Approximate Bayesian Computation: what, why and how

Simon Tavaré

DAMTP and Cambridge Research Institute

AISTATS , May 14 2010

Stochastic Computation in Biology

Biological Motivation

Evolution on different time scales

- Tracking cell lineages in an individual
- Understanding tumour growth, metastasis, response to treatment
- Analysis of genome-wide association studies (GWAS)
- Ancestral history of humans (e.g. phylogeography, admixture with Neanderthals)
- Estimating divergence times of primates from fossil record

Statistical Motivation

HIghly dependent data with nasty likelihoods

- Likelihood-free inference
- ABC: approximate Bayesian computation
- Not just Bayesian: useful in classical setting (e.g. MLEs)

- A gentle overview of 3 basic ABC methods
- An example from stem cell biology

Approach I: Rejection methods

Introduction to Bayesian computation

- Discrete data $\mathcal{D}$, prior $\pi(\theta)$ for parameters θ
- Aim: generate observations from posterior distribution $f(\theta \mid \mathcal{D})$
- We have

$$f(\theta \mid \mathcal{D}) \propto \mathbb{P}(\mathcal{D} \mid \theta)\pi(\theta)$$

Posterior proportional to likelihood × *prior*

Rejection methods I

R1 Generate θ from $\pi(\cdot)$

R2 Accept θ with probability $h = \mathbb{P}(\mathcal{D} \mid \theta)$; repeat

Accepted observations have distribution $f(\theta \mid \mathcal{D})$

R1 Generate θ from $\pi(\cdot)$

R2 Accept θ with probability $h = \mathbb{P}(\mathcal{D} \mid \theta)$; return to R1

Accepted observations have distribution $f(\theta \mid \mathcal{D})$

Can do better: if

$$\mathbb{P}(\mathcal{D} \mid \theta) \leq c \text{ for all } \theta$$

then can replace h above with h/c

The number of runs to get n observations from $f(\theta|\mathcal{D})$ is negative binomial, with mean

$$\frac{nc}{\mathbb{P}(\mathcal{D})}$$

This shows

– the effect of $\mathbb{P}(\mathcal{D})$

– the effect of c

– a way to estimate $\mathbb{P}(\mathcal{D})$

Complex stochastic models

- A stochastic process often underlies the likelihood computation
- This process may be complex, making explicit probability calculations difficult or impossible
- Thus $\mathbb{P}(\mathcal{D} \mid \theta)$ may be uncomputable (either quickly enough or theoretically)

- A stochastic process often underlies the likelihood computation
- This process may be complex, making explicit probability calculations difficult or impossible
- Thus $\mathbb{P}(\mathcal{D} \mid \theta)$ may be uncomputable (either quickly enough or theoretically)

Exploit simulation
Diggle P& R Gratton R (1984) *JRSSB*, **46**, 193–227

Rejection methods II

RS1 Generate θ from $\pi(\cdot)$

RS2 Simulate $\mathcal{D}'$ from stochastic model with parameter θ

RS3 Accept θ if $\mathcal{D}' = \mathcal{D}$; repeat

RS1 Generate θ from $\pi(\cdot)$

RS2 Simulate $\mathcal{D}'$ from stochastic model with parameter θ

RS3 Accept θ if $\mathcal{D}' = \mathcal{D}$; repeat

- Just as before, accepted observations from this algorithm have the density $f(\cdot|\mathcal{D})$
- Despite its appearance, this algorithm is much more general than first one — no need for explicit calculation

Rubin DB (1984) *Ann Statist*, **4**, 1151–1172

Approximate Bayesian Computation I

A1 Generate θ from $\pi(\cdot)$

A2 Simulate $\mathcal{D}'$ from stochastic model with parameter θ

A3 Calculate distance $\rho(\mathcal{D}, \mathcal{D}')$ between $\mathcal{D}'$ and $\mathcal{D}$

A4 Accept θ if $\rho \leq \epsilon$; repeat

- If $\epsilon \to \infty$, generates from prior

  If $\epsilon = 0$, generates from $f(\theta \mid \mathcal{D})$

- Choice of ϵ reflects tension between computability and accuracy

  - PCR — post-computational remorse

- Method is *honest*: you get observations from $f(\theta \mid \rho(\mathcal{D}, \mathcal{D}') \leq \epsilon)$

- Works for continuous data (assuming ρ is a metric)

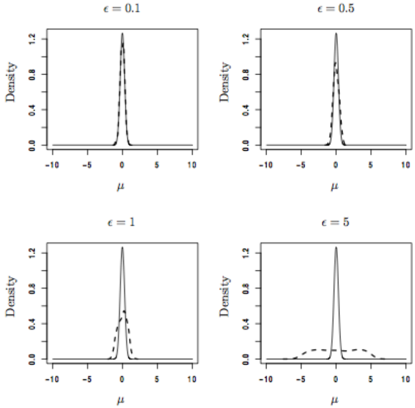
Approximate Bayesian Computation II

- The limit $\epsilon = 0$ reproduces the data precisely
- In many examples the data are too high-dimensional
- . . . so reduce dimension by using summary statistics

An illustrative example (R. Wilkinson)

- $X_1, \ldots, X_n$ iid N(μ, σ^2)
- σ^2 known, improper prior $\pi(\mu) \propto 1$
- Posterior is N$(\bar{X}, \sigma^2/n)$
- Assume $\bar{X} = 0$
- Use ABC, accepting μ if $\bar{X}' \leq \epsilon$

Plots of posterior for μ



Estimating the error

Can calculate

$$d_{TV}(\pi_\epsilon(\mu), \pi(\mu|\bar{X}=0)) = \frac{1}{2}\int |\pi_\epsilon(\mu) - \pi(\mu|\bar{X}=0)| d\mu$$

Get

$$d_{TV} = \frac{cn\epsilon^2}{\sigma^2} + o(\epsilon^2),$$

where $c = \sqrt{2/\pi}\exp(-1/2) \approx 1/2$.

Approximate sufficiency

Recall that $S = S(\mathcal{D})$ is *sufficient* for θ if

$$\mathbb{P}(\mathcal{D} \mid S, \theta) \text{ is independent of } \theta$$

- If S is sufficient for θ, then $f(\theta \mid \mathcal{D}) = f(\theta \mid S)$
- Typically, S is of smaller dimension that $\mathcal{D}$. Inference method can be simplified (and sped up), e.g. rejection method via

$$f(\theta \mid S) \propto \mathbb{P}(S \mid \theta)\pi(\theta)$$

Research problem

Puts a premium on finding decent summary statistics

- Definition of approximate sufficiency?
- A systematic, implementable approach?
- Estimate distance between $f(\theta \mid \mathcal{D})$ and $f(\theta \mid S)$ given a measure of how far from sufficient S is for θ

Combine summaries and rejection

Pritchard J et al (1999) *Mol Biol Evol*, $\mathbf{16}$, 1791–1798

Choose statistics $S = (S_1, \ldots, S_p)$ to summarize $\mathcal{D}$

AS1 Generate θ from $\pi(\cdot)$

AS2 Simulate $\mathcal{D}'$, calculate s'

AS3 Accept θ if $\rho(s', s) \leq \epsilon$; repeat

Generalizations

- Using all simulations

  Beaumont M, Zhang W& Balding D(2002) *Genetics* **162**, 2025–2035

  Weight all simulated observations using distance from target

- Approximately sufficient statistics

  Joyce P & Marjoram P (2008), *SAGMB* **7**, art 26

Advantages and disadvantages of ABC

Pros:

- Usually easy to code
- Generates independent observations (can use embarrassingly parallel computation)
- Can be used to estimate Bayes factors directly
- Usually easy to adapt

Cons:

- May be hard to anticipate effects of summary statistics
- For complex probability models, sampling from prior does not make good use of accepted observations
- Choice of metric matters

Approach II: Markov chain Monte Carlo

The Hastings Markov chain

M1 Now at θ

M2 Propose move to θ' according to $q(\theta \to \theta')$

M3 Calculate the Hastings ratio

$$h = \min\left(1, \frac{\mathbb{P}(\mathcal{D} \mid \theta')\pi(\theta')q(\theta' \to \theta)}{\mathbb{P}(\mathcal{D} \mid \theta)\pi(\theta)q(\theta \to \theta')}\right)$$

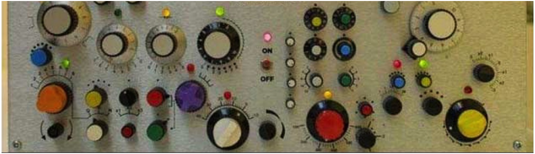
M4 Accept θ' with probability h, else return θ

Basic output analysis

There are more things to check:

- Is the chain ergodic?
- Does it mix well?
- Is the chain stationary?
- Burn in?
- Diagnostics of the run (no free lunches)

MCMC in evolutionary genetics setting



- Small tweaks in the biology often translate into huge changes in algorithm
- Long development time
- All the usual problems with convergence
- Almost all the effort goes into evaluation of likelihood

(Yet) another MCMC approach

Marjoram P et al (2003) *PNAS*, **100**, 15324–15328

MS1 Now at θ

MS2 Propose a move to θ' according to $q(\theta \rightarrow \theta')$

MS3 Generate $\mathcal{D}'$ using θ'

MS4 If $\mathcal{D}' = \mathcal{D}$, go to next step, else return θ

MS5 Calculate

$$h = h(\theta, \theta') = \min\left(1, \frac{\pi(\theta')q(\theta' \rightarrow \theta)}{\pi(\theta)q(\theta \rightarrow \theta')}\right)$$

MS6 Accept θ' with probability h, else return θ

Practical version: ABC

Data $\mathcal{D}$, summary statistics S

[MS4$'$] If $\rho(\mathcal{D}', \mathcal{D}) \leq \epsilon$, go to next step, otherwise return θ

[MS4$''$] If $\rho(S', S) \leq \epsilon$, go to next step, otherwise return θ

for some suitable metric ρ and approximation level ϵ

Observations now from $f(\theta \mid \rho(\mathcal{D}', \mathcal{D}) \leq \epsilon)$ or $f(\theta \mid \rho(S', S) \leq \epsilon)$

Variations on ABC

- Convergence an issue?
- These methods can often be started at stationarity, so no burn-in
- If the underlying probability model is complex, simulating data will not often lead to acceptance. Thus need update for parts of the probability model (data augmentation)
- There are versions with varying ϵ
  Bortot P et al (2007) *JASA*, **104**, 84–92
- There are now many hybrid versions of these approaches (e.g. ABC-within-Gibbs)

Approach III: Population Monte Carlo

ABC-PRC & ABC-PMC

Sisson S et al (2007) *PNAS*, **104**, 1760–1765
Beaumont M et al (2009) *Biometrika*, **96**, 983–990

Start with $\epsilon_1 > \epsilon_2 > \cdots > \epsilon_T$

1. For iteration 1:

   For $i = 1, 2, \ldots, N$,

   - Simulate $\theta_i^{(1)} \sim \pi(\theta)$ and $x \sim f(x|\theta_i^{(1)})$ until $\rho(x, y) < \epsilon_1$
   - Set $w_i^{(1)} = 1/N$

   Take $\tau_2^2 = 2\times$ empirical variance of the $\theta_1^{(i)}$

2. For iterations $2 \leq t \leq T$,

   For $i = 1, 2, \ldots, N$, repeat

   - Generate θ_i^* from $\theta_j^{(t-1)}$ w.p. $w_j^{(t-1)}$
   - Generate $\theta_i^{(t)} | \theta_i^* \sim \mathrm{N}(\theta_i^*, \tau_t^2)$, and $x \sim f(x|\theta_i^{(t)})$ until $\rho(x, y) < \epsilon_t$
   - $w_i^{(t)} \propto \pi(\theta_t^{(i)}) \Big/ \sum_j w_j^{(t-1)} \phi(\sigma_t^{-1} \{\theta_i^{(t)} - \theta_j^{(t-1)}\})$

   Take $\tau_{t+1} = 2\times$ weighted variance of the $\theta_t^{(i)}$

Inference in an agent based model: Stem cell biology

[slides in keynote presentation]

Conclusions

- ABC provides one approach to inference with intractable likelihoods
- There are others:
  - composite likelihood methods
  - model simplification
- Historical examples?
- Many theoretical issues to be resolved
- ABC seems a natural method for inference in agent-based models
- Reviews:
  Marjoram P & Tavaré S (2006) *Nat Rev Genet*, **7**, 759–770
  Beaumont M(2010) AREES, in press

It will be maintained that the end of an era has now been reached, as regards both statistical methods and computational techniques, and an outline of the way in which biometric techniques in genetical demography may be expected to develop will be given. Particular emphasis will be placed on the need to formulate sound methods of 'estimation by simulation' on complex models.

Edwards AWF (1967) *Biometrics*, **23**, 176

The dangers of ABC H. L. Mencken:

For every complex problem, there is an answer that is short, simple and wrong

Why use ABC? J. Galsworthy:

Idealism increases in direct proportion to one's distance from the problem