Bayesian course - problem set 5 (lecture 6)

Ben Lambert

November 30, 2016

1 Stan entry level: discoveries data

The file "prob5_discoveries.csv" contains data on the numbers of "great" inventions and scientific discoveries (X_t) in each year from 1860 to 1959 [1]. In this question you will try to develop a model to explain the variation in scientific inventions across time. The simplest model here is to assume that **a.** one discovery is independent of all others, and **b.** the rate of occurrence of discoveries is the same in all years (λ). Since the data is discrete, these assumptions $\implies$ use a Poisson model:

$$X_t \sim poisson(\lambda) \tag{1}$$

Problem 1.1 *Open a text editor and create a file called "discoveries.stan" in your working directory. In the file create three parameter blocks:*

```
data {

}
parameters {

}
model {

}
```

Problem 1.2 *Fill in the data and parameter blocks for the above model.*

Problem 1.3 *Using a log-normal(2,1) prior for λ code up the "model" block; making sure to save your file afterwards.*

Problem 1.4 *Open your statistical software (R, python, matlab, etc.) and load any packages necessary to use Stan. (Hint: in R this is done by using "library(rstan)'; in Python this is done using "import pystan".)*

Problem 1.5 *Load the "prob5_discoveries.csv" data and graph the data. What does this suggest about our model's assumptions?*

Problem 1.6 *Load the data into your software then put it into a structure that can be passed to Stan. (Hint: in R create a list of the data; in Python create a dictionary where the 'key' for each variable is the desired variable name.)*

Problem 1.7 *Run your model using Stan, with 4 chains, each with a sample size of 1000, and a warm-up of 500 samples. Set seed=1 to allow for reproducibility of your results. Store your result in an object called "fit".*

Problem 1.8 *Diagnose whether your model has converged by printing "'fit".*

Problem 1.9 *For your sample what is the equivalent number of samples for an independent sampler?*

Problem 1.10 *Find the central posterior 80% credible interval for λ.*

Problem 1.11 *Draw a histogram of your posterior sample for λ.*

Problem 1.12 *Create a "generated quantities" block in your Stan file, and use it to sample from the posterior predictive distribution (Hint: use the function "poisson_rng" to generate independent samples from your lambda).*

Problem 1.13 *Carry out appropriate posterior predictive checks to evaluate your model.*

Problem 1.14 *A more robust sampling distribution is a negative binomial model:*

$$X_i \sim neg\_binomial(\mu, \kappa) \tag{2}$$

where μ is the mean number of discoveries per year, and $var(X) = \mu + \frac{\mu^2}{\kappa}$. Here κ measures the degree of over-dispersion of your model; specifically if $\kappa \uparrow$ then over-dispersion$\downarrow$.

Write a new stan file called "discoveries_negbin.stan" that uses this new sampling model (Hint: use the Stan manual section on discrete distributions to search for the correct negative binomial function name; be careful there are two different parameterisations of this function available in Stan!) Assume that we are using the following priors:

$$\mu \sim log\text{-}normal(2, 1) \tag{3}$$

$$\kappa \sim log\text{-}normal(2, 1) \tag{4}$$

$$\tag{5}$$

Draw 1000 samples across 4 chains for your new model. Has it converged to the posterior?

Problem 1.15 *Carry out posterior predictive checks on the new model. What do you conclude about the use of a negative binomial here versus the simpler Poisson?*

Problem 1.16 *Find the central posterior 80% credible interval for the new mean rate of discoveries μ. How does it compare with your results from the Poisson model? Why is this the case?*

Problem 1.17 *(Optional) Calculate the autocorrelation in the residuals between the actual and simulated data series. What do these suggest about our current model?*

Problem 1.18 *(Optional) Following on from the above suggest an alternative model formulation.*

2 Stan entry level : Hungover holiday regressions

The data in file "hangover.csv" contains a series of Google Trends estimates of the search traffic volume for the term "hangover cure" in the UK between February 2012 to January 2016. The idea behind this problem is to determine how much more hungover are people in the "holiday season" period, defined here as the period between 10th December and 7th January, than the average for the rest of the year.

Problem 2.1 *Graph the search volume over time, and try to observe the uplift in search volume around the holiday season.*

Problem 2.2 *The variable "holiday" is a type of indicator variable that takes the value 1 if the given week is all holiday season, 0 if it contains none of it, and $0 < X < 1$ for a week that contains a fraction X of days that fall in the holiday season. Graph this variable over time so that you understand how it works.*

Problem 2.3 *A simple linear regression is proposed of the form,*

$$V_t \sim N(\beta_0 + \beta_1 h_t, \sigma) \tag{6}$$

where V_t is the search volume in week t and h_t is the holiday season indicator variable. Interpret β_0 and β_1 and explain how these can be used to estimate the increased percentage of hangovers in the holiday season.

Problem 2.4 *Assuming $\beta_i \sim N(0, 50)$ and $\sigma \sim$ half-normal$(0, 10)$ priors write a Stan model to estimate the percentage increase in hangoverness over the holiday period.*

3 Stan middle level: coding up a bespoke probability density

In the file "prob5_survival.csv" there is data for a variable Y that we believe comes from a probability distribution $p(Y) = \frac{\sqrt[3]{b}}{\Gamma(\frac{4}{3})} exp(-bY^3)$ where $b > 0$ is a parameter of interest. In this question we are going to write a Stan program to estimate the parameter b even though this distribution is not amongst Stan's implemented distributions!

Problem 3.1 *Explain what is meant by the following statement in Stan,*

```
theta ~ beta(1,1);
```

In particular, explain why this is equivalent to the following,

```
target += beta_lpf(theta|1,1);
```

where `target` *is a Stan variable that stores the overall log-probability, and* += *increments* `target` *by an amount corresponding to the RHS.*

Problem 3.2 *By hand work out an expression for the log-probability for a density* $p(Y) = \frac{\sqrt[3]{b}}{\Gamma(\frac{4}{3})} exp(-bY^3)$.

Problem 3.3 *Write a Stan function that for a given value of* y *and* b *calculates the log probability (ignoring any constant terms). Hint: Stan functions are declared as follows,*

```
functions{
  real anExample(real a, real b){
    ...
    return(something);
  }
}
```

where in this example the function takes two reals as inputs and outputs something of type real.

Problem 3.4 *Use your previously created function to write a Stan program that estimates b, and then use it to do so with the y series contained within "prob5_survival.csv". Hint: Stan functions must be declared at the top of a Stan program.*

4 Stan advanced level: is a tumour benign or malignant?

Suppose that if a tumour is benign the result of a clinical test for the disease for individual i is $X_i \sim \text{binomial}(20, \theta_b)$, whereas if the tumour is malignant $X_i \sim \text{binomial}(20, \theta_m)$, where $\theta_b < \theta_m$. Suppose that we collect data on 10 patients' scores on this clinical test $X = \{4, 18, 6, 4, 5, 6, 4, 6, 16, 7\}$ and would like to infer the disease status for each individual, as well as the parameters (θ_b, θ_m).

Problem 4.1 *Write down in pseudo-code the full model, where we suppose that we use uniform priors on (θ_b, θ_m) and discrete uniform priors on the disease status s_i of individual i.*

Problem 4.2 *Assuming that $s_i \in [1, 2]$ is the disease status of each individual (1 corresponding to a benign growth, and 2 to a malignant one), use the* `transformed parameters` *block to calculate the log probability of each individual's data. Hint: this will be a 10 × 2 matrix, where the 2 corresponds to each possible disease status.*

Problem 4.3 *The disease status of each individual $s_i \in [1, 2]$ is a discrete variable, and because Stan does not support discrete parameters directly it is not as straightforward to code up these problems as for continuous parameter problems. The way that to do this is by marginalising out s_i from the joint distribution,*

$$p(\theta_b, \theta_m | X) = \sum_{s_1=1}^{2} p(\theta_b, \theta_m, s_1 | X) \tag{7}$$

where we have illustrated this for the disease status of individual 1. This then allows us to find an expression for the posterior density which we log to give lp, and then use `target+=lp` *to increment the log probability. However, because we do this on the log density scale we instead do the following,*

$$\begin{aligned} log\, p(\theta_b, \theta_m | X) &= log \sum_{s_1=2}^{K} p(\theta_b, \theta_m, s_1 | X) && (8) \\ &= log \sum_{s_1=1}^{2} exp\left(log\, p(\theta_b, \theta_m, s_1 | X)\right) && (9) \\ &= log\_sum\_exp_{s_1=1}^{2}(log\, p(\theta_b, \theta_m, s_1 | X)) && (10) \end{aligned}$$

where `log_sum_exp(.)` *(a function available in Stan) is defined as,*

$$log\_sum\_exp_{i=1}^{K}\alpha = log \sum_{i=1}^{K} exp(\alpha) \tag{11}$$

and is a more numerically-stable way of doing the above calculation. Using this knowledge, write a full Stan model that implements this marginalisation, and use it to estimate θ_b and θ_m. Hint: use the `binomial_logit_lpmf(X[i]|N,alpha[s])` *function in Stan and define* `ordered[2] alpha`*, then transform from the unconstrained alpha to theta using* `inv_logit`*.*

Problem 4.4 *We use the* `generated quantities` *block to estimate the probabilities of state $s = 1$ in each different experiment by averaging over all L posterior draws,*

$$q(s=1|X) \approx \frac{1}{L}\sum_{i=1}^{L} q(s=1, alpha[s=1]|X) \tag{12}$$

where q(.) is the un-normalised posterior density. The averaging over all posterior draws is necessary to marginalize out the alpha parameter. To normalise the posterior density we therefore divide the above by the sum of the un-normalised probability across both states,

$$Pr(s=1|X) = \frac{q(s=1|X)}{q(s=1|X)+q(s=2|X)} \tag{13}$$

Using the above knowledge add a `generated quantities` block to your Stan model that does this, and hence estimate the probability that each individual's tumour is benign.

Problem 4.5 *(Optional) An alternative way to code this problem is to derive a Gibbs sampler. As a first step in this process write out the full joint posterior numerator. Hint: now use a slightly-altered definition of $s_i \in [0,1]$, where 1 indicates a benign tumour for individual i.*

Problem 4.6 *By removing those terms that don't depend on θ_b derive the conditional distribution $\theta_b|\theta_m, S, X$. Hence write down $\theta_m|\theta_b, S, X$*

Problem 4.7 *Show that the distribution for $s_i|s_{-i}, \theta_b, \theta_m, X$ can be written as,*

$$s_i|s_{-i}, \theta_b, \theta_m, X \sim bernoulli\left(\frac{1}{1+\left[\frac{\theta_m}{1-\theta_m}/\frac{\theta_b}{1-\theta_b}\right]^{X_i}\left[\frac{1-\theta_m}{1-\theta_b}\right]^{20}}\right) \tag{14}$$

Problem 4.8 *Using your three derived conditional distributions create a Gibbs sampler in R, and use it to estimate $(\theta_b, \theta_m, s_1, ..., s_{10})$.*

5 Stan advanced level: how many times did I flip the coin?

Suppose that I have a coin with θ denoting the probability of it landing heads-up. In each experiment I flip the coin N times, where N is unknown to the observer, and record the number of heads obtained Y. I repeat the experiment 10 times, each time flipping the coin the same N times, and record $Y = \{9, 7, 11, 10, 10, 9, 8, 11, 9, 11\}$ heads.

Problem 5.1 *Write down an expression for the likelihood, stating any assumptions you make.*

Problem 5.2 *Suppose that the maximum number of times the coin could be flipped is 20, and that all other (allowed) values we regard a priori as equally probable. Further suppose that based on previous coin flipping fun that we specify a prior $\theta \sim$ beta$(7, 2)$. Write down the model as a whole (i.e. the likelihood and the priors).*

Problem 5.3 *This problem can be coded in Stan by marginalising out the discrete parameter N. The key to doing this is writing down an expression for the log-probability for each result Y_i conditional on an assumed value of N, and θ. Explain why this can be written in Stan as,*

```
log(0.1) + binomial_lpmf(Y[i]|N[s],theta);
```

where `N[s]` *is the sth element of a vector N containing all possible values for this variable.*

Problem 5.4 *In the* `transformed parameters` *block write code that calculates the log probability for each experiment and each possible value of N.*

Problem 5.5 *Write a Stan program to estimate θ. Hint: in the* `model` *block use* `target+= log_sum_exp(lp)` *to marginalise out N and increment the log probability.*

Problem 5.6 *Use the* `generated quantities` *block to estimate the probabilities of each state.*

Problem 5.7 *An alternative way to estimate N and θ is to derive a Gibbs sampler for this problem. To do this first show that the joint (un-normalised) posterior distribution can be written as,*

$$p(\theta, N|Y) \propto \left[\prod_{i=1}^{K} \binom{N}{Y_i} \theta^{Y_i}(1-\theta)^{N-Y_i}\right] \theta^{\alpha-1}(1-\theta^{\beta-1}) \tag{15}$$

where $K = 10$ and $(\alpha, \beta) = (7, 2)$ are the parameters of the prior distribution for θ.

Problem 5.8 *Derive the conditional distribution $\theta|N, Y$. Hint: remove all parts of the joint distribution that do not explicitly depend on θ.*

Problem 5.9 *Write an R function that independently samples from the conditional distribution $\theta|N, Y$.*

Problem 5.10 *Show that the conditional pmf $N|\theta, Y$ can be written as,*

$$p(N|\theta, Y) \propto \left[\prod_{i=1}^{K} \binom{N}{Y_i}\right] (1-\theta)^{NK} \tag{16}$$

Problem 5.11 *Using the previously-derived expression, write a function that calculates the un-normalised conditional $N|\theta, Y$ for $N = 11, ..., 20$, which when normalised can be used to sample a value for N. Hint use the* `sample` *function in R.*

Problem 5.12 *Write a working Gibbs sampler using your two previously-created functions, and use this to estimate the probability distribution over θ and N.*

Problem 5.13 *(Optional) Compare the rate of convergence in the mean of N sampled via Gibbs with that over that estimated from the $p(N)$ distribution that you sampled in HMC. Why is the rate of convergence so much faster for HMC? Hint: this is not due to the standard benefits of HMC that I extolled in the lecture.*

References

[1] *The World Almanac and Book of Facts*. 1975.