Bayesian Inference

Tutorial 6

1. Radio-tagging data involves placing a radio-tag on a number of individuals and (assuming no radio failures) recording the number of deaths that occur at a series of successive "capture" times. We assume that only a single radio-tagging event occurs where a total of n lambs are "tagged". We let x_t denote the number of sheep that are subsequently recorded as having died within the interval $(t-1, t]$ (assuming tagging occurs at time 0), for $t = 1, \ldots, T$. We let x_{T+1} denote the number of individuals that survive until time T (i.e. the end of the study). The corresponding likelihood function is a function of the survival probabilities of the sheep. We assume two distinct survival probabilities: ϕ_1 corresponding to first-year survival probability and ϕ_a the "adult" survival probability (i.e. older than first-years). The likelihood is given by,

   $$f(\boldsymbol{x}|\phi_1, \phi_a) \propto \prod_{i=1}^{T} p_i^{x_i}$$

   where,

   $$p_i = \begin{cases} 1-\phi_1 & i = 1 \\ \phi_1(1-\phi_a) & i = 2 \\ \phi_1\phi_a^{i-2}(1-\phi_a) & i = 3, \ldots, T \\ \phi_1\phi_a^{T-1} & i = T+1. \end{cases}$$

   Without any prior information on ϕ_1 and ϕ_a we specify $\phi_1 \sim U[0,1]$ and $\phi_a \sim U[0,1]$, independently. Describe how the Gibbs sampler can be used to obtain a sample from the posterior distribution, $\pi(\phi_1, \phi_a|\boldsymbol{x})$. Do you notice anything of interest here? (The answer is obviously yes here - so what is it that is interesting?!).

2. (From May 2007 exam) Let $X_1, \ldots, X_n$ be i.i.d. random variables such that $X_i \sim N(\mu, \omega^{-1})$, and let both μ and ω be unknown. μ is given Jeffreys' prior, and ω is given the prior $\Gamma(\alpha, \beta)$, where α and β are known values (*Note that the prior for μ is of the form $p(\mu) \propto 1$*). We observe data $x_1, \ldots, x_n$, which we denote $\boldsymbol{x}$.

   (a) Write down the joint posterior distribution $\pi(\mu, \omega|\boldsymbol{x})$ up to a proportionality constant. [1]

   (b) Now write down the conditional posterior distributions $\pi(\mu|\omega, \boldsymbol{x})$ and $\pi(\omega|\mu, \boldsymbol{x})$, up to proportionality constants. Both are of standard forms, enabling you to obtain the full conditional posterior distributions. Give the names of these distributions and their parameter values. [4]

   (c) There is an algorithm that can be used to simulate dependent draws from the joint posterior distribution in cases like this, where the conditional posterior distributions are of standard form. What is it called? [1]

   (d) Outline the steps required to obtain draws from $\pi(\mu, \omega|\boldsymbol{x})$ using this algorithm. Note that initial draws may not be from the target distribution, so your outline should contain guidance for determining which draws should be used in making inferences about $\pi(\mu, \omega|\boldsymbol{x})$ and which (if any) should be discarded. [5]

   (e) How should the algorithm be used to obtain estimates of the posterior expectations $\mathbb{E}_\pi(\mu|\boldsymbol{x})$ and $\mathbb{E}_\pi(\omega|\boldsymbol{x})$? [1]

3. Suppose that we wish to use the Metropolis-Hastings algorithm to generate a sample from $f(x) = N(0, \sigma^2)$, and that we use the proposal $q(y|x) = N(ax, \tau^2)$ for $-1 < a < 1$.

   (a) What is the corresponding acceptance probability $\alpha(x, y)$?

   (b) For what value of τ^2 would this particular sampler never reject the candidate value?

   (c) What happens when $a = 0$?

4. Show that the Metropolis-Hastings algorithm generates a reversible Markov chain, i.e.,

$$\pi(x)\mathcal{K}(x, y) = \pi(y)\mathcal{K}(y, x),$$

   and show that $\pi(x)$ is the stationary distribution of the algorithm i.e.,

$$\int \pi(x)\mathcal{K}(x, y)dx = \pi(y).$$

   Note: Remember that for the M-H algorithm, $\mathcal{K}(x, y) = q(y|x)\alpha(x, y)$, where $\mathcal{K}(x, y)$ corresponds to the event that we sample y from a current state x.

5. Suppose that we observe data $\boldsymbol{x} = (x_1, ..., x_N)$, such that each $x_i \overset{iid}{\sim} Pareto(\alpha, \beta)$, where,

$$\alpha \sim Exp(m); \quad \beta \sim Exp(n),$$

   for known m and n. The $Pareto(\alpha, \beta)$ distribution has density

$$p(x) = \beta\alpha^{\beta}x^{-(\beta+1)},$$

   where $0 < \alpha \leq x$ and $\beta > 0$. Describe how we can obtain a sample from the posterior distribution of the parameters, given the data.

Bayesian Inference

Tutorial 6: Solutions

1. Substituting the p_i values into the likelihood (and after some algebra) we obtain,

$$f(\boldsymbol{x}|\phi_1,\phi_a) \propto p_1^{x_1} p_2^{x_2} \prod_{i=3}^T p_i^{x_i} = (1-\phi_1)^{x_1} \phi_1^{x_2} (1-\phi_a)^{x_2} \phi_1^{\sum_{i=3}^T x_i} (1-\phi_a)^{\sum_{i=3}^T x_i} \phi_a^{\sum_{i=3}^T (i-2)x_i}$$

$$= (1-\phi_1)^{x_1} \phi_1^{\sum_{i=2}^T x_i} (1-\phi_a)^{\sum_{i=2}^T x_i} \phi_a^{\sum_{i=3}^T (i-2)x_i}$$

The posterior distribution is given by,

$$\begin{aligned} \pi(\phi_1,\phi_a|\boldsymbol{x}) &\propto f(\boldsymbol{x}|\phi_1,\phi_a)p(\phi_1)p(\phi_a) \\ &\propto (1-\phi_1)^{x_1} \phi_1^{\sum_{i=2}^T x_i} (1-\phi_a)^{\sum_{i=2}^T x_i} \phi_a^{\sum_{i=3}^T (i-2)x_i}, \end{aligned}$$

(since $(p(\phi_1) = p(\phi_a) = 1)$). The posterior conditional distributions of ϕ_1 and ϕ_a are given by,

$$\begin{aligned} \pi(\phi_1|\phi_a,\boldsymbol{x}) &\propto (1-\phi_1)^{x_1} \phi_1^{\sum_{i=2}^T x_i} \\ \Rightarrow \quad \phi_1|\phi_a,\boldsymbol{x} &\sim Beta\left(\sum_{i=2}^T x_i + 1, x_1 + 1\right); \end{aligned}$$

and

$$\begin{aligned} \pi(\phi_a|\phi_1,\boldsymbol{x}) &\propto (1-\phi_a)^{\sum_{i=2}^T x_i} \phi_a^{\sum_{i=3}^T (i-2)x_i} \\ \Rightarrow \quad \phi_a|\phi_1,\boldsymbol{x} &\sim Beta\left(\sum_{i=3}^T (i-2)x_i + 1, \sum_{i=2}^T x_i + 1\right). \end{aligned}$$

Thus, for Gibbs sampler, set initial parameter values, denoted by $\{\phi_1^0, \phi_a^0\}$. Given state of Markov chain is $\{\phi_1^t, \phi_a^t\}$, update each parameter in turn from posterior conditional distributions, so that,

$$\begin{aligned} \phi_1^{t+1}|\phi_a^t,\boldsymbol{x} &\sim Beta\left(\sum_{i=2}^T x_i + 1, x_1 + 1\right); \\ \phi_a^{t+1}|\phi_1^{t+1},\boldsymbol{x} &\sim Beta\left(\sum_{i=3}^T (i-2)x_i + 1, \sum_{i=2}^T x_i + 1\right). \end{aligned}$$

Discard initial values of the Markov chain as burn-in (using trace plots and BGR statistic). Further realisations can be regarded as a (dependent) sample from posterior distribution of interest. However, note that here the posterior conditional distribution of $\phi_1|\phi_a,\boldsymbol{x}$ is independent of ϕ_a and similarly, $\phi_a|\phi_1,\boldsymbol{x}$ is independent of ϕ_1. In other words, given the data, ϕ_1 and ϕ_a are independent and hence can be sampled from directly (so no MCMC needs to be used here).

2. (a) First notice that,

$$\sum_i^n (x_i-\bar{x})^2 = \sum_i^n x_i^2 - 2\bar{x}\sum_i^n x_i + \sum_i^n \bar{x}^2 = \sum_i^n x_i^2 - n\bar{x}^2$$

Now, we have,

$$\begin{aligned}
\pi(\mu,\omega|\boldsymbol{x}) &\propto f(\boldsymbol{x}|\mu,\omega)p(\mu,\omega)\\
&= f(\boldsymbol{x}|\mu,\omega)p(\mu)p(\omega)\\
&\propto \prod_{i=1}^n \left[\sqrt{\frac{\omega}{2\pi}}\exp\left(-\frac{\omega}{2}(x_i-\mu)^2\right)\right]\times 1\times\frac{\beta^\alpha}{\Gamma(\alpha)}\omega^{\alpha-1}\exp\left(-\beta\omega\right)\\
&= \left(\sqrt{\frac{\omega}{2\pi}}\right)^n \exp\left(-\frac{\omega}{2}\left(\sum_{i=1}^n (x_i^2) - 2n\bar{x}\mu + n\mu^2\right)\right)\times\frac{\beta^\alpha}{\Gamma(\alpha)}\omega^{\alpha-1}\exp\left(-\beta\omega\right)\\
&\propto \omega^{n/2+\alpha-1}\exp\left(-\frac{\omega}{2}\left(\sum_{i=1}^n (x_i^2) - n\bar{x}^2 + n\bar{x}^2 - 2n\bar{x}\mu + n\mu^2\right) - \omega\beta\right)\\
&= \omega^{n/2+\alpha-1}\exp\left(-\frac{\omega}{2}\left(\sum_{i=1}^n (x_i^2) - n\bar{x}^2 + n(\mu-\bar{x})^2\right) - \omega\beta\right)\\
&= \omega^{n/2+\alpha-1}\exp\left(-\omega\left(\frac{1}{2}\sum_{i=1}^n (x_i-\bar{x})^2 + \frac{n}{2}(\mu-\bar{x})^2 + \beta\right)\right)
\end{aligned}$$

(b) We have,

$$\begin{aligned}
\pi(\mu|\omega,\boldsymbol{x}) &\propto \exp\left(-\frac{\omega n(\mu-\bar{x})^2}{2}\right)\\
\Rightarrow \mu|\omega,\boldsymbol{x} &\sim N\left(\bar{x},\frac{1}{\omega n}\right)
\end{aligned}$$

and

$$\begin{aligned}
\pi(\omega|\mu,\boldsymbol{x}) &\propto \omega^{n/2+\alpha-1}\exp\left(-\omega\left(\frac{\sum_{i=1}^n (x_i-\bar{x})^2)}{2} + \frac{n(\mu-\bar{x})^2}{2} + \beta\right)\right)\\
\Rightarrow \omega|\mu,\boldsymbol{x} &\sim \Gamma\left(n/2+\alpha, \frac{\sum_{i=1}^n (x_i-\bar{x})^2}{2} + \frac{n(\mu-\bar{x})^2}{2} + \beta\right)
\end{aligned}$$

(c) Gibbs sampler.

(d) Basic algorithm [3 marks]:

```
1. Initialize the chain with $(\mu^0,\omega^0)$
2. Given that the chain is currently at $(\mu^t,\omega^t)$
   (a) Generate $\mu^{t+1}$ from $N\left(\bar{x},\frac{1}{\omega^t n}\right)$
   (b) Generate $\omega^{t+1}$ from $\Gamma\left(n/2+\alpha,\quad \sum_{i=1}^n (x_i-\bar{x})^2/2 + n(\mu^{t+1}-\bar{x})^2/2 + \beta\right)$
   (c) Record $(\mu^{t+1},\omega^{t=1})$
```

3. Repeat step 2, $N-1$ more times

Convergence: there is no infallible method for determining whether the chain has converged to the target distribution. Trace plots of the sampled parameter values vs. draw number are often used, but are unreliable on their own. Multiple chains may be run in parallel from widely dispersed start points and convergence diagnostics such as Brooks-Gelman-Rubin used. These rely on comparing the within-chain and between-chain variances for various statistics such as sample means of each parameter. When one is happy that the chain has converged, samples taken before convergence are discarded (these come from the "burn-in" period).

(e) Assuming that the first N_b of the N samples have been discarded during "burn-in", and using previous notation:

$$\begin{aligned}\widehat{E}_\pi(\mu|\boldsymbol{x}) &= \frac{1}{N-N_b}\sum_{t=N_b+1}^{N}\mu^t\\ \widehat{E}_\pi(\omega|\boldsymbol{x}) &= \frac{1}{N-N_b}\sum_{t=N_b+1}^{N}\omega^t\end{aligned}$$

3. (a) We have that,

$$\alpha(x,y)=\min\left(1,\ \frac{f(y)q(x|y)}{f(x)q(y|x)}\right).$$

Here

$$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{x^2}{2\sigma^2}\right),$$

and

$$q(y|x)=\frac{1}{\sqrt{2\pi\tau^2}}\exp\left(-\frac{1}{2\tau^2}(y-ax)^2\right).$$

So,

$$\frac{f(y)q(x|y)}{f(x)q(y|x)}=\frac{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{y^2}{2\sigma^2}\right)\frac{1}{\sqrt{2\pi\tau^2}}\ \exp\left(-\frac{1}{2\tau^2}(x-ay)^2\right)}{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{x^2}{2\sigma^2}\right)\frac{1}{\sqrt{2\pi\tau^2}}\exp\left(-\frac{1}{2\tau^2}(y-ax)^2\right)}$$

$$=\exp\left(-\frac{1}{2\sigma^2}(y^2-x^2)\right)\exp\left(-\frac{1}{2\tau^2}\left((x-ay)^2-(y-ax)^2\right)\right).$$

Note that,

$$(x-ay)^2-(y-ax)^2=x^2-y^2-2axy+2yax+a^2y^2-a^2x^2$$

$$=(x^2-y^2)-a^2(x^2-y^2)=(x^2-y^2)(1-a^2).$$

Therefore

$$\frac{f(y)q(x|y)}{f(x)q(y|x)}=\exp\left(-\frac{1}{2}(y^2-x^2)\left(\frac{1}{\sigma^2}+\frac{a^2-1}{\tau^2}\right)\right)$$

(b) This sampler never rejects if $\alpha(x,y)\equiv 1\ \forall x,y$ and $\alpha(x,y)\equiv 1$ iff $\dfrac{1}{\sigma^2}+\dfrac{a^2-1}{\tau^2}=0$, i.e. $\tau^2=\sigma^2(1-a^2)$.

(c) If $a = 0$ then $q(y|x) = N(0, \tau^2)$ so that we are implementing an independence sampler.

Note - that in this example we would clearly simply sample directly from a $N(0, \sigma^2)$ distribution rather than using a MH algorithm!

4. Remember that for the M-H algorithm, $\mathcal{K}(x,y) = q(y|x)\alpha(x,y)$. Therefore the chain is reversible if,

$$\pi(x)q(y|x)\alpha(x,y) = \pi(y)q(x|y)\alpha(y,x).$$

Take two cases.

Case 1. $y \neq x$.

$$\begin{aligned} \pi(x)\mathcal{K}(x,y) &= \pi(x)q(y|x)\min\left(1, \frac{\pi(y)q(x|y)}{\pi(x)q(y|x)}\right) \\ &= \min\left(\pi(x)\ q(y|x),\ \pi(y)\ q(x|y)\right) \\ &= \min\left(\frac{\pi(x)q(y|x)}{\pi(y)q(x|y)}, 1\right)\pi(y)q(x|y) \\ &= \pi(y)\mathcal{K}(y,x). \end{aligned}$$

Case 2. $y = x$. Trivial. □

To show that π is the stationary distribution of the algorithm, we need to show that $E_x(\mathcal{K}(x,y)) = \pi(y)$. Intuitively, the probability that we sample y should be equal to $\pi(y)$, when we average over all possible current states x. We now have that,

$$\begin{aligned} \int \pi(x)\mathcal{K}(x,y)dx &= \int \pi(y)\mathcal{K}(y,x)dx \text{ since reversible.} \\ &= \pi(y)\int K(y,x)dx \\ &= \pi(y). \end{aligned}$$

□

5. For the $Pareto(\alpha, \beta)$ distribution, we obtain the posterior distribution of the parameters to be,

$$\begin{aligned} \pi(\alpha,\beta|\boldsymbol{x}) &\propto \prod_{i=1}^{N} \beta\alpha^{\beta}x_i^{-(\beta+1)} \times m\exp(-\alpha m)n\exp(-\beta n) \\ &\propto \beta^N\alpha^{N\beta}\prod_{i=1}^{N} x_i^{-(\beta+1)} \times \exp(-\alpha m - \beta n). \end{aligned}$$

Thus, we have that,

$$\begin{aligned} \pi(\alpha|\boldsymbol{x},\beta) &\propto \alpha^{N\beta}\exp(-\alpha m), \qquad \text{such that } \alpha \leq \min x_i; \\ \pi(\beta|\boldsymbol{x},\alpha) &\propto \beta^N\alpha^{N\beta}\exp(-\beta n)\prod_{i=1}^{N} x_i^{-(\beta+1)}. \end{aligned}$$

The posterior (full) conditional distributions of the parameters are of non-standard form.We can use a Metropolis-Hastings step to update each parameters within each iteration of the Markov chain. Suppose that we propose to update α. Then, we could use the procedure of proposing a new value,

$$\alpha' = \alpha + u,$$

where, for some fixed σ^2,

$$u \sim N(0, \sigma^2).$$

Since we need to retain $0 < \alpha \leq x_i$, for each x_i, we automatically reject the proposal if the proposed parameter $\alpha' < 0$ or $\alpha' > \min(\boldsymbol{x})$, where $\min(\boldsymbol{x}) = \min(x_1, ..., x_N)$. Assuming that the proposed parameter is not automatically rejected, we accept the proposed parameter with probability, $\min(1, A)$, where

$$\begin{aligned} A &= \frac{\pi(\alpha', \beta|\boldsymbol{x})q(\alpha|\alpha^{'})}{\pi(\alpha, \beta|\boldsymbol{x})q(\alpha^{'}|\alpha)} \\ &= \frac{\pi(\alpha', \beta|\boldsymbol{x})}{\pi(\alpha, \beta|\boldsymbol{x})} \qquad \text{since the proposal is symmetrical} \\ &= \frac{\pi(\alpha'|\beta, \boldsymbol{x})\pi(\beta|\boldsymbol{x})}{\pi(\alpha|\beta, \boldsymbol{x})\pi(\beta|\boldsymbol{x})} \\ &= \frac{\pi(\alpha'|\beta, \boldsymbol{x})}{\pi(\alpha|\beta, \boldsymbol{x})} \\ &= \frac{\alpha'^{N\beta}\exp(-\alpha' m)}{\alpha^{N\beta}\exp(-\alpha m)}. \end{aligned}$$

else the parameter remains at the same value. We use an analogous procedure for updating β.

We can construct a Markov chain, with this stationary distribution, using the following procedure. Within each step of the Markov chain we update each of the parameters using the Metropolis-Hastings step described above. Then, once the chain has been run long enough, realisations of the chain can be regarded as a sample from the posterior distribution of interest and used to obtain Monte Carlo estimates of the parameters of interest.