SC7/SM6 Bayes Methods HT20

Lecturer: Geoff Nicholls

Lecture 9: ABC

Notes and Problem sheets are available at

http://www.stats.ox.ac.uk/~nicholls/BayesMethods/
Approximate Bayesian Computation is a Monte Carlo scheme targeting an approximate posterior. We use it when the likelihood is intractable, but the generative model

\[(\theta, y) \sim \pi(\theta)p(y|\theta)\]

is easy to simulate. If the observation model \(p(y|\theta) \propto \tilde{p}(y, \theta)\) (as a function of \(y\)) then

\[p(y|\theta) = \frac{\tilde{p}(y, \theta)}{c(\theta)}\]

with

\[c(\theta) = \int \tilde{p}(y, \theta)dy\]

and \(c(\theta)\) may be intractable. These problems are called doubly intractable. Standard MCMC wont work, as we cant calculate the acceptance probability \(\alpha(\theta'|\theta)\).
ABC example: the Ising model - a doubly intractable model.

Denote by $\Omega_Y = \{0, 1\}^{n^2}$ the set of all binary images $Y = (Y_1, Y_2, ..., Y_{n^2})$, $Y_i \in \{0, 1\}$, where $i = 1, 2, ..., n^2$ is the cell index on the square lattice of image cells. Let $#y$ give the number of disagreeing neighbors in the binary image $Y = y$.

The Ising model is the following distribution over $\Omega$:

$$p(y|\theta) = \exp(-\theta #y)/c(\theta).$$

Here $\theta \geq 0$ is a positive smoothing parameter and

$$c(\theta) = \sum_{y \in \Omega_Y} \exp(-\theta #y)$$

is a normalizing constant which we can't compute for $n$ large.*

*There is a formula for $c(\theta)$ for the special case of periodic boundary conditions, but that is not generally our image model of choice.
Suppose we have image data $y$ and we want to estimate $\theta$. If $\pi(\theta)$ is a prior for $\theta$ then the posterior is

$$
\pi(\theta|y) = \frac{p(y|\theta)\pi(\theta)}{p(y)}.
$$

Consider doing MCMC targeting $\pi(\theta|y)$. Choose a simple proposal for the scalar parameter $\theta$, say $\theta' \sim U(\theta - a, \theta + a)$, $a > 0$. The acceptance probability is

$$
\alpha(\theta'|\theta) = \min \left\{ 1, \frac{p(y|\theta')\pi(\theta')}{p(y|\theta)\pi(\theta)} \right\}
$$

$$
= \min \left\{ 1, \frac{c(\theta)}{c(\theta')} \times \text{easy stuff} \right\}
$$

and although $p(y)$ cancels, $c(\theta)/c(\theta')$ does not, and so we are left with an acceptance probability we cannot evaluate.
Rejection sampling, ABC-style

The basic ABC algorithm approximates the following variant of rejection. Suppose $y$ and $\theta$ are both discrete so $p(y|\theta) \leq 1$.

The following algorithm simulates $\theta \sim \pi(\theta|y)$ where (as usual) $$\pi(\theta|y) \propto p(y|\theta)\pi(\theta).$$

(1) Simulate $\theta \sim \pi(\theta)$ and $y' \sim p(y'|\theta)$.

(2) If $y' = y$ return $\theta$ and stop, otherwise goto (1).

Line (2) simulates an event succeeding with probability $p(y|\theta)$. 
Let $\Theta_{\text{rej}}$ be the value returned by this algorithm.

Claim: $\Theta_{\text{rej}} \sim \pi(\cdot | y)$.

Proof: To find the probability $\Theta_{\text{rej}} = \theta$, sum over probabilities for all sequences of rejections ending in the return value $\theta$,

$$
\Pr(\Theta_{\text{rej}} = \theta) = \pi(\theta)p(y|\theta) + \pi(\theta)p(y|\theta) \sum_{\theta'} \pi(\theta')(1 - p(y|\theta')) + \ldots
$$

$$
= \pi(\theta)p(y|\theta)(1 + (1 - p(y)) + (1 - p(y))^2 + \ldots)
$$

$$
= \pi(\theta)p(y|\theta) \times \frac{1}{1 - (1 - p(y))}
$$

$$
= \pi(\theta|y)
$$

with $p(y) = \sum_{\theta} \pi(\theta)p(y|\theta)$. 
Rejection-ABC (where now \((\theta, y) \in \Omega \times \mathcal{Y} - \text{general rv}\)).

ABC assumes that if \(y'\) is “close” to \(y\) then \(\pi(\theta|y')\) is a good approximation to \(\pi(\theta|y)\). We measure “close” using summary statistics \(S(y) = (S_1(y), \ldots, S_p(y))\) chosen so they inform \(\theta\).

Let \(d(s', s)\) be a distance between vectors \(s' = S(y'), s = S(y)\). Let \(\delta \geq 0\) be a threshold distance.

(1) Simulate \(\theta \sim \pi(\theta)\) and \(y' \sim p(y'|\theta)\).

(2) If \(d(S(y'), S(y)) < \delta\) return \((\theta, y')\) and stop, else goto (1).

Claim: If \(\theta, y'\) is the (random) output pair then marginally

\[ \theta \sim \pi(\theta|d(S(Y), S(y)) < \delta) \]

with \(Y \sim p(\cdot)\) and \(p(y'), y' \in \mathcal{Y}\) the prior predictive.
Proof: for \( y \in \mathcal{Y} \) let
\[
\Delta_\delta(y) = \{ y' \in \mathcal{Y} : d(S(y'), S(y)) < \delta \}.
\]
The algorithm returns \((\theta, y') \sim \pi(\theta)p(y'|\theta)\) conditioned on
\[
y' \in \Delta_\delta(y)
\]
so the joint distribution of the output pair is
\[
\theta, y'|y \sim \pi(\theta)p(y'|\theta)I_{y' \in \Delta_\delta(y)}.
\]
The marginal distribution of the output \( y' \) is
\[
y'|y \sim p(y')I_{y' \in \Delta_\delta(y)}
\]
with \( p(y') = \int_{\Omega} \pi(\theta)p(y'|\theta)d\theta \) the original prior predictive distribution. The marginal distribution of the output \( \theta \) is
\[
\theta|y \sim \pi(\theta)P(Y \in \Delta_\delta(y)|\theta),
\]
where \( Y \sim p(\cdot) \). This is just \( \theta|y \sim \pi(\theta|Y \in \Delta_\delta(y)) \). [EOP]
We have made two (more) approximations
- we summarise the data with $y \rightarrow s(y)$
- we replace the data-statement
  
  “the realised value of the data is $Y = y$”

with

  “the realised value of the data is in the ball $Y \in \Delta_\delta(y)$”.

If $S : \mathcal{Y} \rightarrow \mathbb{R}^p$ is a sufficient statistic then

\[
\pi(\theta|Y = y) = \pi(\theta|S(Y) = s(y)),
\]

and in that case

\[
\pi(\theta|Y \in \Delta_\delta(y)) \rightarrow \pi(\theta|y)
\]
as $\delta \rightarrow 0$ may often be verified.
Example

Data model: \( y_i \sim \text{Poisson}(\Lambda)^*, \ i = 1, 2, ..., \ n \) with \( n = 5 \).

Prior: \( \Lambda \sim \Gamma(\alpha = 1, \beta = 1) \).

Summary statistic: \( S(y) = \bar{y} \)

Distance measure: \( d(\bar{y}', \bar{y}) = |\bar{y}' - \bar{y}| \) and \( \delta = 0.5, 1 \).

ABC algorithm

(1) Simulate \( \lambda \sim \Gamma(\alpha, \beta) \) and \( y'_i \sim \text{Poisson}(\Lambda), \ i = 1, 2, ..., \ n \).

(2) If \( |\bar{y}' - \bar{y}| < \delta \) return \( \theta \) and stop, otherwise goto (1).

We do Bayesian inference without calculating \( L \) or \( \pi \), sometimes called “likelihood free” inference. We just specify how to simulate parameters and data.

*True value was \( \Lambda = 2 \).
Regression adjustment of samples

ABC generates pairs \( \theta, y' \sim \pi(\theta)p(y'|\theta) \). Conditional on \( y' \), \( \theta \sim \pi(\theta|y') \). “Shift” this distribution onto the data at \( y' = y \! \).

Assume (1) \( s = S(y) \) is sufficient and let \( s' = S(y') \).

Assume (2) shifting the data, \( y \) to \( y' \), shifts the posterior mean

\[
\mu(s) = E(\theta|S(Y) = s)
\]

but has no other effect on the distribution of \( \theta|y \). If this is true then if \( \theta \sim \pi(\cdot|y) \) we can alternatively write

\[
\theta = \mu(s) + \epsilon
\]

with \( \epsilon \sim \mathcal{F} \) a mean zero r.v. with \( \mathcal{F} \) not depending on \( y \).
Assume (3) for $d(s', s)$ small, the linear approximation

$$
\mu(s') \approx \alpha + (s' - s)\beta
$$

is good. Clearly $\alpha \approx \mu(s)$ is the posterior mean at the data.

If we knew $\mu(s)$, we could simulate $\theta|s$ by simulating $\epsilon$ and setting

$$
\theta = \alpha + \epsilon.
$$

We have lots of pairs $\theta(t), S(y(t))$ and we can regress them to estimate a local linear approximation

$$
\theta(t) = \alpha + (S(y(t)) - s)\beta + \epsilon(t).
$$
We estimate $\hat{\alpha}, \hat{\beta}$ using LS-regression and set

$$
\theta_{\text{adj}}^{(t)} = \theta^{(t)} - (S(y^{(t)}) - s)\hat{\beta} \\
= \left[\alpha + (S(y^{(t)}) - s)\beta + \epsilon^{(t)}\right] - [S(y^{(t)}) - s]\hat{\beta} \\
\simeq \alpha + \epsilon^{(t)},
$$

if $\hat{\beta} \simeq \beta$ is a good approximation. In this case $\theta_{\text{adj}}^{(t)} \sim \pi(\cdot|s)$ approximately, a sample from the posterior, under our assumptions.

The regression correction adjusts the distribution of $\theta$ at $y'$ to move it onto the distribution of $\theta$ at $y$.

Example: We did exactly this for the Poisson example above. Worked well. See the figure and R-code for this lecture.
Example: Radiocarbon dating revisited

We will fit our shrinkage model using ABC and see how it compares to MCMC. ABC is much easier here.

Recall the model.

Data model: \( y_i' \sim N(\mu(\theta_i), \sigma^2 + \sigma_c(\theta_i)^2) \) for \( i = 1, ..., n \).

Prior: uniform span \( v \sim U(L, U) \), \( \psi_1 \sim U(L, U - v) \),
\( \psi_2 = \psi_1 + v \), \( \theta_i \sim U(\psi_1, \psi_2) \) \( i = 1, ..., n \).

Summary statistic \( s(y) = y \) (works here as just \( n = 7 \) dates).

Distance: Euclidian \( d(s(y'), s(y)) = |y' - y| \). We chose \( \delta \) by experimentation (see figure).
for (k in 1:K) {
    span=runif(1,min=0,max=U-L); #uniform span
    lower=runif(1,min=L,max=U-span); #psi[1]
    upper=lower+span; #psi[2]
    dates=round(runif(nd,min=lower,max=upper))
    y.sim=mu[dates]+sqrt(d^2+err[dates]^2)*rnorm(nd)
    D=sqrt(sum((y-y.sim)^2))/1000 #arbitrary scale
    if (D<delta) {
        S=rbind(S,y.sim); psi=rbind(psi,c(lower,upper))
    }
}

Implementation detail: it is common practice to save all the simulation output, not just the ones satisfying \( y' \in \Delta y \), and choose \( \delta \) so some fixed fraction are retained. This allows us to trial different \( \delta \)-values without rerunning. Above is simple rejection-ABC.
ABC example: the Ising Model

The Ising distribution $Y \sim p(\cdot | \theta)$ is easy to sample for moderate $n$-values using MCMC. Here are 3 samples $Y \sim p(y|\theta)$:

These samples are not exactly distributed according to $p(y|\theta)$ (convergence) but we can make them as good as we need by taking long MCMC runs.
We can sample $Y \sim p(\cdot|\theta)$ using MCMC: Suppose $Y^{(t)} = y$.

[Step 1] Choose an update, something simple. Choose a cell $i \sim U\{1, 2, ..., n^2\}$. Set $y'_i = 1 - y_i$ and $y'_j = y_j$ for $j \neq i$. Notice that $q(y'|y) = q(y|y') = 1/n^2$ for $y', y$ differing at exactly one cell.

[Step 2] Write down the algorithm. Let $Y^{(t)} = y$. $Y^{(t+1)}$ is determined in the following way.

1. Simulate $y' \sim q(y'|y)$ as above, and $u \sim U(0, 1)$.

2. If $u < \alpha(y'|y)$ set $Y^{(t+1)} = y'$ and otherwise set $Y^{(t+1)} = y$. 
[Step 3] Calculate $\alpha$. The $q$’s cancel as usual, so

$$\alpha(y'|y) = \min \left\{ 1, \frac{p(y'|\theta)q(y|y')}{p(y|\theta)q(y'|y)} \right\}$$

$$= \min \left\{ 1, \exp(-\theta(#y' - #y)) \right\}$$

It is clear the algorithm is irreducible ($q$ is irreducible and $\alpha$ is never zero) and aperiodic (rejection is possible), so it is ergodic for $p(y)$.

An implementation of this algorithm in R is available in the code for this lecture. Some samples produced using this code are shown above.
ABC inference for $\theta$

Recall the doubly intractable inference for $\theta|y$.

We have data $Y = y$ which is an $n \times n$ binary matrix and our observation model for $Y$, $Y \sim p(y|\theta)$ is the Ising model. Our goal is to estimate $\theta$. The posterior is

$$\pi(\theta|y) = \frac{p(y|\theta)\pi(\theta)}{p(y)}$$

and the likelihood

$$p(y|\theta) = \exp(-\theta \#y)/c(\theta)$$

depends on $c(\theta)$, an intractable function of $\theta$. 
Our ABC algorithm sampling $\theta \sim \pi(\theta|y)$ approximately is as follows. Suppose we have a prior for $\theta$. Given the scale of $\theta$, $\text{Exp}(2)$ is a natural generic prior. We take $S(y) = \#y$, which is actually sufficient for $\theta$.

(1) Simulate $\theta \sim \text{Exp}(2)$ and $y' \sim \exp(-\theta\#y')/c(\theta)$.

(2) If $|\#y' - \#y| < \delta$ return $\theta$, otherwise, goto (1).

We implemented this (see attached R, 8x8 Ising, $\theta = 0.8$) and estimated the posterior densities in the figure.

The distribution converges to something stable as $\delta \to 0$. The regression adjustment for $\delta = 0.1$ corrects its distribution to agree with that for $\delta = 0.05$. 