Homework 6- Stat 5392

The data for this homework consist of pairs $\{(x_i, y_i)\}_{i=1}^n$ given in the file `data4hw6` on the course Blackboard shell. The model for the ith observation is a mixture of two linear regressions, as follows.

$$p(y_i \mid x_i, p_i, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1^2, \sigma_2^2) = p_i \frac{1}{\sigma_1}\phi\Big(\frac{y_i - \boldsymbol{\beta}_1' \boldsymbol{x}_i}{\sigma_1}\Big) + (1 - p_i)\frac{1}{\sigma_2}\phi\Big(\frac{y_i - \boldsymbol{\beta}_2' \boldsymbol{x}_i}{\sigma_2}\Big),$$

where ϕ is the standard normal pdf, $\boldsymbol{x}_i = (1, x_i)'$ and

$$p_i = \frac{\exp(\boldsymbol{\delta}' \boldsymbol{x}_i)}{1 + \exp(\boldsymbol{\delta}' \boldsymbol{x}_i)}.$$

Note that p_i is a function of $\boldsymbol{x}_i$ and of an unknown parameter vector $\boldsymbol{\delta} = (\delta_0, \delta_1)'$.

To simplify the computation, let

$$z_i = \begin{cases} 1 & \text{if observation } i \text{ came from component } 1 \\ 0 & \text{otherwise,} \end{cases}$$

for $i = 1, \ldots, n$.

Prior distributions:

- $\sigma_k^2 \overset{\text{ind}}{\sim} IG(\alpha, \lambda)$, $k = 1, 2$, where α and λ are fixed numbers.
- $\boldsymbol{\beta}_k \overset{\text{ind}}{\sim} N\Big(\mathbf{0}, n_k \sigma_k^2 (X_k' X_k)^{-1}\Big)$, $k = 1, 2$, where X_k is the design matrix for the observations allocated to mixture component k, and $n_1 = \sum_{i=1}^n z_i$, $n_2 = n - n_1$. Note that X_k and n_k will change from one iteration to another.
- $\boldsymbol{\delta} \sim N(\mathbf{0}, \sigma_\delta^2 I_2)$, where σ_δ^2 is a fixed large number.

1. Derive the conditional posterior distributions necessary for Gibbs sampling. Note that the augmented likelihood can be expressed as

$$\prod_{i=1}^n \Big[p_i \frac{1}{\sigma_1}\phi\Big(\frac{y_i - \boldsymbol{\beta}_1' \boldsymbol{x}_i}{\sigma_1}\Big)\Big]^{z_i} \Big[(1 - p_i)\frac{1}{\sigma_2}\phi\Big(\frac{y_i - \boldsymbol{\beta}_2' \boldsymbol{x}_i}{\sigma_2}\Big)\Big]^{1 - z_i}.$$

2. Implement the sampling scheme, using a total of 10,000 iterations with 2000 as burn-in.

3. Make trace plots of the parameters.

4. Plot the data along with a fitted curve. To fit the curve, compute $p_i \boldsymbol{x}_i' \boldsymbol{\beta}_1 + (1 - p_i) \boldsymbol{x}_i' \boldsymbol{\beta}_2$ for $i = 1, \ldots, n$ in each iteration of the Gibbs sampler, and then take the mean across the iterations after a burn-in period.

5. Compute 95% pointwise credible intervals for the regression function by using the appropriate percentiles of the $p_i \boldsymbol{x}_i' \boldsymbol{\beta}_1 + (1 - p_i) \boldsymbol{x}_i' \boldsymbol{\beta}_2$ from the iterations after burn-in. Add the credible intervals to the plot from part 4.