Likelihood Free Bayesian Inference for Stochastic Kinetic Models

Jamie Robert Owen

Thesis submitted for the degree of
Doctor of Philosophy

School of Mathematics & Statistics
Newcastle University
Newcastle upon Tyne
United Kingdom

March 2017

Acknowledgements

I would first like to extend my thanks to my academic supervisors, Dr Colin Gillespie and Professor Darren Wilkinson for their unwavering support throughout my time as a research student. Their wisdom and guidance was of immeasurable value, as were their words of encouragement.

I would also like to acknowledge that the work that contributed to this thesis, was kindly funded by the Engineering and Physical Sciences Research Council (EPSRC).

Finally on a more personal note, I would like to express my gratitude to my mother and father, for their never–ending backing, love and understanding. To my partner, Dr Claire Keeble, I also want to express my appreciation for her unquestionable loyalty in putting up with me and her love, help and support throughout. In addition, my thanks to notable friends and colleagues without whom, my time at Newcastle University would have been very different, in particular Dr Samuel James and Keith Newman.

Abstract

Stochastic kinetic models are used to describe a variety of biological, physical and chemical phenomena. One particularly interesting application is computational systems biology, where models are useful for contributing to the quantitative understanding of cellular processes through in–silico experimentation that would otherwise be difficult to undertake in a laboratory. Interest lies in statistical inference for the parameters which govern the dynamics of the system. Likelihood based inference is typically problematic, as discrete time transition kernels for models of this type are intractable in all but the most trivial systems. However, exact realisations can be drawn using a stochastic simulation algorithm. Techniques that rely only on the ability to forward simulate from the model, so called likelihood free inference methods, such as particle Markov chain Monte Carlo and approximate Bayesian computation (ABC) can be leveraged to infer system rate parameters. What is not clear however is how each technique behaves as the nature of the problem changes.

This thesis explores the likelihood free methodology applied to stochastic kinetic models in a range of scenarios in order to draw comparisons between the various developments in each. A variety of models and data observation regimes on synthetic data are used to examine the effect of the choice of summary statistics and metrics on the inferred posterior distributions, prevalent questions within the ABC framework. Likelihood free techniques are considered computationally expensive hence it is necessary to consider the relative efficiency of the various approaches. The relative strengths and weaknesses of particle Markov chain Monte Carlo and approximate Bayesian computation are explored and utilised to develop a hybrid technique exploiting the stronger elements of each approach.

The thesis concludes with inference of rate parameters for a logistic growth model applied to observations of a fluorescent protein in different strains of the gram–positive bacterium, Bacillus subtilis.

Contents

List of Figures

Chapter 1

Introduction

Bayesian inference for partially observed Markov processes is the primary focus of this thesis. In trivial examples where discrete time transition densities are available this goal is easily obtained. Unfortunately real data problems of interest can rarely be described by such simple models and therefore more complex systems are required to express them. Transition densities for these processes are often unavailable hence traditional likelihood based inference can not proceed. However, given the ability to simulate data from a given model, it is possible to perform statistical inference without explicit evaluation of a likelihood function.

Two techniques in particular have identified themselves as competitive approaches to parameter inference for Markov process models in which the likelihood function is either analytically unavailable or computationally infeasible. Particle Markov chain Monte Carlo (pMCMC) (Andrieu *et al.*, 2010), a Markov chain Monte Carlo approach which uses a sequential Monte Carlo sampler known as a particle filter for the unbiased estimation of likelihood can be used in settings where observations are made with non–negligible measurement error. Alternatively approximate Bayesian computation (ABC), initially developed by Tavare *et al.* (1997), is a framework based on obtaining posterior samples that lead to model simulations that are close to the observations. An outline of the subsequent chapters is as follows.

1.1 Outline of Thesis

Chapter 2 introduces stochastic kinetic models, a tool by which we can describe the dynamics of a biochemical network. A brief discussion of stochastic kinetics in presented with details of a simulation procedure for drawing exact realisations from models of this type (Gillespie, 1977). Three example models of varying complexity are introduced which form the basis for simulation based experiments throughout the thesis. These are an immigration death process, a simple stochastic kinetic model for which an analytic solutions exists, a Lotka–Volterra (Lotka, 1925; Volterra, 1926) predator prey model that exhibits more complex oscillatory behaviour with no analytic solution available and a Schlögl model (Schlögl, 1972), a test system that exhibits bistability in certain areas of parameter space.

Chapter 3 presents a number of Monte Carlo methods, techniques used for drawing random samples from a target distribution of interest. The methods employed in this thesis are rejection sampling, importance sampling and Markov chain Monte Carlo, each of which can be implemented when the density of interest is known only up to a normalising constant. These methods are often used within the setting of Bayesian inference to sample from distributions whose density may be known up to some multiplicative constant. The techniques also form the basis for a number of algorithms to address problems where this density is not known analytically, so called likelihood free techniques.

Computational techniques for inference in problems where no analytic likelihood function is available are discussed in chapter 4. Each method has some elements which must be chosen by the user and optimality conditions for some of these are discussed. The chapter concludes with a host of numerical examples within the context of stochastic kinetic models like those introduced in chapter 2.

Chapter 5 considers an in depth examination of likelihood free techniques applied to stochastic kinetic models where data are partial observations within a single cell tracked over time. Active research questions within the ABC framework such as the effect on the inferred posterior distribution of user defined choices such as summary statistics are explored as well as comparison of commonly used metric functions. ABC methodology is then compared with a pseudo marginal MCMC sampler approach based on a bootstrap particle filter which yields exact inference from the perspective of computational efficiency within the modelling framework

discussed in this thesis. Numerous numerical examples are considered to explore the questions considered.

Further advantages and drawbacks of each of the approaches is discussed in chapter 6 motivating the development of a hybrid approach which utilises the relative strengths and weaknesses of each framework to create a more efficient posterior sampler. Further simulation experiments are conducted to demonstrate the hybrid algorithm and its use for parameter inference for models of this type.

Where previous chapters focus on single cell data, chapter 7 considers data obtained from a collection of cells evolving over time such as that obtained from flow cytometry. Information at the resolution of a single cell is not available here and hence there is no way to directly connect observations at a given time point with those from a previous time point. The order of magnitude of the dimension of the data is often much larger in this scenario which yields it's own computational demands. A variety of techniques within the likelihood free framework are again explored for this problem using simulated data and numerical examples before considering inference for rate parameters in a logistic growth model for different strains of the gram–positive bacterium Bacillus subtilis from real data.

Finally chapter 8 summarises the conclusions reached in the thesis and considers areas of possible future research.

Chapter 2

Markov processes and stochastic kinetic models

2.1 Introduction

Consider a stochastic process $\{X(t), t \geq t_0\}$ that is continuous in time $t \in \mathbb{R}$. Such a process is called a Markov process if $\{X(t)\}$ has the property that the future state is independent of the past history of the process given only the current state. That is for any given sequence of times, $t_0 < t_1 < \ldots < t_n$,

$$P(X(t_n) \leq x \,|\, X(t_{n-1}) = x_{t_{n-1}}, \ldots, X(t_0) = x_{t_0}) = P(X(t_n) \leq x \,|\, X(t_{n-1}) = x_{t_{n-1}}). \tag{2.1}$$

If we consider a process which can take on values from some countable state space, $\mathcal{S}$, given that at time, t, the process is in state, $x_t \in \mathcal{S}$, the future behaviour can be characterised by a transition kernel

$$p(x_t, t, x_{t+\delta t}, \delta t) \equiv P(X(t + \delta t) = x_{t+\delta t} \,|\, X(t) = x_t). \tag{2.2}$$

Note that a process whose transition kernel does not explicitly depend on t, rather only on the increment δt, is said to be homogeneous with transition kernel denoted $p(x_t, x_{t+\delta t}, \delta t)$.

It has traditionally been the case that the evolution of a biochemical network through time has been modelled with a set of coupled differential equations. These equa-

tions have typically been derived using the law of mass action (see section 2.3) and the concentrations of species present. This approach however makes some critical, questionable assumptions, namely that such a system is both continuous and deterministic. In reality this is generally not the case. Chemical reactions are intrinsically stochastic and occur as a set of individual, discrete events that occur due to collisions on a molecular level. It is relatively clear then that a deterministic approach is not suitable as some of the key features of the system may be lost as discussed in Zheng & Ross (1991) and Murray (2002). Such systems can only effectively be described as stochastic processes. The stochastic effect on this scale can have a significant effect on the outcomes (Finch & Kirkwood, 2000; Wilkinson, 2009). A stochastic Markov process presents a natural approach to modelling such a network.

2.2 Stochastic kinetic models

A stochastic kinetic model is a mechanism by which we can describe probabilistic evolution of a dynamical system made up of a network of reactions. Models of this type are increasingly used to describe the evolution of biological systems (Golightly & Wilkinson, 2005; Proctor *et al.*, 2007; Boys *et al.*, 2008; Wilkinson, 2009). Consider a model which describes some general system. The system will typically consist of a set of species and a set of reactions. When a reaction occurs the amount of one or more of the species is changed. To describe such a system we use the following notation

- Set of u species: $\mathcal{X} = \{\mathcal{X}_1, \ldots, \mathcal{X}_u\}$,
- Set of v reactions: $\mathbf{R} = \{R_1, \ldots, R_v\}$.

We denote the current state of the system at time t, $\mathbf{x}_t$, where

$$\mathbf{x}_t = (x_{1,t}, \ldots, x_{u,t})^T,$$

and $x_{j,t}$ represents the number of 'molecules' of $\mathcal{X}_j$ at time t. Each reaction typically involves some combination of species (reactants) and results in a different combination of species (products). A general reaction is represented as

$$R_i : p_{i,q}\mathcal{X}_1 + \ldots + p_{i,u}\mathcal{X}_u \rightarrow q_{i,1}\mathcal{X}_1 + \ldots + q_{i,u}\mathcal{X}_u, \quad i = 1, \ldots, v.$$

Here the $p_{i,j}$ represents the number of 'molecules' of species j required for reaction i, and the $q_{i,j}$ the number produced. We can represent the model more succinctly as

$$P\mathcal{X} \to Q\mathcal{X},$$

where P (pre-reaction) and Q (post-reaction) are $v \times u$ dimensional matrices. When a reaction event of type i occurs, the number of molecules of type $\mathcal{X}_j$ will increase by q_{ij} decrease by p_{ij}. This gives us an overall change of $a_{ij} = q_{ij} - p_{ij}$ or

$$A = Q - P.$$

The key features of this network can then be represented by this net-effect matrix, although commonly one would work with the stoichiometry matrix S defined as

$$S = (Q - P)^T = A^T.$$

See Wilkinson (2011) for a more in depth discussion.

2.3 Mass action stochastic kinetics

Let us first consider a reaction of the form

$$\mathcal{X}_1 + \mathcal{X}_2 \to \mathcal{X}_3. \tag{2.3}$$

This reaction will take place only when the two molecules $\mathcal{X}_1$, $\mathcal{X}_2$ combine or collide to form $\mathcal{X}_3$. These molecules are continuously moving around randomly in space. For these reactions we assume the law of mass action kinetics, given the number of molecules of $\mathcal{X}_1$, x_1 and similarly the number of molecules of $\mathcal{X}_2$. Then the hazard or propensity of this reaction taking place would be directly proportional to $x_1 x_2$. This is because there are $x_1 x_2$ distinct pairs of molecules that are able to collide. (Gillespie, 1992) gives a rigorous derivation, the thrust of the argument being that under fairly weak assumption regarding the container and it's contents, namely that is is well stirred and in thermal equilibrium, it can be demonstrated that the collision hazard is constant. If the molecules are uniformly distributed (well mixed) throughout the volume and if this distribution does not depend on time then the chance that two molecules are close enough to react is also independent of time.

Each reaction in the set has an associated stochastic rate constant, θ_i. It is this rate, along with the current state of the system, $\mathbf{x}_t$, that we can use to define the hazard function. The hazard function, $h_i(\mathbf{x}_t, \theta_i)$, represents the propensity of reaction i occurring. When considering these hazard functions it is important to think about the order of a reaction.

2.3.1 Reaction orders

The hazard $h_i(\mathbf{x}_t, \theta_i)$ of reaction R_i occurring is determined by the order of the reaction.

Zeroth-order reactions

A zeroth-order reaction takes the form

$$\emptyset \xrightarrow{\theta_i} \mathcal{X}_j$$

and its hazard function is given by

$$h_i(\mathbf{x}_t, \theta_i) = \theta_i.$$

Whilst in reality it may appear that the creation of something from nothing is illogical, in terms of modelling our system it can be useful to model a constant rate of production (or influx from another source) of a species.

First-order reactions

A first-order reaction takes the form

$$\mathcal{X}_j \xrightarrow{\theta_i} ?$$

with associated hazard function given by

$$h_i(\mathbf{x}_t, \theta_i) = \theta_i x_j.$$

Here it is only the presence of one of the species which determines the hazard of a reaction.

Second-order reactions

A second-order reaction can take the form of either

$$\mathcal{X}_j + \mathcal{X}_k \xrightarrow{\theta_i} ? \text{ where } \quad h_i(\mathbf{x}_t, \theta_i) = \theta_i x_{t,j} x_{t,k},$$

or

$$2\mathcal{X}_j \xrightarrow{\theta_i} ? \text{ where } \quad h_i(\mathbf{x}_t, \theta_i) = \theta_i \frac{x_{t,j}(x_{t,j} - 1)}{2}.$$

Higher order reactions

More generally:

$$h_i(\mathbf{x}_t, \theta_i) = \theta_i \prod_{j=1}^{u} \begin{pmatrix} x_{t,j} \\ p_{i,j} \end{pmatrix}.$$

Time evolution of the system

Since, given $\mathbf{x}_t$ and θ_i, we know that the hazard of reaction R_i is $h_i(\mathbf{x}_t, \theta i)$, then the hazard of some reaction occurring is $h_0(\mathbf{x}_t, \Theta)$, where

$$h_0(\mathbf{x}_t, \Theta) = \sum_i h_i(\mathbf{x}_t, \theta_i).$$

The time evolution of such a system can be regarded as a stochastic process, the time to the next reaction event is given by

$$\delta t \sim \text{Exp}(h_0(\mathbf{x}_t, \Theta)),$$

and the type of the reaction will be stochastically determined with probabilities proportional to the respective hazards.

2.3.2 The chemical master equation

In the stochastic kinetics literature there is often reference to the chemical master equation. This is in fact the Kolmogorov forward equations,

$$\frac{d}{dt}p(x_0, x_t, t) = \sum_{i=1}^{v} h_i(x_t - S^{[i]}, \theta_i)p(x_0, x_t - S^{[i]}, t) - h_i(x_t, \theta_i)p(x_0, x_t, t) \qquad (2.4)$$

This is a sum over all possible R_i which in the case of the stochastic kinetic models is finite. However, for a (countably) infinite state space, it corresponds to an infinite system of ordinary differential equations.

An extensive discussion of the chemical master equation can be found in Van Kampen (1992). There are few cases in which the chemical master equation can be solved explicitly, an examination of which can be found in McQuarrie (1967). In the case of mass action kinetics as described in section 2.3 an analytic solution to the chemical master equation can only be found for systems that contain equations only of order zero and one (Jahnke & Huisinga, 2007).

2.4 Model simulation

Exact analytical solutions to a stochastic kinetic model are tractable for only the most trivial of systems. The reaction networks therefore are typically examined using discrete simulation algorithms. We have already seen that the time to the next reaction event in the system can be simulated as an exponential random variable with rate equivalent to the total hazard. When a reaction event does occur, it will be of a random type. The probability that this reaction event is of type i is proportional to its hazard $h_i(\mathbf{x}_t, \theta_i)$. Gillespie's direct method (Gillespie, 1977) was developed in the context of chemical kinetics and can be used to provide exact samples from the solution to the stochastic process. The direct method is described in Algorithm 1.

Gillespie's direct method is often thought of as a computationally expensive simulation algorithm. It is of note that there are a number of faster, approximate simulators. Such algorithms typically relax the restrictions imposed by the discrete state space, such as the chemical Langevin equation (CLE), see Gillespie (2000) for

Algorithm 1 Gillespie's direct method (Gillespie, 1977)

1. Initialise the system at $t = 0$, with rate constants, $\theta_1, \ldots, \theta_v$ and initial state $\mathbf{x}_0$.
2. Calculate the hazards $h_i(\mathbf{x}_t, \theta_i)$ for each $i = 1, 2, \ldots, u$ according to the current state of the system, $\mathbf{x}_t$.
3. Calculate $h_0(\mathbf{x}_t, \Theta) = \sum_{i=1}^{v} h_i(\mathbf{x}_t, \theta_i)$. If $h_0 = 0$, set $t = T_{max}$ and go to step 8.
4. Simulate the time until the next reaction event as $\delta t \sim Exp(h_0(\mathbf{x}_t, \Theta))$.
5. Update the time: $t := t + \delta t$.
6. Simulate the reaction index, j, from $i = 1, 2, \ldots, v$ with probabilities $\frac{h_i(\mathbf{x}_t, \theta_i)}{h_0(\mathbf{x}_t, \Theta)}$.
7. Update the state, $\mathbf{x}_t$, according to the reaction index, j,

$$\mathbf{x}_{t+\delta t} := \mathbf{x}_t + S^{[j]},$$

   where $S^{[j]}$ is the j^{th} column of the stoichiometry matrix.
8. Output $\mathbf{x}_t$ and t.
9. If $t < T_{max}$ return to step 2.

details or treat the hazard function as constant over short time intervals, updating all reactions within the time window before updating the propensity functions, see the tau–leap algorithm in Gillespie (2001). Caution must be taken if such approximate algorithms are used. Gillespie (2016) showed that they are not appropriate in all cases and ensuring that approximations are good over the full parameter space can be problematic. For a more comprehensive introduction to stochastic kinetic modelling, see Wilkinson (2011).

2.5 Examples

2.5.1 The Immigration–Death process

The immigration–death (ID) process is one of the most basic stochastic kinetic models. It provides an excellent basis for investigation into models of this type since it is one of the few that has an analytic solution. This allows comparison of any inference schemes that could be applied to a general system with a reference true solution. The model is described by the following pair of reactions concerning a single species

$$R_1 : \emptyset \xrightarrow{\theta_1} \mathcal{X}_1 \text{ and } R_2 : \mathcal{X} \xrightarrow{\theta_2} \emptyset.$$

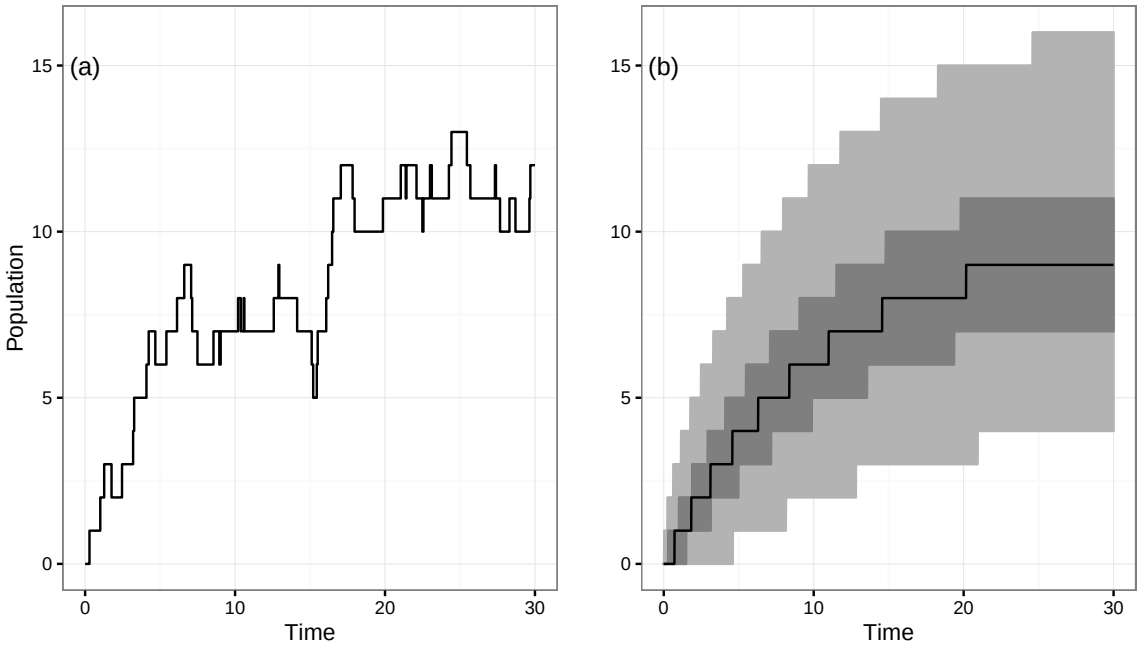


Figure 2.1: (a) a single realisation using Gillespie's direct method (Algorithm 1). (b), median, 50% and 95% point-wise intervals of the true solution of the immigration–death model. For each the initial marking of the system is $\mathbf{x}_0 = 0$ with rate parameters $\Theta = (1.0, 0.1)$.

R_1 is the zeroth order immigration event, individuals arrive into the system from some external source, with associated stochastic rate constant θ_1. Likewise R_2, a first order reaction, represents the death of an individual with constant rate θ_2. The reaction network can be summarised by its stoichiometry matrix

$$S = \begin{pmatrix} 1 & -1 \end{pmatrix}. \tag{2.5}$$

Under the assumption of mass action kinetics we can calculate the hazard function for each reaction,

$$h_1(\mathbf{x}_t, \theta_1) = \theta_1 \text{ and } h_2(\mathbf{x}_t, \theta_2) = \theta_2 x_{t,1}. \tag{2.6}$$

Figure 2.1 (a) shows a single realisation of the immigration–death process using Gillespie's direct method (Algorithm 1) with initial state $x_0 = 0$ and reaction rate parameters $\Theta = (\theta_1 = 1.0, \theta_2 = 0.1)$. Figure 2.1 shows the median, 50% and 95% point-wise intervals of the Poisson distribution representing the true solution given the same initial conditions and rate parameters.

2.5.2 Analytic solution to the immigration death process

Since the immigration death process has reactions that have order zero and one, the solution to the system is tractable. To simplify notation let $p_x(t)$ be the probability that the population of the system at time t is x with conditional notation on x_0 dropped for brevity. Substituting $h_1(\mathbf{x}_t, \theta_1)$, $h_2(\mathbf{x}_t, \theta_2)$ and S into the chemical master equation yields

$$\begin{aligned}\frac{d}{dt}p_x(t) &= \theta_1 \left[p_{x-1}(t) - p_x(t)\right] + \theta_2 \left[(x_t+1)p_{x+1}(t) - x_t p_x(t)\right] \\ &= \theta_1 p_{x-1}(t) - (\theta_1 + \theta_2 x_t)p_x(t) + \theta_2(x_t+1)p_{x+1}(t), \end{aligned} \tag{2.7}$$

for $x = 0, 1, \ldots$.

At $t = 0$ the population of the system is x_0 hence

$$p_{x_0}(0) = 1 \text{ and } p_y(0) = 0 \text{ for } y \neq x_0.$$

Equation 2.7 can be solved using a variety of techniques, one such example being the Lagrange equation method. First introduce the probability generating function (PGF)

$$G(z;t) = \sum_{x=0}^{\infty} z^x p_x(t), \quad |z| \leq 1 \tag{2.8}$$

where probabilities are recovered via

$$p_n(t) = \frac{G^{(n)}(0;t)}{n!} \tag{2.9}$$

where

$$G^{(n)}(0;t) = \left.\frac{d^n G(z;t)}{dz^n}\right|_{z=0}$$

is the n^{th} derivative of the generating function $G(\cdot)$ with respect to z evaluated at $z = 0$. It is also useful to note that

$$G(0;t) = p_0(t) \quad \text{and} \quad G(z;0) = \sum_{x=0}^{\infty} z^x p_x(0) = z^{x_0}. \tag{2.10}$$

On multiplying equation 2.7 by z^x and summing over x, first noting that

$$\frac{\partial G(z;t)}{\partial z} = \sum_{x=0}^{\infty}(x+1)z^x p_{x+1}(t) \text{ and } \frac{\partial G(z;t)}{\partial t} = \sum_{x=0}^{\infty} z^x \frac{dp_x(t)}{dt}, \tag{2.11}$$

we obtain

$$\sum_{x=0}^{\infty} z^x \frac{dp_x}{dt} = \sum_{x=0}^{\infty} \theta_2(x+1)z^x p_{x+1} - (\theta_1 + \theta_2 x)z^x p_x + \theta_1 z^x p_{x-1}$$

$$\frac{\partial G}{\partial t} = \theta_2 \frac{\partial G}{\partial z} - \theta_1 G - \theta_2 z \frac{\partial G}{\partial z} + \theta_1 z G$$

$$\frac{\partial G}{\partial t} + \theta_2(z-1)\frac{\partial G}{\partial z} = \theta_1(z-1)G. \tag{2.12}$$

By applying the standard Lagrange procedure we construct the characteristic equations

$$\frac{dt}{1} = \frac{dz}{\theta_2(z-1)} = \frac{dG}{\theta_1(z-1)G}.$$

From the first pair of characteristic equations we have

$$\int \theta_2 dt = \int \frac{dz}{z-1},$$

which gives solution

$$C = (z-1)e^{-\theta_2 t}$$

for some arbitrary constant C. Using the second pair of equations

$$\int \frac{dG}{G} = \int \frac{\theta_1 dz}{\theta_2}$$

yielding the solution

$$C' = G \exp\left(\frac{-\theta_1 z}{\theta_2}\right),$$

constant C'. The general solution may then be written as $C' = f(C)$ for some arbitrary function $f(\cdot)$,

$$G(z;t)\exp\left(\frac{-\theta_1 z}{\theta_2}\right) = f\left((z-1)e^{-\theta_2 t}\right).$$

Using the initial condition, $t = 0$ and expression 2.10 gives

$$f(z-1) = z^{x_0} \exp\left(\frac{-\theta_1 z}{\theta_2}\right) \Rightarrow$$
$$f(z) = (1+z)^{x_0} \exp\left(\frac{-\theta_1(1+z)}{\theta_2}\right),$$

which on replacing z for $(z-1)e^{-\theta_2 t}$ gives

$$G(z;t) = \left(1 + (z-1)e^{-\theta_2 t}\right)^{x_0} \exp\left(\frac{\theta_1(z-1)(1-e^{-\theta_2 t})}{\theta_2}\right). \tag{2.13}$$

Expression 2.13 then is the probability generating function for the solution to the immigration death model. It is worth noting that this generating function is the product of two factors. The first, a generating function of a binomial distribution with parameters $n = x_0$ and $p = e^{-\theta_2 t}$ whilst the second is the generating function for a Poisson distribution with rate $\lambda = \theta_1(1 - e^{\theta_2 t})/\theta_2$. The two components represent the number of individuals that were present in the system at time $t = 0$ which are still alive at time t, and those that have arrived into the population after $t = 0$ and still present at time t.

One can recover the equilibrium solution by allowing $t \to \infty$,

$$G(z;\infty) = \exp\left(\frac{\theta_1(z-1)}{\theta_2}\right), \tag{2.14}$$

which is the generating function of a Poisson distribution with rate parameter θ_1/θ_2. This limiting distribution is independent of the initial population size x_0.

Since we have the PGF of the solution (equation 2.13), we can recover the probability distribution function. Denote $p_x(t \,|\, x_0; t_0)$ as the probability that the population is x at time t given it was of size x_0 at time t_0 and for convenience let

$$\rho = \frac{\theta_1(1 - e^{-\theta_2 t})}{\theta_2}.$$

From equations 2.9 and 2.13 it follows that

$$p_x(t \,|\, x_0; t_0) = \sum_{i=0}^{x} \binom{x_0}{i} e^{-\theta_2 t i} \left(1 - e^{-\theta_2 t}\right)^{x_0 - i} \frac{\rho^x e^{-\rho}}{x!}. \tag{2.15}$$

Under initial conditions $x_0 = 0$, that is there are no individuals present within the system at time $t = 0$,

$$p_x(t \mid 0; 0) = \frac{\rho^x e^{-\rho}}{x!}. \tag{2.16}$$

I.e the population at time t is distributed as a Poisson random variable with rate $\theta_1(1 - e^{\theta_2 t})/\theta_2$. For a more in depth discussion of this derivation see Gillespie (2003).

Whilst the immigration–death process is a relatively simple model within this framework it does present it's own challenges when it comes to rate parameter inference. Since the limiting distribution of the process as $t \to \infty$ is a Poisson distribution with rate θ_1/θ_2, that is it is stationary and independent of initial conditions, if one were to observe only data from the stationary period then it would not be possible to identify both parameters. In this extreme case only the ratio of the parameters could be inferred. This means that in the posterior distribution the two rate parameters are correlated. The larger the proportion of observed data that comes from this limiting distribution, alternatively the quicker our observed data appears to become stationary or the longer we observe our time series, the stronger the correlation structure in the posterior distribution. It should be acknowledged that inference for highly correlated posterior distributions is challenging.

Simulation of the ID process

The direct method (algorithm 1 can be used to sample exact realisation of any stochastic kinetic model. However one of the by products of having an analytic solution for the immigration death process is that we can sample realisations from it more efficiently. Since the PGF function in equation 2.13 is a product of two generating function we can use the general result that given random variables X and Y with associated probability generating functions G_X and G_Y the generating function of the sum $X + Y$ is

$$G_{X+Y} = G_X G_Y.$$

This implies that we can draw samples from $p_x(t \mid x_0, t_0)$ as the sum of a sample from a binomial distribution with $n = x_0$ and $p = e^{-\theta_2(t-t_0)}$ and a sample from a Poisson distribution with rate $\lambda = \theta_1(1 - e^{-\theta_2(t-t_0)})/\theta_2$.

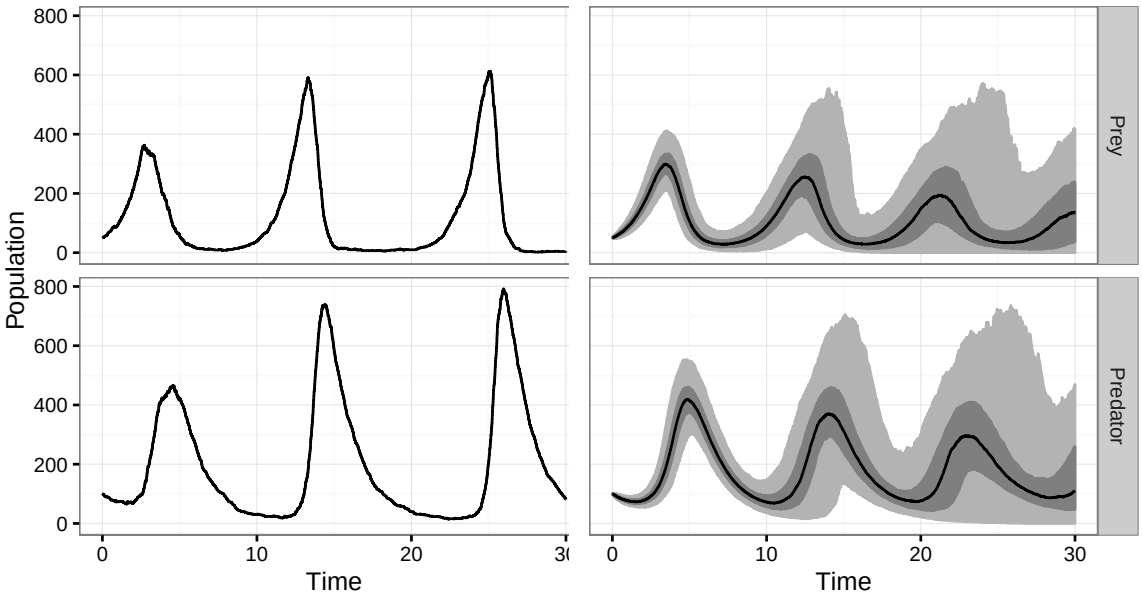


Figure 2.2: Left, a single realisation, right, median, 50% and 95% intervals of 1000 realisations of the Lotka–Volterra model using Gillespie's direct method (Algorithm 1). For each the initial marking of the system is $\mathbf{x}_0 = (50, 100)$ with rate parameters $\Theta = (1.0, 0.005, 0.6)$.

2.5.3 Lotka–Volterra

The Lotka–Volterra model, developed by Lotka (1925) and Volterra (1926) independently is a simple stochastic kinetic model of predator–prey interactions. The model is characterised by a set of three reactions between two species, $\mathcal{X}_1$ and $\mathcal{X}_2$, prey and predators respectively.

$$\begin{aligned} R_1 : &\quad \mathcal{X}_1 &\to &\quad 2\mathcal{X}_1 \\ R_2 : &\quad \mathcal{X}_1 + \mathcal{X}_2 &\to &\quad 2\mathcal{X}_2 \\ R_3 : &\quad \mathcal{X}_2 &\to &\quad \emptyset \end{aligned}$$

Reactions R_1, R_2 and R_3 can be thought of as prey reproduction, predator-prey interaction resulting in prey death and predator birth, and predator death respectively. This is not a true biochemical network, however it is true that $\mathcal{X}_1$ and $\mathcal{X}_2$ could equally be thought of as chemical species. Whilst this is a trivial stochastic kinetic model, it does provide a problem of interest, as it highlights many of the difficulties that arise in more complex systems, demonstrating the sort of auto–regulatory behaviour that is typical of many other biochemical networks.

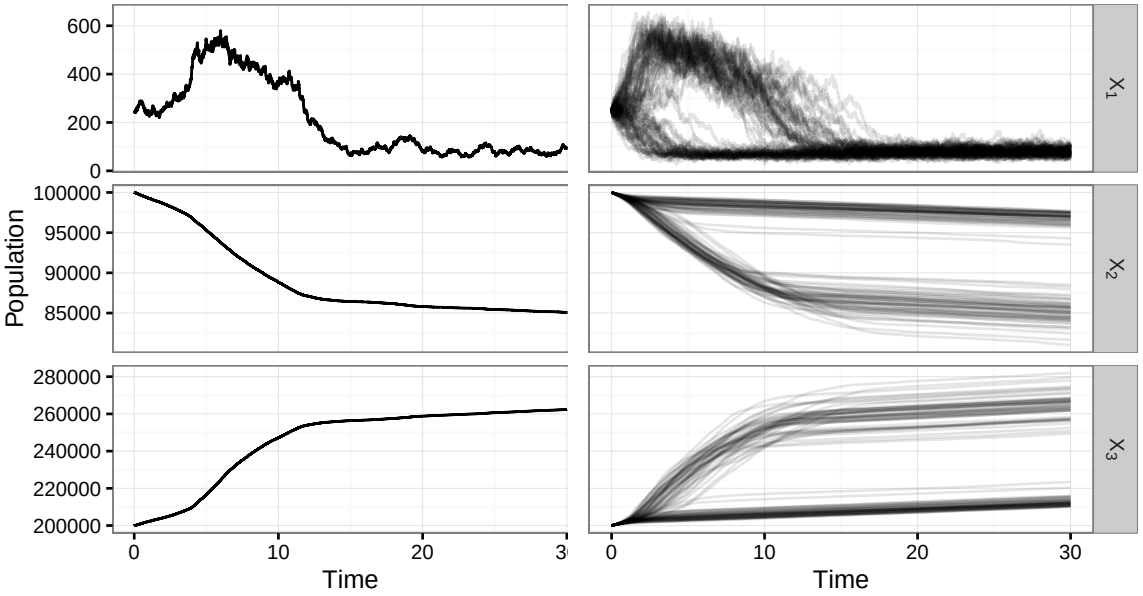


Figure 2.3: Left, a single realisation, right, 100 realisations of the Schlögl using Gillespie's direct method (Algorithm 1). For each the initial marking of the system is $\mathbf{x}_0 = (250, 1 \times 10^5, 2 \times 10^5)$ with rate parameters $\Theta = (3 \times 10^{-7}, 1 \times 10^{-4}, 0.000773, 3.276)$. Given this set of initial conditions the bimodal distribution of the state is clearly visible.

If we now assume that reaction i has stochastic rate constant θ_i and that the system evolves as a Markov process with state $\mathbf{x}_t = (x_1, x_2) = \mathbf{x}$ at time t. Using the law of mass action kinetics the hazards of each reaction can be derived as:

$$h_1(\mathbf{x}, \theta_1) = \theta_1 x_1, \;\; h_2(\mathbf{x}, \theta_2) = \theta_2 x_1 x_2 \;\text{ and }\; h_3(\mathbf{x}, \theta_3) = \theta_3 x_2.$$

Figure 2.2 shows both a single realisation of this model where $\Theta = (1.0, 0.005, 0.6)$ using the Gillespie algorithm, left, and the median, 50% and 95% intervals of a number of realisations, right. Unlike the immigration–process there is no analytically tractable solution for the system dynamics so simulation is necessary to acquire samples.

2.5.4 Schlögl system

The Schlögl model (Schlögl, 1972) is a well known test case in the stochastic kinetics literature, known for the fact that it exhibits bimodal stability in the states for certain regions of parameter space. The set of reactions that make up the system

are

$$\begin{array}{lrcl} R_1: & 2\mathcal{X}_1 + \mathcal{X}_2 & \rightarrow & 3\mathcal{X}_1 \\ R_2: & 3\mathcal{X}_1 & \rightarrow & 2\mathcal{X}_1 + \mathcal{X}_2 \\ R_3: & \mathcal{X}_3 & \rightarrow & \mathcal{X}_1 \\ R_4: & \mathcal{X}_1 & \rightarrow & \mathcal{X}_3. \end{array} \tag{2.17}$$

As with the other models considered we can characterise the reaction network via the stoichiometry matrix

$$S = \begin{pmatrix} 1 & -1 & 1 & -1 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix} \tag{2.18}$$

and hazard function

$$h(\mathbf{x}, \Theta) = \left(\frac{\theta_1 x_1 (x_1 - 1) x_2}{2}, \frac{\theta_2 x_1 (x_1 - 1)(x_1 - 2)}{6}, \theta_3 x_3, \theta_4 x_1 \right). \tag{2.19}$$

A single draw over a 30 unit time course using Gillespie's direct method is shown in figure 2.3(left). Figure 2.3(right) shows 100 such draws, highlighting the bimodal stability in the evolution of the state of the system given $\mathbf{x}_0 = (250, 1 \times 10^5, 2 \times 10^5)$ with rate parameters $\Theta = (3 \times 10^{-7}, 1 \times 10^{-4}, 0.000773, 3.276)$. The system also demonstrates that a deterministic solution is clearly inappropriate since the solution to an ordinary differential equation is unable to capture multiple modes

As is the case with the Lotka–Volterra model introduced in section 2.5.3 the model has no analytic solution, hence model simulation is necessary to proceed with inference.

In order to be able to perform analysis on a biochemical network model it is necessary to obtain each of the parameters (typically reaction rates) for the network (Kitano *et al.*, 2001). The problem of interest here then is the inference of the systems parameters given the observation of some time course data, a process sometimes known as reverse engineering (Bower & Bolouri, 2004). Due to the fact that the chemical master equation, and hence the likelihood function, is tractable for only a handful of simple models inference for models of this type presents a difficult and interesting problem.

Chapter 3

Monte Carlo methods

Monte Carlo methods aim to approximate analytically intractable integrals. Consider a random variable $\Theta = (\theta_1, \ldots \theta_p)$ with density $\pi(\Theta)$ and suppose that we are interested in evaluating the expectation of some function of the random variable, $\mathbf{E}(g(\Theta))$. This is an integration problem,

$$\mathbf{E}(g(\Theta)) = \int_{\Theta} g(\Theta)\pi(\Theta)d\Theta, \tag{3.1}$$

which could well be intractable. A Monte Carlo method attempts to provide an approximation to such an integral by using a sample

$$\{\Theta^{(1)}, \ldots, \Theta^{(n)}\}$$

from the distribution $\pi(\Theta)$ in order to estimate $\mathbf{E}(g(\Theta))$. Usually this is done by replacing the integral in 3.1 with the sample mean,

$$\mathbf{E}(g(\Theta)) \approx \hat{g}(\Theta) = \frac{1}{n}\sum_{i=1}^{n} g(\Theta^{(i)}), \tag{3.2}$$

as this is an unbiased estimator. We can see that this is an unbiased estimator since

$$\begin{aligned} E(\hat{g}(\Theta)) &= \frac{1}{n}\sum_{i=1}^{n} E(g(\Theta^{(i)})) \\ &= E(g(\Theta)). \end{aligned} \tag{3.3}$$

Algorithm 2 Envelope method rejection sampler

Say we wish to draw N samples $\{\Theta^{(1)}, \Theta^{(2)}, \dots \Theta^{(N)}\}$ such that $\Theta^{(i)} \sim \pi(\Theta) \, \forall \, i$.

1. Initialise with $i = 1$.
2. Draw $\Theta^* \sim q(\Theta)$ and $u \sim U(0, Cq(\Theta^*))$.
3. If $u < \pi(\Theta^*)$ set $\Theta^{(i)} = \Theta^*$ else return to 2.
4. If $i < N$ set $i := i + 1$ and return to 2.

As $n \to \infty$, $\hat{g}(\Theta) \to g(\Theta)$ by the law of large numbers and so the estimator is consistent. We can also use this sample to estimate the distribution function of Θ. Replacing $g(\Theta)$ in equation 3.2 with $\mathcal{I}_{\mathcal{A}}(\Theta)$ where $\mathcal{I}_{\mathcal{A}}(\Theta)$, for any subset $\mathcal{A}$, is the indicator function taking the value 1 if $\Theta \in \mathcal{A}$ and 0 otherwise gives $\mathbf{E}(\mathcal{I}_{\mathcal{A}}(\Theta))$ which is an estimate of $P(\Theta \in \mathcal{A})$.

3.1 Bayes theorem

Bayes theorem (equation 3.4) describes the probability of an event conditional on knowledge of some other conditions related to that event.

$$\pi(A \,|\, B) = \frac{\pi(B \,|\, A)\pi(A)}{\pi(B)} \tag{3.4}$$

It is often used within the Bayesian inference framework, introduced in chapter 4, that allows additional evidence to be used to supplement prior beliefs. We introduce Bayes theorem here as it provides the basis for some of the Monte Carlo techniques described in this chapter.

3.2 Rejection sampling

A rejection sampler is a simple method by which we can sample from some target distribution. Suppose interest lies in some arbitrary distribution $\pi(\Theta)$ that is difficult to sample from. Further, let there be some proposal distribution $q(\Theta)$ from which we can easily draw samples where the support of $\pi(\cdot)$ is contained within the support of $q(\cdot)$. Given C such that $Cq(\Theta) > \pi(\Theta)$, $\forall \Theta$ we can draw samples from $\pi(\Theta)$ using proposals from $q(\Theta)$ that we keep with probability $\pi(\Theta)/Cq(\Theta)$.

Algorithm 2 gives a formal description of this rejection sampler, often called the envelope method, that could be used to generate samples $\Theta^{(i)} \sim \pi(\Theta)$. Whilst this is guaranteed to give samples from the desired distribution, in practice performance is dictated by how close $Cq(\Theta)$, often referred to as the envelope function, is to $\pi(\Theta)$, across the support of the target distribution. In fact the acceptance probability can be computed directly as

$$\begin{aligned} P(U < \pi(\Theta)) &= \int_{\Theta} P(U < \pi(\Theta) \,|\, \Theta = \theta) q(\theta) d\theta \\ &= \int_{\Theta} \frac{\pi(\theta)}{Cq(\theta)} q(\theta) d\theta \\ &= \int_{\Theta} \frac{\pi(\theta)}{C} d\theta \\ &= \frac{1}{C}. \end{aligned}$$

Rejection sampling can be used to sample from the area under any curve, regardless of whether or not its area integrates to 1. Scaling by a constant has no effect on the sampled coordinates and hence can be used when a distribution is known only up to a normalising constant.

3.3 Importance sampling

Importance sampling is a technique that allows us to calculate the expectation of a function, say $\mathbf{E}(g(\Theta))$, with respect to a distribution $\pi(\Theta)$, using samples from some other distribution. We begin with a suitable proposal distribution $q(\Theta)$ from which samples are easily obtained. Suitable here also implies that the support of $q(\cdot)$ must contain the support of $\pi(\cdot)$. Then, given a sample of N points, $\{\Theta^{(1)}, \ldots, \Theta^{(N)}\} \sim q(\Theta)$, we define a set of importance weights

$$w^{(i)} = \frac{\pi(\Theta^{(i)})}{q(\Theta^{(i)})}, \qquad 1, \ldots, N. \tag{3.5}$$

We obtain the approximation

$$\begin{aligned}\int_{\Theta} g(\Theta)\pi(\Theta)d\Theta &= \int_{\Theta} \frac{g(\Theta)\pi(\Theta)}{q(\Theta)} q(\Theta)d\Theta \\ &\approx \hat{g}(\Theta) = \frac{1}{N}\sum_{i=1}^{N} g(\Theta^{(i)})w^{(i)}. \end{aligned} \tag{3.6}$$

We can see that the estimator $\hat{g}(\Theta)$ is unbiased,

$$\begin{aligned} E_q(\hat{g}(\Theta)) &= \frac{1}{N}\sum_{i=1}^{N} E(g(\Theta^{(i)})w^{(i)}) \\ &= \frac{1}{N}\sum_{i=1}^{N} \int_{\Theta^{(i)}} g(\Theta^{(i)}) \frac{\pi(\Theta^{(i)})}{q(\Theta^{(i)})} q(\Theta^{(i)})d\Theta^{(i)} \\ &= \int_{\Theta} g(\Theta)\pi(\Theta)d\Theta) \\ &= E_\pi(g(\Theta)), \end{aligned} \tag{3.7}$$

and as $N \to \infty$, $\hat{g}(\Theta) \to E(g(\Theta))$, the estimator converges by the law of large numbers.

If $g(\Theta)$ is the identity, then a sample $\{\Theta_{(1)}, \ldots, \Theta_{(S)}\}$ sampled with probabilities equivalent to the normalised importance weights $\bar{w}_{(i)}$

$$\bar{w}_{(i)} = \frac{w_{(i)}}{\sum_{j=1}^{N} w_{(j)}}, \qquad 1, \ldots N, \tag{3.8}$$

has approximate distribution $\pi(\Theta)$. It is easy to see that we need only know $\pi(\cdot)$ up to a normalising constant, since this drops out of the normalised importance weights. This second sampling step is typically called re-sampling. The sampling importance re-sampling (SIR) technique was initially proposed by Rubin (1987) before being refined in Rubin *et al.* (1988). Provided the number of proposals, N, is sufficiently large, the sample of S points will be a reasonable approximation to $\pi(\Theta)$. Performance here depends on the proposal distribution $q(\cdot)$. If $q(\cdot)$ is too dissimilar to $\pi(\cdot)$ then it is likely that only a few points will have non-negligible weight. An optimal proposal would be $q(\cdot) = \pi(\cdot)$ but obviously this is not possible as if we had access to $\pi(\cdot)$ we could have just sampled from it directly.

3.4 Sequential Monte Carlo

Sequential Monte Carlo methods are a class of algorithms that provide provide weighted samples from a sequence of distributions, typically using importance sampling and re-sampling techniques. The approach was first established in the context of state space models with the aim of finding a solution to what is known as the filtering problem and takes the name particle filter.

3.4.1 The particle filter

The particle filter is a specific type of sequential Monte Carlo scheme, first developed by Gordon *et al.* (1993), applicable for hidden Markov or state space models where interest lies in the distribution of an unobserved or latent state at some discrete time, $\mathbf{x}_t$, of a Markov process given noisy observations of that process $\mathbf{y}_{1:t}$, $\pi(\mathbf{x}_t \,|\, \mathbf{y}_{1:t})$, where each observation is such that there exists some observation density $\pi(\mathbf{y}_t \,|\, \mathbf{x}_t)$. The particle filter is a generalisation of the Kalman filter, (Kalman, 1960), allowing a solution to the filtering problem for non–linear dynamics.

First suppose that at a given time t that the distribution $\pi(\mathbf{x}_{t-1} \,|\, \mathbf{y}_{1:t-1})$ is available. A prior distribution for the state at time t, $\mathbf{x}_t$ given the data so far $\mathbf{y}_{1:t-1}$ could be constructed as

$$\begin{aligned}\pi(\mathbf{x}_t \,|\, \mathbf{y}_{1:t-1}) &= \int \pi(\mathbf{x}_t \,|\, \mathbf{x}_{t-1}, \mathbf{y}_{1:t-1})\pi(\mathbf{x}_{t-1} \,|\, \mathbf{y}_{1:t-1})d\mathbf{x}_{t-1} \\ &= \int \pi(\mathbf{x}_t \,|\, \mathbf{x}_{t-1})\pi(\mathbf{x}_{t-1} \,|\, \mathbf{y}_{1:t-1})d\mathbf{x}_{t-1}. \end{aligned} \tag{3.9}$$

Here, one makes use of the fact that $\pi(\mathbf{x}_t \,|\, \mathbf{x}_{t-1}, \mathbf{y}_{1:t-1}) = \pi(\mathbf{x}_t \,|\, \mathbf{x}_{t-1})$ since the model governing the evolution of the latent state of the system is a Markov process. As the new observation, $\mathbf{y}_t$, becomes available then this prior can be updated using Bayes theorem

$$\pi(\mathbf{x}_t \,|\, \mathbf{y}_{1:t}) = \frac{\pi(\mathbf{y}_t \,|\, \mathbf{x}_t)\pi(\mathbf{x}_t \,|\, \mathbf{y}_{1:t-1})}{\pi(\mathbf{y}_t \,|\, \mathbf{y}_{1:t-1})}. \tag{3.10}$$

Equations 3.9 and 3.10 form a recurrence relation that gives the exact solution to the filtering problem. This recursive propagation of the posterior distribution is rarely available analytically in practice, this is certainly the case for the majority of stochastic kinetic models like those introduced in chapter 2, and hence the solution

is approximated through use of importance sampling techniques.

Suppose then that instead of having access at time point t to the density $\pi(\mathbf{x}_{t-1} \,|\, \mathbf{y}_{1:t-1})$ that we instead have a collection of N weighted samples $\left\{\mathbf{x}_{t-1}^{(i)}, w_{t-1}^{(i)}\right\}$ for $i = 1, \ldots, N$ representing a discrete approximation to $\pi(\mathbf{x}_{t-1} \,|\, \mathbf{y}_{1:t-1})$. If interest lies in $\pi(\mathbf{x}_t \,|\, \mathbf{y}_{1:t})$, that is the updated posterior distribution for the state in light of the new observation, one could create a proposal distribution for an importance sampler of the form

$$q(\mathbf{x}_t \,|\, \mathbf{y}_{1:t}) = q(\mathbf{x}_t \,|\, \mathbf{x}_{t-1}, \mathbf{y}_{1:t}) q(\mathbf{x}_{t-1} \,|\, \mathbf{y}_{1:t-1}). \tag{3.11}$$

That is we could obtain samples $\mathbf{x}_t^{(j)} \sim q(\mathbf{x}_t \,|\, \mathbf{y}_{1:t})$ by taking existing samples from distribution $q(\mathbf{x}_{t-1} \,|\, \mathbf{y}_{1:t-1})$, and augmenting them with the new state $\mathbf{x}_t^{(j)} \sim q(\mathbf{x}_t \,|\, \mathbf{x}_{t-1}, \mathbf{y}_{1:t})$. It is convenient to choose $q(\mathbf{x}_t \,|\, \mathbf{x}_{t-1}, \mathbf{y}_{1:t}) = \pi(\mathbf{x}_t \,|\, \mathbf{x}_{t-1})$ since we can then augment our samples $\mathbf{x}_{t-1}^{(j)} \sim q(\mathbf{x}_{t-1} \,|\, \mathbf{y}_{1:t-1})$ by advancing the state of the system according to the model $\pi(\mathbf{x}_t \,|\, \mathbf{x}_{t-1})$.

The weight update then can be written

$$\begin{aligned} w_t^{(j)} &= \frac{\pi(\mathbf{x}_t^{(j)} \,|\, \mathbf{y}_{1:t})}{q(\mathbf{x}_t^{(j)} \,|\, \mathbf{y}_{1:t})} \\ &\propto \frac{\pi(\mathbf{y}_t \,|\, \mathbf{x}_t^{(j)}) \pi(\mathbf{x}_t^{(j)} \,|\, \mathbf{x}_{t-1}^{(j)}) \pi(\mathbf{x}_{t-1}^{(j)} \,|\, \mathbf{y}_{1:t-1})}{q(\mathbf{x}_t^{(j)} \,|\, \mathbf{x}_{t-1}^{(j)}, \mathbf{y}_{1:t}) q(\mathbf{x}_{t-1}^{(j)} \,|\, \mathbf{y}_{1:t-1})} \\ &= w_{t-1}^{(j)} \frac{\pi(\mathbf{y}_t \,|\, \mathbf{x}_t^{(j)}) \pi(\mathbf{x}_t^{(j)} \,|\, \mathbf{x}_{t-1}^{(j)})}{q(\mathbf{x}_t^{(j)} \,|\, \mathbf{x}_{t-1}^{(j)}, \mathbf{y}_{1:t})}. \end{aligned} \tag{3.12}$$

Since we choose $q(\mathbf{x}_t \,|\, \mathbf{x}_{t-1}, \mathbf{y}_{1:t}) = \pi(\mathbf{x}_t \,|\, \mathbf{x}_{t-1})$ then equation 3.12 reduces to

$$w_t^{(j)} = w_{t-1}^{(j)} \pi(\mathbf{y}_t \,|\, \mathbf{x}_t^{(j)}). \tag{3.13}$$

The full bootstrap particle filtering algorithm is described in Algorithm 3. At step 3 of this algorithm we first resample the weighted points $\left\{\mathbf{x}_{t-1}^{(j)}\right\}$ representing $\pi(\mathbf{x}_{t-1} \,|\, \mathbf{y}_{t-1})$ to obtain an equally weighted sample. In practice this is performed by sampling indicies of particles within the weighted sample, with probabilities equivalent to the normalised weights. Note that throughout this thesis, implementations of the bootstrap particle filter perform the resampling step at every population. This is to tackle something known as the degeneracy problem, the phenomenon that if we were to run a sequential importance sampler, after a few iterations, all but very

Algorithm 3 Bootstrap Particle Filter

1. Initialise with a sample $x_0^{(i)} \sim \pi(x_0), i = 1, \ldots, N$ with weights $\tilde{w}_0^{(i)} = 1/N$.
2. At time $t-1$ we have the weighted sample $\{x_{t-1}^{(i)}, w_{t-1}^{(i)}\}$ from $\pi(x_{t-1} \,|\, \mathbf{y}_{1:t-1})$.
3. Re-sample, generating an equally weighted sample $\{\tilde{x}_{t-1}^{(i)}\}$.
4. Forward simulate each particle using the model to obtain samples $x_t \sim \pi(x_t \,|\, \tilde{x}_{t-1})$.
5. Weight each particle $w_t^{(i)} = \pi(y_t \,|\, x_t^{(i)})$ and calculate normalised weights
$$\tilde{w}_t^{(i)} = \frac{w_t^{(i)}}{\sum_{i=1}^N w_t^{(i)}}.$$
6. Set $t := t+1$ and return to 2.

few particles will have negligible weight. Doucet (1998) showed that it is impossible to avoid the degeneracy problem since the variance of the importance weights can only increase over time. If few particles have non–negligible weight, this implies that we are wasting computational effort updating a large number of particles whose contribution to the approximation of $\pi(\mathbf{x}_t \,|\, \mathbf{y}_{1:t})$ is very small.

A measure of degeneracy in a sequential Monte Carlo sampler is effective sample size, N_{ESS}, see for example Liu & Chen (1998) for an introduction. Effective sample size is a concept used for a collection of samples from a distribution which are correlated or weighted. It describes the equivalent number of independent samples which would represent the same amount of information contained within the data. It is estimated by

$$\hat{N}_{ESS} = \frac{1}{\sum_{i=1}^N (\tilde{w}_t^{(i)})^2} \tag{3.14}$$

where $\tilde{w}_t^{(i)}$ is the normalised weight obtained in the sampler

$$\tilde{w}_t^{(i)} = \frac{w_t^{(i)}}{\sum_{j=1}^N w_t^{(j)}}.$$

$\hat{N}_{ESS} << N$ implies degeneracy. One approach to reduce the effects of degeneracy is to use resampling. The basic idea being that sampling $\{\mathbf{x}_t, w_t\}$ with replacement will concentrate on particles with larger weights, eliminating those with very small weights. The resultant sample is an i.i.d sample from the discrete density approximation and hence all weights are reset to $w_t^{(i)} = 1/N$.

It is important to make a distinction between the particle filter and the generic sequential Monte Carlo sampler as each will be exploited for the purpose of inference for rate parameters in stochastic kinetic models in subsequent chapters. The bootstrap particle filter relies on the increasing dimension of the latent state making it a natural fit to state space models like those introduced in chapter 2 because of the underlying Markov property. The general sequential sampler does not impose such a restriction and can be used to sample from an arbitrary sequence of distributions. For a more comprehensive introduction to sequential Monte Carlo techniques for state space models see Doucet *et al.* (2001*a*).

3.4.2 General SMC sampler

Liu & Chen (1998) showed the sequential Monte Carlo techniques were applicable in a much wider range of problems than state space models, to any sequence of distributions of increasing dimension. The theory was then developed further by Del Moral *et al.* (2006) who made clever use of forward and backward Markov kernels to allow a SMC scheme to target an arbitrary sequence of distributions.

The more general scheme, described in algorithm 4, propagates samples from $\pi_{t-1}(\Theta)$ through some forward kernel $K(\Theta_t \,|\, \Theta_{t-1})$. This defines a joint density for the proposal $q(\Theta_{t-1}, \Theta_t) = \pi_{t-1}(\Theta_{t-1})K(\Theta_t \,|\, \Theta_{t-1})$. A joint target $\pi(\Theta_{t-1}, \Theta_t)$ of the form $\pi_t(\Theta_t)L(\Theta_{t-1} \,|\, \Theta_t)$ where $L(\cdot \,|\, \cdot)$ is commonly referred to as a backward kernel, will admit the correct marginal for the target, $\pi_t(\Theta)$. In practice this backward kernel is entirely arbitrary since it is not needed for simulation throughout the procedure however it does have influence on the efficiency of the sampler due to its appearance in the importance weights

$$w(\Theta_{t-1}, \Theta_t) = \frac{\pi_t(\Theta_t)L(\Theta_{t-1} \,|\, \Theta_t)}{\pi_{t-1}(\Theta)K(\Theta_t \,|\, \Theta_{t-1})}. \tag{3.15}$$

Del Moral *et al.* (2006) showed that an optimal choice for $L(\cdot \,|\, \cdot)$ is

$$L^{\text{opt}}(\Theta_{t-1} \,|\, \Theta_t) = \frac{\pi_{t-1}(\Theta)K(\Theta_t \,|\, \Theta_{t-1})}{\int \pi_{t-1}(\Theta)K(\Theta_t \,|\, \Theta_{t-1})d\Theta_{t-1}}. \tag{3.16}$$

Algorithm 4 The generic SMC algorithm

1. Initialise at time $t = 0$, particle $i = 1$.
2. Sample $\Theta_t^i \sim \pi_t(\cdot)$.
3. Conditional on this sample propose $\Theta_{t+1} \sim K_t(\Theta_{t+1} \,|\, \Theta_t)$.
4. Set
$$w_{t+1}^i = \frac{\pi_{t+1}(\Theta_t^i)}{\int K_t(\Theta_{t+1}^i \,|\, \Theta_t^i)\pi_t(\Theta_t^i)} d\Theta_t.$$
5. If the population indicator $i < N$ set $i := i + 1$ and return to 2.
6. Re-sample $\{\Theta_t^i, w_t^i\}$ to obtain $\{\bar{\Theta}_t^i, \frac{1}{N}\}$.
7. if $t < T_{\max}$ set $t := t + 1$, $i = 1$ and return to 2.

Using this choice for $L(\cdot \,|\, \cdot)$ gives the importance weights as

$$w_{(i)} = \frac{\pi_t(\Theta_t^i)}{\int K(\Theta_t^i \,|\, \Theta_{t-1}^i)\pi_{t-1}(\Theta_{t-1}^i) d\Theta_{t-1}}. \tag{3.17}$$

It is of note that the integral in the denominator of equation 3.17 will often be intractable. If this is the case it can be approximated given a sample $\{\Theta_1, \ldots, \Theta_N\}$ from $\pi_{t-1}(\Theta)$

$$\int K(\Theta_t^i \,|\, \Theta_{t-1}^i)\pi_{t-1}(\Theta_{t-1}^i) d\Theta_{t-1} \approx \frac{1}{N} \sum_{i=1}^{N} K(\Theta_t \,|\, \Theta_i).$$

The scheme as described in algorithm 4 is a construction for an essentially arbitrary sequence of distributions extending the particle filter, developed by Gordon *et al.* (1993).

3.5 Markov chain Monte Carlo (MCMC)

Markov Chain Monte Carlo techniques aim to draw samples from some target distribution $\pi(\cdot)$ by simulating from a carefully constructed Markov chain, whose stationary distribution is the desired target $\pi(\cdot)$. The idea being that once the chain has converged to its stationary distribution, then simulated values will represent samples from the distribution of interest. If the target distribution were multi-dimensional, then our samples will actually represent the appropriate marginal distributions of interest.

3.5.1 Detailed Balance

Consider a discrete time Markov chain $\{\Theta^{(t)}, t = 1, 2, \ldots\}$ in a continuous state space $\mathcal{S} \subseteq \mathbb{R}^p$ with transition kernel $P(\Theta^{(t+1)} \in \mathcal{A} \,|\, \Theta^{(t)} = x) = P_t(x, \mathcal{A})$. Let $p_t(x, y)$ be the transition density from state x to state y at time t.

Note that a transition kernel is said to be homogeneous if $p_t(x, y) = p(x, y)$, that is the transition density has no dependence on t.

Consider that if $\Theta^{(0}, \Theta^{(2)}, \ldots, \Theta^{(N)}$ is a Markov chain, that the reversal of this sequence of states $\Theta^{(N)}, \ldots, \Theta^{(0)}$ is also a Markov chain, see for example Ross (1983). Let $p_t^*(x, y)$ be the transition kernel for the reversed chain. If

$$p_t^*(x, y) = p(x, y)$$

then the chain is said to be reversible and we have

$$\pi(x)p(x, y) = \pi(y)p(y, x), \tag{3.18}$$

the detailed balance equations.

This condition leads to the following result by integrating both sides with respect to x.

Definition 1. A distribution $\pi(\cdot)$ is a stationary distribution of a Markov chain with transition density $p(\cdot, \cdot)$ if

$$\pi(y) = \int_{\mathcal{S}} \pi(x)p(x, y)dx.$$

Essentially this result informs us that once a chain has converged to its stationary distribution, it will remain in that distribution for all time after. See Gamerman (1997) for a rigorous discussion of Markov chain Monte Carlo techniques. Hence if a Markov chain can be constructed for a desired target $\pi(\Theta)$ such that we satisfy detailed balance, then $\pi(\Theta)$ is a stationary distribution of the chain and, on convergence, each value simulated from the Markov chain is a sample from the distribution of interest.

Algorithm 5 Metropolis-Hastings Algorithm

1. Initialise with iteration counter $i = 1$ and $\Theta_{(0)}$.
2. Propose a value $\Theta^* \sim q(\cdot \,|\, \Theta_{(i-1)})$.
3. Evaluate the acceptance probability, $\min\{1, \alpha(\Theta^* \,|\, \Theta_{(i-1)})\}$ where

$$\alpha(\Theta^* \,|\, \Theta_{(i-1)}) = \frac{\pi(\Theta^*)}{\pi(\Theta_{(i-1)})} \times \frac{q(\Theta_{(i-1)} \,|\, \Theta^*)}{q(\Theta^* \,|\, \Theta_{(i-1)})}.$$

4. Set $\Theta_{(i)} = \begin{cases} \Theta^* & \text{with probability } \min\{1, \alpha(\Theta^* \,|\, \Theta_{(i-1)})\} \\ \Theta_{(i-1)} & \text{otherwise} \end{cases}$
5. Set the iteration counter $i := i + 1$ and return to step 2.

3.5.2 Metropolis-Hastings (MH)

The Metropolis-Hastings algorithm is an algorithm used to construct Markov chains with stationary distribution $\pi(\cdot)$. This scheme was first proposed in Metropolis *et al.* (1953) and later adapted and generalized by Hastings (1970).

Say we wish to implement such a scheme to target a distribution $\pi(\Theta)$. We can implement the algorithm by constructing a proposal distribution, $q(\cdot \,|\, \Theta)$, that is easy to sample from. The scheme then follows the form in Algorithm 5.

The choice of the proposal distribution, $q(\Theta^* \,|\, \Theta)$, will strongly influence the performance of the algorithm. Often the proposal takes one of two forms, an independence proposal, or a local random walk. Since the target appears in both the numerator and denominator of the acceptance ratio, it is possible to sample from a distribution known only up to a normalising constant, since this cancels in $\alpha(\cdot, \cdot)$.

Independence Proposal

An independence proposal, rather intuitively, returns proposed values that are independent of the current value,

$$q(\Theta^* \,|\, \Theta) = f(\Theta^*).$$

When being used to integrate, say, a posterior in the Bayesian framework, this could typically be proposed values from the prior density for example. In such a proposal the acceptance probability $\alpha(\Theta^* \,|\, \Theta)$ will be governed by how similar the proposal $f(\cdot)$ is to the target $\pi(\cdot)$.

Random Walk

A common choice for proposal in these algorithms is of the form of a local random walk. Here the next value is proposed conditional on the current state of the chain. Typically this is in the form of either a uniform distribution

$$q(\theta_i^* \,|\, \theta_i) \sim \mathcal{U}(\theta_i - \sigma_i, \theta_i + \sigma_i),$$

with σ_i treated as some tuning parameter to determine how far we can move, or a Gaussian distribution

$$q(\Theta^* \,|\, \Theta) \sim \mathcal{N}(\Theta, \Sigma)$$

with covariance matrix Σ taking the role of the tuning parameter. Too large an innovation will result in a poorly mixing chain where few of the proposals will be accepted. Too small a move and nearly everything will be accepted, however the exploration of the parameter space will be slow and the chain will experience high auto-correlation. Clearly some trade off has to be made here and Roberts *et al.* (1997) show that an acceptance rate of $\alpha(\Theta^* \,|\, \Theta) \approx 0.234$ is optimal under fairly general conditions.

3.5.3 MCMC Analysis

We only obtain samples from the target distribution $\pi(\cdot)$ in the limit as the number of iterations tends to infinity. However, in practice, we assume that a simulated value at a suitable large iteration is drawn from $\pi(\cdot)$. There are a number of key issues which need to be considered when using MCMC methods regarding accuracy and efficiency.

Burn-in Period

By construction, as the number of iterations increases, the distribution of the simulated sample approaches the stationary target of interest $\pi(\cdot)$. After we are deemed to have reached convergence, then all such simulated values have this desired distribution. However, it is necessary to consider that convergence will not be immediate. Usually we would use the scheme to generate a sequence of samples, and then using

some diagnostic checks, determine what is known as a burn-in period, the approximate number of iterations required to achieve convergence. We then discard these initial samples on the basis that all subsequent samples are being drawn from the stationary distribution. Convergence diagnostics for MCMC sampling have been suggested by Heidelberger & Welch (1983); Gelfand & Smith (1990); Gelman (1996); Raftery & Lewis (1996).

Thinning a Chain

Having achieved convergence, it is likely that consecutive simulated values are significantly correlated with one another (auto-correlation). In order that we simulate approximately independent and identically distributed random samples from the target distribution $\pi(\cdot)$ it is necessary to thin the sample. Auto-correlation plots can be used as a diagnostic tool for determining the level of auto-correlation experienced by the chain (influenced by the proposal distribution $q(\cdot, \cdot)$). Having found this value, n say, we could thin our sample by taking only every n^{th} simulated value.

Monte Carlo methods have become of paramount importance for inference within the Bayesian framework as they form the basis for many schemes for sampling from otherwise problematic posterior distributions.

Chapter 4

Bayesian inference

4.1 Introduction

Inferential statistics within the Bayesian framework is typically driven by interest in some posterior distribution $\pi(\Theta \,|\, \mathbf{Y})$. That is the distribution of some variables of interest Θ, given the data observations $\mathbf{Y}$. Bayes theorem tells us that

$$\pi(\Theta \,|\, \mathbf{Y}) = \frac{\pi(\mathbf{Y} \,|\, \Theta)\pi(\Theta)}{\pi(\mathbf{Y})}$$

where $\pi(\Theta)$ is the prior distribution, capturing any information or beliefs we might hold on the parameters of interest before the incorporation of information from observations, on Θ and $\pi(\mathbf{Y} \,|\, \Theta)$ is the likelihood function, the probability of observing data $\mathbf{Y}$ given parameters Θ. Given a prior distribution and a likelihood function, using Bayes theorem, both pieces of information are used to obtain the posterior distribution. Note that the denominator,

$$\pi(\mathbf{Y}) = \int_{\Theta} \pi(\mathbf{Y} \,|\, \Theta)\pi(\Theta) d\Theta,$$

is a normalising constant that is not a function of the parameters Θ. Hence the posterior distribution is often written as proportional to the product of likelihood and prior densities,

$$\pi(\Theta \,|\, \mathbf{Y}) \propto \pi(\mathbf{Y} \,|\, \Theta)\pi(\Theta).$$

If no analytic form for the posterior distribution is available, we can draw samples from the posterior distribution via some Monte Carlo method such as those described in chapter 3 where the posterior $\pi(\mathbf{Y} \mid \Theta)$ takes on the role of the target distribution to be sampled from. Note that for each of the Monte Carlo sampling methods introduced in chapter 3 the target from which we wish to sample need only be known up to a normalising constant. Therefore the ability to sample from any given posterior distribution relies on the ability to evaluate the density of the prior distribution and the likelihood function at proposed values of Θ.

Consider parameter inference for a stochastic kinetic model. As noted at the end of chapter 2 there exists only a handful of reaction systems for which the chemical master equation yields an analytical solution. This in turn means that for more complex systems it is not possible to directly evaluate a likelihood function. Hence traditional Bayesian techniques fall short in this context. Inference for rate parameters in this scenario relies on techniques developed for situations where we can not compute the likelihood directly, other wise known as likelihood-free inference methods.

It should be noted that if one were to observe every reaction that took place in the system during our observation time, where (t_i, ν_i) represent the time and type of the i^{th} reaction event, then we could formulate the likelihood as

$$\pi(\mathbf{Y}|\Theta) = \left\{ \prod_{i=1}^{n} h_{\nu_i}(x_{t_{i-1}}, \theta_{\nu_i}) \right\} \exp\left\{ - \int_0^T h_0(x_t, \Theta) dt \right\}.$$

In this complete-data scenario, the integral here is a finite sum and the likelihood is tractable, irrespective of the complexity of the reaction network. However in practice we would never observe the full trajectory of the species.

4.2 Likelihood free Bayesian inference techniques

When the likelihood function for a model is unavailable, whether that is due to analytic intractability or computational infeasibility, we need consider methods for inference that do not rely on its direct evaluation. Likelihood free techniques rely heavily on the ability to draw realisations from the model $\mathbf{Y}^* \sim \pi(\mathbf{Y} \mid \Theta)$. Such

techniques work by augmenting the target posterior distribution

$$\pi_{LF}(\Theta, \mathbf{Y}^* \mid \mathbf{Y}) \propto \pi(\mathbf{Y} \mid \mathbf{Y}^*, \Theta)\pi(\mathbf{Y}^* \mid \Theta)\pi(\Theta)$$

with simulated data $\mathbf{Y}^*$.

Note that when $\mathbf{Y}^* = \mathbf{Y}$ then $\pi_{LF}(\Theta, \mathbf{Y} \mid \mathbf{Y}) \propto \pi(\mathbf{Y} \mid \Theta)\pi(\Theta)$ and hence samples are drawn from the posterior target exactly. The aim is to marginalise out the simulated data $\mathbf{Y}^*$,

$$\pi_{LF}(\Theta \mid \mathbf{Y}) \propto \pi(\Theta) \int_{\mathbf{Y}^*} \pi(\mathbf{Y} \mid \mathbf{Y}^*, \Theta)\pi(\mathbf{Y}^* \mid \Theta) d\mathbf{Y}^*.$$

The marginal posterior distribution, $\pi_{LF}(\Theta \mid \mathbf{Y})$ estimates the actual posterior distribution, $\pi(\Theta \mid \mathbf{Y})$. Choice comes over the specification of $\pi(\mathbf{Y} \mid \mathbf{Y}^*, \Theta)$. Ideally we want to choose this function such that it takes larger values when $\mathbf{Y}^*$ and $\mathbf{Y}$ are similar.

This section introduces some of the methods that are available in this scenario.

4.2.1 Approximate Bayesian computation (ABC)

Approximate Bayesian computation techniques have increased in popularity in recent years due to their applicability to inference for problems in which the likelihood function $\pi(\mathbf{Y} \mid \Theta)$ is unavailable. Rather than evaluation of the likelihood these methods rely on our ability to be able to simulate from a model given some parameters and all ABC methods follow a common procedure. First we generate a candidate parameter vector, Θ^*, via some proposal mechanism $q(\cdot)$. Conditional on this we simulate some synthetic data, $\mathbf{Y}^* \sim \pi(\cdot \mid \Theta^*)$ and we keep our proposed parameters if this synthetic data is deemed to be "close enough" to our observed data. In ABC, in order to decide how similar $\mathbf{Y}$ and $\mathbf{Y}^*$ are we choose some function of the distance between them,

$$\pi_\epsilon(\mathbf{Y} \mid \mathbf{Y}^*, \Theta) = \frac{1}{\epsilon} K\left(\frac{|\mathbf{Y} - \mathbf{Y}^*|}{\epsilon}\right)$$

where K is some kernel density with scale parameter ϵ.

Algorithm 6 ABC rejection sampler

1. Sample $\Theta^* \sim \pi(\Theta)$.
2. Simulate a dataset, $\mathbf{Y}^* \sim \pi(\cdot|\Theta^*)$.
3. If $\rho(\mathbf{Y}, \mathbf{Y}^*) \leq \epsilon$, accept Θ^*.
4. Return to 1.

A common choice is the uniform kernel, see for example Marjoram *et al.* (2003),

$$\pi_\epsilon(\mathbf{Y} \mid \mathbf{Y}^*, \Theta) \propto \begin{cases} 1 & \text{if } \rho(\mathbf{Y}, \mathbf{Y}^*) \leq \epsilon, \\ 0 & \text{otherwise.} \end{cases} \tag{4.1}$$

Hence closeness here is typically measured by some metric on the candidate and observed datasets, $\rho(\mathbf{Y}, \mathbf{Y}^*)$, falling below some predefined tolerance, ϵ.

The idea of using simulation for inference in situations where the likelihood function is intractable was first proposed in Diggle & Gratton (1984) with modifications by Tavare *et al.* (1997). The algorithm proposed replaced the full data with a summary statistic $\mathbf{s}$ and accepted proposed parameter values with probability proportional to $P(S = \mathbf{s}|\Theta^*)$. Although useful, the approach was limited to use in relatively simple settings where $P(S = \mathbf{s}|\Theta)$ can easily be computed and maximised over Θ. Fu & Li (1997) generalised this by replacing the computation of $P(S = \mathbf{s}|\Theta)$ with a simulation step. This was the now commonly recognised "simulate from the model" step. This algorithm, proposed in the context of population genetics, require that simulated data be equivalent to observed data. Ideally, given a collection of parameter vectors, Θ, we would keep all vectors that gave rise to simulated data which is equivalent to our observed data set. In practice however, the probability that a candidate data set $\mathbf{Y}^* = \mathbf{Y}$, is almost 0. This led to the extension in Weiss & von Haeseler (1998) of introducing a tolerance ϵ and accepting when $\rho(\mathbf{Y}, \mathbf{Y}^*) \leq \epsilon$. A similar rejection sampling method was adopted by Pritchard *et al.* (1999) but with simulation of parameters from the prior. This led to the ABC rejection sampler algorithm as defined in algorithm 6 although the term ABC wasn't coined until 2002, (Beaumont *et al.*, 2002). Hence the posterior distribution is approximated with a collection of parameter vectors that yield data simulation close to that which we have observed, formally $\pi(\Theta \mid \rho(\mathbf{Y}, \mathbf{Y}^*) \leq \epsilon)$.

4.2.2 The role of summary statistics in ABC

Summary statistics have played a crucial role in the formation of ABC algorithms since the first such algorithm proposed by Tavare *et al.* (1997). As the dimension of the data increases the probability of generating a synthetic data set, $\mathbf{Y}^*$, that has a small distance to $\mathbf{Y}$ reduces dramatically. The result is that given a fixed value ϵ the acceptance rates of the rejection sampler become intolerably small. This in turn diminishes the computational efficiency of the sampler. A common approach to reduce the effect of this issue is to replace $\mathbf{Y}$ with a set of lower dimensional summary statistics. The acceptance step in algorithm 6 is then replaced with

$$\rho(S(\mathbf{Y}), S(\mathbf{Y}^*)) \leq \epsilon.$$

If the chosen set of summary statistics are sufficient with respect to Θ, then there is no loss of information and the increase in efficiency does not introduce any error. In this scenario, as $\epsilon \to 0$ then the resultant ABC posterior tends to the true posterior of interest, $\pi(\Theta \,|\, \rho(S(\mathbf{Y}), S(\mathbf{Y}^*)) \leq \epsilon) \to \pi(\Theta \,|\, \mathbf{Y})$.

For most problems, there does not exist a set of sufficient statistics hence summary statistics are chosen in such a way that it is hoped they capture the important features of the data. By utilising a set of non–sufficient statistics we introduce a further layer of approximation due to the information loss in summarising the data in such a way.

ABC methods can be limited by the availability of informative summary statistics for any parameter as discussed in Hey & Machado (2003). A fairly intuitive solution to the lack of informative statistics is to increase the number of statistics used in the construction of the metric function, thereby increasing the amount of information available to determine acceptance. However, increasing the number of summary statistics can lead to poorer inference, Beaumont *et al.* (2002), and certainly causes issues when deciding a sensible metric. Further discussion can be found in Csilléry *et al.* (2010) and Blum *et al.* (2013).

Algorithm 7 ABC MCMC

1. Initialise Θ_0.
2. Propose Θ^* according to some proposal kernel, $q(\Theta^* \mid \Theta_i)$.
3. Simulate a dataset, $\mathbf{Y}^* \sim \pi(\cdot \mid \Theta^*)$.
4. If $\rho(S(\mathbf{Y}), S(\mathbf{Y}^*)) \leq \epsilon$ go to 5, otherwise set $\Theta_{i+1} = \Theta_i$ and go to 6.
5. Set $\Theta_{i+1} = \Theta^*$ with probability

$$\alpha(\Theta_{i+1}|\Theta^*) = \min\left(1, \frac{\pi(\Theta^*)q(\Theta_i|\Theta^*)}{\pi(\Theta_i)q(\Theta^*|\Theta_i)}\right)$$

   and $\Theta_{i+1} = \Theta_i$ with probability $1 - \alpha$.
6. Set $i = i + 1$, and go to 2.

4.2.3 ABC MCMC

One of the issues with the simple rejection sampler algorithm discussed in the previous section is the uninformed exploration of the parameter space. It is possible to use a rejection sampler within an MCMC sampler to create a Markov chain with stationary distribution $\pi(\Theta|\rho(S(\mathbf{Y}), S(\mathbf{Y}^*)) \leq \epsilon)$. Such an approach is defined in algorithm 7. We can then use the last accepted parameter values as the basis for the next proposal using an appropriate random walk kernel.

The ABC MCMC was initially developed by Marjoram *et al.* (2003) who also show that the chain does in fact yield the targeted approximate posterior distribution $\pi(\Theta|\rho(S(\mathbf{Y}), S(\mathbf{Y}^*)) \leq \epsilon)$

Whilst in theory the chain will converge, in practice it is important to consider initialisation of the chain. A random sample from an uninformative prior may lead to a poor initial choice of Θ_0 that has negligible posterior density. If this is the case then it is likely that neighbouring parameter values proposed under the random walk scheme will also have low density and hence the algorithm has a tendency to stick at the initial value. This is due to step 4 where parameter values of low posterior density are unlikely to yield metric values which are below the defined tolerance. It is possible to set up the chain such that the acceptance probability $\alpha(\Theta^*|\Theta) = 1$ by using a symmetric proposal distribution and a uniform prior. In doing so the chain moves only according to whether we simulate data within the tolerance. The advantage of this scheme over the simple rejection sampler is that we are no longer blindly proposing values from the prior distribution $\pi(\Theta)$. Once the chain has converged it is possible to obtain a larger approximate posterior sample

Algorithm 8 ABC SMC (Toni *et al.*, 2009)

1. Initialise $\epsilon_0 > 0$ and set the population indicator, $t = 0$.
2. Set particle indicator, $i = 1$.
3. If t = 0, sample $\theta^{**} \sim \pi(\theta)$
   Else sample θ^* from the previous population $\{\theta_{t-1}^{(i)}\}$ with weights w_{t-1} and perturb to obtain $\theta^{**} \sim K_t(\theta|\theta^*)$
   If $\pi(\theta^{**}) = 0$, return to 3.
   Simulate a candidate dataset $\mathbf{Y}^* \sim \pi(\mathbf{Y}|\theta^{**})$.
   If $\delta_t^* = \rho(\mathbf{Y}, \mathbf{Y}^*) \geq \epsilon_t$, return to 3.
4. Set $\theta_t^{(i)} = \theta^{**}$, $\delta_t^{(i)} = \delta_t^*$ and calculate weight for particle $\theta_t^{(i)}$, $\tilde{w}_t^{(i)}$

$$\tilde{w}_t^{(i)} = \left\{ \begin{array}{ll} 1, & \text{if } t = 0 \\ \frac{\pi(\theta_t^{(i)})}{\sum_{j=1}^N w_{t-1}^{(j)} K_t(\theta_{t-1}^{(j)}, \theta_t^{(i)})}, & \text{if } t > 0 \end{array} \right\}.$$

   If $i < N$, set $i = i + 1$ and go to 3.
5. Normalise the weights, $w_t^{(i)} = \frac{\tilde{w}_t^{(i)}}{\sum_i \tilde{w}_t^{(i)}}$, and select ϵ_{t+1} as the α%–ile of the collection of distances, $\delta_t^{(i)}$ for some chosen $0 < \alpha < 100$.
6. If $t < T$, set $t = t + 1$ and go to 2.

more efficiently than with the rejection only based approach. However the chain is sensitive to a poor initialisation.

4.2.4 ABC SMC

A sequential approach to ABC based on importance sampling was described in Toni *et al.* (2009). The scheme follows the prescription in algorithm 4 where the sequence of bridging distributions, denoted $\pi_{\epsilon_1}(\cdot), \ldots, \pi_{\epsilon_T}(\cdot)$, are ABC posterior distributions based on a decreasing sequence of tolerances $\epsilon_1 > \epsilon_2 > \ldots > \epsilon_T$. By starting with a relatively large tolerance and decreasing, the aim of scheme is to increase acceptance rates of the rejection sampler at the smallest tolerance by gradually moving the distribution toward the desired posterior. The algorithm is described in algorithm 8.

It is of note that a similar idea was proposed in Sisson *et al.* (2007) however the re-weighting step was incorrect resulting in bias. This was acknowledged by the authors and later corrected. The scheme was also proposed by Beaumont *et al.* (2009) independently. The efficiency of the scheme is dependent on the choice of propaga-

tion kernel $K_t(\cdot)$ and the sequence of tolerances $\epsilon_1 > \epsilon_2 > \ldots > \epsilon_T$ in addition to the choice of summary statistics and metric as in other ABC variants.

Optimality of proposal kernel $K_t(\cdot)$

An optimal choice of proposal kernel was the subject of Filippi *et al.* (2013) where optimal criteria were determined to be jointly minimising the Kullback–Leibler divergence between successive distributions and the average acceptance rates. They showed that for a component-wise Gaussian kernel, the optimal choice of parameter σ for the j^{th} component is

$$\sigma_j = \left(\sum_{i=1}^{N_1} \sum_{j=1}^{N_2} w_{t-1}^{(i)} \tilde{w}_{t-1}^{(j)} (\tilde{\Theta}_j^{(t-1)} - \Theta_i^{(t-1)})^2 \right)^{\frac{1}{2}}. \tag{4.2}$$

Here w_{t-1} are the weights that have been calculated at step 5 of algorithm 8, Θ are the N_1 accepted parameters that make up the sample from $\pi_{\epsilon_{t-1}}(\Theta)$, $\tilde{\Theta}$ are the N_2 parameters for which the the calculated distance $\delta_{t-1}^* < \epsilon_t$ and $\tilde{w}_{t-1}$ are the weights of those particles re–normalised such that $\sum \tilde{w}_{t-1} = 1$. This work was an extension of that found in Beaumont *et al.* (2009) that suggested an adaptive Gaussian proposal kernel whose variance is equivalent to twice the empirical variance of the samples $\Theta_{t-1} \sim \pi_{\epsilon_{t-1}}(\Theta)$. This was acknowledged by the authors (Filippi *et al.*, 2013) as a special case of equation 4.2 if one considers $\epsilon_t = \epsilon_{t-1}$. This reliance on previous samples for the improvement of proposals is legitimate from an importance sampling perspective as discussed by Douc *et al.* (2007).

Choosing a sequence of tolerances

Construction of a sequence of tolerances $\epsilon_1 > \epsilon_2 > \ldots > \epsilon_T$ has been the focus of research due to its effect on the overall efficiency of the scheme. A sequence of tolerances that decreases slowly will be expected to have high acceptance rates, the target being close to the proposals, however convergence to the posterior distribution will be slow. Alternatively a tolerance that decreases rapidly will encourage large movement from proposal to target, increasing the rate of convergence to the posterior at the expense of decreased acceptance rates of the sampler. Early implementations of the scheme used a geometric rate of decline for the sequence of tolerances as in

Beaumont *et al.* (2009). Later schemes favour an adaptive choice for the sequence, calculated online. Drovandi & Pettitt (2011) propose a choosing ϵ_t based on the distribution of distances at iteration $t-1$ of the algorithm. Specifically they choose the α percentile of the distances, leaving α as a tuning parameter. Whilst this has been shown to work well Silk *et al.* (2013) state that care must be taken as convergence is not guaranteed under the scheme in all cases. They show that for some values of α, typically one that yields a decrease in tolerance which is too fast, can cause the approximation to get stuck in a local optimum.

4.3 Particle MCMC (pMCMC)

4.3.1 Pseudo–marginal MCMC

Suppose interest lies in the posterior distribution $\pi(\Theta \,|\, \mathbf{Y}^*)$ and that we wish to construct a Metropolis Hastings scheme in order to sample from it. If the likelihood term, $\pi(\mathbf{Y} \,|\, \Theta)$, is unavailable it is intuitive that if we could obtain some estimate of this quantity, $\hat{\pi}(\mathbf{Y} \,|\, \Theta)$, we could plug this into the acceptance ratio in place of the likelihood. Under fairly mild conditions, that the estimate is unbiased, $E(\hat{\pi}(\mathbf{Y} \,|\, \Theta)) = \pi(\mathbf{Y} \,|\, \Theta)$, the stationary distribution of the MCMC scheme is the exact solution that would have been obtained if the likelihood function was readily available despite the fact that a Monte Carlo estimate has been used in the calculation of the acceptance ratio, see for example Beaumont (2003) for further discussion.

If we consider, as before, $\pi(\mathbf{Y} \,|\, \Theta)$ as

$$\pi(\mathbf{Y} \,|\, \Theta) = \int_{\mathbf{Y}^*} \pi(\mathbf{Y} \,|\, \mathbf{Y}^*, \Theta)\pi(\mathbf{Y}^* \,|\, \Theta) d\mathbf{Y}^*, \tag{4.3}$$

the marginal likelihood on integrating out any latent variables in the problem. Provided that we can evaluate $\pi(\mathbf{Y} \,|\, \mathbf{Y}^*, \Theta)$ and sample $\mathbf{Y}^* \sim \pi(\mathbf{Y}^* \,|\, \Theta)$ then there is a simple Monte Carlo approach to the estimation of $\pi(\mathbf{Y} \,|\, \Theta)$. That is simulate $\mathbf{Y}_1^*, \ldots, \mathbf{Y}_N^* \sim \pi(\mathbf{Y}^* \,|\, \Theta)$ and then

$$\hat{\pi}(\mathbf{Y} \,|\, \Theta) = \frac{1}{N} \sum_{i=1}^{N} \pi(\mathbf{Y} \,|\, \mathbf{Y}_i^*, \Theta).$$

By equation 3.3, $\hat{\pi}(\mathbf{Y} \mid \Theta)$ is unbiased and consistent, hence we can use this estimator in place of $\pi(\mathbf{Y} \mid \Theta)$ in an MCMC scheme and maintain the exact target.

Similarly imagine that it is not possible to directly sample $\pi(\mathbf{Y}^* \mid \Theta)$ but can instead draw from some alternative distribution $q(\mathbf{Y}^* \mid \Theta)$. Then by our knowledge of importance sampling, introduced in section 3.3, we could obtain an importance sampling estimate of $\pi(\mathbf{Y} \mid \Theta)$,

$$\pi(\mathbf{Y}^* \mid \Theta) = \int_{\mathbf{Y}^*} \pi(\mathbf{Y} \mid \mathbf{Y}^*, \Theta) \frac{\pi(\mathbf{Y}^* \mid \Theta)}{q(\mathbf{Y}^* \mid \Theta)} q(\mathbf{Y}^* \mid \Theta) d\mathbf{Y}^*. \tag{4.4}$$

So we can, given samples $\mathbf{Y}_1^*, \ldots, \mathbf{Y}_N^*$ from $q(\mathbf{Y}^* \mid \Theta)$, construct the estimate

$$\hat{\pi}(\mathbf{Y} \mid \Theta) = \frac{1}{N} \sum_{i=1}^{N} \pi(\mathbf{Y} \mid \mathbf{Y}_i^*, \Theta) w_i \tag{4.5}$$

where

$$w_i = \frac{\pi(\mathbf{Y}_i^* \mid \Theta)}{q(\mathbf{Y}_i^* \mid \Theta)}$$

are the importance weights. This estimator is also unbiased, see equation 3.7.

This idea was the subject of Andrieu & Roberts (2009) and is in fact a special case of the more general particle marginal Metropolis Hastings (PMMH) that was the subject of Andrieu *et al.* (2010). They suggest the use of a sequential Monte Carlo sampler, namely the bootstrap particle filter, as an estimator for $\pi(\mathbf{Y} \mid \Theta)$. Marjoram *et al.* (2003) are credited with development of the first likelihood free MCMC scheme. In their algorithm the rejection step can be considered as a Monte Carlo estimate, based on a sample size of 1. Andrieu & Roberts (2009) showed that more generally Monte Carlo estimates based on any sample size can be used in the accept-reject step of a Metropolis Hastings algorithm. Provided that the expectation of the Monte Carlo estimate corresponds to the true likelihood the MCMC scheme will converge to the true posterior distribution irrespective of any sampling error. A pseudo-marginal MCMC scheme targeting $\pi(\Theta \mid \mathbf{Y})$ is described in Algorithm 9.

Algorithm 9 Pseudo–marginal MCMC

1. Initialise iteration counter $i = 1$ and $\Theta_{(0)}$.
2. Propose a value $\Theta^* \sim q(\cdot \,|\, \Theta_{(i-1)})$.
3. Evaluate the acceptance probability $\min\left\{1, \alpha(\Theta^* \,|\, \Theta_{(i-1)})\right\}$ where

$$\alpha(\Theta^* \,|\, \Theta_{(i-1)}) = \frac{\hat{\pi}(\mathbf{Y}) \,|\, \Theta^*}{\hat{\pi})(\mathbf{Y} \,|\, \Theta_{(i-1)})} \times \frac{\pi(\Theta^*)}{\pi(\Theta_{(i-1)})} \times \frac{q(\Theta_{(i-1)} \,|\, \Theta^*)}{q(\Theta^* \,|\, \Theta_{(i-1)})},$$

4. where $\hat{\pi}(\mathbf{Y} \,|\, \cdot)$ is an unbiased estimate from a Monte Carlo estimator of $\pi(\mathbf{Y} \,|\, \cdot)$.
5. Set $\Theta_{(i)} = \begin{cases} \Theta^* & \text{with probability } \min\min\left\{1, \alpha(\Theta^* \,|\, \Theta_{(i-1)})\right\} \\ \Theta_{(i-1)} & \text{otherwise.} \end{cases}$
6. Set the iteration counter $i := i + 1$ and return to step 2.

4.3.2 Bootstrap particle filter

When implementing particle MCMC we require some method by which to estimate $\pi(\mathbf{Y} \,|\, \Theta)$. Within the context of state space models, like those discussed in chapter 2 this is easy via a bootstrap particle filter, (Doucet *et al.*, 2001*b*), see section 3.4.1, page 25, for an introduction to the bootstrap particle filter. The bootstrap particle filter is a sequential importance re–sampling scheme that relies on some measurement error distribution. That is there is some distribution for the observed data $\mathbf{Y}$, conditional on the unobserved state of the system $\mathbf{x}$, $\pi(\mathbf{Y} \,|\, \mathbf{x})$.

As with other MCMC type algorithms, efficiency of the sampler is dependent on the variance of the random walk proposal. In traditional likelihood based MCMC sampling, Roberts & Rosenthal (2001) showed that the optimal scaling for a Gaussian random walk kernel, under various assumptions about the target, is

$$\Sigma_q = \frac{2.38^2}{d}\Sigma \tag{4.6}$$

where Σ is the covariance of the posterior distribution and d is the number of parameters to be estimated. They also show that an optimal acceptance rate for proposed parameters is approximately 23.4%. Sherlock *et al.* (2015) showed however that in the context of a pseudo–marginal MCMC scheme that the optimal scaling of a Gaussian random walk is slightly different. They derive

$$\Sigma_q = \gamma\frac{2.56^2}{d}\Sigma \tag{4.7}$$

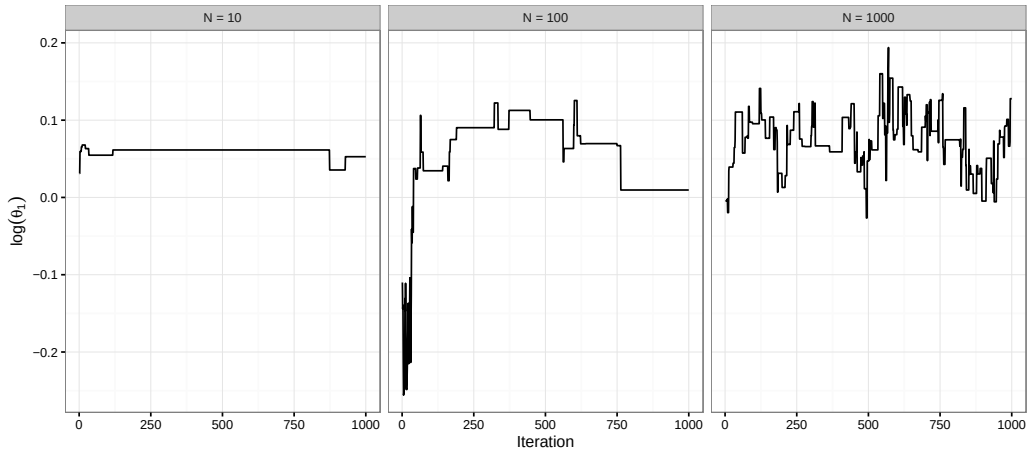


Figure 4.1: Diagnostic plots for pseudo–marginal MCMC using a bootstrap particle filter. The three chains represent differing numbers of particles being used for estimation of likelihood.

as the best variance matrix for the kernel, where γ is an additional multiplicative factor representing the roughness of the target with a value of $\gamma = 1$ being the case for a Gaussian target distribution. The difference between the optimal variance for a pseudo–marginal scheme with that for the traditional MCMC is attributed to the cost in estimation of the likelihood term. In addition the authors show that the optimal acceptance rate is much lower, 7% rather than 23.4%.

Since we are estimating the likelihood using a set of N particles, this is an additional tuning parameter that has bearing on the performance of the algorithm. The effect of the number, N, of particles used in the bootstrap filter on the resultant MCMC sampler will be explored in the following section.

4.3.3 Tuning the bootstrap particle filter

Since we are using the bootstrap particle filter for the estimation of likelihood, the number of particles used will have a bearing on the overall efficiency of the final MCMC sampler. A bootstrap particle filter that uses a very large number of particles will result in estimates of likelihood with small variability, but at a high computational cost. Since the bootstrap particle filter has to be run at every iteration of the MCMC sampler, this increased computational burden results in drastically increased overall runtime. In turn a small number of particles used to estimate the likelihood will yield a smaller computation time for each iteration,

but at the expense of decreased efficiency due to the variability of the likelihood estimator.

Figure 4.1 shows diagnostic analysis of a MCMC run for the Lotka–Volterra predator prey model using a bootstrap particle filter with differing numbers of particles. The plots make it clear that the number of particles is of crucial consequence to the efficacy of posterior sampling. With a small number of particles, calculation of likelihood estimates is relatively fast but mixing of the chain is poor. A larger number of particles improves the mixing of the chain, but each individual iteration takes longer to evaluate.

Optimal choice of the number of particles in the filter was explored by Pitt *et al.* (2012) who suggest that the number of particles should be chosen such that the variance of the log–likelihood estimates is approximately 1. Doucet *et al.* (2015) showed that the efficiency of the particle MCMC scheme is good when the variance of the log–likelihood is between 0.25 and 2.25. In both articles this is done by considering bounds on the integrated auto–correlation time under the assumption that the chain is at stationarity and that the distribution of the noise in the target is independent of it's location in parameter space. In addition it is assumed that the distribution of the additive noise is Gaussian and that the computation time is inversely proportional to the variance in the estimator of the target. Based on these assumption both articles seek to address optimal conditions for the variance of the target. Pitt *et al.* (2012) use the perhaps non-realistic case of an independence sampler where proposals are made from the desired target. Sherlock *et al.* (2015) use slightly differing assumptions and focus on the joint optimisation of the variance in the estimator of the target and the overall acceptance rate of a chain based on a Gaussian random walk kernel (discussed in section 4.3.2) to derive a slightly higher optimal variance of approximately 3.3. The authors similarly work with the assumption that the distribution of the additive noise in the estimated target is independent of location. They do however note that this assumption is a pragmatic choice rather than a reasonably held belief in practice.

The example chains shown in figure 4.1 were generated by initialising in each case a random walk metropolis pseudo marginal scheme with common Gaussian proposal kernel. Each were initialised at the parameter vector used to generate the data where data is that of the partially observed evolution of the two species in a Lotka Volterra process (introduced in section 2.5.3). For each particle filter the variance of the

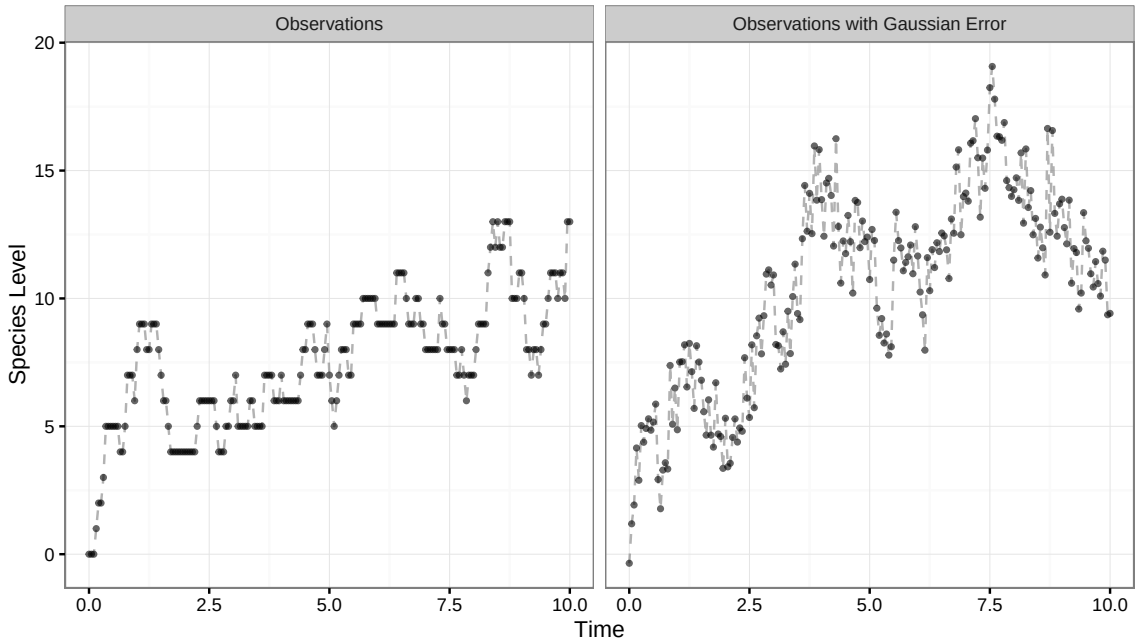


Figure 4.2: Simulated data sets of the Immigration–Death process using Gillespie's direct method (algorithm 1). Left are measurements at a set of discrete time intervals, every 0.05 units of time for a total of 10, which are perfect observations of the system at those times. Right are measurements over the same observation regime, subject to measurement error. The measurement error here was simulated to be a zero mean Gaussian distribution with unit variance. The parameter values chosen for Immigration and Death rate are $\theta_1 = 10.0$ and $\theta_2 = 1.0$ respectively.

estimates of log likelihood was calculated at the true parameter values. When $N = 10$ particles are used, mixing of the chain is very poor. In addition when calculating the variance of the log–likelihood estimates at the true parameter values this particle filter yielded 62% of estimates as negative infinite. This was because all of the weights in step 5 of algorithm 3 in section 3.4.1 were negligible in those particular evaluations of marginal likelihood. Consequently the variance of the estimates of the log target is extremely large when the number of particles is so small. For $N = 100$ and $N = 1000$ particles the filter gave variances of the log target as $\mathrm{Var}\left(\hat{\pi}_{100}(\mathbf{Y} \mid \Theta)\right) \approx 128$ and $\mathrm{Var}\left(\hat{\pi}_{1000}(\mathbf{Y} \mid \Theta)\right) \approx 6$ respectively. The acceptance rates α_{10}, α_{100} and α_{1000} for $N = 10$, $N = 100$ and $N = 1000$, particles were 0.9%, 6.1% and 15% respectively.

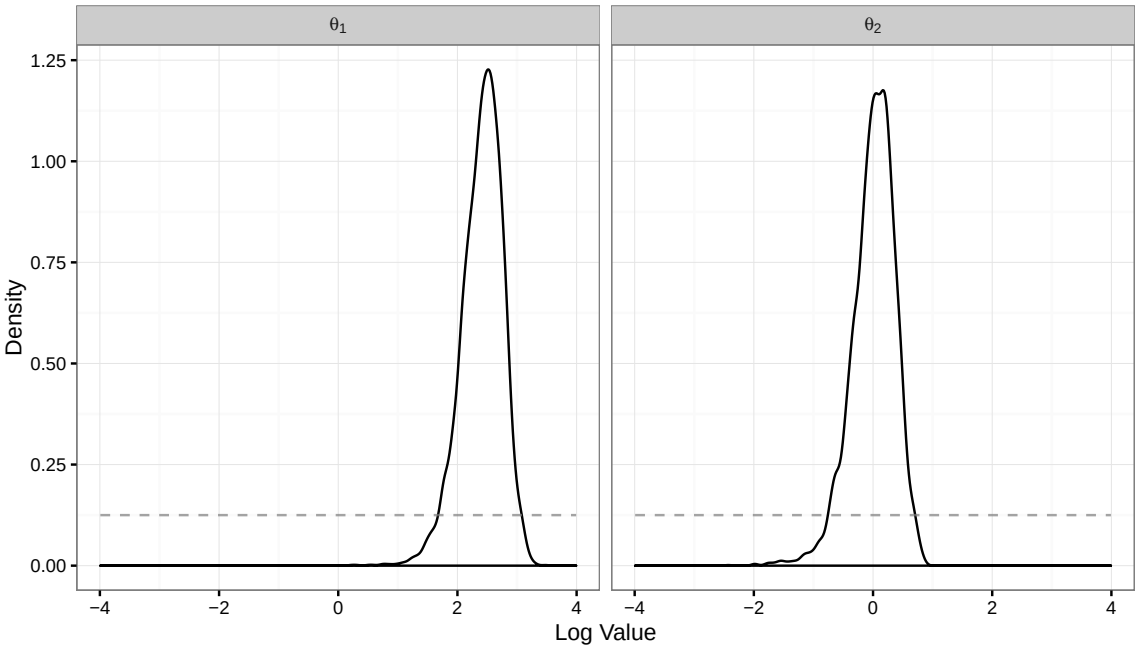


Figure 4.3: Posterior distributions for the noisily observed Immigration–Death data shown in figure 4.2 using the exact likelihood function as defined in equation 4.9. The prior distribution is shown as reference (grey dashed line).

4.4 Numerical examples

4.4.1 Immigration–Death model (ID)

As noted in section 2.5 the immigration–death model is a useful tool for assessing the relative prowess of likelihood–free inference techniques within the stochastic kinetic model setting. This is due to the fact that it is one of the relatively few reaction networks that has an analytic solution under the assumption of mass action kinetics. This allows comparison of any of the techniques that may be useful in more complex systems with a reference true posterior distribution.

Here we demonstrate the use of each of the algorithms discussed in this chapter applied to some artificial data sets from the ID process. The simulated data is shown in figure 4.2.

Prior distributions

For each of posterior inferences that follow a common set of prior distributions for the rate parameters have been employed. That is a vague Uniform prior on the logarithmic scale:

$$\log(\theta_i) \sim \mathcal{U}(-4, 4), \quad i = 1, 2. \tag{4.8}$$

The initial state of the system is assumed to be known, $x_0 = 0$, in each case.

For reference the true posterior distributions, given the noisy observations, are included in figure 4.3. In the case of the perfectly observed data the likelihood is calculated exactly as the product of transition densities

$$\pi(\mathbf{y}_{t_1:t_T} \,|\, \Theta) = \pi(\mathbf{y}_{t_1} \,|\, \Theta) \prod_{i=2}^{T} \pi(\mathbf{y}_{t_i} \,|\, \mathbf{y}_{t_{i-1}}, \Theta), \tag{4.9}$$

where the transition densities can be calculated as in equation 2.15. Where observations are corrupt, subject to a zero mean Gaussian measurement error kernel with variance σ^2 the likelihood function becomes more complicated. Consider a single pair of time points t_1 and t_2. When the process is observed subject to error the true values $\mathbf{x}_{t_1}$ and $\mathbf{x}_{t_2}$ respectively are unobserved latent variables. In order to calculate the transition density of the noisy observations $\mathbf{y}_{t_1}$ to $\mathbf{y}_{t_2}$ these latent variables must be integrated out. Let $\phi(\cdot; \mathbf{x}_t, \sigma^2)$ denote the density of a Gaussian distribution with mean $\mathbf{x}_t$ and variance σ^2, our measurement error kernel. The transition density $\pi(\mathbf{y}_{t_2} \,|\, \mathbf{y}_{t_1}, \Theta)$ can be evaluated as

$$\pi(\mathbf{y}_{t_2} \,|\, \mathbf{y}_{t_1}, \Theta) = \iint \phi(\mathbf{y}_{t_2}; \mathbf{x}_{t_2}, \sigma^2) \pi(\mathbf{x}_{t_2} \,|\, \mathbf{x}_{t_1}, \Theta) \phi(\mathbf{x}_{t_1}; \mathbf{y}_{t_1}, \sigma^2) d\mathbf{x}_{t_1} d\mathbf{x}_{t_2}. \tag{4.10}$$

Expression 4.10 can then be substituted into the factorised likelihood, equation 4.9.

Whilst the integral in equation 4.10 is potentially problematic, due to the space of $\mathbf{X}_t \subset \mathbb{N}$ being positive integers, and the range of the observations being small, the integral can be calculated, for this model, to within negligible error by

$$\pi(\mathbf{y}_{t_2} \,|\, \mathbf{y}_{t_1}, \Theta) \approx \sum_{\mathbf{x}_{t_2}=0}^{\mathbf{X}_{t_2}} \sum_{\mathbf{x}_{t_1}=0}^{\mathbf{X}_{t_1}} \phi(\mathbf{y}_{t_2}; \mathbf{x}_{t_2}, \sigma^2) \pi(\mathbf{x}_{t_2} \,|\, \mathbf{x}_{t_1}, \Theta) \phi(\mathbf{x}_{t_1}; \mathbf{y}_{t_1}, \sigma^2), \tag{4.11}$$

with suitably large $\mathbf{X}_{t_i}$. In practice here $\mathbf{X}_{t_i}$ is taken to be 200, $i = 1, \ldots, T$.

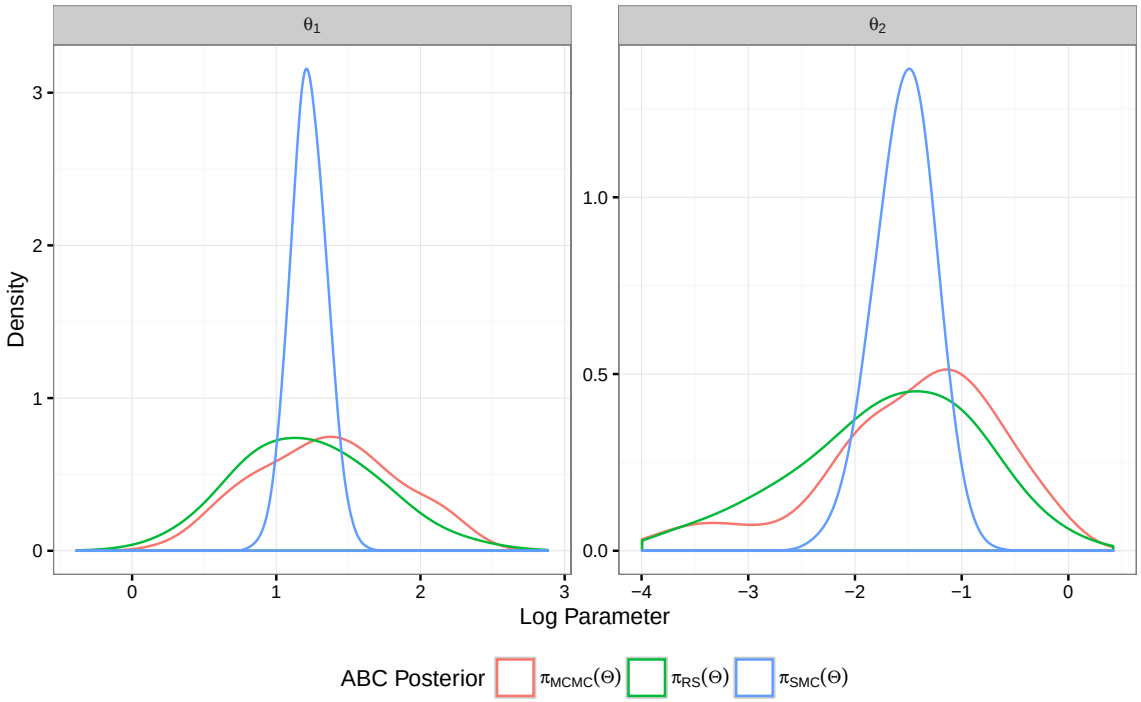


Figure 4.4: Posterior distributions of the immigration–death model rate parameters using each of the ABC algorithms. The rejection sampler and MCMC sampler, yielding π_{RS} and π_{MCMC} respectively, give similar inference of the posterior distribution. The SMC sampler giving posterior π_{SMC} gives a less variable posterior distribution. It should be noted however that in this case the SMC sampler finds the wrong posterior distribution.

Approximate Bayesian Computation

Assessing the ability of ABC techniques to approximate the posterior distribution of interest for the immigration–death model is much easier than other models of this type due to the availability of an analytic likelihood function. Because of this direct comparison between the true posterior distributions of the rate parameters and the approximate ABC posterior distributions can be made. In the following numerical examples the posterior inferences are obtained given observation of the data corrupted by measurement error. This is due to the fact that the bootstrap particle filter requires measurement error to be present in the observations and hence using this data set allows comparisons to be drawn. The distance function $\rho(\cdot)$ is an unweighted Euclidean metric on the full vector of data points. Measurement error in each case is zero mean Gaussian with known variance, $\sigma^2 = 1$.

ABC Rejection Sampler

Figure 4.4 gives the posterior distributions for the rate parameters, denoted $\pi_{RS}(\Theta)$, of the immigration–death model under a simple ABC rejection sampler (algorithm 6). The tolerance ϵ was chosen by first simulating 10,000 distances using the true parameters values and calculating an estimate of the 0.1%-ile of those samples which gave $\epsilon = 37.12$. This choice was made to ensure that the acceptance rate of the rejection sampler was not prohibitively small. Given this choice of ϵ the rejection sampler was run until a sample size of $N = 5000$ was obtained at an acceptance rate of 0.016%. The estimated posterior means are

$$\mathbb{E}_{\pi_{RS}}\left[\log(\theta_1)\right] = 1.22, \qquad \mathbb{E}_{\pi_{RS}}\left[\log(\theta_2)\right] = -1.69.$$

For reference the posterior means when using the analytic likelihood function are

$$\mathbb{E}_{\pi_{TRUE}}\left[\log(\theta_1)\right] = 2.41, \qquad \mathbb{E}_{\pi_{TRUE}}\left[\log(\theta_2)\right] = 0.00.$$

$\mathbb{E}_{\pi_{RS}}\left[\log(\Theta)\right]$ is under estimated for each of the two rate parameters.

ABC MCMC Sampler

Figure 4.4 shows the results of posterior inference using an ABC MCMC scheme, denoted $\pi_{MCMC}(\Theta)$, as described in algorithm 7. The sampler was initialised at the true parameter values, $\log(\Theta) = (2.302, 0)$, and run for a total of 2×10^6 iterations. The final sample was drawn by thinning the resultant chain by a factor of 400 to obtain 5000 samples. It is necessary to choose a value for ϵ prior to execution of the sampling algorithm due to the fact that it plays a crucial role in the acceptance criteria for proposed moves in the Markov Chain. Here ϵ is chosen to be the 0.1%-ile of a simulated distribution of distances given the true model parameter values, $\epsilon = 37.12$. MCMC diagnostic plots are shown in figure 4.5.

Estimated posterior means are

$$\mathbb{E}_{\pi_{MCMC}}\left[\log(\theta_1)\right] = 1.35, \qquad \mathbb{E}_{\pi_{MCMC}}\left[\log(\theta_2)\right] = -1.47.$$

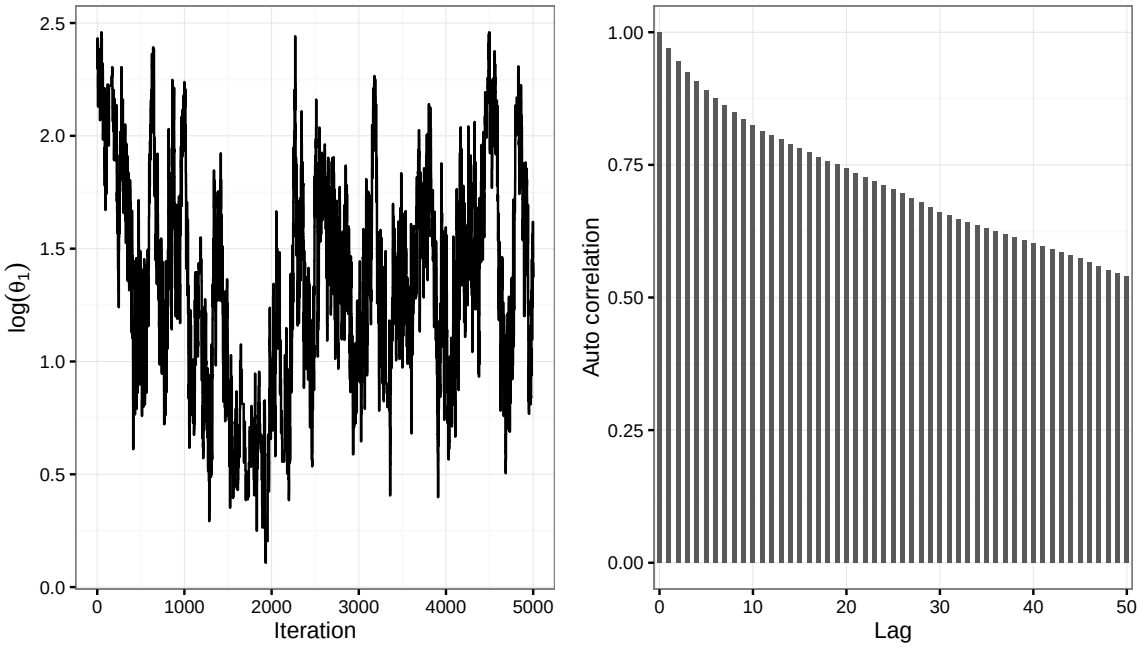


Figure 4.5: Diagnostic plots for the ABC MCMC sampler for the immigration rate parameter θ_1 in the ID process model.

Posterior sampling in this example is poor, with the MCMC chain exhibiting poor mixing. Both a smaller tolerance and a larger random walk proposal variance gave a chain which struggled to move at all, for a large uncorrelated sample here we would ideally like to run the chain for a much greater time. The issues here are symptomatic of common problems with MCMC based ABC sampling and will be discussed in more depth in subsequent chapters.

ABC SMC Sampler

Figure 4.6 and figure 4.4 show the progression of the posterior approximation as the tolerance is decreased and the final posterior distributions respectively. At each given ϵ in the sequence the sampler was run until 5000 samples were obtained. The sequence of tolerances was chosen such that the α%-ile of the calculated distances at stage t was chosen as ϵ_{t+1} where $\alpha = 40\%$. At each stage of the algorithm the optimal proposal kernel was determined as in equation 4.2. The initial tolerance $\epsilon_0 = 62.56$ was chosen as the median of the distribution of distances given the true parameter values, estimated from sample of 10,000. Figure 4.6 shows clear reduction in the variance of the resultant sample as the distribution moves closer to that of the

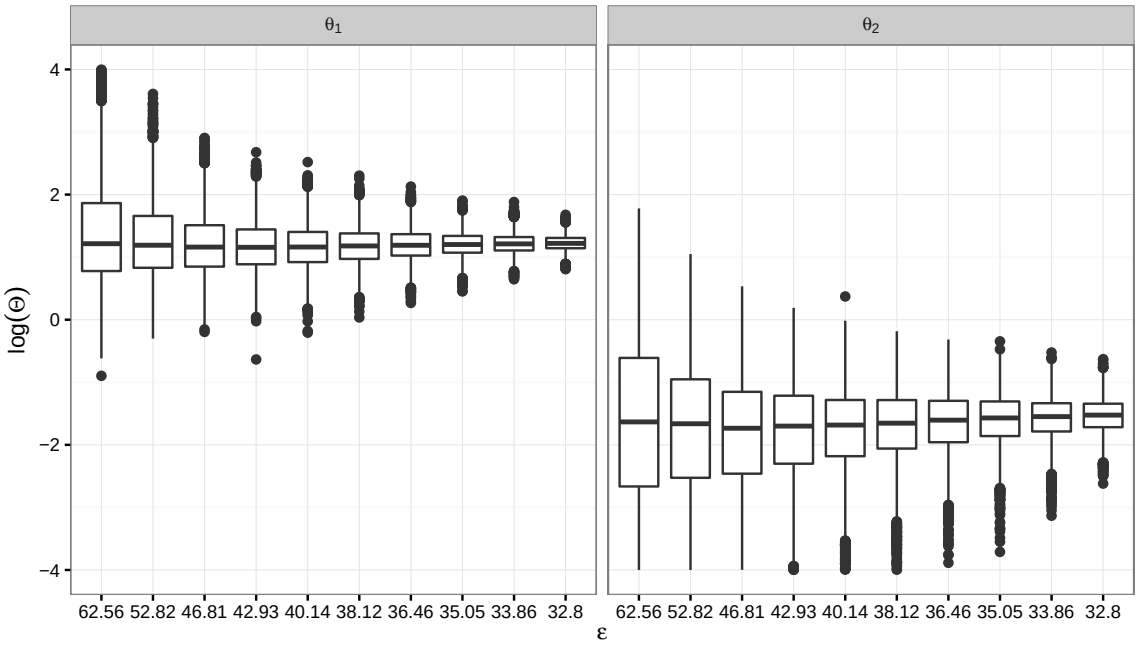
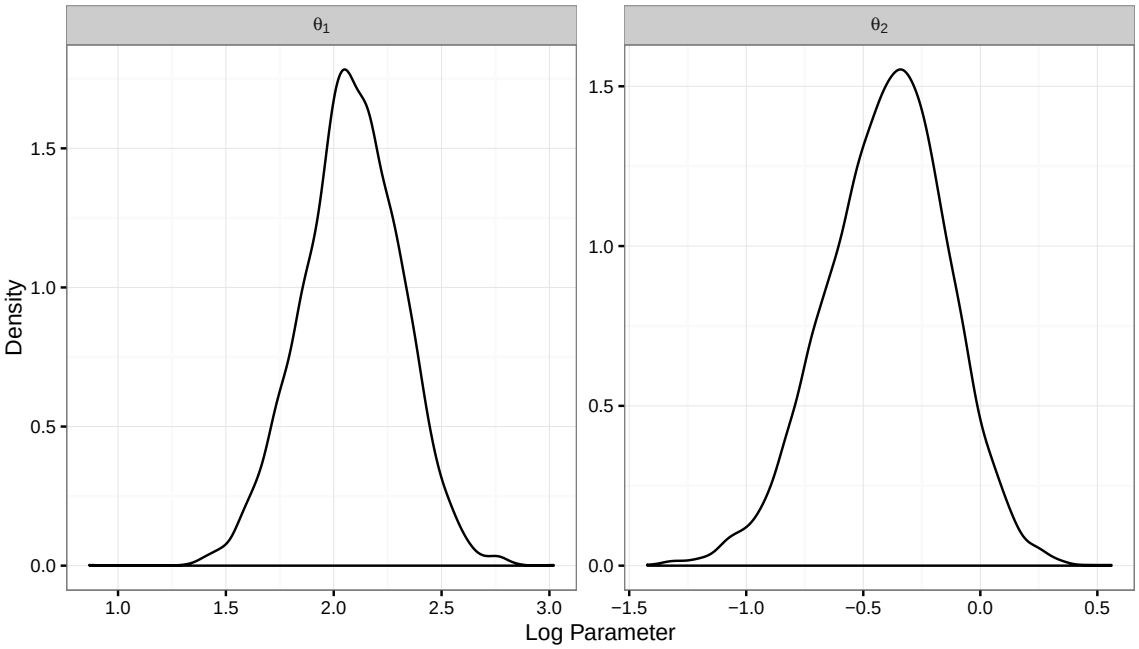


Figure 4.6: Sequence of distributions for the ABC SMC scheme for the Immigration–Death model. As the tolerance decreases the posterior samples get closer to the true posterior distribution. The gradual reduction of ϵ allows sampling from a distribution that is a better approximation to the true posterior distribution.

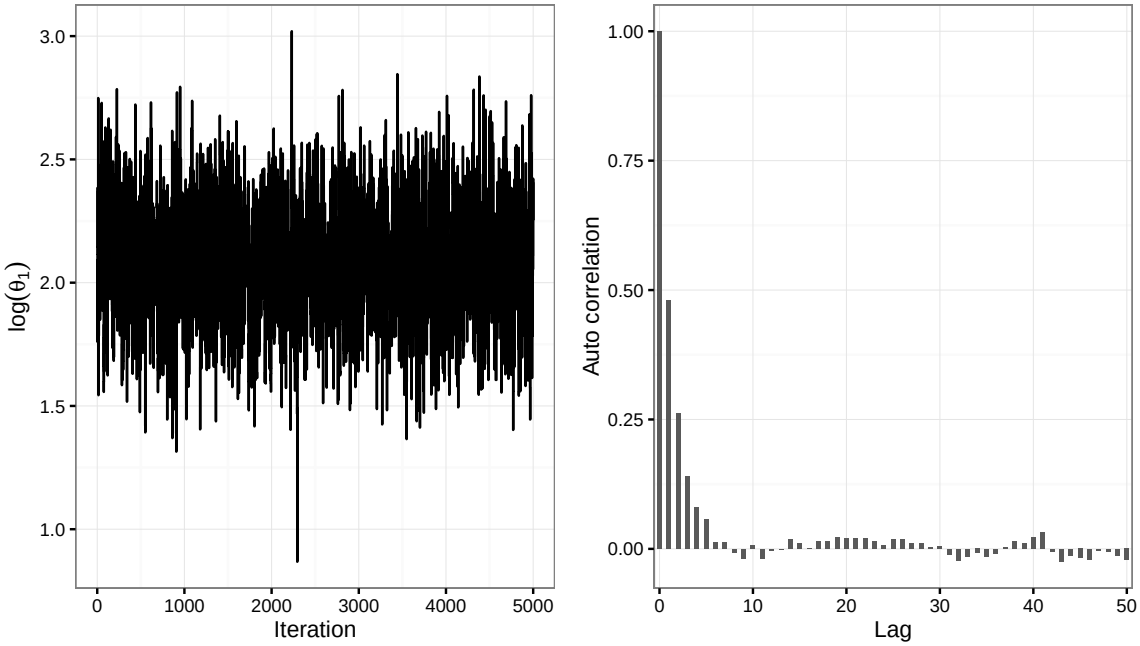
reference true posterior distribution. Posterior expectations of the rate parameters under this sampling scheme are

$$\mathbb{E}_{\pi_{SMC}}\left[\log(\theta_1)\right] = 1.22, \qquad \mathbb{E}_{\pi_{SMC}}\left[\log(\theta_2)\right] = -1.54.$$

One of the advantages that the ABC SMC sampler has over the other two is that due to it's construction, reducing the tolerance is typically more accessible. Because one can start with a higher tolerance, at each iteration proposing new samples constructed from those accepted under the previous tolerance, we typically see higher overall acceptance rates for smaller tolerances. In this example the final tolerance $\epsilon_T = 32.8$ is smaller than in the rejection sampler case, $\epsilon = 37.12$ yet here overall acceptance rates were 0.14% as compared to 0.016%. Further comparison and more in depth investigation of each of the samplers will be the subject of chapter 5.



(a) Posterior distributions of the rate parameters for the immigration–death model using the particle MCMC scheme. The true parameter values, $\log(\Theta) = (2.30, 0)$ are well identified by the posterior distribution.



(b) Diagnostic plots for the pfMCMC sampler yielding posterior distributions shown in figure 4.7a. The trace shows the chain to be well mixing with sample auto-correlations also indicating good posterior sampling.

Particle filter MCMC

Figure 4.7a shows the posterior inference for the two rate parameters of the immigration–death model with samples drawn using a pseudo–marginal MCMC scheme with a

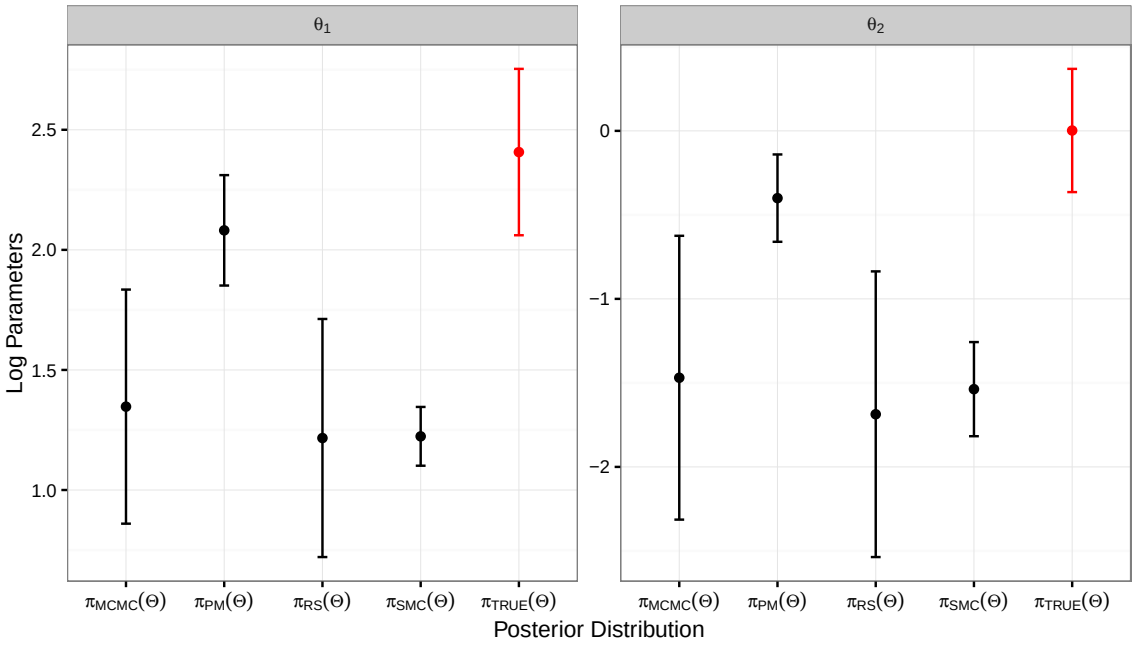


Figure 4.8: Error bar plots for the posterior distributions for the immigration–death rate parameters. The posterior distribution obtained using the analytic likelihood function is shown in red.

bootstrap particle filter being used as the mechanism by which to unbiasedly estimate marginal likelihood as per the description in algorithm 3. In order for computation to be efficient, an intelligent choice of the number of particles in the bootstrap filter must be made. Here the number of particles used to estimate marginal likelihood is chosen such that the variance of log–likelihood estimates at the true parameter values is < 2. This yielded a choice of 100 particles for the filter. The chain was initialised at the true parameter values that lead to the simulated data set. The sampler was run for a total of 500,000 iterations, thinned by a factor of 100 to yield a final sample of size 5000.

Posterior inferences for the immigration–death model

We can examine the posterior inferences for each of the samplers discussed in this chapter for inference of the rate parameters in an immigration–death model. Figure 4.8 shows posterior point estimates of the means of each immigration rate, θ_1, and death rate, θ_2, with error bars showing plus or minus one standard deviation. The reference posterior distribution obtained using the analytic likelihood function

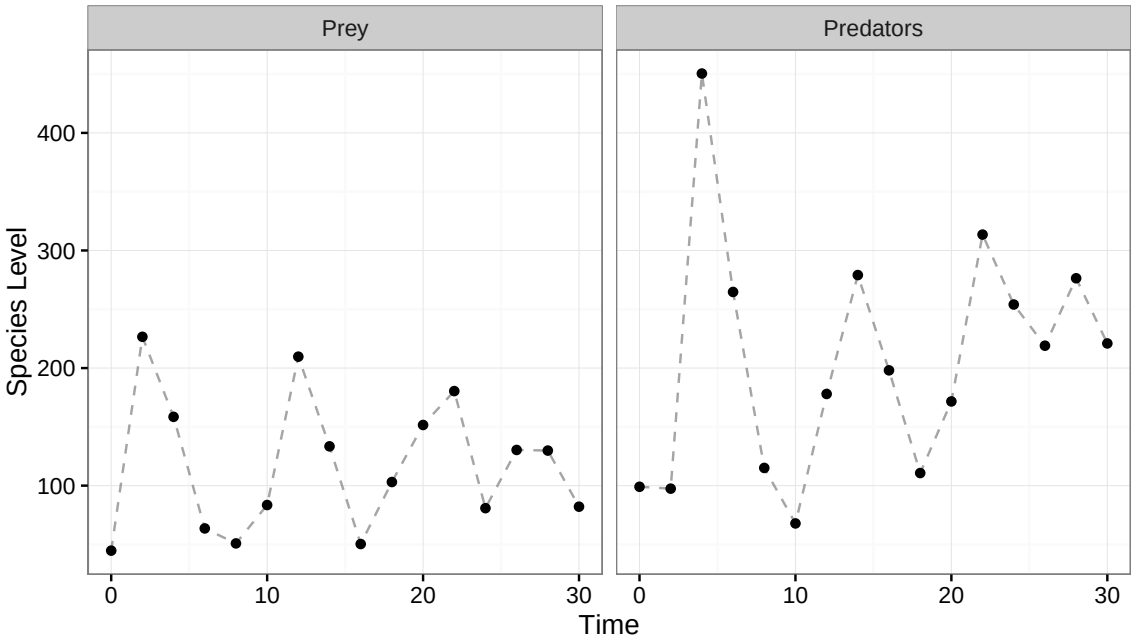


Figure 4.9: A time series of noisily observed species counts of the Lotka–Volterra predator prey system over a regular 2 unit time interval regime for a total of 30 units. This pseudo data has been created using a realisation of the Gillespie algorithm with model rate parameters $\Theta = (1.0, 0.005, 0.6)$. The true counts have then been corrupt with zero mean Gaussian measurement error with standard deviation $\sigma = 10$.

is highlighted in red. Observe that all of the likelihood-free techniques fail to fully identify the support of the reference posterior distribution. In each of the ABC schemes the mean is under estimated, with posterior variability much greater than the reference posterior. The particle filter MCMC has posterior expectation closer to the reference, but over estimates each parameter. The variance is also greater than the reference distribution. Of course each likelihood–free posterior distribution could be improved with greater computational expense. The following chapter (chapter 5) will consider computational efficiency as part of the comparisons between each algorithm.

4.4.2 Lotka–Volterra model (LV)

The Lotka–Volterra predator prey system introduced in section 2.5.3 poses an interesting inference problem due to the intractability of it's the model's likelihood function. This does however present an issue for assessing the performance of different inference schemes as a true posterior distribution is unaccessible. In this section

we present an example of each of the inference tools introduced in this chapter applied to the problem of posterior learning of the underlying rate parameters. The pseudo data to be used for these examples is a time series observed over a window of 30 units with noisy observations of the two species at regular 2 time unit intervals as shown in figure 4.9. The true rate parameter values are $\Theta = (1.0, 0.005, 0.6)$ or $\log(\Theta) = (0, -5.30, -0.51)$, exact species counts at each observation are then corrupted with zero mean Gaussian noise with standard deviation, $\sigma = 10$.

Prior distributions

A common set of prior distributions are used in each example presented here. That is a vague, uniform prior on the log parameter values.

$$\log(\theta_i) = \mathcal{U}(-7, 2), \quad i = 1, 2, 3. \tag{4.12}$$

ABC Rejection Sampler

Figure 4.10 gives the posterior distributions for the rate parameters of the Lotka–Volterra model under a simple ABC rejection sampler (algorithm 6). The posterior distribution, denoted $\pi_{RS}(\Theta)$ was obtained as retaining the first 5000 prior parameter samples that yielded $\rho(\mathbf{y}, \mathbf{y}^*) < \epsilon$ where $\epsilon = 373.90$ was chosen as the 10%-ile of 10,000 simulated distances at the true parameter values. This gave an overall acceptance rate of approximately $10^{-5}\%$.

The density plot of the distribution shows that posterior inference under this scheme is more variable than that obtained using the sequential sampler but similar to the MCMC sampler in this case. Posterior expectations are

$$\mathbb{E}_{\pi_{RS}}[\log(\theta_1)] = 0.13, \qquad \mathbb{E}_{\pi_{RS}}[\log(\theta_2)] = -5.19, \qquad \mathbb{E}_{\pi_{RS}}[\log(\theta_3)] = -0.70,$$

where the true rate parameters are $\Theta = (0, -5.30, -0.51)$.

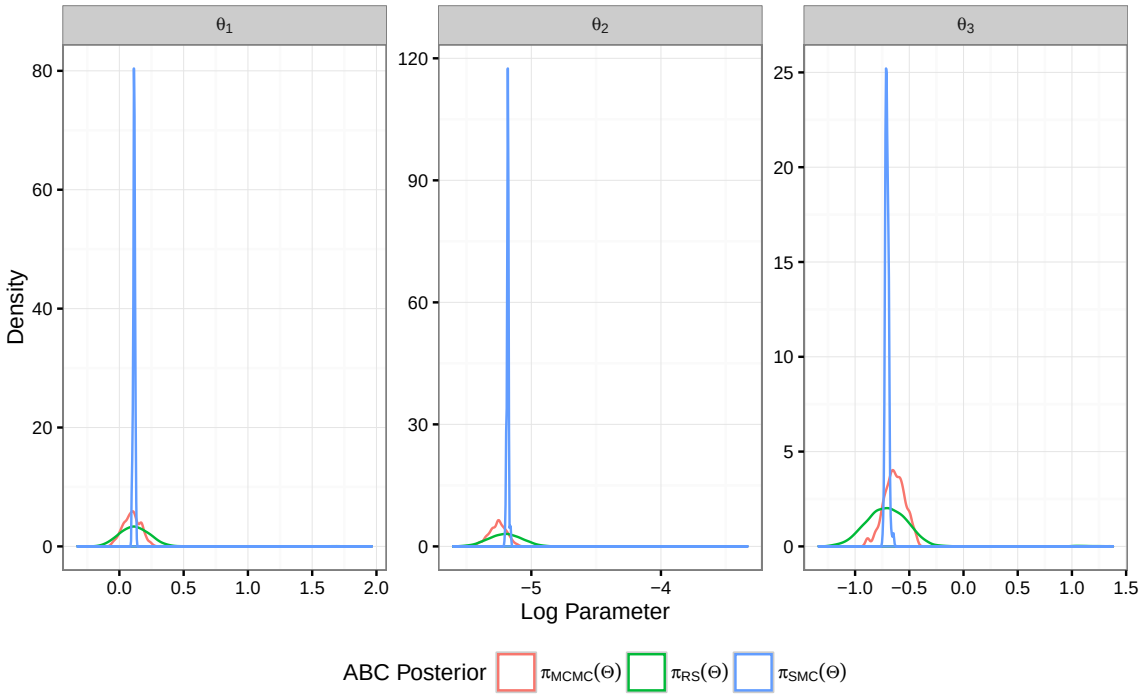


Figure 4.10: Approximate posterior distributions for the three rate parameters of the Lotka–Volterra model using the three difference ABC samplers.

ABC MCMC Sampler

Figure 4.10 shows the results of posterior inference using an ABC MCMC scheme as described in algorithm 7. The sampler was initialised at the true parameter values, $\log(\Theta) = (0, -5.29, -0.51)$, and run for a total of 2×10^6 iterations. The final sample was drawn by thinning the resultant chain by a factor of 400 to obtain 5000 samples yielding the posterior distributions denoted $\pi_{MCMC}(\Theta)$ in figure 4.10. Here ϵ is chosen to be the 0.1%–ile of a simulated distribution of distances given the true model parameter values, $\epsilon = 233.45$. The posterior distributions here show no real improvement over those found using the simple rejection sampler despite having a smaller tolerance value, ϵ. The scheme benefits from the random walk proposal, allowing a much smaller choice of ϵ whilst retaining reasonable acceptance rates than in the rejection sampler. Using the same choice of ϵ in the rejection sampler leads to intolerably small acceptance rates. In the posterior sample $\pi_{RS}(\Theta)$ only 17 proposals yielded $\rho(\mathbf{y}, \mathbf{y}^*) < 233.45$. Posterior expectations using the MCMC based

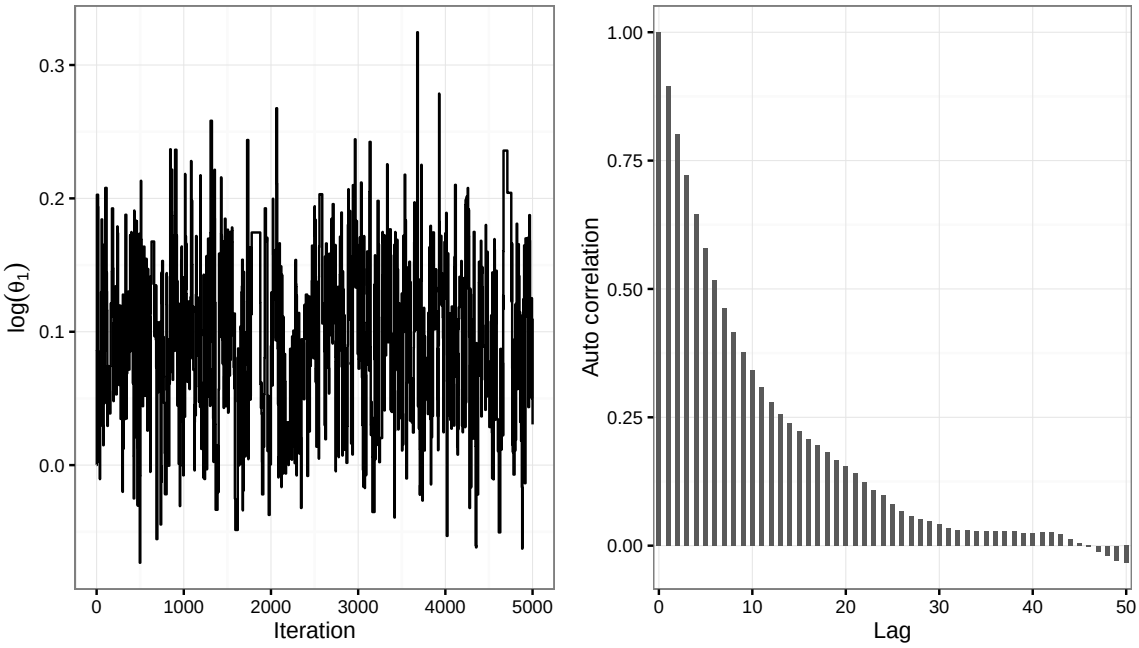


Figure 4.11: MCMC diagnostic plots for the ABC MCMC scheme applied to the Lotka–Volterra predator prey model.

ABC sampling scheme are

$$\mathbb{E}_{\pi_{MCMC}}\left[\log(\theta_1)\right] = 0.10, \quad \mathbb{E}_{\pi_{MCMC}}\left[\log(\theta_2)\right] = -5.26, \qquad \mathbb{E}_{\pi_{MCMC}}\left[\log(\theta_3)\right] = -0.64,$$

similar to those obtained using the rejection sampler. The diagnostic plots in figure 4.11 show the chain mixing well.

ABC SMC Sampler

Figure 4.12 shows the evolution of the approximate posterior distributions as the tolerance decreases in the sequence. The initial tolerance $\epsilon_0 = 525.89$ was chosen as the 40%–ile of a simulated distribution of distances given the true rate parameters. The same regimen was used for the adaptive choice of tolerance as in section 4.4.1. That is given the distances at iteration t, ϵ_{t+1} was chosen as the 40%–ile. In contrast to the immigration–death example, the final tolerance is larger than that used for the MCMC sampler however the density plots in figure 4.10 show a much smaller posterior variance with a sharply peaked density. In the final posterior distribution, $\epsilon = 235.95$, the true parameter values, $\log(\Theta) = (0, -5.30, -0.51)$ are not within the support of the posterior which is not the case for the other ABC based samplers.

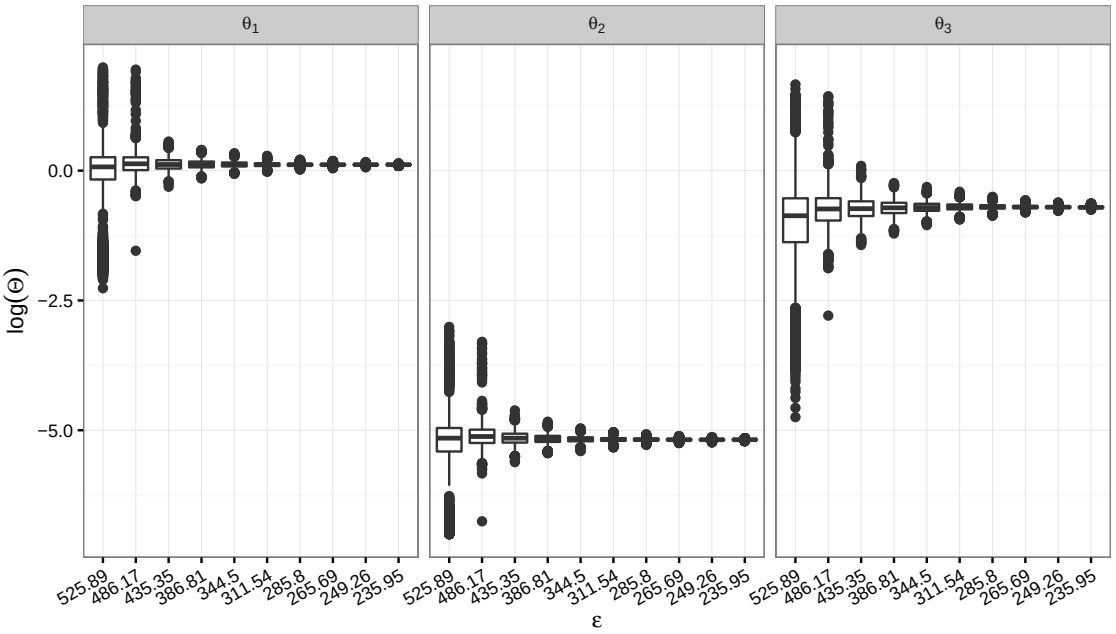


Figure 4.12: Sequence of distributions for the ABC SMC scheme for the Lotka–Volterra model. As the tolerance decreases the posterior samples move about parameter space.

Posterior expectations using the SMC sampler are

$$\mathbb{E}_{\pi_{SMC}}\left[\log(\theta_1)\right] = 0.11, \quad \mathbb{E}_{\pi_{SMC}}\left[\log(\theta_2)\right] = -5.18, \qquad \mathbb{E}_{\pi_{SMC}}\left[\log(\theta_3)\right] = -0.70,$$

Silk *et al.* (2013) showed that under certain conditions the convergence of the posterior distribution in the ABC SMC scheme was not guaranteed. They discuss for values of α that are too large, it is possible that the posterior approximation gets trapped in a local minimum of the metric function surface. The scheme will be examined in greater depth in chapter 5.

Particle filter MCMC

Figure 4.13a shows density plots of 5000 posterior samples obtained using a pseudo–marginal MCMC chain with bootstrap particle filter for the estimation of marginal–likelihood. As with the immigration–death model the number of particles was chosen such that the variance of the log–likelihood estimated at the true parameter values was less than 2. This gave 90 particles as an efficient choice. The final chain was run for a total 500,000 iterations thinned by a factor of 100 to give the final sample. The

diagnostic plots in figure 4.13b show a well mixing chain with low autocorrelation. Posterior expectations were

$$\mathbb{E}_{\pi_{PM}}\left[\log(\theta_1)\right] = 0.07, \quad \mathbb{E}_{\pi_{PM}}\left[\log(\theta_2)\right] = -5.29, \qquad \mathbb{E}_{\pi_{PM}}\left[\log(\theta_3)\right] = -0.51,$$

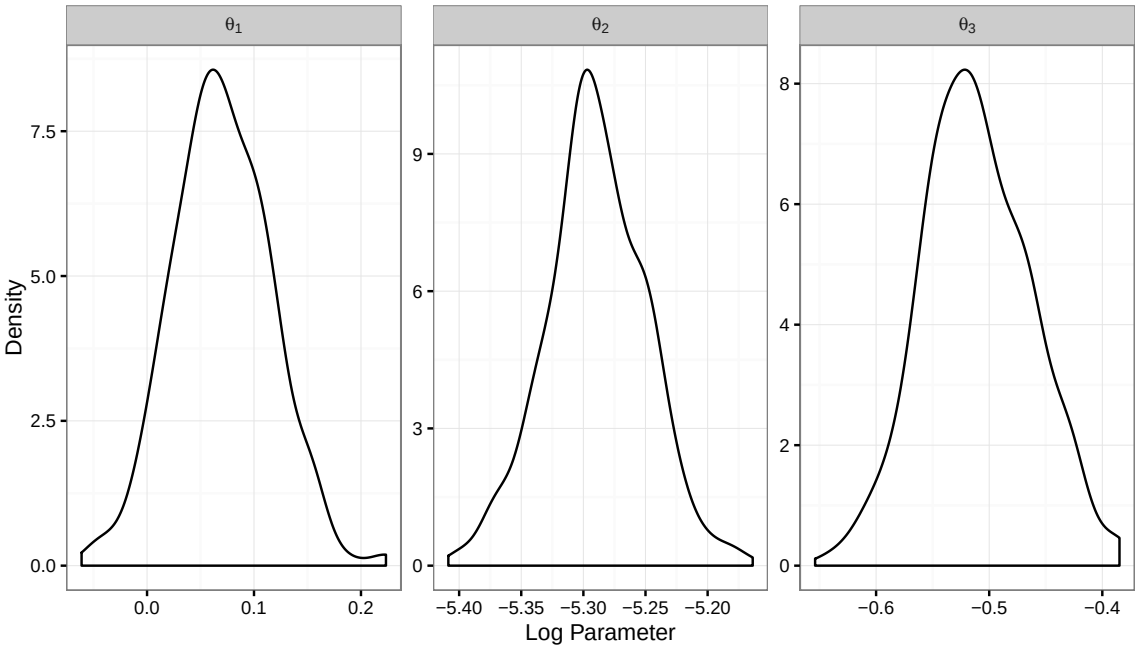
closer to the true parameters that generated the data than the ABC posterior means.

Posterior inferences for the Lotka–Volterra model

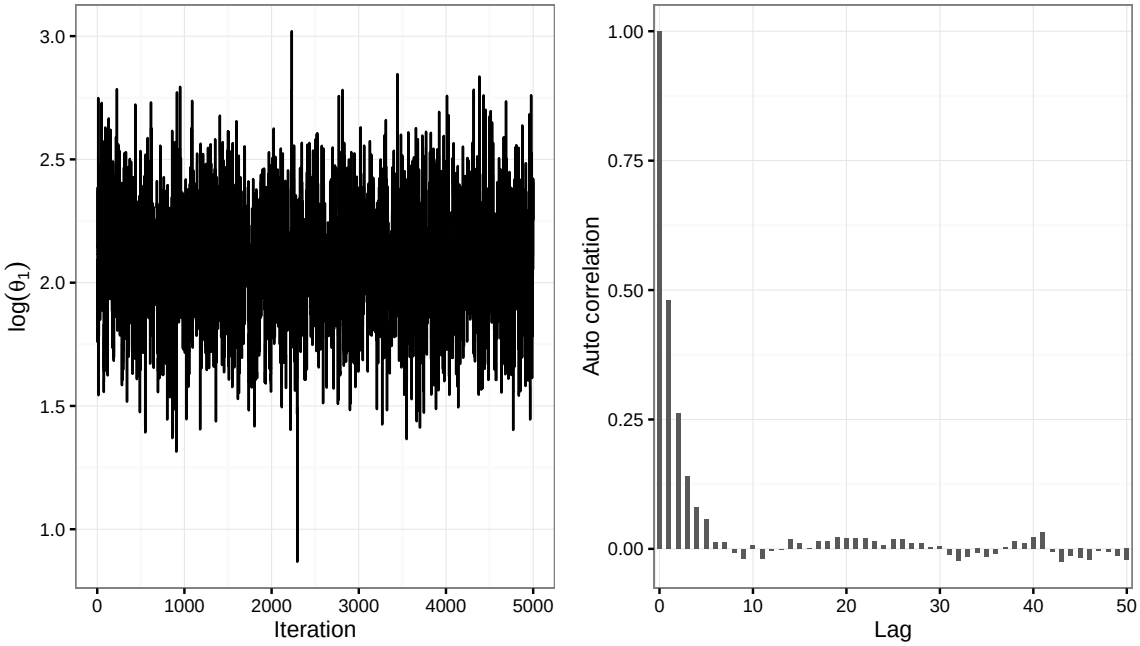
Comparison of the posterior distributions resulting from each of the samplers discussed in this chapter for the Lotka–Volterra model are shown as error bars in figure 4.14. They show point estimates of the mean of each of the rate parameters together with their standard errors. They ABC SMC scheme, $\pi_{SMC}(\Theta)$, yields the smallest posterior variance however the pseudo–marginal MCMC, $\pi_{PM}(\Theta)$ is asymptotically exact as the number of particles used to estimate the likelihood is increased. This would imply that of the ABC samplers the MCMC scheme gives the best posterior inferences.

4.4.3 Conclusion

This chapter introduces a number of techniques which can successfully be applied in the context of inference of rate parameters in stochastic kinetic models given a time course of cell trace data. The performance of each of the algorithms could be improved with the trade off of additional computational expense. In addition there are additional tuning parameters to be considered within the ABC framework which may yield improved posterior inference such as the use of summary statistics for reducing the dimension of the acceptance criteria and more intelligent design of the metric function on those summary statistics. Chapter 5 will focus on a more in depth comparison of the techniques introduced here exposing the relative strengths and weaknesses of the inference schemes.



(a) Posterior distributions of the rate parameters for the Lotka–Volterra model using the particle MCMC scheme. The true parameter values are well identified by the posterior distribution.



(b) Diagnostic plots for the pfMCMC sampler yielding posterior distributions shown in figure 4.13a. The trace shows the chain to be well mixing with sample auto-correlations also indicating good posterior sampling.

Figure 4.13

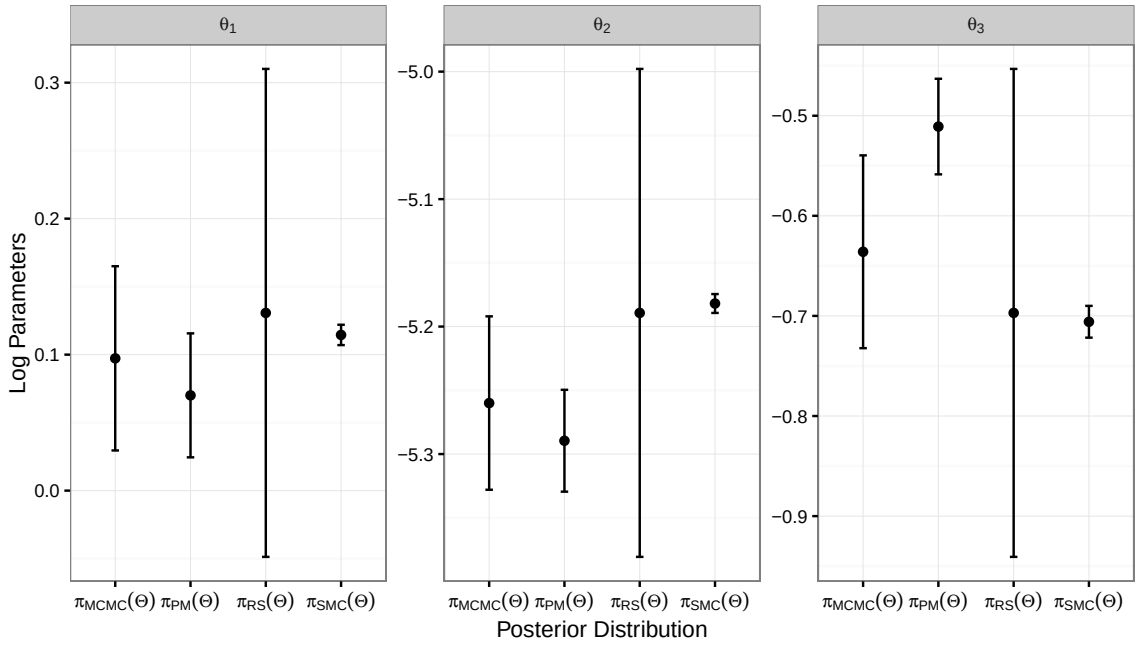


Figure 4.14: Error bar plots for the posterior distributions of the rate parameters for the Lotka–Volterra model.

Chapter 5

Inference for intractable Markov processes

5.1 Introduction

Chapter 4 introduced a number of techniques that can be used to perform inference in the context of an unavailable likelihood function. Exact inference is attainable through the use of pseudo–marginal techniques in the presence of measurement error, whilst approximate inference is available through use of an ABC method. In addition, ABC techniques have a number of tuning options. This chapter will investigate the effect of different sets of summary statistics and metrics within the ABC framework to determine where differences in the rate parameter posterior distributions lie in the context of stochastic kinetic models similar to those introduced in chapter 2. Where appropriate comparisons between particle MCMC and ABC methodology will be made.

5.2 ABC tuning options

ABC techniques are subject to a number of user specified options which determine the definition of the resultant algorithm. These include

- $S(\cdot)$, a function which given observed or simulated data yields a vector of

summary statistics,

- $\rho(\cdot, \cdot)$, the metric function used to define distance between two vectors of summary statistics,
- ϵ, a tolerance which controls how far, as measure by our distance function, a set of simulated data is allowed to be from the observations in order that the parameters that generated them are retained,

which are common to each of the different algorithms. Some of the advancements over the simple rejection sampler (algorithm 6) such as ABC SMC have additional tuning parameters for which optimal choices have been discussed in chapter 4. This section will investigate the effect of different choices for each of these inputs on the posterior distributions that are generated.

5.2.1 Metrics in ABC

In general a metric on a set X is a function $\rho : X \times X \to [0, \infty)$ that satisfies each of the following properties

1. $\rho(x, y) \geq 0$,
2. $\rho(x, y) = 0 \Leftrightarrow x = y$,
3. $\rho(x, y) = \rho(y, x)$,
4. $\rho(x, z) \leq \rho(x, y) + \rho(y, z)$.

Within the context of ABC a metric or distance function $\rho(\cdot, \cdot)$ is used as a means to measure closeness of simulated data to the given observations. Whilst this function often is referred to as a metric it is not implied that $\rho(\cdot, \cdot)$ is a true metric since there is no practical reason that it should satisfy all 4 of the properties above. Jones *et al.* (2015) for example, discuss that there is no requirement for property 3, symmetry to be satisfied.

The metric function does however perform a crucial role within each of the ABC algorithms and will clearly have bearing on the resultant approximate posterior distribution $\pi(\Theta \,|\, \rho(\mathbf{Y}, \mathbf{Y}^*) \leq \epsilon)$. By far the most common choice is the Euclidean metric or L^2 norm

$$\rho_{L^2}(x, y) = \sqrt{\sum_i (x_i - y_i)^2}. \tag{5.1}$$

It is not clear however what the effect of using a different metric would be. In the following sections different metrics applied to a range of different sets of summary statistics will be explored.

5.2.2 Summary statistics

The collection of summary statistics used to represent a set of data on which the distance between two data sets is calculated is a user–defined aspect of an ABC algorithm. The summary statistics themselves can be an essentially arbitrary choice but clearly have bearing on the posterior distribution. The optimal choice would be to have $s(X)$ minimally sufficient but in situations where the likelihood function is intractable sufficient statistics are also unattainable. The aim then is to choose a set of summary statistics that are low in dimension but informative about the data. Choice of summary statistics has been the focus of active research over recent years. Joyce & Marjoram (2008) propose a sequential scheme based on approximate sufficiency. Summary statistics are included if their effect on the posterior distribution is larger than some given threshold. Wegmann *et al.* (2009) suggest Partial Least Squares (PLS) as a tool for choosing summary statistics. The motivation behind this is to seek linear combinations of an original set of summary statistics which are jointly maximally decorrelated with each other and highly correlated with the model parameters Θ. A reduction in dimension is achieved by choosing only the first r components.

One of the issues with attempting to choose a principled set of optimal summary statistics is that the optimal set may well depend on the location of the true, yet unknown, parameters. A criterion for choosing summary statistics that are optimal in a neighborhood close to the true parameters might be preferable. Clearly this neighborhood can not be known in advance. The idea of focusing the choice of summary statistics on some optimal location was the subject of Nunes & Balding (2010) and Fearnhead & Prangle (2012). Nunes & Balding (2010) identify a neighborhood via an algorithm which minimises entropy then choose the summary statistics which minimise mean squared error over a test set. Fearnhead & Prangle (2012) also pro-

pose a two step process based on first defining locality, then optimising within that region. The authors show that for a given loss function there exists an optimal summary statistic. In particular a quadratic loss function yields optimal summary statistic of the posterior mean. Since this information is not available a–prior they develop a heuristic for it's estimation. In a similar two stage approach Aeschbacher *et al.* (2012) propose to choose summary statistics by boosting.

5.3 Framework for comparison of metrics and summary statistics

In order to make meaningful comparisons between different configurations of tuning parameters within an ABC algorithm, it is necessary to construct a rigorous framework for their evaluation. In order to do this here we consider a set of numerical examples using the immigration–death model introduced in section 2.5.1 in the following manner. Each configuration of the ABC tuning parameters to be considered will utilise a common collection of simulated data given random draws from a vague prior distribution. Approaching the comparisons in this way allows precise investigation into the way in which the metrics and summary statistics have bearing on whether or not a proposed parameter vector is accepted into the resultant approximate posterior distribution.

5.3.1 Immigration–death

The immigration death model makes for an ideal model to explore likelihood free posterior inference as it is possible to examine how each of the ABC posteriors fairs with regard to the reference posterior distribution obtained using the analytical likelihood function. Since we have access to the true posterior distribution, subject to Monte Carlo error, it is possible to quantify the dissimilarity between this and any inferred approximate posteriors. One such measure that could be used for this is Kullback–Leibler divergence (Kullback & Leibler, 1951) which is a measure of the difference between two probability distributions. Given a reference "true" distribution P and an approximation to it Q, the divergence of Q from P describes

the amount of information lost by using Q to approximate P

$$D_{KL}(P \,||\, Q) = \int_{\mathbf{X}} p(x) \log \left(\frac{p(x)}{q(x)} \right) dx. \tag{5.2}$$

Since we have no analytic form for the posterior density functions obtained we must estimate the divergence from the obtained samples. Wang *et al.* (2006) developed a k-nearest neighbor approach to the estimation of divergence and showed that as the sample sizes increased the estimator was asymptotically unbiased and consistent. A small value k yields smaller variance at the expense of greater bias but Wang *et al.* (2006) showed that choosing $k = 1$ had good performance for moderate sample sizes and hence we make the same choice here.

5.3.2 Immigration–death inference set up

Common to section 4.4.1 the observations are taken as the same noisily observed time series given true reaction rate parameters $\Theta = (10, 1)$, the simulated data can be seen in figure 4.2 page 49. In keeping with the inference example in section 4.4.1 the prior distributions on the rate parameters are defined in equation 4.8, page 49. This yields an inference problem for which there is vague prior information on the model parameters and a large dimension of observations which proves problematic in the ABC accept/reject stage.

5.3.3 Weighting the Euclidean norm

The Euclidean or L_2 norm is by far the most commonly used metric for calculation of distance between summary statistics in an ABC setting. One of the drawbacks however is that since there is no requirement for summary statistics to be on a common measurement scale one or more statistics could dominate the accept/reject decision in the sampler. One proposed solution, see for example Beaumont *et al.* (2002) to this problem is to instead weight the components of the vector of summary statistics with the inverse of the prior predictive variance of that statistic. By doing so one arrives at a Euclidean metric on a vector of statistics on a common scale,

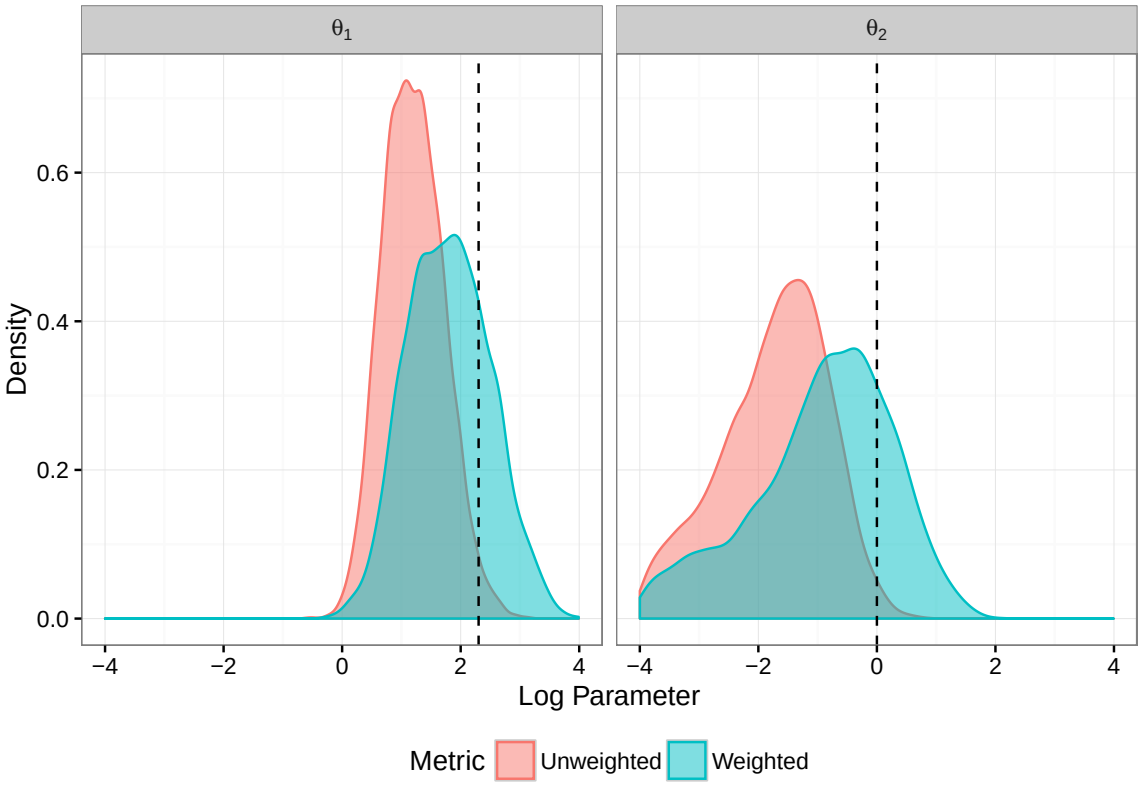


Figure 5.1: Comparison of approximate posteriors obtained using an ABC rejection sampler using a weighted Euclidean metric vs a non-weighted Euclidean metric using the vector of observations as summary statistics.

here denoted $\rho_{L^2_w}(\cdot)$,

$$\rho_{L^2_w}(s(x), s(y)) = \sqrt{\sum_i \left(\frac{s(x)_i - s(y)_i}{w_i} \right)^2}, \tag{5.3}$$

where

$$w = \text{SD}\left(\int_\Theta \pi(s(x) \mid \Theta) \pi(\Theta) d\Theta \right). \tag{5.4}$$

In practice the integral in equation 5.4 is unavailable and hence approximated by taking a sample of parameter vectors from the prior distribution, simulating the summary statistics and taking the standard deviation of each of the statistics. By using the same set of simulated statistics from the model over the prior, and by using a separate prior sample for approximating w we can make a direct comparison between posterior samples under a scheme using an unweighted and weighted metric. Where appropriate the tuning run for calculating prior predictive standard

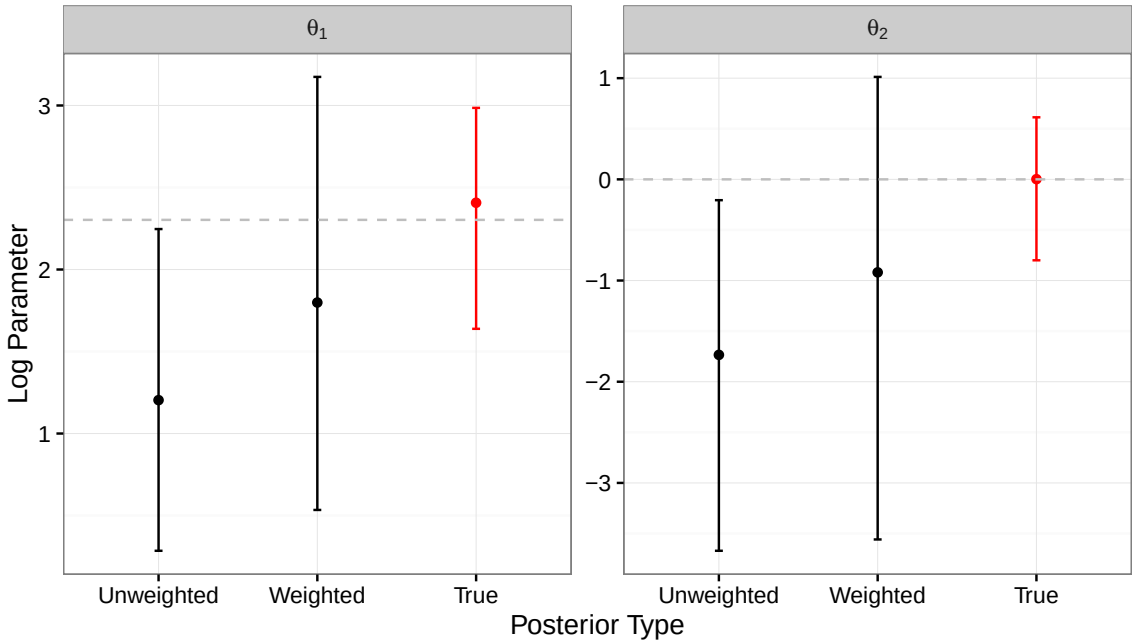


Figure 5.2: Comparison of posterior means and central 95% using the weighted and unweighted Euclidean metric on the vector of observations with respect to the reference posterior using the analytic likelihood function.

deviations uses 100,000 draws from the prior.

Figure 5.1 shows posterior inferences given the same set of simulations from ten million random draws from the prior distribution. In each case the vector of summary statistics is taken to be the raw observations, blue shows posterior samples given distances calculated using a weighted Euclidean metric, $\pi(\Theta \,|\, \rho_{L^2_w}(s(x), s(y)) < \epsilon_{L^2_w})$ whereas red are distances on unscaled summary statistics, $\pi(\Theta \,|\, \rho_{L^2}(s(x), s(y)) < \epsilon_{L^2})$. For each distribution the 5000 samples with smallest distance were retained, it is clear that weighting the metric has a marked effect on the posterior distribution. By using the exact same simulations in each ABC run we can ascertain that of those parameters kept when using the unweighted Euclidean metric, 6.5% are also present in the posterior distribution obtained using the weighted metric.

Figure 5.2 shows error bars for the posterior central 95% given the weighted and unweighted Euclidean metric against the reference true posterior distribution. It is clear that using a tuning run to estimate the standard deviation of the summary statistics over the prior distribution and using this to weight the summary statistics in the metric, putting them on a common scale, improves the posterior means. It, in this instance however, seems to have little effect on the posterior variance. Assessing

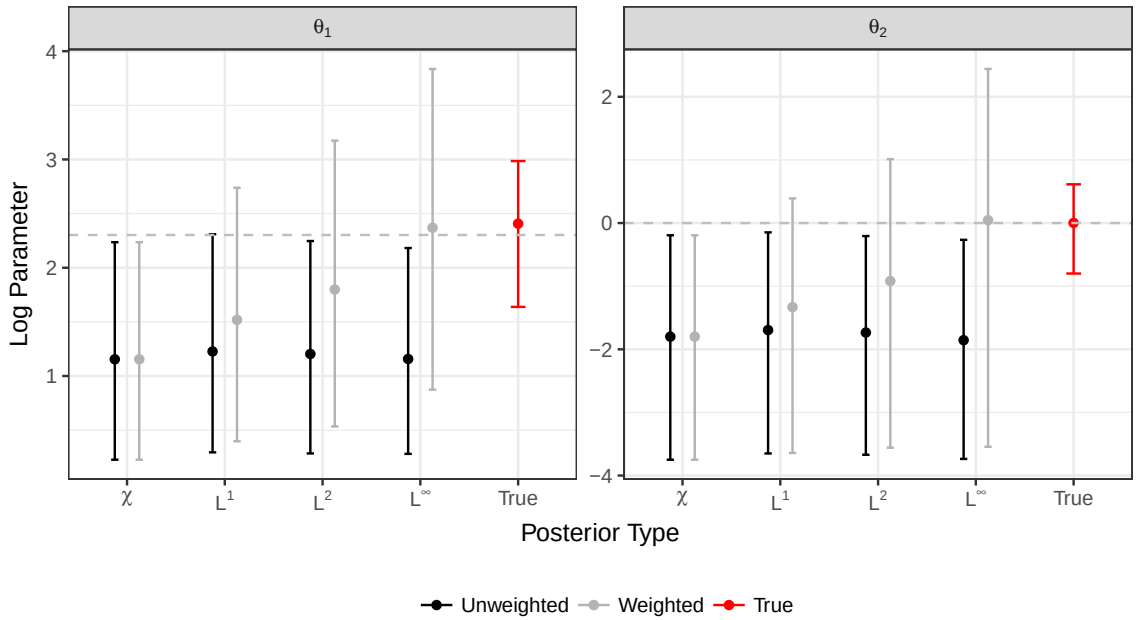


Figure 5.3: ABC posterior distributions for the immigration death model rate parameters using different metric functions to measure distance between simulated and observed summary statistics, $S(\mathbf{Y}) = Y$. Error bars represent the central 95% of the marginal distributions of the parameters. Note that for the χ^2 metric, no weighted version is available, the additional error bar on the graph was to retain common aesthetics across the x axis.

the closeness of the two approximate posteriors to the truth via Kullback–Leibler divergence yields a divergence of 3.08 for the unweighted Euclidean metric and 1.00 for the weighted metric.

Other metrics

There appears to be little literature that directly compares the effect of different metrics. To our knowledge the only published article that does so is McKinley *et al.* (2009). Their example consists of temporal observations for an epidemic model. In a similar experiment we consider here different metrics on the raw observations for the immigration death model. First we note that a general L^p norm is defined as

$$L^p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}} \tag{5.5}$$

Four metrics were considered, the commonly used Euclidean or L^2 norm, the L^1 norm or sum of absolute differences as was considered in McKinley *et al.* (2009), the

L^∞, or maximum, norm (equation 5.6),

$$L^\infty = \operatorname*{Max}_i \{|x_i - y_i|\}, \tag{5.6}$$

and a chi–squared metric (equation 5.7),

$$\rho_{\chi^2}(x, y) = \sum_i \frac{(x_i - y_i)^2}{y_i}. \tag{5.7}$$

Each of the L^p norms can also be weighted, the χ^2 metric has no weighted alternative, and so were also considered with weights equivalent to the inverse of the prior predictive standard deviations. Figure 5.3 shows the central 95% for the marginal posterior distributions of the rate parameters of the immigration death model for each of the 4 considered metrics both weighted and unweighted with comparison to the reference true posterior distribution. Interestingly the effect of weighting the metric is different in each case. With an unweighted metric all of the posterior distributions are very similar. The L^∞ norm shows the biggest change when weights are used and gives point estimates for the posterior expectation closest to those obtained from the true posterior distribution. However the posterior variances are higher than those obtained for the other metrics.

It should be noted that in terms of Kullback–Leibler divergence from the reference distribution the posterior obtained using the weighted Euclidean norm performed best. In comparison to the results of McKinley *et al.* (2009) it was similarly found that where the metrics are not weighted the estimates of posterior expectations are similar for each. The authors of that paper do not consider weighting of the metrics and so no such comparison can be made. They did however find favourable the χ^2 metric which in this experiment performed worst both in terms of both Kullback Leibler divergence from the true distribution and point estimates of posterior means. In addition there is no natural way to incorporate variability of the summary statistics under the prior distribution.

Regression correction

Beaumont *et al.* (2002) proposed an improvement to the ABC posterior through a regression adjustment to weaken the effect of the discrepancy between simulated

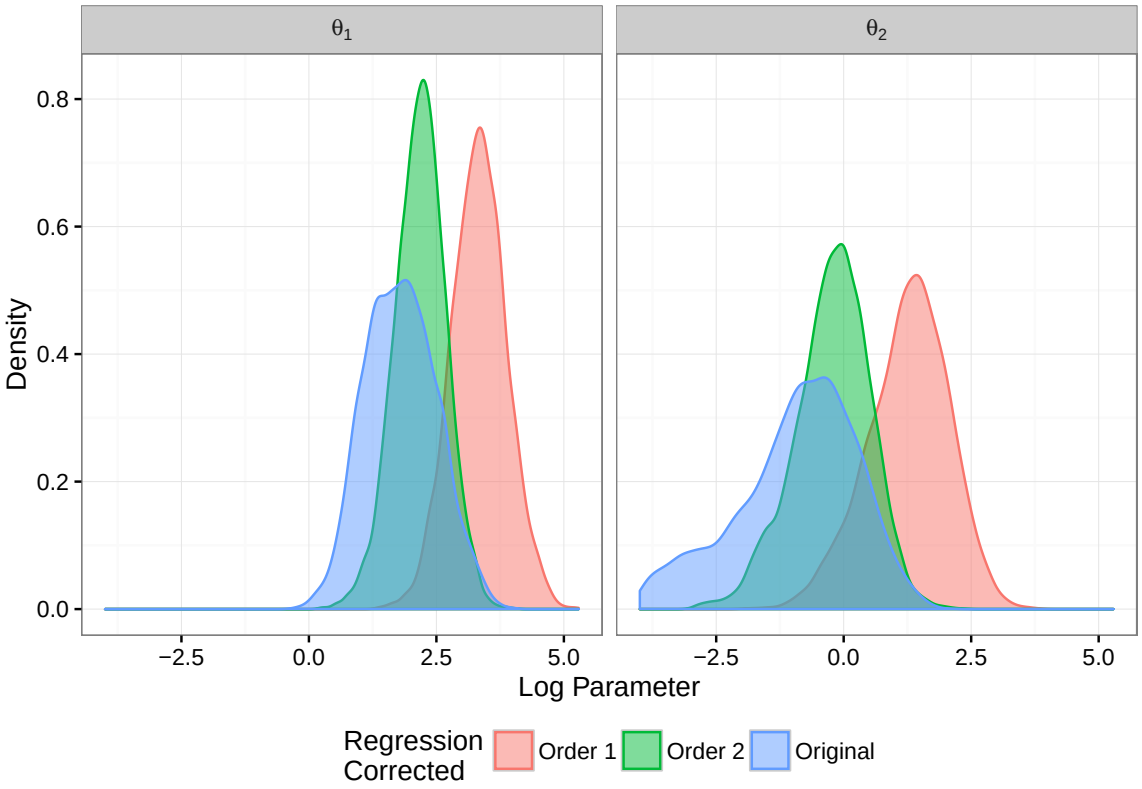


Figure 5.4: Densities for the original, order 1 and order 2 polynomial regression corrected posterior distributions. It is clear that making a regression adjustment has a substantial effect on the inferred distribution. The mode of the distribution shifts significantly and the corrected posterior distributions yield lower posterior variance.

and observed summary statistics. Their recommendation is to use a local linear regression in the vicinity of the observed summary statistics $\mathbf{s} = S(\mathbf{Y})$. Given the linear regression model

$$\Theta_i = \alpha + (\mathbf{s}_i^* - \mathbf{s})^T \beta + \epsilon_i \tag{5.8}$$

the estimates of α and β are found by minimising

$$\sum \left(\Theta_i - \alpha - (\mathbf{s}_i^* - \mathbf{s})^T \beta\right)^2 K_\epsilon \left(\|\mathbf{s}_i^* - \mathbf{s}\|\right) \tag{5.9}$$

where $K_\epsilon(\cdot)$ is an Epanechnikov kernel,

$$K_\epsilon(t) = \begin{cases} \frac{3}{2\epsilon}\left(1 - \left(\frac{t}{\epsilon}\right)^2\right), & t \leq \epsilon \\ 0 & t > \epsilon. \end{cases} \tag{5.10}$$

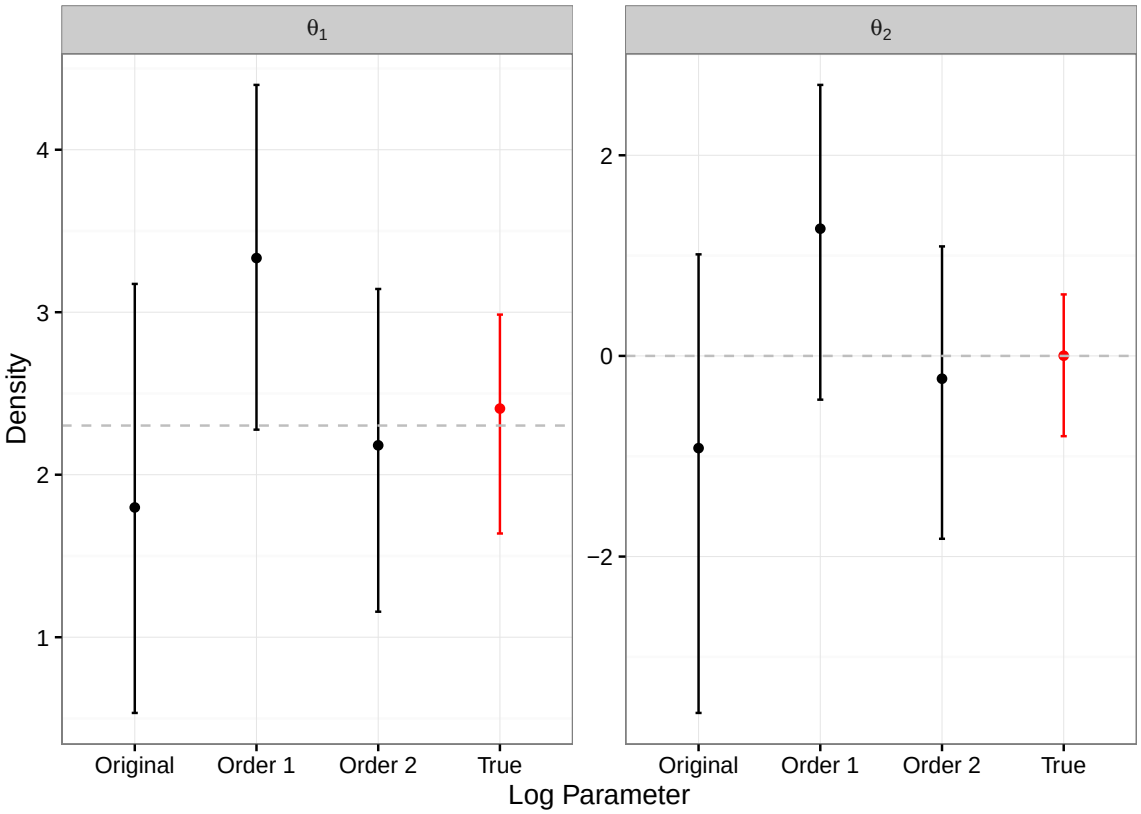


Figure 5.5: Error bars showing the mean and central 95% for each of the uncorrected and regression corrected ABC posterior distributions. The reference true posterior distribution is shown in red. It is clear that the order 2 polynomial regression corrected posterior distribution is much closer to the truth than either the uncorrected or order 1 polynomial corrected posterior.

The corrected parameter values Θ^* are then found as

$$\Theta_i^* = \Theta_i - (\mathbf{s}_i^* - \mathbf{s})^T \hat{\beta}. \tag{5.11}$$

Here we apply the regression correction of Beaumont *et al.* (2002) to the posterior sample using the weighted Euclidean metric from section 5.3.3. Two such regression adjustments are made, first we fit an order 1 polynomial to the $(\mathbf{s}_i - \mathbf{s})$ terms, secondly an order 2 polynomial. Figure 5.4 shows the posterior densities of the original and regression adjusted posterior distributions. It is clear that the inferred distribution is substantially different for each with the corrected distributions exhibiting a different location of the mode and reduced variance. Figure 5.5 compares the means and standard deviations of the 3 ABC posteriors with the reference distribution that

uses the analytic likelihood function. Using an order 2 polynomial on the differences between simulated and observed values on the yields an approximate distribution that is much closer to the true posterior. On examination of the Kullback–Leibler divergences from the reference posterior we find that the order 1 polynomial regression correction gives a distribution further away than no correction at all with a divergence of 2.09 compared to that of 1.00 in the uncorrected case. Fitting a second order polynomial regression model and using that to correct the samples yields a divergence of 0.51.

5.3.4 Dimension reduction of summary statistics

It has long been acknowledged that ABC methodology suffers from the curse of dimensionality. The example in section 5.3.3 examines the effect of weighting the raw observations on the inferred posterior distribution. This was a vector of 201 observations to be considered at the accept reject stage of the sampler. In this section we examine the effect of reducing the dimension of the problem through use of summary statistics. Given a set of summary statistics that consists of the mean, $\bar{y}$, standard deviation, s_y, lag 1, 2 and 3 auto–correlations, where we denote a lag h autocorrelation $a_{y,h}$ and lag 2 and 3 partial auto–correlations, denoted here $p_{y,h}$, we consider various subsets of these. Since we are considering only a small number of summary statistics here it is feasible to check every possible non empty subset for this example. In each case we can also consider the inclusion of weights in the Euclidean metric and regression correction for each inferred posterior. With 127 possible non empty subsets which can be weighted or unweighted by prior predictive standard deviation each of which can be uncorrected or regression corrected using first and second order polynomials there are a total of 762 posterior distributions to consider. We denote posterior distributions that are uncorrected $\pi_\epsilon(\Theta)$, first order polynomial regression corrected $\pi_\epsilon(\Theta')$ and second order polynomial corrected $\pi_\epsilon(\Theta'')$ for the remainder of this section. Conditional notation on $\rho(S(\mathbf{Y}), S(\mathbf{Y}^*))$ has been omitted and is made clear from context. Further, in the following results section reference is made to the true posterior distribution, as obtained using the analytic likelihood function, is denoted $\pi(\Theta)$ and to the best posterior obtained using the full set of observations $\pi_\epsilon(\Theta^{obs})$.

Results

Of the collection of approximate posterior distributions that came from the unweighted Euclidean metrics, $\rho_{L^2}(\cdot)$, both uncorrected and regression corrected, we found that none were closer to $\pi(\Theta)$ as measured by Kullback–Leibler divergence than $\pi_\epsilon(\Theta^{obs})$, the best we got when using all of the raw observations. When using the weighted Euclidean metric, $\rho_{L^2_w}(\cdot)$, of the 381 possible posterior distributions including uncorrected and first and second order regression corrections only 4 gave a Kullback–Leibler divergence smaller than $\pi_\epsilon(\Theta^{obs})$, 3 of which came from the same subset of summary statistics, $S(\mathbf{Y}) = (s_y, a_{y,1})$. The best posterior distribution came from using $S(\mathbf{Y}) = (s_y, a_{y,1})$, using distance function $\rho_{L^2_w}(\cdot)$, with second order polynomial regression correction of the samples retained in the rejection sampler yielding a divergence of $D_{KL}(\pi(\Theta) \,||\, \pi_\epsilon(\Theta'')) = 0.44$, this compares to 0.51 for the divergence of $\pi(\Theta^{obs})$ from $\pi(\Theta)$.

Figure 5.6 shows error bars of the mean and middle 95% of the marginal distributions for the rate parameters in each of the distributions that yielded a Kullback–Leibler divergence smaller than that obtained using the full vector of raw observations. The exact prescription for each of the distributions shown in figure 5.6 is:

- A - The reference true posterior, $\pi(\Theta)$, obtained using the analytic likelihood function,
- B - ABC posterior from the previous section $\pi_\epsilon(\Theta^{obs})$,
- C - $\pi_\epsilon(\Theta'')$ for $S(\mathbf{Y}) = (\bar{y}, s_y, a_{y,1}, p_{y,3})$, distance function $\rho_{L^2_w}(\cdot)$,
- D - $\pi_\epsilon(\Theta)$ for $S(\mathbf{Y}) = (s_y, a_{y,1})$, distance function $\rho_{L^2_w}(\cdot)$,
- E - $\pi_\epsilon(\Theta')$ for $S(\mathbf{Y}) = (s_y, a_{y,1})$, distance function $\rho_{L^2_w}(\cdot)$
- F - $\pi_\epsilon(\Theta'')$ for $S(\mathbf{Y}) = (s_y, a_{y,1})$, distance function $\rho_{L^2_w}(\cdot)$.

The distributions B-F in figure 5.6 are arranged in order of descending Kullback–Leibler divergence. Interestingly point estimates for the means of each rate parameter are closest using the full vector of observations as the set of summary statistics despite the other distributions having smaller divergence from the truth. It is clear from this that using the raw observations is certainly competitive with using summary statistics from the set considered. This is perhaps surprising since the best posterior distribution in terms of KL divergence has summary statistics of dimen-

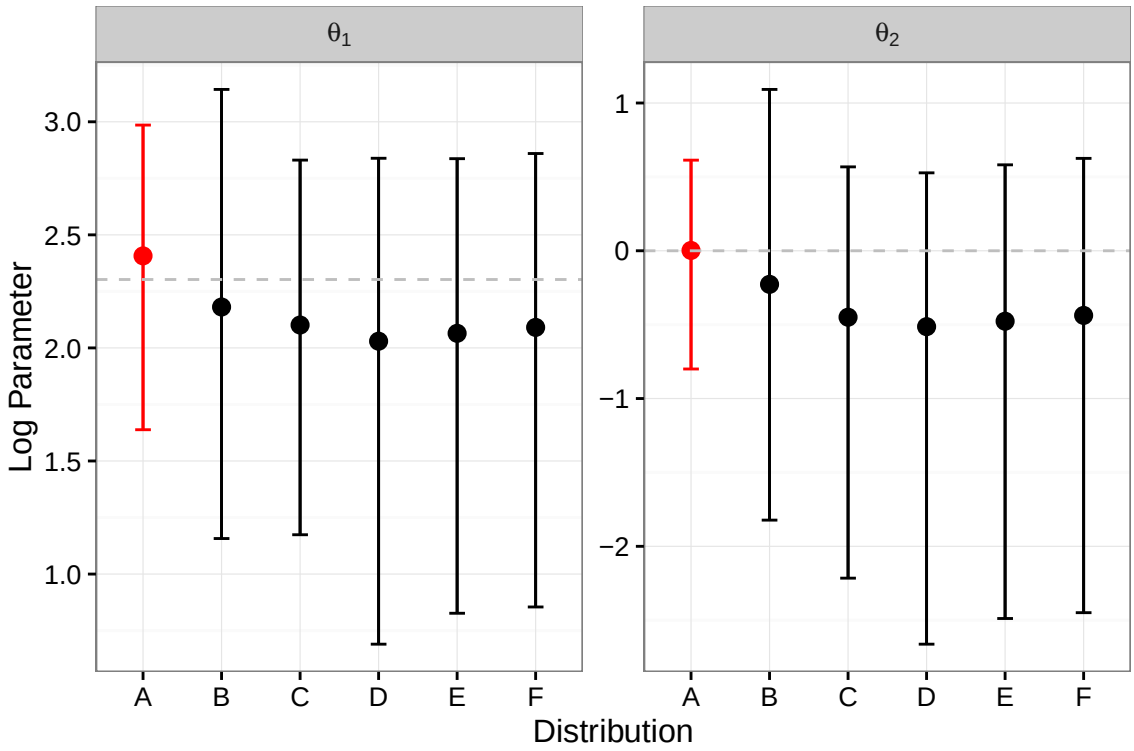


Figure 5.6: Examining posterior inferences using different subsets of summary statistics for the immigration–death model. **A** is the true distribution for reference. **B** is obtained using the full set of raw observations, distances calculated using a weighted Euclidean metric with second order polynomial regression correction. **C** the second order regression corrected posterior using mean, standard deviation, lag 1 autocorrelation and lag 3 partial autocorrelation. **D**, **E** and **F** are uncorrected, order 1 and order 2 polynomial regression corrected distributions respectively using standard deviation and lag 1 autocorrelation as summary statistics.

sion 2, compared to the full vector of raw observation which has dimension 201. Despite a 99% reduction in the dimension of the problem there is no meaningful improvement in the posterior inferences, at least on consideration of the summary statistics that were included.

Considering all possible subsets of the chosen summary statistics, with weighted or unweighted Euclidean metric and applying the regression corrections to each allows us to examine whether improvements that were observed using the full vector of observations are consistent here. In section 5.3.3 it was found that weighting the observations by the inverse of their prior predictive standard deviation improved the posterior inference. Over the collection of 127 subsets of summary statistics weighting them yielded an improvement in only 31.5% of cases. In addition we can make

similar statements for the regression correction procedure. Over the 254 posterior distributions obtained using subsets of summary statistics, a first order polynomial regression correction improved the inference in 47.6% of cases, among the weighted posteriors this was true in 61.4% of cases. In other words more often than not, when weighting the summary statistics in the metric using a regression correction from a fitted single order polynomial makes the posterior distribution better. A second order polynomial regression correction was better than no correction in 51.6% of cases but is better than a first order polynomial correction in 73.2% of cases. Amongst the weighted subsets these percentages are 63.8% and 77.2%. What this is telling us is that weighting the summary statistics in this way, and the regression correction of Beaumont *et al.* (2002) do not consistently improve the posterior inferences and that whether an improvement is found or not is unintuitive. It is also the case that if we were to order the subsets of summary statistics by the divergence of the resultant posterior from the truth, that the order is not retained after regression correction. Whilst in this example the best subset prior to regression adjustment was still the best after correction it is highly possible that this is not always true.

Minimum Entropy

In practical examples in which the true posterior distribution is unavailable one must find another method by which to choose a best subset of summary statistics. Nunes & Balding (2010) define a method to do this using a minimum entropy approach. Entropy of a distribution, (Shannon, 1948), is a measure of information where high entropy corresponds to low information and vice versa. By considering subsets of the available summary statistics and evaluating the entropy of the retained posterior for each they aim to find the subset of summary statistics which provides the most information, denoted $S_{ME}(\cdot)$. They estimate the entropy of the distribution using a k-nearest neighbour approach and we follow that approach here.

As was the case in the previous results section we consider all possible subsets of summary statistics both weighted and unweighted. Of the unweighted summary statistics the subset that would be chosen by the criterion of minimum entropy is $S_{ME}(\cdot) = (\bar{y}, s_y)$. According to the previous analysis which measured the divergence from the truth of all possible subsets this particular set of summary statistics was ranked 38th out of the unweighted 127 subsets. If we consider the weighted subsets

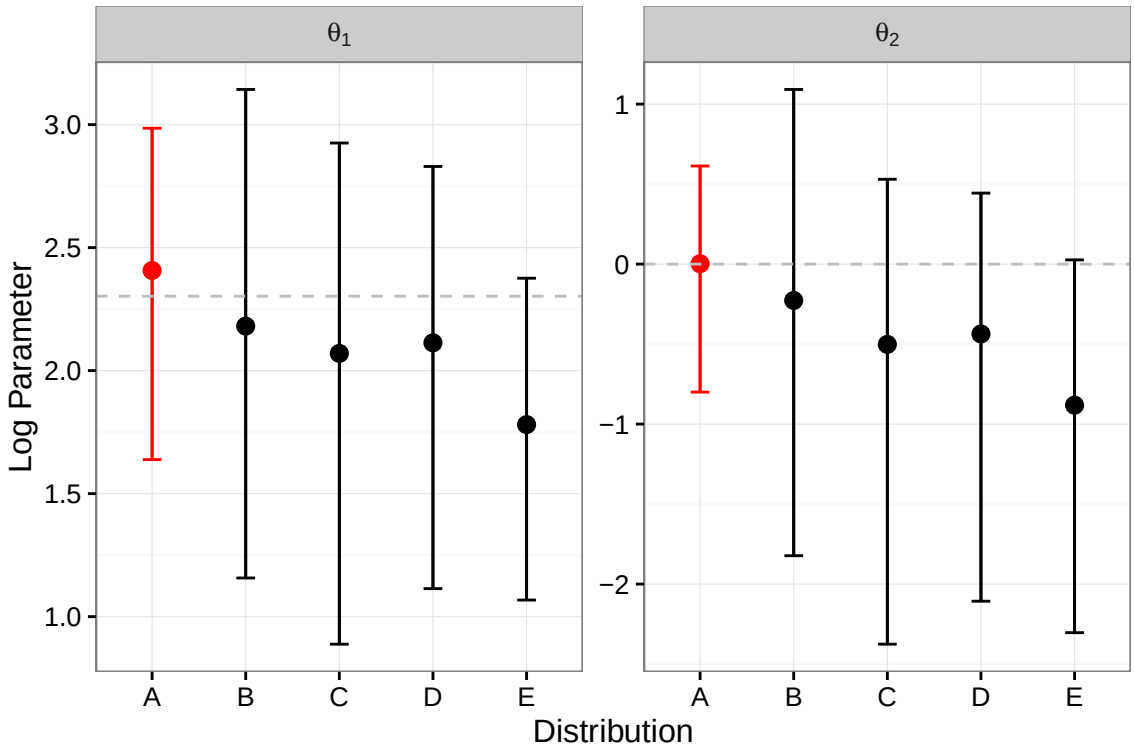


Figure 5.7: Examination of the minimum entropy approach to best subset selection for summary statistics. **A** is the reference true distribution. **B** is the best posterior obtained without any dimension reduction. **C** represents the posterior by using $S_{ME}(\cdot)$ on consideration of non regression corrected posteriors. **D** is the posterior sample **C** after regression correction. **E** was the posterior which had the smallest overall entropy. Error bars represent the mean and middle 95% of each marginal distribution.

$S_{ME}(\cdot) = (\bar{y}, s_y, a_{y,1})$ whose rank when comparing divergence is 2nd. The choice is much better in this case however the approach fails to identify the best subset of the summary statistics considered. Nunes & Balding (2010) also consider a regression correction to their obtained posteriors, if we take $S_{ME}(\cdot)$ with $\rho_{L^2}(\cdot)$ and apply first order and second order regression corrections the rank according to KL divergence is 13 of the equivalently corrected posteriors using other subsets. In fact if we were to run the same minimum entropy optimisation for regression corrected posteriors a different set of summary statistics is chosen. This suggests that if the intention is to perform a regression correction on the obtained posterior that this should be part of the optimisation. The smallest entropy of all of the possible posterior distributions came with $S(\cdot) = (\bar{y}, s_y, a_{y,2}, p_{y,3})$ using an unweighted Euclidean metric after second order regression correction. If we consider distributions that are not regression corrected the best overall choice is $S_{ME}(\cdot) = (\bar{y}, s_y, a_{y,1})$.

Figure 5.7 shows error bars comparing the marginal distributions of rate parameters for posteriors that could feasibly be chosen using the minimum entropy criteria with $\pi(\Theta)$ and $\pi_\epsilon(\Theta^{obs})$. The description of each posterior depicted in figure 5.7 is as follows:

- A - $\pi(\Theta)$, the true posterior distribution, using the analytic likelihood function.
- B - $\pi_\epsilon(\Theta^{obs})$, the approximate posterior distribution, given $S(\mathbf{Y}) = \mathbf{Y}$, after regression correction.
- C - $\pi_\epsilon(\Theta)$ with $S(\cdot) = (\bar{y}, s_y, a_{y,1})$ and $\rho_{L^2_w}(\cdot,\cdot)$. This subset of summary statistics is chosen as $S_{ME}(\cdot)$ when considering all uncorrected posterior distributions.
- D - $\pi_\epsilon(\Theta'')$ with $S(\cdot) = (\bar{y}, s_y, a_{y,1})$ and $\rho_{L^2_w}(\cdot,\cdot)$, that is distribution C after regression correction.
- E - $\pi_\epsilon(\Theta'')$ with $S(\cdot) = (\bar{y}, s_y, a_{y,2}, a_{y,3})$ and $\rho_{L^2_w}(\cdot,\cdot)$. This posterior had the smallest entropy value of all those considered.

Figure 5.7 shows that whilst selecting $S(\cdot)$ via a minimum entropy approach appears to be competitive if selection is made without consideration of regression corrections posteriors, this approach is no improvement on using the raw vector of values with no dimension reduction at all. The distribution E which had the smallest entropy of all distributions considered was relatively poor in comparison to others. This suggests that here, with the set of summary statistics being considered, use of the raw observations is still preferable. The authors then propose an optional second step, which uses the posterior obtained as a result of S_{ME} to assess the choice of $S(\cdot)$ via a measure of average error of the accepted samples in an ABC rejection run. We found however that this second step returned the same choice as S_{ME}.

Partial Least Squares (PLS)

Rather than attempt to find a best subset of summary statistics Wegmann *et al.* (2009) propose to reduce the dimension of the problem through use of a Partial Least Squares regression model. The idea is to use a pilot run of model realisations given parameter draws from the prior distribution. For each of the simulated data sets calculate a large set of hopefully informative summary statistics and fit a PLS

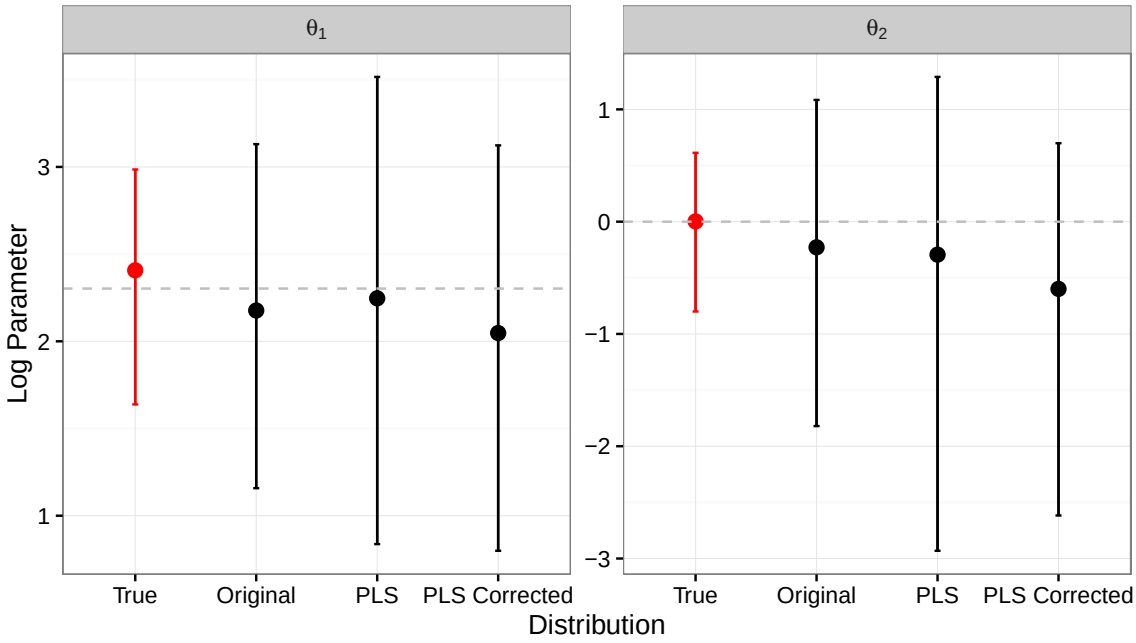


Figure 5.8: Posterior error bars for means and middle 95% of the PLS approach to summary statistics in ABC compared to the true posterior distribution and the original posterior inference using no summary statistics. They show that in the case of the single time series data whilst the PLS approach does similarly well to the original, using the raw observations themselves is just as competitive.

regression model with the summary statistics as predictor variables for the parameters that generated them. They motivate the use of PLS by the desire to have a set of summary statistics that are maximally decorrelated with one another whilst jointly being highly correlated with the model parameters. Dimension reduction is achieved by retaining only the first r components of the fitted model chosen by a cross validation procedure. The values of the r retained components are then taken to be the summary statistics used in the calculation of $\rho(S(\mathbf{Y}), S(\mathbf{Y}^*))$.

Here we use a PLS regression model on the raw vector of observations, scaled to have common mean and variance. Cross validation results lead to choosing the first 3 components as the summary statistics. Figure 5.8 shows error bars for the posterior means with middle 95% of the inferred marginal distributions of the rate parameters, including those where the posterior distributions have been regression corrected compared with the true distribution and $\pi_\epsilon(\Theta^{obs})$. Posterior variances for the PLS approach is greater than in $\pi_\epsilon(\Theta^{obs})$ showing that nothing has been gained over not using summary statistics at all in this case.

Semi–automatic ABC

Similar to the PLS approach Fearnhead & Prangle (2012) propose an approach to dimension reduction by constructing a new set of summary statistics from the observations. The authors show that on consideration of a quadratic loss function for estimation of the parameters the optimal summary statistics are $S(\Theta) = E[\Theta \mid \mathbf{Y}]$, that is the posterior mean. Since the posterior mean is unavailable they suggest using a linear regression model of the form

$$\Theta_i = E[\Theta \mid \mathbf{Y}] = \alpha + \beta f(\mathbf{Y}) \tag{5.12}$$

to estimate it. That is they propose to use a training run of model simulations given draws from the prior and use these to estimate the regression coefficients. Then for the rejection sampler they draw samples from the prior, estimate the summary statistics from the model and then follow the standard accept–reject step of the original algorithm. The authors find that fitting a regression model with $f(\mathbf{Y}) = (\mathbf{Y}, \mathbf{Y}^2, \mathbf{Y}^3, \mathbf{Y}^4)$ was competitive for a similar inference problem and we replicate that in this example.

Figure 5.9 shows that the semi automatic approach to calculation of summary statistics gives poorer performance in this example than using the raw observations directly. A regression corrected posterior for the semi–automatic samples gives better performance with means closer to the true posterior means and slightly smaller posterior variance than in $\pi_\epsilon(\Theta)$.

Fearnhead & Prangle (2012) also suggest an optional pilot run in which they locate a region close to the true parameter values for which to train their regression model. The motivation for such a step is that the fitted model may be more appropriate in the region of posterior mass. On doing this one should then truncate the prior distribution to the same region to avoid model interpolation beyond the space to which the model was fit. They note that a similar idea was used in Blum & François (2010) and can be viewed as weakly using the information from the pilot run in the final algorithm. For the pilot run, one needs to choose a set of summary statistics in order to define closeness for identification of the space on which to train the regression model. Here we use the vector of observations with a weighted Euclidean metric to do so as this has proven to be competitive thus far. Figure 5.10 is the posterior plots found by employing the pilot run and truncating the prior distribution

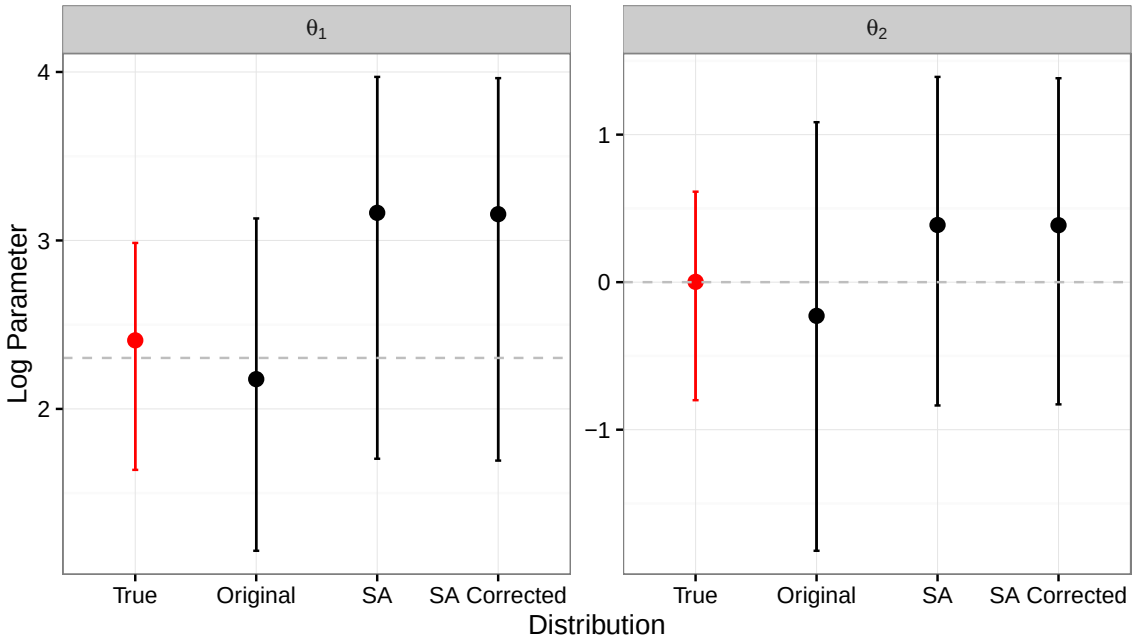


Figure 5.9: Posterior inferences for rate parameters of the immigration–death model using a semi automatic approach to the selection of summary statistics. Distributions are compared to the true posterior $\pi(\Theta)$ and $\pi_\epsilon(\Theta^{obs})$ through error bars showing means and central 95%.

analogous to those in figure 5.9. It appears in this case that this extra step has provided no improvement.

5.3.5 Lotka–Volterra

The Lotka–Volterra model introduced in section 2.5.3 exhibits more complex dynamics than the immigration–death model. It features an additional species and reaction with an oscillatory evolution of the states. On top of this the kinetics of the reaction network are stable only in certain areas of the state space. Here we investigate a similar range of scenarios within an ABC rejection sampler as for the immigration–death model in order to determine whether our findings are consistent.

Since an analytic solution to the Lotka–Volterra model is unavailable, the reasons for which are discussed in chapter 2, the true posterior distribution to be used as a reference is a long run of a pseudo marginal MCMC scheme that uses a large number of particles. As discussed in chapter 4 this sort of approach can be used to obtain an

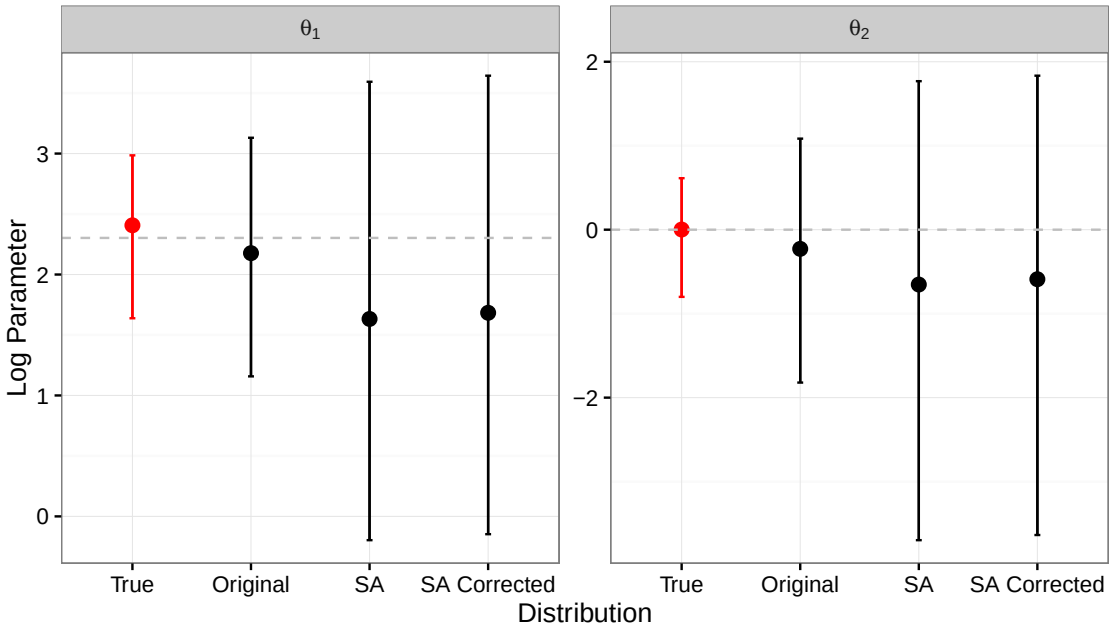


Figure 5.10: Posterior plots for the semi automatic approach to summary statistics where a pilot run has been used to choose an appropriate region for the regression model to be fit and the prior truncated accordingly. This seems to in this example offer no improvement over the results shown in figure 5.9 where no pilot run was used.

exact posterior distribution, albeit at great computational expense. A comparison of the relative efficiency of the two approaches is the focus of section 5.4 later in this chapter.

Weighting the Euclidean metric

In section 5.3.3 we found that when using the raw observations as the collection of summary statistics posterior inferences were improved by weighting them by the prior predictive standard deviations in the metric. We repeat a similar experiment with simulated data for the Lotka–Volterra model here. The observations, shown in figure 5.15, are noisy measurements of the underlying state over a regular time interval. Prior distributions for the rate parameters are vague and described in section 4.4.2.

Figure 5.11 shows marginal posterior densities for the rate parameters of the model which make clear that posterior inferences using an unweighted Euclidean metric is much better than using a metric weighted by the prior predictive standard deviations

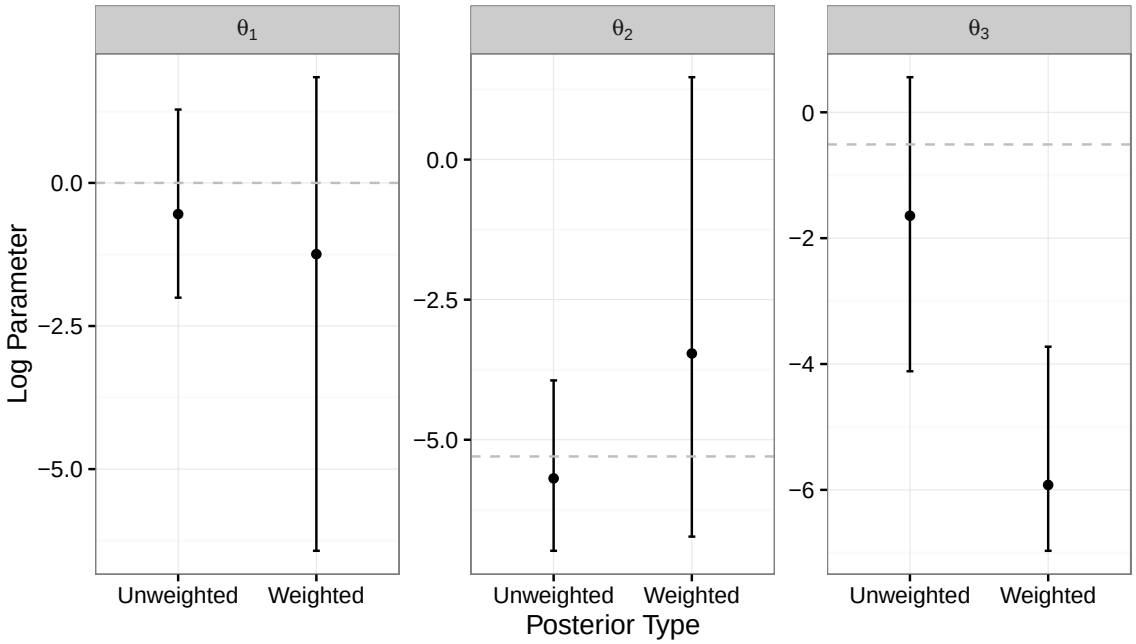


Figure 5.11: Marginal posterior error bars for the middle 95% of the distributions of rate parameters of a Lotka–Volterra predator prey model obtained using an ABC rejection sampler. Each posterior sample is the result of retaining the best 0.05% given ten million draws from the prior. Dashed lines give the parameter values used to generate the data.

of the observations in this case. This is in contrast to what was observed with the immigration–death model. It is not entirely clear why this is the case. One reason for the difference in behavior could be that the Lotka–Volterra system dynamics are unstable for a large region of parameter space with a tendency for either extinction of both species or an explosion of prey. This behaviour in turn leads to very large prior predictive variances for those observations which occur later in the time series. Large variances lead to small weights in the metric meaning we are basing our acceptance criteria on potentially only a few early observations. Examination of the weights in the simulated experiment showed that the values of the predator and prey at time $t = 0$ accounted for approximately 95% of the weight in the calculation of δ. What this tells us is that here we are basing the majority of our accept reject step on our prior information for the initial state of the system.

Csilléry *et al.* (2012) suggest that rather than use an empirical standard deviation over the prior predictive distribution to use median absolute deviation (MAD). The motivation behind such a choice is that this value is much more robust to extreme outliers. On examination of this weighting scheme for the rejection sampler in

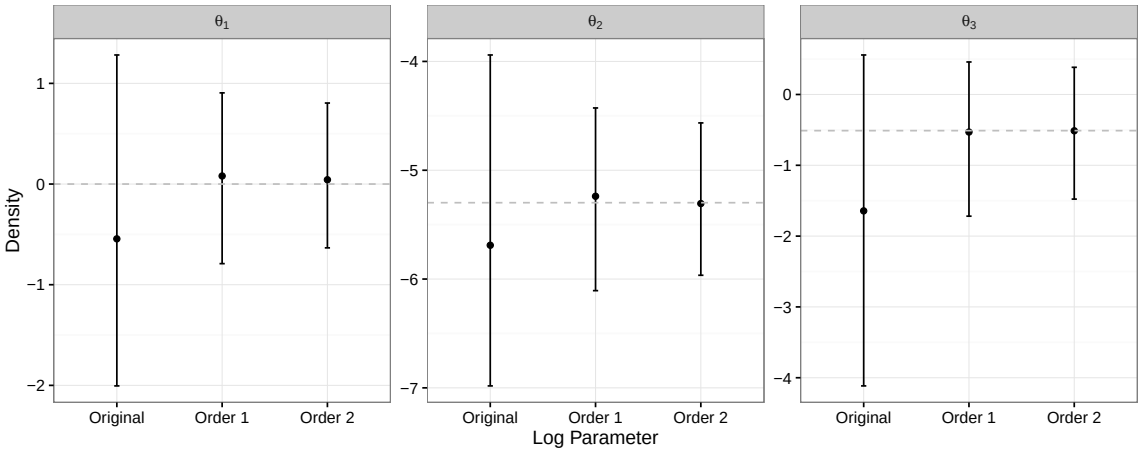


Figure 5.12: Point estimates of mean and central 95% error bars for the inferred posterior distributions of the rate parameters of the Lotka–Volterra predator prey model with $S(\mathbf{Y}) = \mathbf{Y}$, and subject to first and second order posterior regression correction of Beaumont *et al.* (2002).

this example, no improvement over the unweighted summary statistics was found, although it should be noted that it yielded much better posterior distributions than weighting by the inverse of the standard deviations.

Regression correction

As was the case for the immigration death model a significant improvement in the posterior distributions obtained could be found by performing a regression adjustment of Beaumont *et al.* (2002). Both a single order and second order polynomial regression correction yielded a posterior whose mean was very close to the true value with a dramatic reduction in posterior variance. Figure 5.12 shows error bars for the middle 95% of the marginal posterior distributions of the uncorrected and corrected samples.

Dimension reduction

Investigation into the effect of choosing informative summary statistics as a means of reducing the dimension of the problem becomes more difficult as the number of candidate summary statistics increases. This is in part due to the increase in the number of subsets of a given set of summary statistics. Evaluation of posterior approximations under all possible subsets quickly becomes computationally infeasible

as the number of candidate statistics increases. This was acknowledged by Joyce & Marjoram (2008) and Nunes & Balding (2010) among others. Joyce & Marjoram (2008) then suggest an iterative approach to testing the possible inclusion or exclusion of a statistic but do note that the order in which statistics are tested has consequence in the subset that is selected.

Partial Least Squares

As was observed for the immigration death model, a PLS regression approach to reduction of the dimension of summary statistics yielded no improvement in the posterior sampling for the Lotka–Volterra model. A PLS regression was performed using the raw observations as the untransformed summary statistics. The first 7 components were then chosen to use in the ABC rejection sampler, chosen by cross validation. Regression correction of the resultant posterior did improve the samples however performance was still poorer than using raw observations.

Semi automatic ABC

A semi automatic approach to the choice of summary statistics for the Lotka Volterra model yielded similar results to those found for the immigration death model. That is that for the simple rejection sampler there was no improvement in performance over using raw observations. We followed the approach as in Fearnhead & Prangle (2012) by fitting a regression model predicting each parameter using a fourth order polynomial on the simulated values of the process. As with other experiments a posterior regression correction was also performed.

Figure 5.13 shows 95%–ile error bars with posterior means for the approximate posteriors obtained using the summary statistics generated by the regression model. From ten million model simulations given draws from the prior those with $\rho(s(x), s(y)) < \epsilon$ where ϵ is chosen such that the best 0.05% of simulations are retained to mimic the acceptance rate in other experiments. The semi automatic ABC posteriors, both uncorrected and regression corrected show poorer posterior sampling than was obtained using $s(x) = x$. Similar results were found when first using a pilot run to identify a truncated region of the prior distribution for the regression model to be fitted.

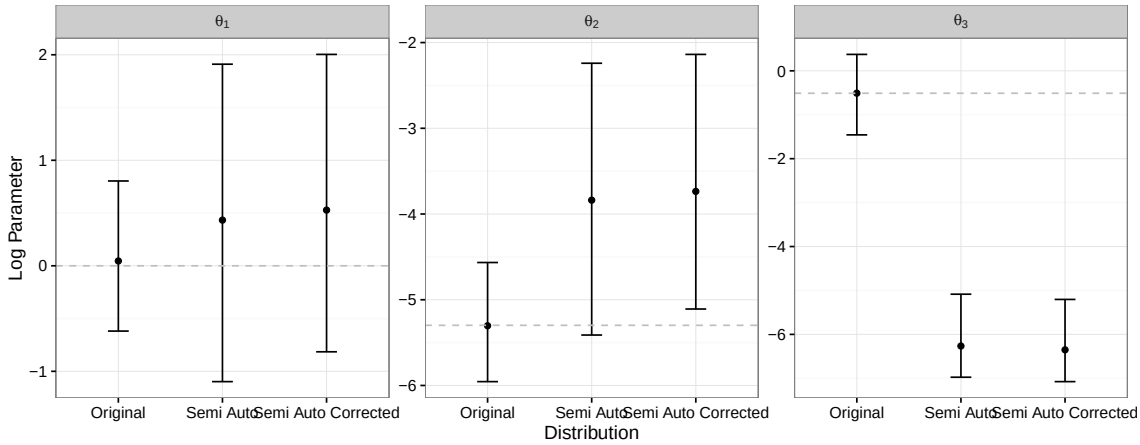


Figure 5.13: Approximate posterior distributions using the semi automatic approach of Fearnhead & Prangle (2012) to the choice of summary statistics for the Lotka Volterra model. Error bars show the middle 95% and posterior means for both regression corrected and non regression corrected posteriors compared to the best ABC posterior obtained with no dimension reduction.

We note that the performance of the semi–automatic approach was, in these examples, heavily influenced by the region on which the model was fit. Using either the full, vague, prior parameter space, or only a short pilot run in which to identify a truncated region for model fitting led to high variability in model parameter estimates over repeated experiments. The pilot run here was kept short in order to be consistent with other examples but we note that we find the approach to be more competitive when the regression model is fit on a more informative region of parameter space.

5.3.6 Conclusions

The mechanism used to measure similarity between simulated and observed data in ABC algorithms has a dramatic effect on the resultant posterior distributions inferred. Perhaps surprisingly it appears, within the context of models of this type at least, that the distance function itself seems to have very little bearing on the outcome when each point is considered with equal weight. A discrepancy is borne out when weighting points by estimates of prior predictive standard deviation and whilst an L^∞ norm gave a posterior with a more accurate mean the increased variance meant that the commonly used Euclidean norm performed more favourably overall.

The often quoted benefits of reducing the dimension of the problem with summary statistics that compress the information in the data in some way, including some loss, do not appear to hold true in this context and using raw data appears competitive in each example. A regression correction of the posterior distribution seems to improve the sample. The effect seems much more pronounced using a second order polynomial on $(\mathbf{s}_i^* - \mathbf{s})$ rather than a single order polynomial in the examples considered here. In general however one can consult standard linear regression model diagnostics to find a model that justifies the underlying assumptions.

Whilst the simple linear regression correction approach of Beaumont *et al.* (2002) works reasonably well in the examples above it is not without issue. Difficulties can arise when observed the relationship between parameters and summary statistics is highly non–linear or when observed summary statistics lie towards the edges of the prior predictive distribution of summary statistics. Blum & François (2010) provide a number of refinements to the regression approach. Rather than linear regression they use a feed forward neural network to estimate $\hat{\mathbb{E}}(\Theta \,|\, \mathbf{s})$. Additionally they model the log of the squared residuals to obtain an estimate of $\hat{\mathrm{SD}}(\Theta \,|\, \mathbf{s})$. This leads to a modified regression correction step

$$\Theta_i^* = \frac{\hat{\mathrm{SD}}(\Theta \,|\, \mathbf{s})}{\hat{\mathrm{SD}}(\Theta \,|\, \mathbf{s}_i)} \Big(\Theta_i - \hat{\mathbb{E}}(\Theta \,|\, \mathbf{s}_i) \Big) + \hat{\mathrm{SD}}(\Theta \,|\, \mathbf{s}) \tag{5.13}$$

where the idea is that by doing this they allow both the mean and variance of Θ to vary with $S(\mathbf{Y})$. Their results show improved accuracy over the earlier approach by Beaumont *et al.* (2002).

The conclusions here have been made with respect to a simple rejection ABC sampler however it is reasonable to assume that they extend to other ABC based samplers. Indeed the relative merits of different metrics and summary statistics should remain constant in each of the ABC variants and examples of regression correction to posterior samples in an ABC MCMC setting and ABC SMC can be found in Lopes & Beaumont (2010) and Blum & François (2010) respectively.

5.4 Comparison of different approaches

This section is based on and is an expansion of the work published by Owen *et al.* (2015). The aim of the following set of computational experiments is to try to gain insight as to the relative efficiency of ABC and particle MCMC to parameters inference for the class of models.

In the context of stochastic kinetic models for which the likelihood function is intractable, inference must proceed using either particle MCMC, which avoids evaluation of the likelihood function through a Monte Carlo estimate of marginal likelihood, or an ABC based technique which uses the rejection sampler approach based on the distance of simulated data from observed. Both particle MCMC, as in Golightly & Wilkinson (2011) and Wilkinson (2011), and ABC methods, (Drovandi & Pettitt, 2011; Fearnhead & Prangle, 2012), have been successfully applied in the context of stochastic kinetics, but it is unclear as to which approach is favorable.

If exact posterior inference is desired, and there is measurement error in the observations, we are limited to particle MCMC. Whilst in theory given a state space that is a countable set exact inference is possible using ABC methodology (as $\epsilon \to 0$ in the case of $s(x) = x$ or $s(x)$ being sufficient), in practice this outcome is rarely achieved. However, if exact inference is not the primary concern and there are computational constraints, perhaps available CPU time, it is not obvious which approach should be employed. Is it the case that increased computational efficiency is an adequate trade–off for the reduction in accuracy? Additionally particle MCMC in this context relies on there being an adequate amount of measurement error. The relative efficiency of each posterior sampling scheme may well depend on the size of the measurement error.

In order to make direct comparisons between the different techniques it is necessary to create some framework by which meaningful conclusions can be made. In order to do this it is important to consider a set of rules by which a fair test can be undertaken. In addition to this we are interested in some measure of efficiency of each sampler and the discrepancy between the resultant posterior in each case and the true posterior. One of the primary motivations for such a comparison was to determine which method is most appropriate with particular consideration to the notion of computational restrictions.

5.4.1 Computational budget

Whilst it is of interest to be able to compare the execution time of each of the inference schemes, experimentation is subject to variation in the implementation of each algorithm. Instead comparisons could be made using the notion of computational units rather than wall clock time. The motivation behind this choice is such that any possible inefficiencies in the code are irrelevant to the conclusions made. This ignores some of the practical elements of the computation that are prevalent in each algorithm, for example the parallelisability, a discussion of which will be made in chapter 6. Any conclusions drawn within the computational budget paradigm are correlated with wall clock time subject to this additional discussion.

5.4.2 Initialisation

Each of the approaches to inference discussed in chapter 4 have tuning parameters which have to be chosen in some way, each of which has a bearing on the efficiency of the sampler. In order to make comparison as fair as possible it is desirable to have each algorithm in some sense optimised using standard published methods. The cost of obtaining such tuning parameters is to be collected and deducted from the allocated computational budget. The motivation behind the inclusion of this cost is that a practitioner would have to undergo this procedure in a "real data setting".

Particle MCMC

It has been well documented that the efficiency of random walk Metropolis algorithms is highly dependent on the choice of proposal kernel. A distribution which yields small deviations from the current state will ensure that a large number of moves are accepted but samples will be highly correlated. Large moves around the space on the other hand will often be rejected leading to the chain spending large amounts of time stuck at the same value. A brief introduction to Metropolis Hastings MCMC was given in section 3.5.2. Optimal tuning parameters for a Gaussian random walk proposal kernel in an Metropolis Hastings MCMC scheme are different in the case where likelihood estimates are obtained using a particle filter than in the standard analytic likelihood case due to the inherent cost in obtaining estimates of

likelihood, see section 4.3.2. The starting parameter vector, Θ_0, of the chain also has an effect on the efficiency of the sampler. A choice of Θ_0 which is far from a region of non–negligible posterior density will lead to a chain which takes a long time to move toward the target distribution, whereas a chain initialised close to stationarity will yield useful samples sooner. This burn–in period can sometimes consume a sizeable fraction of the computational budget. This is of greater concern for particle MCMC than it is for posterior sampling with an analytic likelihood function as particle MCMC, due to the variability of the estimator being used at each iteration can suffer from poor mixing, particularly in the tails of the target distribution. In a region of low posterior density, if the particle filter were to estimate a particularly high density value of the target due to chance in a variable estimator, the chain would be reluctant to move away from this value in future iterations. This phenomena will be examined in more depth during this section.

Particle MCMC as described in section 4.3 relies on a sequential Monte Carlo algorithm for approximation of the likelihood, $\hat{\pi}(\mathcal{D}|\theta)$. The bootstrap filter requires multiple model realisations, via a set of "particles" in order to achieve this approximation. In addition, the approximation has to be calculated at every iteration of the MCMC algorithm, hence clearly the number of particles in the filter will greatly affect the runtime of the resultant algorithm. A small number of particles will result in a shorter computation time for the likelihood approximation but leads to larger variability in the estimated likelihood. This increased variability leads to decreased efficiency of the inference scheme, as noted by Andrieu & Roberts (2009). A large number of particles, useful for consistent estimates of $\hat{\pi}(\mathcal{D}|\theta)$ will lead to slower posterior sampling in the chain. An optimal choice for the number of particles was discussed in section 4.3.3.

In practice, for the purpose of the comparisons made in section 5.5 below we choose an initial parameter vector, Θ_0, as a random sample from the posterior distribution. The number of particles used in the particle filter, N is then chosen by repeated runs of a particle filter with increasing N until $1.5 < \mathrm{Var}(\hat{l}(\theta_0|\mathcal{D})) < 1.8$. We then use the covariance matrix of the posterior, Σ_p to inform our choice for Σ_q, the Gaussian random walk proposal variance, ensuring that the acceptance rate is around 15%, following the practical advice of Sherlock *et al.* (2015).

During initial experiments with the pMCMC algorithm for these models initialisation and tuning of the algorithm was approached under the assumption of no

knowledge of the posterior distribution of interest, reflecting the position of a practitioner in a real scenario. However, this proved to be problematic as finding a sensible choice of Θ_0, number of particles, N, and proposal variance, Σ_q, often used a large proportion of the allocated computational budget. Under the computational restrictions imposed by the budget choice this made pMCMC look completely uncompetitive relative to ABC techniques. This problem itself is interesting as it highlights a potential drawback of using pMCMC in practice. We acknowledge that the initialisation tools used in this comparison are not representative of the full problem that faces users looking to utilise these algorithms but wish to present a comparison on the relative efficiency of the posterior samplers. The outstanding issue of initialising and tuning a particle MCMC scheme under no knowledge of the posterior density that we wish to learn about will be discussed in more depth in the following chapter and was addressed in Owen *et al.* (2014) on which that chapter is based.

ABC rejection sampler

For the ABC rejection sampler, once a metric function and set of summary statistics has been chosen, it only remains that weights for the summary statistics in the metric and a tolerance ϵ need be determined. In section 5.2 we saw that for time course data of the stochastic kinetic models using $S(\mathbf{Y}) = \mathbf{Y}$ was competitive and that weights could be determined by a small (relative to the size of the full experiment) pilot run with draws from the prior distribution. The final tuning parameter to be determined, ϵ, is not necessary as it is possible to retain the best $\alpha\%$ of samples at the end. Whilst there is some question over what α is a reasonable choice it has no bearing on the computation time as it can be chosen after all computation has been performed.

ABC MCMC

Like any MCMC algorithm it is necessary to choose a proposal kernel. Since to our knowledge there is no published literature that addresses optimal proposal kernels for random walk ABC MCMC specifically we rely on the same theory that applies to random walk Metropolis Hastings samplers in general. Unlike the rejection sampler variant of ABC it is necessary to determine a sensible choice of tolerance, ϵ, before

the algorithm is run as crucially this plays a role in the acceptance probability of proposed moves in the Markov Chain. It is therefore necessary to have some sensible approach for choosing this tolerance. There is no trivial answer to the question over what makes a good value of ϵ, and the answer is likely very application specific. In the absence of specialist prior knowledge and understanding of how a process works we can rely only on some pilot simulation information to attempt to choose something sensible.

This proved to be problematic in practice for utilising an ABC MCMC approach competitively. A first approach used a simple rejection sampler to identify a region according to a best $\alpha\%$ principle, denoted $\pi_\alpha(\Theta)$. From that we could take the distance value δ_α that corresponded to the $\alpha\%$-ile of the distribution of distances and use the mean of $\pi_\alpha(\Theta)$, denoted $\bar{\Theta}_{\delta_\alpha}$, to initialise the chain. The reasoning behind this was that by first identifying a region $\pi_\alpha(\Theta)$ that gave a small number of samples in a rejection sampler we could better explore that smaller space with an MCMC sampler with local proposals. Being able to initialise somewhere close to the center of this region seemed intuitive.

Whilst choosing an ϵ and Θ_0 in this way yields a chain that explores the space well it does not exploit the strengths of using a Markov chain with local proposals that motivated the development of ABC MCMC at its inception. That is the increased acceptance rate allowing the use of a smaller tolerance. It is desirable to be able to choose a smaller value of ϵ than that which is borne out of the rejection sampler tuning approach, however a meaningful way to do this is not obvious. One approach could be to take some percentile, q of an empirical distribution of distances given the $\bar{\Theta}_{\delta_\alpha}$. By doing so what we are essentially imposing is that we accept on average $q\%$ of the time at the mean of $\pi_\alpha(\Theta)$. If we then initialise at that mean, we should expect a reasonable acceptance rate in the chain.

To overcome this we took the samples $\pi_\alpha(\Theta)$ and made repeated draws of δ from each sample build a representative sample of the predictive distribution of δ for the region

$$\pi_\alpha(\delta) = \int \pi_(\delta \mid \Theta)\pi_\alpha(\Theta)d\Theta.$$

From this distribution we can choose a tolerance ϵ based on some quantile, q_δ, of the distribution where q_δ now represents the approximate acceptance rate that would be exhibited in a simple random sampler using the original approximate posterior

for proposals. To initialise the chain we use the parameter vector Θ which gave the smallest average distance value in this additional tuning run. This should ensure that we expect the chain to move with a reasonable acceptance probability.

ABC SMC

Initialisation of a sequential ABC algorithm as described in section 4.2.4 is somewhat less involved. This is due to the fact that optimal Gaussian proposal kernels for advancement to subsequent targets can be calculated during execution. In addition the sequence of tolerances is chosen adaptively throughout the algorithm. It remains that there is need to specify an initial tolerance value, ϵ_0. One could argue that tuning the choice of metric and summary statistics to be used is also of interest. Investigation into the effect of different metrics and summary statistics was the focus of section 5.2. Whilst it is likely that the best choice of summary statistics and metrics is highly problem specific, in all of the numerical examples of section 5.3 a Euclidean metric for $S(\mathbf{Y}) = Y$ was found to be competitive. In order to choose a suitable ϵ_0 for the scheme we simply calculate $\rho(\mathcal{D}, \mathcal{D}^*|\theta)$ using a number of samples from $\pi(\theta)$. From this we take ϵ_0 to be the value equivalent to the 1%-ile of the resultant distribution of distances.

5.5 Numerical examples

5.5.1 Comparing the different ABC algorithms

Before comparing ABC with particle MCMC it seems prudent to make some comparison between the ABC algorithms. Whilst it is expected that both ABC SMC and ABC MCMC are more efficient than a simple rejection sampler implementation since they were both conceived with the intention of addressing weaknesses in the rejection sampler, it is not obvious however whether the sequential Monte Carlo variant will be more efficient than the Markov chain based ABC sampler. These preliminary runs were performed using a smaller computational budget of 10^7 model realisations with the motivation of taking only the most efficient ABC sampler forward to compare with particle MCMC. This seemed a reasonable approach since from a practical point of view there is little additional difficulty to

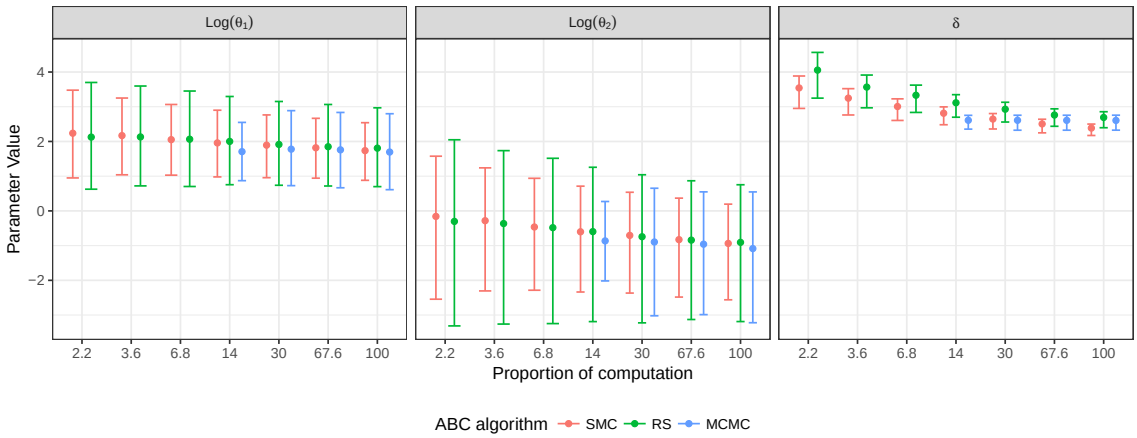


Figure 5.14: Error bars for posterior distributions of rate parameters for an immigration death process for each of the 3 types of ABC algorithm. The grouping along the x axis is such that the cost of obtaining the distribution, in percentage units of model simulations, is constant.

the practitioner in constructing any of the ABC sampling algorithms discussed in chapter 4 on understanding of the theory.

Immigration death

An immigration death process with true rate parameters $\Theta = (10, 1)$ and Gaussian observation noise $\sigma^2 = 1$, as shown in figure 4.2 was the focus of inference for each algorithm. In each case the prior distribution on the rate parameters was

$$log(\theta_i) \sim \mathcal{U}(-4, 4), \quad i = 1, 2$$

where the noise parameter $\sigma = 1$ and the initial state of the system $\mathbf{X}_0 = 0$ are assumed known.

Initialisation of each algorithm follows the specification outlined in section 5.4.2. Specifically for the rejection sampler, from the allocated budget of 10^7 model realisations, one hundred thousand (1%) were used to estimate the prior predictive standard deviations of the summary statistics to deduce the weights. Choice of ϵ was left out of the initialisation process since it can be chosen at the end. For the MCMC sampler an initial set of five hundred thousand model realisations was used for the determination of summary statistic weights and to choose an intermediate tolerance δ_α where $\alpha = 0.1\%$. Given δ_α, to encourage a smaller tolerance, and hence

hopefully a better approximation to the true posterior distribution, calculate $\bar{\Theta}_{\delta_\alpha}$ the mean parameter vector of the samples for which $\delta < \delta_\alpha$. Using $\bar{\Theta}_{\delta_\alpha}$ one hundred thousand distances were calculated and ϵ was taken to be the value corresponding to the 1%-ile of the distribution of distances $\pi(\delta \,|\, \Theta = \bar{\Theta}_{\delta_\alpha})$. Finally to estimate an optimal proposal variance for the random walk kernel one hundred thousand iterations of an MCMC scheme with a very small proposal variance were used to create a sample from which to estimate variance Σ. Given Σ the variance for the Gaussian random walk kernel was chosen as $2.38^2\Sigma/2$. For the ABC SMC sampler the initial tolerance ϵ_0 was chosen as the 1%–ile of given five hundred thousand parameter draws from the prior distributions. The weights to be used in the metric function are also determined using these samples. The initial distribution is then $\pi_0(\Theta) = \pi_{\epsilon_0}(\Theta)$ and choice of the tolerance ϵ_t given distribution $\pi_{\epsilon_{t-1}}(\Theta)$ is chosen as the 30%–ile of distances δ_{t-1}.

The preliminary examples using an immigration death process model found that ABC based on a sequential Monte Carlo sampler was the most efficient of the ABC iterates. It's advantage over the simple rejection sampler is clear, by reducing the tolerance gradually acceptance rates can be maintained and having an iterative approach to updating the posterior distribution given proposals based on the previous target allows guided exploration of the space. It also proved advantageous over ABC MCMC. Whilst it is possible to get good samples using an ABC MCMC scheme one of the drawbacks is that it is not trivial to reduce the tolerance once the chain sampler has begun. In addition the extra effort in ensuring initialisation of the chain is reasonable is unwanted complication. If initialisation proves to be poor, either the chosen tolerance is too large, the random walk kernel is sub–optimal or the starting point of the chain is in a region of negligible support, then substantial computational effort may be expended before ultimately having to restart the chain given different initialisation criteria.

Figure 5.14 shows the evolution of the marginal approximate posterior distributions for the rate parameters for the immigration death process and the distribution of calculated distances against computational cost. The groupings have been chosen to match the computational cost of the successive distributions in the sequential ABC. That is as each successive population of the ABC SMC scheme was obtained, the proportion of the allowed computational budget was noted, and the posterior distribution available given the same computational expense for the other ABC

algorithms was used for comparison. Posterior error bars are not present in the early phase comparisons for the ABC MCMC scheme since at those early stages, the cost of initialising the chain was such that a large enough number of posterior samples had yet to be drawn. Posterior variance is smaller for the ABC SMC scheme than the other ABC approaches, and the distribution of distances is smaller on average for this relatively simple model.

Lotka–Volterra

Each of the three ABC samplers were employed to infer the rate parameters of the Lotka–Volterra model. As with the immigration death example in the previous section the measurement error noise is assumed to be known as $\sigma = 10$. Prior distributions for the rate parameters and initial state $\mathbf{X}_0$ are

$$\log(\theta_i) \sim \mathcal{U}(-6, 2),\ i = 1, 2, 3 \quad X_{0,1} \sim \text{Po}(50) \quad X_{1,1} \sim \text{Po}(100). \tag{5.14}$$

The algorithms were initialised in the same way as with the immigration death example, with the exception that no weighting of the summary statistics was sought since results in section 5.15 found for this particular model with this prior that weighting the summary statistics gave poorer performance than not weighting them.

The results of this investigation were consistent with those found for the immigration death model but more pronounced given the additional complexity of the problem. Given the additional parameter to infer, ABC SMC was clearly the best choice.

5.5.2 Comparison of ABC and PMCMC

Lotka Volterra predator–prey model

Synthetic data

For the purpose of making comparisons we use a number of data sets over different observation regimes simulated using reaction rate vector $\theta = (1.0, 0.005, 0.6)$. In each case we corrupt the X_t with a Gaussian error with mean 0 and variance σ^2, $\pi(d_t|X_t, \sigma) \sim \mathcal{N}(X_t, \sigma^2)$. $X_0 = (50, 100)$ is used throughout. Plots of each of the

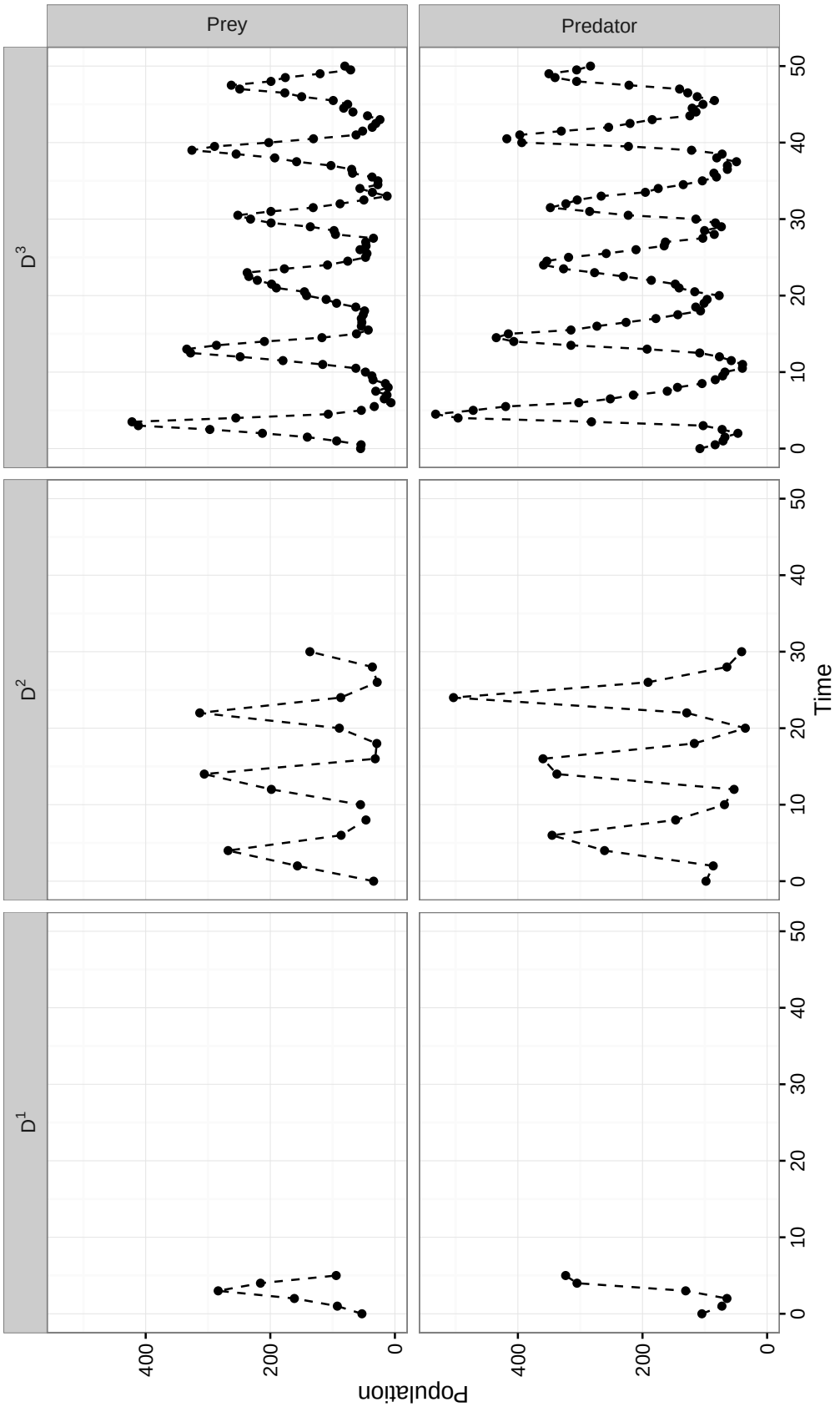


Figure 5.15: Synthetic data sets for the Lotka–Volterra predator prey model given $\log(\theta) = (0, -5.30, -0.51)$, $\log(\sigma) = 2.3$ and $X_0 = (50, 100)$. Dataset $\mathcal{D}^1$ is a short time series with observations at 6 time points at integer frequency. $\mathcal{D}^2$, a time series observed at even time points with 16 time point measurements. Dataset $\mathcal{D}^3$, a long time series of 101 time point measurements observed every 0.5 time units.

data sets considered are in figure 5.15. Given this set of parameter values the model exhibits relatively stable oscillatory behavior for both species and provides an interesting starting point for our investigation. We shall use this model to explore posterior sampling efficiency given data sets of a range of sizes, under full and partial observation regimes, whilst also giving consideration to the effect of assuming known measurement error or including this parameter in the set to be inferred. We consider time series of differing lengths to determine whether the amount of data available has an influence on which inference method may be most appropriate. In each case we also consider a partial observation regime where predator observations are unavailable by discarding these measurements. Data sets shown in figure 5.15 are denoted $\mathcal{D}^1$, $\mathcal{D}^2$ and $\mathcal{D}^3$ respectively. We introduce additional subscript notation such that $\mathcal{D}^1_p$ implies the data set $\mathcal{D}^1$ where predator observations have been discarded and $\mathcal{D}^1_u$ symbolises treatment of $\mathcal{D}^1$ under the assumption of unknown measurement error. In addition $\mathcal{D}^1_*$ will be used as a reference to the collection of data sets $\mathcal{D}^1, \mathcal{D}^1_u, \mathcal{D}^1_p$ and $\mathcal{D}^1_{u,p}$.

Inference set up

We now create a scenario in which prior parameter information is poor. We place uniform prior information on $\log(\theta)$,

$$\log(\theta_i) \sim \mathcal{U}(-6, 2), \quad i = 1, 2, 3, \tag{5.15}$$

and a Poisson prior distribution on the initial state

$$X_1 \sim \text{Pois}(50), \quad X_2 \sim \text{Pois}(100). \tag{5.16}$$

Where σ^2 is unknown, we use

$$\log(\sigma) \sim \mathcal{U}(\log(0.5), \log(50)). \tag{5.17}$$

For each repeat, we allow a computational budget of 10^8 model realisations from the Direct method, algorithm 1. We choose this budget based on the fact that given the $\theta = (1.0, 0.005, 0.6)$ our simulator achieves 10^4 simulations of length equivalent to $\mathcal{D}^2$ every 45-50 secs on our relatively fast Intel core i7-2600 clocked at 3.4 GHz.

This yields an approximate total time spent simulating from the model of 14 hours plus some other comparably negligible computation costs for each individual inference run. Clearly improvement on simulation time can be made by parallelising the simulation of independent realisations from the direct method as well as other computational savings being made by clever optimisations in each algorithm. We have tried to disclude the effect of such algorithmic optimisation in the comparison. We include the information on approximate time here as a rough guide to practical implementation of inference for these types of models as well as the reasoning behind our particular budget choice.

Discussion of results

Data sets D^1_*

Here we present results for the inference experiments for the D^1_* collection of data sets. The average number of particles required in the bootstrap particle filter to satisfy the requirements on $\mathrm{Var}(\hat{l}(\Theta))$ for D^1, D^1_u, D^1_p and $D^1_{u,p}$ are 54, 550, 48 and 114 respectively. The number of populations the ABC SMC sampler achieved within the computational budget is 12 in each of the examples where both species are observed and 9 in each of the prey only observation runs.

Figure 5.16 shows posterior inferences of $\log(\theta_1)$ given the shorter time series D^1 in each of the following scenarios

- Both species observed, measurement error assumed known,
- Both species observed, measurement error to be inferred,
- Prey levels only observed, measurement error known,
- Prey levels only observed, measurement error to be inferred.

The distributions are very similar between the two approaches and across replications of the experiment. In all cases the true value $\log(\theta_1) = 0$ is well identified by the posterior. In the case of only Prey observations being made the ABC approach loses some posterior mass in the tails of the distribution, relative to the reference truth. Results for the other two reaction rate parameters are consistent with those reported here.

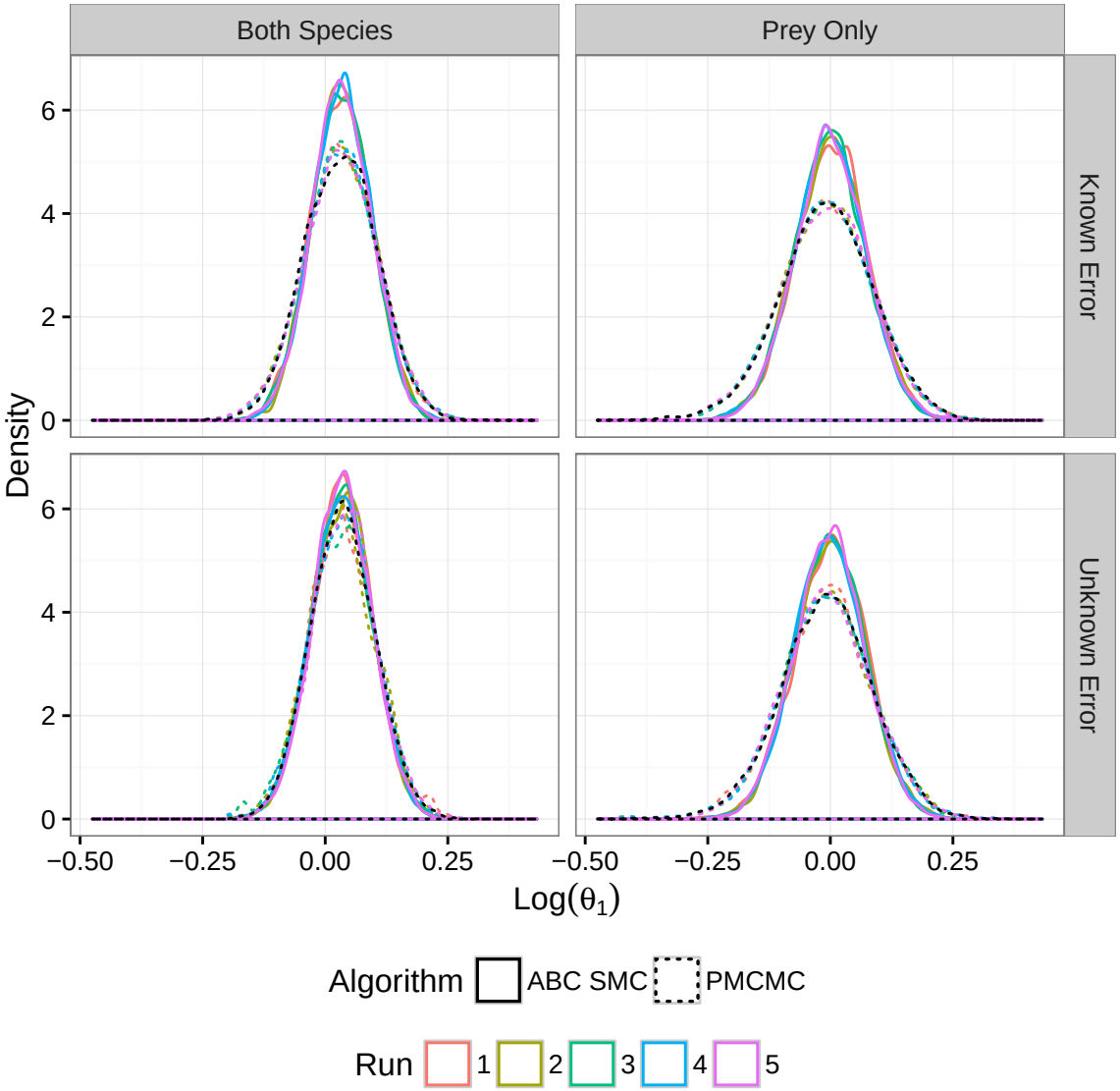


Figure 5.16: Posterior inference for $\log(\theta_1)$, the log of prey birth parameter under each of the sampling schemes. The true posterior distribution is shown in black. Results for other the other reaction rate parameters show results consistent with the ones shown here. The true value $\log(\theta_1) = 0$ used to generate the data set is well represented in each distribution.

Figure 5.17 shows posterior inferences for the noise parameter in the examples where it was not assumed to be known. In each repeat, for both observation regimes and both samplers inference is poor. The reference true distribution also exhibits poor learning of the noise parameter in light of a small amount of data with each PMCMC run performing close to the reference. ABC on the other hand fails to identify the noise parameter at all, with little deviation from the uniform prior.

Posterior densities do not give the full picture of the relative performance of the two

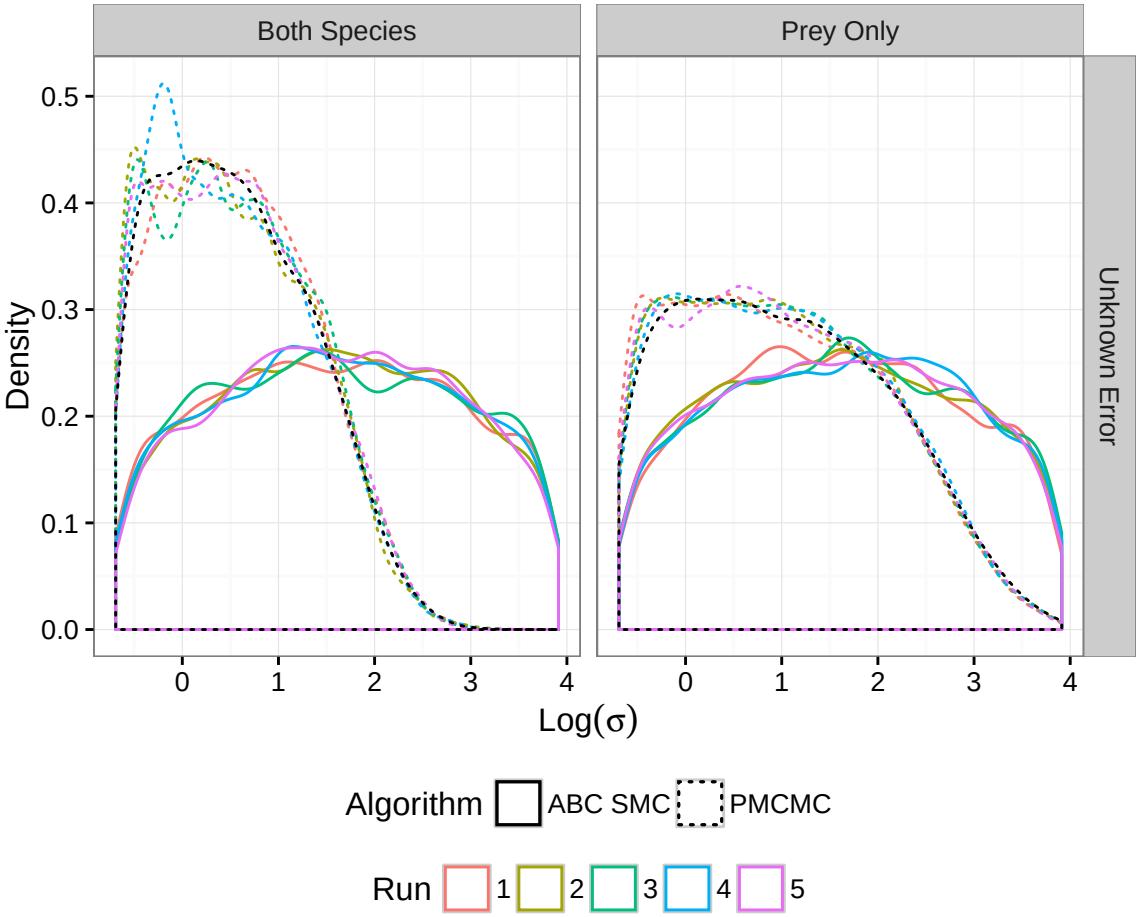


Figure 5.17: Posterior distributions for $\log(\sigma)$ in the D_u^1 and $D_{u,p}^1$ data sets. The true value used in the generation of the data is $\log(\sigma) = 2.30$.

algorithms. Figure 5.18 shows the evolution of the posterior samples in line against the computational expense. The box plots show that the shape of the pMCMC induced posterior distribution changes very little over the course of the computation and within a relatively short time the ABC based scheme gets close to it. From that point onwards the shape of the distribution changes little. As expected the effective sample size of the pMCMC posterior sample increases with computational expense. On consideration of estimates of posterior means and variances, the two posterior sampling algorithms give similar results after a relative short period of time. The fact that the initialisation cost of pMCMC has been omitted however suggests that perhaps ABC is the more efficient choice for learning of reaction rate parameters in this example.

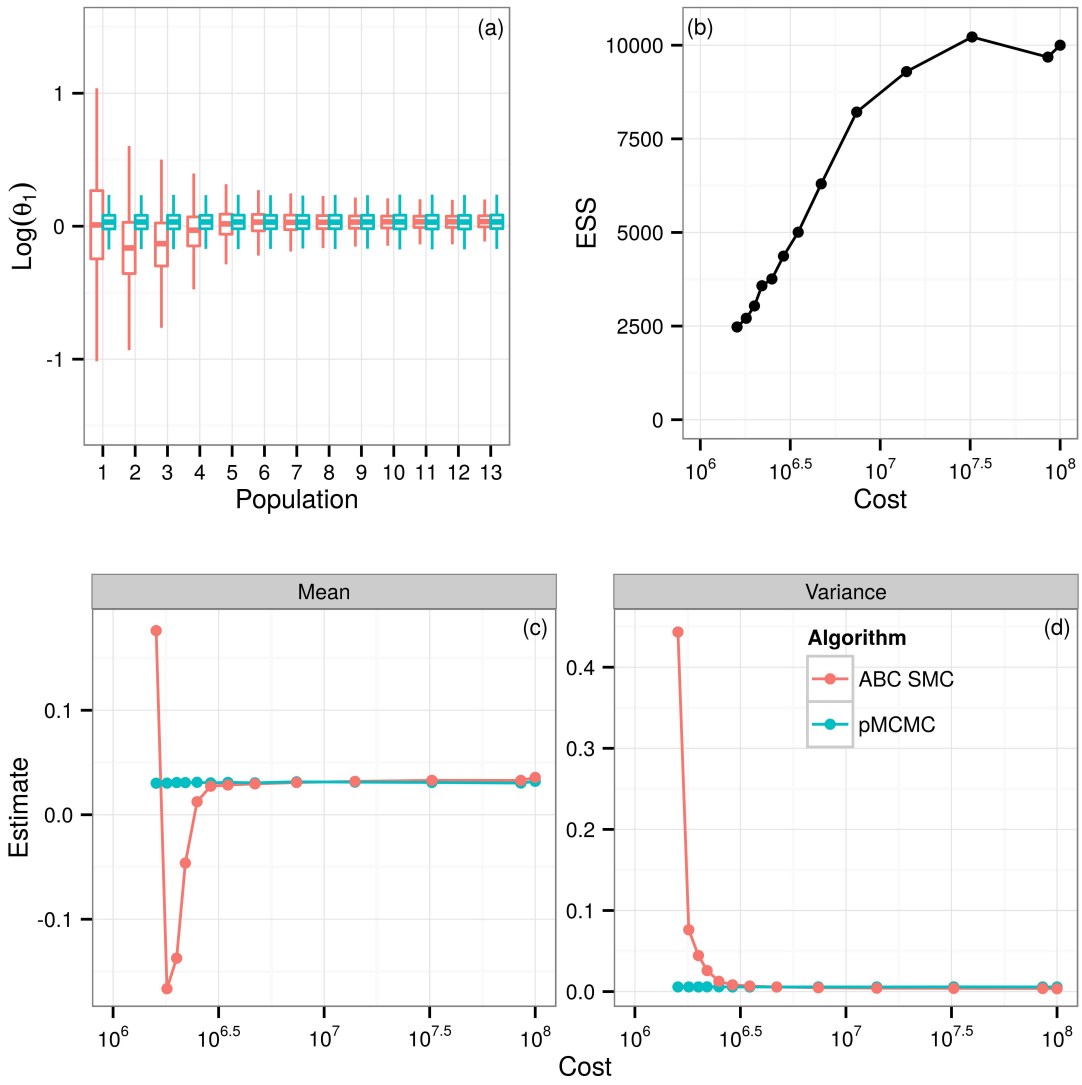


Figure 5.18: Further diagnostic information for assessing the relative performance of ABC SMC and pMCMC for the Lotka–Volterra predator prey example given data set D^1. (a) gives box plots of the distribution, the x axis represents the time t in the sequence of distributions in the sequential sampler. (b) is the effective sample size of the pMCMC posterior. (c-d) are estimates of posterior mean and variance respectively obtained with each scheme.

Data sets D^2_*

Results for each of the D^2_* data sets are reported in the same way as for those in the previous section. In these examples the average number of particles required for D^2, D^2_u, D^2_p and $D^2_{u,p}$ are 132, 370, 70 and 260 respectively and the number of itera-

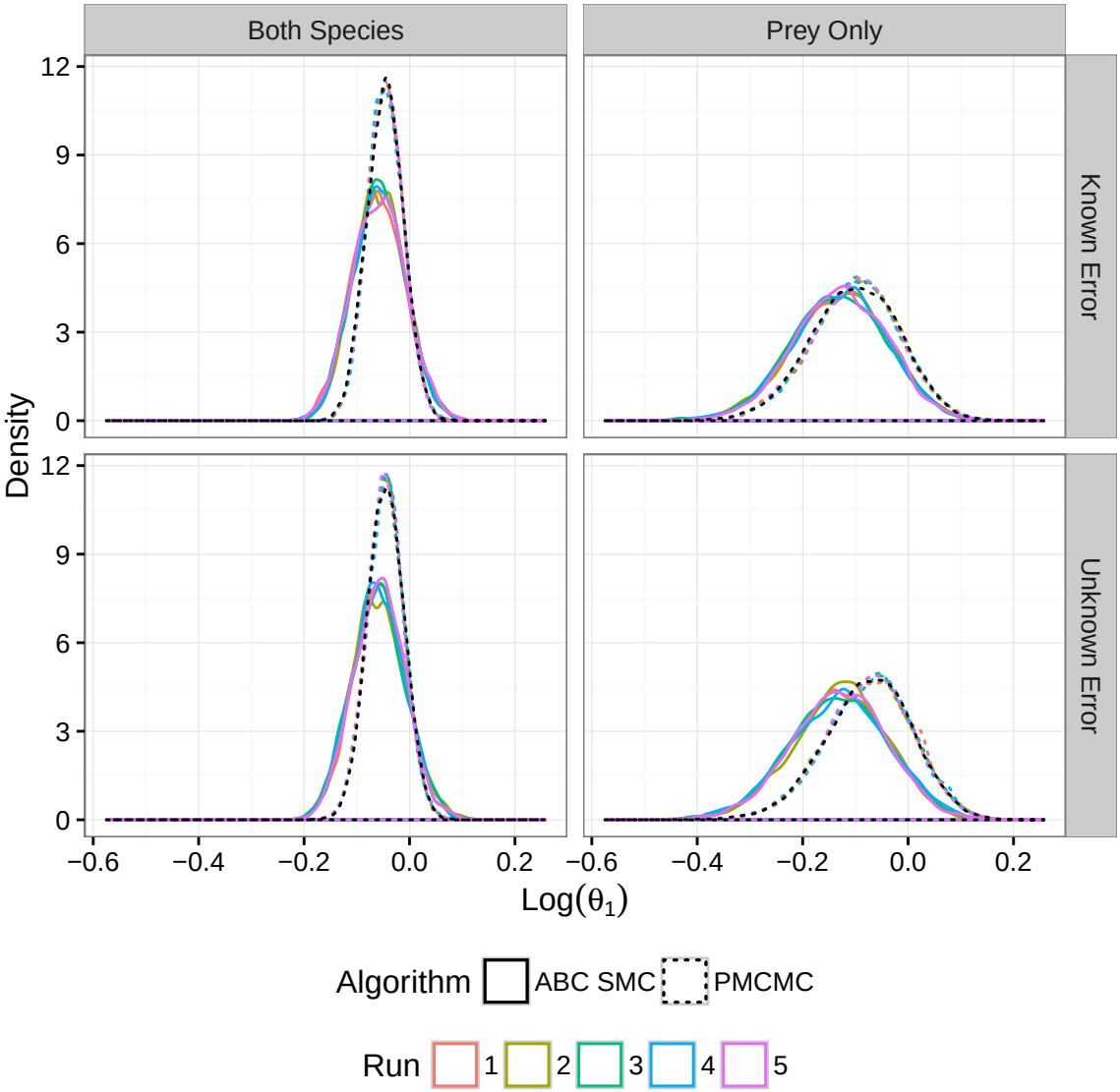


Figure 5.19: Posterior distributions for $\log(\theta_1)$ given 5 repeats for each of the observation regimes using the $\mathcal{D}^2_*$ collection of data sets. A long pMCMC run with a large number of particles to be used as a reference to the truth are in black.

tions that the SMC sampler achieved were 12.2, 11.8, 11 and 10.8 on average.

Figure 5.19 gives the posterior distributions for the D^2_* collection of data sets over 5 repeats for each of the 4 different observation scenarios. In this slightly longer time series there is a notable difference in the comparison between posteriors each of those 4 cases. Each pMCMC run performed well, being almost indistinguishable from the reference true distribution on the plots. Where both species are observed the approximate posterior distributions have a mode which is quite close to the true

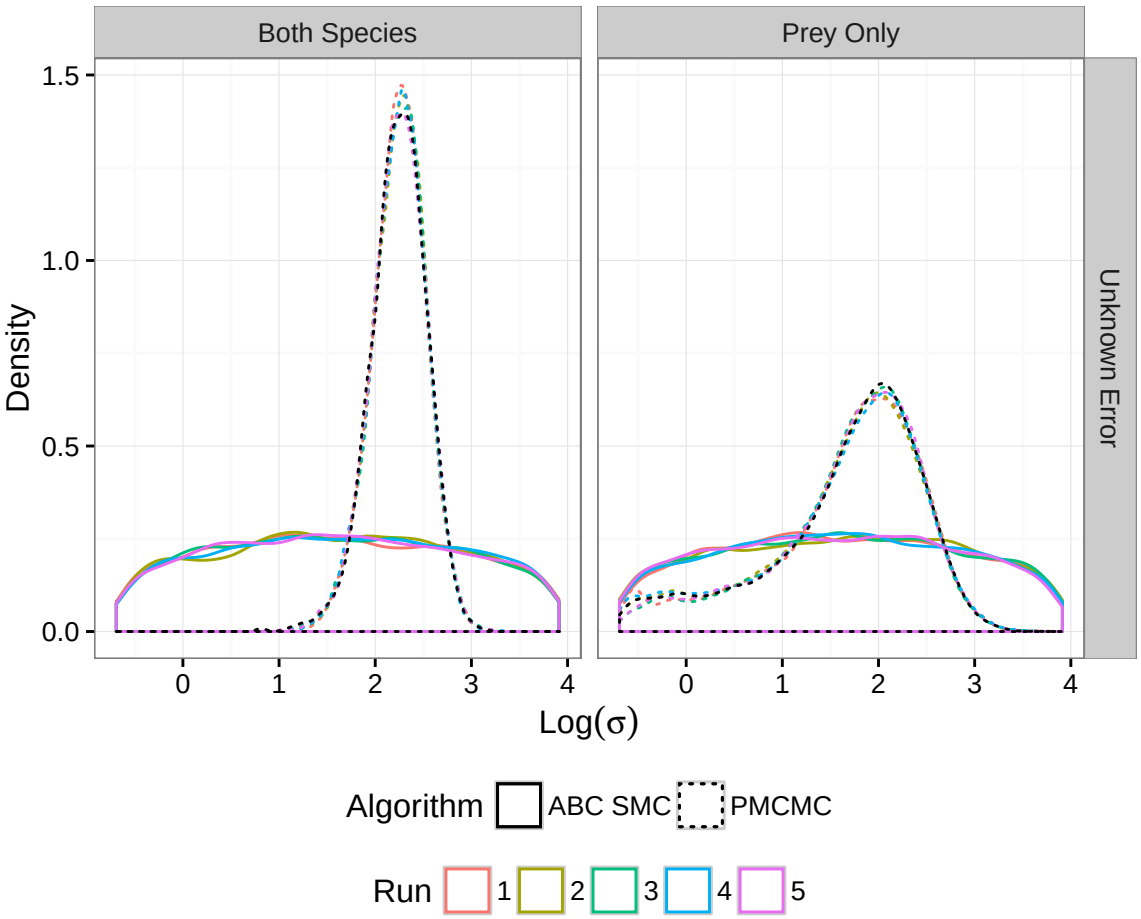


Figure 5.20: Posterior distributions for $\log(\sigma)$ in the D_u^2 and $D_{u,p}^2$ data sets. The true value used in the generation of the data is $\log(\sigma) = 2.30$.

posterior with perhaps a slight tendency to under–estimate, but have notably higher posterior variance. Where only observations of the single species are available the ABC scheme has a tendency to underestimate the parameter, with this behaviour more evident in the case where measurement error is also to be inferred. As with D_*^1 the results for the other reaction rate parameters exhibit similar comparisons to those shown.

Figure 5.20 shows that in light of more observations pMCMC can infer the noise parameter $\log(\sigma)$ well when both species are observed and moderately well with only prey observations. This is in contrast to the results that were found with the shorter time series data sets of D_*^1. However ABC again fails to infer anything meaningful about the measurement error, with little departure from the uniform prior. This can be explained as Wilkinson (2013) showed that in fact ABC targets a

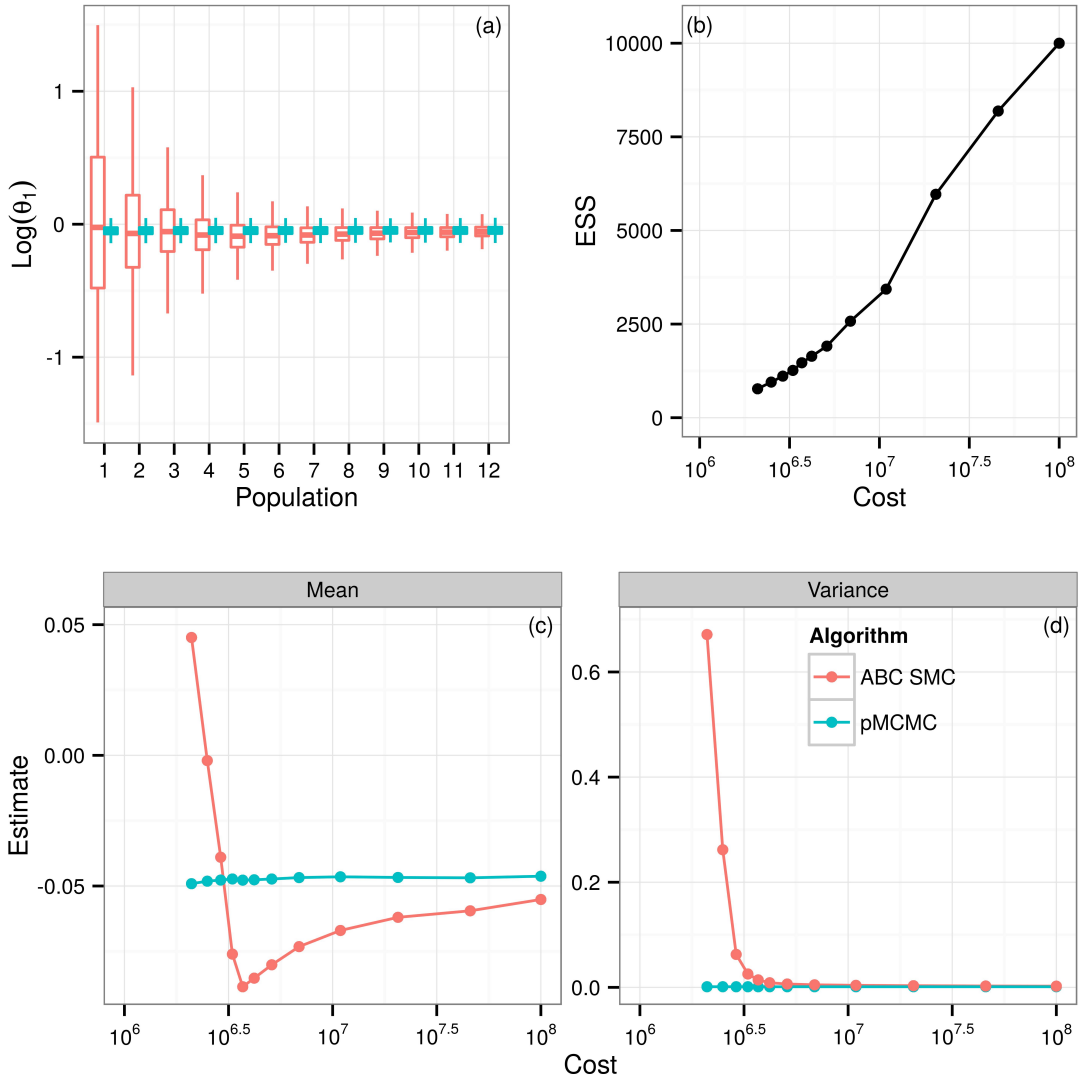


Figure 5.21: Box plots, (a), showing the posterior learning for each of the algorithms broken down by computational units. Each posterior sample through the sequence using ABC SMC is shown. The corresponding pMCMC boxplots show posterior inference using only information gained subject to the same computational budget as the ABC SMC. This gives insight into how the two algorithms compare throughout the experiment. (b) shows the effective sample size of the pMCMC sample when broken into these computational groups, (c-d) are posterior estimates of the mean and variance respectively given the two algorithms. These results are for $\log(\theta_1)$ given one run of each of the algorithms using the $\mathcal{D}^2$ dataset.

misspecified model in this context. That is the model of interest with the inclusion of independent noise whose variance is dependent on the tolerance threshold.

Figure 5.21 gives additional insight into the relative performance of each algorithm. Figure 5.21(a) shows the evolution of the shape of the distribution against the cost to obtain it. The plots are obtained using just a single repeat from each algorithm for the D^2 data set but are representative of the other repeats and data sets. Comparing to the results for the D^1 data set the initial stages of ABC SMC are poorer, compared to the relevant pMCMC posterior and movement towards the same distributional shape is slower. The effective sample sizes for the pMCMC posterior starts smaller and rises at a slower rate with computation time that was the case for D^1, with more data we are requiring to run our algorithm longer to get a good sampler. As before the shape of the pMCMC induced posterior distribution changes little over the course of the sampling run. When considering posterior summaries, the mean and standard deviation estimates are stable in pMCMC from an early stage whereas under ABC SMC sampling, they are less so at the early stages before slowly converging towards those found under particle MCMC.

The overall theme given these data sets is that pMCMC is favourable. Posterior mean and variance estimates remain fairly constant from a relatively early stage in the computation and the shape of the distribution is maintained. Obtaining a large, uncorrelated posterior sample requires that the chain must run for a long time, however running ABC SMC for the same number of model realisations does not yield better results. Whilst given the full budget it is arguable that the approximate posteriors are close, most notably under D_p^2, in the earlier stages of ABC SMC the approximation is much greater and inference is poorer as a result. Having made such statements however it is important to bear in mind that the tuning of pMCMC has been omitted from the computational cost of obtaining the posterior distributions reported. This issue will be addressed in due course.

Data sets D_*^3

The D_*^3 collection of data sets are the longest of the time series considered for the simulations experiments using the Lotka–Volterra example. The difference between the approximate posteriors and those obtained via particle MCMC are more pronounced given the longer time series as shown in figure 5.22. As the number of observations and length of the time series has increased further the ABC SMC scheme struggles to recover a good approximation to the truth more than before. Given

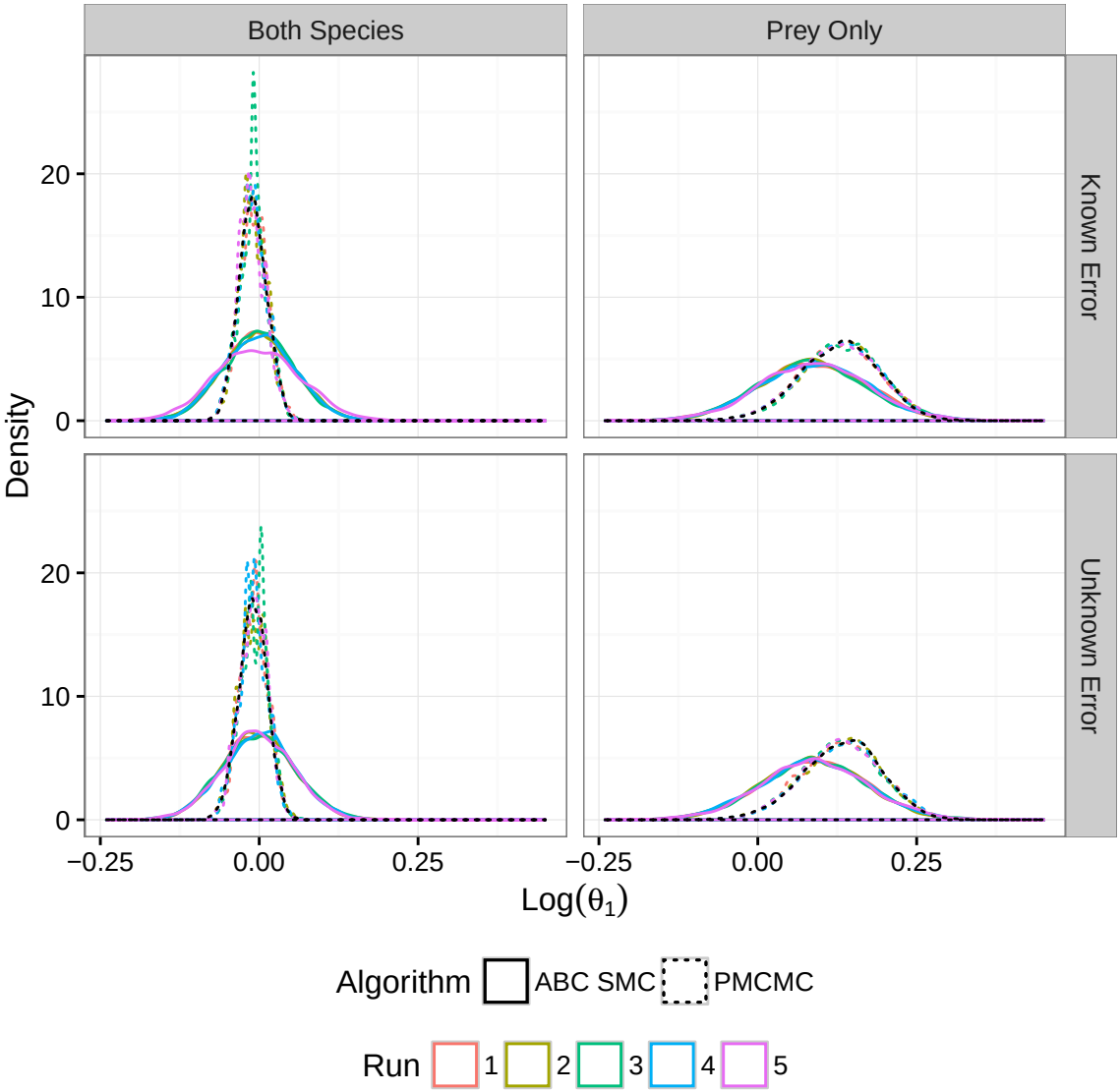


Figure 5.22: Posterior distribution for $\log(\theta_1)$ given the D^3_* collection of data sets. A reference true posterior distribution obtained by a long run of pMCMC with a large number of particles is shown in black.

observations of both species, the posterior mode is similar for both approaches, with both identifying the true parameter value $\log(\theta_1)$, used to generate the data, well. However as was observed when we increased the length of the time series between D^1 and D^2, increasing this further in D^3 has lead to greater posterior variance in the approximate posteriors. We retain a greater level of inaccuracy in our approximation for the larger data sets with the tails of the distribution over represented. With only prey observations available the ABC scheme, as was observed in the previous results, has a tendency to underestimate this parameter.

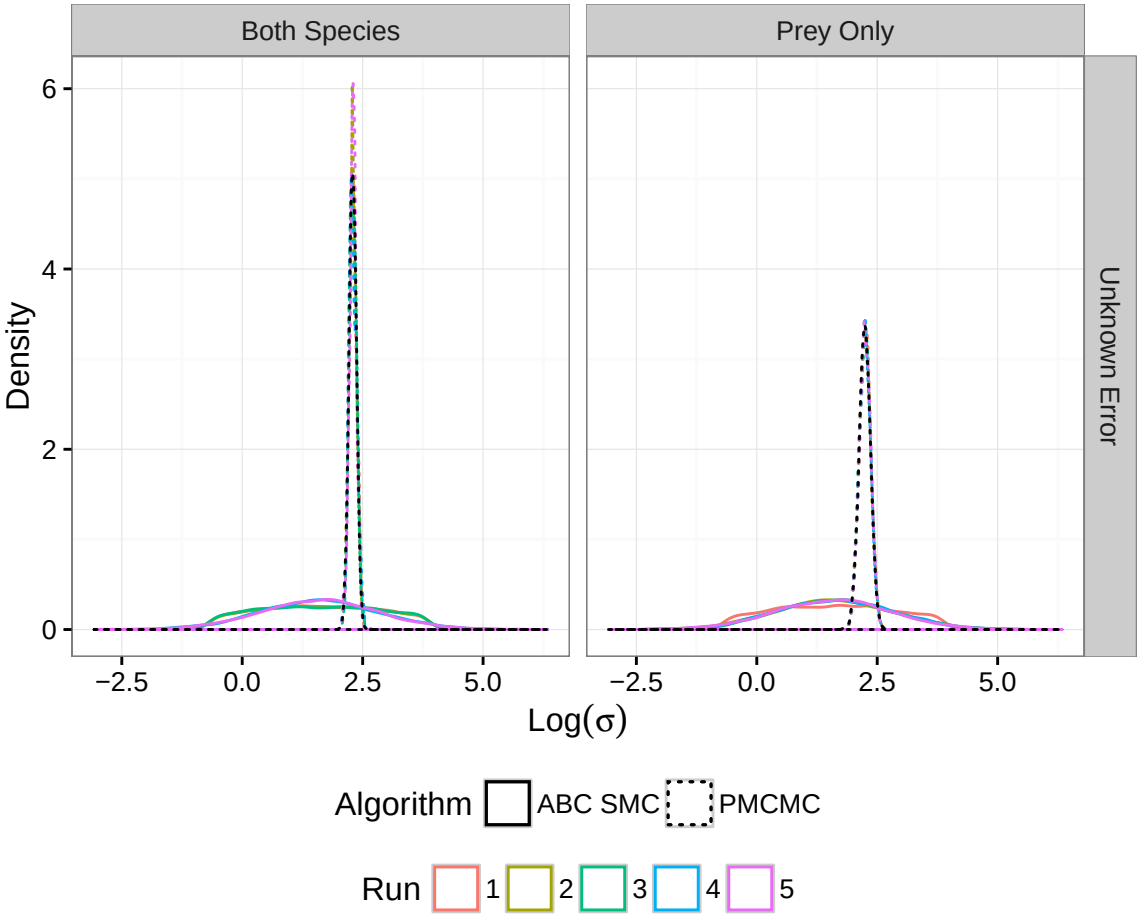


Figure 5.23: Posterior distributions for $\log(\sigma)$, the measurement error noise parameter for the D_*^3 collection of data sets. The true value used to generate the data is $\log(\sigma) = 2.30$ with a reference true posterior distribution, obtained via a long run from a particle MCMC algorithm with a large number of particles shown in black.

It should also be noted that under the computational restrictions imposed for the experiment, given observations on both species, pMCMC also struggles. The density plots are much less smooth, indicating that we have a poorer sample with smaller effective sample size.

Given the more abundant data, pMCMC infers the noise parameter well whilst ABC SMC continues to perform poorly in this respect as shown in figure 5.23. This is consistent with the findings from both the D_*^1 and D_*^2 data sets with ABC SMC showing little departure from the prior distribution for this model parameter. The additional diagnostic information in Figure 5.24 continues the observed trends from the previous results.

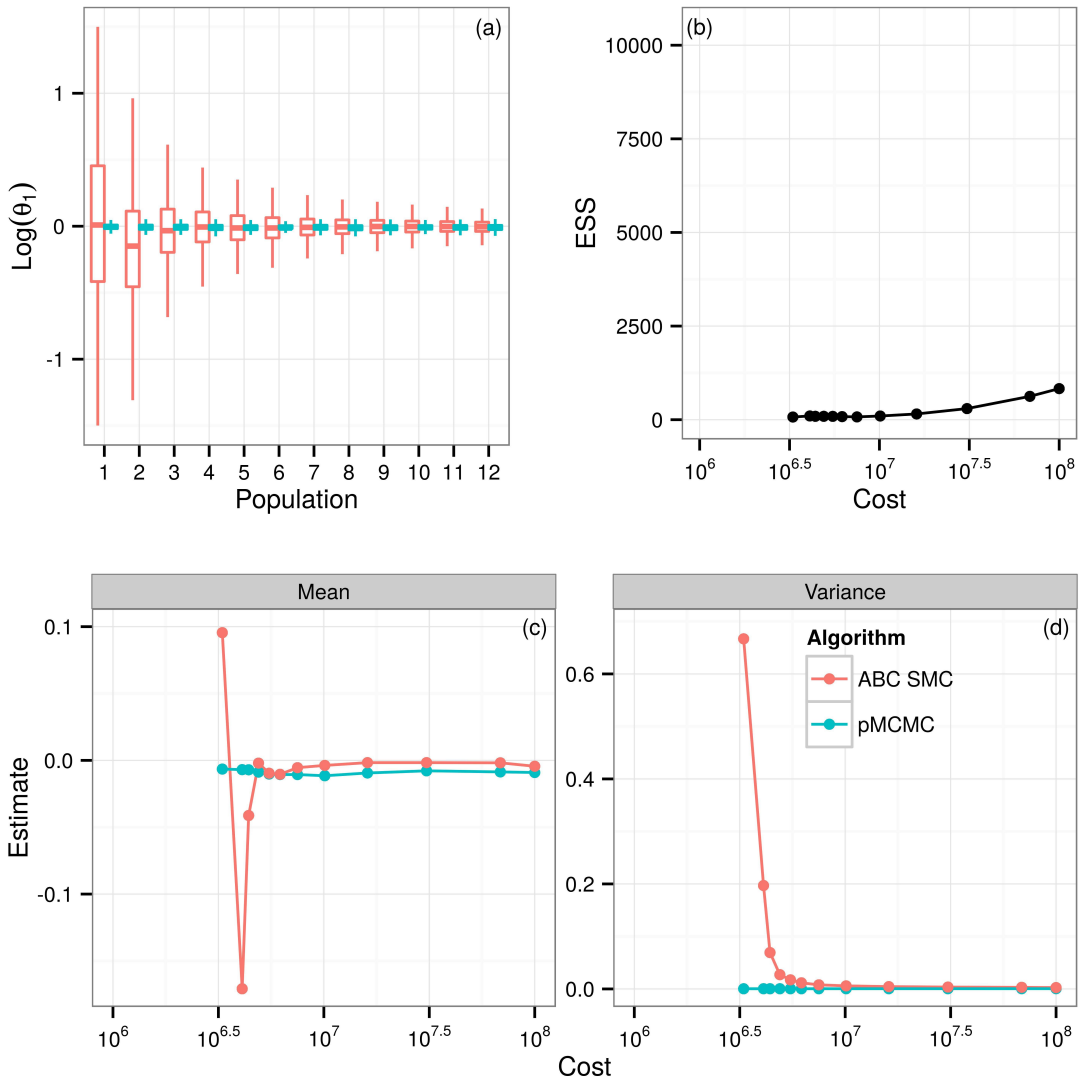


Figure 5.24: Additional analysis of the relative competitiveness of the ABC SMC and particle MCMC schemes for the D^3_* collection of data sets.

The sequence of box plots, Figure 5.24 (a), shows that the level of approximation at the early stages of the ABC algorithm is large, and given the increasing dimension of the data, the length of time and hence the computational expense required for ABC to yield a reasonable approximation to the true posterior as shown by pMCMC is greater. The larger data means that obtaining a diverse sample using pMCMC is more taxing, requiring large numbers of model realisations to obtain reasonable effective sample sizes however the shape of the distribution and estimates of mean and variance remain stable from an early stage.

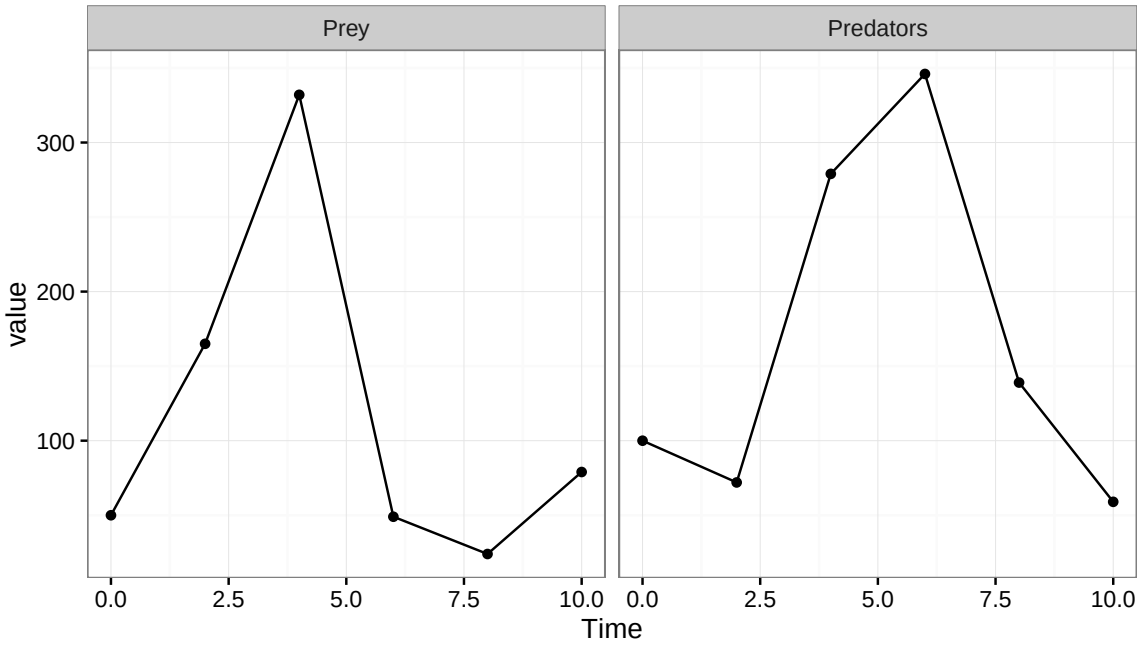


Figure 5.25: A short time series of uncorrupted observations from a Lotka–Volterra predator prey model.

5.5.3 Effect of measurement error

One further consideration to be made is how the comparison is affected by the size of the measurement error. It is well understood that obtaining marginal likelihood estimates using a particle filter on data with smaller measurement error requires a larger number of particles. The relative cost of obtaining a sample from the posterior distribution when measurement error is small is then higher. We can examine how this changing accuracy in the measurement process affects the relative efficiency of the two approaches considered here.

To do so we consider two sets of examples. A short time series of observations from a Lotka–Volterra model with vague prior information on the rate parameters. A further example, based on the Schlögl model introduced in section 2.5.4, considers a longer time series and additional species with an informative set of prior distributions on the rate parameters.

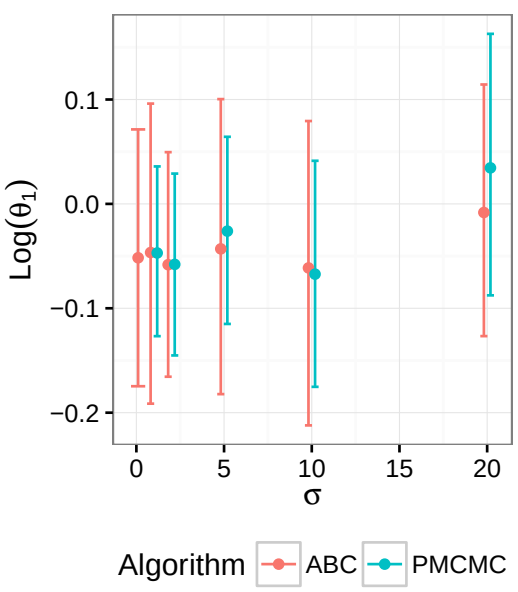
Lotka–Volterra model

For the Lotka–Volterra example we consider a short time series of observations of both species at each of 6 regularly spaced time points. The unobserved true values of the process are shown in figure 5.25. The true observations are then corrupt by a Gaussian measurement error kernel with zero mean and differing standard deviations, $\sigma = (0.1, 1, 2, 10, 20)$.

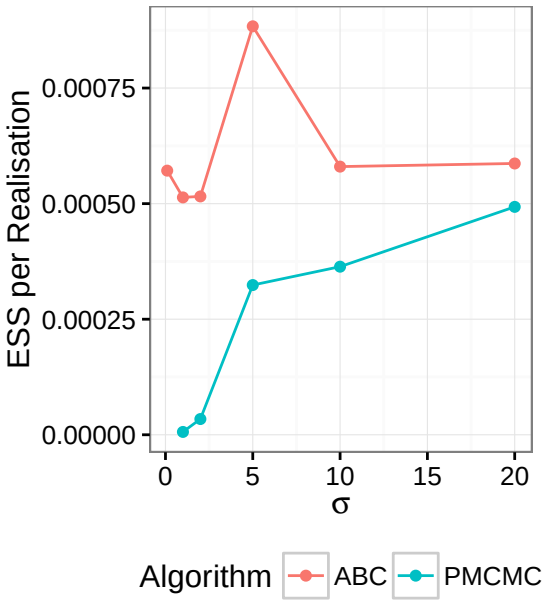
The notion of a computational budget is employed once again for this example where the maximum budget allowed for each algorithm is 10^7 model realisations. The cost of obtaining the number of particles to be used in a particle filter, where appropriate, is discounted from the allocation and particle MCMC chains are initialised at the true parameter values that generated the synthetic data. For each example the same Poisson priors on the initial state of the system are used as in section 5.5.2, the measurement error is assumed to be known in each case. However, since access to the true posterior distribution for each is unavailable a short tuning run of 2000 iterations with a small proposal variance was used to help optimise the exploration of the parameter space. This cost was deducted from the permitted budget.

For the smallest value of σ considered the number of particles required to obtain estimates, $\hat{l}(\Theta)$ such that $\mathrm{Var}(\hat{l}(\Theta))$ was sufficiently small was in excess of 50,000. At this point we stopped searching as the number of particles is sufficiently high that the cost tuning run to optimise proposal variance would have been 10 times greater than the allowed computational expense.

Figure 5.26 shows results of applying the ABC SMC and particle MCMC under these constraints. Each posterior distribution consists of a minimum of 5000 where budget permitted this to be the case. For pMCMC longer chains were thinned such that 5000 samples were retained. The error bars showing the central 95% and means of the resultant posterior distributions are comparable for each value of σ. However the ESS per model realisation shows that the ABC scheme is more computationally efficient in all of the cases shown here. The caveat to this is that with pMCMC we know that the target is exact whereas with ABC we have introduced some error. The ABC approach also permits posterior inference in the case of extremely small measurement error, something that could not be obtained using particle MCMC. The relative efficiency of each scheme seems to converge as the measurement error grows. At $\sigma = 1$, the first for which we can obtain a sample using pMCMC, the ESS



(a) Mean and 95%–ile posterior error bars for θ_1 given the observations in figure 4.2 subject to changing measurement error.



(b) Effective sample sizes per model realisation of the two approaches for the Lotka–Volterra posterior rate parameter distributions.

Figure 5.26: Analysis of the posterior distributions for the prey birth rate parameter in the Lotka–Volterra model for differing observation error.

per model realisation for ABC SMC is approximately 78 times greater, however by $\sigma = 20$ this value is just 1.2.

Schlögl model

The Schlögl model, introduced in section 2.5.4, is a test model well known for the fact that given certain reaction rate parameters and initial conditions, the evolution of the species exhibits bimodal stability. Here we consider a single trace of the process given $\Theta = (3 \times 10^{-7}, 10^{-4}, 0.000773, 3.276)$ and $X_0 = (250, 10^5, 2 \times 10^5)$. Observations are made subject to a Gaussian measurement error kernel with two different values of standard deviation, σ. The uncorrupted synthetic data set is shown in figure 5.27 consisting of 21 observations on each of 3 species. 2 values of σ, $\sigma = 1$ and $\sigma = 10$, assumed to be known in each case, were compared over 5 replicates of each experiment to assess the ability of sequential ABC and particle MCMC to infer the 4 reaction rate parameters.

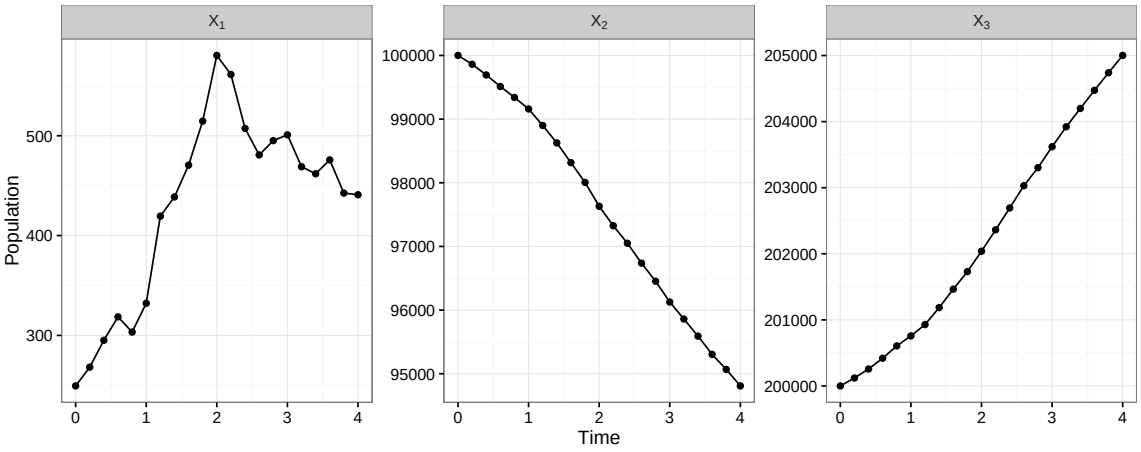


Figure 5.27: A synthetic data set from a Schlögl model. 21 observations on each of 3 species over regular time intervals. Reaction rate parameters are chosen as $\Theta = (3 \times 10^{-7}, 10^{-4}, 0.000773, 3.276)$ with initial state $X_0 = (250, 10^5, 2 \times 10^5)$.

Prior information for this example is much more informative than our previous test case with a larger permitted computational budget of 10^8 model realisations. Specifically

$$\log(\theta_i) \sim \mathcal{N}(\log(\theta_i^{obs}, 0.5^2), \quad i = 1, \ldots, 4, \tag{5.18}$$

a Gaussian distribution on the log scale, centered at the true values that gave rise to the synthetic data with small standard deviation 0.5. In addition, knowledge of the initial state, X_0 is assumed.

Consistent with the results found in a number of other results in previous sections the posterior distributions for this example show that ABC tends to over estimate the posterior variance due to the error introduced by the approximation $\epsilon > 0$.

Figure 5.28 shows posterior distributions for the 4 rate parameters of the Schlögl model using both the ABC SMC and pMCMC schemes for the larger measurement error case, $\sigma = 10$. Whilst the approximation to the true posterior distribution under ABC is not perfect it should be noted that within the time frame of our computational budget (10^8 model realisations which on our fast Intel core i7-2600 powered desktop clocked at 3.4 GHz corresponds to approximately 25 hours of CPU time) the particle MCMC sampler had an average effective sample size of 309 across the replicates. So whilst the location and general shape of the distribution are well inferred using pMCMC, a diverse posterior sample is expensive to obtain.

This problem becomes much more pronounced with smaller measurement error.

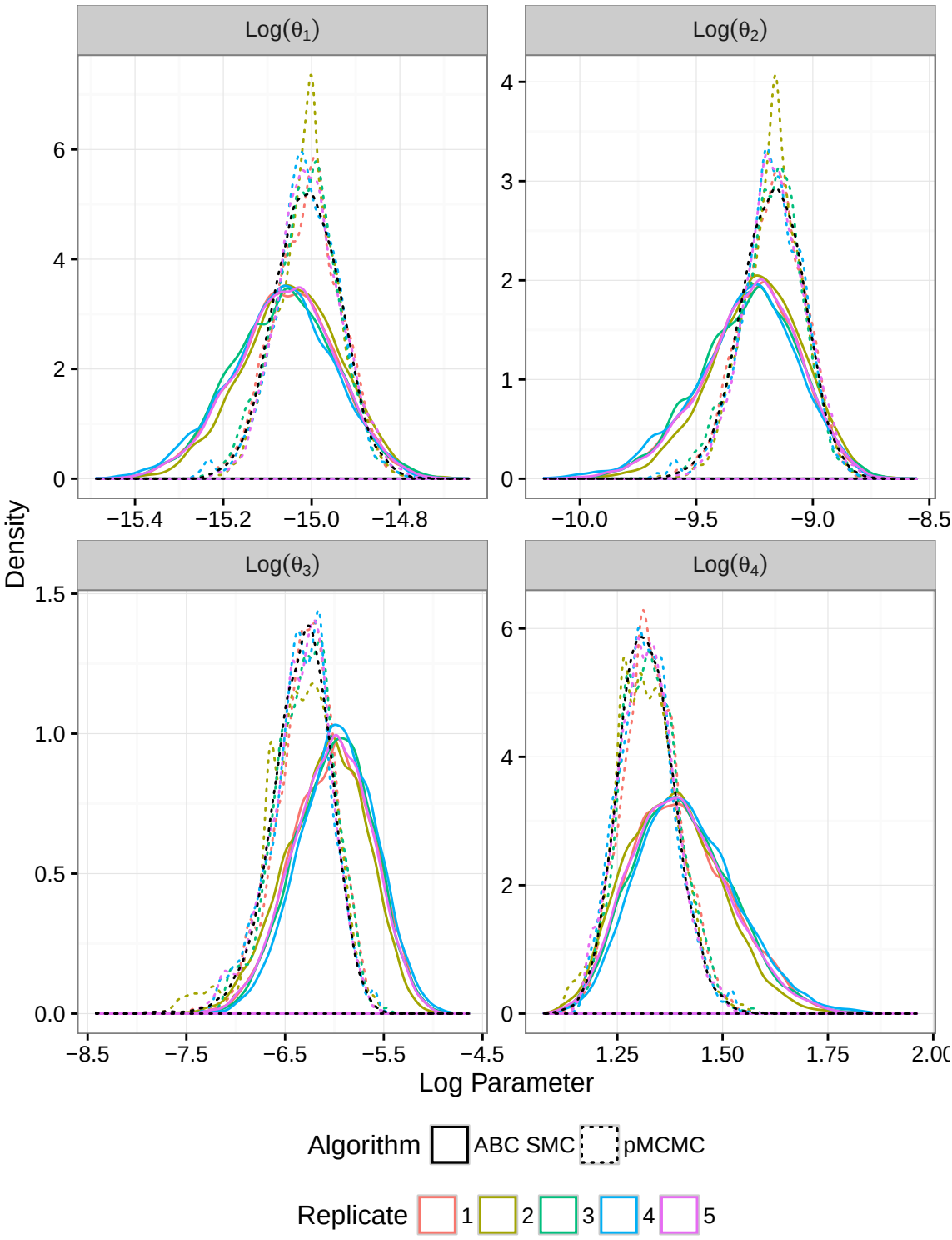


Figure 5.28: Posterior inference for the 4 rate parameters of the Schlögl model given observations with Gaussian measurement error standard deviation $\sigma = 10$. True reaction rate parameters used to simulate the synthetic data are $\Theta = (-15.02, -9.21, -7.17, 1.19)$.

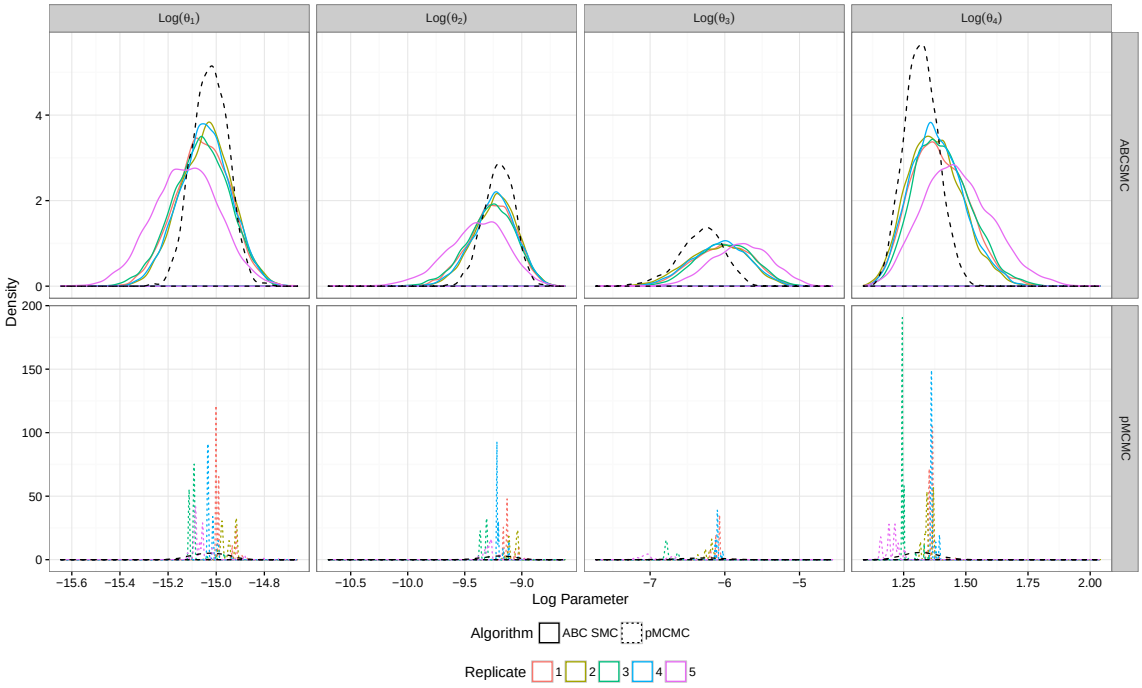


Figure 5.29: Posterior inference for the 4 rate parameters of the Schlögl model given observations with Gaussian measurement error standard deviation $\sigma = 1$. True reaction rate parameters used to simulate the synthetic data are $\Theta = (-15.02, -9.21, -7.17, 1.19)$.

Figure 5.29 shows posterior inferences for each reaction rate parameter for each sampling scheme. Results for the ABC SMC sampler are similar to those shown in Figure 5.28 and in agreement with what we found with the Lotka–Volterra example in that the efficacy of posterior learning using ABC is not affected by the magnitude of the measurement error. The posterior distributions given pMCMC in this example are very poor, because such a large number of particles are required to estimate the log of the marginal likelihood consistently the sampler expends it's budget very quickly. Put another way individual iterations of the MCMC chain take a long time to compute and so the efficiency of the scheme is poor.

These results suggest that in the absence of adequate measurement error ABC presents a more compelling argument. Approaches to improving the relative performance of particle MCMC algorithms have been explored. Golightly *et al.* (2014), on noting that model realisations from a Gillespie algorithm are expensive, consider a delayed acceptance approach. Proposed parameter values are ruled in or out as candidates for the next step in the chain using a fast approximation method for model simulation before using the more expensive Gillespie algorithm in the particle filter.

The aim is that by quickly rejecting proposals that are unlikely to be accepted the chain yields an improved effective sample size per second.

The problem of using a particle filter to estimate likelihood in the case of small measurement error was the focus of Golightly & Wilkinson (2015). The authors consider use of a Gillespie algorithm that simulates across the discrete time transitions conditional on its end points. Each of these approaches could be used to potentially increase the efficiency of the particle MCMC sampler.

5.6 Additional remarks

A recently published article, (Prangle *et al.*, 2016), proposed an improvement to the sequential ABC sampler. The author suggests that the vector of weights associated with the summary statistics in the Euclidean metric is added to the collection of algorithmic tuning parameters which are calculated online. The author uses the same Lotka–Volterra predator prey model example of Owen *et al.* (2015) and the previous section with the same set of prior distributions. It is noted that the variability of each of the summary statistics, $S(\mathbf{Y}) = \mathbf{Y}$, is markedly different under the prior distribution $\pi(\Theta)$ than under the final target approximating the posterior $\pi_{\epsilon_T}(\Theta)$. This is due, in the case of the Lotka–Volterra example, to the fact that the evolutionary behaviour of the species is stable only for a relatively small region of the prior rate parameters, with much of the space characterised by either species extinction, or an explosion of prey levels. It therefore does not necessarily make sense to use the same set of weights in the metric for each iteration of the ABC sampler.

The results in section 5.3.5 found that given the prior distribution specified, weighting the summary statistics led to poorer posterior inference. This was due to the fact that over the prior parameter space, the variability of the summary statistics meant that the subsequent weights chosen yielded an acceptance criteria too heavily influenced by the initial state of the system.

The results of Prangle *et al.* (2016) suggest that ABC SMC can be made to be more efficient than the above comparisons show giving additional favour to ABC SMC.

5.7 Conclusions

In addition to the conclusions made in section 5.3.6 we have now made comparisons between the competing approaches to likelihood free inference within a range of example scenarios using stochastic kinetic models. ABC SMC holds numerous advantages over the other ABC based samplers, namely improved exploration of parameter space through an iterative procedure of informed proposals to target a sequence of bridging distributions from prior to desired posterior. In addition its tuning parameters are largely chosen online. In practice the only one we need worry about is the quantile parameter α at stage 5 of Algorithm 8. This guided exploration through a sequence of posterior distributions whose tolerance decreases gradually helps to retain more favourable acceptance rates than a simple rejection sampler and hence yields more efficient posterior sampling. ABC SMC is also preferable to a MCMC based sampler for similar reasons. ABC MCMC has the additional difficulty that choosing appropriate tuning parameters for the random walk kernel, the initial point to start the chain, Θ_0, and the requirement to choose an ϵ is non–trivial. Whilst ABC MCMC can be more efficient than a simple rejection sampler the sequential importance sampler is more efficient again. A further tipping point against the MCMC based sampler is that there, as far as this author is aware, is no easy way to reduce the tolerance whilst maintaining the stationary distribution of the chain. Bortot *et al.* (2007) did propose an alternative which augments the Markov chain with a tolerance parameter. This allows a chain, with an appropriate prior distribution on the tolerance, which encourages small values of ϵ but will allows for potentially greater mixing, and improved convergence properties by allowing ϵ to vary. However it is not an adaptive approach which aims to reduce the tolerance to further improve posterior approximation. With both the rejection sampler and sequential sampler a reduction in tolerance can be obtained simply by allowing the algorithm to run for a greater length of time.

Whilst the comparisons drawn between the approaches to inference explored within this chapter are of interest, it should be acknowledged that a single simulated draw of the data may lead to bias, which could in turn lead to over interpretation of the findings. Ideally, future work could expand upon the comparisons made here to address this concern and ensure that the conclusions drawn remain valid.

The discussions here disclude consideration of additional aspects of each of the

algorithms. In particular, amenability to parallelisation and the tuning problems experienced with particle MCMC touched upon briefly in some of the examples in this chapter. This discussion is saved for the introduction and motivation for the work in the following chapter. That chapter is based on and is an expansion of the work published by Owen *et al.* (2014).

Chapter 6

Hybrid ABC pMCMC algorithm

6.1 Introduction

One of the issues with particle MCMC, as explored in the previous chapter, is that initialisation and tuning of the algorithm is difficult. The results presented in section 4.4 have largely ignored the issue of tuning the pMCMC algorithm when comparing it's relative efficiency against ABC methods. The posterior inferences suggest that pMCMC is the better choice in most cases, excluding those when the measurement error model variance is small. However in each case the pMCMC algorithm has been tuned prior to consideration of the computational budget, relying on knowledge of the posterior distribution. In reality access to the posterior distribution is unavailable, since that is what we are trying to infer. The true computational cost of the pMCMC inference then is somewhat higher than the comparisons suggest. For ABC SMC the proposal variance and sequence of tolerances are chosen adaptively meaning that the initialisation and tuning costs are small.

The motivation for choosing to tune pMCMC using information from the posterior for the purpose of comparison was due to the fact that early experimentation found that for the models considered, the cost of initialising the algorithm was large, in many cases requiring more model realisations that our budget permitted. The decision was made so as not to detract from being able to draw conclusions between

the relative efficiency of the two approaches but this problem can not be simply ignored as it poses a very real and significant challenge to a practitioner aiming to make use of particle MCMC. Choosing an initial state of the chain Θ_0, number of particles, N, and random walk proposal variance Σ_q such that posterior sampling is productive is difficult with little prior knowledge.

The number of particles, including the region of parameter space on which to tune this, is of crucial interest when it comes to the efficiency of the resultant sampler. In addition if the initial parameter vector in the Markov chain is in a region of low posterior density, the burn–period of the sampler can be prohibitively large. Initial attempts to find Θ_0 involved sampling from the prior distribution, calculating estimates of the likelihood and then choosing the parameter vector which maximised this. At this stage however we do not know what a good choice for N is. Because the particle filter gives estimates of likelihood with a large variability away from the posterior mode and we wish to avoid choosing a poor Θ_0 that had an unusually high likelihood by chance, this step typically involves using either a large number of particles or a number of repeats at each parameter vector sampled from $\pi(\Theta)$. This requires a large number of model realisations and has a high computational cost. Conditional on the hopefully informative choice of Θ_0 we then attempt to tune the number of particles to be used in the filter, N, for the main sampling run. Since what we would like to do is choose N as small as possible, as we shall be running the particle filter at every iteration of our MCMC scheme, we started with small N and steadily increased it until $\mathrm{Var}(\hat{l}(\Theta))$ was suitably small. Again this step has non–negligible expense requiring multiple estimates $\hat{l}(\Theta)$ for each N to assess the variance for a potentially large number of candidate N values. Finally we desire a random walk proposal variance Σ_q that facilitates good properties of the resultant Markov chain. This step typically involves a pilot MCMC run, using a very small proposal variance and using the resultant distribution to inform the choice of Σ_q.

To give some context to this we found that in practice using 1000 particles in a particle filter to estimate likelihoods given the D^2 data set of the Lotka–Volterra examples for 2000 parameter vectors drawn from $\pi(\Theta)$, repeating each 10 times, and maximising over the average to choose Θ_0 was not enough to guarantee that the resulting choice had good posterior support. This alone used up 20% of the allocated computational budget (which corresponds to significant wall clock time),

before consideration of choosing N and Σ_q. A poor choice of Θ_0 then propagates additional expense throughout as the number of particles N required to satisfy the variance criteria is often greater in the tails of the posterior distribution. Add to this estimation of Σ_q and then the appropriate burn–in period of the resulting Markov chain and it is easy to see that this operation becomes very computationally taxing, even for a relatively simple model with a relatively small number of observations. Further justification for omitting this cost from the comparisons is that the resultant comparisons are invariant to the initialisation and tuning mechanism employed. The method by which we first attempted to do this is only one approach and whilst there are a number of ad hoc approaches to tuning a MCMC algorithm to the best of our knowledge there exists no principled approach by which to tune particle MCMC under poor prior knowledge.

Further to this discussion, the conclusions drawn ignore other aspects of each algorithm which may contribute to the overall efficiency of the samplers. One prominent consideration is an algorithms amenability to parallelisation. Algorithms which parallelise well can exploit ever increasing processing power of multi core computer architecture and multi node clusters often resulting in dramatic reduction of real time execution. Again a conscious decision was made to disclude this aspect from the comparisons in the previous chapter, motivating the use of a computational budget rather than wall clock time, to allow the conclusions to be relevant regardless of such optimisations. However it is an important consideration highly relevant to computational statistics in practice. ABC methods are typically trivially parallelisable, where MCMC schemes are extremely difficult to parallelise. In the specific case of particle MCMC it is possible to parallelise the bootstrap particle filter however thread communication overhead can often be high. One of the trade offs with the ABC however is that posterior learning can be poor, and the decision on the summary statistics is non obvious. In this chapter we explore a hybrid algorithm which attempts to exploit the relative strengths of both ABC and particle MCMC. Namely the exactness of a posterior distribution under pMCMC sampling with the ease of tuning and parallel computation of ABC. The motivation for it's construction is to provide a principled approach to tuning and initialisation of a sampler which retains the exact target of interest and to allow increased efficiency by allowing exploitation of parallel computing.

6.2 Likelihood free Bayesian techniques in parallel

6.2.1 ABC methods in parallel

ABC techniques based on rejection or importance sampling, rather than a scheme constructed around a Markov chain, such as those described in algorithm 6 and algorithm 8 are often amenable to parallelisation. All proposed parameters are independent samples from the same distribution, either a prior distribution or a suitable importance density, and the acceptance criteria for a given proposal does not depend on any other parameter vectors. Hence the bulk of computation for an ABC sampler can be run in parallel, greatly reducing overall CPU time required to obtain a sample from the target. On top of this, coding of a parallel ABC algorithm adds little additional complexity over a non-parallelised version. This is particularly true for the simple rejection sampler where a simple parallel implementation effectively just runs a sampler on each of N processors. No communication between threads is needed throughout the computation, except to collate the final collection of samples. For a sequential ABC scheme some thread communication is necessary as we move from one distribution in the sequence to the next, for calculation of the tuning parameters to be updated, but typically this is small in comparison to the work done in forward simulation.

6.2.2 Parallel MCMC

MCMC algorithms are somewhat less amenable to parallelisation due to the reliance of the sampler Markov property. Exploration of a target distribution is dependent on the current state of the Markov chain. Essentially there are two possible options to be considered, parallelisation of a single MCMC chain or the construction of parallel chains. For an in depth discussion see Wilkinson (2006). Parallelisation of a single chain in many scenarios is somewhat difficult due to the inherently iterative nature will not be discussed in detail here. Running parallel chains, however, is straightforward by the same argument that rejection sampling is easy. If separate chains are completely independent of one another, then parallelisation of multiple chains is akin to running a single MCMC chain per available CPU core or cluster

node.

6.2.3 Parallel chains

In practice, chains initialised with an arbitrary starting point that target the same distribution, will be, after an appropriate burn–in period, sampling from the same stationary distribution. The argument as to whether it is best to run a single chain for a long period of time, or to run many short chain, in the context of a serial implementation, is still subject to debate with each approach having it's own merits. A single chain need only suffer any necessary burn in period once, while numerous chains allow the practitioner to better diagnose convergence to stationarity and validate results obtained. The argument changes however on consideration of a parallel implementation. Indeed, if burn–in periods are relatively short, running independent chains on separate processors can be a very time efficient way of learning about the distribution of interest.

Burn–in is a potential limiting factor on the scaling of the performance gain to be had when employing parallel chains with the number of processors. The greater the period spent converging to the stationary distribution of a chain, the more time each processor has to waste computing samples that will eventually be thrown away.

The theoretical speed up given N processors to obtain n stored samples with a burn–in period b is

$$\text{Speed-up}(N) = \frac{b+n}{b+\frac{n}{N}}, \tag{6.1}$$

which is clearly limited for any $b > 0$, as $N \to \infty$ (Wilkinson, 2006). A "perfect" parallelisation of multiple chains then is one in which there is no burn–in period. A chain with that requires no time to converge to the stationary distribution can be obtained by initialising with an independent draw from the target. In this ideal situation, the performance gain when run on N processors is then of factor N. In most practical situations, initialising with samples drawn from the target is not possible, motivating the use of MCMC in the first place. Hence it is typically impossible to implement this perfect situation.

6.2.4 Particle MCMC in parallel

The details of a pMCMC chain using a bootstrap particle filter have already been introduced (section 4.3). The sequential Monte Carlo approach to estimation of likelihood does mean that particle MCMC as a stand alone approach is more suitable to parallelisation than standard, analytic likelihood function based MCMC. At a given iteration of the particle filter, the individual particles are independent meaning that their propagation through the model can be run on separate processors. However outside of this there is need for frequent thread communication in updating weights in the importance sampler and estimating marginal likelihood for example. This overhead typically prevents approaching the perfect theoretical maximum performance gain of N for N processors, particularly if message passing between processes has it's own non negligible expense. This being particularly prevalent on a multi node cluster for example. It is still desirable to be able to run multiple pMCMC chains in parallel if possible.

6.2.5 A principled approach to initialisation

Since it is typically not possible to initialise a MCMC chain with a draw from the desired target, we propose an approach to parallel MCMC by choosing initial parameter vectors according to samples from an approximate posterior distribution. The intuition is that if we have a reasonable approximation to the target of interest, in this context a Bayesian posterior distribution, then samples from the approximation will closely match those from $\pi(\theta|\mathcal{D})$. Because of this we expect that any burn–in period that the chain must be subjected to is very short before we are sampling from the desired target. As the burn in time b decreases, we are approaching the scenario of near perfect parallelisation of MCMC in equation 6.1. It is clear that as the approximation to the desired stationary distribution improves, the shorter we expect the burn in period to be.

The proposition is to first run an ABC scheme targeting an approximation to this distribution. As discussed in section 6.2.1 this allows us to exploit parallel hardware to eliminate a large region of prior parameter space very quickly. Conditional on the approximation, take a set of N independent samples from $\pi(\theta|\rho(\mathcal{D}, \mathcal{D}^*) < \epsilon)$ to initialise N independent, parallel, pMCMC chains each of which targets the exact

posterior distribution $\pi(\theta|\mathcal{D})$.

In some sense we can consider this process of obtaining a sample from an ABC approximation to the posterior can be thought of as an artificial burn–in period. Importantly however, the ABC sampler yields a collection of samples from $\pi(\theta|\rho(\mathcal{D}, \mathcal{D}^*) \leq \epsilon)$. An arbitrary number of MCMC chains can be initialised with draws, Θ_0, from a distribution close to the target. The artificial burn–in period need only be performed once, and can itself be parallelised. Additionally since we utilise the approximation to the target only as a guide to initialisation of the pMCMC sampler the exactness of the target is unaffected. The ABC posterior does not form part of the proposal, nor is it being utilised as a prior.

6.2.6 Random walk pMCMC using ABC

Having chosen the initial points of the chains, Θ_0, specification of a random walk MCMC scheme also requires choice of a proposal distribution, $q(\cdot)$, commonly a zero mean Gaussian distribution whose covariance matrix, Σ_q, dictates the magnitude and direction of deviations from the current state. Optimal choice of proposal variance for Gaussian random walk kernels have been derived, see Roberts & Rosenthal (2001) as proportional to the covariance of the target. For our targeted posterior, since our approach to obtain vectors for initialisation, Θ_0, yields a sample which we hope is a reasonable approximation to the true posterior, we likewise consider that the covariance of the samples from $\pi(\Theta \,|\, \rho(\mathcal{D}, \mathcal{D}^*) < \epsilon)$, denoted $\Sigma_{\texttt{ABC}}$ will be close to the covariance structure of the exact target. Our collection of samples that form candidates for initialisation, double up as a mechanism for calculating a covariance structure that provides an informative tool for the calibration of random walk innovations. By using $\Sigma_{\texttt{ABC}}$ as a proxy for the true posterior covariance, we hope to alleviate the necessity to tune Σ_q. Since in many applications this is done via pilot runs of the chain using a small proposal variance that yields samples that ultimately must be discarded, we aim to cut down on wasted CPU time. In practice we take

$$\Sigma_q = \frac{(2.38)^2}{d}\Sigma_{\texttt{ABC}}. \tag{6.2}$$

Equation 6.2 as published in Owen *et al.* (2014) has subsequently been improved upon within the context of particle MCMC by Sherlock *et al.* (2015), see sec-

tion 4.3.2. Within the context of the results presented here, it is not expected that the alternative proposal variance for a Gaussian random walk kernel will solve the problems addressed within this chapter. It therefore does not detract from the results and developments outlined within this chapter. Where posterior samples are being obtained using the particle MCMC scheme, sampling may be more efficient benefiting the overall efficacy of the hybrid algorithm.

The final tuning parameter that requires to be specified before the algorithm can proceed is the number of particles used to estimate likelihood in the bootstrap particle filter. We have already explored the fact that the number of particles required to satisfy the variance criteria on the log likelihood is typically much greater away from the posterior mode, and hence for efficient iterations of the sampler, in the regions of high density where most time is spent, we want to tune our particles around this modal value.

We turn to our approximation once again. In practice we take the average of the approximate distribution, $\bar{\theta}_{\texttt{ABC}}$ to calculate the number of particles required to satisfy

$$\text{Var}(\log(\hat{\pi}(\mathcal{D}|\bar{\theta}_{\texttt{ABC}}))) \simeq 2, \tag{6.3}$$

in line with Sherlock *et al.* (2015). The motivation being that the mean of our approximation should be close to the mean of the true target. An alternative to this could be a MAP estimate which will be close to the true posterior mode when the approximation is good. We do this by running the particle filter a number of times with varying numbers of particles until the condition is satisfied.

The hybrid ABC pMCMC algorithm, outlined in algorithm 10, provides a principled approach to the intelligent initialisation of a parallel particle MCMC scheme, whose performance gain scales well with the number of available processors and maintains the exact posterior distribution as it's target.

6.3 Applications

We now apply the hybrid approach introduced to a variety of numerical examples.

Algorithm 10 Hybrid ABC pMCMC

1. Run an ABC algorithm targeting $\pi(\theta|\rho(\mathcal{D},\mathcal{D}^*) \leq \epsilon_T)$.
2. Initialise multiple MCMC chain with a sample $\theta_0 \sim \pi(\theta|\rho(\mathcal{D},\mathcal{D}^*) \leq \epsilon_T)$ and set $i = 1$.
3. Propose $\theta^* \sim q(\theta^*|\theta)$ for $q(\cdot)$ a Gaussian random walk kernel, $\mu_q = \theta, \quad \Sigma_q = \frac{(2.38)^2}{d}\Sigma_{\texttt{ABC}}$.
4. Approximate $\pi(\mathcal{D}|\theta^*)$ via a bootstrap particle filter, $\hat{\pi}(\mathcal{D}|\theta^*)$ where the number of particles is chosen as in (6.3).
5. With probability
$$\min\left\{1, \frac{\hat{\pi}(\mathcal{D}|\theta^*)\pi(\theta^*)q(\theta|\theta^*)}{\hat{\pi}(\mathcal{D}|\theta)\pi(\theta)q(\theta^*|\theta)}\right\}$$
set $\theta^{(i)} = \theta^*$ else set $\theta^{(i)} = \theta^{(i-1)}$.
6. Set $i := i + 1$ and return to 3.

6.3.1 The Lotka–Volterra system

Consider again the Lotka–Volterra predator prey example introduced in section 2.5.3. Exact parameter inference is possible for this system using a pMCMC scheme; provided the chain is initialised near the posterior mode as has been seen in chapter 5, also see (Wilkinson, 2011). However under poor initialisation, a pMCMC scheme will perform very badly.

Consider a scenario in which prior information on the reaction rate parameters is poor,

$$\log(\theta_i) \sim \mathcal{U}(-8, 8), \qquad i = 1, 2, 3. \tag{6.4}$$

Prior distributions on X_0, the initial state of the system are taken to be independent Poisson distributions with rate parameters equivalent to the true initial conditions which gave rise to the synthetic data.

$$x_{1,0} \sim \text{Pois}(50), \quad x_{2,0} \sim \text{Pois}(100). \tag{6.5}$$

Further we assume that the variance, $\sigma^2 = 10^2$, of the Gaussian measurement error distribution is known.

Figure 6.1(a) gives a synthetic data set from the Lotka–Volterra model for use in this numerical example. True rate parameters that generated the realisation of the process are $\log(\Theta) = (0, -5.3, -0.51)$ given $X_0 = (50, 100)$.

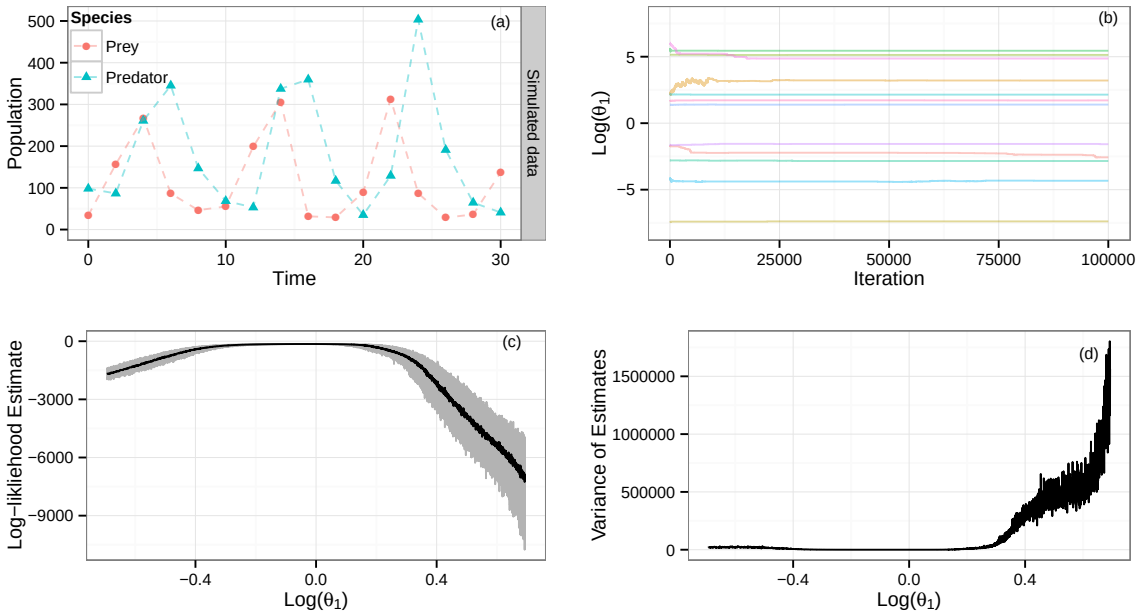


Figure 6.1: Investigation of computational issues with pMCMC for the Lotka–Volterra model defined in section 6.3.1. (a) The true underlying synthetic data set. Species are observed at discrete time points and corrupted with $N(0, 10^2)$ noise. (b) Twelve trace plots of $\log(\theta_1)$ from pMCMC chains initialised with random draws from the prior (see expression 6.4). The chains fail to explore the space. (c) shows the median and 95% interval for estimates of the log–likelihood from the particle filter for varying θ_1 close to the true value, for θ_2 and θ_3 fixed at the true values. (d) shows that the variance of log–likelihood estimates increases away from the true values.

pMCMC for Lotka–Volterra

Using a Gaussian random walk proposal kernel on $\log(\theta)$, $q(\log(\theta)^* | \log(\theta))$ we construct a Metropolis Hastings algorithm using a bootstrap particle filter as a means to estimate likelihood, targeting $\pi(\Theta|\mathcal{D})$.

Figure 6.1(b) shows 12 pMCMC chain trace plots that have been initialised using random draws from the weakly informative prior. The chains do not explore the space, failing to converge. Further investigation shows why this is happening. Figure 6.1(c and d) show that away from the true values and subsequently away from the posterior mode, the variance of the log–likelihood estimates from the bootstrap particle filter increases sharply. Figure 6.1 (c) shows the 95% interval of the estimated log–likelihood given a particle filter using 150 particles against the log of the prey birth parameter. Here, 150 particles is sufficient to satisfy $\text{Var}(\hat{l}(\Theta)) < 2$. Figure 6.1 (d) shows that even small deviations from the data generating parameter

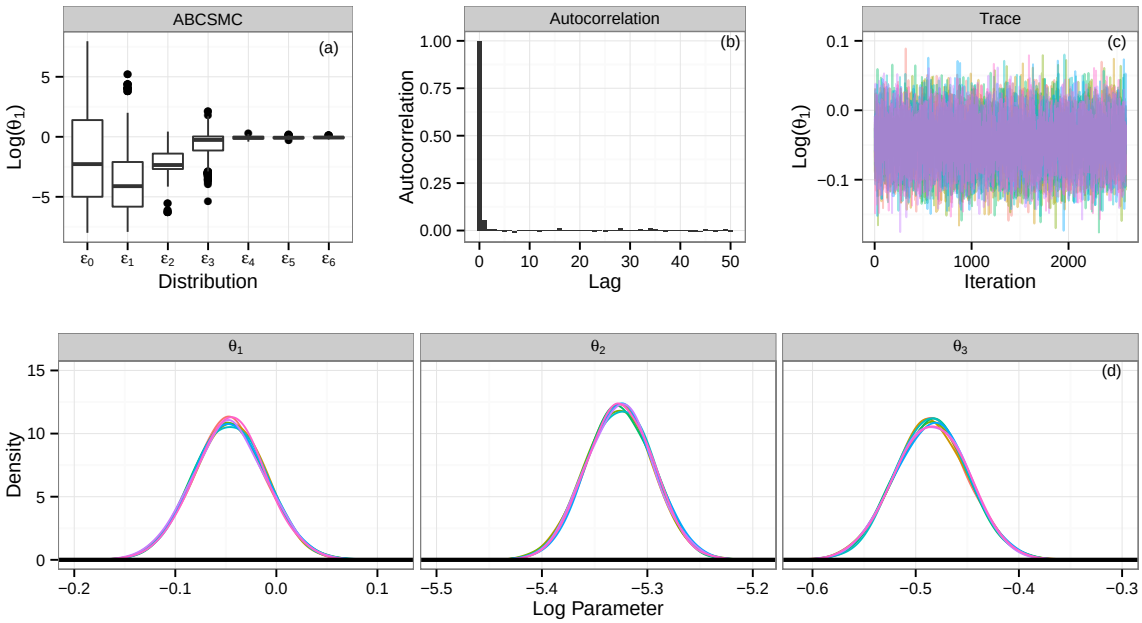


Figure 6.2: Analysis of results for the synthetic data for the Lotka–Volterra model using the hybrid approach. (a) successive distributions of $\log(\theta_1)$ in the sequential ABC scheme, algorithm 8. (b) show autocorrelations for chain 1, representative of each of the parallel chains, and (c) are traces of eight parallel MCMC chains for $\log(\theta_1)$. Note that each chain is sampling from the same stationary distribution and mixing appears good. (d) are the posterior densities for $\log(\theta)$, each chain leads to a posterior density plot that is very close to that of every other chain. True values $\log(\Theta) = (0.0, -5.30, -0.51)$ are well identified.

value, $\log(\theta_1) = 0$, cause the variance of the log-likelihood estimates to increase dramatically. In both figure 6.1(c) and (d) the other two parameters were kept fixed at the true values, on relaxing this the problem is exacerbated. When the state of the Markov chain is in a region with negligible likelihood it has a tendency to stick as a result of the variability in the likelihood estimates. Small proposed moves around parameter space can lead to large variation in the estimated likelihood. This in turn leads to poor exploration of the tails in the posterior distribution without a large number of particles in the bootstrap particle filter. Whilst we are guaranteed to eventually converge to the stationary distribution, the required computational cost, without carefully thought out initialisation, could be very high. Note that this is not a failure of the theory or algorithm, but a consequence of the sensitivity to initialisation of parameter values experienced in this type of model.

We therefore apply the proposed ABC initialisation for a "perfect" parallel pMCMC scheme.

Results for the Lotka–Volterra model using a hybrid approach

A sequence of seven distributions $\pi(\theta|\rho(\mathcal{D}, \mathcal{D}^*) < \epsilon_t)$, $t = 0, \ldots, 6$ are performed at the ABC SMC stage of the hybrid algorithm in order to obtain an approximation to the posterior. For each population, t, in the sequence of bridging distributions we take ϵ_t as the 0.3 quantile of the distribution of distances from the samples at $t-1$. Candidate parameters are proposed from the importance density until a sample of 1000 is obtained for each t. At each stage we perform model realisations given proposed parameter values in parallel. Figure 6.2 (a) shows summaries for log of the prey birth rate parameter $\log(\theta_1)$ for each of the distributions in the series through the sequential ABC. The distributions quickly remove a large region of space that, had we sampled from the prior distribution to initialise the chain, are likely to have been poor starting points. The scheme converges around the true value $\log(\theta_1) = 0$. Given the sample from $\pi(\theta|\rho(\mathcal{D}, \mathcal{D}^*) < \epsilon_6)$ eight MCMC chains are initialised with random draws from the finite sample approximation. Eight were chosen here to allow each chain to run on a separate CPU core. The results in figure 6.2(b,c,d) are then, the 20,000 pooled samples from the eight independent parallel chains, each of which has been thinned by a factor of 100 to give 2,500 samples. Each chain is sampling from the same stationary distribution, giving credence to the fact that we are indeed sampling the correct distribution and that convergence has been achieved, as seen in the trace plot for θ_1, figure 6.2 (c), and mixing is good, figure 6.2 (b). Further the true parameter values, $\log(\Theta) = (0, -5.30, -0.51)$, used to simulate the data are well identified within the posterior densities, figure 6.2 (d).

6.3.2 Real-data problem – aphid model

Next we consider a model of aphid population dynamics as proposed in Matis *et al.* (2007). The system can be represented by the following reactions:

$$\begin{aligned} \mathbf{R}_1 : \quad & N \rightarrow 2N + C \\ \mathbf{R}_2 : \quad & N + C \rightarrow C \end{aligned} \tag{6.6}$$

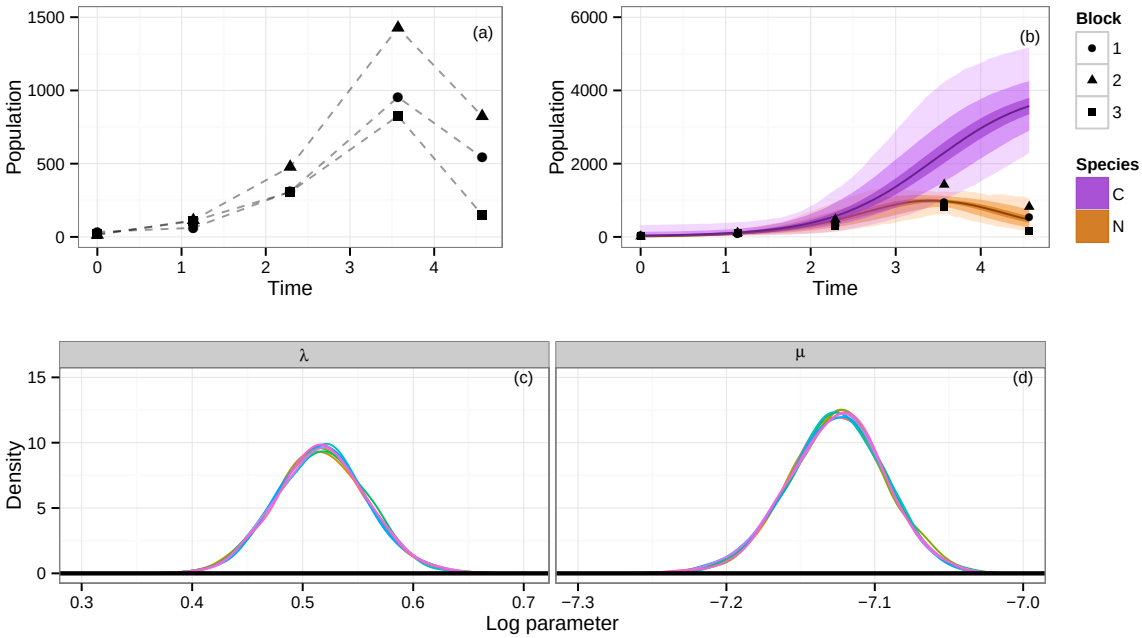


Figure 6.3: Analysis of the real data for the aphid growth model. (a) There are the three data sets of aphid counts each consisting of five observations. (b) The posterior predictive model fit given a sample from the collection of posterior densities. (c,d) Output from each MCMC chain, consistent posterior densities show we are sampling from the same stationary distribution.

and summarised in terms of its stoichiometry matrix S, and hazard function $h(\mathbf{X}_t, \theta)$,

$$S = \begin{pmatrix} 1 & -1 \\ 1 & 0 \end{pmatrix}, \qquad h(\mathbf{X}_t, \theta) = (\lambda N_t, \mu N_t C_t), \tag{6.7}$$

$\mathbf{X}_0 = (N_0, C_0)$, $\theta = (\lambda, \mu)$.

N_t and C_t are the numbers of aphids alive at time t and the cumulative number of aphids that have lived up until time t respectively. In the first reaction, when a new aphid is introduced into the system we get an increase in both the current number of aphids and the cumulative count. When an aphid is removed the number of aphids decreases by one but the cumulative count remains the same. Aphid death is considered to be proportional not only to the number of aphids currently in the system, but also to the cumulative count, representing the idea that over time they are making their own environment less habitable, exhausting resources.

Given initial conditions $\mathbf{X}_0 = (N_0, C_0)$ and a set of reaction rate parameters θ, we can simulate from the model using the Gillespie algorithm. Observations are

noisy, discrete time measurements of just a single species of the system, since the cumulative count C_t can never be observed.

Aphid data

We now consider the data described in Matis *et al.* (2008) consisting of cotton-aphid counts for twenty seven treatment-block combinations. The treatments consisted of three nitrogen levels (blanket, variable and none), three irrigation levels (low, medium and high) and three blocks. The sampling times of the data are $t = 0, 1.14, 2.29, 3.57, 4.57$ weeks, or every seven to eight days. We restrict ourselves to a single treatment combination, three data sets with blanket nitrogen level and low irrigation. If we denote the block by $i \in \{1, 2, 3\}$ then the data $\mathcal{D}_i$ is the number of aphids, N, in block i at each time t. The data are plotted in figure 6.3(a).

We make the assumption that the counts are observed with error such that

$$d_t \sim \text{Pois}(x_t), \tag{6.8}$$

and use a set of weakly informative priors on the rate parameters θ

$$\log(\theta_i) \sim \mathcal{U}(-8, 8), \qquad i = 1, 2. \tag{6.9}$$

We place a prior of the form

$$C_0 = N_0 + g, \quad g \sim \text{Geom}(0.03), \tag{6.10}$$

to reflect the fact that we are unable to measure C_0 but under the knowledge that it must be such that $C_0 \geq N_0$.

We treat the three sets of observations as repeats of the same experiment. Likelihood estimates necessary for pMCMC are obtained by running a particle filter for each of the three data sets and taking the sum of the individual log–likelihood estimates. A full treatment of all twenty-seven data sets using a fixed effects model can be found in Gillespie & Golightly (2010). We consider the initial aphid counts to be the true values consistent with Gillespie & Golightly (2010), on the basis that there should be no error in counting such small populations.

Results for the aphid growth model

We use the same criteria for the choice of ϵ_t for the ABC section of the inference as with the Lotka–Volterra model in 6.3.1, namely the 0.3 quantile of the distribution of distances. A sequence of five distributions gives us 1000 samples from $\pi(\theta|\rho(\mathcal{D}, \mathcal{D}^*) < \epsilon_4)$ which we use to initialise eight parallel chains. We record 20,000 samples from the exact target posterior $\pi(\theta|\mathcal{D})$ after appropriate thinning. Figure 6.3(c and d) shows the analysis of the MCMC chains. Again we find that each chain is sampling from the same target and posterior densities are very close from all eight chains. Figure 6.3(b) shows posterior predictive quantiles given a sample from the posterior samples that result as collating the output of the MCMC chains. The posterior predictive quantiles suggest that model fit appears to be reasonable. The results are consistent with those seen in Gillespie & Golightly (2010) where they assume that observations are made without error and make use of an approximate simulation algorithm for realisations of the model giving us greater confidence in our inferred conclusions.

6.3.3 Gene expression

Finally, we consider a simple gene regulation model characterised by three species (DNA, mRNA, denoted R, and protein, P) and four reactions. The reactions represent transcription, mRNA degradation, translation and protein degradation respectively. The system has been analysed by Komorowski *et al.* (2009) and Golightly *et al.* (2014) among others:

$$\begin{aligned} \mathbf{R}_1 &: \quad DNA \rightarrow DNA + R \\ \mathbf{R}_2 &: \quad R \rightarrow \emptyset \\ \mathbf{R}_3 &: \quad R \rightarrow R + P \\ \mathbf{R}_4 &: \quad P \rightarrow \emptyset \end{aligned} \tag{6.11}$$

with stoichiometry matrix $\mathbf{S}$, and hazard function $h(\mathbf{X}_t, \theta)$

$$\mathbf{S} = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}, \qquad h(\mathbf{X}_t, \theta) = (\kappa_{R,t}, \gamma_R R_t, \kappa_P R_t, \gamma_P P_t) \tag{6.12}$$

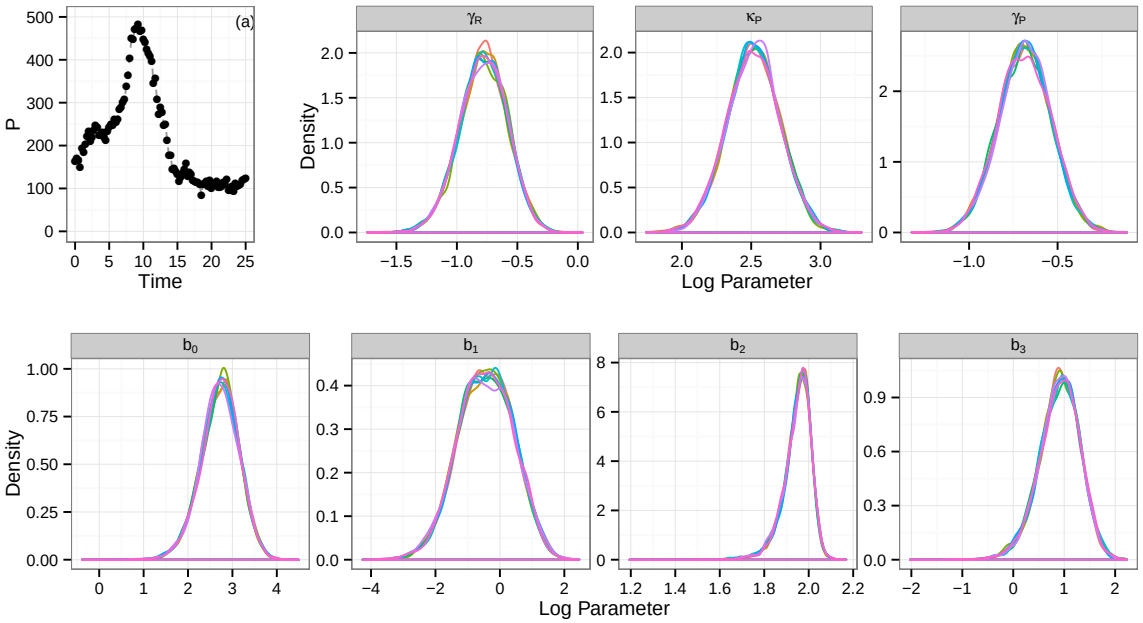


Figure 6.4: (a) is the noisy pseudo-data for the Protein levels in the model. The other plots show the individual densities from pMCMC chains after appropriate thinning having been initialised via an ABC run as described in section 6.2.6. The plots clearly show that each of the chains are in agreement with regard to sampling from the stationary distribution.

where $\mathbf{X}_t = (R_t, P_t)$ and $\theta = (\gamma_R, \kappa_P, \gamma_P, b_0, b_1, b_2, b_3)$ where we note that, as in Komorowski *et al.* (2009), we take $\kappa_{R,T}$ to be the time dependent transcription rate. Specifically,

$$\kappa_{R,t} = b_0 \exp(-b_1(t - b_2)^2) + b_3 \tag{6.13}$$

such that the transcription rate increases for $t < b_2$ and tends towards a baseline value, b_3, for $t > b_2$. As above the goal is inference on the unknown parameter vector, θ. In keeping with inference in Komorowski *et al.* (2009) we create a data poor scenario, 100 observations of synthetic data simulated given initial conditions $X_0 = (10, 150)$ and parameter values $(0.44, 0.52, 10, 15, 0.4, 7, 3)$ corrupted with measurement error, $Y_t \sim \mathcal{N}(X_t, I\sigma^2)$, $\sigma = 10$, with observations on the mRNA discarded. The data is shown in figure 6.4(a).

We follow Komorowski *et al.* (2009) by assuming the same prior distributions, including informative priors for the degradation rates to ensure identifiably. Specifi-

cally

$$\gamma_R \sim \Gamma(19.36, 44) \qquad \gamma_P \sim \Gamma(27.04, 52)$$
$$\kappa_P \sim \text{Exp}(0.01) \qquad b_0 \sim \text{Exp}(0.01)$$
$$b_1 \sim \text{Exp}(1.0) \qquad b_2 \sim \text{Exp}(0.1)$$
$$b_3 \sim \text{Exp}(0.01)$$

where $\Gamma(a, b)$ is the gamma distribution with mean a/b and $\text{Exp}(a)$ is the Exponential distribution with mean 1/a.

For simplicity we assume that both the initial state, $X_0 = (10, 150)$, and the measurement error standard deviation, $\sigma = 10$, are known.

Results for gene expression data

We follow the same procedure as with the two examples above. Using a sequential ABC run to obtain a sample of 1000 parameters vectors distributed according to the approximate posterior. We then use eight random draws from the final ABC sample to initialise the parallel pMCMC chains with tuning parameters chosen as described in section 6.2.6. The posterior densities, a sample of 4000 from each chain having been subjected to appropriate thinning, are shown in figure 6.4. It is clear that each of the chains is sampling from the same target giving us confidence in the resulting densities. The posteriors obtained are consistent with those in Golightly *et al.* (2014) and true parameter values are well identified. Figure 6.5 shows that the sample from the final iteration of the sequential ABC algorithm is markedly different from that in the pMCMC algorithm. There is a measurable improvement in using this type of scheme over using solely ABC in this way. We characterise the difference shown here as an improvement due to the fact that we know that pMCMC is asymptotically exact.

6.4 Discussion

We have proposed an approach to inference for Markov processes that samples exactly from the desired posterior distribution and combines the relative strengths of

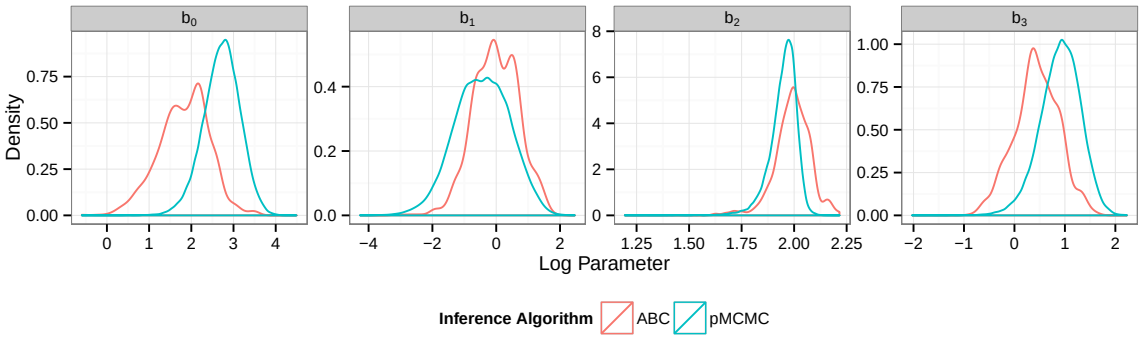


Figure 6.5: A comparison between the final sample using the ABC SMC algorithm and the pMCMC for the gene regulation model for the four parameters in the time dependent hazard. The plot shows that there is a distinct difference between the two posterior samples. Plots for the other three parameters show similar but are omitted here.

ABC and pMCMC methodology to increase computational efficiency through use of parallel hardware. By using an approximation to the posterior distribution of interest, obtained via a sequential ABC algorithm which is easy to parallelise, we can set up a parallel implementation of pMCMC which has numerous desirable properties. By enabling the construction of independent parallel chains initialised close to the stationary distribution, this enables fast convergence and sampling from an exact posterior distribution that scales well with available computational resources. Throughout our analyses we have made use of parallel computation, however we believe that the proposed approach will also be of interest in situations where parallel hardware is not available, as it still addresses the pMCMC initialisation and tuning problem. Algorithmic tuning parameters required for pMCMC, such as the variance of Gaussian random walk proposals and numbers of particles for the particle filter can be chosen without the need for additional pilot runs, as a consequence of having a sample from an ABC posterior. In addition, independent parallel chains allow verification of convergence and the computational saving in burn–in times extends to repeat MCMC analyses.

We have demonstrated this approach by applying it to three stochastic kinetic models of varying complexity. With the Lotka–Volterra predator prey system, a relatively simple model in which both species can be observed, we highlighted clear issues with practical implementation of a pseudo-marginal approach in a scenario in which prior information on reaction rate parameters is poor which concurs with results found in chapter 5. This issue can be alleviated by first obtaining a sample from an approximation to the posterior, then using it to guide an exact pMCMC

scheme. The approach discussed performed similarly well in the application to a set of real data for a model for aphid growth dynamics in which one of the species in the system can never be observed, where we again imposed weak prior conditions on the rate parameters governing the system and had access to repeat data. Finally we applied the scheme to a gene regulation model in which we had partially observed data and rate parameters were not all time homogeneous. The analyses of results show that we can verify that we are sampling from the same target distribution adding to our belief that we have converged to the true posterior of interest in each case.

Chapter 7

Inference for cell population data

7.1 Introduction

The study of dynamic molecular networks is increasingly exploiting technology such as flow cytometry to obtain large scale data on the marginal distributions of species at snapshots in time. Where single cell time course data allows the tracking of molecular species over time, where draws from the transition densities are directly observed, in the snapshot data this resolution is lost. We can not link a given observation at time t in the snapshot data with an observation at time $t-1$.

Inference within this framework is problematic, due to both the volume of data and its resolution coupled with the intractable likelihood function of the models. It has already been recognised that traditional likelihood free techniques are difficult to implement within this framework due to the sheer computational burden of model simulation for data of this size. Lillacci & Khammash (2013) propose an approach to likelihood free inference for problems of this type which aims to minimise the the number of model simulation steps by considering a Kolmogorov distance between observed and simulated empirical distribution functions. On recognising that large numbers of model simulation steps for each proposed parameter set is undesirable they use a hypothesis testing framework to devise a critical number of simulations necessary to determine with high confidence that, under the null hypothesis that

both observed and simulated data are from the same distribution, the Kolmogorov distance is within the predetermined tolerance.

Zechner *et al.* (2012) propose a population moment closure approximation of the system but recognise that determining whether this carries sufficient information to identify the rate parameters is non–trivial.

In this chapter we focus on a more general approach to approximate Bayesian computation for inference on large scale cell population data. Various techniques to reduce the computational expense of model simulation will be explored allowing a wider scope of metric functions to be used than that found in Lillacci & Khammash (2013). In addition this allows more advanced ABC techniques such as those based on sequential Monte Carlo sampling to be leveraged.

The techniques will be explored using the immigration–death process for which an analytic solution is available for reference purposes before being applied to an inference problem for rate parameters of 3 strains of Bacillus Subtilis.

7.2 ABC for cell snapshot data

The increased volume of the data means that dimension reduction is almost certainly necessary, point-wise distance measurements between observations no longer makes sense in this context. As data are sample representations of the marginal distributions of the observed species at a collection of time points, intuitive choices of summary statistics are those that help to describe it's shape such as measures of location and dispersion.

Consider an arbitrary continuous probability density function $f(x)$ with distribution function $F(x)$. If $F(x)$ is a strictly increasing function then there exists a unique inverse $F^{-1}(x) = Q(x)$ where $Q(x)$ is called the quantile function. Consequently sample quantiles from large samples constitute a useful collection of summary statistics to measure the closeness of two univariate distributions.

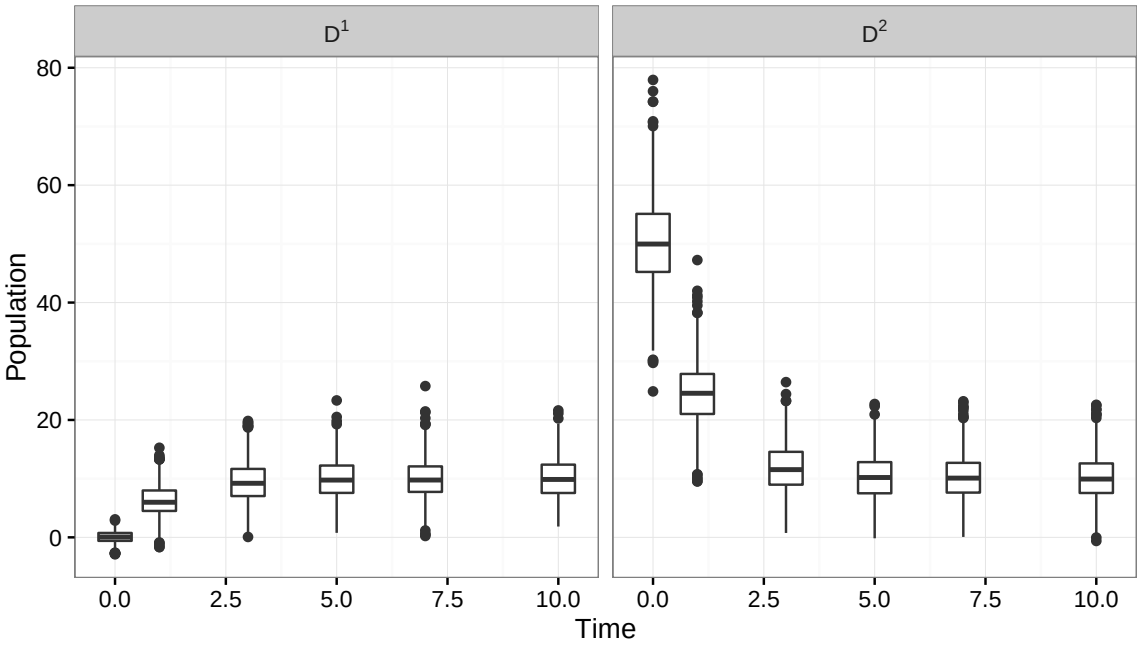


Figure 7.1: Cell population data for an immigration death process. Left are noisy observations of the species levels in 1000 cells at each of time point given initial condition $X_0 = 0$. Right are noisy observations of the species levels in 1000 cells at each time point given $X_0 = 50$. In each case the measurement error model is zero mean Gaussian with $\sigma = 2$. True reaction rate parameters that generated the data are $\log(\Theta) = (2.30, 0)$.

7.2.1 Immigration death process population data

For the purpose of numerical examples in this chapter we revisit the immigration death process introduced in section 2.5.1. We consider two data sets each consisting of levels of the species in 1000 cells at each time point given different initial conditions.

The data sets, shown in Figure 7.1 are noisy observations of the species levels in 1000 cells giving sample approximations to the marginal distributions of the process at each time point given true parameter value $\Theta = (10, 1)$, subject to a zero mean Gaussian measurement error model with $\sigma = 2$. Left, denoted D^1 are observations given initial conditions $X_0 = 0$, treatment of this data set will be made under the assumption that the initial conditions and the measurement error variance are known. Right, denoted D^2 are noisy observations given initial conditions $X_0 = 50$ however treatment of this data set will consider both the initial conditions and the variance of the measurement error model to be unknown.

Prior distributions

For each numerical experiment the prior distributions on the rate parameters are vague, uniform priors on the log scale.

$$\pi(\log(\theta_i)) = \mathcal{U}(-4, 4), i = 1, 2. \tag{7.1}$$

For the numerical examples that treat D^2 we have the additional prior distributions on the measurement error noise parameter and initial conditions

$$\pi(\log(\sigma)) = \mathcal{U}(\log(0.5)\log(5)), \pi(X_0) = \text{Po}(50) \tag{7.2}$$

7.2.2 Summary statistic weighting

An interesting consequence of having data of a large dimension comes when considering the weights of summary statistics to be used in the Euclidean metric. In section 5.3.3 we explored the typical approach of using a prior predictive distribution of the summary statistics to inform choice of these weights. The standard approach when considering a weighted Euclidean distance function between N_s summary statistics on the observations, $\mathbf{s}$ and simulated summary statistics from the model under parameters Θ, $\mathbf{s}_\Theta$,

$$\delta_\Theta = \sqrt{\sum_{i=1}^{N_s} \left(\frac{s_{\Theta,i} - s_i}{w_i} \right)^2} \tag{7.3}$$

is to choose the vector of weights w as to reflect

$$w = \text{SD}\left(\int_\Theta \pi(\mathbf{s}_\Theta \,|\, \Theta)\pi(\Theta) d\Theta \right) \tag{7.4}$$

for prior parameter distribution $\pi(\Theta)$.

Note that the integral in equation 7.4 is intractable since $\pi(\mathbf{s}_\Theta \,|\, \Theta)$ is unavailable and so in practice draws from the sampling distribution of s_Θ given sample size equivalent to that of the observations are used to give empirical estimates of the standard deviations of the summary statistics over the marginal distribution of $\mathbf{s}$.

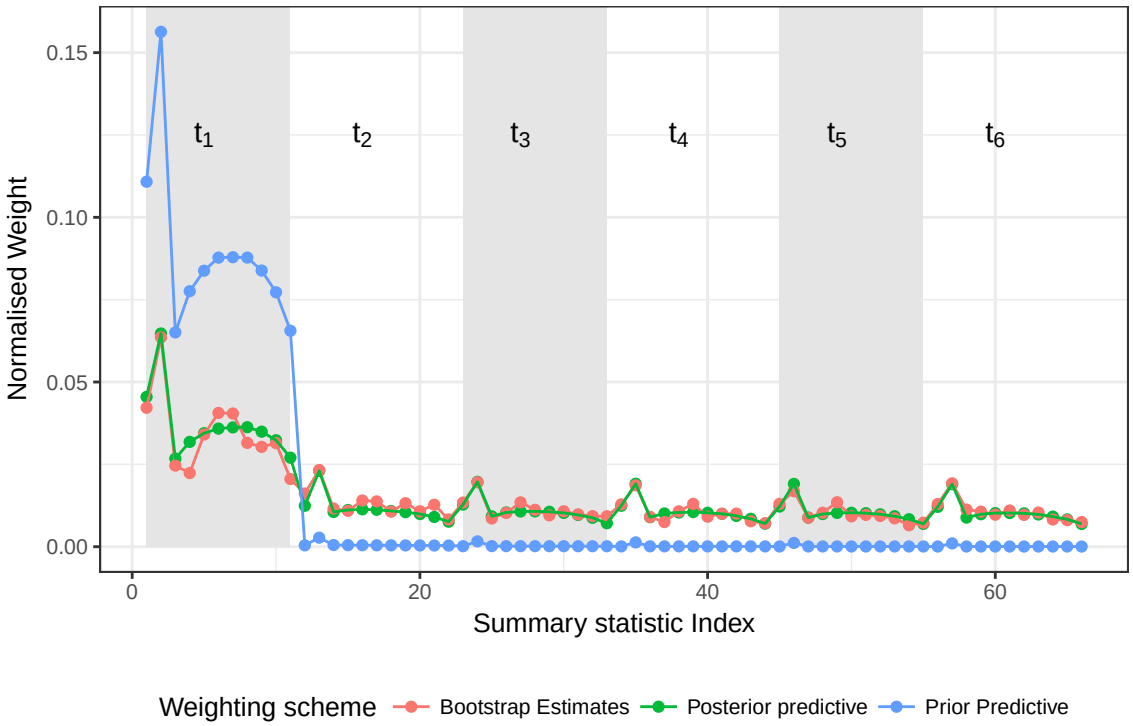


Figure 7.2: Summary statistic weights for a Euclidean distance metric on consideration of different approaches to weight estimation.

Prangle *et al.* (2016) showed that under vague prior information, for summary statistics from a model with complex dynamics, that such a weighting structure may not always be the most appropriate. The authors motivate an adaptive weighting scheme as part of their ABC SMC algorithm such that as the algorithm progresses, and the approximation to the posterior improves, the weights of the summary statistics are modified to more closely reflect the posterior predictive variability of the statistics. As a result they show improvement in the efficiency and accuracy of the resultant posterior distributions.

Considering this approach in the limit as the ABC scheme converges to the true posterior distribution, an ideal weighting mechanism then would be to use directly the posterior predictive distribution of the summary statistics such that the vector of weights

$$w = \text{SD}\left(\int_{\Theta} \pi(\mathbf{s}_{\Theta} \,|\, \Theta)\pi(\Theta \,|\, \mathbf{s})d\Theta\right). \tag{7.5}$$

In practice this is unavailable, since the goal we are trying to achieve is to learn about the parameters in the model that lead to the data observed. If we had access

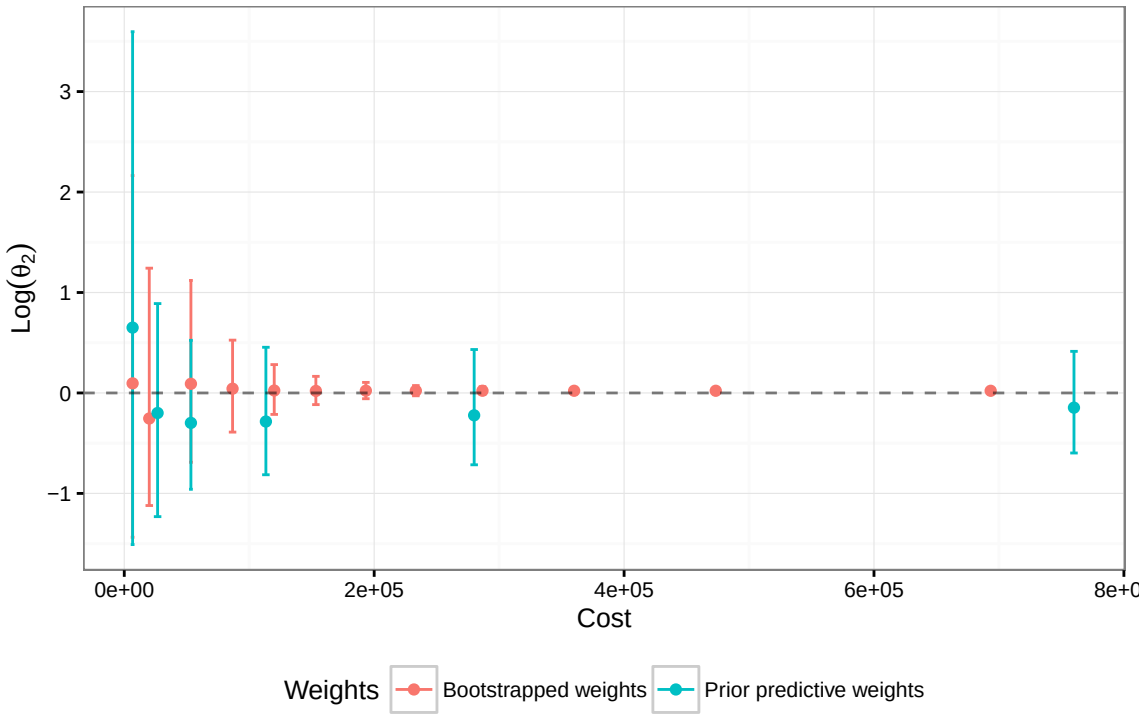


Figure 7.3: Evolution of the posterior distribution across the sequence of bridging distributions for decreasing ϵ under the two weighting schemes.

to $\pi(\Theta \mid \mathbf{s})$ then our inference problem would be solved. Note that even if $\pi(\Theta \mid \mathbf{s})$ were available that the integral in equation 7.5 is intractable but draws from the sampling distribution of $\mathbf{s}_\Theta$ could be used to give an empirical estimate.

We can not hope to calculate w using sample estimates of the summary statistics given the posterior distribution however we can obtain a proxy for it, namely the variability of the summary statistics given the true parameter values.

Since our observations are a large finite sample representation of the marginal distributions of the process at the observed time points given the unknown true parameters Θ^* we can estimate the variability of the sampling distribution of any given summary statistic $\mathbf{s}_\Theta$ via bootstrapping.

The idea is to use directly the variability of sampling distribution of $\mathbf{s}$ given the true parameter values and sample size that generated the data Θ^*,

$$w_i = \mathrm{SD}(\mathbf{s}_{\Theta^*}). \tag{7.6}$$

In this case each term in the summation is an estimate of the number of stan-

dard deviations away from the observed summary statistic $\mathbf{s}_{\Theta^*}$ under the sampling distribution $\pi(\mathbf{s} \mid \Theta^*)$.

Given a set of observations $\mathbf{x}_{\Theta^*} \sim \pi(\mathbf{x} \mid \Theta^*)$ the observed summary statistics $\mathbf{s}_{\Theta^*}$ are a sample estimate of $\mathbb{E}(\mathbf{s} \mid \Theta^*)$. If our sample of size N_{obs} from $\pi(\mathbf{x} \mid \Theta^*)$ is such that N_{obs} is large we can use bootstrap samples to quantify the variability in the sample estimates of $\mathbf{s}_{\Theta^*}$.

Figure 7.2 shows the vector of summary statistic weights as calculated by the standard approach of a pilot run to estimate the prior predictive standard deviations and the bootstrapping described above here. For the example shown the set of summary statistics considered were the means, standard deviations and the quantiles $q_{10\%}, q_{20\%}, \ldots, q_{90\%}$ applied to the D^2 data set. For the immigration death model, the likelihood function is tractable allowing standard likelihood inference to be made. This allows us to compare the proposed weighting scheme with that under the posterior predictive distribution. The green points show the weights that would be calculated if one were able to use this information. The red points show the weights calculated on consideration of bootstrap estimates of the summary statistics whilst for comparison the blue show those calculated under a pilot run from the prior.

Clearly the variance of the bootstrap sample estimates of summary statistics are a much closer match to the posterior predictive variance than those from the prior tuning run. All weights in the plots have been normalised, to sum to 1, such that the relative weights are being compared. The prior predictive distribution attributes the majority of the accept reject decision to the early time points of the observations, mirroring what was found in chapter 5.

In addition to comparing the value of weights directly we can consider the inferred posterior distribution under each weighting approach. Figure 7.3 shows the evolution of the posterior approximation through the sequence of distributions of an ABC SMC scheme of the log of the death parameter of the ID model. In each case the same algorithmic specification was used, namely a reduction of tolerance based on the $30^{\text{th}}\%$–ile of the distribution of distances at each stage, with optimal perturbation kernel of Filippi *et al.* (2013) for propagation of samples. For each iteration of the sequential sampler, particles were proposed until a sample of 2000 were accepted. In figure 7.3 observe that the convergence of the posterior approximation is much faster given the bootstrap sampling approach to calculation of weights.

7.2.3 Alternative distance functions

Since data are effectively independent marginal distributions of species at a collection of time points it opens the way for alternative distance functions beyond the use of descriptive statistics as a means by which to summarise the data. Since cell population data consist of large samples from the marginal distributions of the states of species at given time points, a function which measures distance between two distributions seems a natural fit. Where the dimension of data is suitably large we have an empirical approximation to the true marginal distributions. An f-divergence is a function, $D_f(P||Q)$, which measures the distance between two probability distributions. Examples of f–divergence functions include Hellinger distance, total variation, but perhaps the most natural choice is Kullback Leibler divergence.

7.2.4 Kullback Leibler divergence as a metric

It was noted in section 5.2.1 that there is no practical reason why a distance function used to quantify dissimilarity between two data sets in the context of ABC need be a metric in the formal sense.

Kullback–Leibler divergence (Kullback & Leibler, 1951) is a special case of f-divergence. The motivation for it's use as a metric in ABC is it's direct connection with the likelihood function.

Let

$$\frac{1}{N}\sum_{i=1}^{N}\log(\pi(x_i \,|\, \Theta)$$

be the log likelihood function of the model parameters. Note that by the strong law of large numbers

$$\frac{1}{N}\sum_{i=1}^{N}\log(\pi(x_i \,|\, \Theta) \overset{a.s}{\to} \mathbb{E}(\log(\pi(x \,|\, \Theta)))$$

The expectation of the difference in log likelihood of the true model parameters Θ_0

and a candidate model parameter set Θ then is

$$\begin{aligned}\mathbb{E}(\log(\pi(x \mid \Theta_0) - \log(\pi(x \mid \Theta))) &= \mathbb{E}\left(\log\left(\frac{\pi(x \mid \Theta_0)}{\pi(x \mid \Theta)}\right)\right) \\ &= \int \log\left(\frac{\pi(x \mid \Theta_0)}{\pi(x \mid \Theta)}\right) \pi(x \mid \Theta_0) dx,\end{aligned}$$

the definition Kullback–Leibler divergence of the distribution given the true unknown parameter values $\pi(x \mid \Theta_0)$ from $\pi(x \mid \Theta)$.

Considering an ABC sampler where one retains samples of model parameters Θ which give a distance δ smaller than some tolerance ϵ, when using KL–divergence to calculate δ, this is equivalent to keeping model parameters whose expectation of log–likelihood, $l(\Theta)$ is greater than some value L.

Clearly in order to exactly calculate the Kullback–Leibler divergence it is equivalent that we need be able to evaluate $\pi(x \mid \Theta)$, the likelihood function. This is not possible as we have already seen that the likelihood function for stochastic kinetic models of any complexity is intractable. However given a sample from $\pi(x \mid \Theta)$ is it possible to estimate Kullback–Leibler divergence. In chapter 5 we have already made use of a nearest neighbour estimator for KL–divergence when assessing the closeness of different posterior distributions to the truth in the immigration–death process examples.

Figure 7.4 shows the marginal posterior distribution of the immigration rate parameter given observations D^1 using KL divergence as a metric. The mode of the distribution is close to that of the true posterior distribution however the posterior variance is over estimated. For comparison figure 7.4 also shows the the approximate posterior distribution when using means, standard deviations and deciles as means to measure dissimilarity subject to the same computational expense. Kullback Leibler divergence is not as efficient at targeting the posterior given this example.

An additional note to make about the practical implementation of using the nearest neighbour estimator for KL divergence is that we found that it was not uncommon for estimates to be below zero for moderate sample sizes. Consequently, it was possible for the quantile reduction of tolerance to give a target of 0, which yields samples which do not give a perfect match between simulated and observed distributions. This is a consequence of attempting to estimate KL divergence over a multi dimensional distribution, using only a moderate sample size. This issue was much less

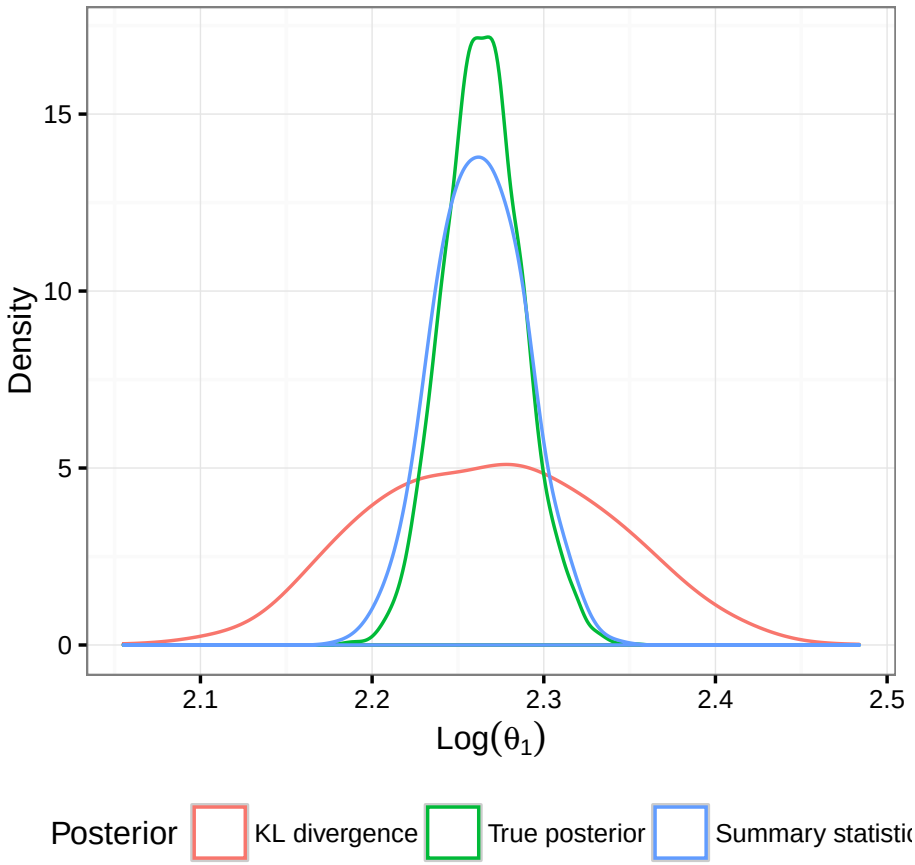


Figure 7.4: Comparison of approximate posterior distributions inferred for the immigration rate parameter under observations D^1 when using Kullback Leibler divergence as a metric.

prevalent if the simulated data were much larger. Unfortunately this is the opposite of what we want when computation cost is already high.

7.3 Reducing the amount of simulation

With a large dimension of observations, having a model simulation step in order to estimate dissimilarity between observed and simulated data sets is expensive. It is desirable, if possible, to be able to obtain summary statistics for the data with fewer realisations than the number of observations in the data. A potential problem with this is that as the number of simulated data points is reduced, the variability of the estimated summary statistics increases. The effect of this on individual summary statistics and the subsequent effect on the posterior distributions inferred is unclear as the variance of a summary statistic is rarely linear in the number

of points used for its estimation. Because of this certain summary statistics may be more amenable to the reduction of simulation steps than others. For example consider the immigration–death process from the previous example. One might imagine that if a vector of means were the summary statistic of choice, that reasonably good estimates of those means may be obtained using only 100 simulations from the model, say. However estimating Kullback Leibler divergence between two multivariate distributions using a nearest neighbour approach is likely to be more problematic.

The variability of the simulated summary statistics will in turn affect the variability of the distances calculated. This then changes the probability that a given parameter vector is accepted in an ABC algorithm.

Our method of weighting the summary statistics, considerate of the variability of the sampling distribution from which they came fits well with our desire to reduce the amount of simulation. We can adjust the weights in the metric to account for the increased variability of the summary statistics. When calculating our bootstrapped estimates of the standard deviations of each summary statistic we simply reduce the sample size of each bootstrap sample. Usefully this also means that if the way in which the variability of the summary statistics changes is not the same for each s_i, and there is no reason to expect this to be the case, our bootstrap sampling approach captures this, with the relative weights in the metric changing accordingly.

Figure 7.5 compares the sequences of posterior approximation to θ_2 in the immigration death model when using simulation size equivalent to the observations, and a second run where far fewer simulations per proposal are used. The plot shows that movement of both posterior distributions are similar. The posterior distribution given 1000 model simulations per proposed parameter takes fewer proposals to converge tightly around the true value. But on consideration that the cost of each proposal using only 25 simulations is a factor of 40 cheaper, the smaller number of realisations per proposal is in fact more efficient, particularly at the early stages, in which the tolerance is still relatively large.

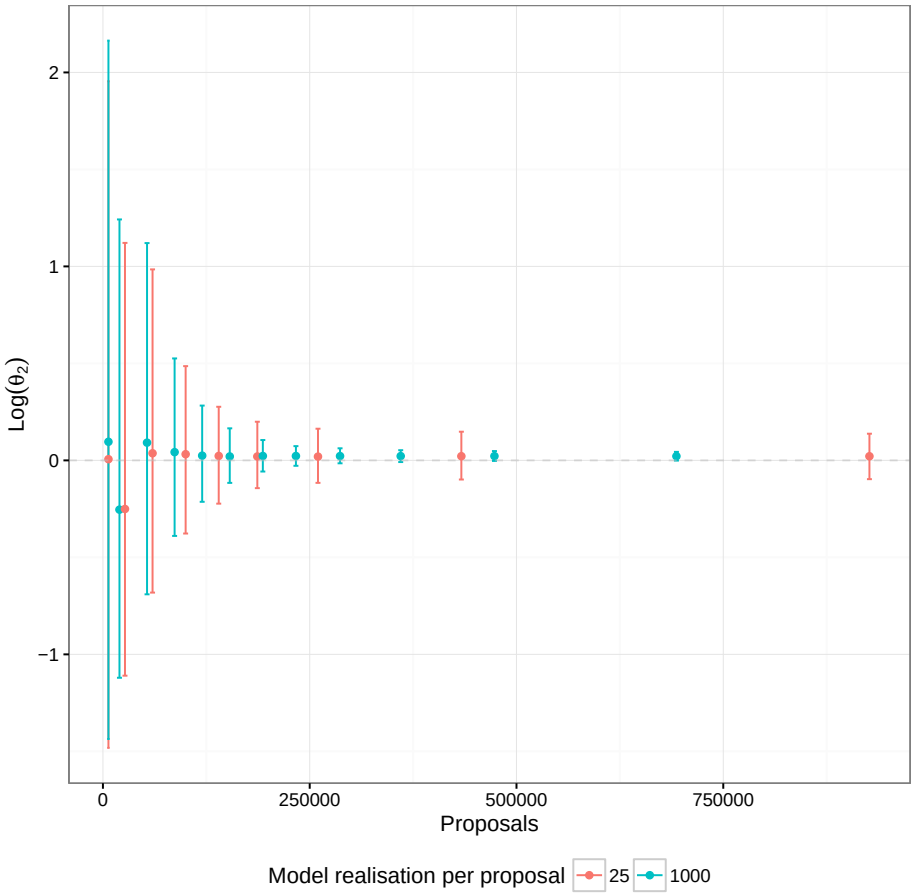


Figure 7.5: Comparison of the relative performance of posterior sampling when using far fewer simulations for each proposed parameter than the dimension of observations.

7.3.1 Early termination of simulation

As the cost of simulating from the model to each time point is particularly expensive in the case of cell population data any algorithmic optimisations that can be made which limit the time wasted on simulations of candidate parameters that will not be accepted is of interest. The calculation of a distance function in ABC lends itself well to employing early stopping criteria for certain sets of summary statistics.

Consider an ABC rejection sampler targeting $\pi(\Theta \,|\, \rho(\mathbf{s}_\Theta, \mathbf{s}) < \epsilon)$ where $\mathbf{s}_\Theta$ is the result of calculating a set of statistics summarising the information in some time series data given model parameters Θ. In the model simulation stage of the algorithm, we naturally have access to the observations at the first time point first. At the arrival of the simulated observations from time point t_1 we can calculate any of the summary statistics compressing information about $\mathbf{x}_{t_1}$ that are independent

of observations at other time points. Let this set of summary statistics be denoted $\mathbf{s}_{t_1}$ and more generally let $\mathbf{s}_{t_i,t_j,t_k,\ldots}$ represent the set of summary statistics whose calculation is dependent on observations from time points $t_i, t_j, t_k, \ldots$.

The Euclidean distance function for measuring dissimilarity between $\mathbf{s}$ and $\mathbf{s}_\Theta$ can be rewritten as

$$\delta_\Theta = \sqrt{\sum_{t_j} \sum_{s_i \in \mathbf{s}_{t_j}} (s_{\Theta,i} - s_i)^2 + \sum_{(t_j,t_k), j \neq k} \sum_{s_i \in \mathbf{s}_{t_j,t_k}} (s_{\Theta,i} - s_i)^2 + \ldots}. \tag{7.7}$$

Where $\mathbf{s}$ consists only of summary statistics $s_i \in \mathbf{s}_{t_j}$ this reduces to

$$\delta_\Theta = \sqrt{\sum_{t_j} \sum_{s_i \in \mathbf{s}_{t_j}} (s_{\Theta,i} - s_i)^2}, \tag{7.8}$$

which is a summation of terms that are strictly greater than or equal to 0.

Consider now that we are at time t_n during the model simulation step of a rejection sampler and that our set of summary statistics to be considered consist only of $s_i \in \mathbf{s}_{t_j}, j = 1, \ldots, T$. We can calculate $\delta_{\Theta,t_1:t_n}$,

$$\delta_{\Theta,t_1:t_n} = \sqrt{\sum_{t=t_1}^{t_n} \sum_{s_i \in \mathbf{s}_t} (s_{\Theta,i} - s_i)^2} \tag{7.9}$$

as the sum of the components of the Euclidean metric for which we have simulated observations available.

Then on rewriting δ_Θ as,

$$\delta_\Theta = \sqrt{(\delta_{\Theta,t_1:t_n})^2 + \sum_{t_j > t_n} \sum_{s_i \in \mathbf{s}_{t_j}} (s_{\Theta,i} - s_i)^2}, \tag{7.10}$$

we can consider that $\delta_{\Theta,t_1:t_n} \leq \delta_\Theta$ since all terms under the summation must be non–negative. Therefore at time t_n in the simulate from the model step we have a lower bound on δ_Θ. Hence if $\delta_{\Theta,t_1:t_n} > \epsilon$ then it must be the case that $\delta_\Theta > \epsilon$ and we can safely reject the proposal Θ. Rejecting at this stage precludes the need to simulate the model over the remaining time course for $t > t_n$.

This early termination step is trivial to add for very little computational overhead

into the algorithm and can be used as part of the accept reject stage of each of the rejection, MCMC and SMC based ABC samplers. Whilst it is true that the computational saving is highly dependent on the model being considered and on ϵ, $\mathbf{s}$ and the closeness of the proposal kernel to the target, where we expect nearly all proposals towards the regions of high mass in the target to almost never be rejected early, the cost of implementation is typically vastly inferior to the cost of simulating from the model. Therefore we find that if even only a few proposals are rejected early we have made some saving.

7.4 Improved proposals for ABC SMC

One of the prevalent issues with simulation based inference where data are large is that the computational expense of the "simulate from the model" step is often a limiting factor. Since the cost of one model simulation step is high it is desirable to have as many proposed parameter vectors in regions of non–negligible support in the posterior distribution. If ones makes a higher percentage of good proposal parameters, one would expect the acceptance rate in turn to be higher and hence the overall computational cost of obtaining the posterior would decrease.

We have seen throughout this thesis that an ABC scheme based on sequential Monte Carlo sampling can be used to explore a large prior parameter space, improving the approximation to the posterior in an iterative fashion. The ABC SMC algorithm, as described by Toni *et al.* (2009) (see Algorithm 8), is based on sequential importance sampling. The importance density, $q_t(\Theta_t)$, used for proposals when targeting $\pi_{\epsilon_t}(\Theta_t)$ is based on $\pi_{\epsilon_{t-1}}(\Theta_{t-1})$,

$$q_t(\Theta_t) = \int K(\Theta_t \,|\, \Theta_{t-1}) \pi_{\epsilon_{t-1}}(\Theta), \tag{7.11}$$

where $K(\cdot)$ is some, typically Gaussian, perturbation kernel. Since $\pi_{\epsilon_{t-1}}(\Theta_{t-1})$ is not available directly a discrete approximation to it, consisting of a collection of weighted samples $\left\{\Theta_{t-1}^{(i)}, w_{t-1}^{(i)}\right\}$, is used instead. Namely a sample from our current particle approximation to $\pi_{\epsilon_{t-1}}(\Theta_{t-1})$ is chosen with probability w_{t-1}, propagated through kernel $K(\cdot)$ and is used as a proposal $\Theta_t^* \sim q_t(\Theta)$.

Intuitively one expects that if the difference between $\pi_{\epsilon_{t-1}}(\Theta)$ and $\pi_{\epsilon}(\Theta)$ is small

that $q_t(\Theta)$ chosen in this way that a large number of proposed parameters will have a relatively high probability of acceptance. However as the difference between the two distributions increases one would expect the acceptance rate to decrease. There are then some trade offs to be made in the specification of the algorithm, choosing a tolerance ϵ_t as the α%–ile of distances δ_{t-1} for small α gives rise to an algorithm which attempts to make larger moves at the potential expense of reduced acceptance rate yielding an algorithm which converges towards the final target slowly. Conversely if α is large then a small move between successive distributions produces higher acceptance rates but because moves are small convergence to the final target is potentially still sub optimal.

Given an importance density constructed as in equation 7.11 optimal choices for the variance of the propagation kernel $K(\cdot)$ was explored in Beaumont *et al.* (2009) and Filippi *et al.* (2013). The latter show that the results of the former given a Gaussian propagation kernel are a special case of their general result when $\epsilon_t = \epsilon_{t-1}$. The optimality criterion on which they base choice of proposal variance is based on jointly maximising the acceptance rate and minimising the Kullback Leibler divergence between the proposal density $q_t(\Theta)$ and the target $\pi_{\epsilon_t}(\Theta)$. The results obtained are conditional on the value α and those for a component wise Gaussian perturbation are shown in equation 4.2.

Whilst this gives some optimality criterion for the proposal kernel given a value of α it does not consider what an optimal choice of α might be. We attempt to solve the same problem here, that is choose an importance density which gives high rates of acceptance for large moves towards the final target, meaning in turn lower total computation cost across a shorter sequence of bridging distributions.

7.4.1 Regression proposal kernel

Beaumont *et al.* (2002) introduced the idea of regression correction to a posterior distribution to improve posterior samples. This was explored in chapter 5 and has been used throughout this thesis for improving ABC posterior distributions. Beaumont *et al.* (2002) make the assumption that the posterior distribution $\pi(\Theta \mid \mathbf{S})$ can be modelled by a regression model

$$\Theta_i = \beta_0 + \beta f(\tilde{\mathbf{s}}_i) + \varepsilon_i, \tag{7.12}$$

where $\tilde{\mathbf{s}}_i = (\mathbf{s}_i - \mathbf{s})$. Specifically they use $f(\tilde{\mathbf{s}}) = \tilde{\mathbf{s}}$. We explored in chapter 5 the use of regression adjustment to the posterior and found that $f(\tilde{\mathbf{s}}) = (\tilde{\mathbf{s}}, \tilde{\mathbf{s}}^2)$ gave better results. This intuitively made sense since for each $\mathbf{s}_i$, $(\mathbf{s}_i - \mathbf{s}) \in \mathbb{R}$ where we expect the best posterior samples to take values close to 0 with increasingly poor samples in both directions.

The logic behind why this sort of correction works is such that if the regression model holds then if $\mathbf{s}_i = \mathbf{s}$ then Θ_i is drawn from the posterior distribution $\pi(\Theta \,|\, \mathbf{S} = \mathbf{s})$ with expectation $E(\Theta \,|\, \mathbf{S} = \mathbf{s}) = \beta_0$. Therefore on finding estimates $\hat{\beta}_0$ and $\hat{\beta}$, the values

$$\Theta_i - \hat{\beta} f(\tilde{\mathbf{s}}_i) = \hat{\beta}_0 + \varepsilon_i \tag{7.13}$$

form an approximate sample from $\pi(\Theta \,|\, \mathbf{S} = \mathbf{s})$.

To return to the problem of choosing an importance density $q_t(\Theta_t)$ in a sequential ABC scheme if α is small then in turn ϵ_t is small and $\pi_{\epsilon_t}(\Theta)$ is close to $\pi_0(\Theta) = \pi(\Theta \,|\, \mathbf{S} = \mathbf{s})$. If a regression corrected posterior, now denoted $\pi'_\epsilon(\Theta)$, approximates $\pi_0(\Theta)$ then parameter proposals based on $\pi'_\epsilon(\Theta)$ should have good support in that target.

We propose an alternative to the optimal Gaussian perturbation kernel of Filippi *et al.* (2013) in which an importance density for target $\pi_{\epsilon_t}(\Theta)$ is instead based on the regression corrected intermediate posterior distribution $\pi'_{\epsilon_{t-1}}(\Theta)$. Given weighted samples $\left\{\Theta^{(i)}_{t-1}, w^{(i)}_{t-1}\right\}$ a proposal distribution $q_t(\Theta_t)$ is chosen by first performing a regression correction to obtain the weighted sample $\left\{\Theta'^{,(i)}_{t-1}, w^{(i)}_{t-1}\right\}$ and a perturbation kernel $K_t(\cdot \,|\, \Theta)$ as a Gaussian with zero mean and variance σ^2_ε, the variance of the residuals in the regression model. A proposal is then made by choosing a Θ'_{t-1} with probability $w^{(i)}_{t-1}$ and perturbing to obtain $\Theta^*_t \sim K_t(\cdot \,|\, \Theta'_{t-1})$. The reweighting step of the sequential importance sampler is equivalent to that of Filippi *et al.* (2013), (step 4 of algorithm 8), for a different K_t.

By doing this, using a small α, we expect the Kullback–Leibler divergence between the proposal $q_t(\Theta)$ and the target $\pi_{\epsilon_t}(\Theta)$ to be small, and the overall acceptance rate to be high, the same goal by which Filippi *et al.* (2013) deduce their optimum criterion. The guided proposals should also allow a smaller choice of α yielding larger moves between successive distributions in the bridging sequence whilst retaining reasonable acceptance rates.

There is a very small computational overhead associated with the regression based kernel as compared with the optimal kernel, namely the fitting of a regression model. Whilst fitting regression models to large numbers of samples, summary statistics and rate parameters can be computationally taxing, on considering that this cost is typically dwarfed by the cost of simulating large amounts of data from the model, the regression model fitting is relatively negligible. It is also of note that one can effectively pause the execution of the algorithm after a sample has been obtained to assess the appropriateness of fitted regression models should that be desired. Additionally fitting multiple different regression models does not require simulation from the Markov Process model and hence the cost of this is relatively small.

For convergence in an importance sampling based algorithm there are important conditions that must be placed on the kernel. That is that the support of the kernel must contain the support of the target and that the density must vanish into the tails of the distribution slowly enough to ensure finite variance in the importance weights. This was noted also in Filippi *et al.* (2013). Because the kernel is based on a Gaussian distribution the support is unbounded and therefore must contain the support of the target. There is some question over whether this kernel facilitates finite variance of the importance weights however. Simulated experiments within this context found that the proposal worked well for the examples considered but this does not guarantee that it will be the case for all examples. One method by which we could aim to ensure that the density vanishes slowly enough into the tails is to introduce some multiplicative constant, $\gamma > 1$, to the variance of the proposal, essentially flattening out the density.

7.4.2 Composite proposal kernel

The above approach should perform well whenever the regression correction is reasonable. However it is possible that the regression correction is unstable and the approximation $\pi'_\epsilon(\Theta)$ is wildly different from $\pi_0(\Theta)$. In addition, where the regression corrected posterior has small variance. In order to address this we can consider use of an importance density based on a mixture of both the regression based kernel and the kernel of Filippi *et al.* (2013). This should give a kernel which has increased mass in the target, due to the proposals coming from the regression corrected kernel part whenever the regression correction is good but one that is robust to the cases

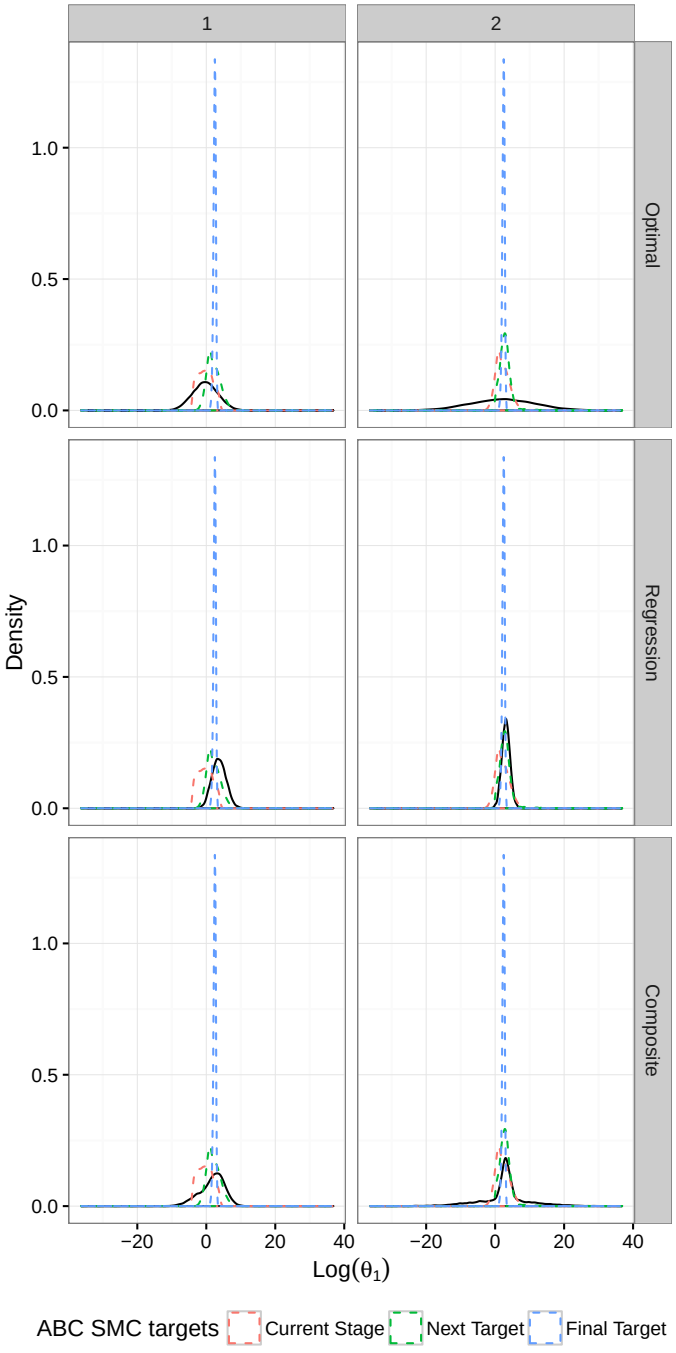


Figure 7.6: Comparison of the distributional shape of the optimal proposal of Filippi *et al.* (2013) and our regression and composite importance densities.

when the regression correction is poor.

Let $K_t(\Theta^* \mid \Theta)$ be the Gaussian kernel whose variance has been chosen according to Filippi *et al.* (2013) and $K'_t(\Theta^* \mid \Theta)$ the Gaussian kernel with variance chosen by the regression correction method. A proposal $\Theta^* \sim q_t(\Theta)$ can then be obtained in the following way,

1. Simulate a Bernoulli random variable to determine which component kernel is to be used $c \sim \text{Bern}(0.5)$.
2. If $c = 0$ sample $\Theta \sim \pi_{\epsilon_{t-1}}(\Theta)$ from the weighted sample approximation otherwise sample $\Theta \sim \pi'_{\epsilon_{t-1}}(\Theta)$.
3. If $c = 0$ sample $\Theta^* \sim K_t(\Theta^* \mid \Theta)$ else $\Theta^* \sim K'_t(\Theta^* \mid \Theta)$.

An estimate of the importance density is obtained as

$$q_t(\Theta) = 0.5 \sum_{i=1}^{N} w^{(i)} K_t(\Theta \mid \Theta_{t-1}^{(i)}) + 0.5 \sum_{i=1}^{N} w^{(i)} K'_t(\Theta \mid \Theta_{t-1}^{\prime,(i)}) \tag{7.14}$$

to be used in the reweighting step of the sequential Monte Carlo scheme.

As above the importance density has unbounded support being a mixture of Gaussians. In addition on utilising the optimal kernel of Filippi et al. we adhere to the same heuristic argument for convergence as in their article.

More generally we could consider the importance density as a mixture of the two kernels with mixture weights $(p, 1-p)$, the example above being the case when $p = 1 - p = 0.5$. We could then, as the ABC SMC scheme progresses keep track of the acceptance rates of proposed parameters conditional on which component of the mixture they came from and adapt the mixture weights to reflect this. The idea being that if we have a case where the regression correction kernel is a poor approximation to the target, with low numbers of proposed parameter vectors from this component being accepted we simply update the importance density such that in future a greater proportion of the proposals come from the optimal Gaussian perturbation kernel. Vice versa, when the regression kernel proposals are good, with high acceptance rates we want a greater proportion of our proposals to come from that component.

Figure 7.6 shows the difference between our regression based proposal kernels and the

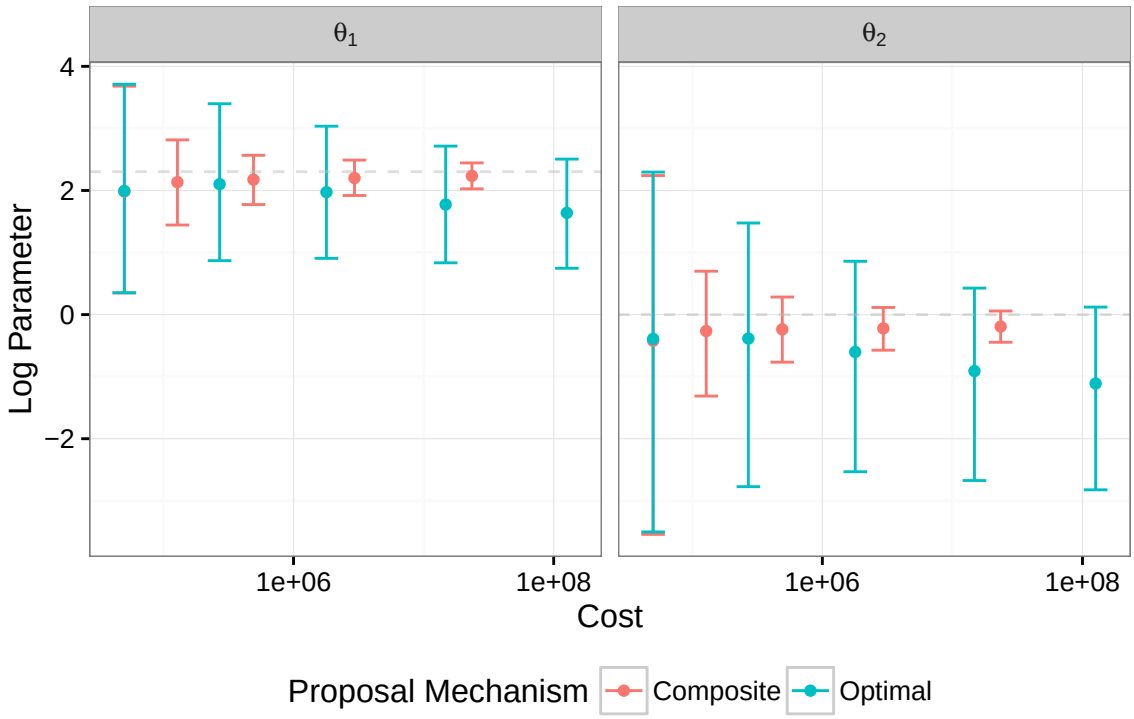


Figure 7.7: Comparison of the sequence of tolerances under the composite proposal and optimal perturbation proposal for inferring rate parameters of the immigration–death process.

optimal kernel of Filippi *et al.* (2013) and two successive stages in a sequential ABC algorithm. The regression based kernels are closer to the next target in the sequence, and also to the final target in both cases. Beaumont *et al.* (2002) acknowledged that the quality of the regression correction was dependent on the size of the tolerance ϵ from which the corresponding samples were taken. This is borne out in figure 7.6 as the regression kernel proposal at the second stage is much closer to the final target.

In addition figure 7.7 demonstrates that the sequence of distributions in a ABC SMC scheme more quickly converges tightly around the true parameter values in an immigration death example given the composite proposal that makes utility of both regression correction and the optimal kernel of Filippi *et al.* (2013). We note that similar results were found in other models, however in an example where regression correction performed poorly, the bistable Schlögl model, it was necessary to incorporate the optimal kernel of Filippi *et al.* (2013) to guarantee convergence.

7.5 Inference for growth parameters in *Bacillus Subtilis* strains

Bacillus Subtilis is a Gram–positive bacterium found in soil and the gastrointestinal tracts of certain mammals and humans. It is a well studied model organism used to study chromosome replication and cell differentiation.

The growth dynamics of the bacteria can be modelled as a stochastic kinetic model. Whilst observations on the number of cells directly is not possible, use of green fluorescent protein (GFP) provides a proxy for tracking the growth.

A simple model to describe the system concerns three species and two reactions,

$$\begin{aligned} N + X &\to 2X \\ X &\to X + G, \end{aligned}$$

where X represents a Bacillus Subtilis cell. N is representative of a source of nutrition for the growing cells, once the supply of N has been exhausted, no further growth of the bacteria can occur. This induces a logistic growth dynamic on the X species, where the rate of growth increases exponentially at first, before resources become limited and the number of cells reaches maximum carrying capacity.

The hazard functions associated with each reaction in this model are

$$h_1(\mathbf{X}, \Theta) = \theta_1 \frac{n \times x}{n + x}, \quad h_2(\mathbf{X}, \Theta) = \theta_2 x, \tag{7.15}$$

where n and x are the species counts of N and X respectively.

In order to make observations on the copy number of X a fluorescent protein is introduced.

Observations of the intensity of green fluorescence are made, such that its level is proportional to the ratio of the copy number of G to the copy number of X.

7.5.1 The data

The data, shown in figure 7.8, consist of a set of discrete time observations of the green fluorescent protein in each of 3 strains of Bacillus Subtilis with and without the

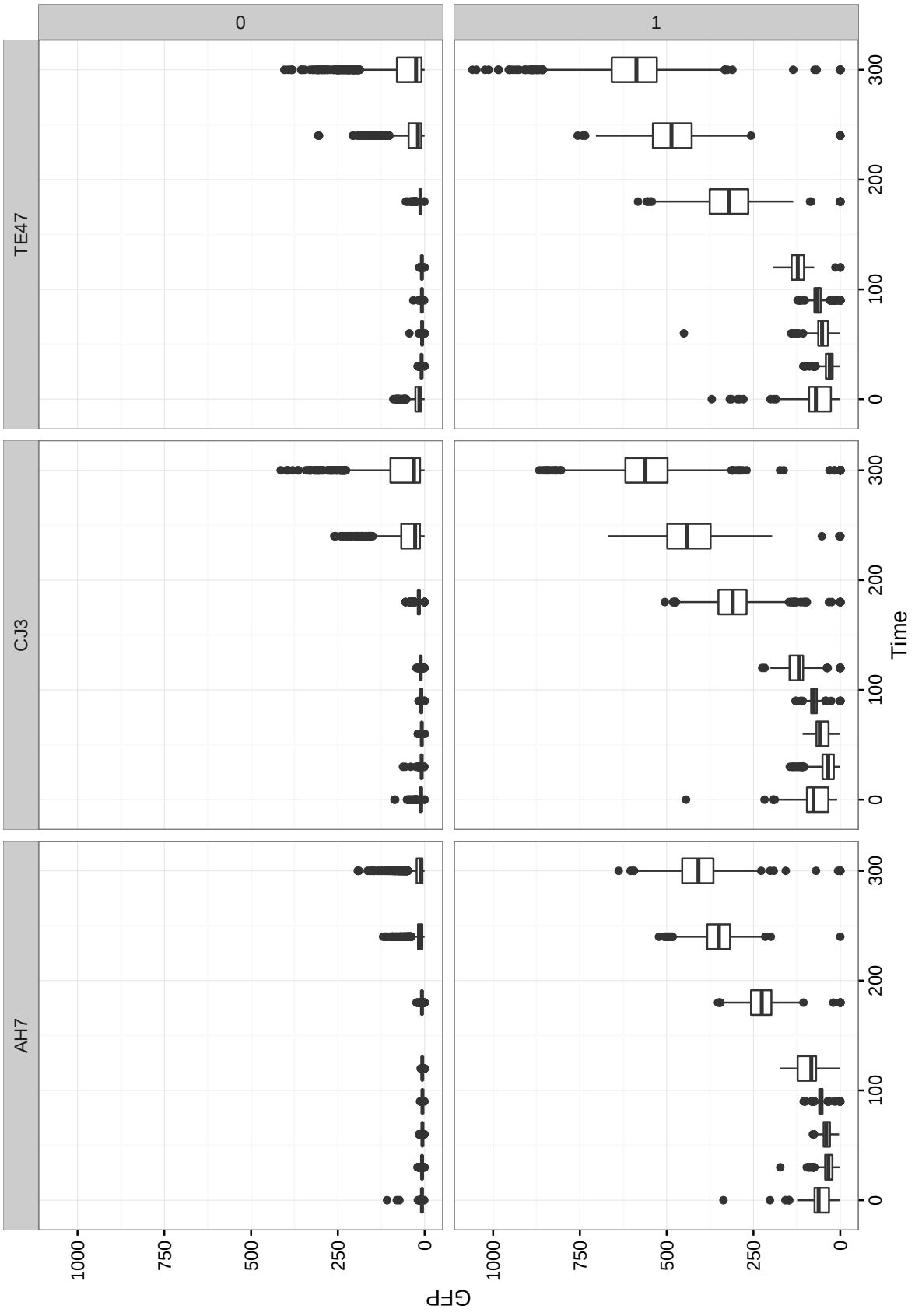


Figure 7.8: Noisy observations of a fluorescent green reporter protein for 3 strains of Bacillus Subtilis with and without an inhibitor.

presence of an inhibitor. We thank Dr Phillip Aldridge of the Centre for Bacterial Cell Biology, Newcastle University, and his PhD student Tom Ewen for the provision of the data. The inhibitor here is expected to allow faster division of the bacterial cells and increased GFP production rate.

7.5.2 Inference

Inferring rate parameters of the Bacillus Subtilis data poses some interesting challenges. Firstly, by the fact that we only have observations that are ratios, a deterministic model would have identifiability issues. Whilst given a stochastic treatment, there could be some information in the intrinsic noise, diffusive cellular dynamics requiring collision between reactants means that processes such as production and degradation are random, there is variation in identically regulated species within a single cell. Without data at the resolution of single cell trajectories there is insufficient information in the noisy data to be able to identify the scale of X.

To proceed then, it is necessary to fix the initial value X_0. Prior information for the other parameters is vague. We know that reaction rates must be positive and the order of magnitude of the rates is expected to be reasonably close to 0.

We therefore place uniform prior distributions on the log reaction rates

$$\log(\Theta) \sim \mathcal{U}(-4, 4) i = 1, 2. \tag{7.16}$$

We have observations proportional G_0/X_0, recordings taken at the time the experiments commenced. We can therefore, given our fixed value of X_0 choose prior distributions for G_0 that have support at the appropriate order of magnitude. Specifically we choose X_0 to be 30, and choose Negative Binomial priors for G_0 such that the mean of the prior distribution is equivalent to the mean of the observations at time 0, rescaled by X_0, and the variance suitably large such that the rescaled observations at time 0 have support within the prior.

Our inference scheme for this model is an ABC SMC scheme with a composite proposal mechanism for propagation of rate parameter samples through the sequential Monte Carlo sampler. This choice was made encouraged by the positive results, increased efficiency in posterior sampling measured as a function of model simula-

tions required, found with the immigration–death toy model. Since simulation from this model was expensive we wished to reduce the amount of model simulation required as much as possible. In addition to using the composite proposal, we use the bootstrapping approach of section 7.2.2 to summary statistic weights and choose summary statistics such that they jointly describe a distributional shape well and facilitate early termination of model simulation as in section 7.3.1.

We choose as summary statistics, at each time point, the mean, standard deviation, and quantiles $q_{20\%}$, $q_{40\%}$, $q_{50\%}$, $q_{60\%}$, $q_{80\%}$ as performance given this choice was good with the toy model. The weights for the summary statistics in the Euclidean metric were calculated as the inverse of the empirical standard deviation of a bootstrap sample representation of the sampling distributions of the observed summary statistics. Each bootstrap sample consisted of 25 observations at each time point such that, for inference, we use only 25 forward simulations to approximate the distribution of G/X for our simulated data.

This choice also means that we can use the early rejection ideas so as to not waste computation time on poor proposals.

7.5.3 Results

Figures 7.9, 7.10, 7.11, 7.12, 7.13 show the inferred approximate posterior distributions for $\log(\theta_1)$, $\log(\theta_2)$ the reaction rate parameters, $\log(\sigma)$ the Gaussian measurement error noise, and G_0 and N_0 the initial state of the reaction network respectively. There is a clear difference in the rate parameters inferred, where the presence of an inhibitor yields an increase in the rate of cell division and production of GFP.

Where observations were low, the inferred observation error is also small, with a greater error variance is inferred when the observed proportion is greater. The posterior distributions, figure 7.11, however suggest that potentially the prior distribution on the error variance was too restrictive.

The inferred initial levels of G_0 are consistent with the magnitude of the ratio observed, given our chosen scaling of X_0, however posterior variance is under estimated.

Finally the inferred distribution of available nutrients is largely consistent across the different strains.

A posterior predictive distribution of the levels of the observable ratio G_t/X_t for the AH7 strain with an inhibitor present is shown in figure 7.14. The posterior predictive distribution shown is representative of the other strains. In each case, whilst the location of the predictive distribution is approximately correct, the posterior predictive variances are under represented. This could be an artifact of fixing the scale at such a small value and highlights a drawback with the current model. Alternatively perhaps less restrictive priors would yield better inference.

The goal of this study was to examine the way in which the fluorescent reporter works. The use of reporter proteins is common amongst practitioners attempting to learn about reactions within a cell. One of the issues with this is that the fluorescence of the reporter protein is itself stochastic and noisy. The results here allow us a better understanding of the way in which the activation of the reporter gene through the inhibitor affects observations of species within the Bacillus Subtilis strains. This allows a better understanding of what's going on when using reporter proteins to examine more complex cellular processes.

7.6 Discussion

The final chapter of this thesis focuses on inference for cell population data using approximate Bayesian techniques. The repercussions of the different order of magnitude of the amount of data, and it's lower resolution at observing the system is discussed. Inference for problems of this nature are typically very computationally taxing due to numerous Gillespie algorithm simulations being required. Hence techniques for reducing the computational cost are explored.

We find that using the observations to make statements about sampling distribution of summary statistics is beneficial to the construction of a metric which yields efficient use of model realisations. Additionally we consider that certain sets of summary statistics allow a practitioner to employ early termination criteria for proposed parameters, allowing faster rejection of proposals that would not be accepted in an ABC scheme. This further optimises the use of available CPU resources by a greater proportion of time being spent considering proposals likely to be accepted.

An alternative to the optimal perturbation kernels of Filippi *et al.* (2013) for propagation of parameter samples through the sequence of distributions in ABC SMC.

This alternative approach considers regression correction as a mechanism to propose a greater proportion of parameters towards the regions of mass in the final target. The technique is made robust to examples where the specified regression model is poor by consideration of a mixture proposal that incorporates the optimal Gaussian perturbation which also ensures that there is adequate mass in the tails of the importance density.

The techniques considered were explored within the context of a toy example, the immigration death process and shown to improve on efficiency over current approaches.

Finally on consideration of the techniques discussed, inference was made on the rate parameters concerning a logistic growth model of different strains of the bacteria Bacillus Subtilis. Inference for this challenging problem seemed reasonable, as confirmed by posterior predictive distributions but it was acknowledged that there are some limitations.

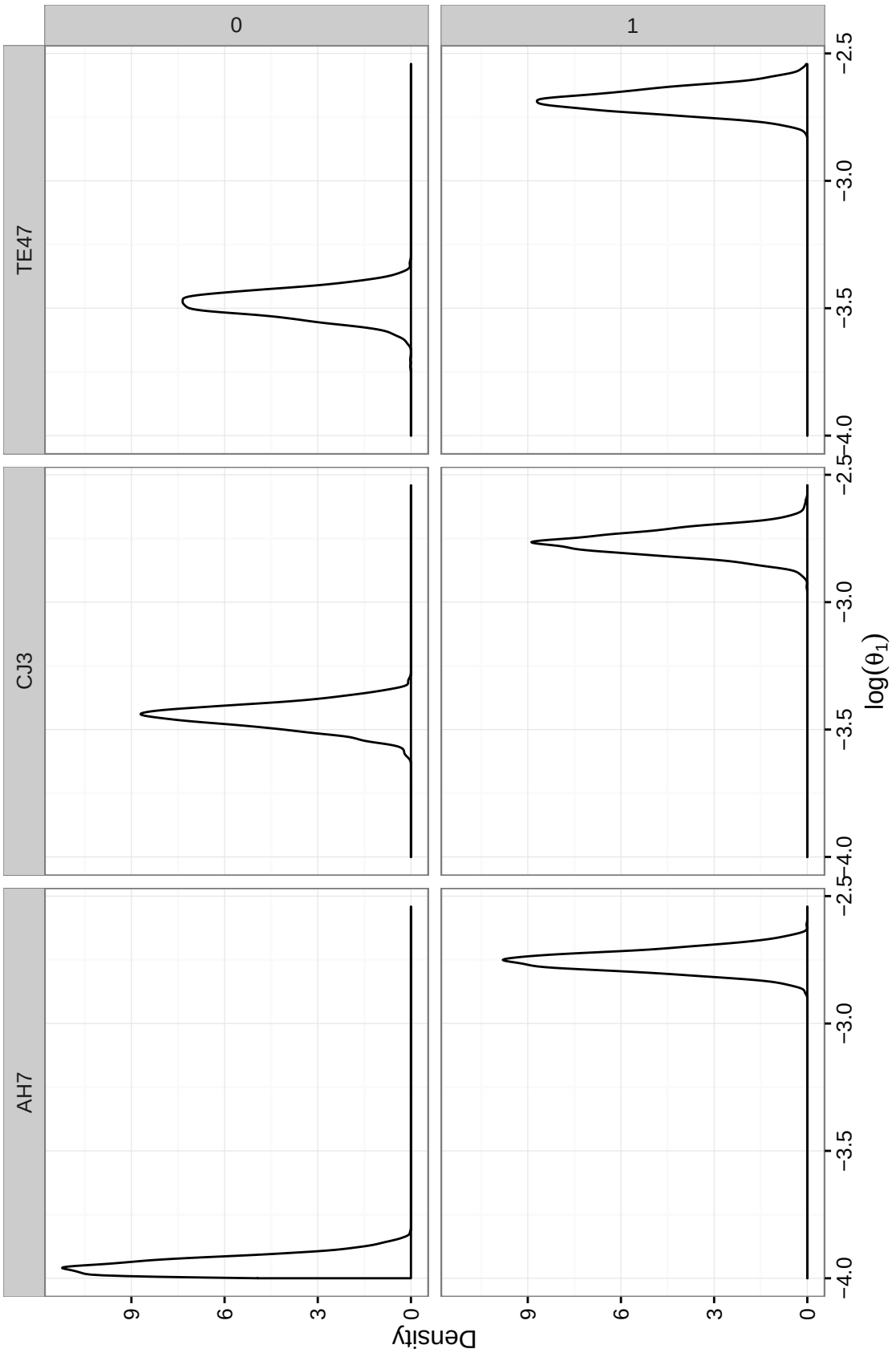


Figure 7.9: Posterior inference of the first rate parameter that controls the rate at which the cells divide, given each strain of the Bacillus Subtilis data.

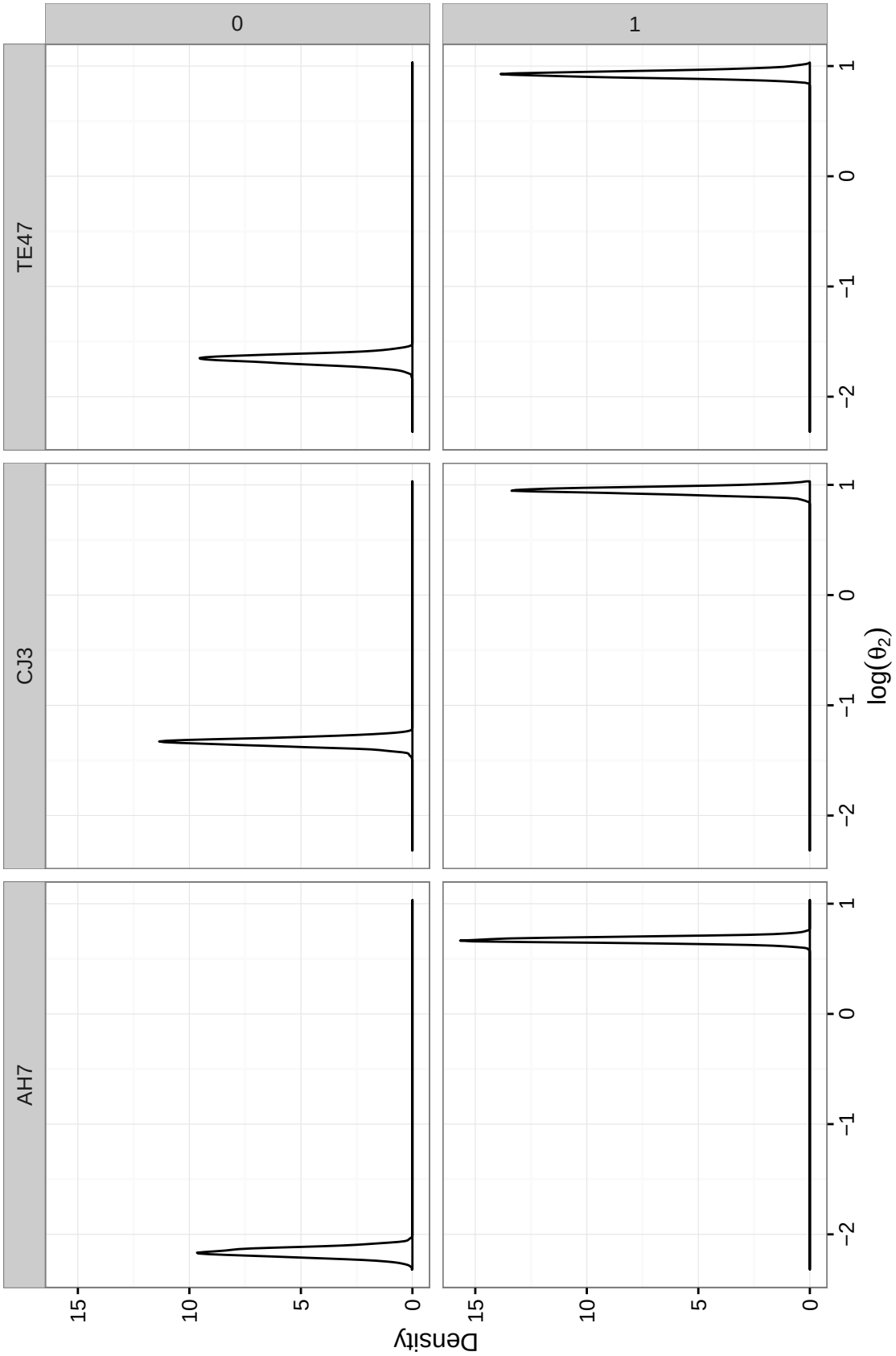


Figure 7.10: Posterior inference of $\log(\theta_2)$. This rate parameter controls the rate that the observable green protein fluoresces.

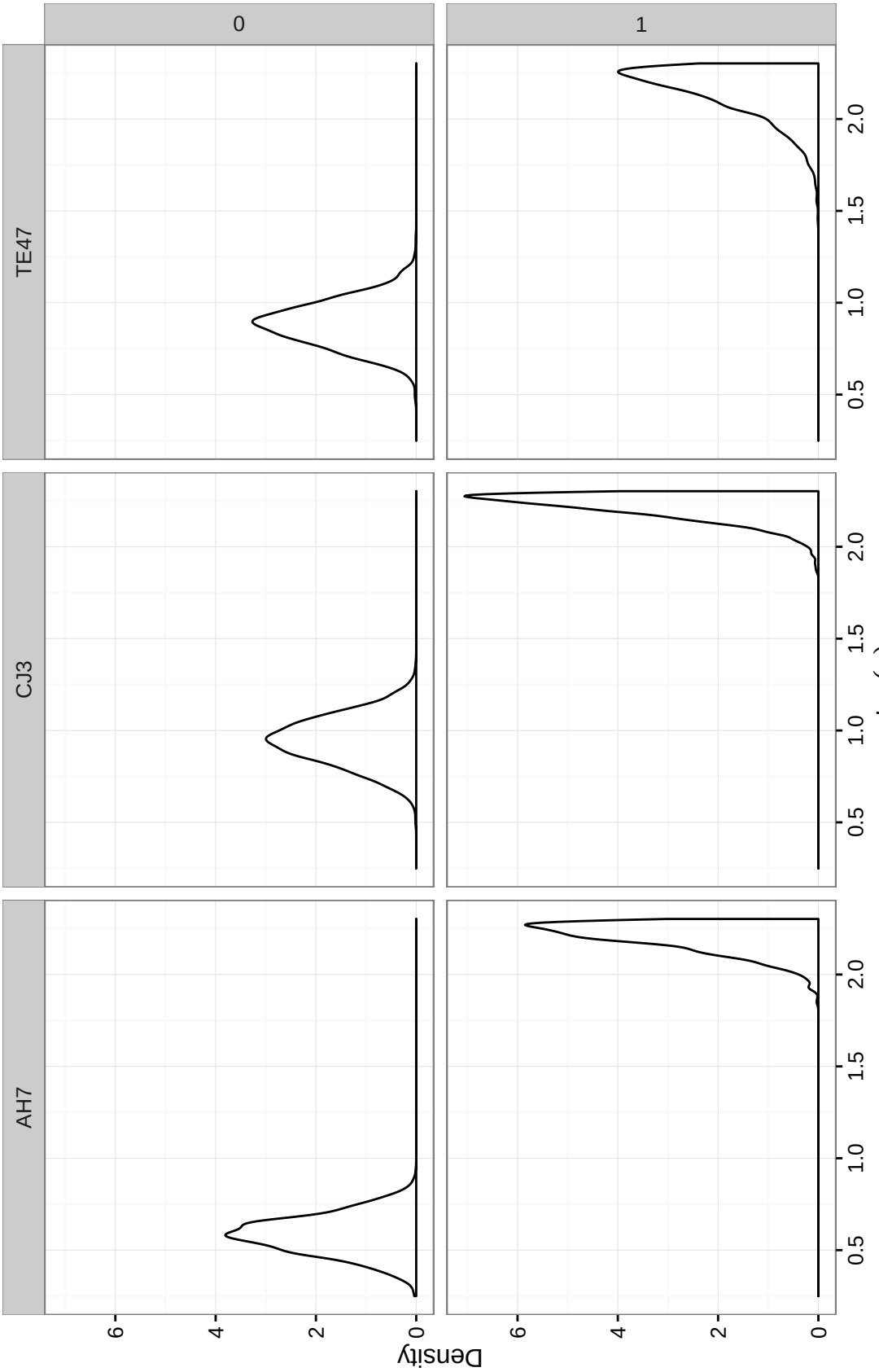


Figure 7.11: Inference of the log of the observation error, $\log(\sigma)$ for the different Bacillus Subtilis strains.

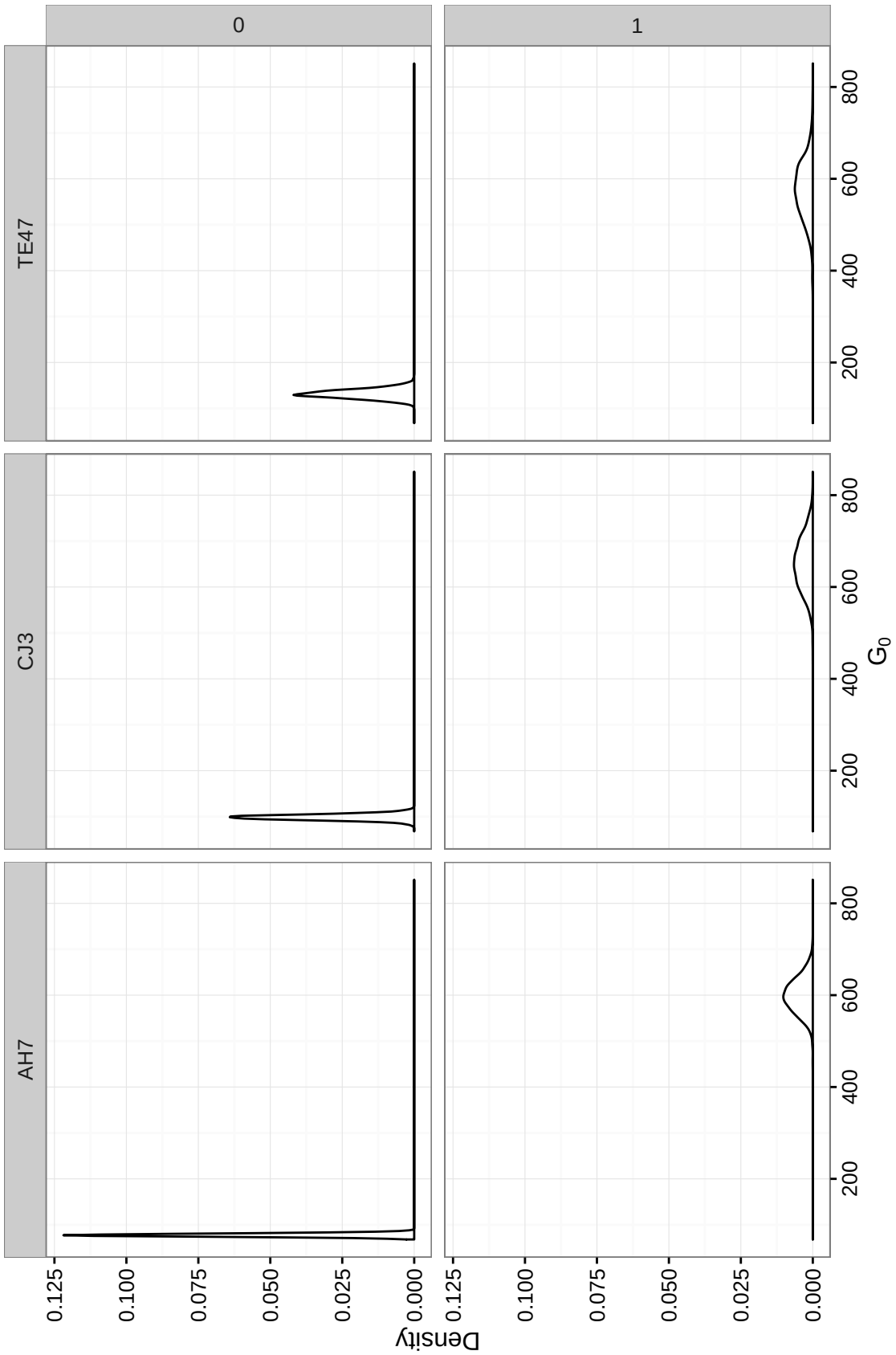


Figure 7.12: Posterior learning of the initial levels of observable Green Fluorescent Protein given the cell population time series for each of the different strains of bacteria.

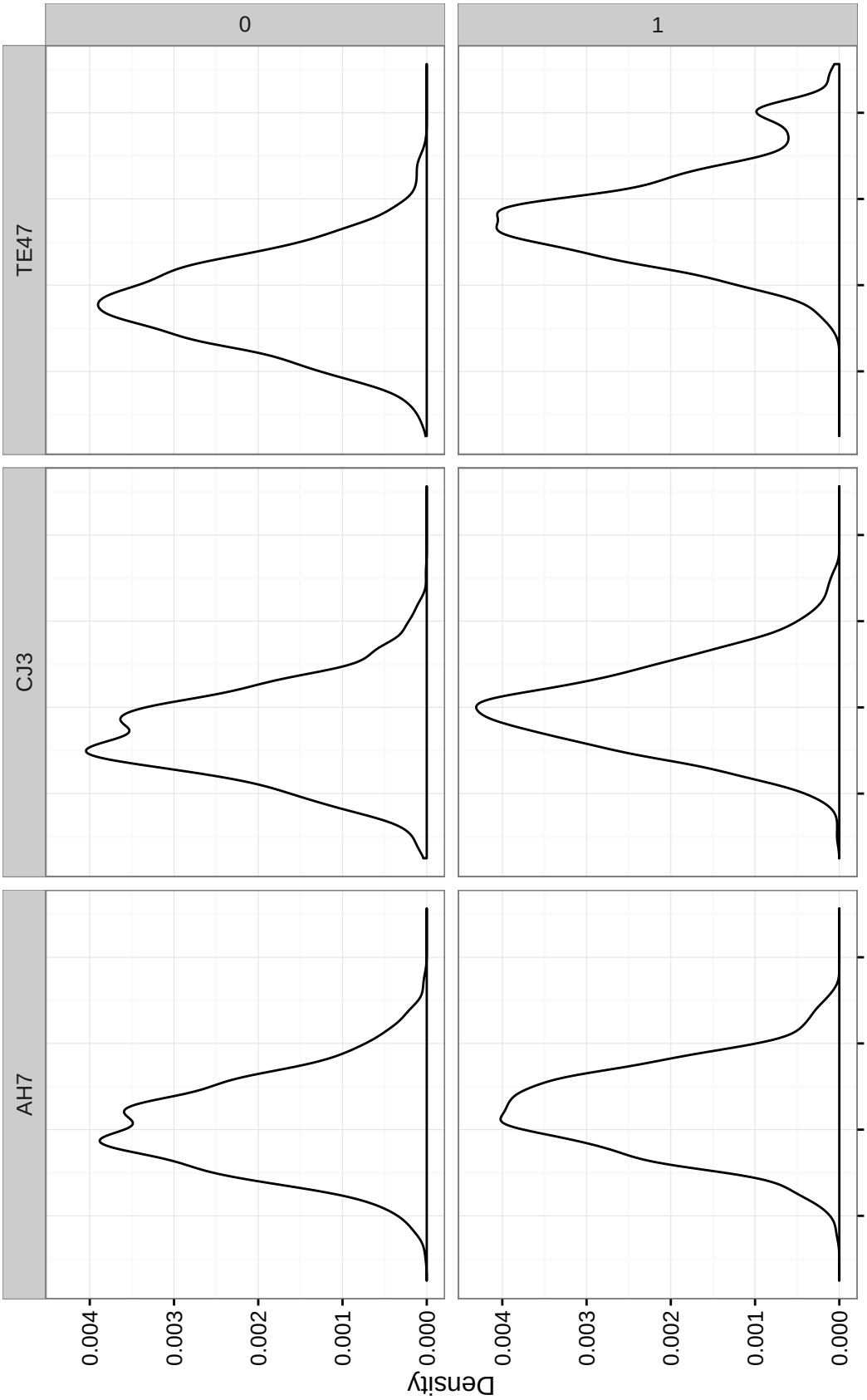


Figure 7.13: Marginal posterior distributions of the initial available nutrients for the growth of the different strains of bacteria.

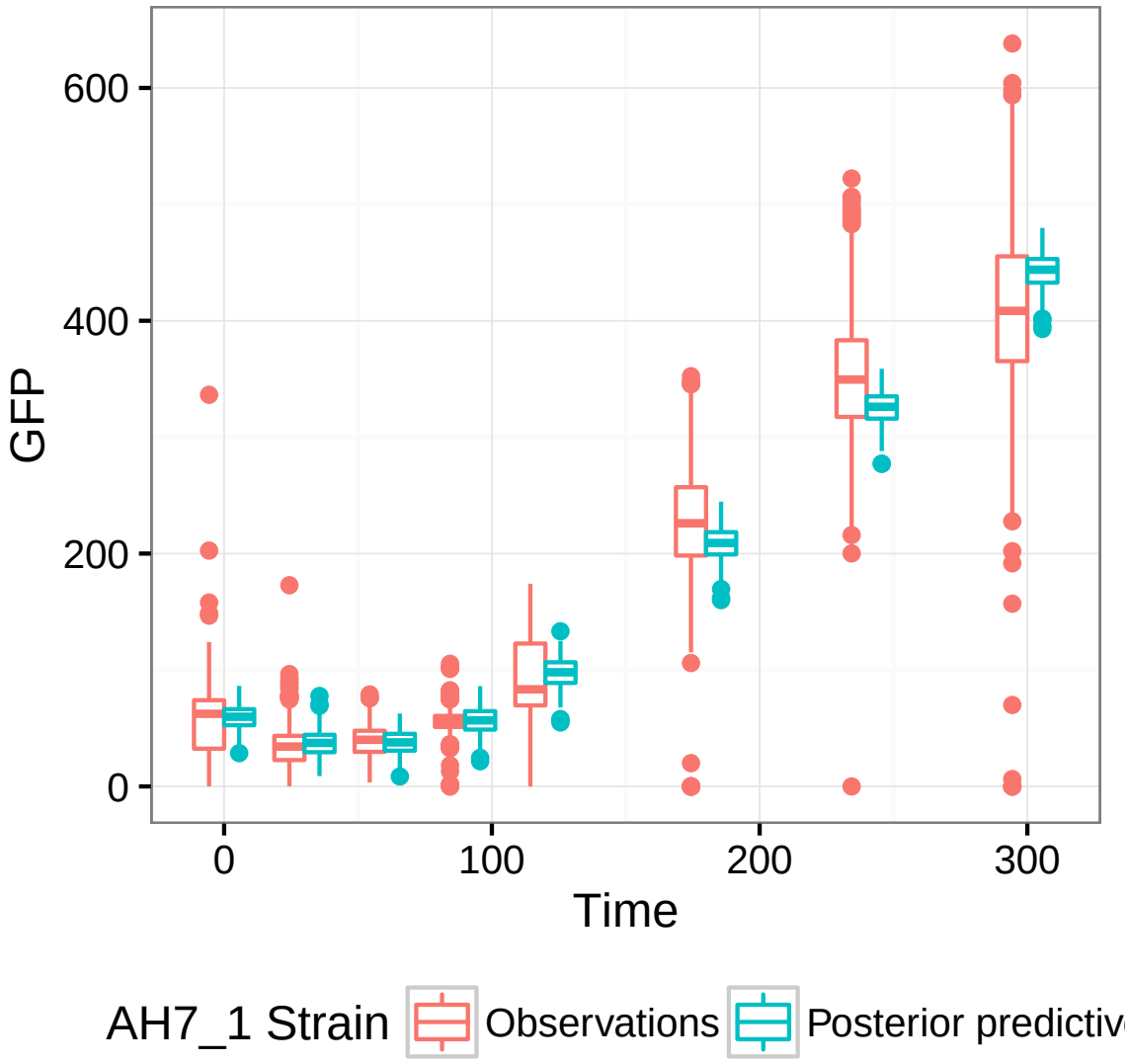


Figure 7.14: A posterior predictive distribution of G_t/X_t for the AH7 Bacillus Subtilis strain in the presence of the inhibitor. Posterior variance is underestimated but this consolidates our belief that inference of the rate parameters is reasonable.

Chapter 8

Concluding discussion

The purpose of this thesis was to provide an in depth, comparative review of the principle methods of Bayesian parameter inference for problems with intractable likelihoods, with particular consideration to posterior learning of rate parameters governing system dynamics of partially observed, continuous time Markov processes. Investigation of the different approaches to approximate Bayesian computation was considered. The rejection sampler, first introduced in the form recognised today by Pritchard *et al.* (1999), and the extensions of it based on Markov chain Monte Carlo and Sequential Monte Carlo were explored within the context of stochastic kinetic models, which find application within population ecology, chemical kinetics and systems biology among others.

Of the different ABC sampling algorithms available to a practitioner it was found that for models of this type a sequential Monte Carlo sampler proved to be the most efficient. For data of the form of a single time series, tracking the evolution of species within a single cell, on reviewing a number of approaches to dimension reduction, it was found that use of the raw time series was competitive, particularly on consideration of posterior regression correction methods like those introduced by Beaumont *et al.* (2002).

Further, ABC was compared with particle Markov chain Monte Carlo (Andrieu & Roberts, 2009; Andrieu *et al.*, 2010), a technique for exact posterior sampling that relies on unbiased Monte Carlo estimation of marginal likelihood in the presence of a measurement error model. Whilst pMCMC is guaranteed to always be able to sample from the posterior distribution, its efficiency is highly dependent on algorithmic

tuning parameters and adequate measurement error. This lead to the conclusion that whilst we typically desire the exact target, good approximations to posterior distributions can be made, with less computational expense, using state of the art ABC sampling.

Each approach has it's own relative merits and deficiencies. ABC techniques are, by and large, amenable to parallel implementation. This allows them to exploit the ever improving and increased abundance of multicore computational hardware, large multi node clusters and cloud computing. Its biggest drawbacks are that inferred posteriors are only approximate and choice of summary statistics and metric functions are still non–obvious in a general setting. However the tuning of other algorithmic parameters, such as the decreasing sequence of tolerances, governing the improving approximation, and exploratative perturbation kernels can be done automatically, as the computation progresses. Particle MCMC on the other hand gives the exact target as it's stationary distribution, under fairly mild conditions of the Monte Carlo estimator. Unfortunately, naive initialisation and tuning can lead to the computational burden of obtaining a diverse set of informative samples becoming prohibitive, even in relatively simple reaction networks. Additionally the difficulty in optimising a particle filter to fully parallel computation, and the iterative nature of a Markov chain mean that performance gains do not scale well with available computing resources. The comparative review of the two approaches within this modelling framework was published by Owen *et al.* (2015).

These conclusions lead to the development of a hybrid algorithm that utilises the strengths of each of the sampling schemes. It provides a principled approach to the determination of initialisation and tuning parameters necessary for efficient sampling in a particle MCMC algorithm by using an ABC approximation to the target from which posterior samples are desired. The use of an ABC step gives an artificial burn in period, replacing potentially long convergence times in pMCMC, that can be performed in parallel and scales well with available hardware. Having an approximation to the target allows an essentially arbitrary number of independent MCMC chains to be constructed, each of which should have short burn in periods allowing parallel execution that scales well to give an efficient manner in which to sample the posterior distribution exactly. The hybrid scheme was shown to work well to a number of example stochastic kinetic models, including application to both synthetic and real data and formed the basis of Owen *et al.* (2014).

Finally the ABC framework was considered for parameter learning for models given large scale, noisy cell population level data. The large dimension of the observations poses it's own computational issues and consideration was given to optimally using available resources. Techniques to minimise the amount of model simulation steps were explored and an approach to improving the proportion importance density proposals in regions of high target density in the ABC SMC scheme was proposed, based on and relying upon the work of Filippi *et al.* (2013) with numerical experiments of a toy model used for discussion.

The concluding example of the thesis made inference of rate parameters of a model governing growth dynamics of different strains of Bacillus Subtilis, a bacteria commonly found in soil, often used to study cell division. Observation of the number of cells directly was unavailable, instead bioluminescent intensity of a green fluorescent reporter protein was used as a proxy.

8.1 Future work

There are numerous avenues of exploration to extend this work. The particle MCMC papers (Andrieu & Roberts, 2009; Andrieu *et al.*, 2010) show that multiple samples can be used in the Monte Carlo estimation of a likelihood function to target an exact posterior distribution, provided that the Monte Carlo estimates are unbiased. The results they derive are not limited to estimation of a likelihood function and in fact hold also in ABC approximations for example. Whilst there are some examples of the use of multiple samples conditional on the same parameter values within ABC, see for example Becquet & Przeworski (2007) or Del Moral *et al.* (2012), as far as this author knows, investigation into an optimal choice for numbers of particles for ABC is lacking.

Treatment of the Bacillus Subtilis data in section 7.4.2 considers treatment of each available strain in isolation. A natural extension to the work presented here would be to extend the methodology to a hierarchical modelling framework which allows each strain to considered jointly. Further, for this example, the model does have limitations. It is possible that the degradation rate of the observable protein is sufficiently high that it should be included in the model. Additionally, because there are fundamental identifiability issues in the deterministic model, it was required to

fix the scale in order to make inference on the rate parameters. It is not clear what the sensitivity of results is to this decision.

In this thesis all model simulation was performed using the direct method of Gillespie (1977) which provides exact samples from the transition densities of the stochastic kinetic models. It was noted in chapter 2 that there exist numerous approximate simulation algorithms however Gillespie (2016) shows that they are not always appropriate and developed an approach for assessing their accuracy. In particular consideration of the large scale cell population data considered in chapter 7 it should be possible to leverage the information contained within the ABC distance calculation to assess whether or not an approximate simulation algorithm is appropriate by using a mixture of exact and approximate samples from the model. This could potentially be used as a mechanism for the construction of a hybrid simulator, that simulates primarily from the exact stochastic simulation algorithm in regions of space in which approximations are poor, but allows faster sampling in regions where approximations are good.

Further related work that could be considered is the use of Gaussian processes emulation to reduce computation time. The idea was briefly considered, although not reported here, on consideration of the synthetic likelihood, to attempt to mitigate the burden of requiring large numbers of model realisations at each iteration. A similar idea was explored in Drovandi *et al.* (2015) and it would be interesting to consider the extension of this to sampling within the ABC framework.

Bibliography

Aeschbacher, S., Beaumont, M. A. & Futschik, A. 2012 A novel approach for choosing summary statistics in approximate Bayesian computation. *Genetics* **192** (3), 1027–1047.

Andrieu, C., Doucet, A. & Holenstein, R. 2010 Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72** (3), 269–342.

Andrieu, C. & Roberts, G. O. 2009 The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics* **37** (2), 697–725.

Beaumont, M. A. 2003 Estimation of population growth or decline in genetically monitored populations. *Genetics* **164** (3), 1139–1160.

Beaumont, M. A., Cornuet, J.-M., Marin, J.-M. & Robert, C. P. 2009 Adaptive approximate Bayesian computation. *Biometrika* **96** (4), 983–990.

Beaumont, M. A., Zhang, W. & Balding, D. J. 2002 Approximate Bayesian computation in population genetics. *Genetics* **162** (4), 2025–2035.

Becquet, C. & Przeworski, M. 2007 A new approach to estimate parameters of speciation models with application to apes. *Genome research* **17** (10), 1505–1519.

Blum, M. G. & François, O. 2010 Non-linear regression models for approximate Bayesian computation. *Statistics and Computing* **20** (1), 63–73.

Blum, M. G., Nunes, M. A., Prangle, D., Sisson, S. A. *et al.* 2013 A comparative review of dimension reduction methods in approximate bayesian computation. *Statistical Science* **28** (2), 189–208.

Bortot, P., Coles, S. G. & Sisson, S. A. 2007 Inference for stereological extremes. *Journal of the American Statistical Association* **102** (477), 84–92.

Bower, J. M. & Bolouri, H. 2004 *Computational modeling of genetic and biochemical networks*. MIT press.

Boys, R. J., Wilkinson, D. J. & Kirkwood, T. B. 2008 Bayesian inference for a discretely observed stochastic kinetic model. *Statistics and Computing* **18** (2), 125–135.

Csilléry, K., Blum, M. G., Gaggiotti, O. E., François, O. *et al.* 2010 Approximate Bayesian computation (abc) in practice. *Trends in ecology & evolution* **25** (7), 410–418.

Csilléry, K., François, O. & Blum, M. G. 2012 abc: an r package for approximate bayesian computation (abc). *Methods in ecology and evolution* **3** (3), 475–479.

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.

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.

Diggle, P. J. & Gratton, R. J. 1984 Monte Carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 193–227.

Douc, R., Guillin, A., Marin, J.-M., Robert, C. P. *et al.* 2007 Convergence of adaptive mixtures of importance sampling schemes. *The Annals of Statistics* **35** (1), 420–448.

Doucet, A. 1998 On sequential monte carlo methods for bayesian filtering, dept. eng., univ. cambridge. *Tech. Rep.*. UK, Tech. Rep.

Doucet, A., De Freitas, N. & Gordon, N. 2001*a* An introduction to sequential monte carlo method. *SMC in Practice* .

Doucet, A., de Freitas, N. & Gordon, N. 2001*b* *Sequential Monte Carlo methods in practice*. Springer.

Doucet, A., Pitt, M. & Kohn, R. 2015 Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika* **102**, 295–313.

Drovandi, C. C., Moores, M. T. & Boys, R. J. 2015 Accelerating pseudo-marginal mcmc using gaussian processes.

Drovandi, C. C. & Pettitt, A. N. 2011 Estimation of parameters for macroparasite population evolution using approximate Bayesian computation. *Biometrics* **67** (1), 225–233.

Fearnhead, P. & Prangle, D. 2012 Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74** (3), 419–474.

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.

Finch, C. E. & Kirkwood, T. B. 2000 *Chance, development, and aging*. Oxford University Press.

Fu, Y.-X. & Li, W.-H. 1997 Estimating the age of the common ancestor of a sample of dna sequences. *Molecular biology and evolution* **14** (2), 195–199.

Gamerman, D. 1997 Markov chain Monte Carlo: Stochastic simulation for Bayesian inference .

Gelfand, A. E. & Smith, A. F. 1990 Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association* **85** (410), 398–409.

Gelman, A. 1996 Inference and monitoring convergence in markov chain monte carlo in practice. wr gilks, s. richardson, and dj spiegelhalter, eds, pp 131-143.

Gillespie, C. S. 2003 *Counting statistics of stochastic processes*. University of Strathclyde.

Gillespie, C. S. 2016 Diagnostics for assessing the accuracy of approximate stochastic simulators. *Statistical Applications in Genetics and Molecular Biology, in press* .

Gillespie, C. S. & Golightly, A. 2010 Bayesian inference for generalized stochastic population growth models with application to aphids. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **59** (2), 341–357.

Gillespie, D. T. 1977 Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry* **81** (25), 2340–2361.

Gillespie, D. T. 1992 A rigorous derivation of the chemical master equation. *Physica A: Statistical Mechanics and its Applications* **188** (1), 404–425.

Gillespie, D. T. 2000 The chemical Langevin equation. *The Journal of Chemical Physics* **113** (1), 297–306.

Gillespie, D. T. 2001 Approximate accelerated stochastic simulation of chemically reacting systems. *The Journal of Chemical Physics* **115** (4), 1716–1733.

Golightly, A., Henderson, D. & Sherlock, C. 2014 Delayed acceptance particle mcmc for exact inference in stochastic kinetic models. *Statistics and Computing* pp. 1–17.

Golightly, A. & Wilkinson, D. J. 2005 Bayesian inference for stochastic kinetic models using a diffusion approximation. *Biometrics* **61** (3), 781–788.

Golightly, A. & Wilkinson, D. J. 2011 Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo. *Interface Focus* **1** (6), 807–820.

Golightly, A. & Wilkinson, D. J. 2015 Bayesian inference for markov jump processes with infromative observations. *Statistical Applications in Genetics and Molecular Biology* **14**, 169–188.

Gordon, N. J., Salmond, D. J. & Smith, A. F. 1993 Novel approach to nonlinear/non-gaussian bayesian state estimation. In *IEE Proceedings F (Radar and Signal Processing)*, , vol. 140, pp. 107–113. IET.

Hastings, W. K. 1970 Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57** (1), 97–109.

Heidelberger, P. & Welch, P. D. 1983 Simulation run length control in the presence of an initial transient. *Operations Research* **31** (6), 1109–1144.

Hey, J. & Machado, C. A. 2003 The study of structured populationsnew hope for a difficult and divided science. *Nature Reviews Genetics* **4** (7), 535–543.

Jahnke, T. & Huisinga, W. 2007 Solving the chemical master equation for monomolecular reaction systems analytically. *Journal of mathematical biology* **54** (1), 1–26.

Jones, P. J., Sim, A., Taylor, H. B., Bugeon, L., Dallman, M. J., Pereira, B., Stumpf, M. P. & Liepe, J. 2015 Inference of random walk models to describe leukocyte migration. *Physical Biology* **12** (6), 066001.

Joyce, P. & Marjoram, P. 2008 Approximately sufficient statistics and Bayesian computation. *Statistical applications in genetics and molecular biology* **7** (1).

Kalman, R. E. 1960 A new approach to linear filtering and prediction problems. *Journal of basic Engineering* **82** (1), 35–45.

Kitano, H. *et al.* 2001 *Foundations of systems biology*. MIT press Cambridge.

Komorowski, M., Finkenstädt, B., Harper, C. V. & Rand, D. A. 2009 Bayesian inference of biochemical kinetic parameters using the linear noise approximation. *BMC Bioinformatics* **10** (1), 343.

Kullback, S. & Leibler, R. A. 1951 On information and sufficiency. *The annals of mathematical statistics* **22** (1), 79–86.

Lillacci, G. & Khammash, M. 2013 The signal within the noise: efficient inference of stochastic gene regulation models using fluorescence histograms and stochastic simulations. *Bioinformatics* **29** (18), 2311–2319.

Liu, J. S. & Chen, R. 1998 Sequential monte carlo methods for dynamic systems. *Journal of the American statistical association* **93** (443), 1032–1044.

Lopes, J. & Beaumont, M. 2010 Abc: a useful bayesian tool for the analysis of population data. *Infection, Genetics and Evolution* **10** (6), 825–832.

Lotka, A. J. 1925 *Elements of physical biology*. Williams & Wilkins Baltimore.

Marjoram, P., Molitor, J., Plagnol, V. & Tavaré, S. 2003 Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences* **100** (26), 15324–15328.

Matis, J. H., Kiffe, T. R., Matis, T. I. & Stevenson, D. E. 2007 Stochastic modeling of aphid population growth with nonlinear, power-law dynamics. *Mathematical Biosciences* **208** (2), 469–494.

Matis, T. I., Parajulee, M. N., Matis, J. H. & Shrestha, R. B. 2008 A mechanistic model based analysis of cotton aphid population dynamics data. *Agricultural and Forest Entomology* **10** (4), 355–362.

McKinley, T., Cook, A. R. & Deardon, R. 2009 Inference in epidemic models without likelihoods. *The International Journal of Biostatistics* **5** (1).

McQuarrie, D. A. 1967 Stochastic approach to chemical kinetics. *Journal of applied probability* **4** (3), 413–478.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. 1953 Equation of state calculations by fast computing machines. *The journal of chemical physics* **21**, 1087.

Murray, J. D. 2002 Mathematical biology i: An introduction, vol. 17 of interdisciplinary applied mathematics.

Nunes, M. A. & Balding, D. J. 2010 On optimal selection of summary statistics for approximate Bayesian computation. *Statistical applications in genetics and molecular biology* **9** (1).

Owen, J., Wilkinson, D. & Gillespie, C. 2014 Scalable inference for markov processes with intractable likelihoods. *Statistics and Computing* pp. 1–12.

Owen, J., Wilkinson, D. J. & Gillespie, C. S. 2015 Likelihood free inference for markov processes: a comparison. *Statistical applications in genetics and molecular biology* **14** (2), 189–209.

Pitt, M. K., Silva, R. d. S., Giordani, P. & Kohn, R. 2012 On some properties of Markov chain Monte Carlo simulation methods based on the particle filter. *Journal of Econometrics* **171** (2), 134–151.

Prangle, D. *et al.* 2016 Adapting the abc distance function. *Bayesian Analysis* .

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A. & Feldman, M. W. 1999 Population growth of human y chromosomes: a study of y chromosome microsatellites. *Molecular Biology and Evolution* **16** (12), 1791–1798.

Proctor, C., Lydall, D., Boys, R., Gillespie, C., Shanley, D., Wilkinson, D. & Kirkwood, T. 2007 Modelling the checkpoint response to telomere uncapping in budding yeast. *Journal of The Royal Society Interface* **4** (12), 73–90.

Raftery, A. & Lewis, S. 1996 Implementing mcmc, in markov chain monte carlo in practice, wr gilks, s. richardson, and dj spiegelhalter, eds, pp 115-130.

Roberts, G. O., Gelman, A. & Gilks, W. R. 1997 Weak convergence and optimal scaling of random walk metropolis algorithms. *The Annals of Applied Probability* **7** (1), 110–120.

Roberts, G. O. & Rosenthal, J. S. 2001 Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science* **16** (4), 351–367.

Ross, S. M. 1983 Stochastic processes, john willy & sons. *New York* .

Rubin, D. B. 1987 A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: the sir algorithm. *Journal of the American Statistical Association* **82** (398), 543–546.

Rubin, D. B. *et al.* 1988 Using the sir algorithm to simulate posterior distributions. *Bayesian statistics* **3** (1), 395–402.

Schlögl, F. 1972 Chemical reaction models for non-equilibrium phase transitions. *Zeitschrift für Physik* **253** (2), 147–161.

Shannon, C. E. 1948 A mathematical theory of communication. *The Bell System Technical Journal* **27** (4), 623–656.

Sherlock, C., Thiery, A. H., Roberts, G. O. & Rosenthal, J. S. 2015 On the efficiency of pseudo-marginal random walk Metropolis algorithms. *The Annals of Statistics* **43** (1), 238–275.

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.

Sisson, S., Fan, Y. & Tanaka, M. M. 2007 Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences* **104** (6), 1760–1765.

Tavare, S., Balding, D. J., Griffiths, R. & Donnelly, P. 1997 Inferring coalescence times from dna sequence data. *Genetics* **145** (2), 505–518.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A. & Stumpf, M. P. 2009 Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface* **6** (31), 187–202.

Van Kampen, N. G. 1992 *Stochastic processes in physics and chemistry*, , vol. 1. Elsevier.

Volterra, V. 1926 Fluctuations in the abundance of a species considered mathematically. *Nature* **118**, 558–560.

Wang, Q., Kulkarni, S. R. & Verdú, S. 2006 A nearest-neighbor approach to estimating divergence between continuous random vectors. *convergence* **1000** (1), 11.

Wegmann, D., Leuenberger, C. & Excoffier, L. 2009 Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics* **182** (4), 1207–1218.

Weiss, G. & von Haeseler, A. 1998 Inference of population history using a likelihood approach. *Genetics* **149** (3), 1539–1546.

Wilkinson, D. J. 2006 *Parallel Bayesian computation*, *Statistics Textbooks and Monographs*, vol. 184. MARCEL DEKKER AG.

Wilkinson, D. J. 2009 Stochastic modelling for quantitative description of heterogeneous biological systems. *Nature Reviews Genetics* **10** (2), 122–133.

Wilkinson, D. J. 2011 *Stochastic modelling for systems biology*, 2nd edn., *Chapman & Hall/CRC mathematical biology and medicine series*, vol. 44. CRC press.

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.

Zechner, C., Ruess, J., Krenn, P., Pelet, S., Peter, M., Lygeros, J. & Koeppl, H. 2012 Moment-based inference predicts bimodality in transient gene expression. *Proceedings of the National Academy of Sciences* **109** (21), 8340–8345.

Zheng, Q. & Ross, J. 1991 Comparison of deterministic and stochastic kinetics for nonlinear systems. *The Journal of chemical physics* **94** (5), 3644–3648.