Statistics 568 - Bayesian Analysis
Final Exam

Due: noon, Wed May 12, 2021, Canvas submission

Note.

- The exam is **open book.** You may consult any study material as you wish. However, you must take the exam **by yourself**, without consulting another living person.
- There are four questions, worth 30 + 50 + 90 + 30 = **200 points** in total. **Read the questions carefully** and follow the instructions.
- Submit your solution to the entire exam as **a single .pdf** document. Clearly label each question and subquestion.
- **Reproducible R source code** should accompany all questions that have a computing component. Include either in markdown format, or as appendix in the end.
- Keep in mind that you will be assessed both on the **execution** and the **communication** of your results. Answer every question as comprehensively as you can, using **full sentences**. Explain your results, decisions, as well as any output (such as figures/tables) which you produce by `R`, whenever applicable.
- You must submit before due time. **Late submissions will not receive credit.**

1. (10+10+10 = 30 pts) A pond is full of two types of fish: the blue fish and the green fish. We want to estimate the number of each type of fish in the pond. On day one, you went and caught 100 fish at random, and tagged them on their tails. Of these 100 fish, 60 were blue and 40 were green. A week later, your friend returned to the pond and caught another 100 fish at random. (That is, assume both times the fish caught constituted simple random samples from the pond.) Your friend noted that 20 of those 100 fish he caught were tagged on their tails. However, he accidentally knocked over the fish bucket and all his fish fell back into the pond, so he did not get a chance to record their colors. Let N_b and N_g be the *total* number of blue fish and green fish in the pond.

   (a) Define the complete data, observed data, as well as all relevant parameters.

   (b) Give a reasonable prior distribution for N_b and N_g.

   (c) Write down the joint posterior distribution.

2. (10+20+10+10 = 50 pts) Consider a model of n independent observations,

$$y_i \sim Gamma(a, bx_i), \quad i = 1, \ldots, n,$$

   with a normal prior on $\log(a)$ with mean $\log 5$ and standard deviation $\log 2$, and a normal prior on $\log(b)$ with mean $\log 0.1$ and standard deviation $\log 10$. (All logs are natural logs.)

   (a) Write down the log posterior distribution. Be sure to appropriately account for any parameter transformations.

   (b) Simulate a synthetic dataset with $n = 100$, the [correction 5/7] log of the x_i's drawn independently from the standard Normal distribution, $a = 2$, and $b = 0.3$. Then, program your own *Markov Chain Monte Carlo* algorithm (i.e. HMC/Metropolis-Hastings or the like – simple Monte Carlo and `Stan` are *not* allowed), to fit the model to this synthetic dataset.

   (c) Demonstrate that your algorithm has approximately converged, using a variety of convergence diagnostics.

   (d) Summarize posterior inference for (a, b), and display a joint scatterplot/contour plot constructed from the posterior samples, overlaying the location of the true parameter values used for simulating the synthetic data. Make sure that your plot is well-positioned, and captures the vast majority of the posterior mass.

3. (10+10+10+20+20+10+10 = 90 pts) The 2006 General Social Survey (GSS) included several questions about social networks, including asking respondents whether they know someone who is gay. In this question, you will model the survey response using an abridged version of the GSS data, available on Canvas as `GSS2006_abridged.csv`. Here are the variables in the dataset:

   - `AGE`: respondent's age;

- `SEX`: 1 if the respondent identifies as male, and 0 if female;
- `RACE_WHITE`: 1 if the respondent identifies as white;
- `RACE_BLACK`: 1 if the respondent identifies as black;
- `DEMOCRAT`: 1 if the respondent is politically affiliated with the Democratic party;
- `REPUBLICAN`: 1 if the respondent is politically affiliated with the Republican party;
- `GAYACQ`: 1 if the respondent is acquainted with at least one gay person, 0 otherwise.

You will explore and build a Bayesian logistic regression model for the percentage of people in the population who believe they know someone gay in 2006, as a function of age, sex (male or female), race (white, black, or other), and political affiliation (Democrat, Republican, or neither).

(a) Using clear, unambiguous notation, write down the full generative Bayesian model, including all your prior specifications.

(b) Explain and justify your choice of priors.

(c) Write down the posterior distribution.

(d) Fit the model in `Stan`, and make sure it has converged.

(e) Conduct an adequate level of posterior predictive checking, to assess the fit of your model.

(f) Graph your posterior estimates (of the percentage of people in the population who believe they know someone gay) along with the data, including some kind of posterior uncertainty quantification in the graph. Interpret your results. You may plot the estimates and the data by age, one figure for every combination of levels of the discrete covariates. Display multiple figures concisely.

(g) Provide a posterior estimate and a 50% credible interval for the percentage of people who believe they know someone gay, among each of the following groups in the population:

i. 30-years-old black men who are Democrats;

ii. 50-years-old white men who are Republicans;

iii. white women over the age of 70 who do not identify with a political party.

4. (10+10+10 = 30 pts) Suppose the data $y = (y_1, \ldots, y_J)$ are independently distributed as normal random variables with unknown mean:

$$y_j \mid \theta_j \sim N\left(\theta_j, \sigma_j^2\right),$$

where the σ_j's are known constants. Consider the following two competing Bayesian models for the data. For some choice of $A > 0$,

H_1: (no pooling) $\theta_j \sim N\left(0, A^2\right)$, i.i.d.;

H_2: (complete pooling) $\theta_1 = \theta_2 = \cdots = \theta_J = \theta$, where $\theta \sim N\left(0, A^2\right)$.

Suppose we want to use the Bayes factor to help decide between the two models.

(a) Derive the marginal distributions of y under either models, $p(y \mid H_1)$ and $p(y \mid H_2)$, and thus the Bayes factor

$$\frac{p(y \mid H_1)}{p(y \mid H_2)},$$

as a function of $y_1, \ldots, y_J$, $\sigma_1, \ldots, \sigma_J$, and A.

(b) Evaluate the Bayes factor in the limit as $A \to \infty$. Which model does the Bayes factor favor?

(c) For fixed A, evaluate the Bayes factor as the sample size J increases, assuming for simplicity that $\sigma_1 = \cdots = \sigma_J = \sigma$, and that the sample mean and variance do not change as a function of J. Discuss the implication of your result.