STAT 206 Homework 8 – Due Wednesday, November 30, 2016, 11:59 PM

General instructions for homework: Homework must be completed as an R Markdown file. Be sure to include your name in the file. Give the commands to answer each question in its own code block, which will also produce plots that will be automatically embedded in the output file. Each answer must be supported by written statements as well as any code used. (Examining your various objects in the "Environment" section of RStudio is insufficient – you must use scripted commands.)

Part I - Metropolis-Hasting algorithm

Suppose $f \sim \Gamma(2, 1)$.

1. Write an independence MH sampler with $g \sim \Gamma(2, \theta)$.
2. What is $R(x_t, X^*)$ for this sampler?
3. Generate 10000 draws from f with $\theta \in \{1/2, 1, 2\}$.
4. Write a random walk MH sampler with $h \sim N(0, \sigma^2)$.
5. What is $R(x_t, X^*)$ for this sampler?
6. Generate 10000 draws from f with $\sigma \in \{.2, 1, 5\}$.
7. In general, do you prefer an independence chain or a random walk MH sampler? Why?
8. Implement the fixed-width stopping rule for you preferred chain.

Part II - Anguilla eel data

Consider the **Anguilla** eel data provided in the `dismo` R package. The data consists of 1,000 observations from a New Zealand survey of site-level presence or absence for the short-finned eel (Anguilla australis). We will use six out of twelve covariates. Five are continuous variables: `SegSumT`, `DSDist`, `USNative`, `DSMaxSlope` and `DSSlope`; one is a categorical variable: `Method`, with five levels `Electric`, `Spo`, `Trap`, `Net` and `Mixture`.

Let x_i be the regression vector of covariates for the ith observation of length k and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_9)$ be the vector regression coefficients. For the ith observation, suppose $Y_i = 1$ denotes presence and $Y_i = 0$ denotes absence of Anguilla australis. Then the Bayesian logistic regression model is given by

$$
\begin{aligned}
Y_i &\sim Bernoulli(p_i)\ , \\
p_i &\sim \frac{\exp(x_i^T \boldsymbol{\beta})}{1 + \exp(x_i^T \boldsymbol{\beta})} \quad \text{and,} \\
\boldsymbol{\beta} &\sim N(\mathbf{0}, \sigma_\beta^2 \mathbf{I}_k)\ ,
\end{aligned}
$$

where $\mathbf{I}_k$ is the $k \times k$ identity matrix. For the analysis, $\sigma_\beta^2 = 100$ was chosen to represent a diffuse prior distribution on $\boldsymbol{\beta}$.

9. Implement an MCMC sampler for the target distribution using the `MCMClogit` function in the `MCMCpack` package.
10. Comment on the mixing properties for your sampler. Include at least one plot in support of your comments.
11. Run your sampler for 100,000 iterations. Estimate the posterior mean along with an 80% Bayesian credible interval for each regression coefficient in the model. Be sure to include uncertainty estimates.
12. Compare your Bayesian estimates to those obtained via maximum likelihood estimation.

Part II - Permutation tests

The Cram'er von Mises statistic estimates the integrated square distance between distributions. It can be computed using the following formula

$$W = \frac{mn}{(m+n)^2}\left[\sum_{i=1}^{n}(F_n(x_i) - G_m(x_i))^2 + \sum_{j=1}^{m}(F_n(y_j) - G_m(y_j))^2\right]$$

where F_n and G_m are the corresponding empirical cdfs.

13. Implement the two sample Cram'er von Mises test for equal distributions as a permutation test. Apply it to the `chickwts` data.
14. How would you implement the bivariate Spearman rank correlation test for independence as a permutation test? The Spearman rank correlation test statistic can be obtained from the function `cor` with `method="spearman"`. Compare the achieved significance level of the permutation test with the p-value reported by `cor.test` on the same samples.