SC7/SM6 Bayes Methods HT19

Lecturer: Geoff Nicholls


Notes and Problem sheets are available at

http://www.stats.ox.ac.uk/~nicholls/BayesMethods/

and via the MSc weblearn pages.
Prior Elicitation

Think about the prior! When Bayesian reasoning leads to nonsensical answers, it is almost always the result of careless prior specification. The issue is obviously important when the data are only weakly informative of the parameter. The problem of prior specification becomes acute in high dimensional problems.

*Things to think about when building a prior..*

- Is there a key scientific hypothesis or parameter? If so we may wish to construct a prior which is non-informative with respect to this hypothesis/parameter. [example - the span variable in Radiocarbon dating]
- Is the parameter $\theta$ generated by some process we can model? If so then the distribution over $\theta$ determined by the process is the prior. [example - ages of undated objects in radiocarbon dating]

- Is there some physically interpretable function $f(\theta)$ of the parameter? The distribution of $f(\theta)$ is determined by the prior so the prior is constrained to realise a priori plausible $f$-values. [example - choosing priors for logistic effects]

- How reliable is the information you are using to build a prior? If it is unreliable, you may wish to downweight it, taking care to ensure that carelessly imposed prior structure doesn't overwhelm data-information for parameters which are poorly informed by the data [this leads to objective Bayes]
is the number of things you don't know one of the things you don't know? In this case you may need to put a prior on the number of unknowns! [example - galaxy data and reversible jump]

And once you have built a prior...

- The prior density you write down is meant to model your prior knowledge. Once you are done, simulate the prior, and check the realised samples and physically meaningful functions of the samples are distributed as intended.

- It isn't necessary (or even sensible) to analyse the data with just one prior. We typically check results are insensitive to a range of priors representing different states of knowledge.
Using Simulation to check a prior

*Example* I used the Challenger data to illustrate bridge estimation and model averaging. The linear predictor is \( \eta_i = \beta_1 + \beta_2 x_i \) with 
\[ x_i = \text{scale}(\text{temp})[i] = (\text{temp}[i] - \text{mean}(\text{temp}))/\text{sd}(\text{temp}) \] 
the scaled temperature. Consider the logistic link setting. The prior I used was \( \beta_1, \beta_2 \sim N(0, 3^2) \). Is this remotely plausible?

One quantity we might have some feeling for is \( \mu(\eta) \), the probability for O-ring failure. We know there was a failure, we know they don't always fail. The probability for failure shouldn't vary too sharply with temperature but must depart substantially from zero and one.

We simulate \( \beta_1, \beta_2 \sim N(0, v) \) for a few values of \( v \) and look at \( \mu(\beta_1 + \beta_2 x) \) as a function of \( x \).
Example: Radiocarbon dating

We dig a hole in the ground. We dig through a habitation layer. Above and below this layer there is no evidence for dwellings on the site. We take \( n \) charcoal samples from the habitation layer and get them radiocarbon dated. The key questions of interest are, when was the site settled, for how long was it settled and when did the settlement cease? There is a suggestion that this camp was settled for just weeks or months rather than years.

The data we have are 7 dates from a ancient settlement in NZ. The strata are very clear. As we dig down we encounter artifact-free sand, then obvious settlement remains, and then below that artifact-free earth. The researchers are very confident (before see the RCD's) there was no settlement prior to 1000 years BP (in fact 700 BP would still be reasonable), and that the settlement had been abandoned prior to 500BP.
An uncalibrated radiocarbon age $y_i$ is, for $i = 1, 2, ..., n$, a noisy biased measurement of the unknown true age $\theta_i$ of the $i$'th dated specimen. The observation model for the data is

$$y_i = [\mu(\theta_i) + \zeta_i] + \epsilon_i$$

with

$$\zeta_i \sim N(0, \sigma_c(\theta_i)^2)$$

modeling uncertainty in $\mu$ and

$$\epsilon_i \sim N(0, \sigma_i^2)$$

the error associated with the measurement of specimen $i$ itself. The likelihood is

$$L(\theta; y) = \prod_i \frac{1}{\sqrt{2\pi \sigma^2}} \exp\left(-\frac{(y_i - \mu(\theta_i))^2}{2(\sigma_c(\theta_i)^2 + \sigma_i^2)}\right)$$
Calibration curve $\mu$ and Likelihood

- Uncalibrated Radiocarbon Age ($Y$)
- Calendar Years Before the Present (Theta)
- Calibration curve $\mu$
- $\mu \pm \text{err}$
- Uncalibrated carbon date $y_6$
- $y_6 \pm \sigma_6$
- (scaled) likelihood for $\theta_6$

Legend:
- calibration curve $\mu$
- $\mu \pm \text{err}$
- uncalibrated carbon date $y_6$
- $y_6 \pm \sigma_6$
- (scaled) likelihood for $\theta_6$
The data

\( n = 7 \) radiocarbon dates from a river-mouth settlement in NZ.

<table>
<thead>
<tr>
<th>Date id</th>
<th>i</th>
<th>( y_i )</th>
<th>( \sigma_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>NZ 7758</td>
<td>1</td>
<td>580</td>
<td>47</td>
</tr>
<tr>
<td>NZ 7761</td>
<td>2</td>
<td>600</td>
<td>50</td>
</tr>
<tr>
<td>NZ 7757</td>
<td>3</td>
<td>537</td>
<td>44</td>
</tr>
<tr>
<td>NZ 7756</td>
<td>4</td>
<td>670</td>
<td>47</td>
</tr>
<tr>
<td>NZ 7755</td>
<td>5</td>
<td>646</td>
<td>47</td>
</tr>
<tr>
<td>WK 2589</td>
<td>6</td>
<td>630</td>
<td>35</td>
</tr>
<tr>
<td>NZ 7771</td>
<td>7</td>
<td>660</td>
<td>46</td>
</tr>
</tbody>
</table>

The likelihood plot in the previous graph showed \( L(\theta_6; y_6) \), where the overall likelihood is

\[
L(\theta; y) = \prod_{i=1}^{n} L(\theta_i; y_i).
\]
Priors

A simple uniform prior $\pi_u(\theta)$ for $\theta = (\theta_1, \ldots, \theta_n)$ might set lower and upper bounds $L = 500$ and $U = 1000$ on the true ages of the specimen, and assert that all other values are equally probable:

$$\pi_u(\theta) \propto \prod_{i} I(L \leq \theta_i \leq U).$$

We could estimate the start and end of settlement using $\theta_{(1)} = \min(\theta)$ and $\theta_{(n)} = \max(\theta)$ the oldest and most recent dates, and estimate the span using

$$S_u = \max(\theta) - \min(\theta),$$
Does this prior meet our elicitation criteria?

(A) can we think of the dates $\theta_i, i = 1, 2, \ldots, n$ as generated by some “physical” process?

(B) arrival date, departure date, length of stay are linked to scientific questions - make sure prior is non-informative on these variables.

Does $\pi_u$ meet these criteria? Certainly not (A). For (B) the marginal prior density of $\theta_{(n)}$ (order statistics of a uniform) is $f(\theta_{(n)}) \propto \left[(\theta_{(n)} - L)/(U - L)\right]^{n-1}$ which is not uniform. You may prefer to check using simulation.
These marginal priors $\pi_u(\theta_{(1)})$, $\pi_u(\theta_{(n)})$ and $\pi_u(S_u)$ are strongly informative and we don't want this.
A prior from process generating $\theta$

The settlement starts at an unknown age $\psi_2$ and ends at a time $\psi_1$. These are some of the things our prior knowledge directly informs so we should control the information we provide by making them an explicit part of our prior. The span is

$$S_s = \psi_2 - \psi_1.$$

Between $\psi_2$ and $\psi_1$ the inhabitants generated datable material in the interval $dt$ with probability $\lambda dt$. This gives a Poisson process generating datable material. The material was thinned (by erosion etc) so it is datable with probability $p$. Assume $p$ is a constant not depending on time (so older material is equally likely to survive). The dates $\theta$ are a realisation of a Poisson process with rate $p\lambda$. We choose the number of dates so we condition
on \( n \). It follows that \( \theta_i \sim U(\psi_1, \psi_2) \) (conditioning a Poisson process on the number of events gives a uniform distribution for the event times). Our prior for \( \theta|\psi \) is therefore

\[
\pi_s(\theta|\psi) = (\psi_2 - \psi_1)^n.
\]

We have to specify the prior for \( \psi \). We would like the span to be uniform. The prior

\[
\pi_s(\psi) \propto \frac{1}{(U - L - (\psi_2 - \psi_1))}
\]

has a uniform distribution on values of \( \psi_2 - \psi_1 \), so this is non-informative with respect to the span. Since \( \pi_s(\theta, \psi) = \pi_s(\theta|\psi)\pi_s(\psi) \) we get

\[
\pi_s(\theta, \psi) \propto \frac{1}{(\psi_2 - \psi_1)^n} \frac{1}{(U - L - (\psi_2 - \psi_1))}
\]

models the prior information we actually have.
These marginal priors $\pi_S(\theta_1)$, $\pi_S(\theta_n)$ and $\pi_S(S_u)$ better represent the prior information we have. The prior on span is uniform, and the priors $\pi_S(\theta_1)$, $\pi_S(\theta_n)$ distribute probability mass more evenly over the parameter domain.
Radiocarbon example continued ... Sampling the posterior

The two posterior distributions have the same likelihood, and prior models $m = 1$ ($\pi_u(\theta)$) and $m = 2$ ($\pi_u(\theta, \psi)$).

$$\pi_u(\theta|y) \propto L(\theta; y)\pi_u(\theta)$$

and

$$\pi_s(\theta, \psi|y) \propto L(\theta; y)\pi_s(\theta, \psi).$$

I used simple random walk MH taking each variable in turn. For example, targeting $\pi_s(\theta, \psi|y)$, propose $\theta^I_i \sim U(\psi_1, \psi_2)$ and accept wp

$$\alpha(\theta', \psi|\theta, \psi) = \min \left\{ 1, \frac{L(\theta'; y)}{L(\theta; y)} \right\}$$

since $\pi_s(\theta, \psi)$ depends only on $\psi$ which hasnt changed.
For $\psi_1$, $\psi'_1 \sim U(L, \text{min}(\theta))$, set $\psi' = (\psi'_1, \psi_2)$ and

$$\alpha(\theta, \psi'|\theta, \psi) = \min \left\{ 1, \frac{\pi_s(\theta, \psi')}{\pi_s(\theta, \psi)} \right\}$$

since the likelihood doesn't depend on $\psi$.

The step for $\psi_2$ proposes $\psi'_2 \sim U(\text{max}(\theta), U)$ and accepts with the same formulae as $\psi_1$ (but now $\psi' = (\psi_1, \psi'_2)$).

See attached R-code for details.
Summarising the results

There are many things we test for. Here I focus on the span, and the two models \( m = 1, 2 \). I ran the MCMC, and checked convergence. I obtained HPD estimates for the span.

Model 2 (shrink/\( \pi_S \)): \([0,160]\)  
Model 1 (unif/\( \pi_U \)): \([70,160]\)

The marginal posterior histograms differ in shape and support.
We would like to compare the two models using a Bayes Factor. This is not testing a pre-defined scientific hypothesis. We are using the test to establish our prior modelling is sound (perhaps for use in other contexts).

We cannot use our preferred Bridge estimator for the Bayes factor itself, because the two models have different parameter spaces $\theta$ and $\theta, \psi$ but we could use it for the marginal likelihoods, as the prior and posterior at least have the same parameter space.

We check the result is stable across multiple runs, with very large sample sizes ($\text{ESS} \gg 1000$).
If in model $m = "u", "s"$ the posterior samples for $\theta$ are $\theta^{(m,t)}$, $t = 1, 2, ..., T$ and the prior samples are $\tilde{\theta}^{(m,t)}$ then the bridge estimator identity (using $h = 1/\sqrt{\pi \pi L}$) for $p(y|m)$ is

$$p(y|m) = \frac{E_{\theta|m}(\pi_m(\theta) L(\theta; y) h(\theta))}{E_{\theta|y,m}(\pi_m(\theta) h(\theta))}$$

$$= \frac{E_{\theta|m}(L(\theta; y)^{1/2})}{E_{\theta|y,m}(L(\theta; y)^{-1/2})}$$

and the estimator itself is

$$\hat{p}_m = \frac{\sum_{t=1}^{T} L(\tilde{\theta}^{(m,t)}; y)^{1/2}}{\sum_{t=1}^{T} L(\theta^{(m,t)}; y)^{-1/2}}$$

We find $\hat{p}_u \simeq 4 \times 10^{-21}$ and $\hat{p}_s \simeq 8 \times 10^{-19}$ so the Bayes factor for shrinking over uniform priors is about $B_{2,1} \simeq 200$, so the shrinkage prior is clearly favoured.
Conclusions

We don’t have much data (7 noisy numbers!) so the conclusions show some sensitivity to the choice of prior. The process-model based prior allows very small spans close to zero, while the uniform prior rules them out. This is clearly a case where we don’t want the prior to impose structure we can’t support on prior grounds.

We might go on to test for $\psi_1 = \theta_1 = \ldots \theta_n = \psi_2$ and consider goodness of fit.