Approximate Bayesian Computation

Jessi Cisewski-Kehe
Yale University

Pennstate Astroinformatics Summer School
June 7, 2018

Some slides were adapted from a presentation by Chad Schafer (CMU)
Approximate Bayesian Computation

- “Likelihood-free” approach to approximating $p(\theta | x_{\text{obs}})$
  ($p(x_{\text{obs}} | \theta)$ not specified)
- Proceeds via simulation of the forward process

The posterior for $\theta$ given observed data $x_{\text{obs}}$:

$$p(\theta | x_{\text{obs}}) = \frac{p(x_{\text{obs}} | \theta)p(\theta)}{\int p(x_{\text{obs}} | \theta)p(\theta)d\theta} \propto p(x_{\text{obs}} | \theta)p(\theta)$$

Why would we not know $p(x_{\text{obs}} | \theta)$?

1. Physical model too complex
2. Strong dependency in data
3. Observational limitations

Some Astronomy ABC examples: Cameron and Pettitt (2012); Schafer and Freeman (2012); Weyant et al. (2013); Akeret et al. (2015); Ishida et al. (2015)
Basic ABC algorithm

For the observed data $x_{obs}$ and prior $p(\theta)$:

1. Sample $\theta_{prop}$ from prior $p(\theta)$
2. Generate $x_{prop}$ from forward process $F(x \mid \theta_{prop})$
3. Accept $\theta_{prop}$ if $x_{obs} = x_{prop}$
4. Return to step 1

*Introduced in Tavaré et al. (1997) and Pritchard et al. (1999)
Binomial illustration

- Data are a sample of 1’s and 0’s coming from $Y_i \sim \text{Bernoulli}(p)$ where $n =$ sample size, $\theta = P(Y = 1)$.

- Likelihood is $p(y | \theta) = \binom{n}{y} \theta^y (1 - \theta)^{n-y}$, where $y = \sum_{i=1}^{n} y_i$ (but we will pretend we do not know this).

Need to determine a distance function, $\rho$. Use the following:

$$\rho(y, x) = \frac{1}{n} |y - x|$$

Hence $\rho(y, x) = 0$ if the generated dataset $x$ has the same number of 1’s as $y$. 
n <- 1000  #number of observations
N <- 1000  #generated sample size
true.p <- .75
data <- rbinom(n,1,true.p)
epsilon <- 0
alpha.hyper <- 1
beta.hyper <- 1
p <- numeric(N)
rho <- function(y,x) abs(sum(y)-sum(x))/n
for(i in 1:N){
  d <- epsilon+1
  while(d>epsilon) {
    proposed.p <- rbeta(1,alpha.hyper,beta.hyper)
x <- rbinom(n,1,proposed.p)
d <- rho(data,x)
  }
p[i] <- proposed.p}

Reference: Turner and Zandt (2012)
Binomial illustration: posterior

Tolerance: 0, N:1000, n:25

- True Posterior
- Estimated Posterior
- True mu
It turns out that $\theta_{acc}$ is a draw from the posterior if

$$P(\text{Accept } \theta_{prop} \mid \theta_{prop} = \theta) \propto p(x_{obs} \mid \theta) \text{ (the likelihood)}$$

- This provides a basis for assessing the quality of the ABC approximation

- To achieve this, we could accept $\theta_{prop}$ if $x_{prop} = x_{obs}$ (i.e. accept $\theta_{prop}$ that reproduce the $x_{obs}$ exactly)
  
  $\rightarrow$ Of course, this is not practical (way too slow!)

- Instead, accept $\theta_{prop}$ if $x_{prop}$ is “close to” $x_{obs}$ using some chosen distance metric $\Delta$. 
Tolerance: $\epsilon$

Define:

$$\phi_\epsilon(x_{\text{prop}}, x_{\text{obs}}) = \begin{cases} 1, & \text{if } \Delta(x_{\text{prop}}, x_{\text{obs}}) < \epsilon \\ 0, & \text{if } \Delta(x_{\text{prop}}, x_{\text{obs}}) \geq \epsilon \end{cases}$$

In other words, $\phi_\epsilon(x_{\text{prop}}, x_{\text{obs}})$ is an indicator as to whether or not $x_{\text{prop}}$ is close to $x_{\text{obs}}$.

Hence,

$$P(\text{Accept } \theta_{\text{prop}} \mid \theta_{\text{prop}} = \theta) = P(\Delta(x_{\text{prop}}, x_{\text{obs}}) < \epsilon \mid \theta_{\text{prop}} = \theta) = \int \phi_\epsilon(x, x_{\text{obs}}) p(x \mid \theta) \, dx$$

$$\rightarrow K p(x_{\text{obs}} \mid \theta) \text{ as } \epsilon \rightarrow 0$$

Hence, for $\epsilon$ small,

$$P(\text{Accept } \theta_{\text{prop}} \mid \theta_{\text{prop}} = \theta) \approx K p(x_{\text{obs}} \mid \theta)$$
**Toy Example:** Assume we have a single observation, $x_{obs}$, from a Gaussian with mean $\theta$ and variance one.

The convolution

$$\int \phi_\epsilon(x, x_{obs}) f(x \mid \theta) \, dx = P(\text{Accept } \theta_{prop} \mid \theta_{prop} = \theta)$$

for case where $x_{obs} = 1$, $\theta = 0$ (left) / $\theta = 1$ (right), $\epsilon = 0.1$. 
Note: Acceptance probability curve has been normalized so the area under the curve is 1.
Comparing $x_{\text{prop}}$ with $x_{\text{obs}}$ is not generally computationally feasible

- For example, when $x$ is high-dimensional, $\epsilon$ will need to be too large in order to keep the acceptance probability reasonable.

- Instead, compare (lower dimensional) summaries, $S(x_{\text{prop}})$ and $S(x_{\text{obs}})$. 
For observations $x_{\text{obs}}$, distance function $\rho$, and (small) tolerance $\epsilon$

Algorithm 1 Basic ABC Algorithm

1: for $i = 1$ to $N$ do
2: while $\rho \left( S(x_{\text{obs}}), S(x_{\text{prop}}) \right) > \epsilon$ do
3: Propose $\theta_{\text{prop}}$ by drawing $\theta_{\text{prop}}$ from prior $p(\theta)$
4: Generate $x_{\text{prop}}$ from forward process $F(x | \theta_{\text{prop}})$
5: Calculate summary statistics $\{S(x_{\text{obs}}), S(x_{\text{prop}})\}$
6: end while
7: $\theta^{(i)} \leftarrow \theta_{\text{prop}}$
8: end for

- ABC posterior based on $\{\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(N)}\} = \{\theta^{(i)}\}_{i=1}^{N}$
- $\{\theta^{(i)}\}_{i=1}^{N}$ are often referred to as particles
“The basic idea behind ABC is that using a representative (enough) summary statistic $\eta$ coupled with a small (enough) tolerance $\epsilon$ should produce a good (enough) approximation to the posterior...”

Marin et al. (2012)
Gaussian illustration

- Data $x_{\text{obs}}$ consists of 25 iid draws from Normal($\mu$, 1)
- Summary statistics $S(x) = \bar{x}$
- Distance function $\Delta(S(x_{\text{prop}}), S(x_{\text{obs}})) = |\bar{x}_{\text{prop}} - \bar{x}_{\text{obs}}|$
- Tolerance $\epsilon = 1$ and 0.08
- Prior $\pi(\mu) = \text{Normal}(0, 10)$
Different tolerances ($\epsilon = 1$ vs $\epsilon = 0.08$)

Choice of $\epsilon$ is important
Different summary statistics (sample mean vs sample median)

→ choice of summary statistic(s) is(are) important
Decisions that need to be made:

1. Select distance function ($\rho$) and summary statistic(s)
2. Tolerance ($\epsilon$)

Finding the “right” $\epsilon$ can be inefficient
$\rightarrow$ we end up throwing away many of the theories proposed from the selected priors

How can we improve this algorithm?
Main idea

Instead of starting the ABC algorithm over with a smaller tolerance ($\epsilon$), use the already sampled particle system as a proposal distribution *rather* than drawing from the prior distribution.

Particle system:

(1) retained sampled values, (2) importance weights

Some references:
Beaumont et al. (2009); Moral et al. (2011); Bonassi and West (2004)
Algorithm 2 ABC - Population Monte Carlo algorithm

1: At iteration $t = 1$
2: Basic ABC sampler to obtain $\{\theta_1^{(i)}\}_{i=1}^N$
3: Set importance weights $W_1^{(i)} = 1/N$ for $i = 1, \ldots, N$
4: for $t = 2$ to $T$ do
5: Set $\tau_t^2 = 2 \cdot \text{var} \left( \{\theta_{t-1}^{(i)}, W_{t-1}^{(i)}\}_{i=1}^N \right)$
6: for $i = 1$ to $N$ do
7: while $\rho(S(x_{\text{obs}}), S(x_{\text{prop}})) > \epsilon_t$ do
8: Draw $\theta_0$ from $\{\theta_{t-1}^{(i)}\}_{i=1}^N$ with probabilities $\{W_{t-1}^{(i)}\}_{i=1}^N$
9: Propose $\theta_{\text{prop}} \sim N(\theta_0, \tau_t)$
10: Generate $x_{\text{prop}}$ from $F(x | \theta_{\text{prop}})$
11: Calculate summary statistics $\{S(x_{\text{obs}}), S(x_{\text{prop}})\}$
12: end while
13: $\theta_t^{(i)} \leftarrow \theta_{\text{prop}}$
14: $\widetilde{W}_{t}^{(i)} \leftarrow \frac{\pi(\theta_t^{(i)})}{\sum_{j=1}^{N} W_{t-1}^{(j)} \phi\left[\tau_{t-1}^{-1}(\theta_t^{(i)} - \theta_t^{(j)})\right]}$
15: end for
16: $\{W_{t}^{(i)}\}_{i=1}^N \leftarrow \{\widetilde{W}_{t}^{(i)}\}_{i=1}^N / \sum_{i=1}^{N} \widetilde{W}_{t}^{(i)}$
17: end for

Decreasing tolerances $\epsilon_1 \geq \cdots \geq \epsilon_T$, $\phi(\cdot)$ is the density function of a $N(0, 1)$
From Beaumont et al. (2009)
Gaussian illustration: sequential posteriors

N:1000, n:25

Tolerance sequence, $\varepsilon_{1:10}$:

1.00 0.75 0.53 0.38 0.27 0.19 0.15 0.11 0.08 0.06
Sequential setting: decisions

1. Determining the sequence of tolerances, $\epsilon_{1:t}$
   One possibility: use a quantile (e.g. 50th percentile) of the distribution of accepted distances from the previous time step

2. Moving the particles between time steps
   Need to ensure any constraints on the parameter space are satisfied

3. Calculating the particle weights
   Relies on ideas from *Importance Sampling*
There are other variations of ABC that may prove useful in your setting (Marin et al., 2012)

Beaumont et al. (2002) introduces a post-processing adjustment (using local regression) to the simulation output in order to use more of the simulated draws (with extensions in Blum and François (2010))
Approximate Bayesian Computation could be a useful tool in astronomy, but it must be handled with care.

There are three main decisions that need to be made in the standard ABC algorithm: summary statistic, distance function, and tolerance.

Considering a sequence of tolerances can lead to more efficient sampling, but results in more decisions: how to decrease the tolerance, when to stop the sampling, how to “move” or “mix” the particles between sampling steps.

**Additional resources**

- Csilléry et al. (2010): Approximate Bayesian Computation (ABC) in practice
- Csillery et al. (2012): abc: an R package for approximate Bayesian computation (ABC)
- Jabot et al. (2013): EasyABC: performing efficient approximate Bayesian computation sampling schemes (R package)


