The Chinese University of Hong Kong, Shenzhen
SDS · School of Data Science

· Fall Semester 2022

DDA 4010 – Bayesian Statistics

Exercise Sheet 6

This exercise is due on **Dec 23rd, 5:00 pm.**

Assignment A6.1 (10.5 in Textbook):
Logistic regression variable selection: Consider a logistic regression model for predicting diabetes as a function of $x_1 =$ number of pregnancies, $x_2 =$ blood pressure, $x_3 =$ body mass index, $x_4 =$ diabetes pedigree and $x_5 =$ age. Using the data in `azdiabetes.dat`, center and scale each of the x variables by subtracting the sample average and dividing by the sample standard deviation for each variable. Consider a logistic regression model of the form $\operatorname{Pr}(Y_i = 1 \mid \boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{z}) = e^{\theta_i} / \left(1 + e^{\theta_i}\right)$ where

$$\theta_i = \beta_0 + \beta_1\gamma_1 x_{i,1} + \beta_2\gamma_2 x_{i,2} + \beta_3\gamma_3 x_{i,3} + \beta_4\gamma_4 x_{i,4} + \beta_5\gamma_5 x_{i,5}.$$

In this model, each γ_j is either 0 or 1 , indicating whether or not variable j is a predictor of diabetes. For example, if it were the case that $\gamma = (1,1,0,0,0)$, then $\theta_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2}$. Obtain posterior distributions for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, using independent prior distributions for the parameters, such that $\gamma_j \sim \operatorname{binary}(1/2), \beta_0 \sim \operatorname{normal}(0,16)$ and $\beta_j \sim \operatorname{normal}(0,4)$ for each $j > 0$.

- Implement a Metropolis-Hastings algorithm for approximating the posterior distribution of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. Examine the sequences $\beta_j^{(s)}$ and $\beta_j^{(s)} \times \gamma_j^{(s)}$ for each j and discuss the mixing of the chain.
- Approximate the posterior probability of the top five most frequently occurring values of γ. How good do you think the MCMC estimates of these posterior probabilities are?
- For each j, plot posterior densities and obtain posterior means for $\beta_j\gamma_j$. Also obtain $\operatorname{Pr}(\gamma_j = 1 \mid \boldsymbol{x}, \boldsymbol{y})$.

Assignment A6.2 (11.2 in Textbook):
Randomized block design: Researchers interested in identifying the optimal planting density for a type of perennial grass performed the following randomized experiment: Ten different plots of land were each divided into eight subplots, and planting densities of $2,4,6$ and 8 plants per square meter were randomly assigned to the subplots, so that there are two subplots at each density in each plot. At the end of the growing season the amount of plant matter yield was recorded in metric tons per hectare. These data appear in the file `pdensity.dat`. The researchers want to fit a model like $y = \beta_1 + \beta_2 x + \beta_3 x^2 + \epsilon$, where y is yield and x is planting density, but worry that since soil conditions vary across plots they should allow for some across-plot heterogeneity in this relationship. To accommodate this possibility we will analyze these data using the hierarchical linear model described in Section 11.1. Randomized block design: Researchers interested in identifying the optimal planting density for a type of perennial grass performed the following randomized experiment: Ten different plots of land were each divided into eight subplots, and planting densities of $2,4,6$ and 8 plants per square meter were randomly assigned to the subplots, so that there are two subplots at each density in each plot. At the end of the growing season the

amount of plant matter yield was recorded in metric tons per hectare. These data appear in the file pdensity. dat. The researchers want to fit a model like $y = \beta_1 + \beta_2 x + \beta_3 x^2 + \epsilon$, where y is yield and x is planting density, but worry that since soil conditions vary across plots they should allow for some across-plot heterogeneity in this relationship. To accommodate this possibility we will analyze these data using the hierarchical linear model described in Section 11.1.

- Before we do a Bayesian analysis we will get some ad hoc estimates of these parameters via least squares regression. Fit the model $y = \beta_1 + \beta_2 x + \beta_3 x^2 + \epsilon$ using OLS for each group, and make a plot showing the heterogeneity of the least squares regression lines. From the least squares coefficients find ad hoc estimates of $\boldsymbol{\theta}$ and Σ. Also obtain an estimate of σ^2 by combining the information from the residuals across the groups.
- Now we will perform an analysis of the data using the following distributions as prior distributions:

$$\begin{aligned} \Sigma^{-1} &\sim \text{ Wishart } \left(4, \hat{\Sigma}^{-1}\right) \\ \boldsymbol{\theta} &\sim \text{ multivariate normal } (\hat{\boldsymbol{\theta}}, \hat{\Sigma}) \\ \sigma^2 &\sim \text{ inverse } - \text{gamma} \left(1, \hat{\sigma}^2\right) \end{aligned}$$

  where $\hat{\boldsymbol{\theta}}, \hat{\Sigma}, \hat{\sigma}^2$ are the estimates you obtained in a). Note that this analysis is not combining prior information with information from the data, as the "prior" distribution is based on the observed data. However, such an analysis can be roughly interpreted as the Bayesian analysis of an individual who has weak but unbiased prior information.
- Use a Gibbs sampler to approximate posterior expectations of $\boldsymbol{\beta}$ for each group j, and plot the resulting regression lines. Compare to the regression lines in a) above and describe why you see any differences between the two sets of regression lines.
- From your posterior samples, plot marginal posterior and prior densities of $\boldsymbol{\theta}$ and the elements of Σ. Discuss the evidence that the slopes or intercepts vary across groups.
- Suppose we want to identify the planting density that maximizes average yield over a random sample of plots. Find the value $x_{\max}$ of x that maximizes expected yield, and provide a 95% posterior predictive interval for the yield of a randomly sampled plot having planting density $x_{\max}$.

Sheet 6 is due on **Dec. 23rd**. Submit your solutions before **Dec. 23rd**, **5:00 pm**.