SC7 Bayes Methods

First problem sheet (Sections 1-2 of lecture notes).

Section A questions

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.]

   (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 $b = 2$.]

   (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.

2. (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.

   (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 $\Theta|X = x$.

Section B questions

3. In the radiocarbon dating example, suppose the dated materials are found in layers (strata) piled up on one another, with $y_{i,j}$ the radiocarbon date for $\theta_{i,j}$, the j'th date in the i'th layer. Let $L < \psi_1 < \psi_2 < ... < \psi_M < U$ be the age parameters for the layer boundaries. If we have n_i dates from the ith layer we know that for $i = 1, 2, ..., M-1$, and $j = 1, 2, ..., n_i$, $\psi_i < \theta_{i,j} < \psi_{i+1}$ (so specimen dates in higher layers are not as old as dates in lower layers). Let $\psi = (\psi_1, ..., \psi_M)$ and $\theta = (\theta_1, ...\theta_{M-1})$ with $\theta_i = (\theta_{i,1}, ..., \theta_{i,n_i})$.

   Derive a prior density $\pi(\theta, \psi)$ for the parameters θ, ψ with reference to the prior elicitation checklist given in lectures. Hint: how are the layer boundary dates $\psi_2, ..., \psi_{M-1}$ generated?

4. Let $\Gamma(x; \alpha, \beta)$ be the Gamma density. Consider Poisson observations $Y = (Y_1, Y_2, ..., Y_n)$ with means $\lambda = (\lambda_1, \lambda_2, ..., \lambda_n)$ given by a mixture of Gamma densities: for shape parameters α_1, α_2 and rate parameters β_1, β_2, a known mixture proportion $0 < p < 1$ and $i = 1, 2, ..., n$, we observe

$$Y_i|\lambda_i \sim \text{Poisson}(\lambda_i)$$

   (all iid) with

$$\lambda_i \sim p\Gamma(\lambda_i; \alpha_1, \beta_1) + (1-p)\Gamma(\lambda_i; \alpha_2, \beta_2).$$

   (a) Denote by $\pi(\alpha_1, \beta_1, \alpha_2, \beta_2)$ a prior for the unknown shape and rate parameters. Write down the joint posterior for $\alpha_1, \beta_1, \alpha_2, \beta_2$ and λ given $Y_1, Y_2, ..., Y_n$. Give an MCMC algorithm sampling $\alpha_1, \beta_1, \alpha_2, \beta_2, \lambda|Y_1, ..., Y_n$.

   (b) Integrate λ out of the joint posterior to obtain a marginal posterior denisty for $\alpha_1, \beta_1, \alpha_2, \beta_2|Y_1, ..., Y_n$. Comment briefly on how you would alter your MCMC algorithm for the new target. What considerations would guide your choice of simulation method (ie, whether to simulate the joint or the marginal posterior density)?

5. Let X be an $n \times p$ design matrix with rows $x_i, i = 1, 2, ..., n$ and $\theta = (\theta_1, \theta_2, ..., \theta_p)^T$ a p-component vector of parameters. Let $z = (z_1, ..., z_n)$ be jointly independent normal random variables, $z \sim N(X\theta, I_n)$ with I_n the $n \times n$ identity. In the probit observation model for $y = (y_1, ..., y_n)$, we observe $y_i = 1$ if $z_i > 0$ and $y_i = 0$ if $z_i \leq 0$.

   Denote by $\pi(\theta, z) = \pi(\theta)\pi(z|\theta)$ the joint density of θ and z with $\pi(\theta) = N(\theta; 0, \Sigma)$ a normal prior for θ and Σ a $p \times p$ covariance matrix.

   (a) Show that $y_i \sim \text{Bernoulli}(\Phi(x_i\theta))$.

   (b) Write the posterior $\pi(\theta, z|y)$ in terms of the model elements.

   (c) Show that

$$p(\theta|z) = N(\theta; \mu, V)$$

   with $\mu = VX^Tz$ and $V = (\Sigma^{-1} + X^TX)^{-1}$.

(d) Show that

$$\pi(z_i|y_i,\theta) \propto \begin{cases} N(z_i; x_i\theta, 1)\mathbb{I}_{z_i \leq 0} & \text{if } y_i = 0 \\ N(z_i; x_i\theta, 1)\mathbb{I}_{z_i > 0} & \text{if } y_i = 1 \end{cases}$$

(e) Give a Gibbs sampler sampling $\pi(\theta|y)$ (Hint: $\pi(\theta, z|y)$ would be easier).

6. Let $\pi(\theta), \theta \in R$ be a prior density for a scalar parameter, let $p(y|\theta), y \in R^n$ be the observation model density and let $\pi(\theta|y) \propto \pi(\theta)p(y|\theta)$ be the posterior density. Consider a Markov chain simulated in the following way. Suppose $\theta^{(0)} \sim \pi(\cdot)$ is a draw from the prior and for $t = 0, 1, 2, ...$ we generate a Markov chain by simulating data $y^{(t)} \sim p(\cdot|\theta^{(t)})$ and then $\theta^{(t+1)} \sim \pi(\cdot|y^{(t)})$.

(a) i. Calculate the joint density, $p(\theta^{(0)}, \theta^{(1)})$ say, for $\theta^{(0)}, \theta^{(1)}$ and show that $p(\theta^{(0)}, \theta^{(1)}) = p(\theta^{(1)}, \theta^{(0)})$ (ie they are exchangeable).

ii. Show that marginally, $\theta^{(t)} \sim \pi(\cdot)$ for all $t = 0, 1, 2, ...$

iii. Give the transition probability density $K(\theta, \theta')$ for the chain and show the chain is reversible with respect to the prior $\pi(\theta)$.

(b) Suppose we are given an MCMC algorithm $\theta^{(T)} = \mathcal{M}(\theta^{(0)}, T, y)$, initialised at $\theta^{(0)}$, and targeting the posterior $\pi(\theta|y) \propto \pi(\theta)p(y|\theta)$, so $\theta^{(T)} \xrightarrow{D} \pi(\cdot|y)$ as $T \to \infty$. Here $\mathcal{M}$ is a function that moves us T steps forward in the MCMC run and this Markov chain is just some MCMC algorithm for simulating $\pi(\theta|y)$ and so not related to the Markov chain in the previous part.

Suppose we think we have chosen T sufficiently large that the chain has converged, and so we believe $\theta^{(T)} \sim \pi(\cdot|y)$ is a good approximation.

Consider the following procedure simulating pairs (ϕ_i, θ_i), $i = 1, 2, ..., K$: (Step 1) parameter $\phi_i \sim \pi(\cdot)$ is an independent draw from the prior; (Step 2) synthetic data $y_i' \sim p(\cdot|\phi_i)$ is an independent draw from the observation model; (Step 3) the MCMC algorithm $\mathcal{M}$ is initialised with a draw $\theta_i^{(0)} \sim \pi^{(0)}$ from an arbitrary fixed initial distribution $\pi^{(0)}$ and (Step 4) we set $\theta_i = \mathcal{M}(\theta_i^{(0)}, T, y_i')$.

Let $\phi = (\phi_1, ..., \phi_K)$ and $\theta = (\theta_1, ..., \theta_K)$ be samples generated in this way.

i. Suppose the chain has indeed converged by T steps for all starting states $\theta^{(0)}$. Let $p(\phi, \theta)$ be the joint distribution of the random vectors ϕ and θ. Show that $p(\phi, \theta) = p(\theta, \phi)$.

ii. Give a non-parametric test for MCMC convergence which makes use of the result in Question 6(b)i. Hint: the null is $\theta^{(T)} \sim \pi(\cdot|y)$.

7. (From Cox and Hinkley *Theoretical Statistics*) For $i = 1, ..., n$, let $\theta_i \in \{0, 1\}$ be the indicator for the event that student i enjoys the course in 2023 and let $\theta = (\theta_1, ..., \theta_n)$. Suppose our prior probability for $\theta_i = 1,\ i = 1, ..., n$ is that they are iid with $P(\theta_i = 1) = p$ with p our prior probability that an individual student enjoys the course and we take a fixed value of p expressing our prior expectation for the proportion enjoying the course (based perhaps on past years).

   Our prior on the function $q(\theta) = n^{-1} \sum_i \theta_i$ has mean p (that's good) and variance $p(1 - p)/n$. If n is large this prior expresses near certainty in the proportion of students enjoying the course.

   Criticise this prior elicitation and suggest an improvement. As a hint, something is wrong with the prior variance of the random variable $Q = q(\theta)$ and we should change the prior to fix this.

Section C questions

8. (MSc 2020 exam - students had a related practical in 2020) A book club with n members wants to decide what book to read next. They have a shortlist of B books with labels $\mathcal{B} = \{1, ..., B\}$. Let $\mathcal{P}_\mathcal{B}$ be the set of all permutations of the labels in $\mathcal{B}$. For $i = 1, ..., n$ the i'th reader gives a ranked list of the books $y_i = (y_{i,1}, ..., y_{i,B}),\ y_i \in \mathcal{P}_\mathcal{B}$, ranking them from most to least interesting. The data are $y = (y_1, ..., y_n)$.

   In a Plackett-Luce model each book $b = 1, ..., B$ has interest measure $\theta_b > 0$. Let $\theta = (\theta_1, ..., \theta_B),\ \theta \in R^B$. Let $Y_i \in \mathcal{P}_\mathcal{B}$ denote the random ranking from the i'th reader. In the Plackett-Luce model, given $Y_{i,1} = y_{i,1}, ..., Y_{i,a-1} = y_{i,a-1}$, the a'th entry (ie, the next entry) is decided by choosing book b with probability proportional to θ_b from the books $\mathcal{B} \setminus \{y_{i,1}, ..., y_{i,a-1}\}$ remaining. The $Y_1, ..., Y_n$ are jointly independent given θ.

   (a) i. Show that the likelihood $L(\theta; y)$ is

$$L(\theta; y) = \prod_{i=1}^{n} \prod_{a=1}^{B} \frac{\theta_{y_{i,a}}}{\sum_{b=a}^{B} \theta_{y_{i,b}}}.$$

   ii. The prior is $\pi_\mathcal{B}(\theta) = \prod_{b=1}^{B} \pi(\theta_i)$ with $\pi(\theta_i) = \Gamma(\theta_i; \alpha', 1)$ with $\alpha' > 0$ given. Write down the posterior density $\pi(\theta|y)$ and give an MCMC algorithm targeting $\pi(\theta|y)$.

   iii. Explain why the scale β' in the prior $\Gamma(\alpha', \beta')$ for $\theta_i,\ i \in \mathcal{B}$ may be set equal one. Suppose odds of $1000 : 1$ for ranking one book above another represent extreme preference and are a priori unlikely for books on the shortlist. Explain how a fixed numerical value of α' might be chosen, noting any assumptions.

(b) Suppose B is large so each reader $i = 1, ..., n$ only reports the first N entries $x_i = (x_{i,1}, ..., x_{i,N})$ in their ranking, with $N \ll B$. Here $x_{i,j} = y_{i,j}$ for $i = 1, ..., n$ and $j = 1, ..., N$. The data are $x = (x_1, ..., x_n)$.

i. Show that the likelihood $L(\theta; x)$ for the new data is

$$L(\theta; x) = \prod_{i=1}^{n} \prod_{a=1}^{N} \frac{\theta_{x_{i,a}}}{\sum_{b=a}^{N} \theta_{x_{i,b}} + \sum_{d \in \mathcal{B} \setminus x_i} \theta_d}.$$

ii. Let $\mathcal{C} = \bigcup_{i=1}^{n} x_i$ give the books appearing in at least one ranking and $\mathcal{D} = \mathcal{B} \setminus \mathcal{C}$ be the books appearing in none. Let $\theta_{\mathcal{C}} = (\theta_b)_{b \in \mathcal{C}}$ and $V = \sum_{d \in \mathcal{D}} \theta_d$.
Write down the prior distribution of V and the likelihood $L(\theta_{\mathcal{C}}, V; x)$, and give the posterior $\pi(\theta_{\mathcal{C}}, V | x)$ as a function of $\theta_{\mathcal{C}}$ and V.

iii. Give an MCMC algorithm targeting $\pi(\theta_{\mathcal{C}}, V | x)$. State briefly why it may be more efficient, for estimation of $\theta_{\mathcal{C}}$ in the case $|\mathcal{C}| \ll B$, than MCMC targeting $\pi(\theta | x)$.

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