Bayesian Phylogenetics

Workshop on Molecular Evolution
Woods Hole, Massachusetts
4 Aug 2019

Paul O. Lewis
Department of Ecology & Evolutionary Biology

See also 20 & 27 June 2018 at http://phyloseminar.org/recorded.html
Bayesian inference
Joint probabilities

10 marbles in a bag
Sampling with replacement

- Pr(B,S) = 0.4
- Pr(W,S) = 0.1
- Pr(B,D) = 0.2
- Pr(W,D) = 0.3
Conditional probabilities

What's the probability that a marble is black given that it is dotted?

5 marbles satisfy the condition (D)

Pr(B|D) = \frac{2}{5}

2 remaining marbles are black (B)
Marginal probabilities

Marginalizing over color yields the total probability that a marble is dotted (D)

\[ \Pr(D) = \Pr(B,D) + \Pr(W,D) \]

\[ = 0.2 + 0.3 \]

\[ = 0.5 \]

Marginalization involves summing all joint probabilities containing D
Marginalization

\[
\begin{array}{cc}
\text{B} & \text{W} \\
\hline
\text{D} & \Pr(D,B) & \Pr(D,W) \\
\text{S} & \Pr(S,B) & \Pr(S,W) \\
\end{array}
\]
Marginalizing over colors

Marginal probability of being dotted is the sum of all joint probabilities involving dotted marbles.
Joint probabilities

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>W</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>Pr(D,B)</td>
<td>Pr(D,W)</td>
</tr>
<tr>
<td>S</td>
<td>Pr(S,B)</td>
<td>Pr(S,W)</td>
</tr>
</tbody>
</table>
Marginalizing over "dottedness"

\[
\begin{array}{cc}
D & B \\
S & W
\end{array}
\]

\[
\begin{align*}
\Pr(D, B) & \quad \Pr(D, W) \\
\Pr(S, B) & \quad \Pr(S, W)
\end{align*}
\]

Marginal probability of being a white marble
Bayes' rule

The joint probability $\Pr(B,D)$ can be written as the product of a conditional probability and the probability of that condition.

\[
\Pr(B,D) = \Pr(B|D) \Pr(D) + \Pr(D|B) \Pr(B)
\]
Equate the two ways of writing $\Pr(B,D)$

$$\Pr(B|D) \Pr(D) = \Pr(D|B) \Pr(B)$$

Divide both sides by $\Pr(D)$

$$\frac{\Pr(B|D) \Pr(D)}{\Pr(D)} = \frac{\Pr(D|B) \Pr(B)}{\Pr(D)}$$

Bayes' rule

$$\Pr(B|D) = \frac{\Pr(D|B) \Pr(B)}{\Pr(D)}$$
Bayes' rule

\[
\frac{2}{5} = \frac{1}{\frac{1}{2}} \times \frac{\sqrt{3}}{5}
\]

\[
\frac{2}{\text{Bayes' rule}} = \frac{2}{\frac{2}{\text{Pr}(B|D) = \frac{\text{Pr}(D|B) \cdot \text{Pr}(B)}{\text{Pr}(D)}}}
\]

Copyright © 2019 Paul O. Lewis
Bayes' rule (variations)

\[
\Pr(B|D) = \frac{\Pr(D|B) \Pr(B)}{\Pr(D)} = \frac{\Pr(D|B) \Pr(B)}{\Pr(B, D) + \Pr(W, D)}
\]

\(\Pr(D)\) is the **marginal probability** of being dotted
To compute it, we **marginalize over colors**
Bayes' rule (variations)

\[
\Pr(B|D) = \frac{\Pr(D|B) \Pr(B)}{\Pr(B, D) + \Pr(W, D)}
\]

\[
= \frac{\Pr(D|B) \Pr(B)}{\Pr(D|B) \Pr(B) + \Pr(D|W) \Pr(W)}
\]

\[
= \frac{\Pr(D|B) \Pr(B)}{\sum_{\theta \in \{B, W\}} \Pr(D|\theta) \Pr(\theta)}
\]
Bayes’ rule in statistics

\[ P(\theta|D) = \frac{P(D|\theta) P(\theta)}{\sum_{\theta} P(D|\theta) P(\theta)} \]

- **Likelihood** of hypothesis \( \theta \)
- **Prior probability** of hypothesis \( \theta \)
- **Posterior probability** of hypothesis \( \theta \)
- **Marginal probability** of the data (marginalizing over hypotheses)
### Paternity example

\[ \Pr(\theta | D) = \frac{\Pr(D | \theta) \Pr(\theta)}{\sum_{\theta} \Pr(D | \theta) \Pr(\theta)} \]

<table>
<thead>
<tr>
<th></th>
<th>( \theta_1 )</th>
<th>( \theta_2 )</th>
<th>Row sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genotypes</td>
<td>AA</td>
<td>Aa</td>
<td>---</td>
</tr>
<tr>
<td>Prior</td>
<td>1/2</td>
<td>1/2</td>
<td>1</td>
</tr>
<tr>
<td>Likelihood</td>
<td>1</td>
<td>1/2</td>
<td>---</td>
</tr>
<tr>
<td>Prior X Likelihood</td>
<td>1/2</td>
<td>1/4</td>
<td>3/4</td>
</tr>
<tr>
<td>Posterior</td>
<td>2/3</td>
<td>1/3</td>
<td>1</td>
</tr>
</tbody>
</table>

(father) \( A^- \)

\( Aa \) (child)

aa (mother)
Bayes’ rule: continuous case

\[ p(\theta | D) = \frac{p(D | \theta)p(\theta)}{\int p(D | \theta)p(\theta)d\theta} \]

- Likelihood
- Prior probability \textit{density}
- Posterior probability \textit{density}
- Marginal probability of the data
If you had to guess...

Not knowing anything about my archery abilities, draw a curve representing your view of the chances of my arrow landing a distance $d$ centimeters from the center of the target.
Case 1: assume I have talent

An informative prior (low variance) that says most of my arrows will fall within 20 cm of the center (thanks for your confidence!)
Case 2: assume I have a talent for missing the target!

Also an *informative* prior, but one that says most of my arrows will fall within a narrow range just outside the entire target!
Case 3: assume I have no talent

This is a vague prior: its high variance reflects nearly total ignorance of my abilities, saying that my arrows could land nearly anywhere!
A matter of scale

Notice that I haven't provided a scale for the vertical axis.

What exactly does the height of this curve mean?

For example, does the height of the dotted line represent the probability that my arrow lands 60 cm from the center of the target?
Probabilities are associated with intervals

**Probabilities** are attached to **intervals** (i.e. ranges of values), **not** individual **values**

The probability of any given point (e.g. $d = 60.0$) is zero!

However, we can ask about the probability that $d$ falls in a particular interval e.g. $50.0 < d < 65.0$
Probabilities vs. probability densities

Note: the height of this curve does not represent a probability (if it did, it would not exceed 1.0)
Densities of various substances

<table>
<thead>
<tr>
<th>Substance</th>
<th>Density (g/cm³)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cork</td>
<td>0.24</td>
</tr>
<tr>
<td>Aluminum</td>
<td>2.7</td>
</tr>
<tr>
<td>Gold</td>
<td>19.3</td>
</tr>
</tbody>
</table>

*Density does not equal mass*

mass = density × volume
A brick with varying density

![Graph showing density variation from left to right end, with gold on the left and aluminum on the right.](image-url)
Integrating a density yields a probability

The density curve is scaled so that the value of this integral (i.e. the total area) equals 1.0
Integrating a density yields a probability

Area of rectangle = $p(\theta)d\theta$

$0.39109 = \int_{1}^{2} p(\theta)d\theta$

The area under the density curve from 1 to 2 is the probability that $\theta$ is between 1 and 2.
These density curves are all variations of a **gamma probability distribution**. We could have used a gamma distribution to specify each of the prior probability distributions for the archery example. Note that **higher variance** means **less informative**.
Usually there are many parameters...

A 2-parameter example

\[
p(\theta, \phi \mid D) = \frac{p(D \mid \theta, \phi) p(\theta) p(\phi)}{\int_{\theta} \int_{\phi} p(D \mid \theta, \phi) p(\theta) p(\phi) \, d\phi \, d\theta}
\]

Posterior probability density

Likelihood

Prior density

Marginal probability of data

An analysis of 100 sequences under the simplest model (JC69) requires 197 branch length parameters. The denominator would require a 197-fold integral inside a sum over all possible tree topologies! It would thus be nice to avoid having to calculate the marginal probability of the data...
Markov chain Monte Carlo (MCMC)
Markov chain Monte Carlo (MCMC)

For more complex problems, we might settle for a **good approximation** to the posterior distribution.
MCMC robot’s rules

- Uphill steps are always accepted
- Slightly downhill steps are usually accepted
- Drastic “off the cliff” downhill steps are almost never accepted

With these rules, it is easy to see why the robot tends to stay near the tops of hills.
Actual rules (Metropolis algorithm)

Uphill steps are always accepted because $R > 1$

Slightly downhill steps are usually accepted because $R$ is near 1

Drastic “off the cliff” downhill steps are almost never accepted because $R$ is near 0

The robot takes a step if a Uniform$(0,1)$ random deviate $\leq R$

Cancellation of marginal likelihood

When calculating the ratio \((R)\) of posterior densities, the marginal probability of the data cancels.

\[
\frac{p(\theta^* | D)}{p(\theta | D)} = \frac{\frac{p(D | \theta^*) p(\theta^*)}{p(D)}}{\frac{p(D | \theta) p(\theta)}{p(D)}} = \frac{p(D | \theta^*) p(\theta^*)}{p(D | \theta) p(\theta)}
\]

Posterior odds

Apply Bayes' rule to both top and bottom

Likelihood ratio

Prior odds
Target vs. Proposal Distributions

"good" proposal distribution

The proposal distribution is used by the robot to choose the next spot to step, and is separate from the target distribution.

The target is usually the posterior distribution

White noise appearance is a sign of good mixing

Tracer (app for generating trace plots from MCMC output): https://github.com/beast-dev/tracer/releases/tag/v1.7.1

trace plot

MCMC iteration

log(posterior)

0 200 400 600 800 1000

-16 -15 -14 -13 -12 -11 -10
Target vs. Proposal Distributions

"baby steps" proposal distribution

target distribution

Big waves in trace plot indicate robot is crawling around

MCMC iteration

log(posterior)

0 200 400 600 800 1000

−16 −15 −14 −13 −12 −11 −10 −9
Target vs. Proposal Distributions

"overly bold" proposal distribution

Plateaus in trace plot indicate robot is often stuck in one place
MCRobot (or "MCMC Robot")

Javascript version used today will run in most web browsers and is available here:

https://phylogeny.uconn.edu/mcmc-robot/

Free app for Windows or iPhone/iPad available from http://mcmicrobot.org/

(also see John Huelsenbeck’s iMCMC app for MacOS: http://cteg.berkeley.edu/software.html)
Metropolis-coupled Markov chain Monte Carlo (MCMC)

Sometimes the robot needs some help,

MCMC introduces helpers in the form of "heated chain" robots that can act as scouts.

Heated chains act as scouts for the cold chain

**Cold chain robot** can easily make this jump because it is uphill

**Hot chain robot** can also make this jump with high probability because it is only slightly downhill
Heated chains act as scouts for the cold chain

Swapping places means both robots can cross the valley, but this is more important for the cold chain because its valley is much deeper.
The Hastings ratio

Suppose the probability of proposing a spot to the right is twice that of proposing a spot to the left.

The Hastings ratio

Example in which proposals were biased toward due east, but Hastings ratio was not used to modify acceptance probabilities.
The Hastings ratio

Accepting half as many right proposals as left proposals serves to balance the proposal probabilities.

Hastings Ratio

\[ R = \frac{p(D | \theta^*) \, p(\theta^*)}{p(D | \theta) \, p(\theta)} \cdot \frac{q(\theta | \theta^*)}{q(\theta^* | \theta)} \]

posterior ratio \quad Hastings ratio

Note that the Hastings ratio is 1.0 if \( q(\theta^* | \theta) = q(\theta | \theta^*) \)