Daniel Silk^a, Sarah Filippi^a and Michael P. H. Stumpf*

Optimizing threshold-schedules for sequential approximate Bayesian computation: applications to molecular systems

Abstract: The likelihood-free sequential Approximate Bayesian Computation (ABC) algorithms are increasingly popular inference tools for complex biological models. Such algorithms proceed by constructing a succession of probability distributions over the parameter space conditional upon the simulated data lying in an ϵ -ball around the observed data, for decreasing values of the threshold ϵ . While in theory, the distributions (starting from a suitably defined prior) will converge towards the unknown posterior as ϵ tends to zero, the exact sequence of thresholds can impact upon the computational efficiency and success of a particular application. In particular, we show here that the current preferred method of choosing thresholds as a pre-determined quantile of the distances between simulated and observed data from the previous population, can lead to the inferred posterior distribution being very different to the true posterior. Threshold selection thus remains an important challenge. Here we propose that the threshold-acceptance rate curve may be used to determine threshold schedules that avoid local optima, while balancing the need to minimise the threshold with computational efficiency. Furthermore, we provide an algorithm based upon the unscented transform, that enables the threshold-acceptance rate curve to be efficiently predicted in the case of deterministic and stochastic state space models.

Keywords: dynamical system; parameter inference; unscented transform.

^aThese authors contributed equally.

*Corresponding author: **Michael P.H. Stumpf**, Centre for Integrative Systems Biology at Imperial College London, UK, e-mail: m.stumpf@imperial.ac.uk

Daniel Silk and Sarah Filippi: Centre for Integrative Systems Biology at Imperial College London, UK

Introduction

Mathematical models have become powerful tools for both summarising our current biological understanding, and generating novel hypotheses. However, as our models become more ambitious in size and complexity, the computational challenges of a number of tasks such as parameter inference and model validation are increasingly demanding. For large, complex or stochastic models, exploring the likelihood surface can be too complicated or numerically too demanding, even though it is possible to simulate the model. For this reason, likelihood-free methods such as Approximate Bayesian Computation (ABC), and its more efficient sequential versions, are becoming increasingly important.

Sequential ABC algorithms proceed by constructing a succession of probability distributions over the parameter space conditional on the model simulations lying in an ϵ -ball around the observed data (or more generally a summary of the data) and use decreasing values of the threshold ϵ to incrementally approximate the true posterior distribution. While in theory the distributions (starting from a suitably defined prior) will converge towards the unknown posterior as ϵ tends to zero, in practice the exact sequence of thresholds can have a great impact on the computational efficiency and success of a particular application. Currently, thresholds are typically chosen as a pre-determined quantile of the distances between simulated and observed data from the previous population, or simply by intuition – the drawbacks of which are made clear in the results below.

Here we present an automated and adaptive method for threshold choice that is based upon the threshold-acceptance rate (TAR) curve. Knowledge of the complete TAR curve allows acceptance thresholds to

be chosen that not only balance the need to minimise ϵ with computational efficiency, but further, enable the detection and avoidance of local optima – a problem that, although perhaps not always acknowledged, has plagued previous threshold schemes. We will show below that this problem is particularly pertinent for schemes that choose threshold schedules from quantiles of the previous population of accepted particles. In the simplified case of deterministic or stochastic state space models, we provide a method for obtaining the TAR curve using the unscented transform (UT). Generally known for its use in extending the Kalman filter to non-linear problems, the UT allows the statistics of a Gaussian random variable that has undergone a non-linear transform to be estimated. In combination with Gaussian mixtures, the UT can be used to predict the ABC acceptance rate for any threshold value with minimal computational expense.

The remainder of the paper is organised as follows: we first introduce sequential ABC and use a simple illustrative model to discuss the challenges of threshold selection, and in particular, the difficulty of avoiding local optima. We then describe how these challenges may be overcome given knowledge of the TAR curve. In the case of deterministic models (with Gaussian measurement error), we suggest an algorithm for predicting the TAR curve and automatically choosing threshold values. Finally we compare the performance of the new adaptive method with various fixed quantile schedules for inference on both the illustrative model, and biological systems including two biochemical oscillators.

Adaptive sequential Monte Carlo methods in approximative Bayesian computation

The aim of ABC is to obtain a good and computationally affordable approximation to the posterior distribution

$$p(\theta|x^*) \propto f(x^*|\theta)\pi(\theta),$$

where $\pi(\theta)$ denotes the prior distribution over the parameter space and $f(x^*|\theta)$ is the likelihood of the observed data x^* for a given parameter, θ . Rather than evaluating the likelihood directly, which for many real-world problems can be intractable, ABC-based approaches use systematic comparisons between real and simulated data. The main principle consists of comparing the simulated data, x , with the real data, x^* , and accepting simulations if a suitable distance measure between them, $\Delta(x, x^*)$, is less than a specified threshold, ϵ . The ABC algorithm thus provides a sample from the approximate posterior of the form,

$$p(\theta|x^*) \approx p_\epsilon(\theta|x^*) \propto \int f(x|\theta) \mathbb{1}(\Delta(x, x^*) \leq \epsilon) \pi(\theta) dx.$$

The simple ABC scheme outlined above suffers from the same shortcomings as other rejection samplers: most of the samples are drawn from regions of parameter space which cannot give rise to simulation outputs that resemble the data. Over the past few years many improvements to these algorithms have been proposed that makes ABC inference more efficient: regression-adjusted ABC (Tallmon et al., 2004; Fagundes et al., 2007; Blum and François, 2010), Markov chain Monte Carlo ABC schemes (Marjoram and Molitor, 2003; Ratmann et al., 2007), and ABC implementing variants of sequential importance sampling (SIS) or sequential Monte Carlo (SMC) (Sisson et al., 2007; Beaumont et al., 2009; Toni et al., 2009; Del Moral et al., 2011). While our focus will be upon this latter class of ABC algorithms, it is worth keeping in mind that the methods and results discussed below are applicable to other sequential ABC frameworks. This will be explored further in the discussion.

ABC methods based on SIS or SMC samplers aim to sample from a sequence of distributions, which increasingly resemble the target posterior; they are constructed by estimating intermediate distributions $p_{\epsilon_t}(\theta|x^*)$ for a decreasing sequence of $\{\epsilon_t\}_{1 \leq t \leq T}$. In this article, we focus on the implementation of Toni et al. (2009) and Beaumont et al. (2009) described in Algorithm 2. This implementation that we will call in the following ABC SMC differs from the ABC SMC algorithm of Del Moral et al. (2011) and Drovandi and Pettitt (2011) in a number of points, and proceeds as follows. The first population of *particles* is constructed using

the rejection ABC algorithm described above with a sufficiently large value of ϵ_1 such that many particles are accepted: the parameters θ are drawn from the prior distribution $\pi(\theta)$, and are accepted only if the distance between the simulated and observed data is smaller than ϵ_1 . We denote by $\{\theta^{(i,t)}\}_{1 \leq i \leq N}$ the set of accepted particles at step t , and by $\{x^{(i,t)}\}_{1 \leq i \leq N}$ the corresponding simulated data. Each particle $\theta^{(i,t)}$ has an associated weight $\omega^{(i,t)}$; in the first population all weights are equal to $1/N$. For each intermediate population t , a parameter $\{\theta^{(i,t-1)}\}_{1 \leq i \leq N}$ is sampled from the previous population, $t-1$, with probability defined by the weights, $\{\omega^{(i,t-1)}\}_{1 \leq i \leq N}$, and perturbed using a *perturbation kernel*, $\tilde{\theta} \sim K_t(\cdot | \theta)$; the parameter $\tilde{\theta}$ is then accepted if and only if the distance between the simulated and the observed data is smaller than ϵ_t . These sample, perturbation, simulation and acceptance/rejection steps are repeated until N particles have been accepted. The ratio of this population size, N , and the number of times the model has been simulated during the t -th round will be called the acceptance rate and denoted $\mathfrak{R}_t$. The weight of each particle is then computed as

$$\omega^{(i,t)} = \frac{\pi(\theta^{(i,t)})}{\sum_{j=1}^N \omega^{(j,t-1)} K_t(\theta^{(i,t)} | \theta^{(j,t-1)})}.$$

The efficiency of the sequential ABC algorithm described above strongly relies on the choice of the perturbation kernel $\{K_t(\cdot | \cdot)\}_t$ as well as the sequence of thresholds $\{\epsilon_t\}_t$. Over the past years adaptive methods to choose perturbation kernels have gained popularity. Beaumont et al. (2009) first suggested to use a componentwise-normal perturbation kernel with an adaptive choice for the variances. Filippi et al. (2013) then generalized this approach to a multivariate normal perturbation kernel and compared the efficiency of the ABC SMC algorithm for a selection of adaptive covariance matrices. In this paper we use such a multivariate normal perturbation kernel based upon the entire previous population.

Until recently, there has been no systematic way of determining the threshold sequence. In practice an ideal threshold scheme ($\epsilon_n \rightarrow 0$) has been thought of as simply the one that minimizes the total number of simulations, since this is typically the most computationally expensive part of any ABC algorithm. This in itself requires a careful balance between a small number of populations i.e., a rapidly decreasing sequence of thresholds, and a high acceptance rate per round which generally happens if the difference between two consecutive thresholds is small enough. Perhaps the most commonly used adaptive scheme for threshold choice in this vein is based on the quantile of the empirical distribution of the distances between the simulated data from the previous population, and the observed data (see Beaumont et al. 2009, Lenormand et al. 2011 and in a different way Del Moral et al. 2011, Drovandi and Pettitt 2011). The method determines ϵ_t at the beginning of the t -th round by sorting the distances $\{\Delta(x^{(i,t-1)}, x^*)\}_{1 \leq i \leq N}$ and setting ϵ_t such that α percent of the simulated data $\{x^{(i,t-1)}\}_{1 \leq i \leq N}$ are below it, for some predetermined α .

Here we show that a severe drawback of this quantile approach for threshold selection is that the final ABC posterior distribution $p_{\epsilon_t}(\theta | x^*)$ may end up being very different to the true posterior $p(\theta | x^*)$. In particular, if particles are sampled from a large region of parameter space that offers negligible or little support for the posterior distribution, there is a risk of getting stuck in this parameter region if the threshold is selected using a quantile method; we will refer to such a run of ABC SMC as having “failed.” As an illustration we consider a simple model where for each θ , the simulated data is $x = g(\theta) = (\theta - 10)^2 - 100 \exp(-100(\theta - 3)^2)$. Moreover, we suppose that the true data are generated using the parameter $\theta^* = 3$, and therefore the true posterior distribution is a Dirac function at $\theta = 3$. The support of the posterior distribution approximated by ABC SMC should then contain this parameter value.

Figure 1A represents the $L1$ distance between the simulated data $g(\theta)$ and $x^* = g(3)$ as a function of θ . In this example, for all the parameter space except in the interval $(2.92, 3.08)$ the distances are larger than 50. ABC SMC is used on this example, selecting the threshold sequences as the 0.8 quantile of the previous population's distances. The prior distribution is Gaussian with mean 10 and variance 10. Particles from successive populations are represented by the red dots in Figure 1A, with population t aligned along the horizontal line $y = d_t$, where d_t is the maximum distance between $g(\theta^{(i,t)})$ and $g(3)$ for all $1 \leq i \leq N$. For example, $\max_i |g(\theta^{(i,1)}) - g(3)|$ being equal to 150, the first population is represented by dots on the line corresponding to $y = 150$. We note that for all t the distributions $p_{\epsilon_t}(\theta | x^*)$ are centred around the local

Figure 1 Global and local minima. (A) Plot of the distance between simulated and true data for different parameter values. The model was designed to produce a severe global optimum at the true parameter value $\theta^*=3$, and a broad local optimum associated with distances no smaller than 50. The red dots arranged in horizontal lines are the members of successive ABC accepted populations. After the sixth population, parameters near the global optimum are no longer sampled. (B) Plots of the failure rates (grey), and total number of simulations (red) for a range of fixed quantile schedules. The red dots indicate the means over 100 ABC SMC runs, while the upper and lower whiskers are the maximum and minimum values respectively.

minimum at 10 and that the true parameter 3 has a very low probability under the final distribution, where the threshold has converged to a non-zero value. Here we assume that convergence has been reached when the threshold sequence has decreased below 10^{-4} , or alternatively when it fails to decrease by more than 10^{-2} for 3 consecutive rounds of ABC.

Repeating this inference for different values of α , leads to very different results (Figure 1B), with the failure rate increasing with the quantile value – for quantiles of 0.3 and higher, the failure rate is $>80\%$. This makes sense as higher quantile values the threshold reduces more slowly, and for each of the corresponding ABC populations it is unnecessary to sample from the global optimum in order to reach N accepted particles, and so an opportunity exists to either miss it entirely, or to sample it sparsely enough that it is lost during perturbation. Evidently, the choice of threshold schedule has implications not only for the efficiency of the inference, but also its success. Furthermore, in this example we find that the number of simulations needed for convergence of the sequence of epsilons, reduces as the quantile increases – that is, the computational expense of failure is lower than success. The optimal (or at least safe) choice of α clearly depends on the data, the model and the prior range. We feel that these problems highlight the potential issues arising in real-world applications.

The threshold–acceptance rate curve

The proposed method centres around understanding the TAR curve – the acceptance rate $\mathfrak{N}_t(\epsilon)$ as a function of the threshold ϵ for the next round of ABC simulations. We ask, if the TAR curve were known, how should ϵ be chosen in order to optimally balance computational efficiency with the need to minimise ϵ , and to avoid getting stuck in regions of parameter space that share little support with the posterior distribution?

To motivate our approach which is developed fully below, we revisit the simple model introduced above, and shown in Figure 1A. For this model and prior distribution, the TAR curve (estimated using an algorithm described below and shown in Figure 2) is sigmoidal with the position of the lower “elbow” (at $\epsilon \approx 50$) coinciding with the minimal distance obtainable from parameters in the local optimum region. Setting ϵ to this value will cause nearly all particles from the local optimum to be rejected, and force particles closer to the

Algorithm 1 ABC SMC algorithm

```

1:   input: a final threshold  $\epsilon_T$ , data  $x^*$ 
2:   output: a weighted sample of particles from  $p_{\epsilon_T}(\theta|x^*)$ 
3:   for all  $1 \leq t \leq T$  do
4:       determine the perturbation kernel  $K_t(\cdot|\cdot)$  and the next threshold  $\epsilon_t$ 
5:        $i \leftarrow 1$ 
6:       repeat
7:           if  $t=1$  then
8:               sample  $\tilde{\theta}$  from  $\pi(\theta)$ 
9:           else
10:              sample  $\theta$  from the previous population  $\{\theta^{(i,t-1)}\}_{1 \leq i \leq N}$  with weights  $\{\omega^{(i,t-1)}\}_{1 \leq i \leq N}$  sample  $\tilde{\theta}$ 
              from  $K_t(\cdot|\theta)$  and such that  $\pi(\tilde{\theta}) > 0$ 
11:          end if
12:          sample  $x$  from  $f(\cdot|\tilde{\theta})$ 
13:          if  $\Delta(x, x^*) \leq \epsilon_t$  then
14:               $\theta^{(i,t)} \leftarrow \tilde{\theta}$ 
15:               $i \leftarrow i+1$ 
16:          end if
17:      until  $i=N+1$ 
18:      calculate the weights: for all  $1 \leq i \leq N$ 
19:      if  $t \neq 1$  then
20:          
$$\omega^{(i,t)} \leftarrow \frac{\pi(\theta^{(i,t)})}{\sum_{j=1}^N \omega^{(j,t-1)} K_t(\theta^{(i,t)}|\theta^{(j,t-1)})}$$

21:      else  $\omega^{(i,1)} \leftarrow 1$ 
22:      end if
23:      normalize the weights
24:  end for

```

true posterior to be sampled. Indeed, using $\epsilon=50$, ABC SMC converges to a distribution about the true parameter value for every repetition of the inference. One such inference is shown in Figure 2B, where the second population is already restricted to a small interval about the true parameter value. While the fixed quantile methods can be similarly successful e.g., for 0.05, in practice this occurs only with the good fortune of choosing a quantile that selects a threshold appropriately with respect to some unknown critical value (here 50) sufficiently quickly.

More generally, we can see examples of TAR curves for a variety of models in Figure 3. Although the structure of the proposal distribution and likelihood surface could give rise to anything monotonic increasing, we tend to encounter three main types of curve (shown in Figure 3); concave, convex and sigmoidal. Further, for their interpretation it helps to consider each curve as a combination of convex and concave parts. A concave shape occurs over a particular range of threshold values, when the majority of particles drawn from the perturbed distribution (with distances relevant for this range) give rise to simulated data that is relatively close to the observed data. The opposite is true for convex shapes.

One way the latter can happen is when the perturbed distribution spans a region of parameter space that includes both a sharp global maximum and a broader local maximum of the likelihood. To see this, one can

Figure 2 Application of our method to the simple illustrative model with a local optimum. (A) The TAR curve for the toy model, with solid and dotted lines indicating the mean, maximum and minimum predicted values over 100 runs of our algorithm.

The red dot indicates the value $\epsilon^* = \arg\max_{\epsilon} \frac{\partial^2 \mathbb{N}_0(\epsilon)}{\partial \epsilon^2}$. (B) Starting from a population spread across both the global and local minimum, a threshold value of approximately 50 rejects all parameters except those situated in a small interval about the true parameter value of 3. Successive populations refine the distribution about this value.

imagine an ϵ -ball expanding about the true data; at first the ball only encompasses a small number of particles that were drawn from very close to the global maximum, corresponding to the low gradient at the foot of the shape. Once ϵ is large enough we are able to accept the relatively large number of particles sitting in the local maximum, which causes a rapid increase in the gradient of the TAR curve and results in convex behaviour. This is exactly the scenario described in the illustrative example above where the ABC SMC algorithm

Figure 3 (A) Predicted TAR curves for different models (clockwise from top left); a Gaussian p.d.f, a quadratic function, the Repressilator, and the smallest biochemical system exhibiting a Hopf bifurcation. Red dots indicate the actual acceptance rate of an ABC run with corresponding threshold values. (B) Typical TAR curve shapes. Red dots indicate the threshold values considered by our method – either at the extreme of a convex section to avoid a possible local optimum, or the point $\arg\min_{\epsilon} \Delta \left(\left(\frac{\epsilon}{\epsilon_{t-1}}, \frac{\mathbb{N}_t(\epsilon)}{\mathbb{N}_t(\epsilon_{t-1})} \right), (0,1) \right)$ that balances reduction of ϵ with computational expense. In the case where multiple threshold values satisfy the latter condition, we select the smallest one.

is seen to fail when using various quantile strategies to select the threshold values (Figures 1 and 2), and is likely to be a common occurrence in biological systems where likelihood surfaces are known to be highly complex (Gutenkunst et al., 2007; Erguler and Stumpf, 2011). Indeed below we present such an example involving the smallest possible biochemical system that can exhibit an oscillation inducing Hopf bifurcation (Wilhelm and Heinrich, 1995; Kirk et al., 2008).

This interpretation of convex shapes as a symptom of sampling from local optima in the posterior suggests the following criterion for threshold selection: the ABC SMC algorithm can get stuck when a threshold schedule allows particles from a relatively broad local optimum to be accepted too frequently and for successive populations – at each round of the algorithm we risk not sampling from the correct posterior (i.e., that part of the posterior that account for “almost all” of the probability mass of the posterior) at all; diffusion of the parameter particles to the incorrect area in parameter space is thus entropically driven. This can be avoided by choosing a threshold that rejects these particles with high probability. From the TAR curve we

can identify such a threshold value as one that lies at the bottom of the steep incline, i.e., $\operatorname{argmax}_{\epsilon} \frac{\partial^2 \mathfrak{K}_t(\epsilon)}{\partial \epsilon^2}$.

For concave shapes the same danger is not apparent from the curve, and we can instead try to balance computational expense against the desired reduction in ϵ . Here we treat the TAR curve in the same spirit as we would treat a ROC curve, identifying the optimal “cut-point” as that which minimises the distance between

$\left(\frac{\epsilon}{\epsilon_{t-1}}, \frac{\mathfrak{K}_t(\epsilon)}{\mathfrak{K}_t(\epsilon_{t-1})} \right)$ and $(0, 1)$. Similar thresholds can be defined when the relative tradeoffs between the need to reduce the threshold and computational expense are weighed differently.

We can now state our proposed threshold selection method given the TAR curve:

1. If $t > 0$, define $d_{\min}$ to be the minimum distance produced by past simulations.
2. Define $\epsilon^* = \operatorname{argmax}_{\epsilon} \frac{\partial^2 \mathfrak{K}_t(\epsilon)}{\partial \epsilon^2}$.
3. If $\mathfrak{K}_t(\epsilon^*) > \delta$ or $\epsilon^* > d_{\min}$ (for the case $t > 0$), set $\epsilon_t = \epsilon^*$. (*Detection of a possible local minimum*)
4. If $\mathfrak{K}_t(\epsilon_t) \leq \delta$ and $\epsilon^* \leq d_{\min}$, choose $\epsilon_t = \operatorname{argmin}_{\epsilon} \Delta \left(\left(\frac{\epsilon}{\epsilon_{t-1}}, \frac{\mathfrak{K}_t(\epsilon)}{\mathfrak{K}_t(\epsilon_{t-1})} \right), (0, 1) \right)$. (*Reduction of threshold v.s. Computational expense*)

where δ is a small tolerance in the UT acceptance-rate estimation, to avoid choosing thresholds with zero acceptance rates. Of course, the shape of the TAR curve is in general unknown but below we show that it is often possible to obtain useful predictions of this curve which allows us to “guess” near-optimal thresholds.

Estimating the threshold–acceptance rate curve

The threshold selection scheme described above relies upon knowledge of the TAR curve. In this section we suggest a computationally inexpensive way in which it can be approximated in the case of ODE models.

We first define formally the acceptance rate of the algorithm for round $t > 0$ and any threshold, ϵ by,

$$\mathfrak{K}_t(\epsilon) := \int p_t(x) \mathbb{1}(\Delta(x, x^*) \leq \epsilon) dx \quad (1)$$

where $p_t(x)$ is the distribution of the simulated data corresponding to parameters sampled from the last population and perturbed by the kernel K_t , i.e.,

$$p_t(x) = \int q_t(\theta) f(x|\theta) d\theta$$

with the perturbed distribution, $q_t(\theta)$, defined via

$$q_t(\theta) := \sum_{i=1}^N \omega^{(i,t-1)} K_t(\theta | \theta^{(i,t-1)}).$$

A simple way to estimate $p_t(x)$ would be via Monte Carlo approximation, i.e., simulating data from a large sample drawn from $q_t(\theta)$. However, the expense of such a naive approach is generally prohibitive. Here we use the so-called unscented transform to approximate the distribution $p_t(x)$ of the model output given the distribution $q_t(\theta)$. The unscented transform (UT) (Julier et al. 2000), tells us how the moments of a random variable, θ , are transformed by a non-linear function, g . Its computational efficiency and flexibility to the form of the non-linear function, has led to its extensive use in filtering (Wan and van der Merwe, 2000) and smoothing (Briers et al., 2010) algorithms, and its increasing popularity as a tool for parameter inference (Quach et al., 2007; Liu and Niranjana, 2012) and uncertainty propagation (Giza et al., 2009). From here we will model our data as a non-linear transformation, g , of the parameter θ with an additive zero-mean noise term. However, the method can be extended to stochastic state space models, with some limitations on the form of the observation model, or without these limitations in an ABC filtering framework.

The unscented transform requires that the perturbed distribution, $q_t(\theta)$, is decomposed into a mixture of Gaussians,

$$q_t(\theta) \approx \sum_i \alpha_i p_i(\theta)$$

with each p_i being a Gaussian density that can be fit by an EM algorithm. The choice of number of components will be discussed below.

For each component of the mixture, we use the UT to approximate,

$$\int f(x|\theta) p_i(\theta) d\theta,$$

as follows. The first step in the UT algorithm is to determine a set of weighted particles (called sigma-points) with the same sample moments up to a desired order as the distribution $p_i(\theta)$. Here we use a scaled sigma-point set $\{\chi_k\}_{k=0, \dots, 2L}$ that captures both means and covariances (Julier, 2002).

$$\begin{aligned} \chi_0 &= \mu_\theta \\ \chi_k &= \mu_\theta + \left[\sqrt{(L+\lambda)\Sigma_\theta} \right]_k \quad k=1, \dots, L \\ \chi_k &= \mu_\theta - \left[\sqrt{(L+\lambda)\Sigma_\theta} \right]_k \quad k=L+1, \dots, 2L \end{aligned}$$

where L is the dimension of θ , μ_θ and Σ_θ are the mean and covariance of $\theta \sim p_i(\cdot)$, $[A]_k$ represents the k th column of a matrix A , and

$$\lambda = \alpha^2(L+k) - L.$$

The sigma-point weights $\{v_k^c, v_k^m\}_{k=0, \dots, 2L}$ are given by,

$$\begin{aligned} v_0^m &= \frac{\lambda}{L+\lambda} \\ v_0^c &= \frac{\lambda}{L+\lambda} + (1-\alpha^2+\beta) \\ v_k^m &= v_k^c = \frac{1}{2(L+\lambda)} \quad k=1, \dots, 2L. \end{aligned}$$

and finally, the parameters κ , α and β may be chosen to control the positive definiteness of covariance matrices, spread of the sigma-points, and error in the kurtosis respectively. While sigma-point selection schemes exist for higher moments (Julier, 1998; Tenne and Singh, 2003) they come with significantly increased computational cost.

Once determined, each sigma-point is propagated individually through the function, g , and the mean and covariance of the transformed variable, $g(\theta)$, can be estimated using the update equations,

$$\mu_{g(\theta)} \approx \sum_{k=0}^{2L} v_k^m g(\chi_k) \quad (2)$$

$$\Sigma_{g(\theta)} \approx \sum_{k=0}^{2L} v_k^c (g(\chi_k) - \mu_{g(\theta)})(g(\chi_k) - \mu_{g(\theta)})^T. \quad (3)$$

We denote the resulting approximate probability density function for $g(\theta)$, for mixture component i by $\mathcal{U}_{p_i}(x)$.

By matching terms in the Taylor expansions of the estimated and true values of these moments, it can be shown that the above algorithm is accurate to second order in the expansion. More generally, if the sigma-point set approximates the moments of θ up to the n th order then the estimates of the mean and covariance of $g(\theta)$ will be accurate up to the n th term (Julier et al., 2000). Crucially, the number of points required ($2L+1$ for this scheme) is much smaller than the number required to reach convergence with Monte-Carlo methods.

Given the $\mathcal{U}_{p_i}(x)$, for each mixture component, we can approximate the distribution of the output x as follows,

$$p_t(x) \approx \sum_i \alpha_i \int f(x|\theta) p_i(\theta) d\theta \quad (4)$$

$$\approx \sum_i \alpha_i \mathcal{U}_{p_i}(x). \quad (5)$$

Samples $\{x_j\}_{j=0, \dots, M}$ from the mixture of Gaussians distribution in equation (5) may then be used as an inexpensive proxy for ABC simulations, and the acceptance rates can be estimated as,

$$\mathfrak{R}_t(\epsilon) \approx \frac{1}{M} \sum_{j=1}^M H_\epsilon(\Delta(x_j, x^*)) \quad (6)$$

where,

$$H_\epsilon(\Delta(x_j, x^*)) = \frac{1}{\left(1 + e^{-k \left(\frac{\Delta(y, y^*)}{\epsilon} - 1\right)}\right)} \quad (7)$$

is used as a smooth approximation to the “accept and reject” indicator function, with k controlling the severity of the step. The smooth approximation is necessary for estimating the critical value,

$$\frac{\partial^2 \mathfrak{R}_t(\epsilon)}{\partial \epsilon^2} = \frac{1}{M} \sum_{j=1}^M \frac{\partial^2 H(\Delta(x_j, x^*))}{\partial \epsilon^2}$$

of the proposed threshold selection scheme.

In summary, the ABC SMC acceptance rate may be approximated for any threshold value at the beginning of each round $t > 0$ of the algorithm using the steps:

1. generate a population of perturbed particles, sampling from $\{\theta^{(i, t-1)}, \omega^{(i, t-1)}\}_{1 \leq i \leq N}$ and perturbing each particle independently with K_t ,
2. fit a Gaussian mixture model to the perturbed population,
3. estimate $p_t(x)$ using the unscented transform independently for each component p_i of the Gaussian mixture,
4. estimate acceptance rates for different threshold values according to equation (6) sampling from $\sum_i \alpha_i \mathcal{U}_{p_i}(x)$.

Applications to biological models

We now contrast our adaptive method to various fixed quantile threshold schedules in the context of two biological dynamical systems. We consider two criteria: firstly, the total number of simulations required to reach a pre-chosen threshold value; and secondly, the proportion of repeat runs that fail, i.e., get stuck in a local minimum, or fail to reach the threshold in a given (very long) period of time.

Computational expense: quantiles vs. UT

The repressilator has become a classic example of a synthetic biological oscillator (Elowitz and Leibler, 2000). It consists of six species [three mRNAs, (m_i), and their protein products, (p_i)], with regulatory links between them forming a single feedback loop – each protein inhibits the production of the next protein's mRNA. The dynamics of the species concentrations are governed by the first order differential equations,

$$\frac{dm_1}{dt} = -m_1 + \frac{\alpha}{1+p_3^n} + \alpha_0, \quad (8)$$

$$\frac{dp_1}{dt} = -\beta(p_1 - m_1), \quad (9)$$

$$\frac{dm_2}{dt} = -m_2 + \frac{\alpha}{1+p_1^n} + \alpha_0, \quad (10)$$

$$\frac{dp_2}{dt} = -\beta(p_2 - m_2), \quad (11)$$

$$\frac{dm_3}{dt} = -m_3 + \frac{\alpha}{1+p_2^n} + \alpha_0, \quad (12)$$

$$\frac{dp_3}{dt} = -\beta(p_3 - m_3), \quad (13)$$

where $\theta = (n, \beta, \alpha, \alpha_0)$ is the parameter vector to be inferred. We set the initial species concentrations to $(m_1, p_1, m_2, p_2, m_3, p_3) = (0.0, 2.0, 0.0, 1.0, 0.0, 3.0)$, and generate some oscillatory data by simulating the model with $\theta = (2.0, 4.0, 100.0, 1.0)$, and “observing” the state of p_1 at time-points $(4.0, 8.0, \dots, 20.0)$, subject to some small added zero-mean Gaussian noise with covariance $0.01I$. With Gaussian prior distributions that encompass the true parameter values, we perform ABC SMC, choosing thresholds according to our method and a range of fixed quantile threshold schedules. The inferences are repeated 10 times for each method, and stopped once a round of ABC SMC has been completed with a threshold value below a pre-chosen challenging threshold – in this case, 35, for which accepted simulations are at least in qualitative agreement with the oscillatory data.

A comparison of the performance of each method is shown in Figure 4A, along with the predicted TAR curves and selected threshold values for the adaptive method. The computational expense of the quantile methods is found to vary significantly and in non-linear fashion, with the best performing quantile (0.3) over three times cheaper than the worst (0.9). This highlights the difficulty of choosing fixed threshold values that perform well. Our method scores similarly to the best fixed schedules (at approximately 4000 simulations), which suggests that it is successfully reducing the computational cost for this inference.

Next we explore the effect of varying the number of components used for the Gaussian mixture approximation. Figure 5A shows the predicted TAR curves and selected thresholds for the first round of ABC SMC, using between 1 and 250 components. Although it is possible to conceive of situations where the opposite would hold, in this example, the use of fewer components is associated with under-estimation of $\mathfrak{K}_0(\epsilon)$, and the point, $\operatorname{argmax}_{\epsilon} \frac{\partial^2 \mathfrak{K}_0(\epsilon)}{\partial \epsilon^2}$. For >50 components, the selected thresholds seem relatively stable at $\epsilon \approx 100$ for this example.

In general, we can say that the greater the number of components used, the more accurate the acceptance rate approximations become, and that this is because we are not only trying to fit the input distributions but allow enough flexibility to approximate a possibly complex, multi-modal output. Indeed we find that increasing degrees of non-linearity in g requires more Gaussian components in order to keep the accuracy at the same level (see Figure 6). However, using more components has the obvious downside of increasing the associated computational burden (Figure 5B). Despite the results achieved here, using 100 components for each system when comparing to fixed quantile selection schemes, an improvement to the method would be to update the number of components automatically with respect to the model behaviour. Similarly to how Gaussian mixtures can be adaptively chosen in particle based simulation of Liouville-type equations (Horenko and Weiser, 2003; Weiße et al., 2006), extra components could be introduced at the next round of ABC SMC in close regions of the perturbed distribution, $q_t(\theta)$, that are found to map to distant areas in output space.

Oscillations and local minima

The model below represents the simplest biochemical reaction system that permits a Hopf bifurcation (Wilhelm and Heinrich, 1995; Kirk et al., 2008). It can be shown that this system, described by,

$$\begin{aligned}\frac{dx}{dt} &= (Ak_1 - k_4)x - k_2xy, \\ \frac{dy}{dt} &= -k_3y + k_5z, \\ \frac{dz}{dt} &= k_4x - k_5z,\end{aligned}$$

Figure 4 (A) Box plots showing the total number of simulations required to reach a threshold value of 35 (with the L2 distance function), for the Repressilator model and different threshold schedules. An order of magnitude difference exists in the computational expense of the best and worst performing quantile methods. Our adaptive method performs comparably to the best fixed quantile schedules. (B) TAR curves and threshold values selected by the adaptive method, with the greyscale (from light to dark) indicating which round of ABC SMC the curve is predicting. The curves change distinctly as the prior mass shifts towards the posterior.

Figure 5 (A) Plot showing how the predicted TAR curve changes as the number of mixture components is varied between 1 and 250. Each curve is for the first round of ABC SMC, with the chosen threshold value marked with a cross. Mixtures with lower numbers of components tend to under-estimate $\mathfrak{X}_0(\epsilon)$, and the point, $\operatorname{argmax}_{\epsilon} \frac{\partial^2 \mathfrak{X}_0(\epsilon)}{\partial \epsilon^2}$. (B) Plot showing the percentage of the Repressilator inference running time used to select the optimal threshold values, for different numbers of mixture components. As expected, the higher the number of components used, the greater the computational cost incurred. Error bars show 1 standard deviation in each direction, and are calculated over 10 repetitions.

where, x, y, z , represent the concentrations of three reactants, k_i , are the reaction rates, and, A , is the fixed concentration of a fourth reactant, displays a limit cycle for $Ak_1 = k_3 + k_4 + k_5$. Further, when the true value of Ak_1 is greater than the critical value, $k_3 + k_4 + k_5$, the bifurcation has an effect on the likelihood of producing a global maximum and broader local maximum, with the regions becoming more defined for larger data sets (Kirk et al., 2008). This is illustrated in the legend of Figure 7, where the squared distance between simulations (for different values of Ak_1) and the data are shown for data sets of size $T = (100, 200, \dots, 500)$, with fixed $k_i = 1$ and the true value of Ak_1 set at 5.5. Values of Ak_1 below the bifurcation point ($Ak_1 = 3$) correspond to distances that,

Figure 6 Prediction errors. Plot showing the mean squared error in predicting acceptance rates over 10 values of ϵ for different numbers of mixture components. Results for simulation models ($g(\theta) = \theta^k$, $k = 1, \dots, 6$) of varying degrees of non-linearity are shown in different colours. Each prediction for each combination of model and number of mixture components was evaluated 100 times, with error bars showing 1 standard deviation in each direction from the mean error value. In general, for <15 components, the higher the degree of non-linearity, the greater the inaccuracy in the predicted TAR curves.

Figure 7 Plots showing the performance of our adaptive method and different fixed quantile schedules for parameter inference on the Hopf bifurcating system. For each data set and method, 10 inferences were performed. Colours indicate the number of data points used for the inference, and crosses mark failures of some inferences to reach a fixed threshold value (scaled according to the number of points used), within 100,000 simulations. The percentage of runs that fail is also given where it is non-zero. The legend also shows the ℓ^2 distance between simulations and data for each dataset, with the trough around the true parameter value narrowing as the set size increases. Note that the computational expense varies least across datasets for our method, which also suffers no failures. In comparison, the 0.9 quantile schedule always fails when using 500 data points.

in the case of larger T , are smaller than those for values of Ak_1 above the bifurcation point that do not belong to a small interval about the true value.

We repeat ABC SMC inferences 10 times for each of the data sets and for each threshold choice scheme. The total number of simulations needed to reach a target threshold value of $\sqrt{80T}$ (scaled according to the size of the data set), is recorded unless this grows above 100,000, in which case the inference is considered to have failed. By varying T we are able to examine the adaptability of our threshold choice method to different likelihood shapes and, moreover, perform a “stress test” by controlling how challenging it is to avoid the local maximum.

Results comparing our strategy to various fixed quantile schedules are shown in Figure 7. We find that in all cases, the expense of our method is comparable or cheaper than the best performing fixed quantile schedule. Furthermore, the variability in cost between different data sets is smallest for our method. Moreover, our method is successful in all cases, while the fixed quantile schedules (with the exception of 0.3) suffer failures for $T \geq 300$; in one case (for $\alpha=0.9$) this happens for every repetition. For the larger quantiles, these failures are caused by the accepted population becoming trapped in the interval (0.3), while for the 0.01 quantile, the reduction in threshold can be too severe, which leads to a very low acceptance rate and computationally overly expensive inference.

Discussion

Reducing the computational expense of sequential ABC algorithms is an important goal in real-world applications of ABC. But while it is clear that careful threshold selection is part of the solution, we have shown

here that current methods with efficiency as their sole objective, can risk the validity of the resulting posterior approximations. Specifically, we have found that gentle reductions in the threshold, ϵ_t , between populations have a tendency to lead populations of particles to diffuse into broad regions of low posterior probability. This should be a compelling argument against the popular quantile-based adaptive schemes, in particular those where the computational cost is fixed, i.e., where the number of simulations is specified and a fixed fraction of the simulated particles with the smallest distances are used to make up the intermediate population. The attraction of fixed (or controlled) computational burden comes with the high risk of convergence to spurious and biased “posterior” distributions.

Knowing the relationship between the acceptance rate and the ϵ threshold schedule applied in ABC SMC has obvious implications for the efficiency and computational affordability of ABC inference in complex inference tasks. In addition to this, we have shown how convex shapes within the TAR curve can indicate the presence of local optima and help identify threshold values that avoid them. It is worth noting that this requires the parameter space to be sufficiently explored; if particles are never sampled or perturbed outside of the local minimum, then there is no way to know that lower distances are achievable. For deterministic and stochastic state space models (the latter with some restrictions upon the form of the observation function), the UT can be employed to predict the TAR curve, and select such safe and (near-)optimal threshold value. The accuracy of TAR predictions was found to be dependant upon both the number of Gaussian mixture components and their placement, with 100 components fit via an EM algorithm more than sufficient for the examples given here. However, it is possible to imagine even less well-behaved mappings between parameter and state space for which the placement of the components must be more carefully considered. Note that the UT is an ancillary or supporting statistical inference step which allows us to fine-tune the technical parameters of the inference process; in no way does this interfere with conventional Bayesian practice or conventions.

In this paper, we have focused on the sequential ABC algorithm proposed by Toni et al. (2009) and Beaumont et al. (2009), but our approach can also be used for other ABC schemes based on SMC such as the algorithms proposed by Drovandi and Pettitt (2011) and Del Moral et al. (2011). Essentially, we have shown that algorithms such as ABC SMC are subject to the same challenges as other local inference and optimisation methods; that inferred solutions can be related to a local optimum of the likelihood or cost function. Such problems are more widely recognised in ABC MCMC (Marjoram et al., 2003), where the inference is based on the trajectory of a single particle as it explores parameter space. This exploration is driven by an iterated two step procedure: the particle is first perturbed, and then accepted or rejected depending (among other factors) on the distance between the observed data and the data simulated conditionally on this perturbed particle. In the original version of ABC MCMC, the particle is rejected if the distance is larger than a fixed threshold; an extension of this algorithm uses a tempering schedule where the threshold value is decreased at each iteration of the algorithm (Ratmann et al., 2007). Our method could be extended to the choice of the tempering schedule in this scheme, which has been observed to strongly affect the convergence of this ABC MCMC algorithm (Ratmann et al., 2007).

We have here focussed on dynamical systems where the observed data are compared directly to the simulations, rather than summary statistics of real and observed data. Extension of this approach to inference using single summaries (Joyce and Marjoram, 2008; Nunes and Balding, 2010; Barnes et al., 2012; Fearnhead and Prangle, 2012) is in principle straightforward as the UT only aims to predict the shape of the TAR curve. When multiple summaries are used that may not be aggregated into a single statistic the same method could still be used to predict the higher dimensional TAR “curve.” How to interpret such an object and identify safe and efficient threshold choices is a topic of future research. The remaining parts of the algorithm are not affected by this in principle, although in practice the Gaussian mixture model and other factors impacting the efficiency and accuracy of the predictions may need to be considered carefully.

It has to be kept in mind that at the moment we adopt a greedy procedure and predict only the next threshold. Providing a global choice of the ϵ_t threshold for all t is a much harder inference task. Although the overall number of simulations in the ABC SMC scheme (once all ϵ_t are determined) may be less than the number of simulations required by our greedy approach, we believe that the computational burden and complications inherent in determining global schedules are prohibitive. We conclude by noting that the choice of

an optimal ϵ_t schedule will depend on the choice of the kernel as well as the model at hand. In order to derive maximum benefit from ABC approaches in the context of challenging real-world problems serious consideration needs to be given to their computational performance, as well as to their theoretical validity.

Acknowledgements: We thank Oliver Ratmann for stimulating discussions on the use of the presented method with different ABC schemes, and we are grateful for the helpful comments received from two anonymous referees. DS is supported by BBSRC grant BB/K003909/1; SF is funded through an MRC Computational Biology Research Fellowship; MPHS is a Royal Society Wolfson Merit Award holder.

References

- Barnes, C., S. Filippi, M. P. H. Stumpf and T. Thorne (2012): "Considerate approaches to achieving sufficiency for ABC model selection," *Stat. Comput.*, 22, 1181–1197.
- Beaumont, M. A., J. M. Cornuet, J. M. Marin and C. P. Robert (2009): "Adaptive approximate Bayesian computation," *Biometrika*, 96, 983–990.
- Blum, M. G. B. and O. François (2010): "Non-linear regression models for Approximate Bayesian Computation," *Stat. Comput.*, 20, 63–73.
- Briers, M., A. Doucet and S. Maskell (2010): "Smoothing algorithms for state-space models," *Ann. Inst. Stat. Math.*, 62, 61–89.
- Del Moral, P., A. Doucet and A. Jasra (2011): "An adaptive sequential Monte Carlo method for approximate Bayesian computation," *Stat. Comput.*, 22, 1009–1020.
- Drovandi, C. C. and A. N. Pettitt (2011): "Estimation of parameters for macroparasite population evolution using approximate Bayesian computation," *Biometrics*, 67, 225–233.
- Elowitz, M. B. and S. Leibler (2000): "A synthetic oscillatory network of transcriptional regulators," *Nature*, 403, 335–338.
- Erguler, K. and M. P. H. Stumpf (2011): "Practical limits for reverse engineering of dynamical systems: a statistical analysis of sensitivity and parameter inferability in systems biology models," *Mol. BioSystems*, 7, 1593–1602.
- Fagundes, N., et al. (2007): "Statistical evaluation of alternative models of human evolution," *Proc. Natl. Acad. Sci.*, 104, 17614.
- Fearnhead, P. and D. Prangle (2012): "Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation," *J. Roy. Stat. Soc. B (Statistical Methodology)*, 74, 419–474.
- Filippi, S., C. Barnes, J. Cornebise and M. Stumpf (2013): "On optimality of kernels for approximate Bayesian computation using sequential Mnote Caril," *Stat. Appl. Genet. Mol. Biol.*, 12, 87–107.
- Giza, D., P. Singla and M. Jah (2009): An approach for nonlinear uncertainty propagation: Application to orbital mechanics. In: *AIAA Guidance, Navigation, and Control Conference*, Chicago, IL.
- Gutenkunst, R., et al. (2007): "Universally sloppy parameter sensitivities in systems biology models," *PLoS Comput. Biol.*, 3, 1871–1878.
- Horenko, I. and M. Weiser (2003): "Adaptive integration of molecular dynamics," *J. Comput. Chem.*, 24, 1921–1929.
- Joyce, P. and P. Marjoram (2008): "Approximately sufficient statistics and Bayesian computation," *Stat. Appl. Genet. Mol. Biol.*, 7, Article26.
- Julier, S. J. (1998): Skewed approach to filtering. In *Aerospace/Defense Sensing and Controls* 271–282. (International Society for Optics and Photonics).
- Julier, S. (2002): The scaled unscented transformation. *American Control Conference*.
- Julier, S., J. Uhlmann and H. Durrant-Whyte (2000): "A new method for the nonlinear transformation of means and covariances in filters and estimators," *IEEE Trans. Automatic Control*, 45, 477–482.
- Kirk, P., T. Toni and M. Stumpf (2008): "Parameter inference for biochemical systems that undergo a Hopf bifurcation," *Biophys. J.*, 95, 540–549.
- Lenormand, M., F. Jabot and G. Deffuant (2011): "Adaptive approximate Bayesian computation for complex models," *arXiv preprint arXiv:1111.1308*.
- Liu, X. and M. Niranjan (2012): "State and parameter estimation of the heat shock response system using Kalman and particle filters," *Bioinformatics*, 28, 1501–1507.
- Marjoram, P. and J. Molitor (2003): "Markov chain Monte Carlo without likelihoods," *Proc. Natl. Acad. Sci. USA*, 100, 15324.
- Marjoram, P., J. Molitor, V. Plagnol and S. Tavaré (2003): "Markov chain Monte Carlo without likelihoods," *Proc. Natl. Acad. Sci.*, 100, 15324–15328.
- Nunes, M. A. and D. J. Balding (2010): "On optimal selection of summary statistics for approximate Bayesian computation," *Stat. Appl. Genet. Mol. Biol.*, 9, 1.
- Quach, M., N. Brunel and F. D'Alché-Buc (2007): "Estimating parameters and hidden variables in non-linear state-space models based on ODEs for biological networks inference," *Bioinformatics*, 23, 3209–3216.
- Ratmann, O., et al. (2007): "Using likelihood-free inference to compare evolutionary dynamics of the protein networks of *H. pylori* and *P. falciparum*," *PLoS Comput. Biol.*, 3, e230.

- Sisson, S. A., Y. Fan and M. M. Tanaka (2007): "Sequential Monte Carlo without likelihoods," *Proc. Natl. Acad. Sci. USA*, 104, 1760–1765.
- Tallmon, D., G. Luikart and M. Beaumont (2004): "Comparative evaluation of a new effective population size estimator based on approximate Bayesian computation," *Genetics* 167, 977–988.
- Tenne, D. and T. Singh (2003): The higher order unscented filter. In: *American Control Conference, 2003. Proceedings of the 2003*, vol. 3, 2441–2446 (IEEE, 2003).
- Toni, T., D. Welch, N. Strelkowa, A. Ipsen and M. P. H. Stumpf (2009): "Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems," *J. Roy. Soc. Interf. Roy. Soc.*, 6, 187–202.
- Wan, E. and R. van der Merwe (2000): "The unscented Kalman filter for nonlinear estimation." *Adaptive Systems for Signal Processing, Communications, and Control Symposium 2000. AS-SPCC. The IEEE 2000* 153–158.
- Weiß, A., I. Horenko and W. Huisinga (2006): "Adaptive approach for modelling variability in pharmacokinetics," *Comput. Life Sci. II*, 194–204.
- Wilhelm, T. and R. Heinrich (1995): "Smallest chemical reaction system with Hopf bifurcation," *J. Math. Chem.*, Springer, 17, 1–14.