Homework 1 (Due on Sept 16)

September 4, 2014 STA 5934

Problem I

Consider the binary regression model $P(y_i = 1 \mid x_i, \beta) = \Phi(x_i'\beta)$, $i = 1, \ldots, n$ where y_i's are binary random variables, x_i's are p-dimensional covariates and β is a p-dimensional coefficient vector. Introduce the auxiliary variable $z_i \sim \mathrm{N}(x_i'\beta, 1)$ and set $y_i = I(z_i > 0)$. Assume $\mathbf{y} = (y_1, \ldots, y_n)'$, $\mathbf{z} = (z_1, \ldots, z_n)'$, X is the $n \times p$ covariate matrix. Consider a point mass prior on β

$$\pi(\beta) = \prod_{j=1}^{p} \{\delta_0(\beta_j) p_{0j} + (1 - p_{0j}) N(\beta_j; 0, c_j^2)\}$$

where p_{0j} is the prior probability of excluding the j-th predictor by setting its coefficient to 0. Show that the conditional posterior of β_j , for $j = 1, \ldots, p$, is given by

$$\pi(\beta_j \mid \beta_{-j}, \mathbf{z}, \mathbf{y}, X) = \hat{p}_j \delta_0(\beta_j) + (1 - \hat{p}_j)\mathrm{N}(\beta_j; E_j, V_j)$$

where $V_j = (c_j^{-2} + X_j'X_j)^{-1}, E_j = V_j X_j'(\mathbf{z} - X_{-j}\beta_{-j}), X_j = j$ th column of X, $X_{-j} = X$ with j th column excluded, $\beta_{-j} = \beta$ with j th element excluded, and

$$\hat{p}_j = \frac{p_{0j}}{p_{0j} + (1 - p_{0j}) \frac{N0;0,c_j^2)}{N(0; E_j, V_j)}}$$

is the conditional probability of $\beta_j = 0$. Here $N(x; \mu, \sigma^2)$ denotes the normal density with mean μ, variance σ^2 evaluated at x.

Problem II

Simulate data from a binary regression model with $p = 7$ (including the intercept), $n = 100$, $x_{ij} \sim \mathrm{U}(0, 1)$ and the intercept and the slope for x_{i2} as the only non-zero coefficients. Using a point-mass mixture prior as in Problem I with $c_j = 1$ and $p_{0j} = 1/2$, run a data augmentation Gibbs sampler for 5,000 iterations after discarding the first 2,000 as burn-in. Summarize posterior mean, median, credible interval and exclusion probabilities of the parameters and the top 10 highest posterior probability models. Calculate the percentage of the visited models. Increase p to 200 with only the first two active predictors (including intercept) and repeat the analysis.