Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems

Tina Toni[1,2,*], David Welch[3,†], Natalja Strelkowa[4],
Andreas Ipsen[5] and Michael P. H. Stumpf[1,2,*]

[1]*Centre for Bioinformatics, Division of Molecular Biosciences,* [2]*Institute of Mathematical Sciences,* [3]*Department of Epidemiology and Public Health,* [4]*Department of Bioengineering, and* [5]*Department of Biomolecular Medicine, Imperial College London, London SW7 2AZ, UK*

Approximate Bayesian computation (ABC) methods can be used to evaluate posterior distributions without having to calculate likelihoods. In this paper, we discuss and apply an ABC method based on sequential Monte Carlo (SMC) to estimate parameters of dynamical models. We show that ABC SMC provides information about the inferability of parameters and model sensitivity to changes in parameters, and tends to perform better than other ABC approaches. The algorithm is applied to several well-known biological systems, for which parameters and their credible intervals are inferred. Moreover, we develop ABC SMC as a tool for model selection; given a range of different mathematical descriptions, ABC SMC is able to choose the best model using the standard Bayesian model selection apparatus.

1. INTRODUCTION

Most dynamical systems studied in the physical, life and social sciences and engineering are modelled by ordinary, delay or stochastic differential equations. However, for the vast majority of systems and particularly for biological systems, we lack reliable information about parameters and frequently have several competing models for the structure of the underlying equations. Moreover, the biological experimental data are often scarce and incomplete, and the likelihood surfaces of large models are complex. The analysis of such dynamical systems therefore requires new, more realistic quantitative and predictive models. Here, we develop novel statistical tools that allow us to analyse such data in terms of dynamical models by (i) providing estimates for model parameter values, and (ii) allowing us to compare the performance of different models in describing the overall data.

In the last decade, extensive research has been conducted on estimating the parameters of deterministic systems. Much attention has been given to local and global nonlinear optimization methods (Mendes & Kell 1998; Moles *et al.* 2003) and, generally, parameter estimation

has been performed by maximum-likelihood estimation (Muller *et al.* 2004; Timmer & Muller 2004; Baker *et al.* 2005; Bortz & Nelson 2006). The methods developed for ordinary differential equations have been extended to ordinary differential equations with time delays (Horbelt *et al.* 2002). Deterministic models have also been parametrized in a Bayesian framework using Bayesian hierarchical models (Putter *et al.* 2002; Banks *et al.* 2005; Huang *et al.* 2006). Simulated annealing, which attempts to avoid getting trapped in local minima, is another well-known optimization algorithm that has been found successful in various applications (Kirkpatrick *et al.* 1983; Mendes & Kell 1998). There are also several Monte Carlo-based approaches applied to the parameter estimation of deterministic (Battogtokh *et al.* 2002; Brown & Sethna 2003) and stochastic (Sisson *et al.* 2007) systems. The parameter estimation for stochastic models has been extensively explored in financial mathematics (Johannes & Polson 2005) and has been applied to biological systems in a frequentist maximum likelihood (Reinker *et al.* 2006) and Bayesian (Golightly & Wilkinson 2005, 2006; Wilkinson 2006) framework.

Most commonly, model selection has been performed by likelihood ratio tests (in the case of nested models) or the Akaike information criterion (in the case of non-nested models). Recently, Bayesian methods have increasingly been coming into use. Vyshemirsky & Girolami (2008) have investigated different ways of computing the Bayes factors for model selection of deterministic differential equation models, and

*Authors and address for correspondence: Centre for Bioinformatics, Imperial College London, London SW7 2AZ, UK (ttoni@imperial.ac.uk; m.stumpf@imperial.ac.uk).
†Present address: Department of Statistics, Pennsylvania State University, University Park, PA 16802, USA.

Brown & Sethna (2003) have used the Bayesian information criterion. In population genetics, model selection has been performed using approximate Bayesian computation (ABC) in its basic rejection form (Zucknick 2004; Wilkinson 2007) and coupled with multinomial logistic regression (Beaumont 2008a; Fagundes *et al.* 2007).

There is thus a wide variety of tools available for parameter estimation and, to a lesser extent, model selection. However, to our knowledge, no method available can be applied to all different kinds of modelling approaches (e.g. ordinary or stochastic differential equations with and without time delay) without substantial modification, estimate credible intervals from incomplete or partially observed data, reliably explore the whole parameter space without getting trapped in local extrema and be employed for model selection.

In this paper, we apply an ABC method based on sequential Monte Carlo (SMC) to the parameter estimation and model selection problem for dynamical models. In ABC methods, the evaluation of the likelihood is replaced by a simulation-based procedure (Pritchard *et al.* 1999; Beaumont *et al.* 2002; Marjoram *et al.* 2003; Sisson *et al.* 2007). We explore the information provided by ABC SMC about the inferability of parameters and the sensitivity of the model to parameter variation. Furthermore, we compare the performance of ABC SMC with other ABC methods. The method is illustrated on two simulated datasets (one from ecology and another from molecular systems biology), and real and simulated epidemiological datasets. As we will show, ABC SMC yields reliable parameter estimates with credible intervals, can be applied to different types of models (e.g. deterministic as well as stochastic models), is relatively computationally efficient (and easily parallelized), allows for discrimination among sets of candidate models in a formal Bayesian model selection sense and gives us an assessment of parameter sensitivity.

2. METHODS

In this section, we review and develop the theory underlying ABC with emphasis on applications to dynamical systems, before introducing a formal Bayesian model selection approach in an ABC context.

2.1. Approximate Bayesian computation

ABC methods have been conceived with the aim of inferring posterior distributions where likelihood functions are computationally intractable or too costly to evaluate. They exploit the computational efficiency of modern simulation techniques by replacing the calculation of the likelihood with a comparison between the observed and simulated data.

Let θ be a parameter vector to be estimated. Given the prior distribution $\pi(\theta)$, the goal is to approximate the posterior distribution, $\pi(\theta|x) \propto f(x|\theta)\pi(\theta)$, where $f(x|\theta)$ is the likelihood of θ given the data x. The ABC methods have the following generic form.

1. Sample a candidate parameter vector θ^* from some proposal distribution $\pi(\theta)$.

2. Simulate a dataset x^* from the model described by a conditional probability distribution $f(x|\theta^*)$.

3. Compare the simulated dataset, x^*, with the experimental data, x_0, using a distance function, d, and tolerance ϵ; if $d(x_0, x^*) \leq \epsilon$, accept θ^*. The tolerance $\epsilon \geq 0$ is the desired level of agreement between x_0 and x^*.

The output of an ABC algorithm is a sample of parameters from a distribution $\pi(\theta|d(x_0, x^*) \leq \epsilon)$. If ϵ is sufficiently small, then the distribution $\pi(\theta|d(x_0, x^*) \leq \epsilon)$ will be a good approximation for the posterior distribution $\pi(\theta|x_0)$. It is often difficult to define a suitable distance function $d(x_0, x^*)$ between the full datasets, so one may instead replace it with a distance defined on summary statistics, $S(x_0)$ and $S(x^*)$, of the datasets. That is, the distance function may be defined as $d(x_0, x^*) = d'(S(x_0), S(x^*))$, where d' is a distance function defined on the summary statistic space. However, here, as we consider values of a dynamical process at a set of time points, we are able to compare the datasets directly without the use of summary statistics. In any case, the algorithms take the same form.

The simplest ABC algorithm is the *ABC rejection sampler* (Pritchard *et al.* 1999), which is as follows.

R1 Sample θ^* from $\pi(\theta)$.
R2 Simulate a dataset x^* from $f(x|\theta^*)$.
R3 If $d(x_0, x^*) \leq \epsilon$, accept θ^*, otherwise reject.
R4 Return to *R1*.

The disadvantage of the ABC rejection sampler is that the acceptance rate is low when the prior distribution is very different from the posterior distribution. To avoid this problem, an ABC method based on Markov chain Monte Carlo was introduced (Marjoram *et al.* 2003). The *ABC MCMC algorithm* proceeds as follows.

M1 Initialize θ_i, $i=0$.
M2 Propose θ^* according to a proposal distribution $q(\theta|\theta_i)$.
M3 Simulate a dataset x^* from $f(x|\theta^*)$.
M4 If $d(x_0, x^*) \leq \epsilon$, go to *M5*, otherwise set $\theta_{i+1} = \theta_i$ and go to *M6*.
M5 Set $\theta_{i+1} = \theta^*$ with probability

$$\alpha = \min\left(1, \frac{\pi(\theta^*)q(\theta_i|\theta^*)}{\pi(\theta_i)q(\theta^*|\theta_i)}\right)$$

and $\theta_{i+1} = \theta_i$ with probability $1 - \alpha$.
M6 Set $i = i+1$, go to *M2*.

The outcome of this algorithm is a Markov chain with the stationary distribution $\pi(\theta|d(x_0, x^*) \leq \epsilon)$ (Marjoram *et al.* 2003). That is, ABC MCMC is guaranteed to converge to the target approximate posterior distribution. Note that if the *proposal distribution* is symmetric, $q(\theta_i|\theta^*) = q(\theta^*|\theta_i)$, then α depends only on the prior distribution. Furthermore, if the prior is uniform, then $\alpha = 1$ in *M5*. Potential disadvantages of the ABC MCMC algorithm are that the correlated nature of samples coupled with the potentially low acceptance probability may result in very long chains and that the chain may get stuck in regions of low probability for long periods of time.

The above-mentioned disadvantages of ABC rejection and ABC MCMC methods can, at least in part, be avoided in ABC algorithms based on SMC methods, first developed by Sisson *et al.* (2007). In this paper, we derive ABC SMC from a sequential importance sampling (SIS) algorithm (Del Moral *et al.* 2006); see appendix A for the derivation and appendix B for a comparison with the algorithm of Sisson *et al.* (2007).

In ABC SMC, a number of sampled parameter values (called particles), $\{\theta^{(1)}, \ldots, \theta^{(N)}\}$, sampled from the prior distribution $\pi(\theta)$, are propagated through a sequence of intermediate distributions, $\pi(\theta|d(x_0, x^*) \leq \epsilon_i)$, $i = 1, \ldots, T-1$, until it represents a sample from the *target distribution* $\pi(\theta|d(x_0, x^*) \leq \epsilon_T)$. The tolerances ϵ_i are chosen such that $\epsilon_1 > \cdots > \epsilon_T \geq 0$, thus the distributions gradually evolve towards the target posterior. For sufficiently large numbers of particles, the population approach can also, in principle, avoid the problem of getting stuck in areas of low probability encountered in ABC MCMC. The ABC SMC algorithm proceeds as follows.[1]

- *S1*   Initialize $\epsilon_1, \ldots, \epsilon_T$.
  Set the population indicator $t = 0$.
- *S2.0* Set the particle indicator $i = 1$.
- *S2.1* If $t = 0$, sample θ^{**} independently from $\pi(\theta)$.
  Else, sample θ^* from the previous population $\{\theta^{(i)}_{t-1}\}$ with weights w_{t-1} and perturb the particle to obtain $\theta^{**} \sim K_t(\theta|\theta^*)$, where K_t is a perturbation kernel.
  If $\pi(\theta^{**}) = 0$, return to *S2.1*.
  Simulate a candidate dataset $x^* \sim f(x|\theta^{**})$.
  If $d(x^*, x_0) \geq \epsilon_t$, return to *S2.1*.
- *S2.2* Set $\theta^{(i)}_t = \theta^{**}$ and calculate the weight for particle $\theta^{(i)}_t$,

$$
w_t^{(i)} = \begin{cases} 1, & \text{if } t = 0, \\ \dfrac{\pi(\theta_t^{(i)})}{\sum\limits_{j=1}^{N} w_{t-1}^{(j)} K_t(\theta_{t-1}^{(j)}, \theta_t^{(i)})}, & \text{if } t > 0. \end{cases}
$$

  If $i < N$, set $i = i+1$, go to *S2.1*.
- *S3*   Normalize the weights.
  If $t < T$, set $t = t+1$, go to *S2.0*.

Particles sampled from the previous distribution are denoted by a single asterisk, and after perturbation these particles are denoted by a double asterisk. Here, we choose the perturbation kernel K_t to be a random walk (uniform or Gaussian). Note that for the special case when $T = 1$, the ABC SMC algorithm corresponds to the ABC rejection algorithm.

2.2. Model selection

Here, we introduce an ABC SMC model selection framework that employs standard concepts from Bayesian model selection, including Bayes factors (a comprehensive review of Bayesian model selection can be found in Kass & Raftery 1995). Let m_1 and

[1]For a more general version of the algorithm, suitable especially for application to stochastic models, see appendix A.

Table 1. Interpretation of the Bayes factor (adapted from Kass & Raftery 1995).

the value of the Bayes factor B_{12}	evidence against m_2 (and in favour of m_1)
1–3	very weak
3–20	positive
20–150	strong
>150	very strong

m_2 be two models; we would like to choose which model explains the data x better. The Bayes factor is defined as

$$
B_{12} = \frac{P(m_1|x)/P(m_2|x)}{P(m_1)/P(m_2)}, \tag{2.1}
$$

where $P(m_i)$ is the prior and $P(m_i|x)$ is the marginal posterior distribution of model m_i, $i = 1, 2$. If the priors are uniform, then (2.1) simplifies to

$$
B_{12} = \frac{P(m_1|x)}{P(m_2|x)}. \tag{2.2}
$$

The Bayes factor is a summary of the evidence provided by the data in favour of one statistical model over another (see table 1 for its interpretation).

There are several advantages of Bayesian model selection when compared with traditional hypothesis testing. First, the models being compared do not need to be nested. Second, the Bayes factors do not only weigh the evidence against a hypothesis (in our case m_2), but can equally well provide evidence in favour of it. This is not the case for traditional hypothesis testing where a small p-value only indicates that the null model has insufficient explanatory power. However, one cannot conclude from a large p-value that the two models are equivalent, or that the null model is superior, but only that there is not enough evidence to distinguish between the two. In other words, 'failing to reject' the null hypothesis cannot be translated into 'accepting' the null hypothesis (Cox & Hinkley 1974; Kass & Raftery 1995). Third, unlike the posterior probability of the model, the p-value does not provide any direct interpretation of the weight of evidence (the p-value is not the probability that the null hypothesis is true).

Here, we approach the model selection problem by including a 'model parameter' $m \in \{1, \ldots, M\}$, where M is the number of models, as an additional discrete parameter and denote the model-specific parameters as $\theta(m) = (\theta(m)^{(1)}, \ldots, \theta(m)^{(k_m)})$, $m = 1, \ldots, M$, where k_m denotes the number of parameters in model m.

In each population, we start by sampling a model indicator m from the prior distribution $\pi(m)$. For model m, we then propose new particles by perturbing the particles from the previous population specific to m; this step is the same as in the parameter estimation algorithm. The weights for particles $\theta(m)$ are also calculated in a similar way as in the parameter estimation algorithm for m.

The *ABC SMC algorithm for model selection* proceeds as follows.[2]

[2]In the stochastic framework, we again suggest using the general form of the algorithm with $B_t > 1$; see appendix A.

MS1   Initialize $\epsilon_1, ..., \epsilon_T$.
       Set the population indicator $t = 0$.

MS2.0 Set the particle indicator $i = 1$.

MS2.1 Sample m^* from $\pi(m)$.
       If $t = 0$, sample θ^{**} from $\pi(\theta(m^*))$.
       If $t > 0$, sample θ^* from the previous population $\{\theta(m^*)_{t-1}\}$ with weights $w(m^*)_{t-1}$.
       Perturb the particle θ^* to obtain $\theta^{**} \sim K_t(\theta|\theta^*)$.
       If $\pi(\theta^{**}) = 0$, return to *MS2.1*.
       Simulate a candidate dataset $x^* \sim f(x|\theta^{**}, m^*)$.
       If $d(x^*, x_0) \geq \epsilon_t$, return to *MS2.1*.

MS2.2 Set $m_t^{(i)} = m^*$ and add θ^{**} to the population of particles $\{\theta(m^*)_t\}$, and calculate its weight as,

$$
w_t^{(i)} = \begin{cases} 1, & \text{if } t = 0, \\[2ex] \dfrac{\pi(\theta^{**})}{\displaystyle\sum_{j=1}^{N} w_{t-1}^{(j)} K_t(\theta_{t-1}^{(j)}, \theta^{**})}, & \text{if } t > 0. \end{cases}
$$

       If $i < N$, set $i = i + 1$, go to *MS2.1*.

MS3   For every m, normalize the weights.
       If $t < T$, set $t = t + 1$, go to *MS2.0*.

The outputs of the ABC SMC algorithm are the approximations of the marginal posterior distribution of the model parameter $P(m|x)$ and the marginal posterior distributions of parameters $P(\theta_i|x, m)$, $m = 1, ..., M$, $i = 1, ..., k_m$. Note that it can happen that a model dies out (i.e. there are no particles left that belong to a particular model) if it offers only a poor description of the data. In this case, the sampling of particles continues from the remaining models only.

The Bayes factors can be obtained directly from $P(m|x)$ using equation (2.2). However, in many cases there will not be a single best and most powerful/explanatory model (Stumpf & Thorne 2006). More commonly, different models explain different parts of the data to a certain extent. One can average over these models to obtain a better inference than from a single model only. The approximation of the marginal posterior distribution of the model, $P(m|x)$, which is the output of the above algorithm, can be used for Bayesian model averaging (Hoeting *et al.* 1999).

The parameter estimation for each of the models is performed simultaneously with the model selection. The model with the highest posterior probability will typically have the greatest number of particles, thereby ensuring a good estimate of the posterior distribution of the parameters. However, some models are poorly represented in the marginal posterior distribution of m (i.e. only a small number of particles belong to these models), and so the small number of particles does not provide a very good estimate of the posterior distributions of the parameters. Therefore, one might wish to estimate parameters for these models independently.

We note that the ABC SMC model selection algorithm implicitly penalizes the models with a large number of parameters; the higher the parameter dimension is, the smaller is the probability that the perturbed particle is accepted.

2.3. Implementation of the algorithm

The algorithm is implemented in C++. For the ODE solver code, the fourth-order classical Runge–Kutta algorithm from the GNU Scientific Library (Galassi 2003) is used; for the simulation of stochastic models, we use the Gillespie algorithm (Gillespie 1977); and for the simulation of delay differential equations, we implemented the algorithm based on the adaptive step-size ODE solvers from Numerical recipes in C (Press *et al.* 1992) extended by code handling the delay part according to Paul (1992) and Enright & Hu (1995).

3. RESULTS

We demonstrate the performance of the ABC algorithms using the simulated data from deterministic and stochastic systems. The data points were obtained by solving the systems for some fixed parameters at chosen time points. The sizes of the input datasets were chosen to reflect what can typically be expected in real-world datasets in ecology, molecular systems biology and epidemiology. The first two examples highlight the computational performance of ABC SMC, the problem of inferability of dynamical models and its relationship to parameter sensitivity. The third example illustrates the use of ABC SMC for model selection, which is then further demonstrated in an application to a real dataset.

3.1. Parameter inference for the deterministic and stochastic Lotka–Volterra model

The first model is the Lotka–Volterra (LV) model (Lotka 1925; Volterra 1926) describing the interaction between prey species, x, and predator species, y, with parameter vector $\theta = (a, b)$,

$$
\frac{\mathrm{d}x}{\mathrm{d}t} = ax - xy, \tag{3.1a}
$$

$$
\frac{\mathrm{d}y}{\mathrm{d}t} = bxy - y. \tag{3.1b}
$$

3.1.1. Computational efficiency of ABC SMC applied to deterministic LV dynamics. The data $\{(x_d, y_d)\}$ are noisy observations of the simulated system with parameter values set at $(a, b) = (1, 1)$. We sample eight data points (for each of the species) from the solution of the system for parameters $(a, b) = (1, 1)$ and add Gaussian noise $\mathcal{N}(0, (0.5)^2)$ (figure 1a). Let the distance function $d((x_d, y_d), (x, y))$ between the data $\{x_d[i], y_d[i]\}$, $i = 1, ..., 8$, and a simulated solution for proposed parameters, $\{(x[i], y[i])\}$, be the sum of squared errors,

$$
d((x, y), (x_d, y_d)) = \sum_i \left((x[i] - x_d[i])^2 + (y[i] - y_d[i])^2 \right). \tag{3.2}
$$

In appendix C we show that this distance function is, in fact, related to the conventional likelihood treatment of ODEs.

The distance between our noisy data and the deterministic solution for $(a, b) = (1, 1)$ from which the data were generated is 4.23, so the lowest distance to be
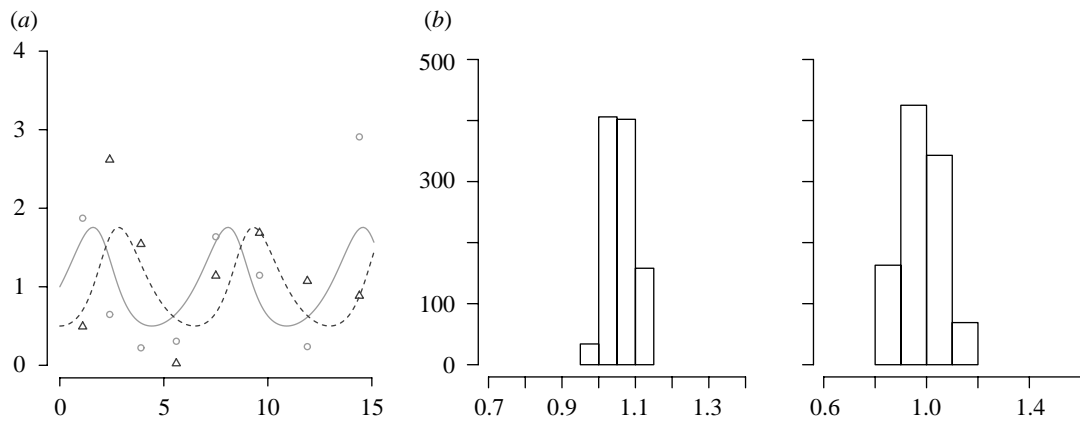
Figure 1. (*a*) Trajectories of prey (solid curve) and predator (dashed curve) populations of the deterministic LV system and the data points (circles, prey data; triangles, predator data). (*b*) Parameters inferred by the ABC rejection sampler.

Table 2. Cumulative number of data generation steps needed to accept 1000 particles in each population for deterministic LV dynamics.

population	1	2	3	4	5
data generation steps	26 228	36 667	46 989	49 271	52 194

reached is expected to be close to this number and we choose the tolerance ϵ accordingly.

First, we apply the ABC rejection sampler approach with $\epsilon = 4.3$. The prior distributions for a and b are taken to be uniform, $a, b \sim U(-10, 10)$. In order to obtain 1000 accepted particles, approximately 14.1 million data generation steps are needed, which means that the acceptance rate (7×10^{-5}) is extremely low. The inferred posterior distributions are shown in figure 1*b*.

Applying the ABC MCMC scheme outlined above yields results comparable to those of ABC rejection, and after a careful calibration of the approach (using an adaptive Gaussian proposal distribution), we manage markedly to reduce the computational cost (including burn-in, we had to generate between 40 000 and 60 000 simulations in order to obtain 1000 independent particles).

Next, we apply the ABC SMC approach. The prior distributions for a and b are taken to be uniform, $a, b \sim U(-10, 10)$, and the perturbation kernels for both parameters are uniform, $K_t = \sigma U(-1, 1)$, with $\sigma = 0.1$. The number of particles in each population is $N = 1000$. To ensure the gradual transition between populations, we take $T = 5$ populations with $\epsilon = (30.0, 16.0, 6.0, 5.0, 4.3)$. The results are summarized in table 2 and figure 2. From the last population (population 5), it can be seen that both parameters are well inferred (*a*: median = 1.05, 95% quantile range = [1.00,1.12]; *b*: median = 1.00, 95% quantile range = [0.87,1.11]). The outcome is virtually the same as previously obtained by the ABC rejection sampler (figure 1*b*); however, there is a substantial reduction in the number of steps needed to reach this result. For this model, the ABC SMC algorithm needs 50 times fewer data generation steps than the ABC rejection sampler, and about the same

number of data generation steps as the adaptive ABC MCMC algorithm.

The analyses were repeated with different distance functions, such as

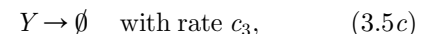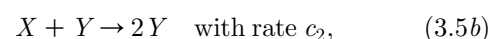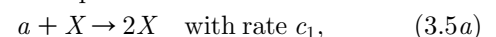$$d((x, y), (x_d, y_d)) = \sum_i (|x[i] - x_d[i]| + |y[i] - y_d[i]|)$$

$$(3.3)$$

and

$$d((x, y), (x_d, y_d)) = 2 - \frac{x \cdot x_d}{\|x\| \|x_d\|} - \frac{y \cdot y_d}{\|y\| \|y_d\|}, \quad (3.4)$$

where the dot denotes the inner product. As expected, the resulting approximations of posterior distributions are very similar (histograms not shown). Replacing the uniform perturbation kernel by a Gaussian kernel also yields the same results, but requires more simulation steps (results not shown).

3.1.2. ABC SMC inference for stochastic LV dynamics. Having obtained good estimates for the deterministic case, next, we try to infer parameters of a stochastic LV model. The predator–prey process can be described by the following rate equations:

$$a + X \rightarrow 2X \quad \text{with rate } c_1, \quad (3.5a)$$

$$X + Y \rightarrow 2Y \quad \text{with rate } c_2, \quad (3.5b)$$

$$Y \rightarrow \emptyset \quad \text{with rate } c_3, \quad (3.5c)$$

where X denotes the prey population; Y is the predator population; and a is the fixed amount of resource available to the prey (we fix it to $a = 1$). These reactions define the stochastic master equation (van Kampen 2007)

$$\frac{\partial P(x, y, t)}{\partial t} = c_1 a(x - 1) P(x - 1, y, t)$$
$$+ c_2(x + 1)(y - 1) P(x + 1, y - 1, t)$$
$$+ c_3(y + 1) P(x, y + 1, t)$$
$$- (c_1 a x + c_2 x y + c_3 y) P(x, y, t). \quad (3.6)$$

The ABC approach can easily be applied to inference problems involving master equations, because there exists an exact simulation algorithm (Gillespie algorithm; Gillespie 1977; Wilkinson 2006).

Our simulated data consist of 19 data points for each species with rates $(c_1, c_2, c_3) = (10, 0.01, 10)$ and initial
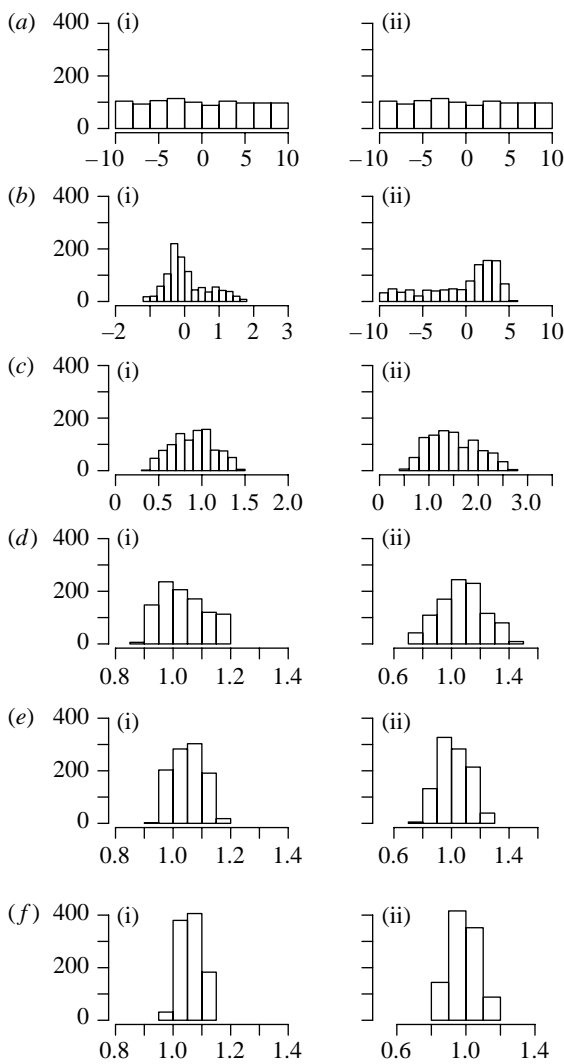
Figure 2. Histograms of populations (*a*) 0 (uniform prior distribution), (*b*) 1, (*c*) 2, (*d*) 3, (*e*) 4 and (*f*) 5 (approximation of posterior distribution) of parameters (i) *a* and (ii) *b* of the LV system.

conditions $(X_0, Y_0) = (1000, 1000)$. The distance function is the square root of the sum of squared errors (3.2) and the SMC algorithm is run for $T = 5$ populations with $\epsilon = (4000, 2900, 2000, 1900, 1800)$. Especially in laboratory settings, the results from several replicate experiments are averaged over; here, we therefore also use data averaged over three independent replicates. The simulated data at every run are then, just as the experimental data, averaged over three runs. For inference purposes, the average over several runs tends to hold more information about the system's mean dynamics than a single stochastic run. If the experimental data consisted of one run only, i.e. if there were no repetitions, then the inference could in principle proceed in the same way, by comparing the experimental data with a single simulated run. This would result in a lower acceptance rate and, consequently, more data generation steps to complete the inference.

We generate $N = 100$ particles per population and assign the prior distributions of the parameters to be $\pi(c_1) = U(0, 28)$, $\pi(c_2) = U(0.0, 0.04)$ and $\pi(c_3) = U(0, 28)$, reflecting the respective orders of magnitude of the simulated parameters (if a larger domain for $\pi(c_2)$
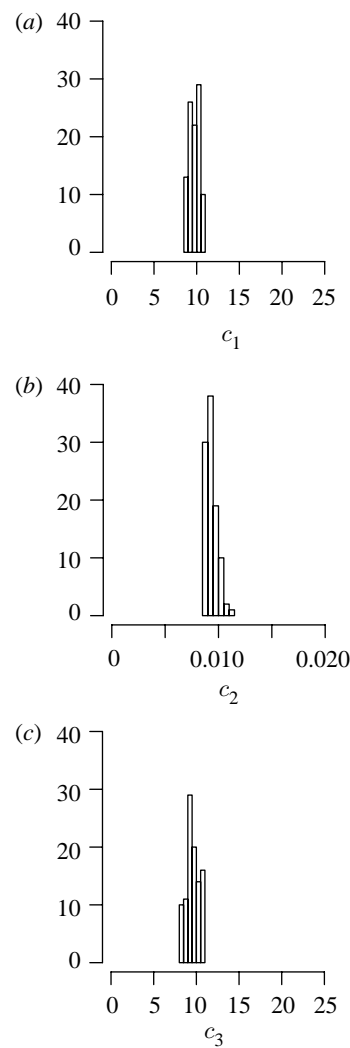

Figure 3. Histograms of the approximated posterior distributions of parameters (*a*) c_1, (*b*) c_2 and (*c*) c_3 of the stochastic LV dynamics.

is taken, there are no accepted instances for $c_2 > 0.04$). Perturbation kernels are uniform with $\sigma_{c_1} = \sigma_{c_3} = 1.0$ and $\sigma_{c_2} = 0.0025$, and $B_t = 10$. The results are summarized in figure 3.

3.2. Parameter inference for the deterministic and stochastic repressilator model

The repressilator (Elowitz & Leibler 2000) is a popular toy model for gene regulatory systems and consists of three genes connected in a feedback loop, where each gene transcribes the repressor protein for the next gene in the loop. The model consists of six ordinary differential equations and four parameters, which are as follows:

$$\frac{dm_1}{dt} = -m_1 + \frac{\alpha}{1 + p_3^n} + \alpha_0, \qquad (3.7a)$$

$$\frac{dp_1}{dt} = -\beta(p_1 - m_1), \qquad (3.7b)$$

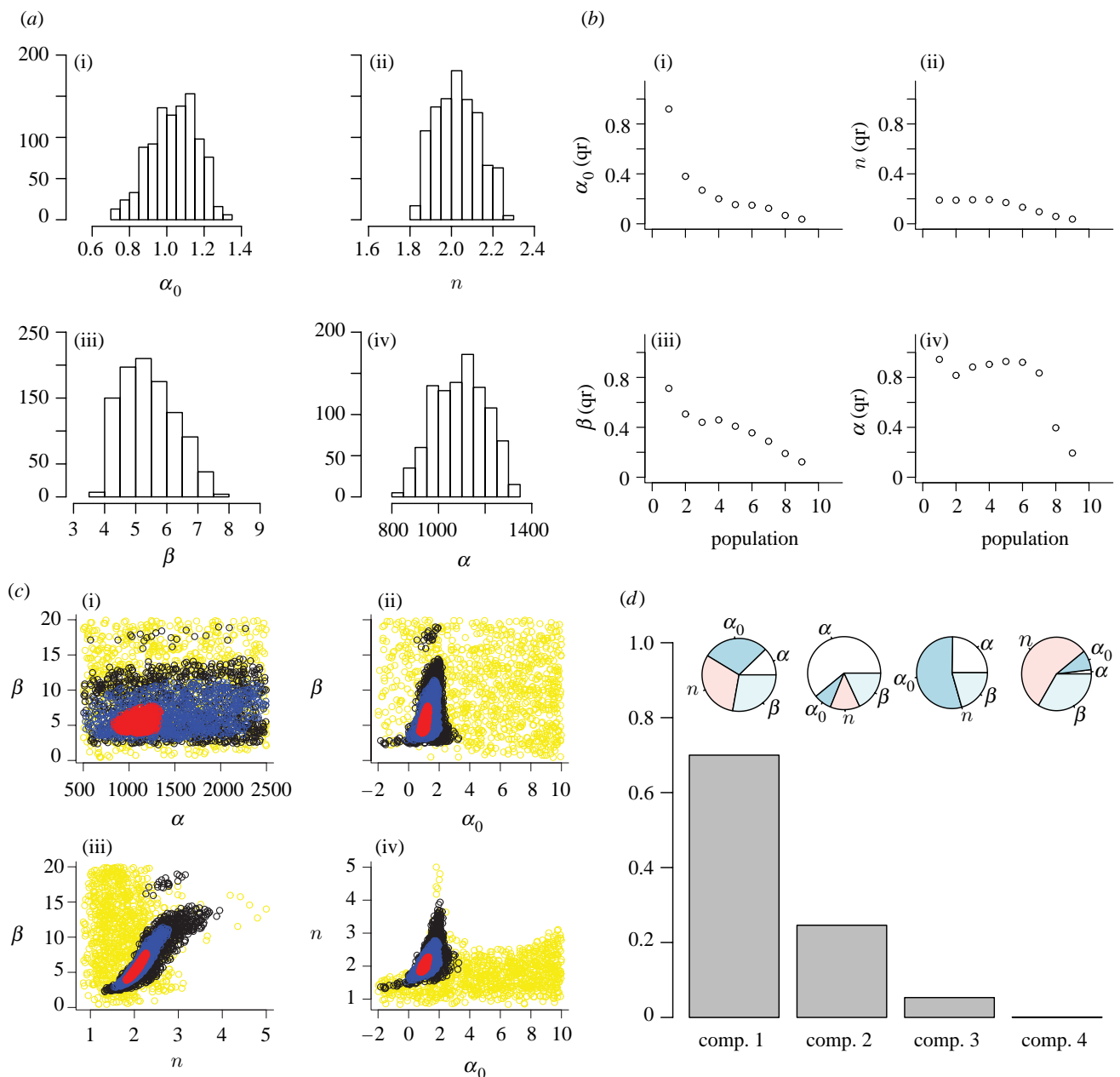$$\frac{dm_2}{dt} = -m_2 + \frac{\alpha}{1 + p_1^n} + \alpha_0, \qquad (3.7c)$$

Figure 4. (*a*) Histograms of the approximate marginal posterior distributions of parameters α_0, n, β and α of the deterministic repressilator model. (*b*) The normalized 95% interquantile ranges (qr) of each population. The narrower the interval for a given tolerance ϵ_t, the more sensitive the model is to the corresponding parameter. The interquantile range reached in population 9 is determined by the added experimental noise. As ϵ_9 was chosen accordingly, one cannot proceed by lowering the tolerance further. The sharp change in the interquantile ranges, which occurs, for example, for parameter α_0 between populations 1 and 2, can be explained by the steep gradient of the likelihood surface along α_0. (*c*) The output (i.e. the accepted particles) of the ABC SMC algorithm as two-dimensional scatterplots. The particles from population 1 are in yellow, particles from population 4 in black, particles from population 7 in blue and those from the last population in red. Islands of particles are observed in population 4 and they can be explained by the multimodality of the fourth intermediate distribution. (*d*) PCA of the set of accepted particles (population 9). Owing to the dependence of the PCA on the scaling of original variables, the PCA was performed on the correlation matrix. The first PC explains 70.0% of the total variance, the second 24.6%, the third 5.3% and the fourth 0.1% of the variance. Pie charts show the fraction of the length of PCs explained by individual parameters.

$$\frac{\mathrm{d}p_2}{\mathrm{d}t} = -\beta(p_2 - m_2), \qquad (3.7d)$$

$$\frac{\mathrm{d}m_3}{\mathrm{d}t} = -m_3 + \frac{\alpha}{1 + p_2^n} + \alpha_0, \qquad (3.7e)$$

$$\frac{\mathrm{d}p_3}{\mathrm{d}t} = -\beta(p_3 - m_3). \qquad (3.7f)$$

3.2.1. Inferability and sensitivity in deterministic repressilator dynamics. Let $\theta = (\alpha_0, n, \beta, \alpha)$ be the parameter vector. For the simulated data, the initial conditions are $(m_1, p_1, m_2, p_2, m_3, p_3) = (0, 2, 0, 1, 0, 3)$ and the values of parameters are $\theta = (1, 2, 5, 1000)$; for these parameters the repressilator displays limit-cycle behaviour. We assume that only the mRNA (m_1, m_2, m_3) measurements are available and the protein concentrations are

considered as missing data. Gaussian noise $\mathcal{N}(0, 5^2)$ is added to the data points. The distance function is defined to be the square root of the sum of squared errors. The prior parameter distributions are chosen as follows: $\pi(\alpha_0) = U(-2, 10)$; $\pi(n) = U(0, 10)$; $\pi(\beta) = U(-5, 20)$; and $\pi(\alpha) = U(500, 2500)$. We assume that the initial conditions are known.

The results are summarized in figure 4*a–d*, where we show the approximate posterior distributions, the changes in 95% interquantile ranges of parameter estimates across the populations and scatterplots of some of the two-dimensional parameter combinations. Each parameter is easily inferred when the other three parameters are fixed (histograms not shown). When the algorithm is applied to all four parameters simultaneously, parameter n is inferred the quickest and has the smallest posterior variance, while parameter α is barely inferable and has large credible intervals (figure 4*a,c*).

We find that ABC SMC recovers the intricate link between model sensitivity to parameter changes and inferability of parameters. The repressilator system is most sensitive to changes in parameter n and least sensitive to changes in α. Hence, the data appear to contain little information about α. Thus, ABC SMC provides us with a global parameter sensitivity analysis (Sanchez & Blower 1997) on the fly as the intermediate distributions are being constructed. Note that the intermediate distributions are nested in one another (as should be the case in SMC; figure 4*c*). An attempt to visualize the four-dimensional posterior distributions can be accessed in the electronic supplementary material where we provide an animation in which the posterior distribution in the four-dimensional parameter space is projected onto two-dimensional planes.

The interquantile ranges and the scatterplots provide an initial impression of parameter sensitivity; however, the first problem with scatterplots is that it is increasingly difficult to visualize the behaviour of the model with increasing parameter dimension. Second, we have to determine the sensitivity when a combination of parameters is varied (and not just individual parameters), and this cannot be visualized via simple one-dimensional interquantile ranges or two-dimensional scatterplots.

We can use principal component analysis (PCA) to quantify the sensitivity of the system (Saltelli *et al.* 2008). The output of the ABC SMC algorithm, which we are going to use for our sensitivity analysis, is the last population of N particles. Associated with the accepted particles is their variance–covariance matrix, Σ, of rank p, where p denotes the dimension of the parameter vector. The principal components (PCs) are the eigenvectors of Σ, which define a set of eigenparameters, $\chi_i = a_{i1}\theta_1 + \cdots + a_{ip}\theta_p$. Here, $a_i = (a_{i1}, \ldots, a_{ip})$ is the normalized eigenvector associated with the ith eigenvalue of Σ, λ_i, and a_{ij} describes the projection of parameter θ_j onto the ith eigenparameter. The PCA provides an orthogonal transformation of the original parameter space and the PCs can be taken to define a p-dimensional ellipsoid that approximates the population of data points (i.e. the accepted particles), and the eigenvalues specify the p corresponding radii. The variance of the ith PC is given by λ_i and the total variance of all PCs equals $\sum_{i=1}^{p} \lambda_i = \text{trace}(\Sigma)$. Therefore, the eigenvalue λ_i

associated with the ith PC explains a proportion
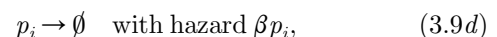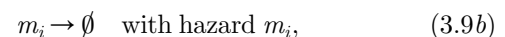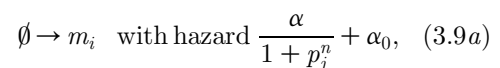
$$\frac{\lambda_i}{\text{trace}(\Sigma)} \quad (3.8)$$

of the variation in the population of points. The smaller the λ_i, the more sensitive the system is to the variation of the eigenparameter χ_i. The PCA yields only an approximate account of sensitivity similar to what would be obtained by computing the Hessian around the mode of the posterior distribution.

Figure 4*d* summarizes the output of the PCA. It shows how much of the variance is explained by each PC, and which parameters contribute the most to these PCs. In contrast to the interest in the first PC in most PCA applications, our main interest lies in the smallest PC. The last PC extends across the narrowest region of the posterior parameter distribution, and therefore provides information on parameters to which the model is the most sensitive. In other words, the smaller PCs correspond to stiff parameter combinations, while the larger PCs may correspond to sloppy parameter combinations (Gutenkunst *et al.* 2007).

The analysis reveals that the last PC mainly extends in the direction of a linear combination of parameters n and β, from which we can conclude that the model is most sensitive to changes in these two parameters. Looking at the third component, the model is somewhat less sensitive to variation in α_0. The model is therefore the least sensitive to changes in parameter α, which is also supported by the composition of the second PC. This outcome agrees with the information obtained from the interquantile ranges and the scatterplots.

3.2.2. Inference of the stochastic repressilator dynamics.
Next, we apply ABC SMC to the stochastic repressilator model. We transformed the deterministic model (3.7) into a set of the following reactions:

$$\emptyset \rightarrow m_i \quad \text{with hazard } \frac{\alpha}{1 + p_j^n} + \alpha_0, \quad (3.9a)$$

$$m_i \rightarrow \emptyset \quad \text{with hazard } m_i, \quad (3.9b)$$

$$m_i \rightarrow m_i + p_i \quad \text{with hazard } \beta m_i, \quad (3.9c)$$

$$p_i \rightarrow \emptyset \quad \text{with hazard } \beta p_i, \quad (3.9d)$$

where $i = 1, 2, 3$ and, correspondingly, $j = 3, 1, 2$. The stochastic process defined by these reactions can be simulated with the Gillespie algorithm. True parameters and initial conditions correspond to those of the deterministic case discussed previously. The data include both mRNA and protein levels at 19 time points. Tolerances are chosen as $\epsilon = \{900, 650, 500, 450, 400\}$, the number of particles $N = 200$ and $B_t = 5$. The prior distributions are chosen as follows: $\pi(\alpha_0) = U(0, 10)$; $\pi(n) = U(0, 10)$; $\pi(\beta) = U(-5, 20)$; and $\pi(\alpha) = U(500, 2500)$.

The inference results for comparing the average over 20 simulations with the data generated from the average of 20 simulations are summarized in figure 5*a,b*. Figure 5*a,b* shows that parameters n and β get reasonably well inferred, while α_0 and α are harder to

infer. It is clearly noticeable that parameters are better inferred in the deterministic case (figure 4a).

3.2.3. Contrasting inferability for the deterministic and stochastic dynamics. Analysing and comparing the results of the deterministic and stochastic repressilator dynamics shows that parameter sensitivity is intimately linked to inferability. If the system is insensitive to a parameter, then this parameter will be hard (or even impossible) to infer, as varying such a parameter does not vary the output—which here is the approximate posterior probability—very much. In stochastic problems, we may furthermore have the scenario where the fluctuations due to small variations in one parameter overwhelm the signals from other parameters.

3.3. Model selection on different SIR models

We illustrate model selection using a range of simple models that can describe the epidemiology of infectious diseases. SIR models describe the spread of such disease in a population of susceptible (S), infected (I) and recovered (R) individuals (Anderson & May 1991). The simplest model assumes that every individual can be infected only once and that there is no time delay between the individual getting infected and their ability to infect other susceptible individuals,

$$\dot{S} = \alpha - \gamma SI - dS, \qquad (3.10a)$$

$$\dot{I} = \gamma SI - vI - dI, \qquad (3.10b)$$

$$\dot{R} = vI - dR, \qquad (3.10c)$$

where $\dot{x}$ denotes the time derivative of x, dx/dt. Individuals, who are born at rate α, are susceptible; d is the death rate (irrespective of the disease class, S, I or R); γ is the infection rate; and v is the recovery rate.

The model can be made more realistic by adding a time delay τ between the time an individual gets infected and the time when they become infectious,

$$\dot{S} = \alpha - \gamma SI(t - \tau) - dS, \qquad (3.11a)$$

$$\dot{I} = \gamma SI(t - \tau) - vI - dI, \qquad (3.11b)$$

$$\dot{R} = vI - dR. \qquad (3.11c)$$

Another way of incorporating the time delay into the model is by including a population of individuals in a latent (L) phase of infection; in this state, they are infected but cannot yet infect others. The equations then become

$$\dot{S} = \alpha - \gamma SI - dS, \qquad (3.12a)$$

$$\dot{L} = \gamma SI - \delta L - dL, \qquad (3.12b)$$

$$\dot{I} = \delta L - vI - dI, \qquad (3.12c)$$

$$\dot{R} = vI - dR. \qquad (3.12d)$$

Here, δ denotes the transition rate from the latent to the infective stage.

Another extension of the basic model (3.10) allows the recovered individuals to become susceptible again

(with rate e),

$$\dot{S} = \alpha - \gamma SI - dS + eR, \qquad (3.13a)$$

$$\dot{I} = \gamma SI - vI - dI, \qquad (3.13b)$$

$$\dot{R} = vI - (d + e)R. \qquad (3.13c)$$

There are obviously many more ways of extending the basic model, but here we restrict ourselves to the four models described above. Given the same initial conditions, the outputs of all models are very similar, which makes it impossible to choose the right model by visual inspection of the data alone. Therefore, some more sophisticated, statistically based methods need to be applied for selecting the best available model. Therefore, we apply the ABC SMC algorithm for model selection, developed in §2.2. We define a model parameter $m \in \{1, 2, 3, 4\}$, representing the above models in the same order, and model-specific parameter vectors $\theta(m)$: $\theta(1) = (\alpha, \gamma, d, v)$; $\theta(2) = (\alpha, \gamma, d, v, \tau)$; $\theta(3) = (\alpha, \gamma, d, v, \delta)$; and $\theta(4) = (\alpha, \gamma, d, v, e)$.

The experimental data consist of 12 data points from each of the three groups (S, I and R). If the data are very noisy (Gaussian noise with standard deviation $\sigma = 1$ was added to the simulated data points), then the algorithm cannot detect a single best model, which is not surprising given the high similarity of model outputs. However, if intermediate noise is added (Gaussian noise with standard deviation $\sigma = 0.2$), then the algorithm produces a posterior estimate with the most weight on the correct model. An example is shown in figure 6, where the experimental data were obtained from model 1 and perturbed by Gaussian noise, $\mathcal{N}(0, (0.2)^2)$. Parameter inference is performed simultaneously with the model selection (posterior parameter distributions not shown).

The Bayes factor can be calculated from the marginal posterior distribution of m, which we take from the final population. From 1000 particles, model 1 (basic model) was selected 664 times, model 2 230 times, model 4 106 times and model 3 was not selected at all in the final population. Therefore, we can conclude from the Bayes factors

$$B_{1,2} = \frac{664}{230} = 2.9, \qquad (3.14)$$

$$B_{1,4} = \frac{664}{106} = 6.3, \qquad (3.15)$$

$$B_{2,4} = \frac{230}{106} = 2.2, \qquad (3.16)$$

that there is weak evidence in favour of model 1 when compared with model 2 and positive evidence in favour of model 1 when compared with model 4. Increasing the amount of data will, however, change the Bayes factors in favour of the true model.

3.4. Application: common-cold outbreaks on Tristan da Cunha

Tristan da Cunha is an isolated island in the Atlantic Ocean with approximately 300 inhabitants, and it was observed that viral diseases, such as common cold, break
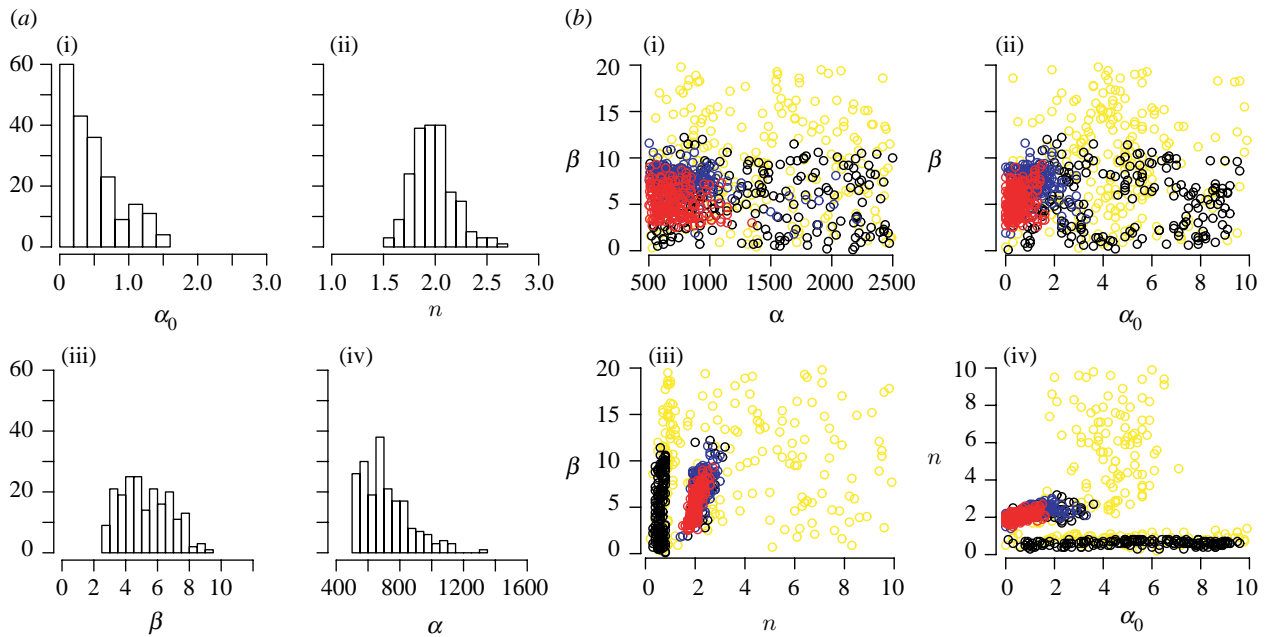
Figure 5. (*a*) Histograms of the approximated posterior distributions of parameters α_0, n, β and α of the stochastic repressilator model. (*b*) The output of the ABC SMC algorithm as two-dimensional scatterplots. The particles from population 1 are in yellow, particles from population 2 in black, particles from population 3 in blue and those from the last population in red. We note that the projection to parameter n in population 2 is a sample from a multimodal intermediate distribution, while this distribution becomes unimodal from population 3 onwards.

out on the island after the arrival of ships from Cape Town. We use the 21-day common-cold data from October 1967. The data, shown in table 3, were obtained from Hammond & Tyrrell (1971) and Shibli *et al.* (1971). The data provide only the numbers of infected and recovered individuals, $I(t)$ and $R(t)$, whereas the size of the initial susceptible population $S(0)$ is not known. Therefore, $S(0)$ is an extra unknown parameter to be estimated.

The four epidemic models from §3.3 are used and because there are no births and deaths expected in the short period of 21 days, parameters α and d are set to 0. The tolerances are set to $\epsilon = \{100, 90, 80, 73, 70, 60, 50, 40, 30, 25, 20, 16, 15, 14, 13.8\}$ and 1000 particles are used. The prior distributions of parameters are chosen as follows: $\gamma \sim U(0, 3)$; $v \sim U(0, 3)$; $\tau \sim U(-0.5, 5)$; $\delta \sim U(-0.5, 5)$; $e \sim U(-0.5, 5)$; $S(0) \sim U(37, 100)$; and $m \sim U(1, 4)$, where $S(0)$ and m are discrete. Perturbation kernels are uniform, $K_t = \sigma U(-1, 1)$, with $\sigma_\gamma = \sigma_v = 0.3$, $\sigma_\tau = \sigma_\delta = \sigma_e = 1.0$ and $\sigma_{S(0)} = 3$.

The target and intermediate distributions of model parameters are shown in figure 7. The model selection algorithm chooses model (3.12), i.e. the model with a latent class of disease carriers, to be the most suitable one for describing the data; however, it is only marginally better than models (3.10) and (3.11). Therefore, to draw reliable conclusions from the inferred parameters, one might wish to use model averaging over models (3.10)–(3.12). The marginal posterior distributions for the parameters of model (3.12) are shown in figure 8. However, the estimated initial susceptible population size $S(0)$ is low compared with the whole population of the island, which suggests that either the majority of islanders were immune to a given strand of cold or (perhaps more plausibly) the system is not well
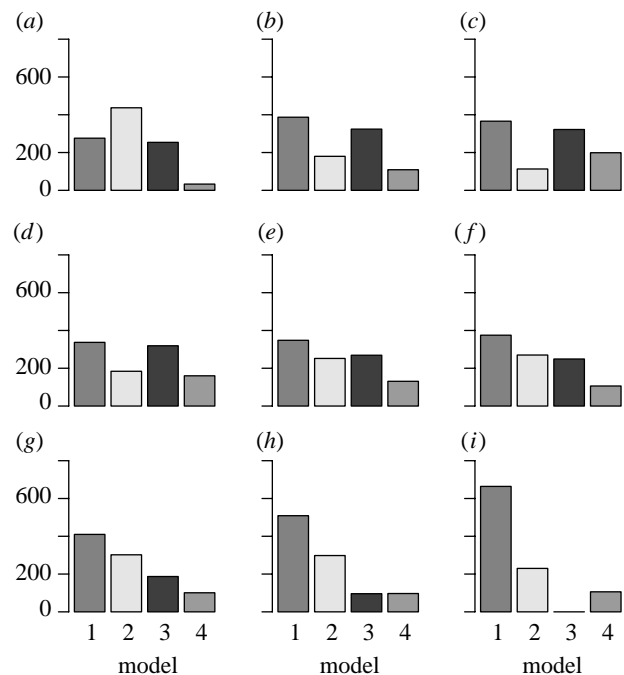

Figure 6. Histograms show populations (*a*−*i*) 1–9 of the model parameter m. Population 9 represents the final posterior marginal estimate of m. Population 0 (discrete uniform prior distribution) is not shown.

represented by our general epidemic models with homogeneous mixing. The estimated durations of the latent period τ (from model (3.11)) and $1/\delta$ (from model (3.12)), however, are broadly in line with the established aetiology of common cold (Fields *et al.* 1996). Thus, within the set of candidate epidemiological models, the ABC SMC approach selects the most plausible model and results in realistic parameter estimates.

Table 3. Common-cold data from Tristan da Cunha collected in October 1967.

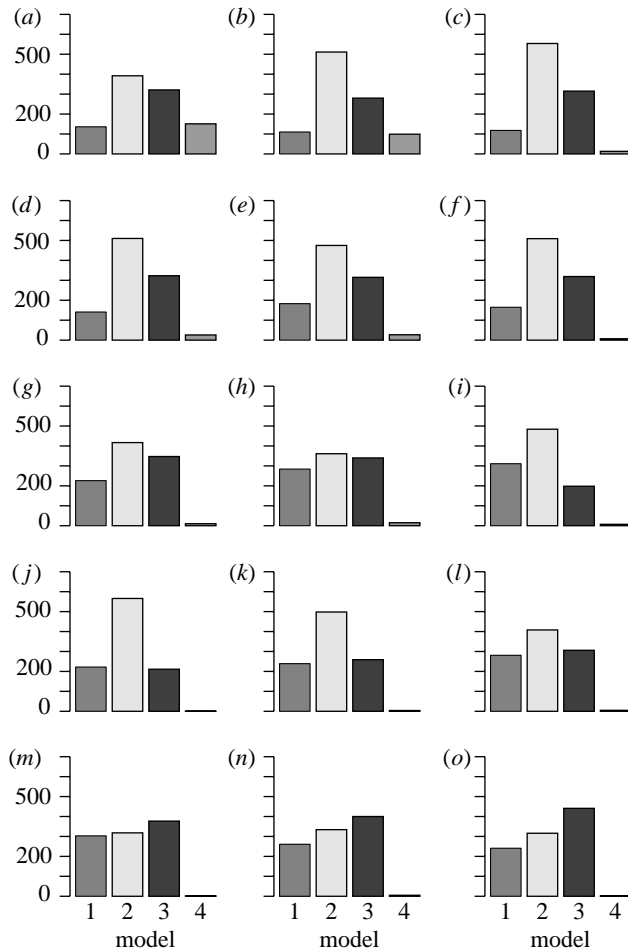| day	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
$I(t)$	1	1	3	7	6	10	13	13	14	14	17	10	6	6	4	3	1	1	1	1	0
$R(t)$	0	0	0	0	5	7	8	13	13	16	16	24	30	31	33	34	36	36	36	36	37

Figure 7. Populations of the marginal posterior distribution of *m*. Models 1–4 correspond to equations (3.10)–(3.13), respectively. An interesting phenomenon is observed in populations 1–12, where model 2 has the highest probability, in contrast to model 3 having the highest inferred probability in the last population. The most probable explanation for this is that a local maximum favouring model 2 has been passed on route to a global maximum of the posterior probability favouring model 3. Populations (*a–o*) 1–15. Population 0 (discrete uniform prior distribution) is not shown.

4. DISCUSSION

Our study suggests that ABC SMC is a promising tool for reliable parameter inference and model selection for models of dynamical systems that can be efficiently simulated. Owing to its simplicity and generality, ABC SMC, unlike most other approaches, can be applied without any change in both deterministic and stochastic contexts (including models with time delay).

The advantage of Bayesian statistical inference, in contrast to most conventional optimization algorithms (Moles *et al.* 2003), is that the whole probability distribution of the parameter can be obtained, rather than merely the point estimates for the optimal parameter values. Moreover, in the context of hypothesis testing, the Bayesian perspective (Cox & Hinkley 1974; Robert & Casella 2004) has a more intuitive meaning than the corresponding frequentist point of view. ABC methods share these characteristics.

Another advantage of our ABC SMC approach is that observing the shape of intermediate and posterior distributions gives (without any further computational cost) information about the sensitivity of the model to different parameters and about the inferability of parameters. All simulations are already part of the parameter estimation, and can be conveniently re-used for the sensitivity analysis via scatterplots or via the analysis of the posterior distribution using, for example, PCA. It can be concluded that the model is sensitive to parameters that are inferred quickly (in earlier populations) and that have narrow credible intervals, while it is less sensitive to those that get inferred in later populations and are not very localized by the posterior distribution. If the distribution does not change much between populations and resembles the uniform distribution from population 0, then it can be concluded that the corresponding parameter is not inferable given the available data.

While the parameter estimation for individual models is straightforward when a suitable number of particles are used, more care should be taken in model selection problems; the domains of the uniform prior distributions should be chosen with care and acceptance rates should be closely monitored. These measures should prevent models being rejected in early populations solely due to inappropriately chosen (e.g. too large) prior domains. Apart from a potentially strong dependence on the chosen prior distributions (which is also inherent in the standard Bayesian model selection; Kass & Raftery 1995), we also observe the dependency of the Bayes factors on changes in the tolerance levels ϵ_t and perturbation kernel variances. Therefore, care needs to be taken when applying the ABC SMC model selection algorithm.

Finally, we want to stress the importance of monitoring convergence in ABC SMC. There are several ways to see whether a good posterior distribution has been obtained: we can use interquantile ranges or tests of goodness of fit between successive intermediate distributions. A further crucial signature is the number of proposals required to obtain a specified number of accepted particles. This will also impose a practical limit on the procedure.

For the problems in this paper, the algorithm was efficient enough. Examples here were chosen to highlight different aspects of ABC SMC's performance and usability. However, for the use on larger systems, the algorithm can be made more computationally efficient by optimizing the number of populations, the distance
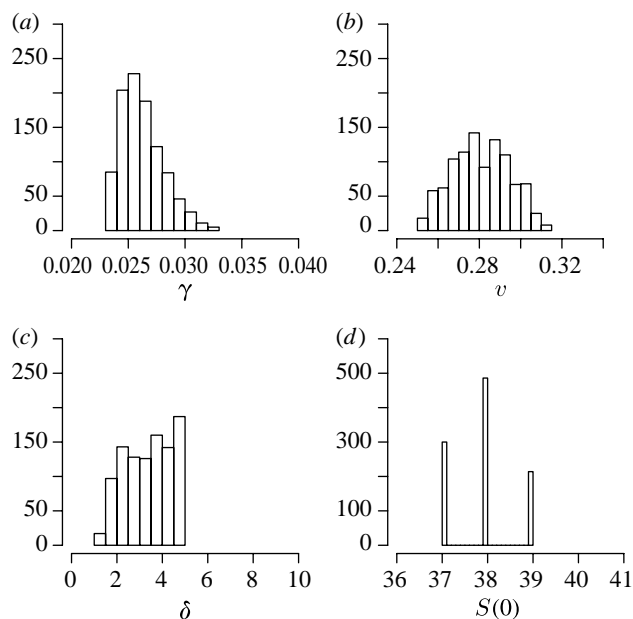
Figure 8. Histograms of the approximated posterior distributions of parameters (*a*) γ, (*b*) v, (*c*) δ and (*d*) $S(0)$ of the SIR model with a latent phase of infection (3.12).

function, the number of particles and perturbation kernels (e.g. adaptive kernels). Moreover, the algorithm is easily parallelized.

5. CONCLUSION

We have developed a SMC ABC algorithm that can be used to estimate model parameters (including their credible intervals) and to distinguish among a set of competing models. This approach can be applied to deterministic and stochastic systems in the physical, chemical and biological sciences, for example biochemical reaction networks or signalling networks. Owing to the link between sensitivity and inferability, ABC SMC can, however, also be applied to larger systems: critical parameters will be identified quickly, while the system is found to be relatively insensitive to parameters that are hard to infer.

APPENDIX A. DERIVATION OF ABC SMC

Here we derive the ABC SMC algorithm from the SIS algorithm of Del Moral *et al.* (2006) and in appendix B we show why this improves on the ABC partial rejection control (PRC) algorithm developed by Sisson *et al.* (2007). We start by briefly explaining the basics of importance sampling and then present the SIS algorithm.

Let π be the distribution we want to sample from, our *target distribution.* If it is impossible to sample from π directly, one can sample from a suitable *proposal distribution*, η, and use importance sampling weights to approximate π. The Monte Carlo approximation of η is

$$\hat{\eta}_N(x) = \frac{1}{N} \sum_{i=1}^{N} \delta_{X^{(i)}}(x),$$

where $X^{(i)} \sim^{iid} \eta$ and $\delta_{x_0}(x)$ is the Dirac delta function, defined by

$$\int_x f(x)\delta_{x_0}(x)\mathrm{d}x = f(x_0).$$

If we assume $\pi(x) > 0 \Rightarrow \eta(x) > 0$, then the target distribution π can be approximated by

$$\hat{\pi}_N(x) = \frac{1}{N} \sum_{i=1}^{N} w(X^{(i)})\delta_{X^{(i)}}(x),$$

with *importance weights* defined as

$$w(x) = \frac{\pi(x)}{\eta(x)}.$$

In other words, to obtain a sample from the target distribution π, one can instead sample from the proposal distribution, η, and weight the samples by importance weights, w.

In SIS, one reaches the target distribution π_T through a series of intermediate distributions, π_t, $t = 1, \ldots, T-1$. If it is hard to sample from these distributions, one can use the idea of importance sampling described above to sample from a series of proposal distributions η_t and weight the obtained samples by importance weights

$$w_t(x_t) = \frac{\pi_t(x_t)}{\eta_t(x_t)}. \tag{A 1}$$

In SIS, the proposal distributions are defined as

$$\eta_t(x_t) = \int \eta_{t-1}(x_{t-1})\kappa_t(x_{t-1}, x_t)\mathrm{d}x_{t-1}, \tag{A 2}$$

where η_{t-1} is the previous proposal distribution and κ_t is a Markov kernel.

In summary, we can write the SIS algorithm as follows.

Initialization Set $t = 1$.
    For $i = 1, \ldots, N$, draw $X_1^{(i)} \sim \eta_1$.
    Evaluate $w_1(X_1^{(i)})$ using (A 1) and normalize.
Iteration Set $t = t+1$, if $t = T+1$, stop.
    For $i = 1, \ldots, N$, draw $X_t \sim \kappa_t(X_{t-1}^{(i)}, \cdot)$.
    Evaluate $w_t(X_t^{(i)})$ using (A 1) with $\eta_t(x_t)$ from (A 2) and normalize.

To apply SIS, one needs to define the intermediate and the proposal distributions. Taking this SIS framework as a base, we now define ABC SMC to be a special case of the SIS algorithm, where we choose the intermediate and proposal distributions in an ABC fashion, as follows. The intermediate distributions are defined as

$$\pi_t(x) = \frac{\pi(x)}{B_t} \sum_{b=1}^{B_t} \mathbb{1}\big(d(D_0, D_{(b)}(x)) \le \epsilon_t\big),$$

where $\pi(x)$ denotes the prior distribution; $D_{(1)}, ..., D_{(B_t)}$ are B_t datasets generated for a fixed parameter, x, $D_{(b)} \sim p(D|x)$; $\mathbb{1}(x)$ is an indicator function; and ϵ_t is the tolerance required from particles contributing to the intermediate distribution π_t. This allows us to define $b_t(x) = \sum_{b=1}^{B_t} \mathbb{1}(d(D_0, D_{(b)}(x)) \leq \epsilon_t)$. We define the first proposal distribution to equal the prior distribution, $\eta_1 = \pi$. The proposal distribution at time t ($t=2, ..., T$), η_t, is defined as the perturbed intermediate distribution at time $t-1$, π_{t-1}, such that for every sample, x, from this distribution we have

(i) $b_t(x) > 0$ (in other words, particle x is accepted at least one out of B_t times), and

(ii) $\pi(x) > 0$ (in order for a condition $\pi_t(x) > 0 \Rightarrow \eta_t(x) > 0$ to be satisfied),

$$\begin{aligned}
\eta_t(x_t) &= \mathbb{1}(\pi_t(x_t) > 0)\mathbb{1}(b_t(x_t) > 0) \\
&\quad \times \int \pi_{t-1}(x_{t-1})K_t(x_{t-1}, x_t)\mathrm{d}x_{t-1} \\
&= \mathbb{1}(\pi_t(x_t) > 0)\mathbb{1}(b_t(x_t) > 0) \\
&\quad \times \int \eta_{t-1}(x_{t-1})w_{t-1}(x_{t-1})K_t(x_{t-1}, x_t)\mathrm{d}x_{t-1},
\end{aligned}$$

(A 3)

where K_t denotes the perturbation kernel (random walk around the particle).

To calculate the weights defined by

$$w_t(x) = \frac{\pi_t(x_t)}{\eta_t(x_t)}, \qquad (A\,4)$$

we need to find an appropriate way of evaluating $\eta_t(x_t)$ defined in equation (A 3). This can be achieved through the standard Monte Carlo approximation,

$$\eta_t(x_t) = \mathbb{1}(\pi_t(x_t) > 0)\mathbb{1}(b_t(x_t) > 0)$$

$$\times \int \pi_{t-1}(x_{t-1})K_t(x_{t-1}, x_t)\mathrm{d}x_{t-1}$$

$$\approx \mathbb{1}(\pi_t(x_t) > 0)\mathbb{1}(b_t(x_t) > 0)\frac{1}{N}\sum_{x_{t-1}^{(i)} \sim \pi_{t-1}} K_t\left(x_{t-1}^{(i)}, x_t\right),$$

where N denotes the number of particles and $\{x_{t-1}^{(i)}\}$, $i=1, ..., N$, are all the particles from the intermediate distribution π_{t-1}. The unnormalized weights (A 4) can then be calculated as

$$w_t(x_t) = \frac{\pi(x_t)b_t(x_t)}{\sum_{x_{t-1}^{(i)} \sim \pi_{t-1}} K_t(x_{t-1}^{(i)}, x_t)},$$

where π is the prior distribution; $b_t(x_t) = \sum_{b=1}^{B_t} \mathbb{1}(d(D_0, D_{(b)}(x_t)) \leq \epsilon_t)$; and $D_{(b)}$, $b=1, ..., B_t$, are the B_t simulated datasets for a given parameter x_t. For $B_t=1$, the weights become

$$w_t(x_t) = \frac{\pi(x_t)}{\sum_{x_{t-1}^{(i)} \sim \pi_{t-1}} K_t\left(x_{t-1}^{(i)}, x_t\right)},$$

for all *accepted* particles x_t.

The ABC SMC algorithm can be written as follows.

S1 Initialize $\epsilon_1, ..., \epsilon_T$.
   Set the population indicator $t=0$.

S2.0 Set the particle indicator $i=1$.

S2.1 If $t=0$, sample x^{**} independently from $\pi(x)$.
   If $t>0$, sample x^* from the previous population $\{x_{t-1}^{(i)}\}$ with weights w_{t-1} and perturb the particle to obtain $x^{**} \sim K_t(x|x^*)$, where K_t is a perturbation kernel.
   If $\pi(x^{**}) = 0$, return to *S2.1*.
   Simulate a candidate dataset $D_{(b)}(x^{**}) \sim f(D|x^{**})$ B_t times ($b=1, ..., B_t$) and calculate $b_t(x^{**})$.
   If $b_t(x^{**}) = 0$, return to *S2.1*.

S2.2 Set $x_t^{(i)} = x^{**}$ and calculate the weight for particle $x_t^{(i)}$,

$$w_t^{(i)} = \begin{cases} b_t(x_t^{(i)}), & \text{if } t = 0, \\[2ex] \dfrac{\pi(x_t^{(i)})b_t(x_t^{(i)})}{\sum_{j=1}^{N} w_{t-1}^{(j)} K_t(x_{t-1}^{(j)}, x_t^{(i)})}, & \text{if } t > 0. \end{cases}$$

   If $i < N$, set $i=i+1$, go to *S2.1*.

S3 Normalize the weights.
   If $t < T$, set $t=t+1$, go to *S2.0*.

When applying ABC SMC to deterministic systems, we take $B_t=1$, i.e. we simulate the dataset for each particle only once.

APPENDIX B. COMPARISON OF THE ABC SMC ALGORITHM WITH THE ABC PRC OF SISSON *et al.*

In this section, we contrast the ABC SMC algorithm, which we developed in appendix A, and the ABC PRC algorithm of Sisson *et al.* (2007). The algorithms are very similar in principle, and the main difference is that we base ABC SMC on an SIS framework whereas Sisson *et al.* use a SMC sampler as a basis for ABC PRC, where the weight calculation is performed through the use of a backward kernel. Both algorithms are explained in detail in Del Moral *et al.* (2006). The disadvantage of the SMC sampler is that it is impossible to use an optimal backward kernel and it is hard to choose a good one. Sisson *et al.* choose a backward kernel that is equal to the forward kernel, which we suggest can be a poor choice.[3] While this highly simplifies the algorithm since in the case of a uniform prior distribution all the weights become equal, the resulting posterior distributions can result in bad approximations to the true posterior. In particular, using equal weights can profoundly affect the posterior credible intervals.[4]

Here, we compare the outputs of both algorithms using the toy example from Sisson *et al.* (2007). The goal is to generate samples from a mixture of two normal distributions, $(1/2)\phi(0, 1/100) + (1/2)\phi(0, 1)$, where $\phi(\mu, \sigma^2)$ is the density function of a $N(\mu, \sigma^2)$ random

---

[3] See also http://web.maths.unsw.edu.au/scott/papers/paper_sm-cabc_optimal.pdf.
[4] This, including the experiment below, was suggested by Beaumont (2008*b*).
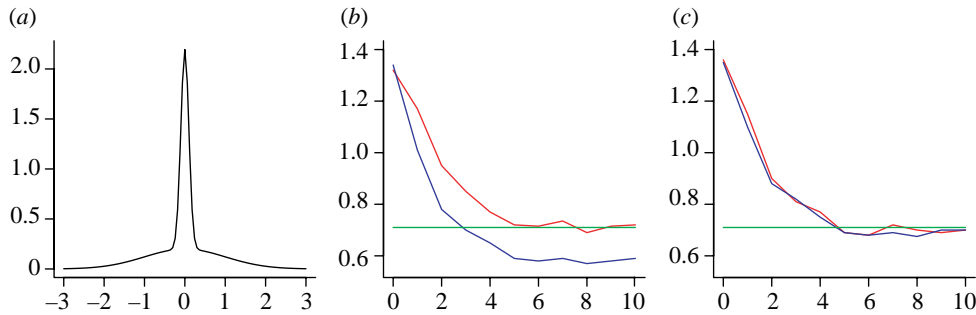
Figure 9. (a) The probability density function of a mixture of two normal distributions, $(1/2)\phi(0, 1/100) + (1/2)\phi(0, 1)$, taken as a toy example used for a comparison of ABC SMC with ABC PRC. (b,c) Plots show how the variance of approximated intermediate distributions π_t changes with populations ($t = 1, \ldots, 10$ on the x-axis). The red curves plot the variance of the ABC SMC population and the blue curves the variance of the non-weighted ABC PRC populations. The perturbation kernel in both algorithms is uniform, $K_t = \sigma U(-1, 1)$. In (b), $\sigma = 1.5$ and this results in poor estimation of the posterior variance with the ABC PRC algorithm. In (c), σ is updated in each population so that it expands over the whole population range. Such σ is big enough for non-weighted ABC PRC to perform equally well as ABC SMC.

variable (figure 9a). Sisson *et al.* approximate the distribution well by using three populations with $\epsilon_1 = 2$, $\epsilon_2 = 0.5$ and $\epsilon_3 = 0.025$, respectively, starting from a uniform prior distribution. However, ABC PRC would perform poorly if they had used more populations. In figure 9b,c, we show how the variance of the approximated posterior distribution changes through populations. We use 100 particles and average over 30 runs of the algorithm, with tolerance schedule $\epsilon = \{2.0, 1.5, 1.0, 0.75, 0.5, 0.2, 0.1, 0.075, 0.05, 0.03, 0.025\}$. The red line shows the variance from ABC SMC, the blue line the variance from the ABC PRC and the green line the variance from the ABC rejection algorithm (figure 9b,c). In the case when the perturbation is relatively small, we see that the variance resulting from improperly weighted particles in ABC PRC is too small (blue line in figure 9b), while the variance resulting from ABC SMC is ultimately comparable to the variance obtained by the ABC rejection algorithm (green line).

However, we also note that when using many populations and too small a perturbation, e.g. uniform perturbation $K_t = \sigma U(-1, 1)$ with $\sigma = 0.15$, the approximation is not very good (results not shown). One therefore needs to be careful to use a sufficiently large perturbation, irrespective of the weighting scheme. However, at the moment, there are no strict guidelines of how best to do this and we decide on the perturbation kernel based on experience and experimentation.

For the one-dimensional deterministic case with a uniform prior distribution and uniform perturbation, one can show that all the weights will be equal when σ of the uniform kernel is equal to the whole parameter range. In this case, the non-weighted ABC PRC yields the same results as ABC SMC (figure 9c). However, when working with complex high-dimensional systems, it simply is not feasible to work with only a small number of populations, or a very big perturbation kernel spreading over the whole parameter range. Therefore, we conclude that it is important to use the ABC SMC algorithm owing to its correct weighting.

Another difference between the two algorithms is that ABC PRC includes the resampling step in order to maintain a sufficiently large effective sample size

(Liu & Chen 1998; Liu 2001; Del Moral *et al.* 2006). In contrast to non-ABC SIS or SMC frameworks, ABC algorithms, by sampling with weights in step *S2.1* prior to perturbation, we suggest, do not require this additional resampling step.

We note that in SIS, the evaluation of weights after each population is computationally more costly, i.e. $O(N^2)$, than the calculation of weights in SMC with a backward kernel, which is $O(N)$, where N denotes the number of particles. However, this cost is negligible in the ABC case, because the vast proportion of computational time in ABC is spent on simulating the model repeatedly. While thousands or millions of simulations are needed, the weights only need to be updated after every population. Therefore, we can easily afford spending a bit more computational time in order to use the correctly weighted version and circumvent the issues related to the backward kernel choice.

APPENDIX C. ABC AND FULL LIKELIHOOD FOR ODE SYSTEMS

We would like to note that the ABC algorithm with the distance function chosen to be squared errors is equivalent to the maximum-likelihood problem for a dynamical system for which Gaussian errors are assumed. In other words, minimizing the distance function

$$\sum_{i=1}^{n} \sum_{j=1}^{m} (x_{ij} - g_j(t_i, \theta))^2, \qquad (C\ 1)$$

where $g(t, \theta) \in \mathbb{R}^m$ is the solution of the m-dimensional dynamical system and $D = (x_i)_{\{i=1,\ldots,n\}}$, $x_i \in \mathbb{R}^m$, are (m-dimensional) data points measured at times $t_1, \ldots, t_n$, is equivalent to maximizing the likelihood function

$$\prod_{i=1}^{n} \frac{1}{(2\pi)^{m/2}|\Sigma|^{1/2}} e^{-(1/2)(x_i - g(t_i, \theta))^T \Sigma^{-1}(x_i - g(t_i, \theta))},$$

where Σ is diagonal and its entries equal. This can straightforwardly be generalized for the case of multiple time-series measurements. Thus, ABC is for deterministic ODE systems closely related to standard Bayesian

inference where the likelihood is evaluated. This is because ODEs are not based on a probability model, and likelihoods are therefore generally defined in a nonlinear regression context (such as assuming that the data are normally distributed around the deterministic solution; e.g. Timmer & Muller 2004; Vyshemirsky & Girolami 2008).

REFERENCES

Anderson, R. M. & May, R. M. 1991 *Infectious diseases of humans: dynamics and control*. New York, NY: Oxford University Press.

Baker, C., Bocharov, G., Ford, J., Lumb, P. S. J., Norton, C. A. H., Paul, T., Junt, P. & Krebs, B. 2005 Ludewig computational approaches to parameter estimation and model selection in immunology. *J. Comput. Appl. Math.* **184**, 50–76. (doi:10.1016/j.cam.2005.02.003)

Banks, H., Grove, S., Hu, S. & Ma, Y. 2005 A hierarchical Bayesian approach for parameter estimation in HIV models. *Inverse Problems* **21**, 1803–1822. (doi:10.1088/0266-5611/21/6/001)

Battogtokh, D., Asch, D. K., Case, M. E., Arnold, J. & Schuttler, H.-B. 2002 An ensemble method for identifying regulatory circuits with special reference to the *qa* gene cluster of Neurospora crassa. *Proc. Natl Acad. Sci. USA* **99**, 16 904–16 909. (doi:10.1073/pnas.262658899)

Beaumont, M. 2008*a Simulations, genetics and human prehistory* (eds S. Matsumura, P. Forester & C. Renfrew). McDonald Institute Monographs, University of Cambridge.

Beaumont, M. A. 2008*b* A note on the ABC-PRC algorithm of Sissons *et al.* (2007). (http://arxiv.org/abs/0805.3079v1).

Beaumont, M. A., Zhang, W. & Balding, D. J. 2002 Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025–2035.

Bortz, D. M. & Nelson, P. W. 2006 Model selection and mixed-effects modeling of HIV infection dynamics. *Bull. Math. Biol.* **68**, 2005–2025. (doi:10.1007/s11538-006-9084-x)

Brown, K. S. & Sethna, J. P. 2003 Statistical mechanical approaches to models with many poorly known parameters. *Phys. Rev. E* **68**, 021 904. (doi:10.1103/PhysRevE.68.021904)

Cox, R. D. & Hinkley, D. V. 1974 *Theoretical statistics*. New York, NY: Chapman & Hall/CRC.

Del Moral, P., Doucet, A. & Jasra, A. 2006 Sequential Monte Carlo samplers. *J. R. Stat. Soc. B* **68**, 411–432. (doi:10.1111/j.1467-9868.2006.00553.x)

Elowitz, M. B. & Leibler, S. 2000 A synthetic oscillatory network of transcriptional regulators. *Nature* **403**, 335–338. (doi:10.1038/35002125)

Enright, W. & Hu, M. 1995 Interpolating Runge–Kutta methods for vanishing delay differential equations. *Computing* **55**, 223–236. (doi:10.1007/BF02238433)

Fagundes, N., Ray, N., Beaumont, M., Neuenschwander, S., Salzano, F. M., Bonatto, S. L. & Excoffier, L. 2007 Statistical evaluation of alternative models of human evolution. *Proc. Natl Acad. Sci. USA* **104**, 17 614–17 619. (doi:10.1073/pnas.0708280104)

Fields, B. N., Knipe, D. M. & Howley, P. M. 1996 *Fundamental virology*. Philadelphia, PA: Lippincott Williams & Wilkins.

Galassi, M. 2003 *GNU scientific library reference manual*, 2nd edn. Cambridge, UK: Cambridge University Press.

Gillespie, D. 1977 Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* **81**, 2340–2361. (doi:10.1021/j100540a008)

Golightly, A. & Wilkinson, D. 2005 Bayesian inference for stochastic kinetic models using a diffusion approximation. *Biometrics* **61**, 781–788. (doi:10.1111/j.1541-0420.2005.00345.x)

Golightly, A. & Wilkinson, D. 2006 Bayesian sequential inference for stochastic kinetic biochemical network models. *J. Comput. Biol.* **13**, 838–851. (doi:10.1089/cmb.2006.13.838)

Gutenkunst, R., Waterfall, J., Casey, F., Brown, K., Myers, C. & Sethna, J. 2007 Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput. Biol.* **3**, e189. (doi:10.1371/journal.pcbi.0030189)

Hammond, B. J. & Tyrrell, D. A. 1971 A mathematical model of common-cold epidemics on Tristan da Cunha. *J. Hyg.* **69**, 423–433.

Hoeting, J., Madigan, D., Raftery, A. & Volinsky, C. 1999 Bayesian model averaging: a tutorial. *Statist. Sci.* **14**, 382–401. (doi:10.1214/ss/1009212519)

Horbelt, W., Timmer, J. & Voss, H. 2002 Parameter estimation in nonlinear delayed feedback systems from noisy data. *Phys. Lett. A* **299**, 513–521. (doi:10.1016/S0375-9601(02)00748-X)

Huang, Y., Liu, D. & Wu, H. 2006 Hierarchical Bayesian methods for estimation of parameters in a longitudinal HIV dynamic system. *Biometrics* **62**, 413–423. (doi:10.1111/j.1541-0420.2005.00447.x)

Johannes, M. & Polson, N. 2005 MCMC methods for continuous-time financial econometrics. In *Handbook of financial econometrics* (eds Y. Ait-Sahalia & L. Hansen). Amsterdam, The Netherlands: Elsevier Science Ltd.

Kass, R. & Raftery, A. 1995 Bayes factors. *J. Am. Stat. Assoc.* **90**, 773–795. (doi:10.2307/2291091)

Kirkpatrick, S., Gelatt, C. & Vecchi, M. 1983 Optimization by simulated annealing. *Science* **220**, 671–680. (doi:10.1126/science.220.4598.671)

Liu, J. S. 2001 *Monte Carlo strategies in scientific computing*. Berlin, Germany: Springer.

Liu, J. S. & Chen, R. 1998 Sequential Monte Carlo methods for dynamic systems. *J. Am. Stat. Assoc.* **93**, 1032–1044. (doi:10.2307/2669847)

Lotka, A. J. 1925 *Elements of physical biology*. Baltimore, MD: Williams & Wilkins Co.

Marjoram, P., Molitor, J., Plagnol, V. & Tavare, S. 2003 Markov chain Monte Carlo without likelihoods. *Proc. Natl Acad. Sci. USA* **100**, 15 324–15 328. (doi:10.1073/pnas.0306899100)

Mendes, P. & Kell, D. 1998 Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. *Bioinformatics* **14**, 869–883. (doi:10.1093/bioinformatics/14.10.869)

Moles, C., Mendes, P. & Banga, J. 2003 Parameter estimation in biochemical pathways: a comparison of global optimization methods. *Genome Res.* **13**, 2467–2474. (doi:10.1101/gr.1262503)

Muller, T. G., Faller, D., Timmer, J., Swameye, I., Sandra, O. & Klingmüller, U. 2004 Tests for cycling in a signalling pathway. *J. R. Stat. Soc. Ser. C* **53**, 557. (doi:10.1111/j.1467-9876.2004.05148.x)

Paul, C. 1992 Developing a delay differential equation solver. *Appl. Numer. Math.* **9**, 403–414. (doi:10.1016/0168-9274(92)90030-H)

Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. 1992 *Numerical recipes in C: the art of scientific computing*, 2nd edn. Cambridge, UK: Cambridge University Press.

Pritchard, J., Seielstad, M. T., Perez-Lezaun, A. & Feldman, M. W. 1999 Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol. Biol. Evol.* **16**, 1791–1798.

Putter, H., Heisterkamp, S. H., Lange, J. M. A. & de Wolf, F. 2002 A Bayesian approach to parameter estimation in HIV dynamical models. *Stat. Med.* **21**, 2199–2214. (doi:10.1002/sim.1211)

Reinker, S., Altman, R. & Timmer, J. 2006 Parameter estimation in stochastic biochemical reactions. *IEE Proc. Syst. Biol.* **153**, 168–178. (doi:10.1049/ip-syb:20050105)

Robert, C. P. & Casella, G. 2004 *Monte Carlo statistical methods*, 2nd edn. New York, NY: Springer.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M. & Tarantola, S. 2008 *Global sensitivity analysis: the primer*. New York, NY: Wiley.

Sanchez, M. & Blower, S. 1997 Uncertainty and sensitivity analysis of the basic reproductive rate: tuberculosis as an example. *Am. J. Epidemiol.* **145**, 1127–1137.

Shibli, M., Gooch, S., Lewis, H. E. & Tyrrell, D. A. 1971 Common colds on Tristan da Cunha. *J. Hyg.* **69**, 255.

Sisson, S. A., Fan, Y. & Tanaka, M. M. 2007 Sequential Monte Carlo without likelihoods. *Proc. Natl Acad. Sci. USA* **104**, 1760–1765. (doi:10.1073/pnas.0607208104)

Stumpf, M. & Thorne, T. 2006 Multi-model inference of network properties from incomplete data. *J. Integr. Bioinform.* **3**, 32.

Timmer, J. & Muller, T. 2004 Modeling the nonlinear dynamics of cellular signal transduction. *Int. J. Bifurcat. Chaos* **14**, 2069–2079. (doi:10.1142/S0218127404010461)

van Kampen, N. G. 2007 *Stochastic processes in physics and chemistry*, 3rd edn. Amsterdam, The Netherlands: North-Holland.

Volterra, V. 1926 Variazioni e fluttuazioni del numero d'individui in specie animali conviventi. *Mem. R. Acad. Naz. dei Lincei* **2**, 31–113.

Vyshemirsky, V. & Girolami, M. A. 2008 Bayesian ranking of biochemical system models. *Bioinformatics* **24**, 833–839. (doi:10.1093/bioinformatics/btm607)

Wilkinson, D. J. 2006 *Stochastic modelling for systems biology*. London, UK: Chapman & Hall/CRC.

Wilkinson, R. D. 2007 Bayesian inference of primate divergence times. PhD thesis, University of Cambridge.

Zucknick, M. 2004 Approximate Bayesian computation in population genetics and genomics. MSc thesis, Imperial College London.