STAT40850 - Bayesian Analysis

Computer Lab 3 – Simple statistical modelling and model comparison

Submission deadline: April 5th at 6pm.

Now that we can run JAGS to sample from the posterior given a likelihood and priors we need to be able to

1. Assess the performance of the algorithms.
2. Compare various models.

1 CODA

In order to assess whether the sampling algorithms built and run by JAGS are performing well we will use the coda (Convergence Diagnosis and Output Analysis) package. In order to use the functions provided by coda we need our MCMC output in a certain format. You can simply replace calls to `jags.samples()` with `coda.samples()`; the format of output will be different but the inference the exact same. The output will now be in the form of a Markov Chain Monte Carlo object (as defined by coda).
Assuming you save the output from a call to `coda.samples()` as "samps" you can then use the following functions (among others):

- `plot(samps)` to create side-by-side trace trace and density plots for the samples from each parameter in the model. Or individually use traceplot for the traces only. Useful to assess mixing and convergence visually.
- `autocorr(samps)` and `acutocorr.plot(samps)` for auto-correlation and plots to identify slow mixing and suggest a thinning interval.
- `crosscor(samps)` and `crosscor.plot(samps)` to identify parameters that are strongly correlated. Correlated parameters is a cause of slow mixing and may suggest that a reparameterization is required.
- `HPDinterval(samps)` gives a Highest Posterior Density interval for all parameters sampled.
- Finally, try `codamenu()` for an interactive setup.

2 Linear regression models

Where we have a response variable y and covariates $x_1, \ldots, x_p$ we may wish to fit a linear regression model, such that:

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + \epsilon_i$$

with $\epsilon_i \sim N(0, \sigma^2)$. The model above defines a simple conditionally independent likelihood. We may assume Jeffreys priors so that $p(\alpha, \beta_1, \ldots, \beta_p, \sigma) \propto \frac{1}{\sigma^2}$. You can achieve this in JAGS by using `dunif()` priors for α, β and `dgamma(0.5,1e-6)` for τ where $\tau = \frac{1}{\sigma^2}$. You can check that this yields a Jeffrey's prior for τ that is proportional to $\tau^{-\frac{1}{2}}$

There are many extensions to such models. For example, polynomial regressions occur when $x_j = x^j$ for $j = 1, \ldots, p$. The parameters themselves might vary for each observation, creating a *hierarchical model* (covered later in the course), or the residual ϵ_i terms might **not** be normally distributed, which in some cases leads to a *generalised linear model*.

Tasks:

- Run the file `lr1.R`. What model is being fitted here? Include `mu` in `variable.names` and plot the mean sampled `mu` against `X`.
- Re-run using the file `lr2.R`. What has changed? Use crosscor() to calculate the correlations between the parameters `alpha`, `beta1`, and `beta2`. What is the difference between the two models? Why might one be preferred over the other?

3 Generalised linear models

In cases where the residuals ϵ are not normally distributed, but are a member of the exponential family, we can still model the mean (or more usually a function of the mean) as a linear function of explanatory variables. There are numerous examples:

- Poisson likelihood: $y_i \sim Po(\lambda_i)$, $\log(\lambda_i) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}$
- Binomial likelihood: $y_i \sim Bin(n, p_i)$, $\text{logit}(p_i) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}$
- Exponential likelihood: $y_i \sim Exp(\gamma_i)$, $\gamma_i^{-1} = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}$

In each case we must specify prior distributions on the parameters $\alpha, \beta_1, \ldots, \beta_p$. The transformation of the mean (eg log, logit, inverse, above) is known as the *link* function. The list above shows the most popular link functions for the specified models, though there are other choices that may be used.

Tasks:

- Open and run the file `glm1.R`. The data are counts of mining disasters between 1851 and 1962. Which of the above models is being fitted here?
- Re-run the model, making sure to include `lambda` in variable.names. Does this affect the running of your model? What use might the values of `lambda` be?
- What does the commented out line create? Remove the comment sign and re-run the model, monitoring the new variable `ystar`. What is the 95% credibility interval for `ystar`?

4 Comparing models: DIC

Later in the course we will meet a model comparison tool known as the *Deviance Information Criterion* or DIC. The DIC is similar to the AIC and BIC you may have met before. These tools allow for comparison between different models with differing structures and numbers of parameters, provided all models use the same data. The DIC is specifically built for Bayesian models and is calculated via the formula:

$$DIC = \bar{D} + p_D$$

where $\bar{D}$ is the mean deviance (a measure of model fit), and p_D is the effective number of parameters. We will discuss the DIC in more detail in lectures, but for now we will assume that lower DICs will lead to better models. The DIC is implemented in JAGS as a separate sampling function dic.samples() or by including "deviance" and "pD" in variable.names in a jags.samples() call. You must run more than 1 chain as the effective number of parameters is calculated across chains.

Tasks:

- Open and run the files `glm1.R` and `glm2.R`. What is the difference between the models?
- Run each model and calculate the DIC. Which model is preferred? What are the values of p_D? Do they relate to the number of parameters in the model?

5 Homework: Putting distances

Some data are available on the number of successful putts from various distances for professional golfers. The data are as follows:

distance(feet)	no of tries	no of successes
x	n	y
2	1443	1346
3	694	577
4	455	337
5	353	208
6	272	149
7	256	136
8	240	111
9	217	69
10	200	67
11	237	75
12	202	52
13	192	46
14	174	54
15	167	28
16	201	27
17	195	31
18	191	33
19	147	20
20	152	24

Choose a suitable linear generalised linear model and fit the data in JAGS. Use DIC to choose between different models you might like to fit. Estimate the proportion of successes from 5, 10, and 30 feet. Write a short report of 2-3 pages with your final model code as an appendix.