STK 4021 - Applied Bayesian Analysis and Numerical Methods
Thordis L. Thorarinsdottir
Fall 2014

Problem Set 3

Problem 1 (mixture model). A population of 532 women living near Phoenix, Arizona were tested for diabetes. Other information was gathered from these women at the time of testing, including the plasma glucose concentration. This data is available in the file `glucose.dat` that can be found on the course website.

(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 $Y_i \sim \mathcal{N}(\theta_1, \sigma_1^2)$, and if $X_i = 2$ then $Y_i \sim \mathcal{N}(\theta_2, \sigma_2^2)$. Let $p \sim \text{Beta}(a, b), \theta_j \sim \mathcal{N}(\mu_0, \tau_0^2)$ and $1/\sigma_j \sim \Gamma(\nu_0/2, \nu_0\sigma_0^2/2)$ for both $j = 1$ and $j = 2$. Obtain full conditional distributions of $(X_1, \ldots, X_n), p, \theta_1, \theta_2, \sigma_1^2$, 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 $\theta_{(1)}^{(k)} = \min\{\theta_1^{(k)}, \theta_2^{(k)}\}$ and $\theta_{(2)}^{(k)} = \max\{\theta_1^{(k)}, \theta_2^{(k)}\}$. Using the `R` function `acf`, compute and plot the autocorrelation functions of $\theta_{(1)}^{(k)}$ and $\theta_{(2)}^{(k)}$. Discuss the mixing of the Markov chain.

(d) For each iteration k of the Gibbs sampler, sample a value $x \sim \text{Bernoulli}(p^{(k)})$, then value $\tilde{Y}^{(k)} \sim \mathcal{N}(\theta_{x+1}^{(k)}, \sigma_{x+1}^{2(k)})$. Plot a histogram or kernel density estimate for the empirical distribution of $\tilde{Y}^{(1)}, \ldots, \tilde{Y}^{(K)}$, and compare to the distribution in part (a). Discuss the adequacy of this two-component mixture model for the glucose data.

Problem 2 (probit regression). A panel study followed 25 married couples over a period of five years. One item of interest is the relationship between divorce reates and the various characteristics of the couples. For example, the researchers would like to model the probability of divorce as a function of age differential, recorded as the man's age minus the women's age. The data can found in the file `divorce.dat` which is available on the course website. We will model these data with probit regression, in which a binary variable Y_i is described in terms of an explanatory variable x_i via the following latent variable model:

$$Z_i = \beta x_i + \varepsilon_i$$
$$Y_i = \mathbb{1}\{Z_i \in (c, \infty)\},$$

where β and c are unkown coefficients and $\varepsilon_1, \ldots, \varepsilon_n \sim \mathcal{N}(0, 1)$.

(a) Assuming $\beta \sim \mathcal{N}(0, \tau_\beta^2)$ obtain the full conditional distribution $p(\beta|\mathbf{y}, \mathbf{x}, \mathbf{z}, c)$.

(b) Assuming $c \sim \mathcal{N}(0, \tau_c^2)$, show that $p(c|\mathbf{y}, \mathbf{x}, \mathbf{z}, \beta)$ is a constrained normal density, i.e. proportional to a normal density but contrained to lie in an interval. Similarly, show that $p(z_i|\mathbf{y}, \mathbf{x}, \mathbf{z}_{-i}, \beta, c)$ is proportional to a normal density but constrained to be either above c or below c, depending on y_i.

(c) Letting $\tau_\beta^2 = \tau_c^2 = 16$, implement a Gibbs sampling scheme that approximates the joint posterior distribution of $\mathbf{Z}, \beta$, and c. Compute the autocorrelation function of the parameters and discuss the mixing of the Markov chain.

(d) Obtain 95% posterior confidence interval for β, as well as $\mathbb{P}(\beta > 0 | \mathbf{y}, \mathbf{x})$.

Solutions will be discussed in class on September 19.