Bayesian Inference

Tutorial 5

1. In the lecture on BUGS, we looked at code that reproduced the analysis for the Example on the lecture notes, where the sampling distribution was Exponential, with a Gamma prior for the exponential parameter. See the code uploaded on Moodle. Comment on any similarity/difference between the posterior mean for $1/\lambda$ and for the predicted observation $x.new$. Justify, intuitively, any similarity or difference. Do the same for the posterior variances.

2. Use NIMBLE to reproduce other conjugate analyses in the lectures notes. Specifically, the Binomial likelihood - Beta prior example on page 19 (Section 1.4.2), and the Normal $N(\mu, \sigma^2)$ likelihood with independent Normal and Inverse Gamma priors on page 31 (Section 1.10), using the data from the example on page 13.

3. (From 2011 exam; edited.) Interest lies in modelling the biological system of the number of canvasback ducks in a given region over time. Let $\boldsymbol{x} = \{x_1, \ldots, x_T\}$ denote the observed size of the population for times $t = 1, \ldots, T$. The following density dependent model is proposed where the x_t are assumed to have a log-Normal distribution of the form,

$$X_t | x_{t-1}, \theta_0, \theta_1, \sigma^2 \sim \log N\left(\log(x_{t-1} \exp(\theta_0 + \theta_1 x_{t-1})), \sigma^2\right),$$

where θ_0, θ_1 and σ^2 are parameters to be estimated. Here, density dependence refers to the dependence of X_t on X_{t-1}, in particular through the exponential effect of X_{t-1} for a non-zero θ_1. We form the joint posterior distribution $\pi(\theta_0, \theta_1, \sigma^2 | \boldsymbol{x})$ and use the following BUGS[1] code:

```
for (t in 2:T) {
  Ex[t] <- log(x[t-1]*exp(param1+param2*x[t-1]))
}
for (t in 2:T) {
  x[t] ~ dlnorm(Ex[t],param4)
}
param1 ~ dnorm(0,0.1)
param2 ~ dnorm(0,0.1)
param3 ~ dunif(0,100)
param4 <- 1/(param3*param3)
```

```
list(T=10, x = c(10.03,5.87,7.52,6.37,7.77,6.81,7.63,7.15,7.14,8.34))
list(param1 = -0.5, param2 = 0.5, param3 = 1)
```

The BUGS code is run for 20,000 iterations with the trace plots for the three parameters, `param1`, `param2` and `param3` provided in Figure 1. The corresponding posterior summary estimates of these parameters are provided in Table 1, after a suitable burn-in has been removed.

[1] Note that we now examine your knowledge of the BUGS language through the assessed practical (for MT4531), and not in the exam. So, if this exam question were set now, we'd describe the statistical model using mathematical equations only, rather than also use raw BUGS code; we wouldn't therefore ask questions like parts (a) and (e) in today's exam. (This also holds for MT5731.)

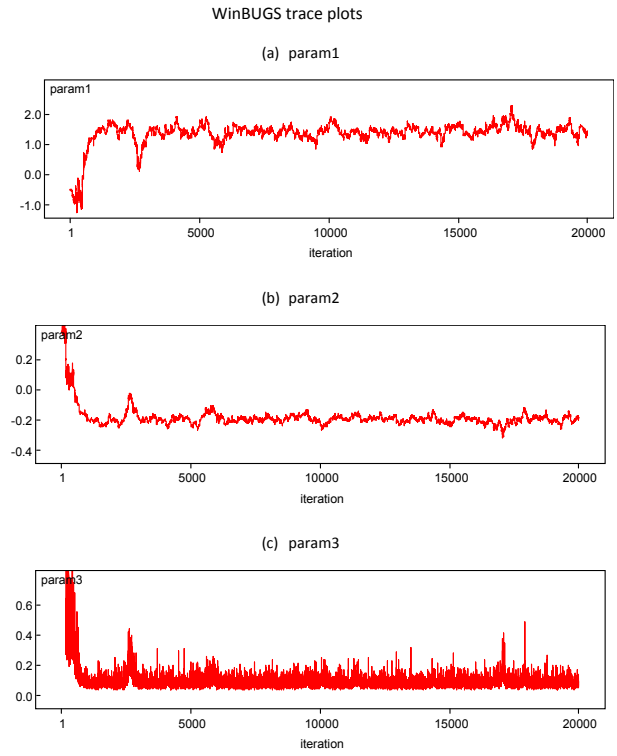


Figure 1: Trace plots and posterior summary statistics

node	**mean**	**sd**	**MC error**	**2.5%**	**median**	**97.5%**
param1	1.426	0.1843	0.01438	1.045	1.422	1.809
param2	-0.1965	0.02486	0.001939	-0.2477	-0.196	-0.146
param3	0.08501	0.02932	8.557E-4	0.04824	0.07876	0.1584

Table 1: Posterior summary statistics.

(a) State the priors specified on the parameters `param1`, `param2` and `param3` and relate these to the parameters, θ_0, θ_1 and σ^2. [2]

(b) Suggest a suitable burn-in, justifying your choice. [1]

(c) State and briefly describe the Brooks-Gelman-Rubin (BGR) convergence diagnostic tool, including how to assess whether convergence has been achieved. [3]

(d) By considering the summary information output in Table 1, suggest whether the chain has been run long enough to obtain reliable posterior estimates or whether the chain should be run for longer, justifying your answer. [1]

(e) We wish to predict the true population size at time $T + 1$ denoted by x_{T+1} (i.e. the one-step ahead population size). Provide the additional BUGS code that is needed in the model component that would allow us to obtain a posterior estimate of x_{T+1}. [3]

The biologist conducting the analysis then wishes to consider the question of whether there is evidence of density dependence, as in the above model, against the alternative model of no density dependence corresponding to adding the parameter restriction $\theta_1 = 0$.

(f) Without conducting a formal analysis[2], but simply given the posterior summary estimates of the model parameters, suggest whether there is evidence of density dependence or not, justifying your answer. [1]

4. (From 2009 exam; edited)[3] A study is undertaken to monitor the feeding choices of 221 alligators. The response measure for each alligator is classified into five categories: fish, invertebrate, reptile, bird, other. Two possible explanatory factors are considered; the length of the alligator (small or large), and the lake (four locations) in which the alligators were monitored. The observed count X_{ijk} then gives the number of alligators of size j, located in lake i, eating food type k. The following BUGS code specifies Model A:

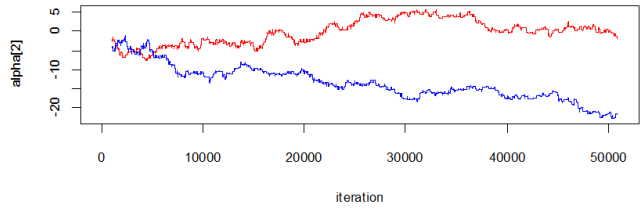
```
model {
  #  PRIORS
  for (k in 1 : K) {
    alpha[k] ~ dnorm(0, 0.00001)
  }
  for (i in 1 : I) {
    for (k in 1 : K){
      beta[i, k] ~ dnorm(0, 0.00001)
    }
  }
  for (j in 1 : J) {
    for ( k in 1 : K){
      gamma[j, k] ~ dnorm(0, 0.00001)
    }
  }
  # LIKELIHOOD
  for (i in 1 : I) {
    for (j in 1 : J) {
      X[i,j,1 : K] ~ dmulti( p[i, j, 1 : K] , n[i, j]  )
      n[i, j] <- sum(X[i, j, 1 : K])
      for (k in 1 : K) {
        p[i, j, k] <- phi[i, j, k] / sum(phi[i, j, 1 : K])
        log(phi[i ,j, k]) <- alpha[k] + beta[i, k]  + gamma[j, k]
      }
    }
  }
}
```

(a) Write down the model that has been fitted ensuring that you specify the form of the likelihood and the priors. [3]

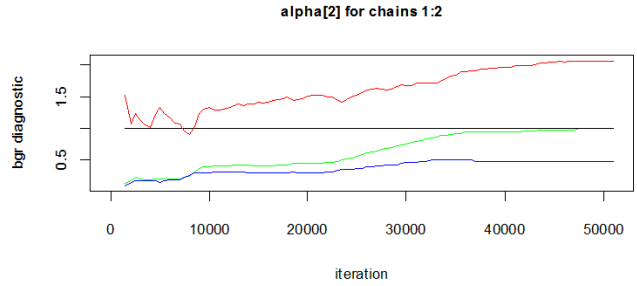
[2] We will discuss formal approaches to answer questions like this one later in the course.
[3] This was quite a hard question!

(b) Another model, called 'Model B' is now fitted to the same data. Based on the output in Figure 2 (obtained with the `OpenBUGS` software), providing the trace plots and Brooks-Gelman-Rubin (BGR) diagnostic plots for parameter α_2 for models A and B, are there any problems present in the fitting process for either model? If so, what are they and what steps would you take to fix them? [5]

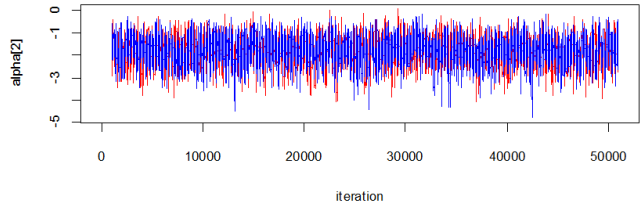
Figure 2: Graphical Information for Question 2



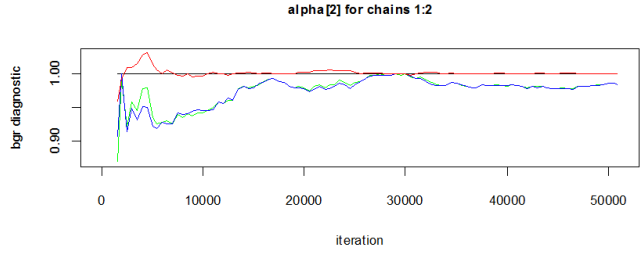
(a) Trace plots for parameter α_2 under model A.



(b) BGR diagnostic plot for α_2 under model A. The R ratio is represented in red, its numerator (pooled) in green and denominator (average) in blue.



(c) Trace plots for parameter α_2 under model B.



(d) BGR diagnostic plot for α_2 under model B. The R ratio is represented in red, its numerator (pooled) in green and denominator (average) in blue.

Bayesian Inference

Tutorial 5: Solutions

1. The posterior expectation of the average lifetime $E_\pi(1/\lambda)$ is the same as the posterior expectation of a new lifetime observation, allowing for Monte Carlo error. This makes sense because, on average, the new observation is generated by an exponential with parameter $E_\pi(\lambda)$.

   On the other hand, the posterior variance of $1/\lambda$ is smaller than the posterior variance for a future lifetime observation. This is because, intuitively, the variability of the average for all laptops is smaller than the variability for a single laptop, as the latter also includes random variation from laptop to laptop.

2. See code uploaded on Moodle.

3. (a) `param1` $= \theta_0 \sim N(0, 10)$; `param2` $= \theta_1 \sim N(0, 10)$; `param3` $= \sigma \sim U[0, 100]$.

   (b) Given the limited available information, a burn-in of 5,000 could be adequate. But to make confident statements on convergence and mixing, at least 2 chains should be fitted.

   (c) BGR statistic, denoted by $\hat{R}$ uses multiple chains and compares the within chain variability and between chain variability. If a chain of $2n$ iterations is used it discards the first n iterations as burn-in and only considers the latter n iterations. `OpenBUGS` then calculates,

   $$\hat{R} = \frac{\text{width of 80\% credible interval of pooled chains}}{\text{mean of width of 80\% credible interval of individual chains}}.$$

   Convergence is assumed when $\hat{R} \approx 1$.

   (d) To consider whether the chain has been run long enough - consider Monte Carlo error. These are all at least a magnitude smaller than the corresponding posterior standard deviations; however, if we apply the rule of thumb suggested in the WinBUGS manual by David Spiegelhalter et al.(https://www.mrc-bsu.cam.ac.uk/wp-content/uploads/manual14.pdf) then we should run the chains for longer as MCerror/PostSD is more than 5% for `param1` and `param2`.

   (e)
```
for (t in 2:T) {
Ex[t] <- log(x[t-1]*exp(param1+param2*x[t-1]))
}
for (t in 2:T) {
x[t] ~ dlnorm(Ex[t],param4)
}

ExTplus1<-log(x[T]*exp(param1+param2*x[T]))
xTplus1~ dlnorm(ExTplus1,param4)

param1 ~ dnorm(0,0.1)
param2 ~ dnorm(0,0.1)
param3 ~ dunif(0,100)
param4 <- 1/(param3*param3)
```

```
list(T=10, x = c(10.03,5.87,7.52,6.37,7.77,6.81,7.63,7.15,7.14,8.34))
list(param1=-0.5,param2=0.5,param3=1, xTplus1=5)
```

(f) Here, density dependence refers to the exponential effect of x_{t-1}, given a non-zero θ_1. There appears to be evidence against $\theta_1 = 0$ since the 95% symmetric CI for this parameter does not include 0.

4. (a) The fitted model takes the form of a multinomial likelihood with a log link function. The vector of observed counts $X_{ij.} = X_{ij1}, X_{ij2}, \ldots, X_{ij5}$ has distribution $X_{ij.}|p_{ij.} \sim Multinomial(n_{ij}, p_{ij.})$ where $n_{ij} = \sum_k X_{ijk}$. The multinomial rates are specified as $p_{ijk} = \frac{\phi_{ijk}}{\sum_k \phi_{ijk}}$ where $\phi_{ijk} = \exp(\alpha_k + \beta_{ik} + \gamma_{jk})$.
The priors for each of the $5(\alpha) + 20(\beta) + 10(\gamma) = 35$ parameters are Normal, but the low precision of 0.00001 effectively makes these vague priors.

(b) The trace plots for models A and B suggest that the fitting process for A suffers from slow mixing. Multiple chains have been used when fitting both models and the chains in A have clearly not converged to the same value; their means and variances appear to be different. By running multiple chains the Brooks-Gelman-Rubin (BGR) convergence diagnostic can be used. This compares within-chain and between-chain variances for the selected parameters and, once convergence is reached, these variances should be approximately equal. From the plots, desired behaviour is to see the red line converge towards 1 and the blue and green lines converge towared stability. This does not happen for model A; the red line increases toward the value 2. For model B the red line converges towards 1 after 10000 iterations and the green and blue lines seem to reach a stable value after 35000 iterations. Hence, model B seems to have returned a better fit than model A. It should be noted that inspection of trace plots is not an infallible method for assessing convergence and that the BGR diagnostic can be sensitive to the choice of initial values.
The apparent lack of convergence in model A due to slow mixing can be investigated using several techniques. The fitting algorithm could be run for a much larger number of iterations and, if convergence is eventually assumed to occur, the samples before convergence can be discarded (the "burn-in" period). Alternatively, posterior correlations can be checked to see if parameter correlation is a problem. If so, a re-parameterisation of the model may reduce correlation between parameters and improve the fitting process. Finally, a more informative prior could lead to improved mixing.