MATH-640: Bayesian Statistics

Final Exam

Due Wednesday May 11, 2011

INSTRUCTIONS

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. 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.

Please fill your name and sign the honor pledge:

I, _____, pledge that I have not violated the Georgetown University honor code (see http://honorcouncil.georgetown.edu). The work I am submitting for this exam is completely my own. I have not communicated with any other person, beside the instructor, and have not allowed any other student to use or borrow portions of my work. I understand that if I violate this honesty pledge, I will be reported for academic dishonesty to the Honor Council.

Signature : _____

- 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 start each problem on a separate page.
- The exam must be submitted **online** by 11:59 pm on Wednesday, May 11, 2016. You can either upload it on Blackboard or e-mail it to me at `mgt26@georgetown.edu`.

1. [**10 points**]

Let X be a random sample from a Pareto(k, θ) distribution with probability density function

$$f(x|\theta) = \frac{\theta k^\theta}{x^{\theta+1}}, \qquad x > k > 0, \quad \theta > 0.$$

(a) Describe in detail how to draw samples from the Pareto distribution using the inverse transformation method. Provide the cdf, $F(x)$, and the inverse cdf, $F^{-1}(x)$.

(b) Using the general procedure you described in (a), generate random numbers from a Pareto$(2, 3)$ distribution.

(c) Plot a histogram of the sampled values with the kernel density estimator overlayed for x-values in the range $(0, 10)$.

(d) Evaluate $P(X < 3)$ analytically and evaluate a Monte Carlo approximation using the simulated samples in (b).

2. [**25 points**]

The AIDS Clinical Trials Group (ACTG) is the largest HIV clinical trials organization in the world and has been pivotal in providing the data necessary for the approval of therapeutic agents and the development of prevention strategies. The data `actg036.dat` come from a double-blind placebo-controlled clinical trial comparing AZT with placebo and contain information on patients' age, treatment (1 if AZT, 0 if placebo), race (1 if white, 0 otherwise), and CD4 count (y_i).

We would like to model

$$y_i = \beta_1 + \beta_2\ AGE_i + \beta_2\ TRT_i + \beta_3\ RACE_i + \varepsilon_i \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

(a) Using $p(\boldsymbol{\beta}, \sigma^2) \propto 1/(\sigma^2)$

(i) evaluate and plot the marginal posterior distributions of $\beta_2, \beta_3, \beta_4$

(ii) compute 95% HPD intervals for $\beta_2, \beta_3, \beta_4$

(iii) what do you conclude from these analyses?

(b) Data from a previous ACTG clinical trial are available in the file `actg019.dat` and will be used to specify the prior for $\boldsymbol{\beta}$

$$\begin{aligned} \boldsymbol{\beta}|\sigma^2, a_0 &\sim \mathcal{N}_4\left(\hat{\boldsymbol{\beta}}_{\mathbf{0}}, a_0^{-1}\sigma^2(X_0^T X_0)^{-1}\right) \qquad \text{where } \hat{\boldsymbol{\beta}}_{\mathbf{0}} = (X_0^T X_0)^{-1} X_0^T y_0 \\ \sigma^2 &\sim \text{Inv-Gamma}\left(\frac{\nu_0}{2}, \frac{\gamma_0}{2}\right) \end{aligned}$$

y_0 and X_0 correspond to the CD4 counts and covariate matrix from the historical data.

(i) Using $\nu_0 = \gamma_0 = 0.1, a_0 = 0.5$, construct 95% HPD intervals for $\beta_2, \beta_3, \beta_4$.

(ii) Perform a sensitivity analysis to a_0 by constructing 95% HPD intervals for $\beta_2, \beta_3, \beta_4$ using $a_0 = 10^{-6}, 0.2, 0.8, 1.0$ with $\nu_0 = \gamma_0 = 0.1$.

(iii) Perform a sensitivity analysis to (ν_0, γ_0) by constructing 95% HPD intervals for $\beta_2, \beta_3, \beta_4$ using $(\nu_0, \gamma_0) = (0.1, 0.1), (1, 1), (5, 5), (10, 10)$ with $a_0 = 0.5$.

(iv) What do you conclude from these analysis?

(c) Using the historical data prior in (b) with $a_0 = 0.5$ and $(\nu_0, \gamma_0) = (0.1, 0.1)$, provide the predictive distribution of $z = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}$ with covariates $X_f = \begin{pmatrix} 40 & 1 & 1 \\ 35 & 0 & 0 \end{pmatrix}$.

(d) Using the historical data prior in (b) with $a_0 = 0.5$ and $(\nu_0, \gamma_0) = (0.1, 0.1)$ and using a uniform prior on the model space, compute the posterior model probabilities for all possible variable subsets. The model space consist of the $2^3 = 8$ possible variable subsets.

(i) Assess the sensitivity of the posterior model probabilities to the choices of a_0 considered in (b.ii).

(ii) Compare your results with those using a stepwise selection with AIC and BIC.

3. [**15 points**]

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:

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

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

(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. Run the MCMC algorithm long enough so that the effective sample sizes of all parameters are greater than 2,000.

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

(i) Plot the marginal posterior distributions of $\theta_1, \ldots, \theta_6$ and compare their posterior estimates 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.