STAT 597 Bayesian Statistics Homework #3

1. Let be the number of sick days that a person takes due to an illness, and let be the number of months that person has been taking part in a treatment program. Assume a model for this data as - a Poisson regression model with two regression parameters *a* and *b*. We will assume that both of these regression parameters have prior distributions that are Gaussian with mean zero and variance equal to 4. The data can be read in as follows:

x=c(8,14,11,7,32,8,28,21,27,15,26,13,19,22,15,

12,15,7,9,15,26,22,16,12,6)

y=c(5,2,5,4,1,3,0,2,1,2,2,5,3,2,1,2,2,8,5,2,1,1,6,4,3)

* 1. Construct a Metropolis Hasting sampler that jointly proposes (*a,b)* from a bivariate normal distribution with the current values of both parameters as the mean. Begin with a proposal distribution that has a diagonal covariance matrix with 0.01 on both diagonal elements. Use Shaby and Wells’ log-adaptive tuning approach to adaptively tune the proposal distribution, with adaptation happening every 100 MCMC iterations. Run this until you are sure that your algorithm converges to the stationary distribution.

1. Specify a reasonable prior distribution for the following situations.
   1. Your data are Bernoulli distributed, with shared probability of success *p.* Your goal for a prior distribution on is that your prior is vague, giving equal probability for any valid value for .
   2. Your data are Bernoulli distributed, with shared probability of success *p.* Your goal for a prior distribution on is to specify a prior distribution that has a 90% probability that is between 0 and 0.5, with a 10% probability that *p* is greater than 0.5.
   3. Your data are Bernoulli distributed, with shared probability of success *p.* Your goal for a prior distribution on is to specify a prior distribution that has a 50% probability that is exactly 0, and a 50% probability that *p* is somewhere between 0 and 1, with equal probability given.
   4. Your data are classic linear regression data, with response and predictor variables. The regression parameter for one parameter of interest has been well studied in the literature. Specify a prior for this distribution which allows for any real number, but which has a 95% prior probability of being between -0.2 and -0.1.
   5. Your data come from a physical process where you know that your parameter must be between 0 and 2. A previous study estimated the parameter as being very close to 2. Specify a prior that respects the required physical constraints, and also places a 75% probability that the parameter is between 1.8 and 2.
2. Read in the “lagos.Rdata” data, using the code in “lagosHwkBayes.r”. The variables in this dataset are:

tn = “total nitrogen”, a measure of the nitrogen concentration in lake water

tp= “total phosphorous”, a measure of the phosphorous concentration in lake water

secchi = “secchi disk depth”, a measure of water clarity

x = longitude of the lake location

y = latitude of the lake location

ag = 1 if surrounding land is agricultural

forest = 1 if surrounding land is forested

(if both ag and forest are 0, then the surrounding land is something else)

Fit a linear regression model with log(tn) as the response, and with log(tp), log(secchi), ag, forest, x, y, x^2, y^2 as predictor variables. Use either a ridge regression prior or a lasso regression prior on the regression parameters, but use a vague Gaussian prior for the intercept. Clearly write out your model, including all priors. Fit the model using MCMC (you may use any method, including nimble or coding your own sampler). Report your results by giving 95% credible intervals for all parameters. For the parameters associated with x, y, x^2 and y^2, show a 95% credible interval for the function and a 95% credible interval for the function where b,c,d, and e are the regression parameters associated with x, y, x^2 and y^2.

1. Assume that you have the following count data, which are the number of calls to a help line each hour.

y=c(9, 15, 14, 5, 6, 4, 4, 4, 8, 0, 10, 22, 7, 2, 6, 2, 18, 9, 4, 2, 5, 7, 7, 5, 8, 7, 2, 15, 17, 7, 1, 4, 8, 5, 8, 9, 25, 6, 6, 4, 22, 3, 2, 5, 3, 4, 8, 4, 17, 14)

The assumed model is that there are two latent “states”, one with high call rate and one with lower call rate. The assumed model is

Where the gamma distributions are parameterized with “shape” and “rate” parameters, and have mean=10 and variance=100.

* 1. Show that you will need to use MH steps for and , using the above model as written. Implement an MCMC sampler to draw samples from the posterior distribution, and report posterior means and 95% credible intervals of the mean number of calls in the “high” state which has rate (
  2. Now consider a data augmentation approach where we split into two pieces.

Under this new model, show that and have conjugate updates, and implement an MCMC sampler to draw samples from the posterior distribution, and report posterior means and 95% credible intervals of the mean number of calls in the “high” state which has rate (

1. **This problem will NOT be graded for homework #3. It may be put on Homework #4!** Now fit a similar linear regression model with log(tn) as the response, and with ag and forest as standard predictor variables. Specify diffuse Gaussian priors for the regression parameters associated with ag and forest. In addition, include in your linear predictor a flexible function f(log(tp)) and g(log(secchi)) where f(.) and g(.) are flexible functions modeled using penalized B-splines. Create a set of B-spline basis functions that spans all values of log(tp) and a separate set of basis functions that spans all values of log(secchi). For the regression parameters associated with these basis functions, specify a prior which seeks to penalize the sum of the square of the 2nd derivative of the functions, thus penalizing towards a linear function. Report your results by showing an estimated 95% credible interval for each smooth function and providing an interpretation of the estimated relationships between log(secchi) and log(tn), as well as between log(tp) and log(tn).