GR8201: An introduction to Approximate Bayesian Computation

Simon Tavaré

June 2019

What is ABC about?

- Stochastic models of physical phenomena are often complicated
- Statistical inference for such models is a challenge
- Likelihood-based methods are commonly used . . .
- . . . but likelihoods are often intractable
- What do we do?

3

Overview of Course

- Bayesian preliminaries: Rejection methods, Population genetics example, Likelihood-free inference
- ABC: Regression-based methods, Summary statistics, Model choice
- MCMC methods, Sequential MC methods, Indirect inference
- ABC samplers, Substantive examples.

4

Notation (1)

$\mathcal{D}$ — data (usually discrete)

θ — parameters (often high-dimensional)

$\pi(\theta)$ — *prior* for θ

Aim is to study the *posterior* $f(\theta|\mathcal{D})$ given by

$$f(\theta|\mathcal{D}) = \frac{\mathbb{P}(\mathcal{D}|\theta)\,\pi(\theta)}{\mathbb{P}(\mathcal{D})}$$

where

$$\mathbb{P}(\mathcal{D}) = \int \mathbb{P}(\mathcal{D}|\theta)\,\pi(\theta)\,d\theta$$

is the *normalising constant*

Posterior is proportional to likelihood times prior

6

Notation (2)

The *marginal likelihood* is

$$f(\mathcal{D}) = \int f(\mathcal{D}|\theta)\,\pi(\theta)\,d\theta$$

The *prior predictive distribution* of a random variable $Y = h(\mathcal{D})$ is

$$f_{\text{prior}}(y) = \int f_Y(y|\theta)\,\pi(\theta)\,d\theta$$

The *posterior predictive distribution* of Y is

$$f_{\text{post}}(y) = \int f_Y(y|\theta)\,f(\theta|\mathcal{D}_0)\,d\theta$$

where $\mathcal{D}_0$ denotes the observed data, and f_Y the distribution of Y.

7

Example (1)

Suppose X is a Poisson random variable with mean θ, so that

$$\mathbb{P}(X = j) := f(j|\theta) = \frac{e^{-\theta}\theta^j}{j!}, j = 0, 1, \ldots$$

We write $X \sim \mathrm{Po}(\theta)$.
Recall that $\mathbb{E}X = \theta = \mathrm{Var}(X)$

Assume π is the gamma density with parameters r and λ

$$\pi(\theta) = \frac{\lambda^r\, \theta^{r-1} e^{-\lambda\theta}}{\Gamma(r)}, \quad \theta > 0$$

We write $\theta \sim \mathrm{Gamma}(r, \lambda)$.
Recall that $\mathbb{E}\theta = r/\lambda$ and $\mathrm{Var}(\theta) = r/\lambda^2$.

8

Example (2)

It is easy to show that

$$\mathcal{L}(\theta|j) \sim \mathrm{Gamma}(j + r, \lambda + 1)$$

and that the normalising constant is

$$\mathbb{P}(j) = \frac{\Gamma(r + j)}{\Gamma(r)\, j!} \left(\frac{1}{\lambda + 1}\right)^j \left(\frac{\lambda}{\lambda + 1}\right)^r \tag{1}$$

We say X has a Negative Binomial distribution with parameters r and p if

$$\mathbb{P}(X = j) = \binom{j + r - 1}{j} (1 - p)^r\, p^j, \quad j = 0, 1, 2, \ldots$$

9

Example (3)

We write $X \sim \mathrm{NegBin}(r, p)$.
Recall that $\mathbb{E}X = rp/(1-p)$ and $\mathrm{Var}(X) = rp/(1-p)^2$.

(1) shows that the prior predictive distribution is Negative Binomial with parameters r and $p = 1/(1+\lambda)$.

You should check that the posterior predictive distribution is Negative Binomial with parameters $r + j_0$ and $p = 1/(\lambda + 2)$, where j_0 is the observed value.

10

The Rejection Algorithm 11

Rejection method

We now turn to methods for simulating observations from the posterior $f(\theta|\mathcal{D})$ The simplest is the *Rejection Method*:

1. Generate $\theta \sim \pi(\cdot)$
2. Accept θ with probability $h = \mathbb{P}(\mathcal{D}|\theta)$; return to [1.]

Observations accepted by this algorithm have density

$$\propto \pi(\theta)\, \mathbb{P}(\mathcal{D}|\theta) = f(\theta|\mathcal{D})$$

12

Hitting the target

How long does it take to get an accepted observation?

$$\begin{aligned}\mathbb{P}(\text{ accept first observation}) &= \int_\theta \pi(\theta)\,\mathbb{P}(\mathcal{D}|\theta)d\theta \\ &= \mathbb{P}(\mathcal{D}) := p\end{aligned}$$

Because the simulations are independent, it follows that

$$\mathbb{P}(\text{first observation accepted on the } r\text{th trial}) = (1-p)^{r-1}p, r = 1, 2, \ldots$$

The expected number of trials to get n accepted observations is n/p

13

Hitting the target quicker

If you can find a constant c such that

$$\mathbb{P}(\mathcal{D}|\theta) \leq c, \quad \forall\theta \tag{2}$$

then can replace step [2.] with

2. Accept θ with probability h/c

The mean number of trials to get n accepted observations is then nc/p

Note: the acceptance rate can be used to estimate the normalizing constant $\mathbb{P}(\mathcal{D})$.

14

Example from Population Genetics 15

The coalescent (1)

The setting: a random sample of n sequences is taken at random from a population and the locations of the segregating sites (or SNPs) are recorded.

Think of the sequences as copies of the unit interval. SNPs in them arise as a consequence of mutation. We will ignore all sorts of things, such as recombination, variable population size, and selection.

For a sample from a stationary population of constant size, the genealogy of the sample is provided by Kingman's *coalescent*.

We model the ancestry of the n sequences as a random tree. It starts from n tips, waits a time T_n and then chooses two sequences at random to join. There are now $n-1$ ancestors of the sample.

16

The coalescent (2)

We then wait time T_{n-1} and choose two of the ancestral sequences to merge. Continuing in this way, the sample spends a time T_2 with two ancestors, finally tracing back to the most recent common ancestor (MRCA).

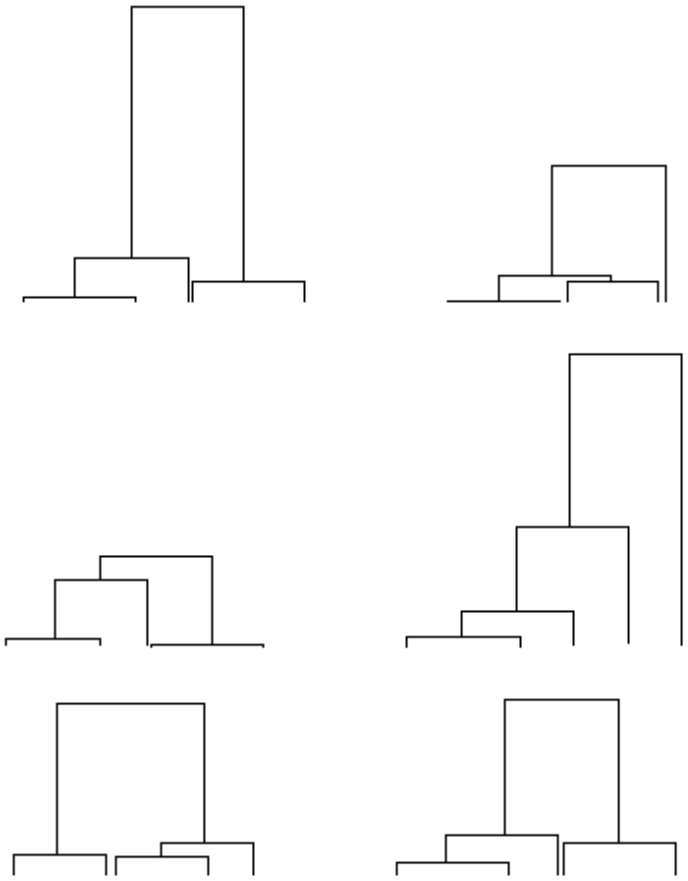
In this simple model, the random variables $T_n, T_{n-1}, \ldots, T_2$ are independent and exponentially distributed, with

$$\mathbb{E}T_j = \frac{2}{j(j-1)}$$

The time scale is measured in units of $2N$ generations, N being the population size.

17

Coalescent trees (1)



18

Coalescent trees (2)

Fig. 4.2. Coalescent trees for samples of size 6 and 32 from a population of constant size

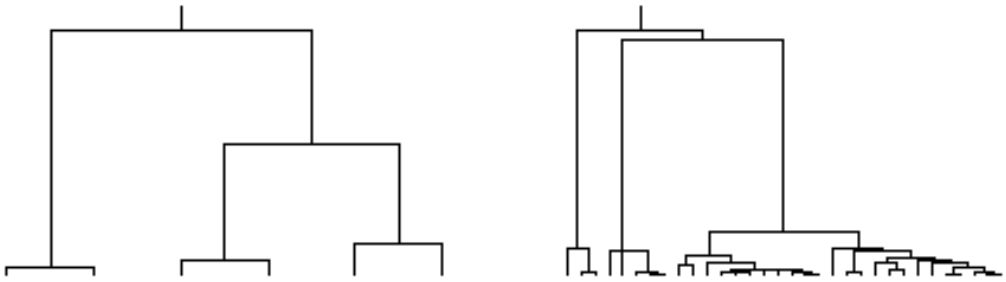
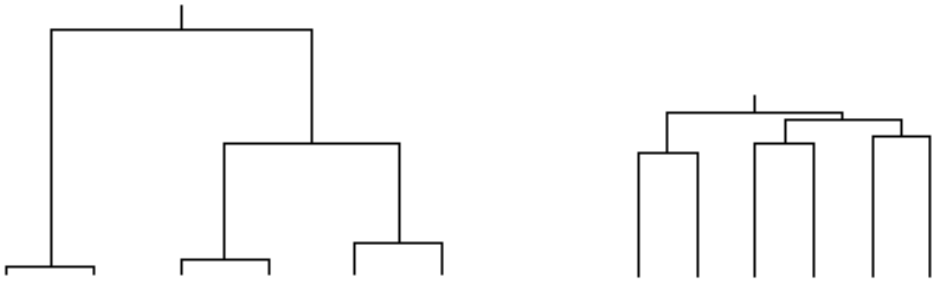


Fig. 4.3. The coalescent tree of a sample of size 6 (constant population size in left panel, exponentially growing population in right panel)



19

Mutations in the coalescent (1)

Mutations are superimposed on the coalescent tree according to points of independent Poisson processes of rate $\theta/2$. In the *infinitely-many sites model*, each mutation introduces a segregating site into the sample. In this setting, θ is the compound parameter $\theta = 4Nu$, where u is the per generation mutation rate.

Note that, given the times $T_n, T_{n-1}, \ldots, T_2$, the number of segregating sites introduced while the sample has $n, n-1, \ldots, 2$ distinct ancestors have independent Poisson distributions with means

$$n\,\frac{\theta}{2}\,T_n, (n-1)\,\frac{\theta}{2}\,T_{n-1}, \ldots, 2\,\frac{\theta}{2}\,T_2.$$

20

Mutations in the coalescent (2)

It follows that, given $T_n, \ldots, T_2$, the total number of SNPs, S_n, in the sample satisfies

$$\mathcal{L}(S_n|T_n, \ldots, T_2) \sim \mathrm{Po}\left(\frac{\theta}{2}\sum_{j=2}^{n} jT_j\right)$$

This gives what we need to find the posterior distribution of $\theta, T_n, \ldots, T_2$ given $S_n = s$, the observed number of segregating sites.

21

The posterior distribution of T_{MRCA} (1)

Our rejection algorithm is

1. Generate $\theta \sim \pi(\cdot)$
2. Generate $T_n, \ldots, T_2$ from model. Calculate $L_n = \sum_{j=2}^{n} jT_j$
3. Accept $\theta, T_n, \ldots, T_2$ with probability

$$h = \mathrm{Po}\left(\frac{\theta}{2} L_n\right)\{s\}$$

   Return to [1.]

We get the posterior of $T_{\mathrm{MRCA}} = T_2 + \cdots + T_n$ from the accepted values in this algorithm

22

Likelihood-free Inference 23

What, no likelihood?

In our earlier examples, we were able to calculate the *likelihood* $\mathbb{P}(\mathcal{D}|\theta)$

What if we can't?

This leads us to the field of likelihood-free inference, and this relies on our ability to simulate observations from the underlying model.

We begin with a result from Don Rubin (1984)

24

Rejection method, revisited (1)

The analogue of the rejection method is:

1. Generate $\theta \sim \pi(\cdot)$
2. Generate $\mathcal{D}'$ from the model with parameter θ
3. Accept θ if $\mathcal{D}' = \mathcal{D}$; return to [1.]

Observations accepted by this algorithm have density

$$\begin{aligned} &\propto \pi(\theta)\,\mathbb{P}(\mathcal{D}' = \mathcal{D}|\theta) \\ &= \pi(\theta)\mathbb{P}(\mathcal{D}|\theta) \\ &\propto f(\theta|\mathcal{D}) \end{aligned}$$

25

Rejection method, revisited (2)

- No likelihoods . . .
- No analogue of c (?)
- Rubin DB (1984) Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann Statist* 12: 1151-1172

26

The posterior distribution of T_{MRCA} (2)

To implement this in our genetics example we replace the h step with a simulation step:

1. Generate $\theta \sim \pi(\cdot)$
2. Generate $T_n, \ldots, T_2$ from model. Calculate $L_n = \sum_{j=2}^{n} jT_j$
3. Simulate S' from $\mathrm{Po}(\theta L_n/2)$
4. Accept $\theta, T_n, \ldots, T_2$ if $S' = s$

27

The first ABC method

In the earlier methods we have assumed that we can hit the target. What happens if the acceptance probability is very small?
This gives us the first ABC method:

1. Generate $\theta \sim \pi(\cdot)$
2. Generate $\mathcal{D}'$ from the model with parameter θ
3. Accept θ if $\rho(\mathcal{D}', \mathcal{D}) < \epsilon$, where
   - ρ is a metric on the space of $\mathcal{D}$s
   - $\epsilon \geq 0$ is a parameter to be chosen

   Return to [1.]

28

Notes

- We can choose ρ to compare the data sets in a useful way
- If ρ is a metric, then $\rho(\mathcal{D}', \mathcal{D}) = 0 \iff \mathcal{D} = \mathcal{D}'$.
  Hence
  - $\epsilon = 0$ gives the exact answer
  - $\epsilon \to \infty$ reproduces the prior
  - $0 < \epsilon < \infty$ gives trade-off between accuracy and computability
- This method works for continuous data – Weiss and von Haeseler (1998) treated the frequentist case
- For the population genetics example we could use neighbourhoods of $\{S_n = s\}$ as the region.

29

ABC

If we were to use the full sequence data in that example, the target is very hard to hit. This suggests summarising the data.

This gives us the following ABC method (Pritchard et al, 1999):

1. Generate $\theta \sim \pi(\cdot)$
2. Generate $\mathcal{D}'$ from the model with parameter θ
3. Choose a set of summary statistics S of the data
   - Compute $S \equiv S(\mathcal{D})$, and $S' = S(\mathcal{D}')$
   - Accept θ if $\rho(S', S) < \epsilon$, where
     - ρ is a metric on the space of Ss

   Return to [1.]

30