University of Edinburgh
School of Mathematics
Bayesian Data Analysis, 2018/2019, Semester 2
Jonathan Gair & Rubén Amorós Salvador

April 8, 2019

Assignment 2

- To be uploaded to Learn by 23:59, Sunday 21 April, 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 using a type size no smaller than 11 point and with 1.5-2.0 line space. This includes figures but excludes the appended code. Document your code so that someone can read it without too much guesswork.
- 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. You will want to be judicious with what you include in the written report—not every figure and table you constructed needs to be included. 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.
- We recommend that you examine the advice given during lecture 6 about both writing assignments and carrying out Bayesian analysis.
- Briefly indicate the technical details of the MCMC analyses you perform (number of iterations, convergence checks, etc.).
- You should write the hierarchical expressions (in terms of probability distributions) of all your models.

1. *Modelling abundance of gulls* **(44 marks)**

   The Columbretes Islands[1] is a group of uninhabited islets in the east coast of Spain. Two species of gulls (Audouin's[2] and Yellow-legged[3]) nest there every year. The cycle of reproduction of the Yellow-legged gulls starts earlier in the year so they start nesting in the islets earlier than the Audouin's gulls, and ecologists theorize that the Yellow-legged gulls are displacing the Audouin's gulls. The file `gulls_data.csv` contains the complete census counts of the nesting couples of both species of gulls during the years 1987-2012.

   (a) **(3 marks)** Perform some exploratory data analysis including a graph of the temporal evolution of the abundance of Audouin's gulls and a graph showing the relation between the abundance of the two species of gulls. Briefly comment this analysis.

[1] https://en.wikipedia.org/wiki/Columbretes_Islands
[2] https://en.wikipedia.org/wiki/Audouin%27s_gull
[3] https://en.wikipedia.org/wiki/Yellow-legged_gull

(b) **(10 marks)** Fit a Bayesian Poisson model with a logarithm link function where the abundance of the Audouin's gull couples is considered to be dependent on the year. Summarize the estimated parameters and briefly interpret them.

(c) **(9 marks)** Expand the previous model by including an extra variance term in the regressor. Set a Gaussian prior for this extra variance term with zero mean and standard deviation with a uniform hyper-prior distribution between 0 and 10.

(d) **(5 marks)** Further expand the previous extra variance model by including the number of Yellow-legged gull couples as an explanatory variable. Discuss the posterior estimation for the parameter of the Yellow-legged gull. Based on the exploratory analysis, is it what you expected to obtain? Consider in your discussion the variability of abundance of resources (food) among years.

(e) **(12 marks)** Compute random samples of the predictive distributions for replicates of your database (same years and abundance of Yellow-legged gull as covariates) for the three models. Plot the abundance of Audouin's gulls with the posterior predictive expected values and 90% credible interval bands of the replicates for the three models.

NOTE: Take into account that you should NOT compute the posterior predictive distribution of the replicates conditional on the extra variance term for each particular observation ($Pr(Y_i^{rep}|\boldsymbol{Y}, \varepsilon_i)$). This extra variance term should instead be integrated out in the predictive distribution (you should compute $Pr(Y_i^{rep}|\boldsymbol{Y})$).

(f) **(5 marks)** Compute the Deviance Information Criterion (DIC) of the three models and compare the performance of the three models (according to DIC and all the previous analyses). In particular you should discuss the relevance of the extra variance term and of the covariate term. Which model or models do you consider adequate for modelling this data?

2. *Modelling presence of parasites in cows* **(44 marks)**

Hairworms are a common parasite that can infect the gastrointestinal tract of cows. A study was performed to assess risk factors related to the presence of this parasite in cattle. The file `cows.csv` contains data on 279 cows (one per rows) from 18 different farms with the following variables:

cowID Identification number of the cow.

farmID Identification number of the farm of the cow.

temp Average daily maximum temperature (Celsius) measured at the nearest weather station to the farm.

rain Average yearly rain (millimetres) measured at the nearest weather station to the farm.

permeab Permeability of the soil around the farm (1-low permeability, water can stagnate, 2-high permeability, water is absorbed by the soil easily)

height Height of the location of the farm (meters).

slope Average slope of the fields around the farm (percentage).

age Age of the cow (years)

parasite Presence (1) or absence (0) of the parasite in the gastrointestinal tract of the cow.

It should be noted that only a sub-sample of the cows of each farm was analysed.

(a) **(3 marks)** Perform some exploratory analysis of the database, including an analysis of the correlation between the variables of the environment of the farm (temperature, rain, permeability, height and slope). Briefly comment this analysis.

(b) **(12 marks)** Fit a Bayesian hierarchical Bernoulli logistic model where the probability for each cow of having hairworm parasites is explained by their own covariates (age), the covariates of the environment of the farm (temperature, rain, permeability, height and slope) and a random effect on the farms themselves. Discuss the estimates for the posterior distribution of the parameters.

(c) **(7 marks)** Considering the results of the correlation between the variables of the environment of the farm, try some simplifications of the model and select one (you can use DIC as a criterion of selection). Have the parameters of the variables remaining in the model changed substantially? Briefly discuss this.

(d) **(13 marks)** Based on the model you have selected, estimate the posterior expected value and 95% credible interval of the proportion of cows in farm 1 that have parasites (assume age equal to the mean age in the study). A farm is declared in epidemic state if the proportion of cows with the parasite in that farm is larger than 20%. What is the probability of farm 1 to be in the epidemic state? Perform the same analysis for farm 6.

(e) **(9 marks)** Modify the last model substituting the hierarchical random effect of the farm for a fixed effect of the farm. Estimate the posterior expected value and 95% credible interval of the proportion of cows in farm 1 and in farm 6 that have parasites, as well as the probability for each farm to be in the epidemic state. Discuss and compare the results for the fixed effects model and the random effects hierarchical model.

3. *Proposing a new model* **(12 marks)** Propose an extension, modification or an alternate model for either the gulls data or the cows data. Do the Bayesian inference for the proposed model and discuss the results comparing them to the models fitted in question 1 or 2.

   Some possible ideas are: using state space models, changing the likelihood distribution that models the data, changing the link function, considering interactions between variables or considering another variable as the response variable. Only one of those changes –or any other change in the same line– is enough for full marks (as long as it makes sense and the inference is adequately performed and discussed).