STAT 452 Assignment 4

Due 11:59pm Friday 31 May 2024

1. In a study that uses Bayesian methods to forecast the number of species that will be discovered in future years, Edie, Smits and Jablonski (2017) report that the number of marine bivalve species discovered each year from 2010 to 2015 was 64, 13, 33, 18, 30 and 20. Denote Y_t as the number of species discovered in the year $2009 + t$, so eg $Y_1 = 64$ is the count for 2010). Assuming

$$Y_t|\alpha, \beta \sim \text{Poisson}(\lambda_t),$$

where

$$\lambda_t = \exp(\alpha + \beta t), \text{ and } \alpha, \beta \sim \text{Normal}(0, 10^2)$$

Write and implement your own Metropolis-Hastings sampler to approximate the posterior distributions of α and β. Use random-walker samplers for both parameters with proposal distributions:

$$\alpha^*|\alpha^{(s)} \sim \text{Normal}(\alpha^{(s)}, 0.2^2), \text{ and } \beta^*|\beta^{(s)} \sim \text{Normal}(\beta^{(s)}, 0.05^2)$$

Submit commented code and the corresponding acceptance rate for each parameter. Also provide trace plots for each parameter to show the chains have converged.

2. Use the 2016 US Presidential election data to perform Bayesian linear regression with the response variable for county i being the difference between the percentage of the vote for the Republican candidate in 2016 minus 2012 and all variables in the object 'X' as covariates. You'll find the response variable in the object `Y` once you load the file `election_2008_2016.RData`. You may wish to use the following code to standardise and rename the columns in the object `X`:

```
load("election_2008_2016.RData")

X     <- scale(X)   # standardize covariates
X     <- cbind(1,X) # add intercept
short <- c("Intercept", "Pop change", "65+","African American",
           "Hispanic","HS grad","Bachelor's",
           "Homeownership rate","Home value",
           "Median income", "Poverty")
names <- c("Intercept", as.character(names[1:11,2]))
colnames(X) <- short
```

   (a) Fit a Bayesian linear regression model with uninformative Gaussian priors for the regression coefficients and summarize the posterior distribution of all regression coefficients.

   (b) Compute the residuals $R_i = Y_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}$ where $\hat{\boldsymbol{\beta}}$ is the posterior mean of the regression coefficients. Do the residuals follow a normal distribution?

   (c) Include a random effect for the state, that is, for a county in state $l = 1, \ldots, 50$,

$$Y_i|\alpha_l, \boldsymbol{\beta} \sim \text{Normal}(\alpha_l + \mathbf{X}_i\boldsymbol{\beta}, \sigma^2),$$

   where $\alpha_l \sim \text{Normal}(0, \tau^2)$ and τ^2 has an uninformative prior. Use the following code to prepare the 'state' variable and assign a numeric id to each state.

```
state <- as.character(all_dat[,3])
  AKHI  <- state=="AK" | state=="HI" | state=="DC"
  fips  <- fips[!AKHI]
  Y     <- Y[!AKHI]
  X     <- X[!AKHI,]
  state <- state[!AKHI]

  # Assign a numeric id to the counties in each state
  st    <- unique(state)
  id    <- rep(NA,length(Y))
  for(j in 1:48){
    id[state==st[j]]<-j
  }
```

Why might adding random effects be necessary? Which states have the highest and lowest posterior mean random effect, and what might this imply about these states?

3. Download and prepare the 'titanic' dataset from R as follows:

```
library(titanic)
  dat    <- titanic_train
  Y      <- dat[,2]
  age    <- dat[,6]
  gender <- dat[,5]
  class  <- dat[,3]
  X      <- cbind(1,scale(age),
            ifelse(gender=="male",1,0),
            ifelse(class==2,1,0),
            ifelse(class==3,1,0))
  colnames(X) <- c("Intercept","Age","Gender","Class=2","Class=3")
  miss <- is.na(rowSums(X))
  X    <- X[!miss,]
  Y    <- Y[!miss]
```

NOTE: You'll need to run the code below to install the `titanic` package first before running the code above:

```
install.packages("titanic")
```

Let $Y_i = 1$ if passenger i survived and $Y_i = 0$ otherwise. Perform a Bayesian logistic regression of the survival probability onto the passenger's 'age', 'gender' (dummy variable) and 'class' (two dummy variables). Use uninformative priors $\beta_j \sim \text{Normal}(0, 1000)$. Provide trace and density plots for the regression coefficients. Based on the 95% posterior credible intervals are there any covariates that may not be important predictors of survival probability?

4. Download the Boston Housing Data in R from the 'Boston' dataset found in the package 'MASS'. The response is 'medv', the median value of owner-occupied homes, and the other 13 variables are covariates that describe the neighborhood. Use stochastic search variable selection (SSVS) to compute the most likely subset of the 13 covariates to include in the model and the marginal probability that each variable is included in the model. Clearly state the model you fit including all prior distributions.