Supplementary material: Derivation of ABC SMC model selection algorithms

Tina Toni, Michael P. H. Stumpf

We start this section by briefly reviewing the building blocks of the ABC SMC algorithm of Toni *et al.* [1], which is based on sequential importance sampling (SIS). The main idea of importance sampling is to sample from the desired target distribution π (which can be impossible or hard to sample from) indirectly through sampling from a proposal distribution η [2]. To get a sample from π, one can instead sample from η and weight the samples by importance weights

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

In SIS one reaches the target distribution π_T through a series of intermediate distributions, π_t, $t = 1, \ldots, T-1$ [3,4]. If it is hard to sample from these distributions one can use the idea of importance sampling described above to sample from a series of proposal distributions η_t and weight the obtained samples by importance weights

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

In SIS the proposal distributions are defined as

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

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

To apply SIS, we need to define the intermediate and the proposal distributions. In an ABC framework [5,6], which is based on comparisons between simulated and experimental datasets, we define the intermediate distributions as [1,7]

$$\pi_t(x) = \frac{P(x)}{B_t} \sum_{b=1}^{B_t} \mathbb{1}\left(d(D_0, D_{(b)}(x)) \leq \epsilon_t\right),$$

where $P(x)$ denotes the prior distribution and $D_{(1)}, \ldots, D_{(B_t)}$ are $B_t \geq 1$ data sets generated for a fixed parameter x, $D_{(b)} \sim f(D|x)$. $\mathbb{1}(x)$ is an indicator function and ϵ_t is the tolerance required from particles contributing to the intermediate distribution t. To simplify the notation we define $b_t(x) = \frac{1}{B_t} \sum_{b=1}^{B_t} \mathbb{1}\left(d(D_0, D_{(b)}(x)) \leq \epsilon_t\right)$.

We define the first proposal distribution to equal the prior distribution, $\eta_1(x) = P(x)$. The proposal distribution at time t ($t = 2, \ldots, T$), η_t, is defined as

$$\eta_t(x_t) \quad = \quad \mathbb{1}\left(P(x_t) > 0\right) \mathbb{1}\left(b_t(x_t) > 0\right) \int \pi_{t-1}(x_{t-1}) K_t(x_t|x_{t-1}) dx_{t-1}, \qquad (3)$$

where K_t denotes the perturbation kernel (e.g. random walk around the particle). For details of how this proposal distribution was obtained, see [1].

In the remainder of this section we introduce three different ways in which ABC SMC ideas presented above can be used in the model selection framework. We start by proposing a simple and naive incorporation of the above building blocks for model selection. We then continue by deriving an ABC SMC model selection algorithm on the joint model and parameter space, which is presented in the methods section of the paper. In the end we present ABC SMC algorithm for approximation of the marginal likelihood, which can also be employed for model selection.

The only of these three algorithms that we present in the main part of the paper and use in examples is algorithm II (ABC SMC model selection on the joint space), since the other two algorithms (I and III) are computationally too expensive and impractical to use.

I) *ABC SMC$_m$ REJ$_\theta$ model selection algorithm*

Very naively and stragihtforwardly the intermediate distributions can be defined as

$$\pi_t(m) = P(m) bm_t(m),$$

where

$$bm_t(m) := \frac{\sum_{\theta \sim P(\theta|m)} \mathbb{1}(d(D_0, D(\theta, m)) < \epsilon_t)}{\sum_{\theta \sim P(\theta|m)} \mathbb{1}(P(\theta|m) > 0)}.$$

For each model m we calculate $bm_t(m)$ as the ratio between the number of accepted particles (where the distance falls below ϵ_t) and all sampled particles, where parameters θ of model m are sampled from the prior distribution $P(\theta|m)$.

If a set of candidate models $\mathcal{M}$ of a finite size $|\mathcal{M}|$ is being considered, and N denotes the number of particles, then we can write the algorithm as follows:

MS1 Initialize $\epsilon_1, \ldots, \epsilon_T$.
Set the population indicator $t = 1$.

MS2 For $i = 1, \ldots, |\mathcal{M}|$, calculate the weights as

$$w_t^{(i)}(m_t^{(i)}) = \begin{cases} bm_t(m_t^{(i)}), & \text{if } t = 1 \\ \frac{P(m_t^{(i)}) bm_t(m_t^{(i)})}{\sum_{j=1}^{N} w_{t-1}^{(j)} KM_t(m_t^{(i)}|m_{t-1}^{(j)})}, & \text{if } t > 1. \end{cases}$$

MS3 Normalize the weights.
If $t < T$, set $t = t + 1$, go to **MS2**.

In this algorithm we estimate the posterior distribution of the model indicator sequentially, but the integration over model parameters is not sequential; we always sample them from the prior distribution $P(\theta|m)$. This algorithm is therefore computationally very expensive. It would be computationally more efficient to generate θ_t by exploiting the knowledge about θ that is contained in $\{\theta\}_{t-1}$. In addition to learning m sequentially, i.e. by exploiting $\{m\}_{t-1}$ for generating m_t, we would also like to learn θ sequentially.

In order to do this, we define

II) *ABC SMC model selection on the joint space*

Let (m, θ) denote a particle from a joint space, where m corresponds to the model indicator and θ are the parameters of model m. We define the intermediate distributions by

$$\pi_t(m, \theta) = P(m, \theta) b_t(m, \theta),$$

where

$$b_t(m, \theta) = \frac{1}{B_t} \sum_{b=1}^{B_t} \mathbb{1}(d(D_0, D_{(b)}(m, \theta)) \leq \epsilon_t).$$

In the following equations KM_t denotes the perturbation kernel for the model parameter, $KP_{t,m}$ denotes the perturbation kernel for the parameters of model m, and t is the population number. Now we derive the sequential importance sampling weights

$$w_t(m_t, \theta_t) = \frac{\pi_t(m_t, \theta_t)}{\eta_t(m_t, \theta_t)}.$$

For a particle (m_t, θ_t) from population t, we define the proposal distribution $\eta_t(m_t, \theta_t)$

as

$$\begin{aligned}
\eta_t(m_t,\theta_t) &= \mathbb{1}\left(P(m_t,\theta_t)>0\right)\mathbb{1}\left(b_t(m_t,\theta_t)>0\right) \\
&\quad \times \int_{m_{t-1}} \pi_{t-1}(m_{t-1}) KM_t(m_t|m_{t-1}) dm_{t-1} \\
&\quad \times \int_{\theta_{t-1}|m_{t-1}=m_t} \pi_{t-1}(\theta_{t-1}) KP_t(\theta_t|\theta_{t-1}) d\theta_{t-1} \\
&\propto \mathbb{1}\left(P(m_t,\theta_t)>0\right)\mathbb{1}\left(b_t(m_t,\theta_t)>0\right) \\
&\quad \times \sum_{j=1}^{|\mathcal{M}|} P_{t-1}(m_{t-1}^{(j)}) KM_t(m_t|m_{t-1}^{(j)}) \\
&\quad \times \sum_{k;m_{t-1}=m_t} \frac{w_{t-1}^{(k)}}{P_{t-1}(m_{t-1}=m_t)} KP_{t,m_t}(\theta_t|\theta_{t-1}^{(k)}) \\
&\propto \mathbb{1}\left(P(m_t,\theta_t)>0\right)\mathbb{1}\left(b_t(m_t,\theta_t)>0\right) \\
&\quad \times \sum_{j=1}^{|\mathcal{M}|} P_{t-1}(m_{t-1}^{(j)}) KM_t(m_t|m_{t-1}^{(j)}) \sum_{k;m_{t-1}=m_t} \frac{w_{t-1}^{(k)}}{P_{t-1}(m_{t-1}=m_t)} KP_{t,m_t}(\theta_t|\theta_{t-1}^{(k)}),
\end{aligned}$$

The weights for all *accepted* particles are

$$w_t(m_t,\theta_t) = \frac{P(m_t,\theta_t) b_t(m_t,\theta_t)}{\sum_{j=1}^{|\mathcal{M}|} P_{t-1}^{(j)}(m_{t-1}^{(j)}) KM_t(m_t|m_{t-1}^{(j)}) \sum_{k;m_{t-1}=m_t} \frac{w_{t-1}^{(k)}}{P_{t-1}(m_{t-1}=m_t)} KP_{t,m_t}(\theta_t|\theta_{t-1}^{(k)})},$$

and intermediate marginal model probabilities $P_t(m)$ are defined as

$$P_t(m_t=m) = \sum_{m_t=m} w_t(m_t,\theta_t).$$

The resulting ABC SMC algorithm is presented in the methodology section of the main part of the paper.

III) *ABC SMC approximation of the marginal likelihood* $P(D_0|m)$

If we can calculate the marginal likelihood $P(D_0|m)$ for each of the candidate models that we consider in the model selection problem, then we can calculate the marginal posterior distribution of a model m as

$$P(m|D_0) = \frac{P(D_0|m)P(m)}{\sum_{m'} P(D_0|m')P(m')}. \tag{4}$$

We now explain how to calculate $P(D_0|m)$ for model m. In the ABC rejection-based approach the posterior distribution of the parameters for each model m are estimated independently by employing ABC rejection; the marginal likelihood then equals the acceptance rate,

$$P(D_0|m) \approx \frac{\#\text{accepted particles given model m}}{N_m}, \tag{5}$$

i.e. the ratio between the number of accepted versus the number of *proposed* particles N_m. We can use this marginal likelihood estimate to calculate $P(m|D_0)$ using equation (4). This approach has been used in [8].

We now derive how ABC SMC can be used for estimating the marginal likelihood, which can be then used for model selection. In a usual ABC SMC setting for drawing samples from the posterior parameter distribution $P(\theta|m, D_0)$ for a given model m, we define intermediate distributions as

$$\pi_t(\theta) = P(\theta)\mathbb{1}(d(D_0, D(\theta)) \leq \epsilon_t). \tag{6}$$

The target distribution π_T is an unnormalized approximation of the posterior distribution $P(\theta|m, D_0)$. We are now interested in its normalization constant, i.e. the marginal likelihood,

$$P(D_0|m) \approx \int_\theta \pi_T(\theta) d\theta.$$

Let us call the integrals of $\pi_t(\theta)$, $\int_\theta \pi_t(\theta)d\theta$, the *intermediate marginal likelihoods*.

In the usual ABC SMC parameter estimation setting, our goal is to obtain samples from distribution $\pi_T(\theta)$, whereas our goal here is to obtain its normalization constant. While this distribution as defined in equation (6) is in general unnormalized, the ABC SMC parameter estimation algorithm performs normalization at every t and therefore returns its normalized version [1]. So we cannot use the usual output of ABC SMC directly. Instead we proceed as follows.

We would like to draw particles from the following target distribution:

$$\mathcal{T}_T(\theta) = P(\theta)\mathbb{1}[d(D_0, D(\theta)) \leq \epsilon_T] + P(\theta)\mathbb{1}[d(D_0, D(\theta)) > \epsilon_T],$$

where $P(\theta)$ is the prior distribution. To draw samples from $\mathcal{T}_T$ we can use ABC SMC, where we define the intermediate distributions as

$$\begin{aligned} \mathcal{T}_t(\theta) &= P(\theta)\mathbb{1}[d(D_0, D(\theta)) \leq \epsilon_t] + P(\theta)\mathbb{1}[d(D_0, D(\theta)) > \epsilon_t] \\ &= \mathcal{T}_t^1(\theta) + \mathcal{T}_t^2(\theta). \end{aligned}$$

In each population we accept N particles, and a particle is only rejected if it falls outside the boundaries of $\mathcal{T}_t$. We classify the accepted particles in two sets, $\Theta_t^1 := \{\theta; d(D_0, D(\theta)) \leq \epsilon_t\}$ and $\Theta_t^2 := \{\theta; d(D_0, D(\theta)) > \epsilon_t\}$, depending on the distance reached. In each population t we can then calculate the intermediate marginal likelihoods by

$$\int_\theta \mathcal{T}_t^1(\theta) d\theta = \sum_{\theta \in \Theta_t^1} w_t(\theta).$$

The target marginal likelihood, $\int_\theta \mathcal{T}_T^1(\theta)d\theta$, is our approximation of $P(D_0|m)$. In an ABC rejection setting, where $T = 1$ and all weights are equal, this result corresponds to (5).

After calculating $P(D_0|m)$ for each m, we can use equation (4) to calculate the marginal posterior distributions for model m,

$$P(m|D_0) \approx \frac{P(m)\sum_{\theta\in\Theta_T^1} w_T(\theta)}{P(m')\sum_{m'}\sum_{\theta'\in\Theta_T'^1} w'_T(\theta')}.$$

The model selection algorithm based on approximating the marginal likelihood proceeds as follows:

Algorithm

M1 For model m_j, $j = 1, \ldots, |\mathcal{M}|$ do steps S1 to S4. Then go to **M2**.

S1 Initialize $\epsilon_1, \ldots, \epsilon_T$.
Set the population indicator $t = 1$.

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

S2.1 If $t = 1$, sample θ^{**} independently from $P(\theta)$.
If $t > 1$, sample θ^* from the previous population $\{\theta_{t-1}^{(i)}\}$ with weights w_{t-1} and perturb the particle to obtain $\theta^{**} \sim K_t(\theta|\theta^*)$, where K_t is a perturbation kernel.
If $P(\theta^{**}) = 0$, return to **S2.1**.
For a particle θ^{**} simulate a candidate data set D and calculate $d(D_0, D(\theta^{**}))$.
If $d(D_0, D(\theta^{**})) \leq \epsilon_t$, add θ^{**} to $\Theta_t^1(m_j)$. If $d(D_0, D(\theta^{**})) > \epsilon_t$, add θ^{**} to $\Theta_t^2(m_j)$.

S2.1 Calculate the weight for particle $\theta_t^{(i)} = \theta^{**}$:

$$w_t^{(i)}(\theta_t^{(i)}) = \begin{cases} 1, & \text{if } t = 1 \\ P(\theta_t^{(i)}) / \left(\sum_{j=1}^N w_{t-1}^{(j)} K_t(\theta_t^{(i)}|\theta_{t-1}^{(j)})\right), & \text{if } t > 1. \end{cases}$$

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

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

S4 Calculate

$$P(D_0|m_j) \approx \frac{P(m_j)\sum_{\theta\in\Theta_T^1(m_j)} w_T(\theta)}{\sum_{m'} P(m')\sum_{\theta'\in\Theta_T'^1(m)} w'_T(\theta')}.$$

M2 For each m_j calculate $P(m_j|D_0)$ using equation (4).

The advantages of this model selection algorithm compared to the marginal likelihood model selection based on ABC rejection can be obtained by (i) starting with a small number of particles N in population 1 and increasing it in each subsequent

population. This way not much computational effort is spent on simulations in earlier populations, but we nevertheless have a big enough sample set in the last population to obtain a reliable estimate; (ii) exploiting the property that intermediate distributions in the parameter estimation framework should be included in one another, and so

$$\text{range } \Theta_t^1 \geq \text{range } \Theta_{t+1}^1, \qquad t = 1, \ldots, T-1.$$

In other words, a proposed particle in population t cannot belong to Θ_t^1 if it cannot be obtained by perturbing any of the particles in Θ_{t-1}^1. We can therefore reject some of the proposed particles without simulation. This means a huge saving in computational time, since simulations are the most expensive part of ABC based algortihms. However, one of the obvious ways to exploit this property would be to use a truncated perturbation kernel with ranges they cover being smaller than the range of prior distribution. But we find this unsatisfactory and, in the present form, feel that evaluating the marginal model likelihood directly is not practical.

References

[1] Toni T, Welch D, Strelkowa N, Ipsen A and Stumpf MPH. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface*, 6:187–202, 2009.

[2] Robert CP and Casella G. *Monte Carlo Statistical Methods*. Springer, 2004.

[3] Doucet A, Freitas ND and Gordon N. *Sequential Monte Carlo Methods in Practice*. Springer, 2001.

[4] Moral PD, Doucet A and Jasra A. Sequential Monte Carlo samplers. *J. Royal Statist. Soc. B*, 2006.

[5] Beaumont MA, Zhang W and Balding DJ. Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

[6] Marjoram P, Molitor J, Plagnol V and Tavare S. Markov chain Monte Carlo without likelihoods. *Proc Natl Acad Sci USA*, 100(26):15324–8, 2003.

[7] Sisson SA, Fan Y and Tanaka MM. Sequential Monte Carlo without likelihoods. *Proc Natl Acad Sci USA*, 104(6):1760–5, 2007.

[8] Wilkinson RD. Bayesian inference of primate divergence times. *PhD thesis, University of Cambridge*, 2007.