An Adaptive Sequential Monte Carlo Method for Approximate Bayesian Computation

Ajay Jasra

ajay.jasra@imperial.ac.uk

Department of Mathematics

Imperial College London

ABC in Paris, June 26th

Joint with Arnaud Doucet & Pierre Del Moral

Introduction

- Consider a standard Bayesian model

$$\pi(\theta|y) \propto f(y|\theta)\pi(\theta)$$

  where $\pi(\theta)$ denotes the prior and $f(y|\theta)$ is the likelihood.

- Assume $f(y|\theta)$ is expensive/impossible to calculate.

- Therefore, can be difficult to compute posterior expectations:

$$\int_{\Theta} h(\theta)\pi(\theta|y)d\theta,$$

  e.g. via MCMC methods.

ABC Approximation

- Consider the approximated posterior:

$$\pi_\epsilon(\theta, x|y) = \frac{\pi(\theta) f(x|\theta) \mathbb{I}_{A_{\epsilon,y}}(x)}{\int_{A_{\epsilon,y} \times \Theta} \pi(\theta) f(x|\theta) dx d\theta} \tag{1}$$

- with
  - $\epsilon > 0$ a tolerance level
  - $\mathbb{I}_B(\cdot)$ the indicator function
  - $A_{\epsilon,y} = \{z \in \mathcal{D} : \rho(\eta(z), \eta(y)) < \epsilon\}$
  - $\eta : \mathcal{D} \to \mathcal{S}$ represents some summary statistics
  - $\rho : \mathcal{S} \times \mathcal{S} \to \mathbb{R}^+$ a distance metric.
- For ϵ small and η sufficient, integrating x, we recover the posterior.

- The choice of summary statistic is important; see the later talks. It is assumed η, ρ are given.
- We focus on trying to simulate from (1) with ϵ as small as possible. This is to recover a good approximation of the true posterior distribution.
- There exist other approximations, e.g. based upon kernel methods. The issue of which may be preferable is not discussed.

Simulation-Based Methods

- Sampling from (1), when ϵ is small, can be difficult.
- Many methods proposed: rejection, MCMC, sequential Monte Carlo (SMC).
- Focus upon SMC, which has been particularly controversial (Sisson et al. 2007).
- Given a set $\{\epsilon_1, \ldots, \epsilon_T : \infty > \epsilon_1 > \cdots > \epsilon_T > 0\}$;
- Sequentially sample from π_{ϵ_1} (as in (1)), then π_{ϵ_2} until π_{ϵ_T}.
- At ϵ_1 it is easy simulate; π_{ϵ_1} is well approximated. Then the idea is to try to use these good approximations, to get to π_{ϵ_2} etc, and ultimately π_{ϵ_T}.

Simulation-Based Methods: Issues

- SMC simulates a collection of N samples, in parallel, sampling from π_{ϵ_1} then π_{ϵ_2} and so on.
- Previous approaches have:
  - Computational complexity, in N, that is $O(N^2)$ (e.g. Beaumont et al. 2009; Toni et al. 2008).
  - A deterministic sequence of $\epsilon_1, \ldots, \epsilon_T$.
- The first issue is computationally prohibitive. The second may be difficult to select in realistic scenarios.

- We present an SMC method that has the properties:
  - a computational complexity that is $O(N)$.
  - it determines, in an automatic fashion, the sequence of tolerance levels to be used.
  - it determines, in an automatic fashion, the parameters of some proposals.

  the latter is present in the method of Beaumont et al. (2009).

SMC Samplers

- Given a sequence of target distributions $\pi_1, \ldots, \pi_T$, on the same space.
- It is possible to sample from them, indirectly, using SMC samplers.
- Use the sequence

$$\widetilde{\pi}_n(z_{0:n}) = \pi_n(z_n) \prod_{j=0}^{n-1} L_j(z_{j+1}, z_j) \tag{2}$$

  where $z_{0:n} := (z_0, \ldots, z_n)$.

- $\{L_n\}_{0 \leq n \leq T-1}$ are a sequence of Markov kernels that act backward in time.
- It is clear from Eq. (2) that $\{\widetilde{\pi}_n\}$ admit $\{\pi_n\}$ as marginals.

SMC Samplers: Algorithm

- Step 0. Set $n = 0$; for $i = 1, \ldots, N$ sample $Z_0^{(i)} \sim \eta_0$ and compute $W_0^{(i)} \propto \pi_0(Z_0^{(i)})/\eta_0(Z_0^{(i)})$, $\sum_{j=1}^N W_0^{(j)} = 1$.

- Step 1. If ESS$(\{W_n^{(i)}\}) < N_T$ then resample N particles, also denoted $\{Z_n^{(i)}\}$ and set $W_n^{(i)} = \frac{1}{N}$. Set $n = n + 1$, if $n = T + 1$ stop.

- Step 2. For $i = 1, \ldots, N$, sample $Z_n^{(i)} \sim K_n(Z_{n-1}^{(i)}, \cdot)$, compute

$$W_n^{(i)} \propto W_{n-1}^{(i)} \frac{\pi_n(Z_n^{(i)}) L_{n-1}(Z_{n-1}^{(i)}, Z_n^{(i)})}{\pi_{n-1}(Z_{n-1}^{(i)}) K_n(Z_{n-1}^{(i)}, Z_n^{(i)})} \tag{3}$$

  and return to Step 1.

SMC Samplers: Algorithm Settings

- For ABC set

$$\pi_{\epsilon_n}(\theta, x_{1:M}|y) \propto \left(\frac{1}{M}\sum_{k=1}^{M}\mathbb{I}_{A_{\epsilon_n,y}}(x_k)\right)\left(\prod_{k=1}^{M} f(x_k|\theta)\right)\pi(\theta) \qquad (4)$$

  for a given $M \in \mathbb{N}$.

- This sequence admits the same marginal in θ for any M.
- It is more expensive to sample from $\pi_{\epsilon_n}(\theta, x_{1:M}|y)$ than $\pi_{\epsilon_n}(\theta, x|y)$ when $M > 1$, but has advantages, illustrated later.
- Let ϵ^* be the smallest value of ϵ we can select.

SMC Samplers: Algorithm Settings

- The performance of SMC samplers depends upon
  - $\{\epsilon_n\}$
  - the transition kernels $\{K_n\}$
  - backward transition kernels $\{L_n\}$.
- Let $\{\epsilon_n\}$ be fixed and K_n, an MCMC kernel of invariant density π_{ϵ_n}. Take L_{n-1} as

$$L_{n-1}(z, z') = \frac{\pi_{\epsilon_n}(z')K_n(z', z)}{\pi_{\epsilon_n}(z)}.$$

- Then Eq. (3) becomes

$$W_n^{(i)} \propto W_{n-1}^{(i)} \frac{\pi_{\epsilon_n}(Z_{n-1}^{(i)})}{\pi_{\epsilon_{n-1}}(Z_{n-1}^{(i)})} \propto W_{n-1}^{(i)} \frac{\sum_{k=1}^{M} \mathbb{I}_{A_{\epsilon_n, y}}(X_{k,n-1}^{(i)})}{\sum_{k=1}^{M} \mathbb{I}_{A_{\epsilon_{n-1}, y}}(X_{k,n-1}^{(i)})}. \qquad (5)$$

- $W_n^{(i)}$ doesn't depend on $\{Z_n^{(i)}\} = \{(\theta_n^{(i)}, X_{1:M,n}^{(i)})\}$: order of sampling and resampling can be swapped.
- A sample is 'dead' if $\forall k$, $\rho(\eta(X_{k,n-1}^{(i)}), \eta(y)) > \epsilon_{n-1}$ ($W_{n-1}^{(i)} = 0$); there are a random number of samples 'alive'.
- Resampling will bring dead particles back to life.
- As $\epsilon_n < \epsilon_{n-1}$, some samples will die when sampling from π_{ϵ_n}.
- Selecting the decay for $\{\epsilon_n\}$ is important:
  - Too fast, and the algorithm will collapse, with no alive particles
  - Too slow, and the algorithm will take a long time.

SMC Samplers: Adaptive $\{\epsilon_n\}$

- Approach based upon fact that (5) for $\{W_n^{(i)}\}$ does not depend on $\{Z_n^{(i)}\}$.
- Hence ϵ_n is selected before the weight is calculated.
- Define

$$\text{PA}(\{W_n^{(i)}\}, \epsilon_n) := \frac{\sum_{i=1}^N \mathbb{I}_{\{W_n>0\}}(W_n^{(i)})}{N}$$

  the proportion of alive particles.

- The value of ϵ_n is set via:

$$\text{PA}(\{W_n^{(i)}\}, \epsilon_n) = \alpha \text{PA}(\{W_{n-1}^{(i)}\}, \epsilon_{n-1})$$

  with $\alpha \in (0, 1)$.

- The $\{K_n\}$ can be adapted using adaptive MCMC ideas.

Adaptive SMC Algorithm

- Step 0. Set $n = 0$; for $i = 1, \ldots, N$, sample $\theta_0^{(i)} \sim \pi(\cdot)$ and $X_{k,0}^{(i)} \sim f(\cdot|\theta_0^{(i)})$, $k = 1, \ldots, M$.
- Step 1. Set $n = n + 1$, if $\epsilon_{n-1} = \epsilon^*$ stop, otherwise determine ϵ_n, with $W_n^{(i)}$ as (5). If $\epsilon_n < \epsilon^*$ then set $\epsilon_n = \epsilon^*$.
- Step 2. If ESS$(\{W_n^{(i)}\}) < N_T$ then resample N particles denoted $\{\theta_{n-1}^{(i)}, X_{1:M,n-1}^{(i)}\}$ and set $W_n^{(i)} = \frac{1}{N}$.
- Step 3. For $i = 1, \ldots, N$, sample $(\theta_n^{(i)}, X_{1:M,n}^{(i)}) \sim K_n((\theta_{n-1}^{(i)}, X_{1:M,n-1}^{(i)}), \cdot)$ if $W_n^{(i)} > 0$ and return to Step 1.

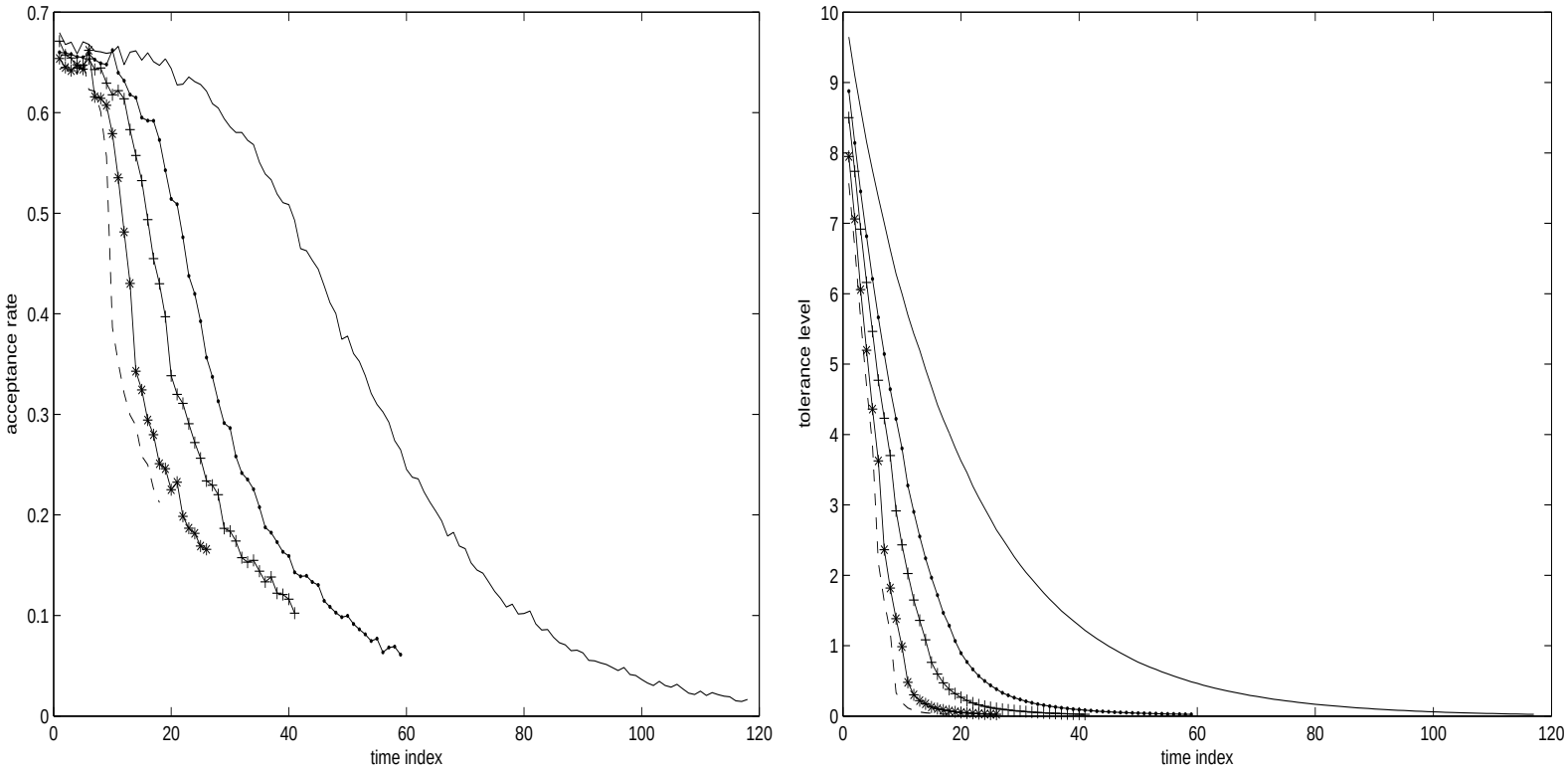
Toy Example

- Some parameters to be set and understood: M, α, N.
- Consider:

$$\theta \sim \mathcal{U}_{[-10,10]},\ f\left(x|\,\theta\right) = 0.5\phi\left(x;\theta,1\right) + 0.5\phi\left(x;\theta,1/100\right).$$

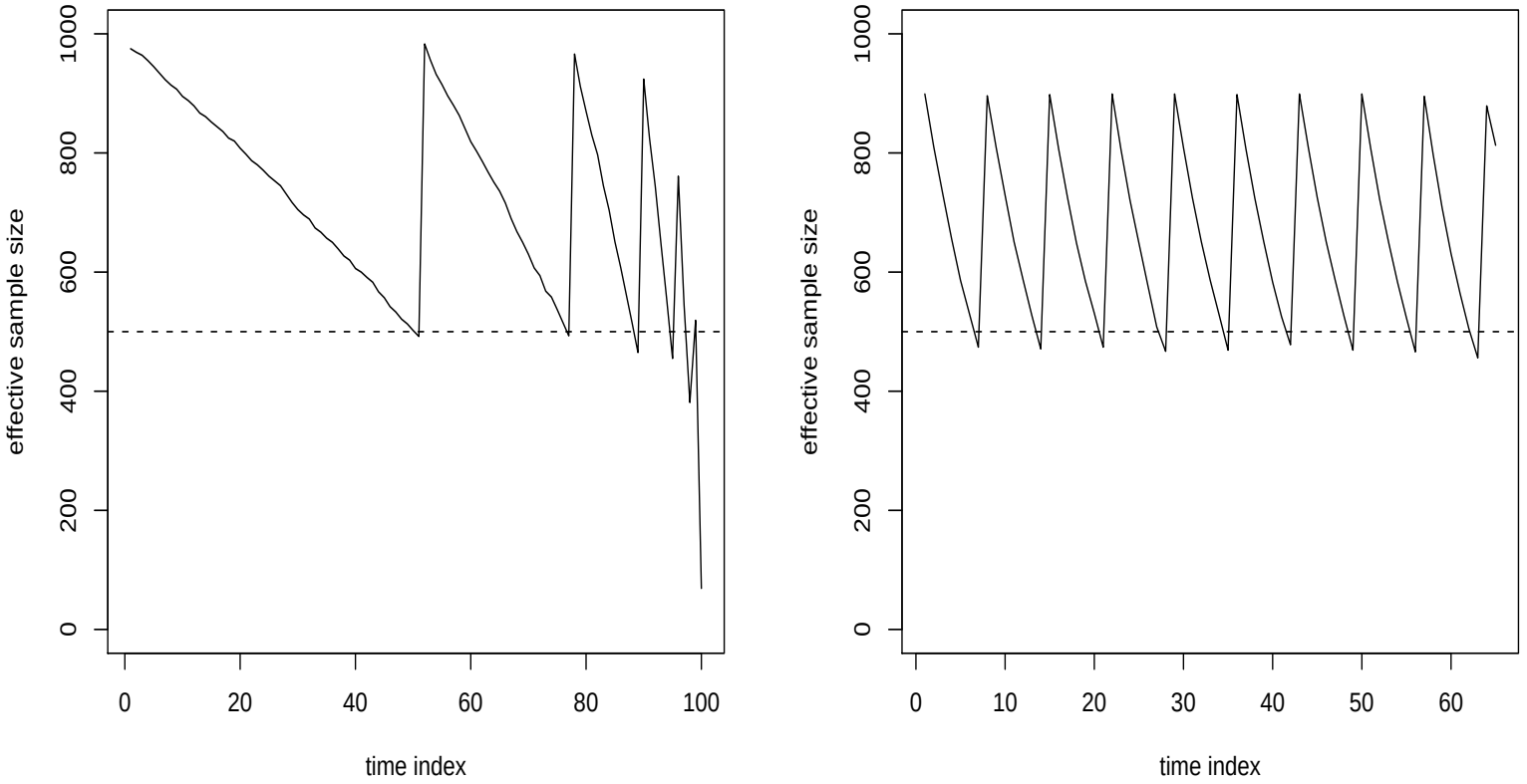
  $\mathcal{U}_{[a,b]}$ is the uniform on $[a,b]$ and $\phi\left(x;m,\sigma^2\right)$ is the normal density.

- Random walk Metropolis kernels used for $\{K_n\}$.
- Found, unsurprisingly, better approximations of π_{ϵ_T}, for large M, N and α.



Average acceptance rate of the MH step (left) and sequence of tolerance levels $\{\epsilon_n\}$ (right) as a function of n for $M = 1$ (solid), 5 (dots), 10 (crosses), 25 (stars) and 50 (dashed dots)

- Also looked at the trade-off of choosing M, N and α.
- For complex problems, want α, N to be 'large' and M to be moderate.
- Otherwise, one can keep α 'small', keeping M and N large.
- Also compare a deterministic ϵ schedule versus an adaptive one:
  - Deterministic: $\epsilon_1 = 10$ and then falling linearly by 0.1 until $\epsilon_n < 0.01$; $\epsilon^* = 0.01$.
  - Adaptive: $\alpha = 0.9$ and $\epsilon^* = 0.01$.
- Whilst reasonable, the deterministic schedule leads to the ESS crashing close to ϵ^*.
- Adaptive procedure has a consistent resampling rate and inferences are likely to be more reliable.



Effective sample size (left deterministic, right adaptive).

Summary

- Presented an SMC method which is fully adaptive and computationally efficient.
- Have compared with population Monte Carlo. Found that for a fixed computational effort SMC samplers can perform better.
- For real data and complex model (in epidemiology), found that SMC samplers can outperform MCMC (Bortot et al, 2007) when sampling from the model with small ϵ^*.
- This latter aspect is the most important part of the algorithm.