Stat 425: MIDTERM #2, due date: 11/29/18, in class

This is a take-home exam. No collaboration or discussion is permitted on this exam. If you need clarifications or have R programming questions, you can contact me, the instructor, but are not allowed to ask anyone else (including the TA). Please note that clarifications are limited to ambiguities in the wording of questions. This exam is intended to demonstrate your grasp of the material, so no help will be provided.

- Show the details of your work in order to get full credit for correct answers, and partial credit for incorrect answers if you are on the right track.
- Include all code used to generate results.

Please fill your name and sign the honor pledge. With your signature, you pledge that you have not been in communication about this exam with anybody, neither with any other student in the class, nor with any other person, other than the instructor, that the work you are submitting for this exam is completely your own; that you have not allowed any other student to use or borrow portions of your work; that you have complied to the guidelines given to you by the Instructor, namely that you have consulted the textbook, your notes and the material on the course website only, including homeworks, and have not consulted any general on-line resources, including Wikipedia; that you understand that if you violate this honesty pledge, you will be reported for academic dishonesty to the Honor Council.

Your Name:

The pledge is: "I pledge my honor that I have not violated the honor code during this exam and have followed all instructions".

Signature:

PROBLEM 1: Consider the balanced, additive, one-way ANOVA model

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \; i = 1, \ldots, I, \; j = 1, \ldots, J \tag{1}$$

where $\epsilon_{ij} \sim N(0, \sigma^2_\epsilon)$ i.i.d, with $\mu \in R$, $\alpha_i \in R$ and $\sigma^2_\epsilon > 0$. We adopt a prior structure that is the product of independent conjugate priors, wherein μ has a flat prior, $\alpha_i \sim N(0, \sigma^2_\alpha)$. Assume that σ^2_α and σ^2_ϵ are known.

5pt(a) Write down the full conditional distributions for μ and α_i necessary for implementing the Gibbs sampler in this problem.

5pt(b) What is meant by "convergence diagnosis"? Describe some tools you might use to assist in this regard. What might you do to improve a sampler suffering from slow convergence?

5pt(c) Now let $\eta_i = \mu + \alpha_i$ so that η_i "centers" α_i. Then we can consider two possible parametrizations: (i) (μ, α) and (ii) (μ, η). Generate a sample of data from likelihood (1), assuming $I = 5, J = 1$ and $\sigma^2_\epsilon = \sigma^2_\alpha = 1$. Write a program to investigate the sample crosscorrelations and autocorrelations produced by Gibbs samplers operating on parametrizations (i) and (ii) above. Which performs better?

5pt(d) Rerun the program for the case $\sigma^2_\epsilon = 1$ and $\sigma^2_\alpha = 10$. Now which parametrization performs better? What does this suggest about the benefits of hierarchical centering reparametrizations?

PROBLEM #2: Heart and Estrogen/progestin Replacement Study

The Heart and Estrogen/progestin Replacement Study (HERS) was a randomized, double-blind, placebo-controlled trial designed to test the efficacy and safety of hormone therapy (HT) for prevention of recurrent coronary heart disease (CHD) events in women. $2,763$ postmenopausal women less than 79 years old were considered for trial. Eligible participants were assigned with equal probability to either the placebo or hormone therapy group. Here, we are interested in a secondary outcome, i.e. the effect of HT and statin use on low-density lipoprotein (LDL) cholesterol, a risk factor for CHD.

A .csv data file can be downloaded from the course website and contains the following data:

```
[1.] HT: random assignment to HT
[2.] Age: age in years
[3.] smoking: current smoker (1) or not(0)
[4.] drinkany: any current alcohol consumption (yes=1)
[5.] exercise: exercise at least 3 times per week (yes=1)
[6.] statins: statin use (yes=1)
[7.] diabetes: presence of diabetes (yes=1)
[8.] BMI index
[9.] LDL1 year 1 cholesterol levels (mg/dl)
```

Please answer the following questions:

5pt(a) Let's first consider a regression model with all predictors. Discuss the choice between an independent prior $\beta_j \sim N(0, b)$, $j = 1, \ldots, p$, for a given value of $b > 0$ and the so-called g-prior for inference on p regression coefficients in a regression model. What are the pros and cons of each choice? How do they differ?

5pt(b) Fit the model [you might consider standardizing the continuous predictors] and present posterior inference for the regression parameters in tables and/or figures. In particular, quantify the effects and assert if there is statistical relevance and/or practical importance of your findings for the physicians. Based on your analysis, is the use of statins associated with LDL? How about smoking status?

5pt(c) Now consider a 60 years-old woman that is undergoing HT, doesn't smoke nor consume alcohol, does some exercise, doesn't use statins, doesn't have diabetes, and has a BMI of 25.8. What is the median level of LDL1 and a range of values that you can predict for this woman based on that information? How would you change your answer if she used statins instead? How would your answer change if the woman were assigned to the placebo group instead?

5pt(d) Now explicitly consider a g-prior (if you haven't done so already) with 3 values of g: a value that puts equal weight to prior and likelihood, a value that has the equivalent weight of one observation, and the value suggested by Fernandez et al (2001), i.e. $g = \max(n, r^2)$. Which of the choices would you prefer, if any?

Note: if you have used an independent prior instead, for the question above, perform sensitivity with that prior, for different values of the parameters. Motivate your final choice.

5pt(e) Consider a model that (possibly in addition to the other covariates) includes BMI (standardized) and the interaction between BMI and statin use. In R, be careful that statins is a "yes/no" factor (you may get an error if you don't model the interaction appropriately, or you may need to recode your data). Based on your analysis, is BMI associated with LDL? What about the interaction? Please, discuss the statistical and practical relevance of your findings.

5pt(f) Let's now consider the problem of variable selection and the Bayes Factor as a criterion for model selection. What are the main features, advantages and possible disadvantages of this model selection criterion? What rule of thumb has been suggested to determine strength of evidence in favor of one model by using the Bayes Factor?

5pt(g) Consider a model where diabetes is one of the predictors and one where it is not, possibly in addition to the other predictors. Compare the two models using the Bayes Factor.

PROBLEM #3:

The number of occurrences of a rare, nongenetic birth defect in a five-year period of six neighboring counties is

$$\boldsymbol{y} = (1, 3, 2, 12, 1, 1).$$

The counties have populations of

$$\boldsymbol{x} = (33, 14, 27, 90, 12, 17) \quad \text{in thousands.}$$

The second county has higher rates of toxic chemicals present in soil samples, and it is of interest to know if this county has a high disease rate as well.

The following Poisson model and prior distributions are considered:

$$\begin{aligned} y_i|\theta_i, x_i &\sim \text{Poisson}(\theta_i\ x_i) \qquad i = 1, \ldots, n \\ \theta_i|\alpha, \beta &\sim \text{Gamma}(\alpha, \beta) \\ \alpha &\sim \text{Gamma}(1, 1) \\ \beta &\sim \text{Gamma}(10, 1) \end{aligned}$$

5pt(a) Write down the full conditional distribution of the rate for each county, $p(\theta_i|\boldsymbol{\theta}_{-i}, \alpha, \beta, \boldsymbol{x}, \boldsymbol{y})$.

5pt(b) Obtain posterior samples of $(\alpha, \beta, \boldsymbol{\theta})$ using a combined Metropolis-Hastings and Gibbs algorithms by iterating the following steps:

1. given a current value $(\alpha^{(t)}, \beta^{(t)}, \boldsymbol{\theta}^{(t)})$, generate a proposal $(\alpha^*, \beta^*, \boldsymbol{\theta}^{(t)})$ by sampling α^* and β^* from a proposal distribution centered around $\alpha^{(t)}$ and $\beta^{(t)}$. Accept the proposal with the appropriate probability.
2. sample new values of the θ_i's from their full conditional distributions.

Perform appropriate diagnostic tests on your chain and make necessary adjustments.

15(c) Draw posterior inference on the infection rates using the samples from the Markov chain. In particular,

(i) Determine the marginal posterior distributions of $\theta_1, \ldots, \theta_6$ and compare them to $y_1/x_1, \ldots, y_6/x_6$.

(ii) Examine the posterior distribution of α/β and compare it to the corresponding prior distribution, as well as to the average of y_i/x_i across the six counties.

(iii) Plot samples of θ_2 versus θ_i for each $i \neq 2$ and overlay a line of slope 1. Also, estimate $P(\theta_2 > \theta_i | \boldsymbol{x}, \boldsymbol{y})$ for each i and $P(\theta_2 = \max\{\theta_1, \ldots, \theta_6\} | \boldsymbol{x}, \boldsymbol{y})$. Interpret these results and compare them to the conclusions one might obtain by just examining y_i/x_i for each county i.