Introduction to Bayesian Data Analysis

Tutorial 6

(1) Problem 6.1 (Hoff) Poisson population comparisons: The data files `menchild30bach.dat` and `menchild30nobach.dat` contain data on the number of children born to men in their 30s with and without bachelor's degrees respectively. Let the average number of children in each group be θ_A and θ_B respectively. Assume Poisson sampling models for the two groups and let $\theta_A = \theta$ and $\theta_B = \theta \times \gamma$. In this parameterization, γ represents the relative rate θ_B/θ_A. Let $\theta \sim \text{gamma}(a_\theta, b_\theta)$ and let $\gamma \sim \text{gamma}(a_\gamma, b_\gamma)$.

(a) Are θ_A and θ_B independent or dependent under this prior distribution? In what situations is such a joint prior distribution justified?

(b) Obtain the form of the full conditional distribution of θ given $\mathbf{y}_A$, $\mathbf{y}_B$ and γ.

(c) Obtain the form of the full conditional distribution of γ given $\mathbf{y}_A$, $\mathbf{y}_B$ and θ.

(d) Set $a_\theta = 2$ and $b_\theta = 1$. Let $a_\gamma = b_\gamma \in \{8, 16, 32, 64, 128\}$. For each of these five values, run a Gibbs sampler of at least 5,000 iterations and obtain $E[\theta_B - \theta_A | \mathbf{y}_A, \mathbf{y}_B]$. Describe the effects of the prior distribution for γ on the results.

(2) Problem 6.2 (Hoff) Mixture model: The file `glucose.dat` contains the plasma glucose concentration of 532 females from a study on diabetes.

(a) Make a histogram or kernel density estimate of the data. Describe how this empirical distribution deviates from the shape of a normal distribution.

(b) Consider the following mixture model for these data: For each study participant there is an unobserved group membership variable X_i which is equal to 1 or 2 with probability p and $1 - p$. If $X_i = 1$ then $Y_i \sim \text{normal}(\theta_1, \sigma_1^2)$, and if $X_i = 2$ then $Y_i \sim \text{normal}(\theta_2, \sigma_2^2)$. Let $p \sim \text{beta}(a, b)$,

$\theta_j \sim \text{normal}(\mu_0, \tau_0^2)$ and $1/\sigma_j \sim \text{gamma}(\nu_0/2, \nu_0\sigma_0^2/2)$ for both $j = 1$ and $j = 2$. Obtain the full conditional distributions of $X_1, ..., X_n$, p, θ_1, θ_2, σ_1^2 and σ_2^2.

(c) Setting $a = b = 1$, $\mu_0 = 120$, $\tau_0^2 = 200$, $\sigma_0^2 = 1000$ and $\nu_0 = 10$, implement the Gibbs sampler for at least 10,000 iterations. Let $\theta_{(1)}^{(s)} = \min\left\{\theta_1^{(s)}, \theta_2^{(s)}\right\}$ and $\theta_{(2)}^{(s)} = \max\left\{\theta_1^{(s)}, \theta_2^{(s)}\right\}$. Compute and plot the autocorrelation functions of $\theta_{(1)}^{(s)}$ and $\theta_{(2)}^{(s)}$, as well as their effective sample sizes.

(d) For each iteration s of the Gibbs sampler, sample a value $x \sim \text{binary}(p^{(s)})$, then sample $\tilde{Y}^{(s)} \sim \text{normal}(\theta_x^{(s)}, \sigma_x^{2(s)})$. Plot a histogram or kernel density estimate for the empirical distritribution of $\tilde{Y}^{(1)}, ..., \tilde{Y}^{(S)}$, and compare to the distribution in part a). Discuss the adequacy of this two-component mixture model for the glucose data.