current observation processes is unclear, rather than assuming that the historical and current observation processes are the same, a better approach may be to use historical parameter values to inform values of the current observation process. For example, different outbreaks of the same pathogen may occur in different locations and in different populations, and it is unclear how observation biases translate across outbreaks. When we know the direction in which a parameter will change but not by how much – for example, assuming that testing rates will increase in a pandemic – a historical parameter value could be used as the lower bound for a prior distribution on the testing rate or, more conservatively, to construct a prior distribution where only a small proportion of probability mass is below this value. Another example is when an intervention increases the testing probability. This scenario requires particular care, as increased testing may increase observed prevalence even when an intervention is effective, and inference ignoring increased testing may incorrectly conclude that the intervention increases prevalence [ALI15]. Conversely, increased testing (motivated, say, with the aim of ascertaining every case possible) could decrease test-positivity, if specimens are collected indiscriminately and testing denominators are unavailable, and this could lead to under-estimates of prevalence.

When inference is conducted on "incidentally available" data rather than data collected for the particular inference study, extra attention has to be paid to modelling of the observation process. This is especially an issue in outbreaks, as surveillance protocols are developed alongside the unfolding of the outbreak. Hay et al. [HAY18] documented that case definitions for microcephaly became more stringent as the Zika outbreak in 2015-2016 developed, and that many cases were reclassified. Either a combination of behavioural change and overreporting of cases under early definitions, or increased Zika surveillance between the two epidemic waves, were required to explain changes in reported microcephaly incidence. Communication between field workers, policymakers and modellers becomes especially important in this context, and local modellers have the opportunity to inform the data collection process (see Policy and Communication section).

On the other hand, closer ties between designers of data collection protocols and developers of inference methods have enabled the collection of data sets designed to infer particular model parameters, as identifiability of model parameters may depend on the study design (see Identifiability section). For example, in the 2009 influenza pandemic, first few hundred (FF100) studies were conducted specifically to understand the transmissibility and severity of the disease during the early stages of the epidemic [McL10]. Since the collection of this dataset, model-based inference methods have been developed to infer hospitalisation rates and within-household transmission in real time [BLA17].

Challenges when using multiple epidemiological datasets

When modelling transmission of an infectious disease there are often multiple different epidemiological data types available with which to infer model parameters. For example, when inferring key characteristics of a nascent epidemic, we might have access to confirmed case counts, syndromic surveillance data and special studies (such as FF100), but each carries its own underlying biases and there is no guarantee that they provide a self-consistent view of disease activity [e.g. THO15]. While there are established approaches to performing inference with each of these types of data, we rarely use methods that can simultaneously consider all available data although there are notable exceptions [COR19]. Consequently,

typically the result is either multiple competing parameter estimates – which leads to obvious challenges for decision-makers – or a single estimate that ignores some of the available information. The alternative is to analyse all available, relevant data using a single joint model. Ideally, the joint model is able to integrate the different data sources in a way that retains the strengths of each, without losing information. De Angelis et al. [DeA15] note that the motivation to combine information from multiple data sources arises from both a perception that this will produce more `defendable', robust outputs and a recognition that comprehensive outbreak analysis requires multiple data types. Here we will focus on data-integrating models for multiple epidemiological time series data but note that models for integrating these and other epidemiological data types and/or phylogenetic datasets have recently gained traction in outbreak analysis [DEM18, CAM19].

Integrating multiple different datasets can increase the precision and accuracy of parameter estimates and enable a greater range of relevant (unobservable) quantities to be estimated [BIR18]. For example, sharing information across abundant, low-quality surveillance data (*i.e.*, high volume, but unknown or poorly characterised observation processes) and a subset of high-quality surveillance data (*i.e.*, well-characterised observation processes, but low volume) can enable the estimation of nuisance parameters, like reporting biases. This approach has previously been demonstrated in ecology [FIT15], where data structures are not dissimilar to disease data. Another advantage of joint inference is the automatic weighting of information from different data sources. When writing out a joint model, explicitly describing the observation models for each data type (including parameters for reporting biases) provides an objective way to weight their respective utility.

An important challenge when constructing a joint model is the handling of dependencies between data sources. Here we separate these dependencies into two types: 1) data sources observe the same underlying epidemiological process, and 2) observation processes themselves are dependent (for example, individuals may be captured by two or more surveillance systems).

Understanding dependencies in the observation process will require close collaboration with data collectors and public health policy-makers [MUS17, DOM18]. In order to understand the magnitude of dependencies in the observation processes between outbreak surveillance datasets, it is important to know precisely how each dataset is assembled and to map out all possible research and health system pathways that could lead to an individual being counted in one or more dataset(s). For example, during an influenza pandemic, households participating in FF100 studies are likely to be recruited from routine case notification systems and would therefore be counted in both FF100 data and case notifications. Further, if we wanted to add information from community survey data such as from Flutracking [CAR10, MOSS19a], we must consider how duplications of these data may arise in FF100 studies and/or case notification datasets.

While it may be possible to write out a single model that links all available datasets, there could be practical hurdles to performing inference for such models. It is important to consider whether inferring parameters from multiple datasets simultaneously, and thus adding substantial model complexity, is worthwhile from both a computational and a decision-making perspective, particularly if supporting decision-making in real-time is a goal. For example, Shubin and colleagues [SHU16] made simultaneous use of data from community and hospital surveillance systems in their transmission model of pandemic A(H1N1)pdm09 influenza, but the computation time for inference is reported in months, which is not practical

for real-time use (note that this was not the goal of their analysis). Moreover, if the goal of such modelling studies is to inform public policy (in real-time or otherwise), the model structure and appropriate interpretation of its outputs will need to be clearly communicated to decision-makers, and more complicated models may be more difficult to translate (see Section 6).

It should also be noted that integrating multiple, low-quality data sources with a single high-quality dataset, will not necessarily provide benefit over analysing the high-quality dataset alone. For example, Moss and colleagues [MOSS17] found that simultaneously using data from three different surveillance systems only improved retrospective seasonal influenza forecasts under certain circumstances, and could even reduce forecasting performance, when compared to forecasts generated using a single data source. They hypothesised that the synthesis of data from multiple surveillance systems may only provide benefit if each data source captures distinct, but complementary, aspects of the epidemiological or observation process.

5. Computational methodology

In common with other areas of mathematical biology, methods for fitting complex epidemic models have progressed a lot recently, driven by advances in Bayesian computational statistics [GRE18]. These methods can be classified in a number of ways and our taxonomy reflects our personal biases and preference for mechanistic, stochastic, models. While non-mechanistic models are useful for some forecasting problems where relatively large amount of historical data are available [BRO18], small data sets, as would be available in the event of an outbreak, can only be interpreted in a mechanistic setting, and likewise the testing and forecasting of various intervention strategies. A fundamental difficulty with inference from outbreak data is that most of the underlying process is unobservable, hence the need to infer or integrate over a large amount of missing data to sample from the parameter posterior. One way of classifying existing algorithms is according to which part of the calculation handles the missing data. This impacts how suitable they are to be parallelised and hence handle larger problems as well as incorporate other evidence, including multiple datasets [BRO11, BIR18].

For models of small, closed populations such as households, continuous-time Markov chain (CTMC) models have found success, due to the size of the state-space being small enough to leverage numerical solutions for calculating the likelihood [BLA17]. For most models, in larger populations, these methods break down due to the increased size of the state space. The oldest methods for exact inference are so called data-augmented (or auxiliary variable) MCMC (DA-MCMC) [ONE99]. These typically infer the missing data as well as the parameters as a single Markov chain from which an expression for the likelihood is trivial to evaluate. Samplers are also easy to construct using a combination of Gibbs and Metropolis-Hastings steps. Data-augmented methods are highly flexible, allowing the use of non-Markovian models, non-homogeneous mixing and detailed spatial information [TOU18], [STO17]. The downsides are common difficulties with convergence and mixing that get worse as the amount of missing data to be inferred grows [McK14]. Efficient use is reliant upon conjugate priors (allowing the posterior to be specified explicitly), so incorporating more general, informative priors (as discussed earlier) can be challenging. Finally, DA-MCMC is fundamentally a serial algorithm, so its use on large datasets becomes slow and parallelism is not easily exploited, beyond running multiple chains.

Although almost all useful epidemic models are analytically intractable, they are typically very simple to simulate. Approximate Bayesian computation (ABC) uses simulations for fitting models where the likelihood is intractable [KYP17], but where the simulated data can be compared with summary statistics. This is probably the most simply implemented method in this class but comes at the cost of introducing some approximation into the posterior. Other methods are exact in that they use an estimate of the likelihood, but still target the correct posterior. The use of sequential Monte Carlo (SMC) methods (which perform estimation sequentially through data) in epidemic modelling is very natural due to the prevalence of time series data and the need to fit dynamical models [DOU01]. Pseudo-Marginal methods such as particle marginal Metropolis-Hastings exploit the unbiased likelihood estimate obtained from a particle filter to also perform inference for the underlying parameters [AND10, BRO15]. In comparison with data-augmented methods, these methods can be seen as integrating over the missing data in the estimation of the likelihood, so the Markov chain targeting the parameter posterior is greatly simplified.

The key to the efficient operation of these pseudo-marginal algorithms is keeping the variance of the likelihood estimate within tight bounds [DOU15, SHE15], otherwise the mixing of the Markov chain targeting the posterior can become very poor. An advantage of pseudo-marginal methods is that they are parallelised quite naturally, so modern computing hardware can be leveraged to reduce the variance in the likelihood estimate by simply using larger numbers of particles of averaging the estimates of independent particles. Although these methods represent the current state of the art in this area, there are still challenges to be overcome. They are not ‘online’ in that the computational expense (and hence run-time) increases as the length of the time series increases [KAN15]. Non-Markovian models remain challenging and the overall efficacy is restricted by the ability to produce simulations that are in some way close to the observed data. Current research employs variance reduction techniques to reduce the variance of the likelihood estimate, and in particular importance sampling has been used to produce realisations that match data closely [BLA18, McK14]. In the case where the model is not strictly non-identifiable, but there is a complex posterior distribution over the parameters, then MCMC methods based on Riemannian geometry can be used [HOU16].

Model Selection

Using data to accurately infer parameters of epidemic models is an important step for informing public health policy, forecasting and understanding the dynamics of diseases. However, if the models are inappropriate these tools generate misinformation. Information criteria such as AIC, BIC, DIC are the most common method for deciding on the best model. These information criteria are used widely by both frequentists and Bayesians due to their asymptotic properties and their often ease of calculation. However, in cases where there is little data and reasonable prior understanding of the epidemic process, these criteria may fall short. Simply put, this is because AIC, BIC, and DIC are intrinsically non-Bayesian as their formulations do not account for prior knowledge of model parameters.

The gold standard for selecting between models while accounting for prior information is to either calculate Bayes factors or the model evidence [KAS95]. These approaches are sometimes avoided due to their computational difficulty; they either require calculation of the normalising constant of the posterior distribution, or calculation of a ratio of normalising constants. Although this problem is difficult via classic methods such as reversible-jump

MCMC [GRE95], there are increasingly efficient methods for performing this kind of model selection. One approach is SMC^2 [DRO16, CHO13], which allows for model selection to be performed during the inference process. Although this is an attractive and efficient method, the stochastic error in model selection estimates is not well understood. An alternative is importance sampling-based methods [GEL94, TOU18], which give unbiased estimates for model selection along with estimates of error. While these methods are computationally intensive, they are made efficient if parameter inference is performed *a priori*. Further, they are embarrassingly (i.e. trivially) parallelisable, that is, they are able to take advantage of modern computational architecture.

Another recent approach to model selection has been through the use of classification methods, and in particular Random Forests [PUD16]. This approach has the benefit of needing only to simulate from the model, and efficiency of the classification algorithms themselves. This is particularly important if one is to consider the optimal design for model selection [HAI18], a particularly computationally-expensive pursuit, and for which heuristics based upon the Random Forest approach have been proposed recently [COP19].

6. Policy and communication

Until now, we have concerned ourselves with some of the key technical challenges in modern day mathematical epidemiology. But for our work to contribute to public health policy, we must consider the broader scientific, social and political environment in which it exists. This is a broad topic and one not unique to health, for example ecology provides a highly-relevant exemplar discipline in which modelling has had a sustained and meaningful impact on decision making [e.g. BAL09]. A central tenet of the approach in that field is that we must distinguish, from the outset, between science for knowledge and discovery's sake, and science for the express purpose of contributing to the decision-making process. Within this context, we make the following observations.

Mathematical models are developed by researchers from a broad range of backgrounds, many of whom do not necessarily have the knowledge required to translate the intricacies of structural identifiability analysis and Bayesian approaches to parameter estimation into practice. In general, more complex models tend to be favoured by policy makers as they are perceived as being more "realistic", and indeed such models likely have more internal validity. Yet they are, in all likelihood, less general and have weaker external validity, in terms of providing unbiased predictions for other related scenarios or situations. Complexity can be a desired quality of a model even if the data are not available to support the model structure. Thus, there is a tension between transparency (which correlates with incompleteness but also with generality) and realism (which correlates with complexity and opacity, but also completeness and lack of generality). Highly complex models are in danger of projecting a false sense of accuracy, with the portrayed accuracy enhancing their attractiveness to policymaking stakeholders, but with untenable policy recommendations as a potential outcome of this cycle [COO06]. And even if accurate for the scenario at hand, findings drawn from them will be less generalisable (because they are more specific). Clearer communication of the pitfalls of the "unconscious use of a non-identifiable model" [SIE12] will help to avoid this trap.

As the epidemiological modelling discipline globalises and nascent modelling groups form in countries previously lacking this capacity, our community should consider not only training modellers to model, but also explore how models are used in the policy environment. It is

important to acknowledge that this environment will differ, and in strong ways, across different social, economic and political contexts. We propose that such topics are included in future training activities offered by established groups. We encourage success stories of models influencing public health policy (or otherwise) be published. The practice of modelling goes far beyond the technical details of the mathematics, statistics and biology.

Policy makers are not the only ones with whom the modellers need to communicate effectively. During the lifecycle of the model development, we need to understand the intricacies of the problem, the available data and the existing views on reasonable values for parameters. Therefore, communicating and involving the experts in the respective biological, medical and health fields should lead to better calibrated and validated models.

Effective interactions among these three key groups (modellers, researchers from other disciplines and policymakers) relies on effective communication, with several themes and foci for improvement (Figure 2).

Developing a shared understanding of the problem

Modelling that strives to inform policy is best conducted in consultation with policymakers. Such engagement enables modellers and policymakers to gain a mutual understanding of the policy question to be addressed and define specific modelling objectives. It is also an opportunity to improve modeller understanding of the policy context and stakeholder understanding of the capabilities and limitations of mathematical models within this context. Examples of productive engagement between policymakers and modellers exist in many fields, including infectious disease epidemiology [LEE13, PRO16, KNI16, QUA17, MOSS19]. Recently, researchers have explored the use of participatory approaches to modelling, where policymakers and modellers "co-develop" models and applications [GAY19]. These examples suggest that a variety of approaches to stakeholder engagement, e.g., in terms of the frequency and timing of consultations within the cycle of model development, can lead to successful outcomes. Modelling conducted in the absence of stakeholder engagement is at risk of being underused, mis-used or inappropriate for addressing a specific policy question [GLA11, MUS17, DOM18].

Understanding the available data

Data is required to calibrate, validate or fit the mathematical models. As part of the planning phase of the model development, knowledge of what data are or will become available is essential. If the data is not publicly available, it is essential that communications are made with the proprietor of the data and have prior agreement with them on data usage.

Modellers have to interact with researchers from other disciplines such as biology, medicine, pharmacology etc. to make more sense of the available data. In these instances, improved communication of modelling ideas to a non-mathematical audience would allow for better discussion on whether the important biological aspects are captured from the model and how the data can be used to estimate the model parameters.

If multiple datasets are available, the use of all datasets in a single data-integrating model may be appealing to both modellers and policymakers – under the assumption that more data will result in greater precision. As discussed earlier, a data-integrating model may not provide

benefit over the use of any single dataset unless each data source captures distinct, but complementary, aspects of the epidemiological or observation process [MOSS17]. When deciding whether data integration is an appropriate and feasible approach to addressing a specific policy question, it is important for modellers to understand how datasets may be dependant and/or complementary. This process will be most insightful when done in close consultation with data-collectors and data-users.

Prior distributions informed by experts

Constructing models to aid decision making in public health requires striking an appropriate balance between accurately representing the system being modelled, and making simplifying assumptions. It is often helpful to construct these models using domain knowledge. The Bayesian framework, with its concept of a prior distribution, provides a natural way to incorporate this information [GEL04]. However, representing this knowledge into parameter values is not always straight-forward [KUH10]. The process of comprehensively eliciting prior knowledge requires substantial effort from both the elicitor and the responder. However, as seen in the field of ecology, this can be a worthwhile approach [KUH05, MAR05, CHO09]. Care must be taken to mitigate the impact of psychological biases such as 'anchoring' and the 'conjunction fallacy' which can lead to poor representation of prior knowledge [KYN08, KUH10]. Moreover, if the responder is unfamiliar with probabilistic concepts this can also hamper the elicitation process [MOR13]. Substantial effort in the fields of both statistics and psychology has produced a number of guidelines to assist in carrying out successful elicitations [SPE75, KYN08, CHO09, KUH10, MOR13, OHA19].

As discussed in Section 3 above, there are also technical issues relating to the incorporation of the elicited information.

Sharable methodologies and best-practice for open and transparent modelling

To prevent reinventing previous work, modellers can provide adaptable, user-friendly routines or packages with worked examples supported by published theory and community-wide consensus. Reproducible methodologies and model source code can be shared on freely available code repositories such as GitHub [https://github.com/], Bitbucket [https://bitbucket.org], et cetera. There should be sufficient documentation provided with the source code in order to ensure reproducibility. Anonymised or simulated data could also be shared if necessary. Knowledge on the various types of licenses that are available to protect intellectual property while maintaining reproducibility is essential for modellers.

Model predictions could be shared through interactive web-applications such as Shiny apps [http://www.rstudio.com/shiny/] and plotly/Dash apps [https://plot.ly]. These web-applications can also help during the iterative process of communicating the model to policymakers and end-users, and improving the model based on their feedback. Computationally expensive models may not be able to run in real-time. In such cases, results could be pre-generated and stored as data which can then be retrieved for a specific scenario. Examples of some shiny apps based on models, together with their source code can be found here [TUN17, CEL19].

Communicating model predictions in policymaking

When communicating model-based insights to policymakers, it is important to present those results in scientific and statistically rigorous language but also in a clear and transparent way to a non-technical audience. Ideally, results should be presented in a way that policymakers and stakeholders can quickly understand to assist their decisions. Various types of communication such as written briefings, informal meetings, and technical interfaces could be used as necessary. Uncertainty in the predictions should also be communicated adequately. The key to successful communication in this context however is to recall that policy development is primarily concerned with **decision-making**, rather than scientific discovery *per se*. With this in mind, scientific findings can be presented in a way that focuses on the decision to be made, providing the scientific evidence as the (rigorous and transparent) basis on which advice is provided.

7. Conclusions

From the very earliest attempts to represent disease dynamics mathematically, the relationship between epidemic models and data has been both of clear importance and a major challenge [ABB52]. Herein we have focused on key challenges with a particular focus on using modelling to inform public health policy:

(i) Model parameters may not be able to be identified uniquely, either in general or for the specific data available in an application. Awareness of this is critical to ensure that robust policy conclusions are drawn from models;

(ii) Prior knowledge can be highly valuable, but it must pertain appropriately to the disease dynamics of current interest. Specification of prior knowledge on disease process observables assists in consistent and readily-interpretable specification; and,

(iii) Data must be modelled appropriately accounting for the observation process. Increasingly we have access, and computational resources, to exploit multiple datasets. However, the dependencies that arise through the underlying epidemiological system and in the sampling process in the observation models must be accounted for to draw robust conclusions.

We have discussed recent progress and new perspectives on each of these challenges, along with recent computational advances in methods with a particular focus on inference. Each of these areas remain active research topics, where advances are critical to improve the robustness, appropriateness and sophistication of model-based, policy-relevant outputs.

Acknowledgements

We thank the mathematical research institute MATRIX in Australia where part of this research was performed.

References

[ABB52] Abbey H. An examination of the Reed-Frost theory of epidemics. Hum Biol. 1952 Sep;24(3):201-33.

[ADL14] Adler AJ, Eames KT, Funk S, Edmunds WJ. Incidence and risk factors for influenza-like-illness in the UK: online surveillance using Flusurvey. BMC Infect

Dis. 2014 May 1;14:232. doi: 10.1186/1471-2334-14-232.

[ALI15] Ali H, Cameron E, Drovandi CC, McCaw JM, Guy RJ, Middleton M, El-Hayek C, Hocking JS, Kaldor JM, Donovan B, Wilson DP; Australian chlamydia incidence estimation group. A new approach to estimating trends in chlamydia incidence. Sex Transm Infect. 2015 Nov;91(7):513-9. doi: 10.1136/sextrans-2014-051631.

[AND10] Andrieu, C. , Doucet, A. and 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

[AZM14] Azmon A, Faes C, Hens N. On the estimation of the reproduction number based on misreported epidemic data. Stat Med. 2014 Mar 30;33(7):1176-92. doi: 10.1002/sim.6015.

[BAL09] Ball, I.R., H.P. Possingham, and M. Watts. 2009. Marxan and relatives: Software for spatial conservation prioritisation. Chapter 14: Pages 185-195 in Spatial conservation prioritisation: Quantitative methods and computational tools. Eds Moilanen, A., K.A. Wilson, and H.P. Possingham. Oxford University Press, Oxford, UK.

[BEL70] R. Bellman, and K. J. Åström. "On structural identifiability", *Mathematical Biosciences* **7**, no. 3-4 (1970) 329-339. https://doi.org/10.1016/0025-5564(70)90132-X

[BEL07] Bellu G, Saccomani MP, Audoly S, D'Angiò L. DAISY: a new software tool to test global identifiability of biological and physiological systems. Comput Methods Programs Biomed. 2007 Oct;88(1):52-61.

[BIR18] Birrell, Paul J.; De Angelis, Daniela; Presanis, Anne M. Evidence Synthesis for Stochastic Epidemic Models. Statist. Sci. 33 (2018), no. 1, 34--43. doi:10.1214/17-STS631.

[BLA17] Black AJ, Geard N, McCaw JM, McVernon J, Ross JV. Characterising pandemic severity and transmissibility from data collected during first few hundred studies. Epidemics. 2017 Jun;19:61-73. doi: 10.1016/j.epidem.2017.01.004.

[BLA18] Black, A.J., Importance sampling for partially observed temporal epidemic models. Stat Comput (2019) 29: 617. https://doi.org/10.1007/s11222-018-9827-1

[BOI15] Boianelli, A., Nguyen, V., Ebensen, T., Schulze, K., Wilk, E., Sharma, N., ... & Meyer-Hermann, M. (2015). Modeling influenza virus infection: a roadmap for influenza research. *Viruses*, *7*(10), 5274-5304. ; https://doi.org/10.3390/v7102875

[BRO11] Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng, *Handbook of Markov Chain Monte Carlo*, Chapman and Hall/CRC, New York, 2011.

[BRO15] Brooks LC, Farrow DC, Hyun S, Tibshirani RJ, Rosenfeld R. Flexible Modeling of Epidemics with an Empirical Bayes Framework. PLoS Comput Biol. 2015 Aug 28;11(8):e1004382. doi: 10.1371/journal.pcbi.1004382.

[BRO18] Brooks LC, Farrow DC, Hyun S, Tibshirani RJ, Rosenfeld R. Nonmechanistic forecasts of seasonal influenza with iterative one-week-ahead distributions. PLoS Comput Biol. 2018 Jun 15; 14(6): e1006134. doi: https://doi.org/10.1371/journal.pcbi.1006134.

[CAM19] Campbell F, Cori A, Ferguson N, Jombart T. Bayesian inference of transmission chains using timing of symptoms, pathogen genomes and contact data. PLoS Comput Biol. 2019 Mar 29;15(3):e1006930. doi: 10.1371/journal.pcbi.1006930.

[CAR10] Carlson SJ, Dalton CB, Durrheim DN, Fejsa J. Online Flutracking survey of influenza-like illness during pandemic (H1N1) 2009, Australia. Emerg Infect Dis. 2010 Dec;16(12):1960-2. doi: 10.3201/eid1612.100935.

[CHA83] Chaloner K.M., Duncan G.T. Assessment of a beta prior distribution: PM elicitation. Journal of the Royal Statistical Society, Series D (The Statistician). 1983;32(1):174-180.

[CHI11] Chis OT, Banga JR, Balsa-Canto E. Structural identifiability of systems biology models: a critical comparison of methods. PLoS One. 2011;6(11):e27755. doi: 10.1371/journal.pone.0027755.

[CHO13] Chopin N, Jacob PE, Papaspiliopoulos O (2013) SMC^2: An efficient algorithm for sequential analysis of state space models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75(3):397–426.

[CHO09] Choy, S. L., O'Leary, R. and Mengersen, K. (2009), Elicitation by design in ecology: using expert opinion to inform priors for Bayesian statistical models. Ecology, 90: 265-277. doi:10.1890/07-1886.1

[CEL19] Celhay OJ, Silal SP, Maude RJ *et al.* An interactive application for malaria elimination transmission and costing in the Asia-Pacific [version 2; peer review: 1 approved, 1 approved with reservations]. *Wellcome Open Res* 2019, **4**:61 (https://doi.org/10.12688/wellcomeopenres.14770.2)

[COO06] Cooper B. Poxy models and rash decisions. Proc Natl Acad Sci U S A. 2006 Aug 15;103(33):12221-2. Epub 2006 Aug 7.

[COP19] Robert C. Cope, Joshua V. Ross, Identification of the relative timing of infectiousness and symptom onset for outbreak control (2019) bioRxiv 571547; doi: https://doi.org/10.1101/571547

[COR19] Alice Corbella, Statistical inference in stochastic/deterministic epidemic models to jointly estimate transmission and severity (Doctoral thesis). https://doi.org/10.17863/CAM.41539

[DeA15] De Angelis, A. M. Presanis, P. J. Birrell, G. Scalia Tomba, T. House, "Four key challenges in infectious disease modelling using data from multiple sources," *Epidemics* **10**(2015) 83-87.

[DeM18] De Maio N, Worby CJ, Wilson DJ, Stoesser N. Bayesian reconstruction of transmission within outbreaks using genomic variants. PLoS Comput Biol. 2018 Apr

18;14(4):e1006117. doi: 10.1371/journal.pcbi.1006117.

[DOM18] Doms C, Kramer SC, Shaman J. Assessing the Use of Influenza Forecasts and Epidemiological Modeling in Public Health Decision Making in the United States. Sci Rep. 2018 Aug 17;8(1):12406. doi: 10.1038/s41598-018-30378-w.

[DOU01] Doucet, Arnaud, Freitas, Nando de, Gordon, Neil (Eds.) Sequential Monte Carlo Methods in Practice (2001) Springer-Verlag New York DOI 10.1007/978-1-4757-3437-9

[DOU15] A. Doucet, M. K. Pitt, G. Deligiannidis, R. Kohn, Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator, *Biometrika*, Volume 102, Issue 2, June 2015, Pages 295–313, https://doi.org/10.1093/biomet/asu075

[DRE13] Drescher M., Perera A.H., Johnson C.J., Buse L.J., Drew A., Burgman M.A. Towards rigorous use of expert knowledge in ecological research. Ecosphere. 2013 Jul 13;4(7):1-26.doi:10.1890/ES12-00415.1

[DRO16] Drovandi, C. C. and McCutchan, R. A. (2016), Alive SMC^2: Bayesian model selection for low-count time series models with intractable likelihoods. Biom, 72: 344-353. doi:10.1111/biom.12449

[ERT18] Ertem Z, Raymond D, Meyers LA. Optimal multi-source forecasting of seasonal influenza. PLoS Comput Biol. 2018 Sep 4;14(9):e1006236. doi: 10.1371/journal.pcbi.1006236.

[FIT15] Fithian, W. , Elith, J. , Hastie, T. and Keith, D. A. (2015), Bias correction in species distribution models: pooling survey and collection data for multiple species. Methods Ecol Evol, 6: 424-438. doi:10.1111/2041-210X.12242

[GEL94] Gelfand AE, Dey DK (1994) Bayesian model choice: Asymptotics and exact calculations. Journal of the Royal Statistical Society Series B (Methodological) 56(3):501–514

[GEL04] Gelman, Andrew, Carlin, John B., Stern, Hal S. and Rubin, Donald B.. Bayesian Data Analysis. 2nd ed. : Chapman and Hall/CRC, 2004.

[GEL17] Gelman, A.; Simpson, D.; Betancourt, M. The Prior Can Often Only Be Understood in the Context of the Likelihood. *Entropy* **2017**, *19*, 555.

[GOD83] Godfrey, Keith. "Compartmental models and their application." *Compartmental models and their application*. Academic Press, 1983.

[GOL11] Golightly A, Wilkinson DJ. Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo. Interface Focus. 2011 Dec 6;1(6):807-20. doi: 10.1098/rsfs.2011.0047.

[GOL18] Golightly, A. & Kypraios, T. Efficient SMC^2 schemes for stochastic kinetic models. Stat Comput (2018) 28: 1215. https://doi.org/10.1007/s11222-017-9789-8

[GRE95] Green PJ (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82(4):711–732, (doi:10.1093/biomet/82.4.711)

[GRE15] Green, P.J., Łatuszyński, K., Pereyra, M., Robert, C.P., Bayesian computation: a summary of the current state, and samples backwards and forwards, Stat Comput (2015) 25:835. https://doi.org/10.1007/s11222-015-9574-5

[HAI18] Markus Hainy, David J. Price, Olivier Restif, Christopher Drovandi, Optimal Bayesian design for model discrimination via classification (2018) [arXiv:1809.05301].

[HAY18] Hay JA, Nouvellet P, Donnelly CA, Riley S. Potential inconsistencies in Zika surveillance data and our understanding of risk during pregnancy. PLoS Negl Trop Dis. 2018 Dec 10;12(12):e0006991. doi: 10.1371/journal.pntd.0006991.

[HEA15] H. Heesterbeek, R. M. Anderson, V. Andreasen, S. Bansal, D. De Angelis, C. Dye, K. T. D. Eames, W. J. Edmunds, S. D. W. Frost, S. Funk, T. D. Hollingsworth, T. House, V. Isham, P. Klepac, J. Lessler, J. O. Lloyd-Smith, C. J. E. Metcalf, D. Mollison, L. Pellis, J. R. C. Pulliam, M. G. Roberts, C. Viboud, and Isaac Newton Institute IDD Collaboration, "Modeling infectious disease dynamics in the complex landscape of global health," *Science* **347**:6227 (2015) aaa4339.

[HEM17] Hemming V., Burgman M.A., Hanea A.M., McBridge M.F., Wintle B.C.. A practical guide to structured expert elicitation using the IDEA protocol. Methods Ecol. Evol. 2017 Jul 30; 9:169-180.

[HOU16] House T, Ford A, Lan S, Bilson S, Buckingham-Jeffery E, Girolami M. Bayesian uncertainty quantification for transmissibility of influenza, norovirus and Ebola using information geometry. *Journal of the Royal Society Interface*. 2016 **13**(121):20160279.

[JAC85] Jacquez, John A., and Peter Greif. "Numerical parameter identifiability and estimability: Integrating identifiability, estimability, and optimal sampling design." *Mathematical Biosciences* 77.1-2 (1985): 201-227

[KAD80] Kadane J.B., Dickey J.M., Winkler R.L., Smith W.S., Peters S.C. Interactive elicitation of opinion for a normal linear model. Journal of the American Statistical Association. 1980;75(372):845-854.

[KAD98] Kadane J.B. and Wolfson L.J. Experiences in elicitation. Journal of the Royal Statistical Society Series D (The Statistician). 1998;47(1):3-19.

[KAN15] Kantas, Nikolas; Doucet, Arnaud; Singh, Sumeetpal S.; Maciejowski, Jan; Chopin, Nicolas. On Particle Methods for Parameter Estimation in State-Space Models. Statist. Sci. 30 (2015), no. 3, 328--351. doi:10.1214/14-STS511.

[KAR12] Karlsson, J., Anguelova, M., & Jirstrand, M. (2012). An efficient method for structural identifiability analysis of large dynamic systems. *IFAC Proceedings Volumes*, *45*(16), 941-946.

[KAS95] Kass RE, Raftery AE (1995) Bayes factors. Journal of the American Statistical Association 90(430):773–795, (doi:10.1080/01621459.1995.10476572)

[KUH05] Kuhnert, P. M., Martin, T. G., Mengersen, K. and Possingham, H. P. (2005), Assessing the impacts of grazing levels on bird density in woodland habitat: a Bayesian approach using expert opinion. Environmetrics, 16: 717-747. doi:10.1002/env.732

[KUH10] Kuhnert, P. M., Martin, T. G. and Griffiths, S. P. (2010), A guide to eliciting and using expert knowledge in Bayesian ecological models. Ecology Letters, 13: 900-914. doi:10.1111/j.1461-0248.2010.01477.x

[KYN08] Kynn, Mary. "The 'Heuristics and Biases' Bias in Expert Elicitation." *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 171, no. 1 (2008): 239-64.

[KYP16] Kypraios T, Neal P, Prangle D. A tutorial introduction to Bayesian inference for stochastic epidemic models using Approximate Bayesian Computation. Math Biosci. 2017 May;287:42-53. doi: 10.1016/j.mbs.2016.07.001.

[LAM10] Lambert SB, Faux CE, Grant KA, Williams SH, Bletchly C, Catton MG, Smith DW,
Kelly HA. Influenza surveillance in Australia: we need to do more than count. Med J Aust. 2010 Jul 5;193(1):43-5.

[LOW09] Low Choy S., O'Leary R., Mengersen K. Elicitation by design in ecology: using expert opinion to inform priors for Bayesian statistical models. Ecology. 2009 Jan 1;90(1):265-277.doi:10.1890/07-1886.1

[MAR05] Martin, T. G., Kuhnert, P. M., Mengersen, K. and Possingham, H. P. (2005), The power of expert opinion in ecological models using Bayesian methods: Impact of grazing on birds. Ecological Applications, 15: 266-280. doi:10.1890/03-5400

[MAR12] Martin T.G., Burgman M.A., Fidler F., Kuhnert P.M., Low-Choy S., McBride M., Mengersen K. Eliciting expert knowledge in conservation science. Conservation Biology. 2012 Jan 26;26(1):29-38.doi:10.1111/j.1523-1739.2011.01806.x

[MAT19] MATRIX Institute, 2019. https://www.matrix-inst.org.au/events/influencing-public-health-policy-with-data-informed-mathematical-models-of-infectious-diseases/ (accessed July 2019)

[McK03] MacKay DJC, *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, 2003.

[McK14] McKinley T.J., Ross J.V., Deardon R., Cook A.R., Simulation-based Bayesian inference for epidemic models, Computational Statistics & Data Analysis, 71 (2014) 434-447 https://doi.org/10.1016/j.csda.2012.12.012.

[McL10] McLean E, Pebody RG, Campbell C, Chamberland M, Hawkins C, Nguyen-Van-Tam JS, Oliver I, Smith GE, Ihekweazu C, Bracebridge S, Maguire H, Harris R, Kafatos G, White PJ, Wynne-Evans E, Green J, Myers R, Underwood A, Dallman T, Wreghitt T, Zambon M, Ellis J, Phin N, Smyth B, McMenamin J, Watson JM. Pandemic (H1N1) 2009 influenza in the UK: clinical and epidemiological findings from the first few hundred (FF100) cases. Epidemiol Infect. 2010 Nov;138(11):1531-41. doi:

10.1017/S0950268810001366.

[MES14] Meshkat, N., Kuo, C. E., & DiStefano, J., 3rd (2014). On finding and using identifiable parameter combinations in nonlinear dynamic systems biology models and COMBOS: a novel web implementation. *PloS one*, *9*(10), e110261. doi:10.1371/journal.pone.0110261

[MOR13] Morgan M.G. Use (and abuse) of expert elicitation in support of decision making for public policy. Proc Natl Acad Sci U S A. 2014 May 20;111(20):7176-84. doi: 10.1073/pnas.1319946111. Epub 2014 May 12.

[MOSS17] Moss R, Zarebski A, Dawson P, McCAW JM. Retrospective forecasting of the 2010-2014 Melbourne influenza seasons using multiple surveillance systems. Epidemiol Infect. 2017 Jan;145(1):156-169. Epub 2016 Sep 27. PubMed PMID: 27671159.

[MOSS19] Moss R, Zarebski AE, Dawson P, Franklin LJ, Birrell FA, McCaw JM. Anatomy of a seasonal influenza epidemic forecast. Commun Dis Intell. 2019 Mar 15;43. doi: 10.33321/cdi.2019.43.7

[MOSS19a] Moss R, Zarebski AE, Carlson SJ, McCaw JM. Accounting for Healthcare-Seeking Behaviours and Testing Practices in Real-Time Influenza Forecasts. Trop Med Infect Dis. 2019 Jan 11;4(1). pii: E12. doi: 10.3390/tropicalmed4010012.

[MUS17] Muscatello DJ, Chughtai AA, Heywood A, Gardner LM, Heslop DJ, MacIntyre CR. Translation of Real-Time Infectious Disease Modeling into Routine Public Health Practice. Emerg Infect Dis. 2017 May;23(5). doi: 10.3201/eid2305.161720.

[NGU16] Nguyen, Van Kinh et al. "Analysis of Practical Identifiability of a Viral Infection Model." *PLOS One* vol. 11,12 e0167568. 30 Dec. 2016, doi:10.1371/journal.pone.0167568

[OHA98] O'Hagan A. Eliciting expert beliefs in substantial practical applications. Journal of the Royal Statistical Society, Series D (The Statistician). 1998;47(1):21-35.

[OHA19] O'Hagan T., Oakley J. SHELF: The Sheffield Elicitation Framework. http://www.tonyohagan.co.uk/shelf/. Visited Jul 31 2019.

[ONE99] O'Neill, P. D. and Roberts, G. O. (1999), Bayesian inference for partially observed stochastic epidemics. Journal of the Royal Statistical Society: Series A (Statistics in Society), 162: 121-129. doi:10.1111/1467-985X.00125

[PE01] Pearl J. (2001) Bayesianism and Causality, or, Why I am Only a Half-Bayesian. In: Corfield D., Williamson J. (eds) Foundations of Bayesianism. Applied Logic Series, vol 24. Springer, Dordrecht

[PEL15] Pellis L, Spencer SE, House T. Real-time growth rate for general stochastic SIR epidemics on unclustered networks. *Mathematical Biosciences* 2015 **265**:65-81.

[PEP17] Peppa M, John Edmunds W, Funk S. Disease severity determines health-seeking behaviour amongst individuals with influenza-like illness in an internet-based

cohort. BMC Infect Dis. 2017 Mar 31;17(1):238. doi: 10.1186/s12879-017-2337-5.

[PUD16] Pierre Pudlo, Jean-Michel Marin, Arnaud Estoup, Jean-Marie Cornuet, Mathieu Gautier, Christian P. Robert, Reliable ABC model choice via random forests, *Bioinformatics*, Volume 32, Issue 6, 15 March 2016, Pages 859-866, https://doi.org/10.1093/bioinformatics/btv684

[RAU09] Raue A, Kreutz C, Maiwald T, Bachmann J, Schilling M, Klingmüller U, Timmer J. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. Bioinformatics. 2009 Aug 1;25(15):1923-9. doi: 10.1093/bioinformatics/btp358.

[REI11] Reich NG, Perl TM, Cummings DAT, Lessler J (2011) Visualizing Clinical Evidence: Citation Networks for the Incubation Periods of Respiratory Viral Infections. PLoS ONE 6(4): e19496. https://doi.org/10.1371/journal.pone.0019496

[ROS15] Rosello A, Mossoko M, Flasche S, Van Hoek AJ, Mbala P, Camacho A, Funk S, Kucharski A, Ilunga BK, Edmunds WJ, Piot P, Baguelin M, Tamfum JJ. Ebola virus disease in the Democratic Republic of the Congo, 1976-2014. Elife. 2015 Nov 3;4. pii: e09015. doi: 10.7554/eLife.09015.

[SHU16] Shubin M, Lebedev A, Lyytikäinen O, Auranen K. Revealing the True Incidence of Pandemic A(H1N1)pdm09 Influenza in Finland during the First Two Seasons - An Analysis Based on a Dynamic Transmission Model. PLoS Comput Biol. 2016 Mar 24;12(3):e1004803. doi: 10.1371/journal.pcbi.1004803.

[SHE15] Sherlock, Chris; Thiery, Alexandre H.; Roberts, Gareth O.; Rosenthal, Jeffrey S. On the efficiency of pseudo-marginal random walk Metropolis algorithms. Ann. Statist. 43 (2015), no. 1, 238--275. doi:10.1214/14-AOS1278

[SIE12] Siekmann I, Sneyd J, Crampin EJ. MCMC can detect nonidentifiable models. Biophys J. 2012 Dec 5;103(11):2275-86. doi: 10.1016/j.bpj.2012.10.024.

[SPE75] Carl S. Spetzler and Carl-Axel S. Staël Von Holstein, Exceptional Paper—Probability Encoding in Decision Analysis, Management Science 1975 22:3, 340-358.

[STO17] Stockdale JE, Kypraios T, O'Neill PD. Modelling and Bayesian analysis of the Abakaliki smallpox data. Epidemics. 2017 Jun;19:13-23. doi: 10.1016/j.epidem.2016.11.005.

[THO15] Thomas EG, McCaw JM, Kelly HA, Grant KA, McVernon J. Quantifying differences
in the epidemic curves from three influenza surveillance systems: a nonlinear regression analysis. Epidemiol Infect. 2015 Jan;143(2):427-39. doi: 10.1017/S0950268814000764.

[TOU18] Touloupou P, Alzahrani N, Neal P, Spencer SEF, McKinley TJ (2018) Efficient model comparison techniques for models requiring large scale data augmentation. Bayesian Anal 13(2):437–459, (doi:10.1214/17-BA1057)

[TUN17] Tun STT, von Seidlein L, Pongvongsa T, Mayxay M, Saralamba S, Kyaw SS,

Chanthavilay P, Celhay O, Nguyen TD, Tran TN, Parker DM, Boni MF, Dondorp AM, White LJ. Towards malaria elimination in Savannakhet, Lao PDR: mathematical modelling driven strategy design. Malar J. 2017 Nov 28;16(1):483. doi: 10.1186/s12936-017-2130-3.

[VAJ84] Vajda, Sàndor. "Structural equivalence and exhaustive compartmental modeling." *Mathematical biosciences* 69.1 (1984): 57-75.

[WHY13] J. M. Whyte, "Inferring global a priori identifiability of optical biosensor experiment models." In *2013 IEEE International Conference on Bioinformatics and Biomedicine*, pp. 17-22. IEEE, (2013). doi: 10.1109/BIBM.2013.6732453

[YAN19] Yan, A., Zaloumis, S. G., Simpson, J. A., & McCaw, J. M. (2019). Sequential infection experiments for quantifying innate and adaptive immunity during influenza infection. *PLoS computational biology*, *15*(1), e1006568. doi:10.1371/journal.pcbi.1006568

Figures

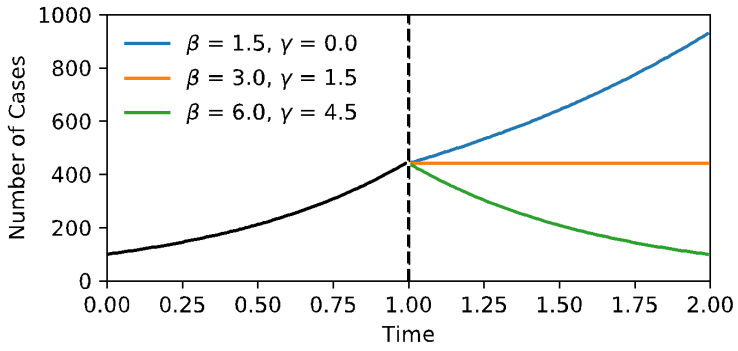

Figure 1. Solutions of Equation (1) giving the number of infectious people, with an intervention at Time = 1 that halves the pre-intervention β for the post-intervention period. Multiple parameter pairs reproduce the pre-intervention data (black line), yet distinct parameter pairs (β, γ) produce differing post-intervention predictions.

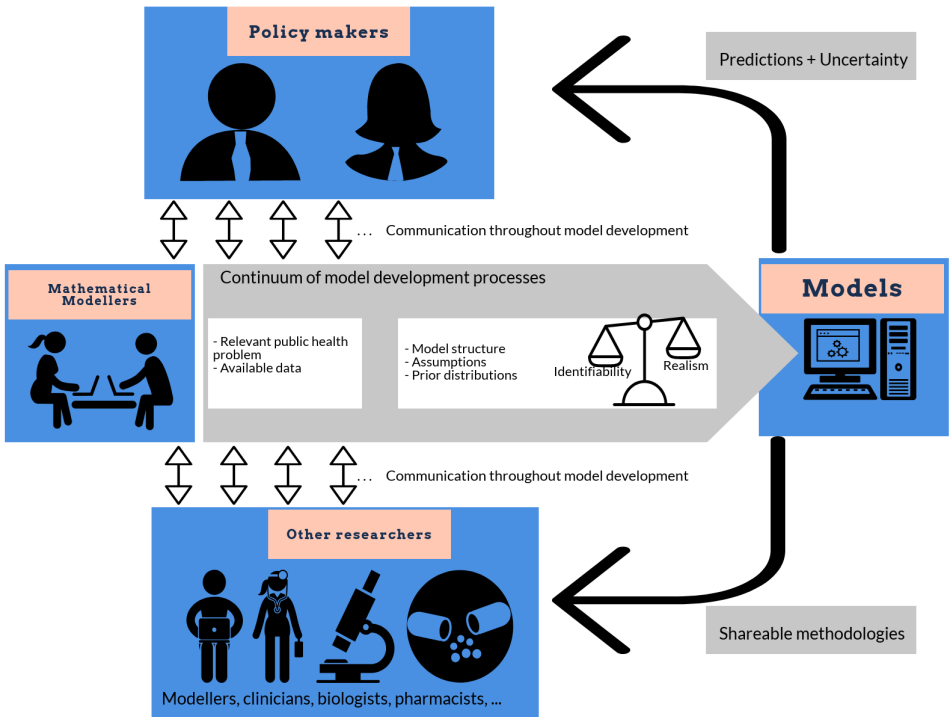

Figure 2. Essential communications between three key groups in the model development and outcome dissemination.

References

Abdessalem, A. B., Dervilis, N., Wagg, D., & Worden, K. (2018). Model selection and parameter estimation in structural dynamics using approximate bayesian computation. *Mechanical Systems and Signal Processing*, *99*, 306–325. 29, 61

Aguas, R., White, L., Snow, R., & Gomes, M. (2008). Prospects for malaria eradication in sub-saharan africa. *PLoS ONE*, *3*(3). (e1767) 72, 119

Alahmadi, A. A., Flegg, J. A., Cochrane, D. G., Drovandi, C. C., & Keith, J. M. (2020). A comparison of approximate versus exact techniques for bayesian parameter inference in nonlinear ordinary differential equation models. *Royal Society Open Science*, *7*(3), 191315. 56

Atkinson, K., Han, W., & Stewart, D. E. (2011). *Numerical solution of ordinary differential equations* (Vol. 108). John Wiley & Sons. 11

Barber, S., Voss, J., Webster, M., et al. (2015). The rate of convergence for approximate bayesian computation. *Electronic Journal of Statistics*, *9*, 80–105. 30, 55

Beaumont, M., Cornuet, J., Marin, J., & Rober, C. (2009). Adaptive approximate bayesian computation. *Biometrika*, *96*(4), 983–990. 29, 61

Beaumont, M. A. (2010). Approximate bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematic*, *41*, 379–406. 25

Bhaumik, P., & Ghosal, S. (2015). Bayesian inference for higher order ordinary differential equation models. *arXiv preprint arXiv:1505.04242*. 81

Bhaumik, P., Ghosal, S., et al. (2015). Bayesian two-step estimation in differential equation models. *Electronic Journal of Statistics*, *9*(2), 3124–3154. 80

Brooks, S., Gelman, A., Jones, G., & Meng, X.-L. (2011). *Handbook of markov chain monte carlo*. CRC press. 16

Brown, P. T. (2019). *Computational bayesian inference using low discrepancy*

sequences (Unpublished doctoral dissertation). The University of Waikato. 88

Buchholz, A., & Chopin, N. (2019). Improving approximate bayesian computation via quasi-monte carlo. *Journal of Computational and Graphical Statistics*, *28*(1), 205–219. 19, 22, 87

Caflisch, R. E. (1998). Monte carlo and quasi-monte carlo methods. *Acta Numerica*, *7*, 1–49. 18, 85

Calderhead, B., Girolami, M., & Lawrence, N. D. (2008). Accelerating bayesian inference over nonlinear differential equations with gaussian processes. *In Advances in Neural Information Processing Systems (NIPS)*, *22*. 1

Calderhead, B., Girolami, M., & Lawrence, N. D. (2009). Accelerating bayesian inference over nonlinear differential equations with gaussian processes. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in neural information processing systems 21.* Curran Associates, Inc. 80

Campbell, D., & Lele, S. (2014). An anova test for parameter estimability using data cloning with application to statistical inference for dynamic systems. *Computational Statistics and Data Analysis*, *70*, 257–267. 99

Campbell, D., & Steele, R. J. (2012). Smooth functional tempering for nonlinear differential equation models. *Statistics and Computing*, *22*(2), 429–443. 80

Celeux, G., Hurn, M., & Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, *95*(451), 957–970. 17, 82

Chi, H., & Mascagni, M. (2007). Efficient generation of parallel quasirandom faure sequences via scrambling. In *International conference on computational science* (pp. 723–730). 22

Chumbley, J. R., Friston, K. J., Fearn, T., & Kiebel, S. J. (2007). A metropolis–hastings algorithm for dynamic causal models. *Neuroimage*, *38*(3), 478–487. 3, 17, 82

Cook, S. R., Gelman, A., & Rubin, D. B. (2006, Sep). Validation of software for Bayesian models using posterior quantiles. *Journal of Computational and*

Graphical Statistics, *15*(3), 675–692. doi: 10.1198/106186006x136976 117, 118

Coulibaly, I., & Lécot, C. (1999). A quasi-randomized runge-kutta method. *Mathematics of Computation of the American Mathematical Society*, *68*(226), 651–659. 82

Craiu, R. V., Rosenthal, J., & Yang, C. (2009). Learn from thy neighbor: Parallel-chain and regional adaptive mcmc. *Journal of the American Statistical Association*, *104*(488), 1454–1466. 17, 82

Dass, S. C., Lee, J., Lee, K., & Park, J. (2016). Laplace based approximate posterior inference for differential equation models. *Statistics and Computing*, 1–20. 81

Del Moral, P., Doucet, A., & Jasra, A. (2006). Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(3), 411–436. 2, 26, 27, 59, 60

Del Moral, P., Doucet, A., & Jasra, A. (2012). An adaptive sequential monte carlo method for approximate bayesian computation. *Statistics and Computing*, *22*(5), 1009–1020. 26, 61

Dick, J., Gantner, R. N., Le Gia, Q. T., & Schwab, C. (2019). Higher order quasi-monte carlo integration for bayesian pde inversion. *Computers and Mathematics with Applications*, *77*(1), 144–172. 82

Dick, J., Kuo, F. Y., & Sloan, I. H. (2013). High-dimensional integration: the quasi-monte carlo way. *Acta Numerica*, *22*, 133–288. 18, 86

Didelot, X., Everitt, R. G., Johansen, A. M., Lawson, D. J., et al. (2011). Likelihood-free estimation of model evidence. *Bayesian Analysis*, *6*(1), 49–76. 29, 61

Dondelinger, F., Rogers, S., & Husmeier, D. (2013). Ode parameter inference using adaptive gradient matching with gaussian processes. In *Sixteenth international conference on artificial intelligence and statistics; aistats.* 1, 80

Drovandi, C. C., & Pettitt, A. N. (2011). Estimation of parameters for macropar-

asite population evolution using approximate bayesian computation. *Biometrics*, *67*(1), 225–233. 26, 61

Feng, Z., & Li, J. (2015). An adaptive independence sampler mcmc algorithm for infinite dimensional bayesian inferences. *arXiv preprint arXiv:1508.03283*. 17, 82

Filippi, S., Barnes, C. P., Cornebise, J., & Stumpf, M. P. (2013). On optimality of kernels for approximate bayesian computation using sequential monte carlo. *Statistical Applications in Genetics and Molecular Biology*, *12*(1), 87–107. 26, 29, 55, 61

Forbes, C., Evans, M., Hastings, N., & Peacock, B. (2011). *Statistical distributions*. John Wiley & Sons. 64

Frazier, D. T., Martin, G. M., Robert, C. P., & Rousseau, J. (2018). Asymptotic properties of approximate bayesian computation. *Biometrika*, *105*(3), 593–607. 23, 24, 30, 55

Frazier, D. T., Robert, C. P., & Rousseau, J. (2017). Model misspecification in abc: Consequences and diagnostics. *arXiv preprint arXiv:1708.01974*. 56

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. Chapman and Hall/CRC. 93

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (Vol. 2). CRC press Boca Raton, FL. 14, 16, 72

Gelman, A., & Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457-472. 15, 67

Ghosh, S., Dasmahapatra, S., & Maharatna, K. (2017). Fast approximate bayesian computation for estimating parameters in differential equations. *Statistics and Computing*, *27*(1), 19–38. 1, 114

Golchi, S., & Campbell, D. A. (2016). Sequentially constrained monte carlo. *Computational Statistics and Data Analysis*, *97*, 98–113. 99

Gu, W., Killeen, G., Mbogo, C., et al. (2003). An individual based model of plasmodium falciparum malaria transmission on the coast of kenya. *Trans-*

actions of the Royal Society of Tropical Medicine and Hygiene, *97*, 43-50. 69

Hardy, G. H. (1905). On double fourier series, and especially those which represent the double zeta-function with real and in commensurable parameters. *Of Pure and Applied Mathematics*, *37*, 53–79. 19, 87

Hastings, W. (1970). Monte carlo sampling methods using markov chain and their applications. *Biometrika*, *57*, 97-109. 14

Hickernell, F. J. (2014). Koksma-hlawka inequality. In *Wiley statsref: Statistics reference online*. American Cancer Society. doi: 10.1002/9781118445112 .stat03070 19, 87

Hoffman, M. D., & Gelman, A. (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, *15*(1), 1593–1623. 16

Jonoska Stojkova, B. (2017). *Bayesian methods for multi-modal posterior topologies* (Unpublished doctoral dissertation). Science: Department of Statistics and Actuarial Science. 99, 100

Joshi, C., Brown, P. T., & Joe, S. (2016, September). Improving Grid Based Bayesian Methods. *ArXiv e-prints*. 81

Joshi, C., & Wilson, S. (2011). Grid based bayesian inference for stochastic differential equation models. *Technical Paper, Trinity College Dublin*. 81

Kim, S., & Li, L. (2012). A switching markov chain monte carlo method for statistical identifiability of nonlinear pharmacokinetics models. *Statistica Sinica*, 1199–1215. 17, 82

Kruschke, J. (2014). *Doing bayesian data analysis: A tutorial with r, jags, and stan*. Academic Press. 17

Lan, S., Streets, J., & Shahbaba, B. (2014). Wormhole hamiltonian monte carlo. In *Twenty-eighth aaai conference on artificial intelligence*. 17, 82

Lécot, C., & Koudiraty, A. (2001). Numerical analysis of runge-kutta quasi-monte carlo methods. *Mathematics and Computers in Simulation, submitted for publication*. 82

Lemieux, C. (2009). Quasi–monte carlo constructions. In *Monte carlo and quasi-monte carlo sampling* (pp. 1–61). Springer. 85

Leobacher, G., & Pillichshammer, F. (2014). *Introduction to quasi-monte carlo integration and applications*. Springer. 21, 85

Levy, D., Hoffman, M. D., & Sohl-Dickstein, J. (2017). Generalizing hamiltonian monte carlo with neural networks. *arXiv preprint arXiv:1711.09268*. 17

Li, W., & Fearnhead, P. (2018). On the asymptotic efficiency of approximate bayesian computation estimators. *Biometrika*, *105*(2), 285–299. 30, 55

Liepe, J., Barnes, C., Cule, E., Erguler, K., Kirk, P., Toni, T., & Stumpf, M. P. (2010). Abc-sysbioÑapproximate bayesian computation in python with gpu support. *Bioinformatics*, *26*(14), 1797–1799. 55

Liepe, J., Kirk, P., Filippi, S., Toni, T., Barnes, C. P., & Stumpf, M. P. (2014). A framework for parameter estimation and model selection from experimental data in systems biology using approximate bayesian computation. *Nature Protocols*, *9*(2), 439. 28, 61

Macdonald, B., Higham, C., & Husmeier, D. (2015). Controversy in mechanistic modemodel with guassian processes. *Journal of Machine Learning Research - Workshop and Conference Proceedings (ICML)*, *37*, 1539-1547. 79, 80

Maire, N., Tediosi, F., Ross, A., & Smith, T. (2006). Predictions of the epidemiological impact of introducing a pre-erythocytic vaccine into the expanded program on immunization in sub-saharan africa. *American Journal of Tropical Medicine and Hygiene*, *75*, 111-118. 69

Marjoram, P., Molitor, J., Plagnol, V., & Tavare, S. (2003). Markov chain monte carlo without likelihoods. *Proceedings of the National Academy of Sciences USA*, *100*, 15324–15328. 24

Massad, E., Coutinho, F. A. B., Burattini, M. N., & Lopez, L. (2004). The eyam plague revisited: did the village isolation change transmission from fleas to pulmonary? *Medical Hypotheses*, *63*(5), 911–915. 99

Maude, R., Pongtavornpinyo, W., et al. (2009). The last man standing is the most resistant: eliminating artemisinin-resistant malaria in cambodia. *Malaria*

Journal, *8*(31). 72, 119

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equations of state by fast computing machines. *The Journal of Chemical Physics*, *21*(1087). 2, 14

Neal, R. M. (1993). *Probabilistic inference using markov chain monte carlo methods*. Department of Computer Science, University of Toronto Toronto, Ontario, Canada. 17, 82

Neal, R. M. (2001). Annealed importance sampling. *Statistics and Computing*, *11*(2), 125–139. 17, 82

Neal, R. M. (2012). *Bayesian learning for neural networks* (Vol. 118). Springer Science & Business Media. 17, 82

Niederreiter, H. (1992). *Random number generation and quasi-monte carlo methods* (Vol. 63). Siam. 20

Owen, A. B. (1998). Scrambling sobol'and niederreiter–xing points. *Journal of complexity*, *14*(4), 466–489. 22

Owen, A. B., et al. (1997). Scrambled net variance for integrals of smooth functions. *The Annals of Statistics*, *25*(4), 1541–1562. 22, 87

Owen, A. B., et al. (2008). Local antithetic sampling with scrambled nets. *The Annals of Statistics*, *36*(5), 2319–2343. 22, 87

Owen, A. B., & Tribble, S. D. (2005). A quasi-monte carlo metropolis algorithm. *Proceedings of the National Academy of Sciences*, *102*(25), 8844–8849. 19, 87

Peng, H., Li, L., Yang, Y., & Wang, C. (2009). Parameter estimation of nonlinear dynamical systems based on integrator theory. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *19*(3), 033130. 80

Picchini, U., & Forman, J. (2014). Accelerating inference for diffusions observed with measurement error and large sample sizes using approximate bayesian computation. *Journal of Statistical Computation and Simulation*. 24, 25

Prangle, D., et al. (2017). Adapting the abc distance function. *Bayesian Analysis*, *12*(1), 289–309. 1, 29, 55, 62

Pritchard, J., Seielstad, M., Perez-Lezaun, A., & Feldman, M. (1999). Population growth of human y chromosomes: a study of y chromosome microsatellites. *Moluecular Biology and Evolution*, *16*, 1791-1798. 2, 22, 55

Qi, X., & Zhao, H. (2010). Asymptotic efficiency and finite-sample properties of the generalized profiling estimation of parameters in ordinary differential equations. *The Annals of Statistics*, 435–481. 79, 80

Raggett, G. (1982). A stochastic model of the eyam plague. *Journal of Applied Statistics*, *9*(2), 212–225. 100

Ramsay, J. O. (1996). Principal differential analysis. *Encyclopedia of Statistical Sciences*. 79

Ramsay, J. O. (2006). *Functional data analysis*. Wiley Online Library. 79

Ramsay, J. O., Hooker, G., Campbell, D., & Cao, J. (2007). Parameter estimation for differential equations: a generalized smoothing approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*(5), 741–796. 79

Roberts, G., Gelman, A., & Gilks, W. (1997). Weak convergence and optimal scaling of random walk metropolis algorithms. *The Annals of Applied Probability*, *7*(1), 110-120. 14

Rudoy, D., & Wolfe, P. J. (2006). Monte carlo methods for multi-modal distributions. In *2006 fortieth asilomar conference on signals, systems and computers* (pp. 2019–2023). 17, 82

Rue, H., Martino, S., & Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series b (Statistical Methodology)*, *71*(2), 319–392. 81

Schillings, C., & Schwab, C. (2016). Scaling limits in computational bayesian inversion. *ESAIM: Mathematical Modelling and Numerical Analysis*, *50*(6), 1825–1856. 82, 83

Sengupta, B., Friston, K. J., & Penny, W. D. (2016). Gradient-based mcmc samplers for dynamic causal modelling. *NeuroImage*, *125*, 1107–1118. 17,

82

Shaw, J. (1988). A quasirandom approach to integration in bayesian statistics. *The Annals of Statistics*, 895–914. 83

Silk, D., Filippi, S., & Stumpf, M. P. (2013). Optimizing threshold-schedules for sequential approximate bayesian computation: applications to molecular systems. *Statistical Applications in Genetics and Molecular Biology*, *12*(5), 603–618. 12, 31, 56

Sisson, S. A., Fan, Y., & Beaumont, M. (2018). *Handbook of approximate bayesian computation*. Chapman and Hall/CRC. 23

Sisson, S. A., Fan, Y., & Tanaka, M. M. (2007). Sequential monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, *104*(6), 1760–1765. 2, 26, 27, 59, 60

Sminchisescu, C., & Welling, M. (2007). Generalized darting monte carlo. In *Artificial intelligence and statistics* (pp. 516–523). 17, 82

Smith, A. F., & Gelfand, A. E. (1992). Bayesian statistics without tears: a sampling–resampling perspective. *The American Statistician*, *46*(2), 84–88. 83

Smith, A. F., Skene, A., Shaw, J., & Naylor, J. (1987). Progress with numerical and graphical methods for practical bayesian statistics. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *36*(2-3), 75–82. 83

Smith, A. F., Skene, A., Shaw, J., Naylor, J., & Dransfield, M. (1985). The implementation of the bayesian paradigm. *Communications in Statistics-Theory and Methods*, *14*(5), 1079–1102. 83

Sobol, I. M. (1976). Uniformly distributed sequences with an additional uniform property. *USSR Computational Mathematics and Mathematical Physics*, *16*(5), 236–242. 21

Stan Development Team. (2019). *RStan: the R interface to Stan.* (R package version 2.19.2) 16, 93

Tavaré, S., Balding, D. J., Griffiths, R. C., & Donnelly, P. (1997). Inferring coalescence times from dna sequence data. *Genetics*, *145*(2), 505–518. 2, 55

Tierney, L., & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, *81*(393), 82–86. 83

Toni, T. (2010). *Approximate bayesian computation for parameter inference and model selection in systems biology* (Unpublished doctoral dissertation). Imperial College London. 30, 57

Toni, T., & Stumpf, M. P. (2009). Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics*, *26*(1), 104–110. 30, 32, 56

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., & Stumpf, M. P. (2009, Feb). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of The Royal Society Interface*, *6*(31), 187–202. doi: 10.1098/rsif.2008.0172 1, 2, 12, 27, 28, 55, 59

Vaart, E., Prangle, D., & Sibly, R. M. (2018). Taking error into account when fitting models using approximate bayesian computation. *Ecological Applications*, *28*(2), 267–274. 1, 12, 13, 25, 30, 56

Vandewoestyne, B., Chi, H., Mascagni, M., & Cools, R. (2007). An empirical investigation of different scrambling methods for faure sequences. In *Sixth imacs seminar on monte carlo methods, reading, uk.* 22

Van Ravenzwaaij, D., Cassey, P., & Brown, S. D. (2018). A simple introduction to markov chain monte–carlo sampling. *Psychonomic Bulletin and Review*, *25*(1), 143–154. 13, 14, 15

Varah, J. (1982). A spline least squares method for numerical parameter estimation in differential equations. *SIAM Journal on Scientific and Statistical Computing*, *3*(1), 28–46. 79

Walter, W. (1998). *Ordinary differential equations* (Vol. 1). Springer. 11

Wang, Y., & Barber, D. (2014). Gaussian processes for bayesian parameter estimation in ordinary differential equations. *Journal of Machine Learning Research - Workshop and Conference Proceedings (ICML)*, *32*, 1485-1493. 1, 17, 80, 82

Wenzel, F., Buchholz, A., & Mandt, S. (2018). *Quasi-monte carlo flows* (Tech. Rep.). EasyChair. 22, 87

White, L., Maude, R., Pongtavornpinyo, W., et al. (2009). The role of simple mathematical model in malaria elimination strategy design. *Malaria Journal*, *8*(212), 1-10. 69, 72, 109, 119

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*(2), 129–141. 30, 56