UNIVERSITY OF EDINBURGH
SCHOOL OF MATHEMATICS
BAYESIAN DATA ANALYSIS

Assignment 1

- To be uploaded to Learn by 23:59, Sunday March 10, 2019.
- This assignment is worth 50% of your final grade for the course.
- Assignments should be typed (LaTeX, word, etc.) and should be no more than 10 pages (including figures but excluding the appended code).
- Answers to questions should be in full sentences.
- Any output (e.g., graphs, tables) from R/JAGS that you use to answer questions must be included with the assignment. Also, please append your R/JAGS code at the end of the assignment.
- The assignment is out of 100 marks.
- You are expected to work independently and not discuss the assignment with others.

1. (**30 marks**) A study is conducted to measure the log-concentration of a particular chemical in soil. It can be assumed that the log-concentration follows a Normal distribution with mean μ and variance σ^2. A set, $\{y_i\}$, of $n = 10$ measurements are made of the log-concentration in soils in a certain region of the UK with the following results

$$-0.566, 3.74, 5.55, -1.90, -3.54, 5.16, -1.76, 4.08, 4.62, 0.732.$$

The primary parameter of interest is μ, the unknown mean of the distribution. We assume initially that $\sigma^2 = 30$ is known.

   (a) (**5 marks**) We consult a panel of 10 UK experts and they believe that $\mu \approx 1$, but could take values in the range $[0, 2]$. Construct a suitable conjugate prior based on this information and obtain the posterior distribution from the experimental data. Include a brief justification of your choice of prior.

   (b) (**5 marks**) Suppose now that we had also consulted a panel of 5 US experts, and their opinion was that $\mu \approx 5$ but could be between 3 and 7. Derive a suitable mixture prior that combines the opinions of both sets of experts and obtain the posterior distribution for this new prior.

   (c) (**8 marks**) Now suppose that σ^2 is also unknown. Using a suitable prior on the precision, $\tau = 1/\sigma^2$, and the same mixture prior on μ, obtain samples from the posterior distribution numerically, using JAGS or another package of your choice. You may wish to use the various convergence diagnostics that were discussed in lectures to verify the accuracy of your posterior distribution, but you do not need to report the results of such checks in your solution. Obtain estimates of the posterior mean, median, and the lower and upper quartiles from your samples.

(d) (**6 marks**) Based on your results, what is the posterior probability that $\mu < 1$? If we take 5 additional measurements, what is the probability that at least one of them will return a negative log-concentration?

(e) (**6 marks**) In building the mixture prior in part (b) you would have used some weighting between the two groups of experts, $p(\mu) = wp_1(\mu) + (1-w)p_2(\mu)$. We will now include the weight w as an additional model parameter. In the case that $\sigma^2 = 30$ is known, obtain the combined posterior distribution on (μ, w), using a flat prior for $w \in [0, 1]$. Obtain also the marginal distributions on μ and w and comment on the result.

2. (**35 marks**) The Laser Interferometer Ground Observatory (LIGO) detected gravitational waves for the first time in September 2015. LIGO has now completed two observing runs. The first run (O1) lasted 3 months, during which time 3 signals from binary black hole mergers were observed. The second observing run (O2) lasted 6 months. In the first 5 months, one additional merger was observed, and then in the last month 5 further signals were detected. We may assume that the events are distributed according to a homogeneous Poisson distribution with parameter λ with units of yr^{-1}. We are interested in inferring the value of λ. Prior to O1 the value of λ was poorly constrained, with estimated values ranging from 0.01 to 1000.

(a) (**6 marks**) Consider the information available prior to the first observing run and construct a conjugate prior for the rate parameter. Briefly justify your reasons for constructing a prior in this way.

(b) (**5 marks**) Derive the posterior distribution for λ using the O1 observations. Report the posterior mean, standard deviation, a 95% symmetric confidence interval and plot the posterior distribution.

(c) (**4 marks**) What is the posterior probability that the rate is > 15?

(d) (**5 marks**) Re-analyse the O1 data using a Jeffreys prior; how do your results in (b) and (c) change?

(e) (**7 marks**) Based on the posterior from the O1 data, what is the posterior predictive probability that we would see 6 or more events in O2? What is the posterior predictive probability that we would see 1 or fewer events in the first 5 months of O2? What is the posterior predictive probability that we would see 5 or more events in the last month of O2/during at least one month of O2? How are these results affected by the choice of prior? Some authors have claimed that the LIGO results provide evidence that the rate is not homogeneous in time. Based on these results, do you agree?

(f) (**8 marks**) The third observing run, O3, will start in April 2019 and will last for one year. Update your posterior distribution to use all of the events from O1 and O2, using one of the previous prior choices. Obtain the posterior predictive distribution for the difference, $|n_2 - n_1|$, in the number of events observed in the first 6 months, n_1, and the last 6 months, n_2, of O3. How large would this difference have to be to provide evidence that the rate is changing with time? Discuss other possible ways to address the question 'is the rate changing with time?' within a Bayesian framework.

3. (**35 marks**) Data on rocks samples from a petroleum reservoir give the permeability of the rock (in milli-Darcies), the area of pore space (in pixels), the perimeter of the pores (in pixels) and the shape of the pores (computed by dividing the perimeter by the square root of the area). The data are available in the `R` dataset `rock`. We would like to identify the factors that affect the permeability of the rock. You'll use Bayesian linear regression to answer this question. In order to do that you should:

   - Carry out exploratory data analyses.
   - Decide on a Bayesian linear regression model for analysing this dataset (with permeability as the response). Describe the likelihood as well as the priors for all parameters.
   - Fit the model and report posterior summaries for all quantities of interest.
   - Check convergence diagnostics.
   - Check the sensitivity to the prior distribution (this might include changing the hyperparameter values and/or the distribution used).
   - Perform model checks.

   Briefly, in words, summarise your main conclusions from analysing this dataset.