SC7 Bayes Methods

Problem sheet 0 (background).

I expect you can solve these problems using the knowledge you bring into the course, except perhaps question 1(d), question 2 and question 4. We will review all this in the course, but fairly briefly, so if this is new I recommend some reading. You will find the Decision Theory material in Q2 is covered in SB2.1 Foundations of Statistical Inference (and see Sec 1.3.3 of my lecture notes) and the MCMC material in Q4 is covered in A12 Simulation and Statistical Programming (and in Sec 5.1.1-5.1.6 of my notes).

I include some R in my answers - this is just for illustration. Programming is not assessed in any way in this course.

1. (a) Consider tossing a drawing pin [see figure at end]. Define the result of a toss to be "heads" if the point lands downwards, and "tails" otherwise. Write p for the probability that a toss will land point downwards. Think about p, and choose a, b, so that a Beta(a, b) prior distribution approximates your subjective prior distribution for p. [I used $a = 2$ and $b = 3$ but you may differ.]

> **Solution:** Subjective, but I think it will mainly land point up - on a hard table it bounces and I think typically finds the minimum energy configuration. So I will take a prior with a mean around $E(p) = 0.4 = a/(a+b)$. I think the prior should go to zero as p goes to 0 or 1 so I want $a, b > 1$. Finally then $p \sim Beta(2, 3)$ seems about right. Checking, the prior 95% equal tail credible interval with $a = 2$ and $b = 3$ is $[0.07, 0.8]$ which is perhaps a bit tight compared to my prior beliefs but more or less OK.

   (b) Now collect data. Toss a drawing pin 100 times and keep track of the number of heads after 10, 50, and 100 tosses. You may find the result depends on the surface you use. [I got 4, 16 and 26 heads after 10, 50 and 100 tosses.]

   (c) Ask someone else what prior they chose. Think of your respective priors as a hypotheses about p. Who's beliefs were better supported by the data? Compute a Bayes factor comparing your priors. [for me the other person used $a = 3$ and $b = 2$.]

> **Solution:** My friend had the "opposite" prior with $a = 3$ and $b = 2$ (so they were expecting pin down). The marginal likelihood $p(y|a, b)$ (ie the prior predictive

distribution) for a model with parameters a, b is

$$\begin{aligned} p(y|a,b) &= \int_0^1 p(y|p)\pi(p)dp \\ &= \int_0^1 \binom{n}{y} p^y (1-p)^{n-y} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1} dp \\ &= \binom{n}{y} \frac{B(a+n, b+n-y)}{B(a,b)}. \end{aligned}$$

where $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ is the Beta-function. The Bayes factor comparing the models with a, b against a', b' is the ratio of the marginal likelihoods,

$$\begin{aligned} BF_{\text{me v. you}} &= p(y|a,b)/p(y|a',b') \\ &= \frac{B(a+n, b+n-y)B(a',b')}{B(a'+n, b'+n-y)B(a,b)}, \end{aligned}$$

which is bigger than one if model a, b beats model a', b'. If I wanted to check my calculation I could simply estimate the marginal likelihoods (the 3-component `ml.me` and `ml.you`vectors below) using the naive estimator (averaging the likelihood $L(p; y)$ in samples from the prior, $p \sim \pi(p)$ (the `rbeta(N,a,b)` and `rbeta(N,ap,bp)` vectors below)).

```
> a=2; b=3; ap=3; bp=2;
> BF.me.v.you=(beta(a+y,b+n-y)*beta(ap,bp))/(beta(ap+y,bp+n-y)*beta(a,b))
> BF.me.v.you
[1] 1.333333 2.000000 2.714286
```

The Bayes factor for my model against theirs is only 2.7 even at $n = 100$, so there is no strong evidence in favor of my model over theirs.

(d) Estimate a 95% HPD credible interval for p for each of the two priors you are considering, for the case when $n = 10$ trials. Write down the posterior averaged over models, stating any assumptions you make, and estimate a 95% HPD credible interval for p from the model averaged posterior.

Solution: For the single model cases we could get the HPD from the quantiles. A simple approach that will work for the model averaged posterior is to sample the distribution and estimate the HPD numerically.

We now write down the model average posterior. Let $m = 1, 2$ be the model index

(so the model is $a = 2, b = 3$ when $m = 1$ and $a' = 3, b' = 2$ when $m = 2$) and $N_m = 2$ be number of models under consideration. The model average posterior is

$$\begin{aligned}\pi(p|y) &= \sum_{m=1}^{N_m} \pi(p, m|y) \\ &= \sum_{m=1}^{N_m} \pi(p|y, m)\pi_M(m|y) \\ &= \pi(p|y, a, b)\pi_M(m = 1|y) + \pi(p|y, a', b')\pi_M(m = 2|y)\end{aligned}$$

where

$$\pi_M(m|y) \propto p(y|m)\pi_{M,m}$$

is the posterior probability for model m, $p(y|m)$ is the marginal likelihood for model m and $\pi_{M,m}$ is the prior probability for model m. Notice that $\pi(p|y, m)$ is just the posterior we had before, $\pi(p|y, m = 1) = \pi(p|y, a, b)$ *etc* and that the model average distribution $\pi(p|y)$ is a mixture distribution which puts more weight on the component with a higher marginal likelihood.

Our strategy for estimating the HPD for the model average will be to sample $p \sim \pi(p|y)$ and estimate the HPD numerically as before. Now $\pi(p|y)$ is a mixture so to sample $p \sim \pi(p|y)$ we sample $m \sim \pi_M(m|y)$ and then sample $p \sim \pi(p|y, m)$. The second bit is easy as $\pi(p|y, m)$ is just the posterior we started with. We need to define and compute $\pi_M(m|y)$.

In our case it is natural to say that the two models are equally likely a priori, so $\pi_{M,1} = \pi_{M,2} = 0.5$. We computed the marginal likelihoods earlier, $p(y|m = 1) = p(y|a, b)$ and $p(y|m = 2) = p(y|a', b')$, so for example

$$\pi_M(m = 1|y) = \frac{p(y|a, b)\pi_{M,1}}{p(y|a, b)\pi_{M,1} + p(y|a', b')\pi_{M,2}}$$

Putting in the expressions above (cancelling the equal $\pi_{M,m}$ and $\binom{n}{y}$ values and simplifying)

$$\pi_M(m = 1|y) = \frac{B(a + n, b + n - y)}{B(a + n, b + n - y) + B(a' + n, b' + n - y)\frac{B(a,b)}{B(a',b')}}$$

and similarly for $\pi_M(m = 2|y)$ (replace a, b with a', b' in the numerator).

```
> N=1000000;
> ps.me=rbeta(N,a+y[1],b+n[1]-y[1])  #first entry as that's n=10
```

```
> ps.you=rbeta(N,ap+y[1],bp+n[1]-y[1])
>
> #sample the mixture (for the model average
> #first compute pi_M(m|y) using the formula in the text
> py=(beta(a+y,b+n-y)+beta(ap+y,bp+n-y)*(beta(a,b)/beta(ap,bp)))
> pm1=beta(a+y,b+n-y)/py; pm1
[1] 0.5714286 0.6666667 0.7307692
>
> pm2=beta(ap+y,bp+n-y)/py; pm2
[1] 0.4285714 0.3333333 0.2692308
>
> #OK lets do the HPD for the case where we have n=10 samples
> #That means we use the first entry pm1[1]=pi_M(1|y[1]) and
> #pm2[1]=pi_M(2|y[1]), and select samples from ps.me vector with
> #probability pm1[1] and samples from ps.you vector with probability pm2[1]
> ps.mix=ps.me;                          #initialise sample vector all ps.me
> choose.you=which(runif(N)<pm2[1]);    #sample from ps.you wp pm2[1]
> ps.mix[choose.you]=ps.you[choose.you] #write m=2-samples into ps.mix vector
>
> #estimate the HPD credible interval
> library(coda)
> round(HPDinterval(as.mcmc(cbind(ps.me,ps.you,ps.mix))),2)
       lower upper
ps.me   0.17  0.64
ps.you  0.23  0.71
ps.mix  0.19  0.68
attr(,"Probability")
[1] 0.95
```

After only 10 samples the HPD interval is still pretty broad. The model average sits between the HPD intervals for the components.

2. Prof Wynn has a hole in his pocket. He walks home from work and finds his keys are not in his pocket. He is $p \times 100\%$ certain he left his keys on his desk (at $\theta = 0$). If they are not there then they could have fallen out anywhere between home and work (uniformly distributed between $\theta = 0$ and $\theta = 1$).

   (a) Let $\theta \in [0, 1]$ be the unknown key location and let $U, V \sim U(0, 1)$ be independent

uniform random variables. Verify that Prof Wynn's prior for the key location has the same distribution as $\theta = V\mathbb{I}_{U>p}$.

> **Solution:** With probability p we have $U < p$ and $\theta = 0$. With probability $(1-p)$ we have $U > p$ and $\theta = V$ which is uniform, matching Prof Wynn's prior. This is like taking $\pi(d\theta) = p\delta_0(d\theta) + (1-p)d\theta$ for $\theta \in [0,1]$ (with $\delta_0(\theta)$ a delta-function).

(b) Show that the prior variance for θ is $\text{var}(\theta) = (1-p)(1/12 + p/4)$ and hence show that as p approaches one (desk-certainty) the prior variance $\text{var}(\theta)$ approaches zero.

> **Solution:** $\text{var}(\theta) = E(\theta^2) - E(\theta)^2$. Substitute $\theta = V\mathbb{I}_{U>p}$ and use the independence of U and V to verify $E(\theta^2) = (1-p)/3$ and $E(\theta) = (1-p)/2$. This gives the stated result and behaviour at $p \to 1$.

(c) Prof Wynn returns to work and finds the keys are not on the desk. Explain why the posterior for θ given this new data has variance $1/12$.

> **Solution:** If the keys are not on the desk then they are uniformly distributed over $\theta \in [0,1]$ so the variance of θ is $1/12$.

(d) Can the posterior have greater variance than the prior? Can conditioning on data *increase* uncertainty?

> **Solution:** Yes. If the data conflicts the prior then the parameter is forced into the tails of the prior. In the example, if p is close to one then the prior variance can be arbitrarily small, smaller than the posterior variance of $1/12$.

3. Let $\theta \sim \pi(\cdot)$ and $y \sim p(\cdot|\theta)$ be a prior for a scalar parameter and the observation model for data $y \in \mathbb{R}^n$ respectively. Let $\hat{\theta}(y)$ be an estimator for θ.

(a) Suppose our loss function for estimating $\hat{\theta}$ when the truth is θ is $l(\theta, \hat{\theta}) = (\hat{\theta} - \theta)^2$. Show that the Bayes estimator is the posterior mean.

> **Solution:** The Bayes estimator minimises the Bayes risk, so it minimises the expected posterior loss (see Section 1.3.2 of notes for a reminder),
>
> $$\rho(\pi, \hat{\theta}|y) = E_{\theta|y}((\hat{\theta}(y) - \theta)^2).$$

Recall that y is fixed. Differentiating WRT to $\hat{\theta}(y)$ and setting $d\rho/d\hat{\theta} = 0$ we get

$$2E_{\theta|y}((\hat{\theta}(y) - \theta)) = 0$$

and solving for $\hat{\theta}$ gives $\hat{\theta}(y) = E_{\theta|y}(\theta)$, the posterior mean. Since this function of y minimises $\rho(\pi, \hat{\theta}|y)$ at every y it minimises the Bayes risk $\rho(\pi, \hat{\theta}) = E_y(\rho(\pi, \hat{\theta}|y))$.

(b) Suppose θ is discrete and we have the zero-one loss $l(\theta, \hat{\theta}) = \mathbb{I}_{\hat{\theta}(y)\neq\theta}$. Find the Bayes estimator for θ.
Hint: If you are not familiar with Decision Theory - just the basics are needed - then please read Section 1.3.3 of the lecture notes first.

Solution:

$$\rho(\pi, \hat{\theta}|y) = E_{\theta|y}(\mathbb{I}_{\hat{\theta}(y)\neq\theta}),$$

so $\rho = 1 - \pi(\hat{\theta}|y)$. Since $\hat{\theta} = \arg\min_\delta \rho(\pi, \delta|y)$ we minimise the EPL by taking $\hat{\theta}$ which maximises $\pi(\hat{\theta}|y)$, in other words, the posterior mode.

4. Let $\theta \sim \pi(\cdot|M = m)$ and $y \sim p(\cdot|\theta, M = m)$ be the prior and observation model when the model is $M = m$ and suppose there are just two models, so $M \in \{1, 2\}$. When the model is m the parameter space is $\theta \in \Omega_m$. Consider model selection.

(a) Write down the Bayes factor $B_{1,2}$ in terms of the model elements.

Solution:

$$B_{1,2} = \frac{p(y|M = 1)}{p(y|M = 2)}$$

where $p(y|M = m) = \int_{\Omega_m} p(y|\theta, M = m)\pi(\theta|M = m)\, d\theta$.

(b) Is it necessary for the models to be nested in order that the Bayes factor (which is after all a likelihood rato) is a model selection criterion?

Solution: No. The Bayes factor is the ratio of the posterior probabilities for the models when the prior probabilities for each model are equal, $\pi_M(1) = \pi_M(2)$. With equal prior model probabilities, model 1 is simply $B_{1,2}$ times more probable than than model 2 a posteriori:

$$\frac{\pi(M = 1|y)}{\pi(M = 2|y)} = B_{1,2}$$

as $\pi(M = m|y) = \pi_M(m)p(y|M = m)/p(y)$ with $p(y) = \pi_M(1)p(y|M = 1) + \pi_M(2)p(y|M = 2)$ and cancelling.

(c) Suppose the models *are* nested with $\Omega_1 \subseteq \Omega_2$, so we get the $M = 1$ prior by taking the $M = 2$ prior and conditioning on $\theta \in \Omega_1$,

$$\pi(\theta|M = 1) = \frac{\pi(\theta|M = 2, \theta \in \Omega_1)}{\pi(\Omega_1|M = 2)}$$

where

$$\pi(\Omega_1|M = 2) = \int_{\Omega_1} \pi(\theta|M = 2)\, d\theta,$$

and $p(y|\theta, M = 1) = p(y|\theta, M = 2)$ for $\theta \in \Omega_1$. Show that

$$B_{1,2} = \frac{\pi(\Omega_1|y, M = 2)}{\pi(\Omega_1|M = 2)}$$

and briefly interpret.

Solution:

$$\begin{aligned}
B_{1,2} &= \frac{p(y|M = 1)}{p(y|M = 2)} \\
&= \frac{\int_{\Omega_1} p(y|\theta, M = 1)\pi(\theta|M = 1)\, d\theta}{\int_{\Omega_2} p(y|\theta, M = 2)\pi(\theta|M = 2)\, d\theta} \\
&= \frac{\int_{\Omega_1} p(y|\theta, M = 2)\pi(\theta|M = 2)\, d\theta}{p(y|M = 2)} \times \frac{1}{\pi(\Omega_1|M = 2)} \\
&= \frac{\int_{\Omega_1} \pi(\theta|y, M = 2)\, d\theta}{\pi(\Omega_1|M = 2)} \\
&= \frac{\pi(\Omega_1|y, M = 2)}{\pi(\Omega_1|M = 2)},
\end{aligned}$$

where we identified

$$\frac{p(y|\theta, M = 2)\pi(\theta|M = 2)}{p(y1|M = 2)} = \pi(\theta|y, M = 2).$$

The Bayes factor tells us how much more probable the event $\theta \in \Omega_1$ becomes after the data arrives than it was a priori.

5. (for those who know some MCMC - we cover this in the course so not strictly background)

(a) Specify a Metropolis-Hastings Markov chain Monte Carlo algorithm targeting $p(x|\theta)$ where $x \in \{0, 1, ..., n\}$ and

$$p(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}.$$

Prove that your chain is irreducible and aperiodic.

Solution: Let $X_0 = 0$. Define $p(x|\theta) = 0$ for $x = -1, n+1$. Let $X_t = x$. X_{t+1} is determined in the following way.

1 Set $y = x + 1$ w.p. $1/2$ and otherwise $y = x - 1$.

2 wp $\alpha(y|x)$ set $X_{t+1} = y$ and otherwise set $X_{t+1} = x$. Here

$$\begin{aligned} \alpha(y|x) &= \min\left(1, \frac{p(y|\theta)q(x|y)}{p(x|\theta)q(y|x)}\right) \\ &= \begin{cases} \min\left(1, \frac{\theta(n-x)}{(x+1)(1-\theta)}\right) & y = x+1 \\ \min\left(1, \frac{(1-\theta)x}{\theta(n-x+1)}\right) & y = x-1 \\ 0 & y \notin \{0, 1, ..., n\}. \end{cases} \end{aligned}$$

and the algorithm rejects proposals which lie outside $\{0, 1, ..., n\}$. Note also that the proposal probability $q(y|x)$ for y given x at step 1. is symmetric, that is $q(x+1|x) = q(x|x+1) = 1/2$ for each value of x which can actually arise.

This is irreducible because the proposal itself is irreducible and the acceptance probability for $y \in x \pm 1$ is never zero for $y = 0, 1, 2, ...$ and it is aperiodic because it is irreducible and can reject (so there is x such that the transition probability $P_{x,x} > 0$).

(b) Suppose now that the unknown true success probability for the Binomial random variable X in part (a) is a random variable Θ which can take values in $\{1/2, 1/4, 1/8, ...\}$ only. The prior is

$$\pi(\theta) = \begin{cases} \theta & \text{for } \theta \in \{1/2, 1/4, 1/8, ...\}, \text{ and} \\ 0 & \text{for } \theta \text{ otherwise.} \end{cases}$$

An observed value $X = x$ of the Binomial variable in part (a) is generated by simulating $\Theta \sim \pi(\cdot)$ to get $\Theta = \theta^*$ say, and then $X \sim p(x|\theta^*)$ as before. Specify a Metropolis-Hastings Markov chain Monte Carlo algorithm simulating a Markov chain targeting the posterior $\pi(\theta|x)$ for $\Theta|X = x$.

Solution: The target distribution is

$$\pi(\theta|x) \propto \theta^{x+1}(1-\theta)^{n-x}$$

for $\theta \in \{1/2, 1/4, 1/8, ...\}$. Extend this by setting $\pi(1|x) = 0$.
Let $\theta_0 = 1/2$. Let $\theta_t = \theta$. θ_{t+1} is determined in the following way.

1 Set $\xi = 2$ w.p. $1/2$ and otherwise $\xi = 1/2$. Set $\theta' = \xi\theta$.

2 wp $\alpha(\theta'|\theta)$ set $\theta_{t+1} = \theta'$ and otherwise set $\theta_{t+1} = \theta$. Here

$$\begin{aligned} \alpha(\theta'|\theta) &= \min\left(1, \frac{\pi(\theta'|x)q(\theta|\theta')}{\pi(\theta|x)q(\theta'|\theta)}\right) \\ &= \min\left(1, \frac{(\xi\theta)^{x+1}(1-\xi\theta)^{n-x}}{\theta^{x+1}(1-\theta)^{n-x}}\right), \end{aligned}$$

and the algorithm rejects any proposal $\theta' = 1$ since $\alpha(1|1/2) = 0$. The proposal probability $q(\theta'|\theta)$ at step 1. is again symmetric.

Statistics Department, University of Oxford
Geoff Nicholls: nicholls@stats.ox.ac.uk