Missing data & hidden Markov models

Alex Cook

Week 10
Bayesian secret weapon

Point 1:
Bayesians can write probability statements about anything: 
\( f(A|B) \) makes sense if
- \( A \) is the data, \( B \) an unknown
- \( B \) is an unknown, \( A \) the data
Bayesian secret weapon

Point 2:

\[
f(A|B) = \frac{f(B|A)f(A)}{f(B)}
\]

\[
f(A) = \int f(A|B)f(B) \, dB
\]

\[
f(A, B) = f(A|B)f(B)
\]
Bayesian secret weapon

Extending the conversation:

- You want to evaluate the likelihood \( f(\text{data}|\theta) \)
- No nice form: cannot evaluate directly
- But you can introduce imaginary data \( E \) to get

\[
    f(\text{data}|\theta) = \int f(\text{data}|E, \theta) f(E|\theta) \, dE
\]

- If do it wisely, both \( f(\text{data}|E, \theta) \) and \( f(E|\theta) \) are easy to evaluate
- This leaves an integral :-(
Bayesian secret weapon

The integral:

- When you do MCMC you are integrating already!
- Think back

\[ f(\theta | \text{data}) \propto f(\text{data} | \theta) f(\theta) \]

- Actually it’s

\[ f(\theta | \text{data}) = \frac{f(\text{data} | \theta) f(\theta)}{f(\text{data})} \]

- I.e.

\[ f(\theta | \text{data}) = \frac{f(\text{data} | \theta) f(\theta)}{\int f(\text{data} | \vartheta) f(\vartheta) d\vartheta} \]
Bayesian secret weapon

Extending the conversation:

- Introduce imaginary data $E$ to get

  \[
  f(\text{data}|\theta) = \int f(\text{data}|E, \theta)f(E|\theta) \, dE
  \]

- Do MCMC (say) on $(E, \theta)$ and then look only at the sample of $\theta$
Eg1: fitting stochastic processes

Simple epidemic model

- $S(t) =$ number of people susceptible at time $t$
- $I(t) =$ number of people infectious at time $t$
- $R(t) =$ number of people recovered at time $t$
- $S(t) + I(t) + R(t) = n$, $S(t) \in \mathbb{N}$, $I(t) \in \mathbb{N}$, $R(t) \in \mathbb{N}$
Eg1: fitting stochastic processes

\[
\begin{pmatrix}
S(t + \delta) &= S(t) - 1 \\
I(t + \delta) &= I(t) + 1 \\
R(t + \delta) &= R(t)
\end{pmatrix}
\begin{pmatrix}
S(t) \\
I(t) \\
R(t)
\end{pmatrix}
= \beta S(t)I(t)\delta + o(\delta)
\]

\[
\begin{pmatrix}
S(t + \delta) &= S(t) \\
I(t + \delta) &= I(t) - 1 \\
R(t + \delta) &= R(t) + 1
\end{pmatrix}
\begin{pmatrix}
S(t) \\
I(t) \\
R(t)
\end{pmatrix}
= \gamma I(t)\delta + o(\delta)
\]
Eg1: fitting stochastic processes

SIR model with transition arrows between S, I, and R. Graphs show the dynamics of S(t), I(t), and R(t) over time.

Alex Cook, ST5219, Bayesian Hierarchical Modelling
Example model

San Francisco 1918 flu
(thousands)

S(t)
I(t)
R(t)

weeks

S(t)
I(t)
R(t)

time
### Approach

- **B symptoms**
  - A → B
  - B → C
  - B → D

- **B to GP**
  - B isolated
  - B recovers

- **B symptoms**
  - B → C
  - B to GP
  - B isolated
  - B recovers

#### Ideally:

- A → B: 12.13 Wed
- B symptoms: 06.59 Thu
- B → C: 15.43 Thu
- B → D: 13.05 Sat
- B visits GP: 09.20 Mon
- B isolated: 11.48 Mon
- B recovers: 12.15 Wed

#### Realistically:

- B symptoms: Thu/Fri
- B → C: ???
- B visits GP: Mon
- B isolated: Mon
- B recovers: Wed
Approach

\( \theta \) becomes \((\theta, E)\):

- \( \theta = \) unknown parameters
- \( E = \) unknown event times

\[
f(\theta|\text{data}) \propto \int f(\text{data}|E, \theta) f(E|\theta) f(\theta) \, dE
\]

- If the set of event times are consistent with the data
  \( f(\text{data}|E, \theta) = 1 \) and so disappears
- \( f(E|\theta) \) can be derived explicitly by calculating hazard and survival functions for infection, and straightforwardly for within host events (recovery in this model)
Application: Tristan da Cunha

Population: 271
Nearest population: 2000km
Application: Tristan da Cunha

Why do epidemics differ?

- Just chance (same $\theta$)?
- Differences in pathogen (different $\theta$)?
Application: Tristan da Cunha

\[ \beta \text{ per SI infection rate, } \gamma^{-1} \text{ mean infectious period} \]
Application: Tristan da Cunha

$\beta$ per SI infection rate, $\gamma^{-1}$ mean infectious period
Application: Tristan da Cunha

\[ \beta \] per SI infection rate, \( \gamma^{-1} \) mean infectious period
Application: Tristan da Cunha

\[ \beta \text{ per SI infection rate, } \gamma^{-1} \text{ mean infectious period} \]
Opinion poll in Slovenia
Opinion poll in Slovenia
Slovenia and Croatia were the first Yugoslav republics to break off from Yugoslavia.

In June 1991 the Slovene government declared independence, followed by a relatively bloodless 10d war.

Slovenia now increasingly prosperous, EU, eurozone and NATO member state.

Critical in the run up to declaration of independence was a plebiscite on independence in Dec 1990.
Opinion poll in Slovenia

Rules of the plebiscite:

- All 1.46M electorate could vote
- A majority of 50% of eligible voters needed to vote yes for it to pass
- I.e. if you didn’t vote, you were counted as voting against independence
Opinion poll in Slovenia

Before the plebiscite:

- Slovene government inserted Qs on voting intentions in the Slovenian Public Opinion Survey (SPOS) 1mo before.
- SPOS is a face to face survey of around 2000 adults.
- Nonresponse fairly low (about 10%).
- Actual design is quite complex—we will assume it’s a simple random sample for this lecture (see Rubin et al, 1995, J Am Stat Assoc 90:822–8 for a fuller analysis).

3 questions:

1. Do you want independence?
2. Do you want secession?
3. Will you vote?

3 answers:

1. Yes
2. No
3. Dunno!
Data: Yes, I’ll vote

<table>
<thead>
<tr>
<th>Independence</th>
<th>No</th>
<th>?</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>68</td>
<td>29</td>
<td>158</td>
</tr>
<tr>
<td>Secession</td>
<td>8</td>
<td>109</td>
<td>90</td>
</tr>
<tr>
<td>Yes</td>
<td>8</td>
<td>21</td>
<td>1191</td>
</tr>
</tbody>
</table>
## Data: Don’t know if I’ll vote

<table>
<thead>
<tr>
<th>Seccession</th>
<th>Independence</th>
<th>No</th>
<th>?</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td></td>
<td>43</td>
<td>31</td>
<td>18</td>
</tr>
<tr>
<td>Yes</td>
<td></td>
<td>3</td>
<td>9</td>
<td>107</td>
</tr>
</tbody>
</table>
Data: No, I won’t vote

<table>
<thead>
<tr>
<th>Seccession</th>
<th>Independence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
</tr>
<tr>
<td>No</td>
<td>14</td>
</tr>
<tr>
<td>?</td>
<td>2</td>
</tr>
<tr>
<td>Yes</td>
<td>0</td>
</tr>
</tbody>
</table>
Data:

<table>
<thead>
<tr>
<th>V: No</th>
<th>I</th>
<th></th>
<th>V: ?</th>
<th>I</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
<td>14</td>
<td>3</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>?</td>
<td>2</td>
<td>25</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>0</td>
<td>4</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>43</td>
<td>31</td>
<td>18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>?</td>
<td>8</td>
<td>96</td>
<td>19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>3</td>
<td>9</td>
<td>107</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

V: Yes | I     |   | V: ?  | I    |   |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
<td>68</td>
<td>29</td>
<td>158</td>
<td></td>
</tr>
<tr>
<td></td>
<td>?</td>
<td>8</td>
<td>109</td>
<td>90</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>8</td>
<td>21</td>
<td>1191</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Analysis 1

- Throw away the secession Q
- Throw away the “don’t knows” for independence and voting Qs
- 1549 left
- 1439 said they will vote and support independence
- $X \sim \text{Bin}(1549, p), \ p \sim \text{Be}(1, 1), \ p|X \sim \text{Be}(1440, 111)$
- Posterior mean 93%, 95%CI (92%,94%)
Analysis 2

- Throw away the secession Q
- Merge the “don’t knows” for independence and voting Qs into “Noes”
- 2074 interviewees
- 1439 said they will vote and support independence
- \( X \sim \text{Bin}(2074, p), \ p \sim \text{Be}(1, 1), \ p|X \sim \text{Be}(1440, 635) \)
- Posterior mean 69%, 95%CI (67%,71%)
Don’t throw away or merge data

- Conceptualise the eventual decision (vote, don’t vote, support, don’t support)

- Assume there is a probability for every individual on every Q not to have reached that eventual decision yet

- Assume probability $u_i$ of being unsure of independence, $u_s$ of secession, and $u_v$ of voting: assumed to be independent of final decision and other intentions
If you knew everyone’s eventual decision, $D$, and questionnaire answers, $A$, likelihood would be easy:

$$f(D, A|p, u) = f(A|D, p, u)f(D|p, u) = f(A|D, u)f(D|p)$$

where

- $f(A|D, U)$ is trivial product of binomials
- $f(D|p)$ is multinomial

So, posterior:

$$f(p, u|A) \propto \int f(p, u)f(A|D, u)f(D|p) \, dD$$
Note: sufficient to know the numbers of people in each of 27 \times 8 = 216 categories

<table>
<thead>
<tr>
<th></th>
<th>nnn</th>
<th>nyn</th>
<th>ynn</th>
<th>yyn</th>
<th>nny</th>
<th>nyy</th>
<th>yny</th>
<th>yyy</th>
</tr>
</thead>
<tbody>
<tr>
<td>yyn</td>
<td></td>
<td></td>
<td></td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td>1191</td>
</tr>
<tr>
<td>yyn</td>
<td></td>
<td></td>
<td></td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td>99</td>
</tr>
<tr>
<td>yyn</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>yyn</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>
Snippets of the code

# for truth
current$loglikelihood =
dmultinom(colSums(current$truth), prob=current$p, log=TRUE)
# for observation
for(col in 1:8)
{
  rows = (1:27)[current$possible[,col]]
  v = current$v[col]; s = current$s[col]; i = current$i[col]
  uv = current$u_v; ui = current$u_i; us = current$u_s
  repv = obdata$v[rows]
  reps = obdata$s[rows]
  repi = obdata$i[rows]
  n = current$truth[rows, col]
  p = (((repv==0)*uv) + ((repv!=0)*(1-uv)))*(...
  current$loglikelihood = current$loglikelihood +
  dmultinom(n, prob=p, log=TRUE)
}
for(k in 1:7)
{
    old=current
    current$p[k]=rnorm(1,current$p[k],pmutator[k])
    current=mh(current, data, old)
}
Snippets of the code

for(row in 1:27)
{
    cols=(1:8)[current$possible[row,]]
    ncols=length(cols)
    if(ncols>1)
    {
        for(index in 1:(ncols-1))
        {
            old=current
            col=cols[index]
            al=round(rnorm(1,0,0.05*obsdata$n[row]))
            current$truth[row,col]=current$truth[row,col]+a
            current$truth[row,cols[ncols]]=current$truth[row,cols[ncols]]-a
            current=mh(current,data,old)
        }
    }
}
- $p_{000}, p_{001}, \ldots, p_{111}$: in particular $\pi = p_{110} + p_{111}$ is the proportion eventually to vote in favour of independence.
- $u_i, u_s, u_v$: proportions uncertain by question type.
- $p_i$ estimated to be 88% with 95% CI 86% to 89%.
- $u$ about 15% for all Qs.
Comparison

Pretend:

- “don’t know”’s typical: 93% (92%, 94%)
- “don’t know”’s against: 69% (67%, 71%)
- \( p(“don’t know”) \) constant: 88% (86%, 89%)

Results

88.5% if voters turned up and voted yes
Eg3: detecting homogeneous DNA

This example is derived from Boys et al (2000) *Appl Stat* 49:269–85—an excellent paper!

**Brief (wrong?) intro to DNA**

- DNA represented by letters ACGT: these code up what happens in your body
- Triplets of bases are important
- Your code has exons—that code for stuff—and introns—that don’t do anything
- Introns are good for recording mutations and hence phylogenies
Eg3: detecting homogeneous DNA

- Boys et al want to be able to detect segments in introns that are inhomogeneous: i.e. the proportions of different bases or the resulting triplets are different.
- They consider intron 7 of the chimpanzee’s α-fetoprotein gene which has an important role in embryo development.

```plaintext
gtgaagagtcctgcttcttaaaaaagatgttttcactccctttttctttcttttgtctcattctaaagggagaa
gtgggttttgacttgaattttggtacagaggtatgtaaaaactaggtgattctcttaaatggcagaaattctc
gatagcagagttaacatcttttgttgatcctgctttcacttttagctataaccccttttttgtgaacaaagggc
catctcttaaaaaaccatggggaaccttctcaagaggccttctccatagttacttggaggagggaggaactaagt
ttaatgtatattatattttcattcattttattcttttccattgacaaataataatataattaaataattttctat
tcgctagccacatagcagacactttgtgttaaaagcacagggctgacctcggaggaatttacagtcgtataggagagataaga
cagtgacctcctttgagttgggactgcctttgctttactgttatctccatatgcacaatgcctggcacatggaggc
atctataatagtttggattaatgaacgaatacaataaaatagcagaaagtgaactgtcctaccaggtaaaaagctag
catgccaaagacaaagtgtgaataaagtggtctcggagaatttagaaaaaacaatattaatcccaacgcatggagtttctttgagtgtgtctctagtcagtgacttaagcagtgtggcatttggtttatatgctaatccctagactttcctaga...
```
Data: moving averages

- Proportion A
- Proportion C
Data: moving averages

- **Proportion G**
  - Values range from 0.0 to 0.6
  - Observations at 0, 500, 1000, 1500, 2000

- **Proportion T**
  - Values range from 0.0 to 0.6
  - Observations at 0, 500, 1000, 1500, 2000

Alex Cook, ST5219, Bayesian Hierarchical Modelling
Data: cusums
Imagine no heterogeneous sections

\( Y_i \in \{a, c, g, t\} \) is the base at site \( i \)

Assume a first-order Markov chain between \( Y_i \) and \( Y_{i+1} \) with transition matrix

\[
P = \begin{pmatrix}
  p_{aa} & p_{ac} & p_{ag} & p_{at} \\
  p_{ca} & p_{cc} & p_{cg} & p_{ct} \\
  p_{ga} & p_{gc} & p_{gg} & p_{gt} \\
  p_{ta} & p_{tc} & p_{tg} & p_{tt}
\end{pmatrix}
\]
Imagine $r$ heterogeneous sections

- $S_i \in \{1, 2, \ldots, r\}$ is the segment type at site $i$

Assume a first-order Markov chain between $S_i$ and $S_{i+1}$ with transition matrix

$$\Lambda = (\lambda_{st})$$
Imagine $r$ heterogeneous sections

$Y_i \in \{a, c, g, t\}$ is the base at site $i$

Assume a first-order Markov chain between $Y_i$ and $Y_{i+1}$ with transition matrix

$$P^S_i = \begin{pmatrix}
p^{S_i}_{aa} & p^{S_i}_{ac} & p^{S_i}_{ag} & p^{S_i}_{at} \\
p^{S_i}_{ca} & p^{S_i}_{cc} & p^{S_i}_{cg} & p^{S_i}_{ct} \\
p^{S_i}_{ga} & p^{S_i}_{gc} & p^{S_i}_{gg} & p^{S_i}_{gt} \\
p^{S_i}_{ta} & p^{S_i}_{tc} & p^{S_i}_{tg} & p^{S_i}_{tt}
\end{pmatrix}$$
Model: DAG

- If you knew \( S \), likelihood trivial
- Don’t know \( S \)
- Extend the conversation to include \( S \) and integrate it out
Priors

Assume:

\[
P_{S_i} = \begin{pmatrix}
p_{aa}^{S_i} & p_{ac}^{S_i} & p_{ag}^{S_i} & p_{at}^{S_i} \\
p_{ca}^{S_i} & p_{cc}^{S_i} & p_{cg}^{S_i} & p_{ct}^{S_i} \\
p_{ga}^{S_i} & p_{gc}^{S_i} & p_{gg}^{S_i} & p_{gt}^{S_i} \\
p_{ta}^{S_i} & p_{tc}^{S_i} & p_{tg}^{S_i} & p_{tt}^{S_i}
\end{pmatrix}
\]

is Dirichlet for each row a priori (ditto for \(\lambda\))
- Assume $\lambda$'s rows are Dirichlet: but *informative* as we expect the probability $S_i = S_{i+1}$ to be high
- Suggest doing it arbitrarily and comparing sensitivity of results
- (Read paper for details)
Posterior

- **Conditional** on the unobserved states $S$, both lots of transition matrices have simple Dirichlet posteriors
- Boys et al use a Gibbs sampler: iteratively
  - Simulating $p$ and $\lambda$ from their conditional distributions given $S$
  - Simulating $S$ given $Y$ and $\lambda$
- Sampling $S$ is challenging: their solution is to use a clever backward sweep starting at the end and working back to the start of the sequence
- Because it’s Gibbs sampling, proposals are accepted automatically, so there is no direct calculation of the posterior
Results: 2 sections

Type 1:

```
intro eg1 eg2 eg3
```

```
<table>
<thead>
<tr>
<th>Type 1:</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
</tr>
<tr>
<td>T</td>
</tr>
<tr>
<td>A</td>
</tr>
<tr>
<td>C</td>
</tr>
<tr>
<td>G</td>
</tr>
<tr>
<td>A</td>
</tr>
<tr>
<td>G</td>
</tr>
<tr>
<td>T</td>
</tr>
<tr>
<td>C</td>
</tr>
<tr>
<td>G</td>
</tr>
<tr>
<td>A</td>
</tr>
<tr>
<td>T</td>
</tr>
<tr>
<td>T</td>
</tr>
<tr>
<td>C</td>
</tr>
<tr>
<td>G</td>
</tr>
</tbody>
</table>

---

Alex Cook, ST5219, Bayesian Hierarchical Modelling
Results: 2 sections

Type 2:
Section locations

![Graph showing probability and change point over time]

- Probability state 1
- Change point?
Section locations

![Graph showing CUSUM and location](image)
λ prior sensitivity

![Graph showing the sensitivity of the prior parameter λ.

The graph plots the probability segment 1 against a range of values on the x-axis, with distinct curves representing different levels of prior sensitivity. The y-axis indicates the probability, ranging from 0.0 to 1.0. The graph visually demonstrates how varying levels of prior sensitivity affect the probability distribution.](image-url)
Lots of extra difficulties

- How many segments to allow?
- How to deal with multimodality?
- ...
- (Read paper to learn more!)