CrossMark

A rare event approach to high-dimensional approximate Bayesian computation

Dennis Prangle[1] · Richard G. Everitt[2] · Theodore Kypraios[3]

Abstract Approximate Bayesian computation (ABC) methods permit approximate inference for intractable likelihoods when it is possible to simulate from the model. However, they perform poorly for high-dimensional data and in practice must usually be used in conjunction with dimension reduction methods, resulting in a loss of accuracy which is hard to quantify or control. We propose a new ABC method for high-dimensional data based on rare event methods which we refer to as RE-ABC. This uses a latent variable representation of the model. For a given parameter value, we estimate the probability of the rare event that the latent variables correspond to data roughly consistent with the observations. This is performed using sequential Monte Carlo and slice sampling to systematically search the space of latent variables. In contrast, standard ABC can be viewed as using a more naive Monte Carlo estimate. We use our rare event probability estimator as a likelihood estimate within the pseudo-marginal Metropolis–Hastings algorithm for parameter inference. We provide asymptotics showing that RE-ABC has a lower computational cost for high-dimensional data than standard ABC methods. We also illustrate our approach empirically, on a Gaussian distribution and an application in infectious disease modelling.

✉ Dennis Prangle
dennis.prangle@newcastle.ac.uk

1   Newcastle University, Newcastle upon Tyne, UK

2   University of Reading, Reading, UK

3   University of Nottingham, Nottingham, UK

1 Introduction

Approximate Bayesian computation (ABC) is a family of methods for approximate inference, used when likelihoods are impossible or impractical to evaluate numerically but simulating datasets from the model of interest is straightforward. ABC can be viewed as a *nearest neighbours* method. It simulates datasets given various parameter values, and finds the closest matches, in some sense, to the observed dataset. The corresponding parameters are used as the basis for inference. Various Monte Carlo methods have been adapted to implement this idea, including rejection sampling (Beaumont et al. 2002), Markov chain Monte Carlo (MCMC) (Marjoram et al. 2003) and sequential Monte Carlo (SMC) (Sisson et al. 2009). However, it is well known that nearest neighbours approaches becomes less effective for higher-dimensional data, a phenomenon referred to as the *curse of dimensionality*. The problem is that even under the best parameter values, it is rare for a high-dimensional simulation to match a fixed target well, essentially because there are many random components all of which must be close matches to observations.

In this paper, we propose a method to deal with this issue and permit higher-dimensional data or summary statistics to be used in ABC. The idea involves introducing latent variables x. We assume data are a deterministic function $y(\theta, x)$, where θ is a vector of parameters. Hence, x encapsulates all the randomness which occurs in the simulation process. Our approach is, for a particular θ value, to use rare event methods to estimate the probability of x values occurring which produce $y(\theta, x) \approx y_{\text{obs}}$. As discussed later, this probability equals, up to proportionality, the approximate likelihood of

🖄 Springer

θ used in existing ABC algorithms. We estimate this probability using SMC algorithms for rare events from Cérou et al. (2012). The resulting estimates are unbiased or low bias, depending on the algorithm, and can be used by many inference methods. We concentrate on the pseudo-marginal Metropolis Hastings algorithm (Andrieu and Roberts 2009), which outputs a sample from a distribution approximating the Bayesian posterior.

The intuition for the rare event probability estimates we use is as follows. Given θ, standard ABC methods effectively simulate one or several x values from their prior and calculate a Monte Carlo estimate of $\Pr(y(\theta, x) \approx y_{\text{obs}})$. This relative error of this estimate has high variance when the probability is small, as is the case when we require close matches. The rare event technique of *splitting* uses nested sets of latent variables $A_1 \supset A_2 \supset \ldots \supset A_T$, representing increasingly close matches. We aim to estimate $\Pr(A_1)$, $\Pr(A_2|A_1)$, $\Pr(A_3|A_2)$, ... and take the product. If these probabilities are all relatively large then the variance of the final estimator's relative error is smaller than using a single stage of Monte Carlo [for a crude variance analysis justifying this see L'Ecuyer et al. (2007). Cérou et al. (2012) prove more detailed results for their SMC algorithms which we summarise later]. We can estimate $\Pr(A_1)$ using Monte Carlo with N samples. Next, we reuse the x samples with $x \in A_1$. We sample randomly from these N times and, to avoid duplicates, perturb each appropriately. We found a good perturbation method was a slice sampling algorithm from Murray and Graham (2016). The resulting sample is used to find a Monte Carlo estimate of $\Pr(A_2|A_1)$. We carry on similarly to estimate the remaining conditional probabilities.

For this approach to work well, a small perturbation of the xs must produce a corresponding small perturbation of the ys. Hence, the mapping $y(\theta, x)$ must be well chosen. This requirement is explored in Sect. 6.1.

We consider two rare event SMC algorithms proposed by Cérou et al. (2012). In one, the nested sets must be fixed in advance and in the other they are selected adaptively during the algorithm. A contribution of this paper is to compare the efficiency of these algorithms within the setting of ABC. Our recommendation, discussed in Sect. 6.2, is a combination of the two approaches: a single run of the adaptive algorithm to select the nested sets, followed by using these in the fixed algorithm.

1.1 Related literature

First, we highlight the difference between our approach and ABC-SMC (Sisson et al. 2009; Moral et al. 2012). These methods find parameter values which are most likely to produce simulations closely matching the observations. We argue that for high-dimensional observations, such simulations are rare even for the best parameter values. Instead, we

use SMC in a different way, to find latent variables which produce successful simulations. In Sect. 6.3, we discuss the possibility of combining these two approaches. Another method that seeks to find promising parameter values is ABC subset simulation (Chiachio et al. 2014). To our knowledge, this is the only other approach to ABC using rare event methods. Again, our approach differs from this by instead searching a space of latent variables.

The most popular approach to deal with the curse of dimensionality in ABC is *dimension reduction*. Here, high-dimensional datasets are mapped to lower dimensional vectors of features, often referred to as *summary statistics*. The quality of a match between simulated and observed data is then judged based only on their corresponding summary vectors. However, using summary statistics involves some loss of information about the posterior which is hard to quantify. Low-dimensional sufficient statistics would avoid this problem but generally do not exist, and there are many competing methods to choose summaries which make a good trade-off between low dimension and informativeness (Blum et al. 2013; Prangle 2017). An alternative approach of Nott et al. (2014) is to improve ABC output by adjusting each parameter's margin to agree with a separate marginal ABC analysis. These analyses can each use different low-dimensional summary statistics, so that the effect of the curse of dimensionality on the margins is reduced. However, there are still issues in selecting these summaries and dealing with approximation error in the dependence structure. Recently, an extension has looked at assuming a Gaussian copula dependence structure (Li et al. 2017). More high-dimensional ABC methods are reviewed in Nott et al. (2017).

Several other authors have recently investigated latent variable approaches to ABC. Neal (2012) introduced *coupled ABC* for household epidemics. This simulates latent variable vectors from their prior and, for each, finds one or many parameter vectors leading to closely matching simulated datasets. These parameters, weighted appropriately, form a sample from an approximate posterior. A similar strategy is employed for more general applications in Meeds and Welling (2015)'s *optimisation Monte Carlo* and the *reverse sampler* of Forneron and Ng (2016). Alternatively, Moreno et al. (2016) perform variational inference, using latent variable vectors drawn from their prior in the estimation of loss function gradients. Another related method is Graham and Storkey (2016), who sample from the (θ, x) space conditioned exactly on the observations using constrained Hamiltonian Monte Carlo (HMC). A limitation is that the $y(\theta, x)$ mapping must be differentiable with respect to both arguments.

A similar SMC approach to ours is outlined, but not implemented, by Andrieu et al. (2012). Analogous methods have been implemented for ABC inference of state space mod-

els, using ABC particle filtering to estimate likelihoods for a sequence of observations (Jasra 2015).

Targino et al. (2015) use similar methods to us in a non-ABC context. They use SMC to estimate posterior quantities for a copula model conditional on a rare event. Like us, they use increasingly rare events as intermediate targets and use slice sampling for perturbation moves. A difference is our focus on estimating the probability of the rare event, and providing results on the asymptotic efficiency of this. Also their perturbation updates each component of x in turn with a univariate slice sampler, while we use truly multivariate updates.

1.2 Contributions and overview

We provide an approximate inference method for the same class of intractable problems as ABC. Our algorithm samples from the same family of posterior approximations as ABC, but can reach more accurate approximations for the same computational cost. In particular, its cost rises more slowly with the data dimension. Therefore, it is feasible to perform inference using a larger, and hence more informative, set of summary statistics. In some cases, it is even feasible to use the full data.

Our method has various differences to competing methods using latent variables. Unlike the majority of these, it does not rely solely on randomly sampling latent variables, but instead searches their space more efficiently. Also unlike HMC approaches, we do not require differentiability assumptions for $y(\theta, x)$.

Typically, SMC methods have many tuning choices. Another benefit of our approach is that these can all be automated. The tuning choices required are simply those for the ABC and PMMH algorithms.

Section 2 describes background information on the methods we use. Section 3 presents our algorithm to estimate the likelihood given a particular parameter vector, and how we use this within a MCMC inference algorithm. Asymptotic results on computational cost are also given here, quantifying the improvement over standard ABC. The method is evaluated on a simple Gaussian example in Sect. 4 and used in an infectious disease application in Sect. 5. Code for these examples is available at https://github.com/dennisprangle/RareEventABC.jl. Section 6 gives a concluding discussion, including when we expect our scheme to work well. "Appendix A" contains technical details of our asymptotics.

2 Background

2.1 Approximate Bayesian computation

Suppose observations y_{obs} are available, and we wish to learn the parameters θ of a model $\pi(y|\theta)$ (a density with respect

to a probability measure dy) given a prior $\pi(\theta)$ (a density with respect to probability measure $d\theta$). Algorithm 1 is an ABC rejection sampling algorithm which performs approximate Bayesian inference. It requires three tuning choices: the number of simulations N, a threshold $\epsilon \geq 0$, and a distance function $d(\cdot, \cdot)$. The latter is typically Euclidean distance or a variation. It is usually sensible to scale data y appropriately so that all components make contributions of similar size to the distance, and we will assume that this has already been done.

Algorithm 1 ABC rejection sampling

Loop over $i = 1, 2, \ldots, N$.
1. Sample θ_i from $\pi(\theta)$.
2. Sample y_i from $\pi(y|\theta_i)$.
3. Accept if $d(y_i, y_{\text{obs}}) \leq \epsilon$.
   End loop
4. **Return**: accepted θ_i values.

The output of Algorithm 1 is a sample from the following approximate posterior density

$$\pi_{\text{ABC}}(\theta|y_{\text{obs}}; \epsilon) \propto \pi(\theta) L_{\text{ABC}}(\theta; \epsilon), \tag{1}$$

where

$$L_{\text{ABC}}(\theta; \epsilon) = V(\epsilon)^{-1} \int \pi(y|\theta) \mathbb{1}[d(y, y_{\text{obs}}) \leq \epsilon] dy, \tag{2}$$

$$V(\epsilon) = \int \mathbb{1}[d(y, y_{\text{obs}}) \leq \epsilon] dy. \tag{3}$$

The *ABC likelihood* L_{ABC} is a convolution of the exact likelihood function and the *kernel*

$$k(y; \epsilon) = V(\epsilon)^{-1} \mathbb{1}[d(y, y_{\text{obs}}) \leq \epsilon], \tag{4}$$

a uniform density on y values close to y_{obs}. Under some weak conditions, as $\epsilon \to 0$ the ABC likelihood converges to the exact likelihood and π_{ABC} to the exact posterior [this is shown by Eq. (6) in "Appendix A", which describes some sufficient conditions]. However, this causes acceptances to become rare. Thus there is a trade-off, controlled by ϵ, between output sample size and the accuracy of π_{ABC}.

ABC rejection sampling is inefficient in the common situation where the prior is much more diffuse than the posterior, as a lot of time is spent on simulations that have very little chance of being accepted. Several more sophisticated ABC algorithms have been proposed which concentrate on performing simulations for θ values believed to have high posterior density. These include versions of importance sampling, MCMC and SMC. These also output samples (sometimes weighted) from an approximation to the posterior, usually π_{ABC} as given in (1). See Marin et al. (2012)

for a review of ABC, including these algorithms and related theory.

As mentioned earlier, ABC suffers from a *curse of dimensionality* issue. Intuitively, the problem is that simulations producing good matches of all summaries simultaneously become increasingly unlikely as $\dim(y)$ grows. For Algorithm 1, it has been proved (Blum 2010; Barber et al. 2015; Biau et al. 2015) that for a fixed value of N the quality of the output sample as an approximation of the posterior deteriorates as d increases, even taking into account the possibility of adjusting ϵ. See Fearnhead and Prangle (2012) for heuristic arguments that the problem also applies to other ABC algorithms.

2.2 Pseudo-marginal Metropolis–Hastings

The approach of this paper is to estimate the ABC likelihood (2) more accurately than standard ABC methods. This section reviews one approach for how such estimates can be used to sample from π_{ABC}.

The Metropolis–Hastings (MH) algorithm samples from a Markov chain with stationary distribution proportional to an unnormalised density $\psi(\theta)$. It is often used in Bayesian inference to produce samples from a close approximation to the posterior distribution. Despite the non-independence of these samples, they can still be used to produce highly accurate Monte Carlo estimates of functions of the posterior. Simulating θ_t, the tth state of the Markov chain is based on sampling a state θ' from a proposal density $q(\theta'|\theta_{t-1})$, typically centred on the preceding state θ_{t-1}. This proposal is accepted as θ_t with probability $\min\left(1, \frac{\psi(\theta')q(\theta_{t-1}|\theta')}{\psi(\theta_t)q(\theta'|\theta_{t-1})}\right)$. Otherwise $\theta_t = \theta_{t-1}$.

This algorithm remains valid if likelihood evaluations are replaced with unbiased nonnegative estimates as follows (Andrieu and Roberts 2009). The state of the Markov chain is now $(\theta_t, \hat{\psi}_t)$, where $\hat{\psi}_t$ is an estimate of $\psi(\theta_t)$, and the acceptance probability must be $\min\left(1, \frac{\hat{\psi}'q(\theta_{t-1}|\theta')}{\hat{\psi}_{t-1}q(\theta'|\theta_{t-1})}\right)$. Crucially, upon acceptance $\hat{\psi}_t$ is set to the estimate $\hat{\psi}'$ for the proposal θ'. So, rather than being recalculated in every iteration, this estimate is used in all future iterations until another proposal is accepted. A version of the resulting *pseudo-marginal Metropolis–Hastings* (PMMH) algorithm, specialised to this paper's setting, is presented below as Algorithm 5.

Optimal tuning of PMMH has been examined theoretically by Pitt et al. (2012), Doucet et al. (2015) and Sherlock et al. (2015), covering the case where each $\hat{\psi}'$ estimate is generated by an SMC algorithm. A central issue is how many SMC particles should be used to optimise the computational efficiency of PMMH. All the authors conclude that this number should be tuned to achieve a particular variance of $\log \hat{\psi}$ (it's assumed, unrealistically, that this variance does not depend on θ. In practice it's typical to investigate the variance

at a fixed value of θ believed to have high posterior density). The value derived for this optimal variance differs between the authors due to their different assumptions, but all values lie in the range 0.8–3.3. Sherlock et al. (2015) also investigate tuning the proposal distribution q, and suggest using proposal variance $\frac{2.562^2}{\dim(\theta)}\Sigma$ where Σ is the posterior variance. They perform simulation studies generally supporting both these results. One key assumption made by all the authors is that $\log \hat{\psi}$ follows a normal distribution. The validity of this assumption in our setting will be investigated later. It's also assumed that the computational cost of SMC is proportional to the number of particles used and does not depend on θ, which is generally true for SMC algorithms.

2.3 Rare event sequential Monte Carlo

To estimate the ABC likelihood (2) in Sect. 3, we will use two algorithms of Cérou et al. (2012) for estimating rare event probabilities using a SMC approach. This section reviews existing work on these algorithms. A few novel remarks which are relevant later are given at the end.

The aim is to estimate a small probability, $P = \Pr(\Phi(x) \leq \epsilon|\theta)$. Here, x is a random variable, θ is a vector of parameters, Φ maps x values to $\mathbb{R}$, and ϵ is a threshold. In the ABC setting of later sections, P will be an estimate of $L_{\mathrm{ABC}}(\theta; \epsilon)$ up to proportionality. As discussed informally in Sect. 1, both algorithms act by estimating conditional probabilities $\Pr(\Phi(x) \leq \epsilon_{k+1}|\theta, \Phi(x) \leq \epsilon_k)$ for a decreasing sequence of ϵ values. In Algorithm 2 (FIXED-RE-SMC), a fixed ϵ sequence must be prespecified. In Algorithm 3 (ADAPT-RE-SMC), the sequence is selected adaptively. Whenever we use RE-SMC without an additional prefix, we are referring to both algorithms.

Cérou et al. (2012) prove that FIXED-RE-SMC produces an unbiased estimator of P, but ADAPT-RE-SMC gives an estimator with $O(N^{-1})$ bias. They also analyse the asymptotic variance of the estimators' relative errors for large N under various assumptions. This variance is generally smaller for ADAPT-RE-SMC. Equality occurs only when FIXED-RE-SMC uses an ϵ sequence such that $\Pr(\Phi(x) \leq \epsilon_{k+1}|\theta, \Phi(x) \leq \epsilon_k)$ is constant as k varies. An approximation to this sequence can be generated by running ADAPT-RE-SMC. We discuss which RE-SMC algorithm to use within our method later. Under optimal conditions, the relative error variances decrease as T, the number of iterations, grows, so that the estimates are more accurate than using plain Monte Carlo, which corresponds to $T = 1$. This result could be extended to take computational cost into account. However, instead we will analyse the overall efficiency of our proposed approach in Sect. 3.3.

Remark 1. Step 2 of ADAPT-RE-SMC selects a threshold sequence in the same way as the ABC-SMC algorithm

Algorithm 2 Rare event SMC algorithm, with fixed ϵ sequence (FIXED-RE-SMC)

Input: Parameters θ, number of particles N, thresholds $\epsilon_1, \epsilon_2, \ldots, \epsilon_T$, Markov kernels for step 3.

1. For $i = 1, \ldots, N$ sample $x_0^{(i)}$ from $\pi(x|\theta)$.
   Loop over $1 \leq t \leq T$:
2. Calculate $I_t = \{i | \Phi(x_{t-1}^{(i)}) \leq \epsilon_t\}$. Let $\hat{P}_t = |I_t|/N$.
   (If $\hat{P}_t = 0$ terminate algorithm returning $\hat{P} = 0$.)
3. For $i = 1, \ldots, N$ sample $x_t^{(i)}$ by drawing j uniformly from I_t and applying a Markov kernel to $x_{t-1}^{(j)}$ with invariant density $\pi(x|\theta, \Phi(x) \leq \epsilon_{t-1})$ (taking $\epsilon_0 = \infty$).
   End loop
4. **Return**: $\hat{P} = \prod_{t=1}^{T} \hat{P}_t$.

Algorithm 3 Rare event SMC algorithm, with adaptive ϵ sequence (ADAPT-RE-SMC)

Input: Parameters θ, number of particles N, target number to accept N_{acc}, acceptance thresholds ϵ, rule to generate Markov kernels for step 3.

1. For $i = 1, \ldots, N$ sample $x_0^{(i)}$ from $\pi(x|\theta)$.
   Loop over $t = 1, 2, \ldots$:
2. Let ϵ_t be the maximum of (a) the N_{acc}th smallest $\Phi(x_{t-1}^{(i)})$ value and (b) ϵ.
   Calculate $I_t = \{i | \Phi(x_{t-1}^{(i)}) \leq \epsilon_t\}$ and $\hat{P}_t = |I_t|/N$.
3. For $i = 1, \ldots, N$ sample $x_t^{(i)}$ by drawing j uniformly from I_t and applying a Markov kernel to $x_{t-1}^{(j)}$ with invariant density $\pi(x|\theta, \Phi(x) \leq \epsilon_{t-1})$ (taking $\epsilon_0 = \infty$).
4. If $\epsilon_t = \epsilon$ break loop and go to step 5, setting $T = t$.
   End loop
5. **Return**: $\hat{P} = \prod_{t=1}^{T} \hat{P}_t$.

---

of Moral et al. (2012). Unlike that work however, this sequence is specialised to one particular θ value rather than being used for many proposed θs.

2. In ADAPT-RE-SMC, typically N_{acc} particles are accepted so that $|I_t| = N_{\mathrm{acc}}$. However, there may be more acceptances in the final iteration or if ties in distance are possible.

3. For $t \leq T$, $\prod_{\tau=1}^{t} \hat{P}_\tau$ is an upper bound on $\hat{P}$ in either RE-SMC algorithm. This bound can be calculated during the tth iteration of the algorithms. This will be used below to terminate the algorithms early once the estimate is guaranteed to be below some prespecified bound.

4. The $x_T^{(i)}$ values can be used for inference of $x|\theta, \Phi(x) \leq \epsilon$. When this is not of interest, as in this paper, then the computational cost can be reduced by omitting step 3 (resampling and Markov kernel propagation) in the final iteration of either algorithm.

5. It's possible for ADAPT-RE-SMC not to terminate. This could occur if the $x_t^{(i)}$ particles become stuck near a mode where $\Phi(x) > \epsilon$ and the Markov kernel is unable to move them to other modes. In Sect. 3.2, we will discuss how our proposed method can avoid this problem by terminating

once it becomes clear the final likelihood estimate will be very low.

6. When ties in the distance are possible, ADAPT-RE-SMC iterations can fail to reduce the threshold. That is, sometimes step 2 can give $\epsilon_{t+1} = \epsilon_t$. This can produce very long run times. Possible improvements to deal with this are discussed in Sect. 6.3 (note that when ADAPT-RE-SMC is being used to select a sequence of thresholds then repeated values should be removed).

7. These algorithms use multinomial resampling. More efficient schemes exist, but are not investigated by the theoretical results of Cérou et al. (2012).

2.4 Slice sampling

We require a suitable Markov kernel to use within the RE-SMC algorithms. This must have invariant density $\pi(x|\theta, \Phi(x) \leq \epsilon_{t-1})$. As discussed below in Sect. 3.2, our ABC setting will assume $\pi(x|\theta)$ is uniform on $[0, 1]^m$. Hence, the required invariant distribution is uniform on the subset of $[0, 1]^m$ such that $\Phi(x) \leq \epsilon_{t-1}$. We will use *slice sampling* as the Markov kernel. This section outlines the general idea of slice sampling and a particular algorithm. We also include some novel material on how it can be adapted to our setting and advantages over alternative choices.

Slice sampling is a family of MCMC methods to sample from an unnormalised target density $\gamma(x)$. The general idea is to sample uniformly from the set $\{(x, h) | h \leq \gamma(x)\}$ and marginalise. We will concentrate on an algorithm of Murray and Graham (2016) for the case where the support of $\gamma(x)$ is $[0, 1]^m$, or a subset of this. Their algorithm updates the current state x by first drawing h from Uniform$(0, \gamma(x))$, then proposing x' values, accepting the first one for which $\gamma(x') \geq h$. The proposal scheme initially considers large changes from x in a randomly chosen direction, and then, if these are rejected, progressively smaller changes.

For use within RE-SMC, $\gamma(x)$ can be taken to be the indicator function $\mathbb{1}(\Phi(x) \leq \epsilon_{t-1})$. This means the condition $\gamma(x') \geq h$ simplifies to $\gamma(x') > 0$, so sampling h can be omitted. The resulting slice sampling update is given by Algorithm 4, which is a special case of the Murray and Graham (2016) algorithm mentioned above (and similar to the *hit-and-run sampler*; see Smith 1996). See their paper for details of the proof that $\gamma(x)$ is the invariant density of this Markov kernel.

Next, we describe two advantages of using slice sampling within RE-SMC, particularly in relation to the alternative of using a Metropolis–Hastings kernel. Firstly, slice sampling requires little tuning. If tuning choices were required, for example a proposal distribution for Metropolis–Hastings, then RE-SMC would need to include rules to make a good choice automatically, which may be difficult. Another advantage of slice sampling is that each iteration outputs a unique x

Algorithm 4 Slice sampling update for RE-SMC

Input: current state x of dimension p, map $\Phi(x)$, threshold ϵ, initial search width w. It's assumed that $\Phi(x) \leq \epsilon$.

1. Sample $v \sim N(0, I_p)$
2. Sample $u \sim \text{Uniform}(0, w)$. Let $a = -u$, $b = w - u$.
   Loop:
3. Sample $z \sim \text{Uniform}(a, b)$.
4. Define a vector x' by $x'_i = r(x_i + zv_i)$ using the *reflection function*:

$$r(y) = \begin{cases} m & m < 1 \\ 2 - m & m \geq 1 \end{cases}$$

   where m is the remainder of y modulo 2.
5. If $\Phi(x') \leq \epsilon$ then **return** x'.
6. If $z < 0$ let $a = z$, otherwise let $b = z$.
   End loop

---

value. On the other hand, Metropolis–Hastings rejections can lead to duplicates, which is problematic within SMC because it leads to increased variance of probability estimates.

The only tuning choice required by Algorithm 4 is the initial search width w. A default choice is $w = 1$, but this means that the number of loops required will increase for small ϵ. To deal with this, we choose $w = 1$ in the first SMC iteration and then select w adaptively, as $\min(1, 2\bar{z})$ where $\bar{z}$ is the maximum final value of $|z|$ from all slice sampling calls in the previous SMC iteration. This choice generally shrinks w based on the most recent value of $\bar{z}$, while avoiding some unwanted behaviours. Firstly, it avoids forcing w to decrease at a fixed rate, so that eventually only very small steps would be attempted. Secondly, it avoids w growing above 1, which would make slice sampling expensive when local moves are required. The effect of our choice is investigated empirically later (see Fig. 3).

3 High-dimensional ABC

This section presents our approach to inference in the ABC setting, using the algorithms reviewed in Sect. 2. Section 3.1 describes how the RE-SMC algorithms can estimate the ABC likelihood given values of θ and ϵ, and a latent variable structure. Such likelihood estimators can be used within several inference algorithms to produce approximate Bayesian inference. In this paper, we concentrate on PMMH. Section 3.2 presents the resulting method. Section 3.3 discusses the computational cost of the resulting *RE-ABC algorithm* in comparison with standard ABC, with particular note of the high-dimensional case.

Two versions of RE-ABC are possible, depending on whether likelihood estimates are produced using FIXED-RE-SMC or ADAPT-RE-SMC. We present both and compare them throughout the remainder of the paper. As will be

explained, in Sects. 3.2 and 6.2, we conclude by arguing in favour of using FIXED-RE-SMC together with an initial run of ADAPT-RE-SMC to select the ϵ sequence.

3.1 Likelihood estimation

For now, suppose θ and $\epsilon > 0$ are fixed. We aim to produce an unbiased estimate of $L_{\text{ABC}}(\theta; \epsilon)$, as defined in (2).

Suppose there exist latent variables x such that the observations can be written as a deterministic function $y = y(\theta, x)$. The idea is that x and θ suffice to specify a complete realisation of the simulation process, even including details such as observation error, and $y(\theta, x)$ is a vector of partial observations. Neglecting θ, which is fixed for now, $y(\theta, x)$ will be written below as simply $y = y(x)$. See Sect. 6.1 for a discussion of properties of $y(\theta, x)$ which help our approach work well.

We specify a density $\pi(x|\theta)$ (with respect to Lebesgue measure) for the latent variables. This is part of the specification of the model, but it can also be viewed as representing prior beliefs about the latent variables. Throughout the paper, we take $\pi(x|\theta)$ to be uniform on $[0, 1]^m$ regardless of θ. Under this interpretation, x is a vector of m independent standard uniform random variables which suffice to carry out the simulation process.

Now, we simply apply one of the RE-SMC algorithms using $\Phi(x) = d(y(x), y_{\text{obs}})$. The small probability estimated by these algorithms is

$$\Pr(\Phi(x) \leq \epsilon | \theta) = \int \pi(x|\theta) \mathbb{1}[d(y(x), y_{\text{obs}}) \leq \epsilon_t] \mathrm{d}x$$

$$= \int \pi(y|\theta) \mathbb{1}[d(y, y_{\text{obs}}) \leq \epsilon_t] \mathrm{d}y,$$

which equals $L_{\text{ABC}}(\theta; \epsilon)$ multiplied by the constant $V(\epsilon)$. Hence, using FIXED-RE-SMC we can obtain an estimate of $L_{\text{ABC}}(\theta; \epsilon)$ which is unbiased, as required by PMMH. Using ADAPT-RE-SMC produces a slightly biased estimate, and we comment on the effect of using this within PMMH in the next section.

Note that we assume $\pi(x|\theta)$ to be uniform simply for convenience. Firstly, many latent variable representations can easily be re-expressed in this form. Secondly, given this assumption, the slice sampling method of Algorithm 4 is well suited to be the Markov kernel within RE-SMC. Our methodology could be adapted to use other $\pi(x|\theta)$ distributions if desired. The main change needed would be to use alternative Markov kernels, for example elliptical slice sampling (see Murray and Graham 2016) for the Gaussian case, or Gibbs updates for the discrete case. These changes could well improve performance for particular applications.

3.2 Inference

Algorithm 5 shows the PMMH algorithm for our setting, which we refer to as RE-ABC. It can use either FIXED-RE-SMC or ADAPT-RE-SMC when estimates of the ABC likelihood are required. We'll use the prefixes FIXED and ADAPT to refer to the version of RE-ABC based on the corresponding RE-SMC algorithm.

Algorithm 5 Pseudo-marginal Metropolis–Hastings using RE-SMC (RE-ABC)

Input: initial state θ_1, number of iterations M, number of SMC particles N and tuning choices for PMMH, RE-SMC and ABC.

1. Let $t = 1$ and calculate $\hat{L}_{ABC,1}$ using FIXED-RE-SMC (or ADAPT-RE-SMC) with slice sampling as the Markov kernel.
  Loop over $2 \leq t \leq M$:
2. Propose new state θ' from $q(\cdot|\theta)$ and sample u from Uniform$(0, 1)$.
3. Let $\hat{L}'_{ABC}$ be the output of FIXED-RE-SMC (or ADAPT-RE-SMC) with slice sampling as the Markov kernel. This algorithm can be stopped early once rejection in the next step is guaranteed.
4. If $u > \frac{\pi(\theta')\hat{L}'_{ABC}q(\theta_{t-1}|\theta')}{\pi(\theta_{t-1})\hat{L}_{ABC,t-1}q(\theta'|\theta_{t-1})}$:

   Reject Let $\theta_t = \theta_{t-1}$ and $\hat{L}_{ABC,t} = \hat{L}_{ABC,t-1}$.

   Else:

   Accept Let $\theta_t = \theta'$ and $\hat{L}_{ABC,t} = \hat{L}'_{ABC}$.

  End loop
  Return: $\theta_1, \theta_2, \ldots, \theta_M$.

For FIXED-RE-ABC, the likelihood estimates are unbiased estimates of $L_{ABC}(\theta; \epsilon)$ up to proportionality. Therefore, the probability of acceptance in step 4 corresponds to a target density proportional to $\pi(\theta)L_{ABC}(\theta; \epsilon)$, i.e., the standard ABC posterior (1). ADAPT-RE-ABC involves biased likelihood estimates so does not sample from exactly this density. However, the bias introduced is small and may have little effect compared to the efficiency benefits of the variance reduction which ADAPT-RE-SMC provides (theoretical and practical aspects of MCMC algorithms that have this character are discussed in Alquier et al. 2016). We investigate this empirically in Sects. 4 and 5 and find no noticeable effect of bias. However, we find ADAPT-RE-SMC to sometimes be less computationally efficient in practice, and so we recommend using the FIXED-RE-SMC algorithm, together with a single run of ADAPT-RE-SMC to select a ϵ sequence. Reasons for this are described shortly, and discussed in more detail in Sect. 6, together with possibilities for improvement.

To reduce computational costs RE-SMC can be terminated as soon as rejection is guaranteed. To implement this, after step 2 of RE-SMC check whether

$$\prod_{\tau=1}^{t_{SMC}} \hat{P}_\tau < \frac{u\pi(\theta_{t-1})\hat{L}_{ABC,t-1}q(\theta'|\theta_{t-1})}{\pi(\theta')q(\theta_{t-1}|\theta')},$$

where t_{SMC} is the current value of the t variable within the RE-SMC algorithm. If this is true, terminate the RE-SMC algorithm and reject the current proposal in the PMMH algorithm. The MCMC algorithm remains valid since the final RE-SMC likelihood estimate is guaranteed to be smaller than $\prod_{\tau=1}^{t_{SMC}} \hat{P}_\tau$ and therefore lead to rejection in PMMH. Early termination prevents extremely long runs of RE-SMC for θ values with low posterior densities. It is most efficient for FIXED-RE-SMC, where it is always possible to terminate in any iteration if the $\hat{P}_\tau$ values are small enough. For ADAPT-RE-SMC, $\hat{P}_\tau \geq \frac{N_{acc}}{N}$ so there is a lower bound of how many iterations are required before termination. This argument suggests ADAPT-RE-SMC is less computationally efficient and agrees with later empirical findings (see Fig. 5).

Earlier we commented that ADAPT-RE-SMC can fail to terminate in some situations. When ties in the distance are not possible, then this is usually not a problem within RE-ABC due to the early termination rule just outlined. However, care is still required the first time ADAPT-RE-SMC is run, and when it is used in pilot runs. Ties in the distance are potentially more problematic and are discussed further in Sect. 6.3.

There are numerous tuning choices required in this PMMH algorithm. Most of these can be based on the output of a pilot analysis, for example an ABC analysis or a short initial run of PMMH. The estimated posterior mean $\hat{\mu}$ can be used as an initial PMMH state. The estimated posterior variance $\hat{\Sigma}$ can be used to tune the PMMH proposal density. Following the PMMH theory discussed in Sect. 2.2, we sample proposal increments from $N\left(0, \frac{2.562^2}{\dim(\theta)}\hat{\Sigma}\right)$ (note that the early termination rule avoids SMC calls having very long run times for some θ values, approximately meeting the assumptions of the PMMH tuning literature). The threshold sequence for FIXED-RE-SMC can be selected by running ADAPT-RE-SMC with $\theta = \hat{\mu}$. To select the number of particles, a few preliminary runs of FIXED-RE-SMC (or ADAPT-RE-SMC) can be performed with $\theta = \hat{\mu}$, aiming to produce a log-likelihood variance of roughly 1. This is at the more conservative end of the range suggested by the theory reviewed earlier.

A crucial tuning choice which remains is ϵ. As in other ABC methods, we suggest tuning this pragmatically based on the computational resources available. This can be done by running ADAPT-RE-SMC with $\theta = \hat{\mu}$ and $\epsilon = 0$ and stopping after a prespecified time, corresponding to how long is available for an iteration of PMMH. The value of ϵ_t when the algorithm is stopped can be used as ϵ. It is still possible for the SMC algorithms to take much longer to run for other θ values. However, the early termination rule will usually mitigate this. Diagnostic plots can be used to investigate whether the ϵ value selected produces simulations judged to be sufficiently similar to the observations. For example, see Figure 1 of the supplementary material.

3.3 Cost

Here, we summarise results on the cost of ABC and RE-ABC in terms of time per samples produced (or effective sample size for PMMH algorithms), in the asymptotic case of small ϵ. Arguments supporting these results are given in "Appendix A". Several assumptions are required, principally that $\pi(y|\theta)$ is a density with respect to Lebesgue measure—informally, the observations must be continuous. Weakening these assumptions is discussed in supplementary material. Note that the results are the same whether FIXED-RE-ABC and ADAPT-RE-ABC is used.

The time per sample is asymptotic to $1/V(\epsilon)$ for ABC and $[\log V(\epsilon)]^2$ for RE-ABC (see (3) for definition of $V(\epsilon)$.) So, asymptotically, RE-ABC has a significantly lower cost to reach the same target density. To illustrate the effect of $D = \dim(y)$, we can consider the asymptotic case of large D (n.b. as shown in the supplementary material, when some observations are non-continuous then D can be replaced with the dimension of $\{y|d(y, y_{\text{obs}}) < \epsilon\}$ for small ϵ.) Under the Lebesgue assumption, (3) gives that $V(\epsilon) \propto \epsilon^D$. Hence, the time per sample is asymptotic to the following expressions, written in terms of $\tau = 1/\epsilon$ for interpretability: $C_1 = \tau^D$ for ABC and $C_2 = D^2[\log \tau]^2 = [\log C_1]^2$ for RE-ABC. Hence, ABC has an exponential cost in D, while RE-ABC has only a quadratic cost. This makes high-dimensional inference more tractable for RE-ABC but dimension reduction via summary statistics will remain useful in controlling the cost when D is large.

These results assume the algorithms are run sequentially. The PMMH stage of RE-ABC is innately sequential, but particle updates can be run in parallel, providing a benefit from parallelisation. Compared to the most efficient ABC algorithms, this is an advantage over ABC-MCMC and seems roughly comparable to that of ABC-SMC algorithms.

4 Gaussian example

In this section, we compare ABC (Algorithm 1) and RE-ABC (Algorithm 5) on a simple Gaussian model. The model is $Y_i \sim N(0, \sigma^2)$ independently for $1 \leq i \leq 25$. We use the prior $\sigma \sim \text{Uniform}(0, 10)$. This is an interesting test case because $\dim(y)$ is large enough to cause difficulties for ABC methods but calculations are quick, and the results can be compared to those of likelihood-based methods.

4.1 Comparison of ABC and RE-ABC

We compared ABC and RE-ABC for observations drawn from the model using $\sigma = 3$. For each of $\epsilon = 8, 9, \ldots, 30$, we ran ABC until $N = 500$ simulations were accepted and calculated the root-mean-squared error and time per accep-

tance. Both FIXED-RE-ABC and ADAPT-RE-ABC were run for 2000 iterations with $\epsilon = 3, 5, 10, 15, 20, 25$. As described in Sect. 3.2, pilot runs were used to tune the number of particles, the Metropolis–Hastings proposal standard deviation and, where necessary, the threshold sequence. We chose the number of acceptances in all ADAPT-RE-ABC analyses to be half the number of particles. To avoid dealing with burn-in, we started the PMMH chains at $\sigma = 3$. For comparison, we also ran ABC-MCMC (Marjoram et al. 2003) and MCMC using the exact likelihood.

Figure 1 shows the results. The left panel illustrates that accuracy improves as the acceptance threshold ϵ is reduced below roughly 15, and, as expected, all methods produce very similar results. In particular, the biased likelihood estimates in ADAPT-RE-ABC have a negligible effect overall. The right panel investigates the time taken per sample by ABC. For MCMC output, this is time divided by the effective sample size (the IMSE estimate of Geyer 1992.) Under ABC and ABC-MCMC, time per sample increases rapidly as ϵ is reduced. For both RE-ABC algorithms, the increase is slower, allowing smaller values of ϵ to be investigated. Neither RE-ABC algorithm is obviously more efficient than the other. This difference between ABC and RE-ABC is consistent with the asymptotics on computational cost described in Sect. 3.3. However, for large ϵ values ABC and ABC-MCMC are cheaper. Overall, RE-ABC permits smaller ϵ values to be investigated at a reasonable computational cost, producing more accurate approximations.

Figure 2 provides some further insight into the efficiency of the RE-ABC algorithms, by looking at the times taken for calls to the RE-SMC algorithms. These have similar distributions for FIXED-RE-ABC and ADAPT-RE-ABC, indicating that there is little difference in their efficiency. One point of interest is that ADAPT-RE-SMC takes a minimum time of 0.095 seconds even when it stops early, while FIXED-RE-SMC sometimes stops early in a much shorter time. However, this happens too rarely to have much effect on overall efficiency.

4.2 Validity of assumptions

We also used the Gaussian example to investigate the validity of various assumptions about RE-ABC used in this paper. First, we considered the cost of slice sampling calls in RE-SMC. Figure 3 shows the mean number of iterations that slice sampling requires during an illustrative FIXED-RE-SMC run. Two cases are shown: non-adaptive slice sampling tuning ($w = 1$ in Algorithm 4) or adaptive tuning (w updated as described in Sect. 2.4). This gives empirical evidence that adaptive tuning prevents the slice sampling cost from increasing during the algorithm, as desired. Repeated trials show that both methods produce very similar mean likelihoods. How-
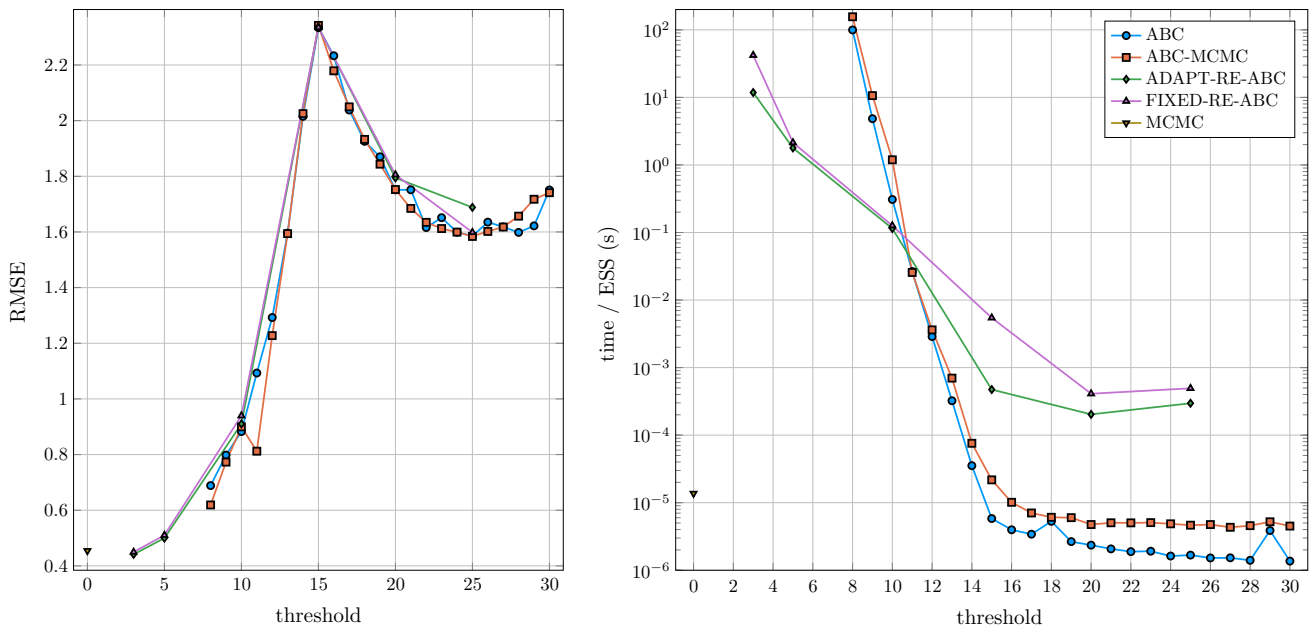
Fig. 1 Simulation study comparing ABC, ABC-MCMC, FIXED-RE-ABC, ADAPT-RE-ABC and exact likelihood MCMC on IID Gaussian data
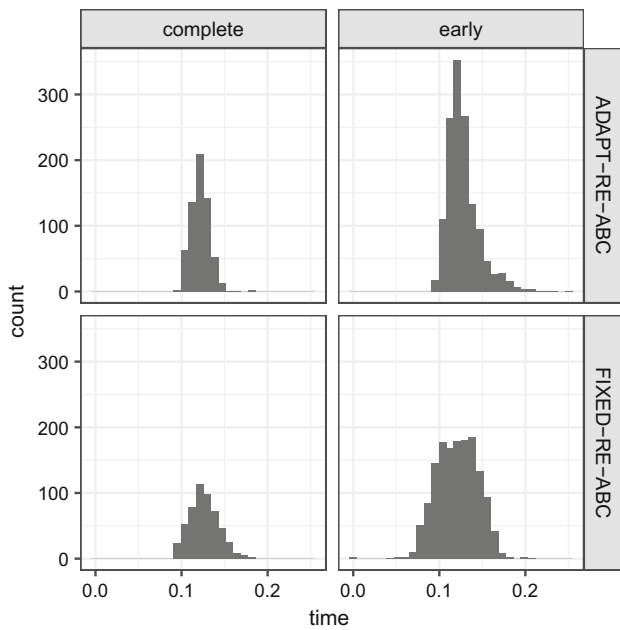

Fig. 2 Histograms of times (in s) taken by calls to RE-SMC within FIXED-RE-ABC and ADAPT-RE-ABC analyses of IID Gaussian data. Both analyses used $\epsilon = 5$ and the same tuning details, chosen using a pilot run. The *left column* is for those calls in which RE-SMC was completed, while the *right* shows those where it was terminated early
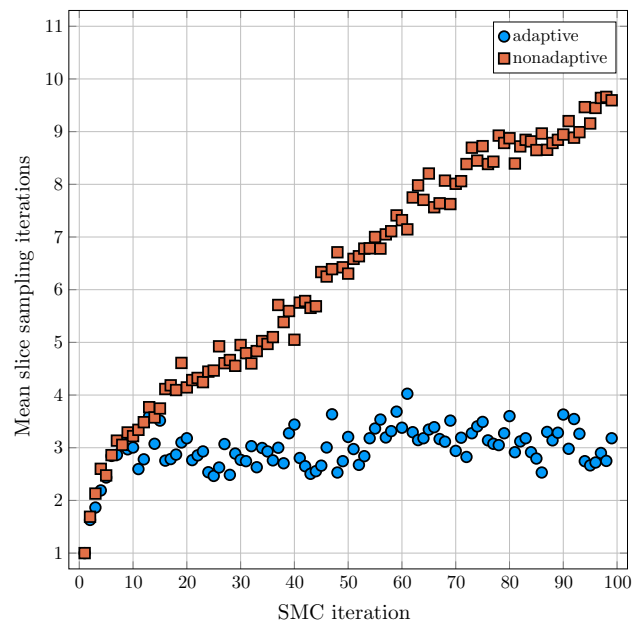
Fig. 3 Number of slice sampling iterations required under adaptive and non-adaptive rules for selecting the tuning parameter w within a run of FIXED-RE-SMC on IID Gaussian data

ever, adaptive tuning did increase the log-likelihood variance slightly so there is a small trade-off in its use.

Secondly, we investigated the distribution of likelihood estimates produced by FIXED-RE-SMC given a particular θ value. Recall that the theoretical literature on PMMH

assumes that these follow a log-normal distribution. Figure 4 shows quantile-quantile plots comparing log-likelihood estimates to normal quantiles. The estimates are approximately normal when a sufficient number of particles are used, but become increasingly skewed as this shrinks. A major departure from normality is that for a small number of particles many likelihood estimates are zero. The corresponding
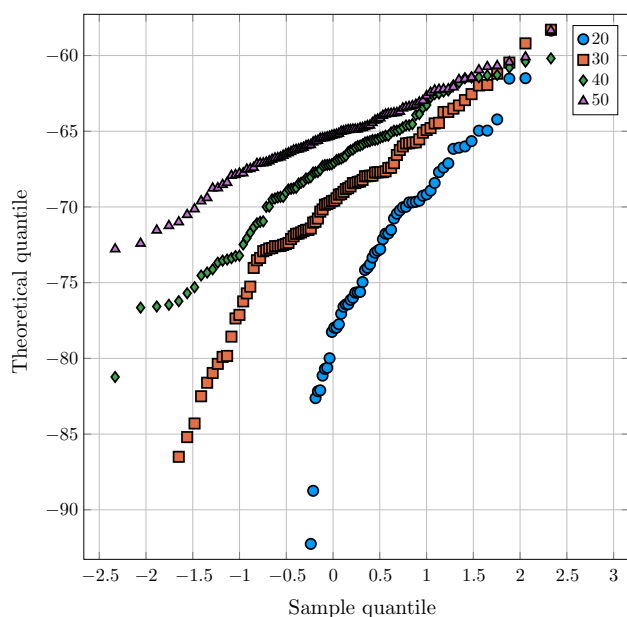
Fig. 4 Normal quantile-quantile plots of log-likelihood estimates for an FIXED-RE-SMC example under various numbers of particles. Omitted points correspond to likelihood estimates of zero

points are omitted from the plot. In conclusion, the normality assumption seems reasonable if a sufficient number of particles are used.

5 Epidemic application

Infectious disease data are often modelled using *compartment models* where members of a population pass through several stages. We will consider a model with susceptible, infectious and removed stages – the so-called *SIR model* (Andersson and Britton 2000). A susceptible individual has not yet been infected with the disease but is vulnerable. An infectious individual has been infected and may spread the disease to others. A removed individual can no longer spread the disease. Depending on the disease this may be due to immunity following recovery, or death.

We will use a stochastic version of this model based on a continuous time stochastic process $\{S(t), I(t) : t \geq 0\}$ for numbers susceptible and infectious at time t. The total population size is fixed at n so the number removed at time t can be derived as $R(t) = n - S(t) - I(t)$. The initial conditions are $(S(0), I(0)) = (n - 1, 1)$. Two jump transitions are possible: infection $(i, j) \mapsto (i - 1, j + 1)$ and removal $(i, j) \mapsto (i, j - 1)$. The simplest version of the model is Markovian and is defined by the instantaneous hazard functions of the two transitions, which are $\frac{\lambda}{n} S(t) I(t)$ for infection and $\gamma I(t)$ for removal. The unknown parameters are λ, controlling infection rates and γ, the removal rate. A goal of inference is often to learn about the basic reproduction

number $R_0 = \lambda/\gamma$. This is the expected number of further infections caused by an initial infected individual in a large susceptible population. When $R_0 < 1$, most epidemics will infect an insignificant proportion of a large population. Many variations on the Markovian SIR model are possible, some of which are outlined below.

Likelihood-based inference is straightforward for fully observed data from an SIR model. However, in practice only partial and possibly noisy observations of removal times are available, producing an intractable likelihood. For many models near-exact inference is possible by MCMC methods (summarised by McKinley et al. 2014), but small changes to the details require new and model-specific algorithms. Approximate inference can be performed by ABC (summarised by Kypraios et al. 2016), which is more adaptable but does not scale well to high-dimensional data. Here, we illustrate how RE-ABC can, without modification, perform inference for several variations on the SIR model, and do so more efficiently than standard ABC methods. As we concentrate on a classic and well-studied dataset, our analysis does not provide any novel subject-area insights.

Section 5.1 describes a method of simulating from SIR models. Section 5.2 discusses the distance function we use to implement RE-ABC. Data analysis is performed in Sect. 5.3.

5.1 Sellke construction

The Sellke construction (Sellke 1983) for an SIR model provides an appealing way to simulate epidemic models. It introduces latent *infectious periods* $g_i \sim F_{\text{inf}}$ and *pressure thresholds* $p_i \sim F_{\text{press}}$ for $1 \leq i \leq n$, all independent. For the Markovian SIR model, F_{inf} is $Exp(\gamma)$ and F_{press} is $Exp(1)$, but other choices are possible and may be more biologically plausible. We condition on $g_1 = 0$ so that the first infection occurs at time 0. Algorithm 6 shows how these variables and the parameter λ are converted to simulated removal times. To use slice sampling, we require the latent variables to be uniformly distributed a priori. Therefore, we use quantiles of the g_is and p_is as the latent variables.

The cost of Algorithm 6 is $O(n \log n)$, where n is the population size. This is because the main loop runs at most $2n - 1$ times and involves finding the minimum of a set of up to $n - 1$ removal times, which requires $O(\log n)$ steps (this is the case if the set is stored as an ordered vector. The cost of adding a new item is $O(\log n)$).

Alternative simulation methods exist, principally the Gillespie algorithm (described in Kypraios et al. 2016, forexample). Here, the latent variables form a sequence controlling the behaviour of each successive jump event. The Gillespie algorithm has the advantage of $O(n)$ cost. However, it seems hard for slice sampling to explore the space of latent variables due to the behaviour of the mapping $y(\theta, x)$. In particular, a small change in latent variables which alters the type of one

jump will typically have a large and unpredictable effect on all the subsequent jumps. For more discussion on desirable properties of $y(\theta, x)$, see Sect. 6.

Note that when F_{press} is $Exp(1)$ then $R_0 = \lambda E(F_{\text{inf}})$ (Andersson and Britton 2000). However, to our knowledge the definition of R_0 has not been extended to cover general F_{press}.

Algorithm 6 Sellke construction epidemic simulator

Input: population size n, scaled infection rate parameter $\beta = \lambda/n$, infectious periods $g_1, \ldots, g_n$ and pressure thresholds $p_2, \ldots, p_n$.

1. Set $r_1 \mapsto g_1$ (assumes individual 1 has infection time 0).
2. Set $r_i \mapsto \infty$ for $i > 2$.
3. Set $I \mapsto 1$ (current number infected), $t \mapsto 0$ (current time), $p \mapsto 0$ (current pressure).
4. **While** $I > 0$**:**
5. Find $p_a = \min\{p_i | p_i > p\}$. If this set is empty use $p_{n+1} = \infty$.
6. Find $r_b = \min\{r_i | r_i > t\}$.
7. Set $p' \mapsto p + \beta I (r_b - t)$ (pressure at time r_b if I does not change)
8. If $p_a < p'$:

   (a) Set $I \mapsto I + 1, t \mapsto t + \frac{p_a - p}{\beta I}, r_a \mapsto t + g_a, p \mapsto p_a$.

9. Else:

   (a) Set $I \mapsto I - 1, t \mapsto r_b, p \mapsto p'$.

10. **End while**

   Output: Removal times $r_1, r_2, \ldots, r_n$. Infinite removal time represents an individual who is never infected.

5.2 Distance function

Recall that the data are the inter-removal times, or equivalently the times since the first removal. For a simulated dataset, let $r_{(1)} \leq r_{(2)} \leq \ldots \leq r_{(\nu)}$ denote the ordered removal times of a dataset with ν removals. The times since first removal are then $s_{(i)} = r_{(i)} - r_{(1)}$ for $1 \leq i \leq \nu$. Similar notation, with the addition of a subscript obs will be used for the observed dataset. We define the distance between a simulated and observed dataset as:

$$
\left[\sum_{i \leq \min(\nu_{\text{obs}}, \nu)} (s_{\text{obs},(i)} - s_{(i)})^2 \right]^{1/2}
$$
$$
+ \sum_{\nu_{\text{obs}} < i \leq \nu} [k + \bar{\rho} - \rho_{(i)}] + \sum_{\nu < i \leq \nu_{\text{obs}}} [k + \rho_{(i)}].
\tag{5}
$$

Here, k is a tuning parameter penalising mismatches between ν and ν_{obs}. We take $k = 1000$. The $\rho_{(i)}$ terms are the sorted simulated pressure thresholds and $\bar{\rho}$ is the total simulated pressure (which equals β times the sum of the infectious periods for removed individuals). They are included to encourage these pressures to increase or decreasing appropriately to match ν and ν_{obs}. Without the pressure terms RE-SMC performed poorly due to the discrete nature of ν. See Sect. 6.3 for further discussion.

5.3 Analysis of Abakaliki data

The Abakaliki dataset contains times between removals from a smallpox epidemic in which 30 individuals were infected from a closed population of 120. It has been studied by many authors under many variations to the basic SIR model. We study three models. The first model uses a Gamma(k, γ) infectious period (similar to Neal and Roberts 2005). The second assumes pressure thresholds are distributed by a Weibull($k, 1$) distribution (as in Streftaris and Gibson 2012.) The third is the Markovian SIR model, but with removal times only recorded within 5 day bins. This is realised by altering the $s_{\text{obs},(i)} - s_{(i)}$ term (difference between simulated and observed day of removal) in (5) to $f(s_{\text{obs},(i)}) - f(s_{(i)})$ where $f(s) = 5\lfloor s/5 \rfloor$, the greatest multiple of 5 less than or equal to s. In each model, there are two or three unknown parameters: λ, controlling infection rates; γ, infectious period scale; k, a shape parameter. These are all assigned independent exponential prior distributions with rate 0.1, representing weakly informative prior beliefs that these parameters are less likely to be large.

We chose the acceptance threshold to be $\epsilon = 15$ on the pragmatic grounds that this produced run times of no more than 6 hours on a desktop PC. Tuning was performed using pilot runs as described in Sect. 3.2. Of particular note is the number of particles required: 300 (Gamma infectious period), 200 (Weibull pressure thresholds) and 400 (binned removal times). First we present results for FIXED-RE-ABC, with discussion on ADAPT-RE-ABC to follow shortly. Table 1 summarises the approximate posterior results. As the parameters differ between models, we don't present parameter estimates. Instead, we give several quantities of interest for each: the R_0 estimate (where defined) and the means and standard deviations of (a) the pressure thresholds and (b) the infectious period. Most quantities are similar to each other and previous analyses (see McKinley et al. 2014 for a summary of many of these) despite the different modelling assumptions. A noticeable difference is that the infectious period is less variable in the model where it follows a Gamma distribution.

Figure 1 of the supplementary material shows simulated epidemics from each model. This shows that our choice of ϵ produces epidemics reasonably close to the observed data for every model. Formal model choice is not straightforward in our framework (see discussion in Sect. 6), but it is easy to explore whether the models produced large differences in log-likelihood. In this case, differences were modest, as shown by Figure 2 in the supplementary material, and within what would be explained, using BIC type arguments, by the

Table 1 Approximate posterior estimates of basic reproduction number R_0 and the means and standard deviations of pressure thresholds and infectious periods for the Abakaliki data under three models computed using FIXED-RE-ABC

Model	R_0	Pressure thresholds		Infectious period	
		Mean	SD	Mean	SD
5 day bins	1.16 (0.30)	0.11 (0.03)	0.11 (0.03)	11.1 (3.0)	11.1 (3.0)
Gamma infectious period	1.18 (0.24)	0.09 (0.03)	0.09 (0.03)	13.6 (3.8)	6.8 (2.2)
Weibull pressure thresholds	–	0.10 (0.04)	0.11 (0.03)	12.4 (3.3)	12.4 (3.3)

The table contains Monte Carlo estimates along with standard deviations in brackets. The R_0 value is not given for the Weibull pressure threshold model as no definition is available for this model

differing number of parameters in the models. So we conclude qualitatively that are no clear differences in fit between the models.

ADAPT-RE-ABC was also tried and returned parameter inference results extremely similar to those for FIXED-RE-ABC – see Table 1 in the supplementary material. This shows that, as in Sect. 4, the bias in its likelihood estimates has a negligible effect on the final results. However, for some analyses the run times were longer. For example, the Gamma infectious period model took 263 minutes for FIXED-RE-ABC and 323 minutes for ADAPT-RE-ABC. Figure 5 investigates this in more detail. It shows that the run time difference is because most calls to RE-SMC terminate early, and these are generally quicker under FIXED-RE-SMC. It is also interesting that ADAPT-RE-SMC is typically faster for completed

RE-SMC calls. These findings are discussed in the next section.

We also ran ABC-MCMC for comparison, using the same MCMC and ϵ tuning choices as for RE-ABC. For run times of comparable length to RE-ABC, ABC-MCMC produced too few acceptances to calculate effective sample sizes accurately. Instead, we consider the time per acceptance. For ABC-MCMC this was at least 12 minutes for all models. For RE-ABC, this value was always less than 2 minutes.

6 Discussion

We have presented a method for approximate inference under an intractable likelihood when simulation of data is possible. It uses the same posterior approximation as ABC, (1), which is controlled by a tuning parameter ϵ. The advantage of our method is that smaller values of ϵ can be achieved for the same computational cost, resulting in more accurate inference. We have shown this is the case through asymptotics (Sect. 3.3) and empirically (Sects. 4, 5.) This increased accuracy allows higher-dimensional data or summary statistics to be analysed in practice.

6.1 Latent variable considerations

Our method represents the model of interest with latent variables x and uses SMC and slice sampling to search for promising x values. For this search strategy to work well, it seems necessary that:

– Evaluating $y(\theta, x)$ is reasonably cheap.
– Sets of the form $\{x | d(y_{\text{obs}}, y(\theta, x)) \leq \epsilon\}$ are easy to explore using slice sampling. This would be difficult for sets made up of many disconnected components, or which are lower dimensional manifolds. Smoothness of y to changes in x will help meet this condition.

Furthermore, our current implementation requires that the number of latent variables is fixed. However, the method could be adapted to the case of a variable number by altering
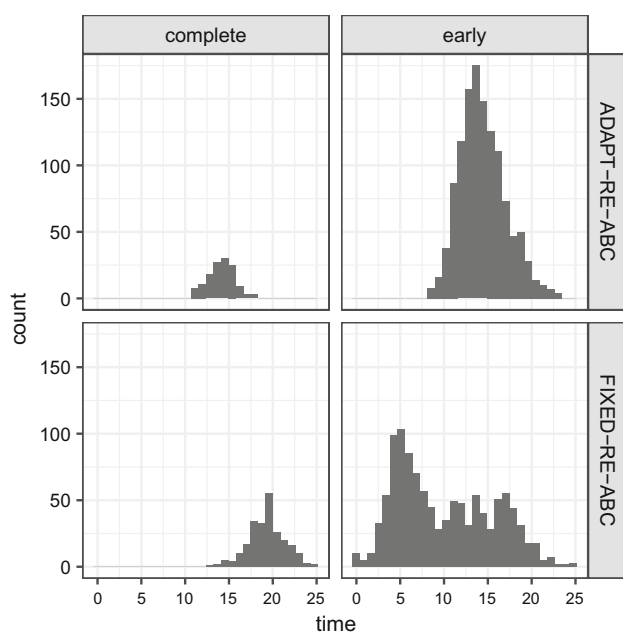

Fig. 5 Histograms of times (in s) taken by calls to RE-SMC within FIXED-RE-ABC and ADAPT-RE-ABC analyses of Abakaliki data. Both analyses used $\epsilon = 15$ and the same tuning details, chosen using a pilot run. The *left column* is for those calls in which RE-SMC was completed, while the *right* shows those where it was terminated early

the slice sampling algorithm (see Section 4.2 of Murray and Graham 2016).

6.2 Adaptive and non-adaptive algorithms

The RE-ABC algorithm can use RE-SMC with a fixed ϵ sequence (FIXED-RE-SMC) or one that is chosen adaptively (ADAPT-RE-SMC). FIXED-RE-SMC provides unbiased estimates of the ABC likelihood, as required by the PMMH algorithm, while ADAPT-RE-SMC has a small bias. In practice, we observe very little difference in the posterior results between the two algorithms, suggesting that this bias has a negligible effect in practice. We also note that, if desired, a bias correction approach from Cérou et al. (2012) could be applied.

Nonetheless, we recommend using the FIXED-RE-SMC algorithm within RE-ABC (together with a pilot run of ADAPT-RE-SMC to choose the ϵ sequence.) The main reason is that it is faster to run in practice, as found in Sect. 5. Figure 5 shows that this is because FIXED-RE-SMC can terminate more quickly for poor proposed θ values. Interestingly, in the iterations where early termination is not required ADAPT-RE-SMC is slightly quicker. We speculate that this is because it often finds a shorter ϵ sequence. Furthermore, the theory of Cérou et al. (2012) suggests that ADAPT-RE-SMC produces less variable ABC likelihood estimates, which would improve PMMH efficiency. Therefore, there may be some scope for a more efficient RE-SMC algorithm which combines the best features of the adaptive and non-adaptive approaches.

6.3 Possible extensions

More efficient ϵ sequence adaptation ADAPT-RE-ABC adapts the ϵ sequence for each θ value separately. One alternative is to instead update the sequence based on information from SMC runs at previous θ values used by PMMH. This could be done using stochastic approximation (see, e.g., Andrieu and Thoms 2008; Garthwaite et al. 2016), with the aim of making the $\hat{P}_t$ values in Algorithm 2 as similar as possible—which minimises asymptotic variance of the likelihood estimates, as discussed in Sect. 2.3. The result would be an adaptive MCMC algorithm, and it may be theoretically challenging to prove it has desirable convergence properties (Andrieu and Thoms 2008).

Joint exploration of (θ, x) Many θ values proposed by RE-ABC are rejected after calculating an expensive likelihood estimate. An appealing alternative is to update the parameters θ conditional on sampled x values, for example through a Gibbs sampler with state (θ, x). Unfortunately in exploratory analyses of such methods, we found the θ updates generally did not mix well. The reason is that x is much more infor-

mative for θ than the observations y_{obs}. This results in small θ moves compared to the posterior's scale.

Alternatively, one could consider nesting an SMC algorithm to explore x within one to explore θ, following Chopin et al. (2013) and Crisan and Miguez (2016). Exploring θ could proceed by reducing ϵ at each iteration. This might avoid the time penalty of ADAPT-RE-SMC when used in PMMH, discussed in Sect. 6.2.

Discrete data RE-SMC can struggle if there is a discrete data variable x^*. It can be hard for SMC to move from accepting a set of latent variables A to another A' in which the range of possible x^* values is smaller, because $\Pr(x \in A' | x \in A, \theta)$ may be very small. The issue is particularly obvious for ADAPT-RE-SMC as the ϵ sequence may fail to move below some threshold for a large number of iterations. For FIXED-RE-SMC, it would instead result in high-variance likelihood estimates. In Sect. 5.2, this problem occurs for ν, the number of removals. There we adopt an application-specific solution by introducing continuous latent variables (pressure thresholds) into the distance function (5). It would be useful to investigate more general solutions from the rare event literature (e.g., Walter 2015). Despite these potential issues, RE-ABC can perform well with discrete data in practice, for example in the binned data model of Sect. 5.3.

Non-uniform ABC kernels In this paper, the ABC likelihood (2) is a convolution of the exact likelihood and a uniform kernel (4). Alternative kernel functions have also been used in ABC (e.g., Wilkinson 2013) such as a Gaussian: $k(y; \epsilon) \propto \exp[-\frac{1}{2\epsilon^2} d(y, y_{\text{obs}})]$. RE-ABC could easily be adapted to make use of these, but it is not clear what effect it would have on our asymptotic results.

Estimating log-likelihood gradients Where log-likelihood gradients can be estimated they allow more efficient inference schemes based on stochastic gradient descent (Poyiadjis et al. 2011) or MCMC (Dahlin et al. 2015). Estimating such gradients from SMC algorithms is possible using the Fisher identity (Poyiadjis et al. 2011). However, the calculation would involve evaluating $\nabla_\theta y(\theta, x)$, which may be demanding for complicated y functions. Moreno et al. (2016) use automatic differentiation to evaluate this for some models. Alternatively, Andrieu et al. (2012) propose using infinitesimal perturbation analysis methods. It would be interesting to use either approach with RE-ABC.

Model choice A desirable extension to RE-ABC would be methods for model choice. Possible methods to extend our PMMH approach include reversible jump MCMC or using a deviance information criterion. See Chkrebtii et al. (2015) and François and Laval (2011) for versions of these methods in the ABC context. Alternatively, it may be more fruitful to

use our likelihood estimate in algorithms which directly output model evidence estimates, such as importance sampling or population Monte Carlo (Cappé et al. 2004).

Appendix A: Computational cost

This appendix justifies the computational costs of ABC and RE-ABC stated in Sect. 3.3. The argument for ABC is rigorous, while that for RE-ABC is more heuristic. Note that throughout this appendix there is no need to distinguish between FIXED-RE-ABC and ADAPT-RE-ABC.

The results are for the asymptotic regime of small ϵ and hold for almost all y_{obs}. We make several assumptions:

A1 The density $\pi(y|\theta)$ is with respect to Lebesgue measure dy of dimension D.
A2 The distance function is Euclidean distance.
A3 Running slice sampling once requires $O(1)$ function evaluations.
A4 RE-SMC uses $O(-\log \Pr(d(y, y_{\text{obs}}) \leq \epsilon|\theta))$ iterations.
A5 The time required to evaluate $y(\theta, x)$ is bounded above and below by nonzero constants which do not depend on θ or x.

Also, we will usually focus on the case where D is asymptotically large.

Informally, A1 requires that all components of y have continuous distributions. Under A2 a key mathematical result below, (6), follows easily. Also, a consequence of A2 which we will use is that, from (3), $V(\epsilon) \propto \epsilon^{-D}$. A3 states that the cost of slice sampling does not increase as ϵ shrinks. This is plausible due to our adaptive choice of w (see Sect. 2.4) and was empirically verified above (see Fig. 3.) It follows that running RE-SMC requires $O(NT)$ function evaluations: the number is asymptotic to the number of particles multiplied by the number of SMC iterations. A4 states that the number of iterations used by RE-SMC is asymptotically proportional to the log of the rare probability being estimated. This follows from a result of Cérou et al. (2012), reviewed in Sect. 2.3, that when the RE-SMC algorithm is tuned optimally $\Pr(A_{k+1}|\theta, x \in A_k)$ is constant, say α, where A_k denotes the event $d(y(\theta, x), y_{\text{obs}}) \leq \epsilon_k$. Therefore, $\Pr(d(y, y_{\text{obs}}) \leq \epsilon|\theta) = \alpha^T$, and taking logs gives A4. So the

assumption is that RE-SMC is tuned to perform similarly to optimal tuning. Assumption A5 states that performing a simulation has a minimum and maximum time requirement regardless of the inputs, which is usually reasonable. This ensures that computation time is asymptotic to the number of simulations performed.

Many of these assumptions can be weakened. This is discussed in the supplementary material, especially for the case of the epidemic model of Sect. 5.

A.1 ABC

Consider the probability of a simulation being accepted given θ:

$$\Pr(d(y, y_{\text{obs}}) \leq \epsilon|\theta) = \int \pi(y|\theta) \mathbb{1}(d(y, y_{\text{obs}}) \leq \epsilon) dy.$$

By the Lebesgue differentiation theorem (see Stein and Shakarchi 2009 for example) for almost all y_{obs}:

$$\lim_{\epsilon \to 0} V(\epsilon)^{-1} \int \pi(y|\theta) \mathbb{1}(d(y, y_{\text{obs}}) \leq \epsilon) dy = \pi(y_{\text{obs}}|\theta),$$

$$(6)$$

Hence for small ϵ:

$$\Pr(d(y, y_{\text{obs}}) \leq \epsilon|\theta) \sim V(\epsilon), \qquad (7)$$

where $\sim$ represents an asymptotic relation (note that while $\pi(y_{\text{obs}}|\theta)$ does not affect this asymptotic relationship, the acceptance probability will decrease for small $\pi(y_{\text{obs}}|\theta)$, i.e., for poor θ choices).

By assumption A5 the time per accepted sample is asymptotic to the number of simulations per accepted sample. Using (7), the latter is asymptotic to $1/V(\epsilon)$. In the case of large D assumption A2 gives that this is $O(\tau^D)$, where $\tau = 1/\epsilon$. For ABC versions of MCMC and SMC, time per accepted sample (or effective sample) is also bounded below by $\min_\theta \Pr(d(y, y_{\text{obs}}) \leq \epsilon|\theta)^{-1}$, so the same result applies.

A.2 RE-ABC

For simplicity, we analyse RE-ABC without the possibility of early termination in the RE-SMC algorithm. An algorithm including early termination will give the same output for a smaller computational cost, although we suspect the gain is only likely to be a $O(1)$ factor. Using the asymptotic results reviewed in Sect. 2 on SMC likelihood estimation and PMMH, we conclude the following. The number of particles in RE-SMC should be $N = O(T)$ to give a likelihood estimator whose log has variance $O(1)$, which optimises efficiency when these estimates are used in PMMH. So, using A3, the

number of simulations required by an iteration of RE-ABC is $O(T^2)$. Using A4 and (7) gives $T = O(-\log V(\epsilon))$.

So the number of simulations required per iteration of RE-ABC is $O([\log V(\epsilon)]^2)$. In the case of large D using A2 gives that this is $O(D^2[\log \tau]^2)$. As in the previous section, assumption A5 implies these expressions also give the time per sample of RE-ABC. They are also valid for the more relevant quantity of time per effective sample since effective sample size is proportional to the actual sample size.

References

Alquier, P., Friel, N., Everitt, R., Boland, A.: Noisy Monte Carlo: convergence of Markov chains with approximate transition kernels. Stat. Comput. **26**(1), 29–47 (2016)

Andersson, H., Britton, T.: Stochastic Epidemic Models and Their Statistical Analysis. Springer, Berlin (2000)

Andrieu, C., Doucet, A., Lee, A.: Contribution to the discussion of Fearnhead and Prangle (2012). J. R. Stat. Soc. B **74**, 451–452 (2012)

Andrieu, C., Roberts, G.O.: The pseudo-marginal approach for efficient Monte Carlo computations. Ann. Stat. **39**, 697–725 (2009)

Andrieu, C., Thoms, J.: A tutorial on adaptive MCMC. Stat. Comput. **18**(4), 343–373 (2008)

Barber, S., Voss, J., Webster, M.: The rate of convergence for approximate Bayesian computation. Electron. J. Stat. **9**, 80–105 (2015)

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

Biau, G., Cérou, F., Guyader, A.: New insights into approximate Bayesian computation. Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques **51**(1), 376–403 (2015)

Blum, M.G.B.: Approximate Bayesian computation: a nonparametric perspective. J. Am. Stat. Assoc. **105**(491), 1178–1187 (2010)

Blum, M.G.B., Nunes, M.A., Prangle, D., Sisson, S.A.: A comparative review of dimension reduction methods in approximate Bayesian computation. Stat. Sci. **28**, 189–208 (2013)

Cappé, O., Guillin, A., Marin, J.-M., Robert, C.P.: Population Monte Carlo. J. Comput. Gr. Stat. **13**(4), 907–929 (2004)

Cérou, F., Del Moral, P., Furon, T., Guyader, A.: Sequential Monte Carlo for rare event estimation. Stat. Comput. **22**(3), 795–808 (2012)

Chiachio, M., Beck, J.L., Chiachio, J., Rus, G.: Approximate Bayesian computation by subset simulation. SIAM J. Sci. Comput. **36**(3), A1339–A1358 (2014)

Chkrebtii, O.A., Cameron, E.K., Campbell, D.A., Bayne, E.M.: Trans-dimensional approximate Bayesian computation for inference on invasive species models with latent variables of unknown dimension. Comput. Stat. Data Anal. **86**, 97–110 (2015)

Chopin, N., Jacob, P.E., Papaspiliopoulos, O.: SMC^2: an efficient algorithm for sequential analysis of state space models. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **75**(3), 397–426 (2013)

Crisan, D., Miguez, J.: Nested particle filters for online parameter estimation in discrete-time state-space Markov models. arXiv:1308.1883 (2016)

Dahlin, J., Lindsten, F., Schön, T.B.: Particle Metropolis–Hastings using gradient and Hessian information. Stat. Comput. **25**(1), 81–92 (2015)

Del Moral, P., Doucet, A., Jasra, A.: An adaptive sequential Monte Carlo method for approximate Bayesian computation. Stat. Comput. **22**(5), 1009–1020 (2012)

Doucet, A., Pitt, M.K., Deligiannidis, G., Kohn, R.: Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. Biometrika **102**(2), 295–313 (2015)

Fearnhead, P., Prangle, D.: Constructing summary statistics for approximate Bayesian computation: semi-automatic ABC. J. R. Stat. Soc. B **74**, 419–474 (2012)

Forneron, J.-J., Ng, S.: A likelihood-free reverse sampler of the posterior distribution. In: GonzÁlez-Rivera, G., Hill, R. C., Lee, T.-H. (eds.) Essays in Honor of Aman Ullah, pp. 389–415. Emerald Group Publishing Limited (2016)

François, O., Laval, G.: Deviance information criteria for model selection in approximate Bayesian computation. Stat. Appl. Genet. Mol. Biol. **10**(1) (2011). doi:10.2202/1544-6115.1678

Garthwaite, P.H., Fan, Y., Sisson, S.A.: Adaptive optimal scaling of Metropolis–Hastings algorithms using the Robbins–Monro process. Commun. Stat. Theory Methods **45**(17), 5098–5111 (2016)

Geyer, C.J.: Practical Markov chain Monte Carlo. Stat. Sci. **7**, 473–483 (1992)

Graham, M.M., Storkey, A.: Asymptotically exact conditional inference in deep generative models and differentiable simulators. arXiv:1605.07826 (2016)

Jasra, A.: Approximate Bayesian computation for a class of time series models. Int. Stat. Rev. **83**, 405–435 (2015)

Kypraios, T., Neal, P., Prangle, D.: A tutorial introduction to Bayesian inference for stochastic epidemic models using Approximate Bayesian Computation. Math. Biosci. **287**, 42–53 (2016)

L'Ecuyer, P., Demers, V., Tuffin, B.: Rare events, splitting, and quasi-Monte Carlo. ACM Trans. Model. Comput. Simul. (TOMACS) **17**(2), 9 (2007)

Li, J., Nott, D.J., Fan, Y., Sisson, S.A.: Extending approximate Bayesian computation methods to high dimensions via a Gaussian copula model. Comput. Stat. Data Anal. **106**, 77–89 (2017)

Marin, J.-M., Pudlo, P., Robert, C.P., Ryder, R.J.: Approximate Bayesian computational methods. Stat. Comput. **22**(6), 1167–1180 (2012)

Marjoram, P., Molitor, J., Plagnol, V., Tavaré, S.: Markov chain Monte Carlo without likelihoods. Proc. Natl. Acad. Sci. **100**(26), 15324–15328 (2003)

McKinley, T.J., Ross, J.V., Deardon, R., Cook, A.R.: Simulation-based Bayesian inference for epidemic models. Comput. Stat. Data Anal. **71**, 434–447 (2014)

Meeds, T., Welling, M.: Optimization Monte Carlo: efficient and embarrassingly parallel likelihood-free inference. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems, pp. 2071–2079. Curran Associates, Inc. (2015)

Moreno, A., Adel, T., Meeds, E., Rehg, J.M., Welling, M.: Automatic variational ABC. arXiv:1606.08549 (2016)

Murray, I., Graham, M.M.: Pseudo-marginal slice sampling. J. Mach. Learn. Res. **51**, 911–919 (2016)

Neal, P.: Efficient likelihood-free Bayesian computation for household epidemics. Stat. Comput. **22**(6), 1239–1256 (2012)

Neal, P., Roberts, G.: A case study in non-centering for data augmentation: stochastic epidemics. Stat. Comput. **15**(4), 315–327 (2005)

Nott, D.J., Fan, Y., Marshall, L., Sisson, S.A.: Approximate Bayesian computation and Bayes linear analysis: toward high-dimensional ABC. J. Comput. Gr. Stat. **23**(1), 65–86 (2014)

Nott, D.J., Ong, V.M.-H., Fan, Y., Sisson, S.A.: High-dimensional ABC. In: Scott, A., Sisson, Y.E., Beaumont, M. (eds.) Handbook of Approximate Bayesian Computation (Forthcoming). Chapman and Hall/CRC Press, Boca Raton (2017)

Pitt, M.K., Silva, R.D.S., Giordani, P., Kohn, R.: On some properties of Markov chain Monte Carlo simulation methods based on the particle filter. J. Econ. **171**(2), 134–151 (2012)

Poyiadjis, G., Doucet, A., Singh, S.S.: Particle approximations of the score and observed information matrix in state space models

with application to parameter estimation. Biometrika **98**(1), 65–80 (2011)

Prangle, D.: Summary statistics. In: Scott, A., Sisson, Y.E., Beaumont, M. (eds.) Handbook of Approximate Bayesian Computation (Forthcoming). Chapman and Hall/CRC Press, Boca Raton (2017)

Sellke, T.: On the asymptotic distribution of the size of a stochastic epidemic. J. Appl. Prob. **20**, 390–394 (1983)

Sherlock, C., Thiery, A.H., Roberts, G.O., Rosenthal, J.S.: On the efficiency of pseudo-marginal random walk Metropolis algorithms. Ann. Stat. **43**(1), 238–275 (2015)

Sisson, S.A., Fan, Y., Tanaka, M.M.: Correction: sequential Monte Carlo without likelihoods. Proc. Natl. Acad. Sci. **106**(39), 16889–16890 (2009)

Smith, R.L.: The hit-and-run sampler: a globally reaching Markov chain sampler for generating arbitrary multivariate distributions. In: Proceedings of the 28th Conference on Winter Simulation, pp. 260–264. IEEE Computer Society (1996)

Stein, E.M., Shakarchi, R.: Real Analysis: Measure Theory, Integration, and Hilbert Spaces. Princeton University Press, Princeton (2009)

Streftaris, G., Gibson, G.J.: Non-exponential tolerance to infection in epidemic systems—modeling, inference, and assessment. Biostatistics **13**(4), 580–593 (2012)

Targino, R.S., Peters, G.W., Shevchenko, P.V.: Sequential Monte Carlo samplers for capital allocation under copula-dependent risk models. Insur. Math. Econ. **61**, 206–226 (2015)

Walter, C.: Rare event simulation and splitting for discontinuous random variables. ESAIM: Prob. Stat. **19**, 794–811 (2015)

Wilkinson, R.D.: Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. Stat. Appl. Genet. Mol. Biol. **12**(2), 129–141 (2013)