Lecture 5.
Brief introduction to simulation-based Bayesian inference

Summary

1. Single parameter models
2. Multi-parameter models
3. Gibbs sampling
4. Relation to directed graphical model (and therefore BUGS program)

Huge topic, can only touch surface

Bayesian analysis

Why is computation important?

• Bayesian inference centres around the posterior distribution
  \[ p(\theta|y) \propto p(y|\theta) \times p(\theta) \]
  where \( \theta \) is typically a large vector of parameters
  \( \theta = \{\theta_1, \theta_2, \ldots, \theta_k\} \)
• \( p(y|\theta) \) and \( p(\theta) \) will often be available in closed form, but \( p(\theta|y) \)
  is usually not analytically tractable, and we want to
  - obtain marginal posterior \( p(\theta_i|y) = \int \cdots \int p(\theta|y) \, d\theta_{(-i)} \) where
    \( \theta_{(-i)} \) denotes the vector of \( \theta \)'s excluding \( \theta_i \)
  - calculate properties of \( p(\theta_i|y) \), such as mean
    \( (= \int \theta_i p(\theta_i|y)d\theta_i) \), tail areas \( (= \int_{-\infty}^\infty p(\theta_i|y)d\theta_i) \) etc.
  \( \rightarrow \) numerical integration becomes vital

Broad principles of Bayesian analysis in graphical modelling software (e.g. BUGS)

1. Specify directed acyclic graph as set of conditional distributions or assignments
2. Specify data: ‘clamps’ specified nodes to those values
3. Program (maybe with some control) works out how to sample
   from conditional distributions of other nodes, given data
4. Specify initial values, decide what to monitor etc
5. Run simulation, checking convergence etc,
6. Report results
7. Check appropriateness of model, sensitivity of conclusions to
   prior assumptions etc
Direct sampling in single parameter models

- We need to directly sample from a distribution proportional to the product of the likelihood and the prior.
- In some situations we can work out the full posterior: e.g.
  - Discrete parameter
  - Conjugate inference, then use standard sampling algorithms, e.g. inversion
- Otherwise we can use sampling algorithms that only require a function proportional to the desired distribution. e.g.
  - Ratio-of-uniforms
  - Metropolis-Hastings
  - Adaptive rejection sampling
  - Slice sampling

Example: use of Bayes theorem in diagnostic testing

A new HIV test is claimed to have “95% sensitivity and 98% specificity”. In a population with an HIV prevalence of 1/1000, what is the chance that patient testing positive actually has HIV?

- Let $A$ be the event that patient is truly HIV positive, $\overline{A}$ be the event that they are truly HIV negative.
- Let $B$ be the event that they test positive.
- We want $p(A|B)$.
- “95% sensitivity” means that $p(B|A) = .95$.
- “98% specificity” means that $p(B|\overline{A}) = .02$.
- Now Bayes theorem says
  $$p(A|B) = \frac{.95 \times .001}{.95 \times .001 + .02 \times .999} = .045.$$  
- Thus over 95% of those testing positive will, in fact, not have HIV.

Easier to explain as expected outcomes in a large number of cases

<table>
<thead>
<tr>
<th></th>
<th>HIV - ($\overline{A}$)</th>
<th>HIV + ($A$)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test - ($\overline{B}$)</td>
<td>97,902</td>
<td>5</td>
<td>97,907</td>
</tr>
<tr>
<td>Test + ($B$)</td>
<td>1998</td>
<td>95</td>
<td>2093</td>
</tr>
<tr>
<td>Total</td>
<td>99,900</td>
<td>100</td>
<td>100,000</td>
</tr>
</tbody>
</table>

$$p(A|B) = \frac{95}{2093} = 0.045$$

Can (rather unnecessarily) use BUGS

```r
B <- 1

###############
A ~ dbern(0.001)
Aplus1 <- A+1 # change 0/1 to 1/2
B ~ dbern(p[Aplus1]) # Aplus1 'picks' the appropriate cell of p[

p[1] <- 0.02
p[2] <- 0.95
```

- Cannot use `dbern(p[A+1])`, as functions of random quantities not allowed as indices
- `B` appears twice on the left-hand side of a statement - BUGS allows data transformations

<table>
<thead>
<tr>
<th>node</th>
<th>mean</th>
<th>sd</th>
<th>MC error</th>
<th>2.5%</th>
<th>median</th>
<th>97.5%</th>
<th>start</th>
<th>sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.04607</td>
<td>0.2096</td>
<td>6.373E-4</td>
<td>0.00</td>
<td>0.00</td>
<td>1.0</td>
<td>1</td>
<td>100000</td>
</tr>
</tbody>
</table>

5-5

5-6

5-7

5-8
Bayesian analysis

Inference about a discrete parameter

Suppose I have 3 coins in my pocket,

1. biased 3:1 in favour of tails
2. a fair coin,
3. biased 3:1 in favour of heads

I randomly select one coin and toss it once, observing a head. What is the probability that I have chosen coin 3?

- Let \( y = 1 \) denote the event that I observe a head
- \( \theta \) denote the probability of a head: \( \theta \in (0.25, 0.5, 0.75) \)
- Prior: \( p(\theta = 0.25) = p(\theta = 0.5) = p(\theta = 0.75) = 0.33 \)
- Sampling distribution: \( p(y|\theta) = \theta^y(1 - \theta)^{1-y} \)

| Coin | \( p(\theta) \) | \( p(y = 1|\theta) \) | \( p(y = 1|\theta)p(\theta) \) | \( \frac{p(y = 1|\theta)p(\theta)}{p(y)} \) |
|------|----------------|----------------|----------------|------------------|
| 1    | 0.25           | 0.33           | 0.0825         | 0.167            |
| 2    | 0.50           | 0.33           | 0.1650         | 0.333            |
| 3    | 0.75           | 0.33           | 0.2475         | 0.500            |
| Sum  | 1.00           | 1.50           | 0.495          | 1.000            |

\( \dagger \) The normalising constant can be calculated as \( p(y) = \sum_i p(y|\theta_i)p(\theta_i) \)

So observing a head on a single toss of the coin means that there is now a 50% probability that the chance of heads is 0.75 and only a 16.7% probability that the chance of heads in 0.25.

In BUGS

```r
  y <- 1
  ###############
  coin ~ dcat(p[])  # categorical variable taking on values 1,2,...
  theta.true <- theta[coin]
  y ~ dbern(theta.true)  # could use theta[coin] directly instead
  for(i in 1:3){
    p[i] <- 1/3
    theta[i] <- 0.25*i
    coin.prob[i] <- equals(coin,i)
  }
```

Doodle

```
node mean sd MC error 2.5% median 97.5% start sample
coin.prob[1] 0.1662 0.3723 0.001141 0.0 0.0 1.0 1 100000
coin.prob[2] 0.3342 0.4717 0.001435 0.0 0.0 1.0 1 100000
coin.prob[3] 0.4997 0.5 0.001491 0.0 0.0 1.0 1 100000
```
Bayesian analysis

Surgery: beta-binomial analysis using BUGS

Remember: a beta(3,27) prior, then observing 0/10 deaths, then wanting the probability of at least 2 deaths in the next 20 patients.

The BUGS syntax for sampling of the surgery posterior and predictive distribution for the next 20 patients is

\[
y \sim 0 \\
\theta \sim \text{dbeta}(3,27) \quad \# \text{prior distribution} \\
y \sim \text{dbin}(\theta,10) \quad \# \text{sampling distribution} \\
Y_{\text{pred}} \sim \text{dbin}(\theta,20) \quad \# \text{predictive distribution} \\
P_{\text{crit}} \leftarrow \text{step}(Y_{\text{pred}}-1.5) \quad \# = 1 \text{ if } y_{\text{pred}} \geq 2, \ 0 \text{ otherwise}
\]

Alternatively, the data and prior parameters could be included in a data statement

```
list(a=3, b=27, y=0, n=10, n.pred=20, n.crit=2)
```

Node statistics

<table>
<thead>
<tr>
<th>node</th>
<th>mean</th>
<th>sd</th>
<th>MC error</th>
<th>2.5%</th>
<th>median</th>
<th>97.5%</th>
<th>start</th>
<th>sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>P.crit</td>
<td>0.4174</td>
<td>0.4931</td>
<td>0.001565</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>1001</td>
<td>100000</td>
</tr>
<tr>
<td>Y.prd</td>
<td>1.498</td>
<td>1.427</td>
<td>0.004439</td>
<td>0.0</td>
<td>1.0</td>
<td>5.0</td>
<td>1001</td>
<td>100000</td>
</tr>
<tr>
<td>eq[1]</td>
<td>0.2827</td>
<td>0.4503</td>
<td>0.001493</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>1001</td>
<td>100000</td>
</tr>
<tr>
<td>eq[2]</td>
<td>0.2999</td>
<td>0.4582</td>
<td>0.001413</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>1001</td>
<td>100000</td>
</tr>
<tr>
<td>eq[3]</td>
<td>0.2076</td>
<td>0.4056</td>
<td>0.001207</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>1001</td>
<td>100000</td>
</tr>
<tr>
<td>....</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>theta</td>
<td>0.07513</td>
<td>0.04132</td>
<td>1.319E-4</td>
<td>0.01612</td>
<td>0.06794</td>
<td>0.1738</td>
<td>1001</td>
<td>100000</td>
</tr>
</tbody>
</table>

Coin tossing: Suppose we are told that a fair coin has come up heads \( y = 10 \) times. How many times has the coin been tossed?

The likelihood for \( N \) is

\[
p(y|N) = \text{Binomial}(0.5, N) \propto \frac{N!}{(N-y)!} \cdot 0.5^N
\]

Suppose we wish to assign a uniform prior over integer values from 1 to 100, i.e. \( \Pr(N = n) = 1/100, \ n = 1, ..., 100 \) (will consider other prior distributions later)

Then the posterior for \( N \) is proportional to the likelihood, and its expectation, for example, is given by

\[
E[N|y] = \sum_{n=1}^{100} n \Pr(N = n|y) = A \sum_{n=1}^{100} \frac{n \times n!}{(n-y)!}0.5^n,
\]

where \( A \) is the posterior normalising constant
Bayesian analysis

Can solve this numerically or reasonably straightforward to sample directly from this

\begin{align*}
    y &\sim \text{dbin}(0.5, N) \\
    N &\sim \text{dcat}(p[i])
\end{align*}

for (i in 1:100) {p[i] <- 1/100}

The mode is 20, which is the intuitive answer (and the MLE)

<table>
<thead>
<tr>
<th>mode</th>
<th>mean</th>
<th>sd</th>
<th>MC error</th>
<th>2.5%</th>
<th>median</th>
<th>97.5%</th>
<th>start</th>
<th>sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>21.07</td>
<td>4.761</td>
<td>0.03929</td>
<td>13.0</td>
<td>21.0</td>
<td>32.0</td>
<td>1001</td>
<td>10000</td>
</tr>
</tbody>
</table>

Bayesian analysis

Sampling from multivariate distributions

- We want samples from joint posterior distribution \( p(\theta|y) \)
- General ‘independent’ sampling not straightforward
- **Dependent** sampling from a Markov chain with \( p(\theta|y) \) as its stationary (equilibrium) distribution
- A sequence of random variables \( \theta(0), \theta(1), \theta(2), \ldots \) forms a Markov chain if \( \theta(i+1) \sim p(\theta|\theta(i)) \) i.e. conditional on the value of \( \theta(i) \), \( \theta(i+1) \) is independent of \( \theta(i-1), \ldots, \theta(0) \)
- Several standard ‘recipes’ available for designing Markov chains with required stationary distribution \( p(\theta|y) \)
  - Metropolis et al. (1953); generalised by Hastings (1970)
  - **Gibbs Sampling** (see Geman and Geman (1984), Gelfand and Smith (1990), Casella and George (1992)) is a special case of the Metropolis-Hastings algorithm
  - See Gilks, Richardson and Spiegelhalter (1996)

Bayesian analysis

Gibbs sampling

Let our vector of unknowns \( \theta \) consist of \( k \) sub-components \( \theta = (\theta_1, \theta_2, \ldots, \theta_k) \)

1) Choose starting values \( \theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_k^{(0)} \)

2) Sample \( \theta_1^{(1)} \) from \( p(\theta_1|\theta_2^{(0)}, \theta_3^{(0)}, \ldots, \theta_k^{(0)}, y) \)
   Sample \( \theta_2^{(1)} \) from \( p(\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, \ldots, \theta_k^{(0)}, x) \)
   ..... 
   Sample \( \theta_k^{(1)} \) from \( p(\theta_k|\theta_1^{(1)}, \theta_2^{(1)}, \ldots, \theta_{k-1}^{(1)}, y) \)

3) Repeat step 2 many 1000s of times
   - eventually obtain sample from \( p(\theta|y) \)

The conditional distributions are called ‘full conditionals’ as they condition on all other parameters
**Gibbs sampling and directed graphical models**

By structuring a DAG out of parameters \( \theta \) and data \( y \) we are assuming

\[
p(\theta, y) = \prod_{v \in G} p(v | \text{pa}[v])
\]

Now \( p(\theta | y) = p(\theta, y) / p(y) \propto p(\theta, y) \) when considered as a function of \( \theta \)

Similarly, any conditional distribution involving all nodes in the graph is also proportional to \( p(\theta, y) \)

**In particular**

\[
p(\theta_i | \theta \setminus \theta_i, y) \propto p(\theta_i | \text{pa}[\theta_i]) \times \prod_{v \in \text{ch}[\theta_i]} p(v | \text{pa}[v]) :
\]

as only terms in the RHS containing \( \theta_i \) are relevant

so the full conditional is dependent only on \( \text{pa}[\theta_i], \text{ch}[\theta_i] \) and all co-parents of \( \theta_i \)'s children: the Markov blanket

\( \theta_i \) is conditionally independent of all other nodes in the graph, given the Markov blanket

BUGS is based on identifying nodes in the Markov blanket for each node \( \theta_i \), and sampling \( \theta_i \) conditional on their current values using the product form above

**Gibbs sampling for normal data** Observe \( y_1, \ldots, y_n \) from a Normal distribution with unknown mean \( \mu \) and unknown precision \( \omega \). Further suppose that we specify independent priors on \( \mu \) and \( \omega \) as follows:

\[
\mu \sim \text{Normal}(\gamma, \tau^2); \quad \omega \sim \text{Gamma}(\alpha, \beta).
\]

These are the conjugate priors for the cases in which the precision and the mean, respectively, have known values, but are not jointly conjugate.

Gibbs sampler.

1. Choose arbitrary starting values, \( \mu^{(0)} = 5 \) and \( \omega^{(0)} = 10 \), say
2. Choose an equally arbitrary updating order, \( \mu \) then \( \omega \), say.
3. At iteration \( t \) of the Gibbs sampler we draw \( \mu^{(t)} \sim p(\mu | \omega^{(t-1)}, y) \) and \( \omega^{(t)} \sim p(\omega | \mu^{(t)}, y) \)
The full conditional for $\mu$ is proportional to the prior for $\mu$ multiplied by the distribution of each child of $\mu$ conditional on that child’s parents. From the graph, $\text{ch}[\mu] = \{ y_i, i = 1, \ldots, n \}$, and so

$$p(\mu | \omega^{(t-1)}, y) \propto \exp\left\{ -\frac{1}{2\tau^2}(\mu - \gamma)^2 \right\} \times \exp\left\{ -\frac{\omega^{(t-1)}}{2} \sum_{i=1}^{n} (y_i - \mu)^2 \right\}.$$  

Conditioning on $\omega = \omega^{(t-1)}$ is essentially the same as assuming $\omega$ to be known. Hence this is exactly the same type of calculation as is required for deriving the single-parameter posterior (for $\mu$) in the unknown mean, known precision/variance case.

Therefore,

$$p(\mu | \omega^{(t-1)}, y) = \text{Normal} \left( \frac{\omega^{(t-1)} \sum y_i + \tau^{-2} \gamma}{n\omega^{(t-1)} + \tau^{-2}}, \frac{1}{n\omega^{(t-1)} + \tau^{-2}} \right).$$

Similarly, $\text{ch}[\omega] = \{ y_i, i = 1, \ldots, n \}$, and so

$$p(\omega | \mu^{(t)}, y) \propto \omega^{\alpha - 1} \exp\left\{ -\beta \omega \right\} \times \omega^{n/2} \exp\left\{ -\frac{\omega}{2} \sum_{i=1}^{n} (y_i - \mu^{(t)})^2 \right\},$$

which is the same as in the known mean, unknown precision case discussed previously.

Hence

$$p(\omega | \mu^{(t)}, y) = \text{Gamma} \left( \alpha + \frac{n}{2}, \beta + \frac{1}{2} \sum_{i=1}^{n} (y_i - \mu^{(t)})^2 \right).$$

Initial values

- The Markov chain needs to start somewhere: initial values
- Should not influence equilibrium distribution of chain
- But may need to be carefully chosen for numerical efficiency reasons
- If the results are influenced by the choice of initial values, then the chain has not converged!

Convergence

- Cannot prove convergence, only detect lack of convergence
- Monitoring a single chain is useful but can be misleading (if ‘stuck’)
- Best to run multiple chains starting from a diverse set of initial values
- Formal diagnostics exist to check if chains end up in essentially same place
- Look for fat hairy caterpillars
- Parameterisation may make a big difference
Formal detection

- CODA and BOA software contain large number of diagnostics
- Brooks-Gelman-Rubin statistic implemented in WinBUGS
- Based on ratio of between to within variances of multiple chains: (ANOVA)

WinBUGS produces plots of

- Average 80% interval within-chains (blue) and pooled 80% interval between-chains (green)
- Ratio green/blue should converge to 1 (red)

Last example of previous plot

BGR diagnostics calculated for ranges 51–100, 101–200, 151–300, 201–400, ......, 501–1000

plotted against starting iteration of each range