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 $p(\theta \mid x_{\text{obs}})$ ($p(x_{\text{obs}} \mid \theta)$ not specified)
- Proceeds via simulation of the forward process

The posterior for θ given observed data x_{obs}:

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

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

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 $p(\theta)$:

Algorithm*

1. Sample θ_{prop} from prior $p(\theta)$
2. Generate x_{prop} from forward process $F(x \mid \theta_{\text{prop}})$
3. Accept θ_{prop} if $x_{\text{obs}} = x_{\text{prop}}$
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 $p(y \mid \theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$, where $y = \sum_{i=1}^n y_i$ (but we will pretend we do not know this).

Need to determine a distance function, ρ. Use the following:

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

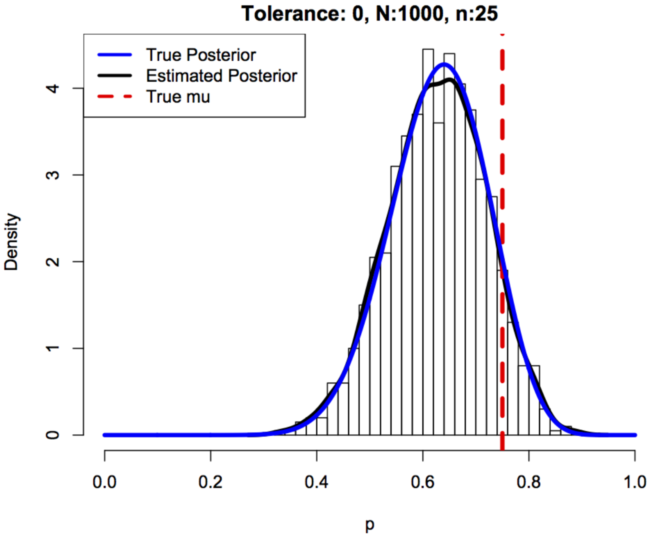
Hence $\rho(y, x) = 0$ if the generated dataset x has the same number of 1's as y.

Binomial illustration: R code

```
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



It turns out that θ_{acc} is a draw from the posterior if

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

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

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

- Instead, accept θ_{prop} if x_{prop} is "close to" x_{obs} using some chosen distance metric Δ.

Tolerance: ϵ

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_{prop} is close to x_{obs}.

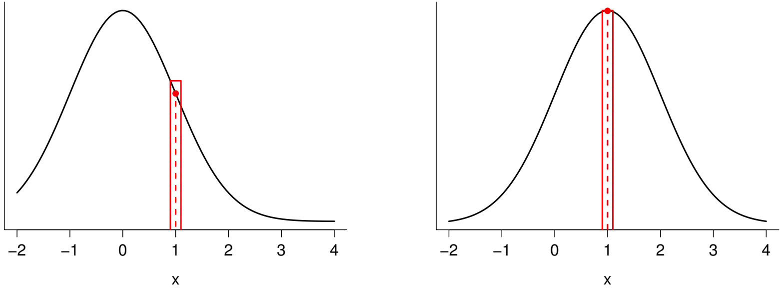
Hence,

$$\begin{aligned} P(\text{Accept } \theta_{\text{prop}} \mid \theta_{\text{prop}} = \theta) &= P(\Delta(x_{\text{prop}}, x_{\text{obs}}) < \epsilon \mid \theta_{\text{prop}} = \theta) \\ &= \int \phi_\epsilon(x, x_{\text{obs}}) p(x \mid \theta)\, dx \\ &\longrightarrow K p(x_{\text{obs}} \mid \theta) \quad \text{as } \epsilon \to 0 \end{aligned}$$

Hence, for ϵ small,

$$P(\text{Accept } \theta_{\text{prop}} \mid \theta_{\text{prop}} = \theta) \approx K p(x_{\text{obs}} \mid \theta)$$

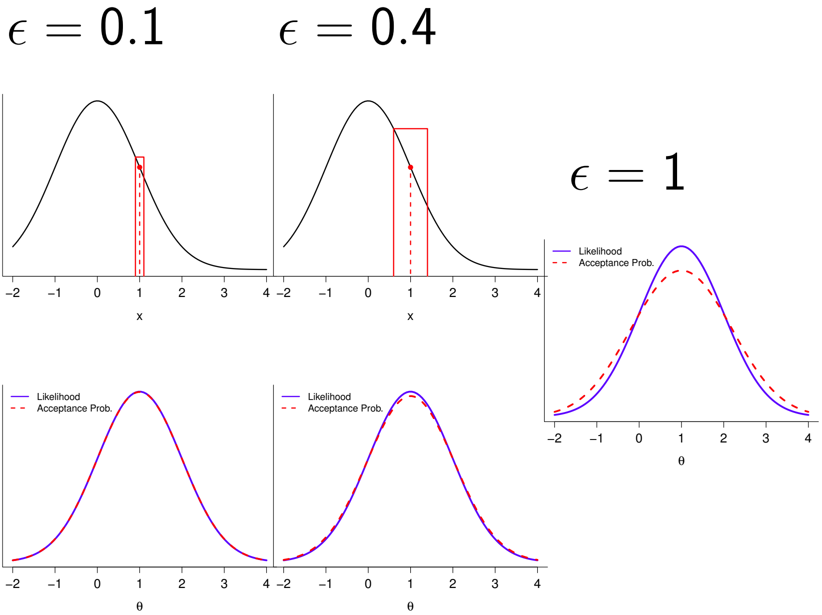
Toy Example: Assume we have a single observation, x_{obs}, from a Gaussian with mean θ and variance one.



Depicts the convolution

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

for case where $x_{\text{obs}} = 1$, $\theta = 0$ (left) / $\theta = 1$ (right), $\epsilon = 0.1$.



Note: Acceptance probability curve has been normalized so the area under the curve is 1.

Summary statistics

Comparing x_{prop} with x_{obs} is not generally computationally feasible

- For example, when x is high-dimensional, ϵ 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_{obs}, distance function ρ, and (small) tolerance ϵ

Algorithm 1 Basic ABC Algorithm

```
for i = 1 to N do
    while ρ(S(x_obs), S(x_prop)) > ϵ do
        Propose θ_prop by drawing θ_prop from prior p(θ)
        Generate x_prop from forward process F(x | θ_prop)
        Calculate summary statistics {S(x_obs), S(x_prop)}
    end while
    θ^(i) ← θ_prop
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*

ABC in a nutshell

"The basic idea behind ABC is that using a representative (enough) summary statistic η coupled with a small (enough) tolerance ϵ should produce a good (enough) approximation to the posterior..."

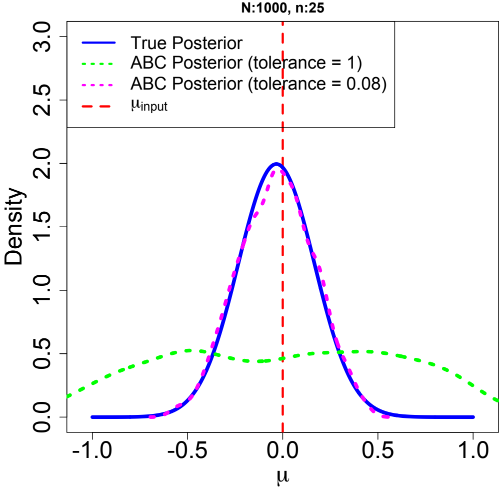
Marin et al. (2012)

Gaussian illustration

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

Gaussian illustration: posteriors for μ

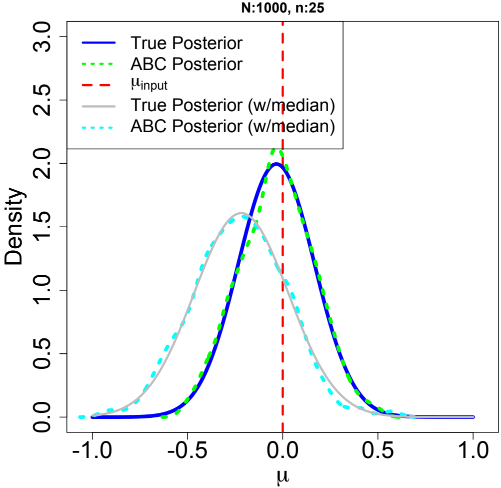
$\longrightarrow$ Different tolerances ($\epsilon = 1$ vs $\epsilon = 0.08$)



$\longrightarrow$ **choice of ϵ is important**

Gaussian illustration: posteriors for μ

$\longrightarrow$ Different summary statistics (sample mean vs sample median)



$\longrightarrow$ **choice of summary statistic(s) is(are) important**

Summary of basic ABC

- Decisions that need to be made:
  1. Select distance function (ρ) and summary statistic(s)
  2. Tolerance (ϵ)

- Finding the "right" ϵ can be inefficient
  $\longrightarrow$ we end up throwing away many of the theories proposed from the selected priors

- How can we improve this algorithm?

Sequential ABC

Main idea

Instead of starting the ABC algorithm over with a smaller tolerance (ϵ), 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

At iteration $t = 1$

Basic ABC sampler to obtain $\{\theta_1^{(i)}\}_{i=1}^N$

Set importance weights $W_1^{(i)} = 1/N$ for $i = 1, \dots, N$

for $t = 2$ to T **do**

    Set $\tau_t^2 = 2 \cdot \mathrm{var}\left(\{\theta_{t-1}^{(i)}, W_{t-1}^{(i)}\}_{i=1}^N\right)$

    for $i = 1$ to N **do**

        while $\rho\left(S(x_{\text{obs}}), S(x_{\text{prop}})\right) > \epsilon_t$ **do**

            Draw θ_0 from $\{\theta_{t-1}^{(i)}\}_{i=1}^N$ with probabilities $\{W_{t-1}^{(i)}\}_{i=1}^N$

            Propose $\theta_{\text{prop}} \sim N(\theta_0, \tau_t)$

            Generate x_{prop} from $F(x \mid \theta_{\text{prop}})$

            Calculate summary statistics $\{S(x_{\text{obs}}), S(x_{\text{prop}})\}$

        end while

        $\theta_t^{(i)} \leftarrow \theta_{\text{prop}}$

        $\widetilde{W}_t^{(i)} \leftarrow \dfrac{\pi\left(\theta_t^{(i)}\right)}{\sum_{j=1}^N W_{t-1}^{(j)} \phi\left[\tau_t^{-1}(\theta_t^{(i)} - \theta_{t-1}^{(j)})\right]}$

    end for

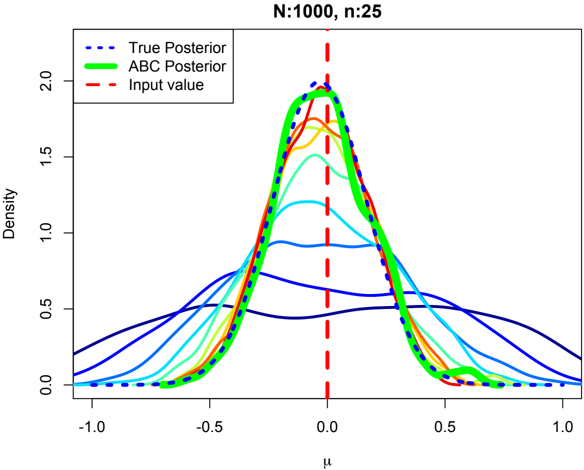
    $\{W_t^{(i)}\}_{i=1}^N \leftarrow \{\widetilde{W}_t^{(i)}\}_{i=1}^N / \sum_{i=1}^N \widetilde{W}_t^{(i)}$

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



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

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))

Concluding remarks

1. Approximate Bayesian Computation could be a useful tool in astronomy, but it must be handled with care
2. There are three main decisions that need to be made in the standard ABC algorithm: summary statistic, distance function, and tolerance
3. 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)

Akeret, J., Refregier, A., Amara, A., Seehars, S., and Hasner, C. (2015), "Approximate Bayesian computation for forward modeling in cosmology," *Journal of Cosmology and Astroparticle Physics*, 2015, 043.

Beaumont, M. A., Cornuet, J.-M., Marin, J.-M., and Robert, C. P. (2009), "Adaptive approximate Bayesian computation," *Biometrika*, 96, 983 – 990.

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002), "Approximate Bayesian Computation in Population Genetics," *Genetics*, 162, 2025 – 2035.

Blum, M. G. B. and François, O. (2010), "Non-linear regression models for Approximate Bayesian Computation," *Statistics and Computing*, 20, 63 – 73.

Bonassi, F. V. and West, M. (2004), "Sequential Monte Carlo with Adaptive Weights for Approximate Bayesian Computation," *Bayesian Analysis*, 1, 1–19.

Cameron, E. and Pettitt, A. N. (2012), "Approximate Bayesian Computation for Astronomical Model Analysis: A Case Study in Galaxy Demographics and Morphological Transformation at High Redshift," *Monthly Notices of the Royal Astronomical Society*, 425, 44–65.

Csilléry, K., Blum, M. G., Gaggiotti, O. E., and François, O. (2010), "Approximate Bayesian Computation (ABC) in practice," *Trends in ecology & evolution*, 25, 410 – 418.

Csillery, K., Francois, O., and Blum, M. G. B. (2012), "abc: an R package for approximate Bayesian computation (ABC)," *Methods in Ecology and Evolution*.

Ishida, E. E. O., Vitenti, S. D. P., Penna-Lima, M., Cisewski, J., de Souza, R. S., Trindade, A. M. M., Cameron, E., and Busti, V. C. (2015), "cosmoabc: Likelihood-free inference via Population Monte Carlo Approximate Bayesian Computation," *Astronomy and Computing*, 13, 1 – 11.

Jabot, F., Faure, T., and Dumoullin, N. (2013), *EasyABC: performing efficient approximate Bayesian computation*.

Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. J. (2012), "Approximate Bayesian computational methods," *Statistics and Computing*, 22, 1167 – 1180.

Moral, P. D., Doucet, A., and Jasra, A. (2011), "An adaptive sequential Monte Carlo method for approximate Bayesian computation," *Statistics and Computing*, 22, 1009–1020.

Pritchard, J. K., Seielstad, M. T., and Perez-Lezaun, A. (1999), "Population Growth of Human Y Chromosomes: A study of Y Chromosome Microsatellites," *Molecular Biology and Evolution*, 16, 1791 – 1798.

Schafer, C. M. and Freeman, P. E. (2012), *Statistical Challenges in Modern Astronomy V*, Springer, chap. 1, Lecture Notes in Statistics, pp. 3 – 19.

Tavaré, S., Balding, D. J., Griffiths, R., and Donnelly, P. (1997), "Inferring coalescence times from DNA sequence data," *Genetics*, 145, 505 – 518.

Turner, B. M. and Zandt, T. V. (2012), "A tutorial on approximate Bayesian computation," *Journal of Mathematical Psychology*, 56, 69 – 85.

Weyant, A., Schafer, C., and Wood-Vasey, W. M. (2013), "Likeihood-free cosmological inference with type Ia supernovae: approximate Bayesian computation for a complete treatment of uncertainty," *The Astrophysical Journal*, 764, 116.