CHL 5223H Applied Bayesian Methods

Homework 3
Monday, March 4, 2019
DUE: Monday, April 1, 2019
If you are graduating in June 2019,
YOU NEED TO SEND ME AN EMAIL TELLING ME

1. In the file “SmokeHyperHwkQuesR.txt”, there is a preliminary analysis of a dataset. The analysis looks at the relationship between smoking and hypertension. For the population under study, there are a series of binary variables that are recorded. They record the following traits: smoker, obesity, snoring, gender as well as the outcome of interest which is hypertension. A table is created where the number of people with hypertension is recorded as well as the total number in that category. (Aside: there were was no one who was a a smoker and obese who did not snore. So that group is not included in the table.) Also, note in that file, there is a basic analysis of this data.

   Starting with the analysis in this file, run the MCMC with more iterations and see if the chain has “converged”. For the purpose of this excercise, just look at three parameters. Use either “sd.b” or “tau.b” for one of the parameters and use one of the “beta” parameters for the other two. Fit this model with either WinBugs, OpenBugs, or Jags (and provide the model in your answer.) Run the iteration with at least three chains in this model. Do the following:

   (a) Run between 10,000 to 30,000 iterations of the MCMC. For the three parameters, provide a copy of the trace plot for a portion of the iteration, the autocorrelation plot, and the statistics.

   (b) Input the MCMC values into R (but not yet into coda). Plot the trace plot in R. Remove some of the early values of the chain (throwing away a part that is “burned-in”) and then plot the estimate of the densities for the parameters with a different density estimate for each of the three chains. What effect does “burning in” the chain have?

   (c) If you thinned the chain, what would be the advantages? Is it necessary to thin a chain?

   (d) Provide the estimate of the posterior mean of the three parameters for each chain and also give the Monte Carlo accuracy of your estimate. For the Monte Carlo accuracy, compute by batch means and by using the autocorrelation function.

   (e) Using the coda (or boa) package, use the Geweke and Brooks-Gelman-Rubin diagnostic procedures to assess how well the MCMC algorithm has converged.

   (f) Using the information from this question, state if you feel that the MCMC algorithm has converged. Justify your answer.

2. Consider the following data which contains information on harvesting dates versus crop yields. The x variable is data on date of harvesting (which is the number of days after flowering) and yield y (kg/ha) of paddy, a grain farmed in India. (Data from Devore pg 518 and originally from *J. Agricultural Eng. Research*, 1975, pp 353-363.) Here is the data:

```
list( x=c(16,18,20,22,24,26,28,30,32,34,36,38,40,42,44,46),
y=c(2508,2518,3304,3423,3057,3190,3500,3883,3823,3646,3708,
3333,3517,3241,3103,2776))
```

Consider two models for this data. The first model predicts the population as a linear function of the date and the second predicts with a quadratic function. The following WinBugs/OpenBugs code will fit these models (This information is in a txt file that is sent with this file. You can get the the data and model codes from that file and you won't have to type them in):

```
Model 1:

model{
for(i in 1:16){
  y[i]~dnorm(mu[i],tau)
  mu[i]<- b[1] + b[2]*(x[i]-31)
}
b[1]~dnorm(0,.000001)
b[2]~dnorm(0,.000001)
tau~dgamma(.0001,.0001)
}

Model 2:

model{
for(i in 1:16){
  y[i]~dnorm(mu[i],tau)
  mu[i]<- b[1] + b[2]*(x[i]-31)+ b[3]*pow((x[i]-31),2)
}
b[1]~dnorm(0,.000001)
b[2]~dnorm(0,.000001)
b[3]~dnorm(0,.01)
tau~dgamma(.0001,.0001)
}
```

Do the following two parts:

(a) Compare these two models. First, compare these two models by looking at the "deviance" measures and the DIC. Calculate these values for each model and comment on them. Then, compare these two models by calculating the Bayes Factor. To calculate the Bayes factor, run an MCMC algorithm which switches between the two models using a method similar (in which you might have to somewhat change the model code as well as the model) to the method proposed by Kuo and Mallick. Comment on your belief between the two models. (That is, which model do you prefer and justify your answer.)

(b) For model 1, look at the residuals for the model using functions 1, 2, and 3. That is, calculate 1) the residuals, 2) the standardized residuals, and 3) the chance of getting a more extreme observation. For the residual and the standardize residual, please calculate the distribution of these statistics under the predictive distribution. Comment on the results of these statistics for the observation. Also, comment on how well you think the model fits the data.

3. Simulating random variables

Introduction: For this problem, you are required to estimate the mean and variance for a random variable using some of the methods discussed in class. Consider a random variable X with density $g(x)$. For each of the questions below, you are allowed to get random uniform variables using the R command `runif`. Also, you are just allowed to sample a total of 1000 uniform random variables. In each question, you will use a different technique which is described in the notes. For a complete answer, you should not simple provide the numeric value for your calculation. You need to describe why the algorithm you used is the correct algorithm.

Details of the random variable Let X be a sample from the triangle distribution. That is, if $g(x)$ is the density function for X, then define $g(x)$ as:

$$g(x) = \begin{cases} 4x & \text{for } 0.0 \leq x \leq 0.5 \\ 4 - 4x & \text{for } 0.5 < x \leq 1.0 \\ 0 & \text{otherwise} \end{cases}$$

In the plot, this density is the solid line. Note, you could define the function $g(x)$ in R as:

```
g=function(x){(x>0)*(x<1)*((x<=0.5)*4*x+ (x>0.5)*(4-4*x))}.
```

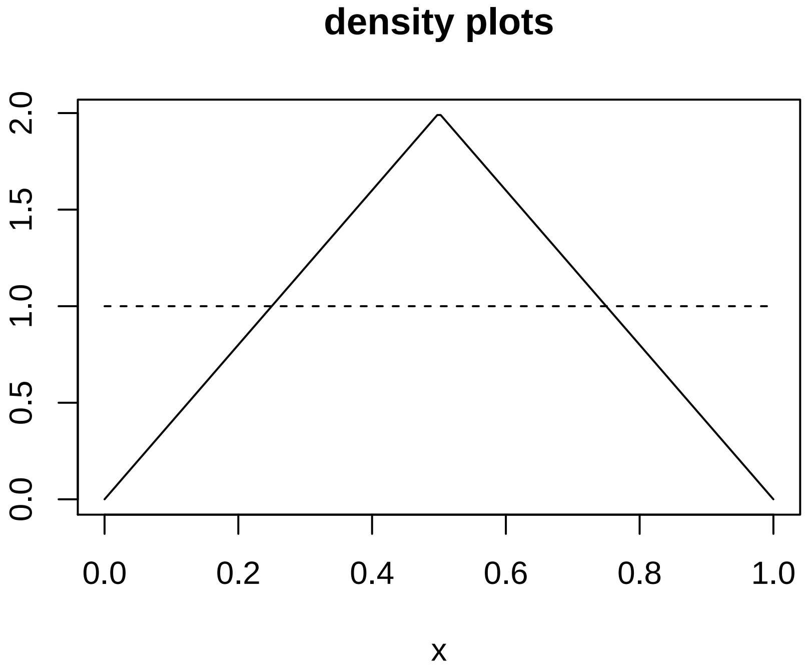

Figure 1: The solid line is the density of the triangle distribution and the dashed line is for the uniform distribution.

Also, let $f(x)$ be the density function for the uniform distribution from 0 to 1. Then, $f(x)$ is defined as:

$$f(x) = \begin{cases} 1 & \text{for } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

In the plot, this density is the dashed line. Note that $2f(x) \geq g(x)$.

Questions Do the following:

(a) Let U1 and U2 be two independent samples from a uniform distribution on $[0, 1]$. Define a sample random variable Z by (U1+U2)/2. Note that Z is then a random sample from the triangle distribution and has density function g. Create a series of independent Z_i's in this manner in order to estimate $\mathrm{E}(X)$ and $\mathrm{Var}(X)$.

(b) Just using sample from a uniform distribution on $[0, 1]$, use the importance sampler method. That is, do the following:

i. Provide the weight function for the importance sampler when using sampled values from the uniform distribution.

ii. Give the estimates for $\mathrm{E}(X)$ and $\mathrm{Var}(X)$. (Note: $\mathrm{Var}(X) = (\mathrm{E}(X^2) - [\mathrm{E}(X)]^2.)$

(c) Using just samples from the uniform distribution, use the acceptance-rejection method to estimate $\mathrm{E}(X)$ and $\mathrm{Var}(X)$. To do this, do the following:

i. Generate a random variable X using the acceptance-rejection method. State how your algorithm works and that the acceptance test function is.

ii. Give the rate of acceptance. That is, what percentage of proposed values is accepted.

iii. Provide your estimates of $\mathrm{E}(X)$ and $\mathrm{Var}(X)$ for this method.

(d) Use the Metropolis-Hasting algorithm to generate an MCMC sequence of X_i's which have the triangle distribution as the invariant and the limiting distribution. Do this by doing the following:

- Let the transition function, $q(x, y)$, be uniform density and have it not depend on the value of x. (That is, the new proposed move is always a sample from the uniform distribution on $[0, 1]$ and this does not depend on the previous location. Therefore, $q(x, y) = 1$ for all values of x and for $y \in [0, 1]$.
- The invariant distribution is the triangle distribution. So, $u(x) = g(x)$.
- Don't burn in the chain and don't thin the chain.

To answer this question provide the following:

i. What is the test function, $\alpha(x, y)$, for this chain given the above information?

ii. Provide the R code which samples the chain.

iii. What is the acceptance rate for the proposed moves in this chain?

iv. What are your estimates of $\mathrm{E}(X)$ and $\mathrm{Var}(X)$?