University of Victoria | Mathematics & Statistics

Statistics 460/560 Final Project

December 3^{rd}, 2020

Instructions: This project consists of two parts. In PART I, you will write the steps to a Gibbs sampler algorithm to sample from the joint posterior in the context of a multiple linear regression model. The assumptions include a particular type of spike-and-slab prior for the coefficients of the model. The spike-and-slab prior is useful for variable selection. In PART II, you will use a dataset that will be emailed to you to perform inference for the coefficients of the linear model. Please follow the steps carefully.

PART I

Consider the multiple linear regression model:

$$y = \boldsymbol{X}\beta + \boldsymbol{E}$$

where y is a vector of size n containing the response variable, $\boldsymbol{X}$ is a matrix of size $n \times J$ of fixed covariates, and β is a vector of size J containing the coefficients that characterize the linear relationship between y and X. Let $\boldsymbol{E}$ be a vector of of size n of random noise terms. We assume $\boldsymbol{E} \sim N_n(0, \Sigma)$, with known $\Sigma = I_n$. Now assume that for each $j = 1, \ldots, J$:

$$\beta_j|\delta_j, \tau, \epsilon \sim \delta_j N(0, \tau^2) + (1 - \delta_j)N(0, \epsilon)$$

$$\delta_j|\pi \sim \text{Bernoulli}(\pi)$$

$$\pi|a_\pi, b_\pi \sim \text{Beta}\left(\frac{a_\pi}{2}, \frac{b_\pi}{2}\right)$$

Let $\theta = (\beta, \delta, \pi)$, then the prior distribution of θ is $p(\theta) = p(\pi|a_\pi, b_\pi)\prod_{j=1}^{J} p(\beta_j|\delta_j, \tau^2, \epsilon)p(\delta_j|\pi)$.

a. Write down $p(\beta_j|\delta_j, \tau^2, \epsilon)$, the prior of β_j, up to a constant of proportionality. **Hint**: start by writing it as the product of two normals with the appropriate exponent for each.

b. Use (a) to find the full conditional distribution of β_j, i.e., $p(\beta_j|\delta_j, \tau^2, \epsilon, y)$.
Hint 1: consider two separate distributions, $p(\beta_j|\delta_j = 0, \tau^2, \epsilon, y)$ and $p(\beta_j|\delta_j = 1, \tau^2, \epsilon, y)$.
Hint 2: If it helps, use the fact that $y_i - \sum_{j=1}^{J} X_{ij}\beta_j = \tilde{y}_i - X_{ij}\beta_j$, where $\tilde{y}_i = y_i - \sum_{l \neq j} X_{il}\beta_l$.

c. Show that the full conditional distribution of δ_j is Bernoulli $\left(\frac{p_1}{p_0+p_1}\right)$ with $p_1 = \pi\exp\{-\frac{1}{2\tau^2}\beta_j^2\}$ and $p_0 = \frac{(1-\pi)\tau}{\sqrt{\epsilon}}\exp\{-\frac{1}{2\epsilon}\beta_j^2\}$.

d. Write down the full conditional distribution of π.

e. Write down a Gibbs sampler algorithm to sample from the joint posterior distribution of θ.

PART II

This part is based on the dataset emailed to you.

f. Let $\epsilon = 10^{-4}$ and $\tau^2 = 10^2$. Explain heuristically how the spike-and-slab prior allows for variable selection in the multiple linear regression model context.

g. Let $a_\pi = b_\pi = 1$, the bathtub prior distribution for π. Using ϵ and τ as in item (f), obtain posterior estimates for the coefficient β.

h. Check for convergence of the MCMC chains using trace plots and compute $\hat{R}$.

i. If the MCMC is converging, present the results including the posterior mean, posterior variance, and a 95% credible interval for each coefficient. Based on these results, which covariates are important to predict the response variable?

j. Sensitivity analysis. Consider four different prior distributions for π by choosing values of a_π and b_π that change the shape of the beta distribution. Plot the prior if π for each of these values. Is the posterior distribution of β sensitive to these new prior distributions?

k. Model checking. Generate 10000 replications of the data y^{rep} using the same x_i as the original data. Compare the posterior mean and median. Based on that, does the model generates predicted results similar to the observed data in the study?