Biostat 682 Homework 4

Due: Tuesday, November 14th, 2017 (in class)

1. Denote by $N_+(\mu, \mu^-, \sigma^2)$ the truncated normal distribution with left truncation point μ^-, i.e. the distribution with density

$$f(x \mid \mu, \mu^-, \sigma^2) = \frac{\exp(-(x-\mu)^2/2\sigma^2)}{\sqrt{2\pi}\sigma[1-\Phi((\mu^- - \mu)/\sigma)]} I[x \geq \mu^-]$$

   (a) Using the classical CDF inversion technique, design and implement an algorithm to simulate the truncated normal distribution

   (b) Let

$$g(x \mid \alpha, \mu^-) = \alpha \exp(-\alpha(x - \mu^-)) I[x \geq \mu^-].$$

      i. Show that there is a constant $M(\alpha, \mu^-)$, such that

$$f(x \mid \mu, \mu^-, \sigma^2) \leq M(\alpha, \mu^-) g(x \mid \alpha, \mu^-).$$

      ii. Using the accept-reject method, design and implement an algorithm to simulate the truncated normal distribution.

      iii. Derive the closed form of the acceptance probability in your designed algorithm and provide the numerical justification of your results.

      iv. Find the best choice of α by maximizing the acceptance probability. Verify your results by numerical experiments.

   (c) Perform a simulation study to compare the two algorithms that you developed in Part 1 and Part 2. For the comparison, we mainly focus on the computational time and the accuracy of the density estimations.

2. Consider a finite Normal mixture model with K components, where K is fixed and pre-specified. Suppose the data are $y_1, \ldots, y_n$ then for $i = 1, 2, \ldots, n$,

$$y_i \mid \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\sigma}^{-2} \sim \sum_{k=1}^{K} \lambda_k \mathrm{N}(\mu_k, \sigma_k^2), \tag{1}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)^T$, $\boldsymbol{\sigma}^{-2} = (\sigma_1^{-2}, \ldots, \sigma_K^{-2})^T$ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)^T$. For priors and hyperpriors, we assume that

$$\begin{aligned}
\mu_k &\overset{\text{i.i.d.}}{\sim} \mathrm{U}[y_{\min}, y_{\max}], \\
\sigma_k^{-2} \mid \beta &\overset{\text{i.i.d.}}{\sim} \mathrm{G}(\alpha, \beta), \\
\beta &\sim \mathrm{G}(a, b), \\
\boldsymbol{\lambda} &\sim \mathrm{Dirichlet}(\theta_1, \ldots, \theta_K),
\end{aligned}$$

where $y_{\min} = \min\{y_1, \ldots, y_n\}$ and $y_{\max} = \max\{y_1, \ldots, y_n\}$; and α, a, b and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)^T$ are fixed and pre-specified hyperparameters.

(a) Suppose $K = 2$. Using the Laplace method, design and implement an algorithm to estimate the marginal posterior distribution of μ_1, μ_2, σ_1^{-2}, σ_2^{-2} and λ_1.

(b) Suppose $K \geq 2$. Without introducing any auxiliary variables, design and implement an MCMC algorithm to simulate the joint posterior distribution of $\boldsymbol{\mu}$, $\boldsymbol{\sigma}^{-2}$ and $\boldsymbol{\lambda}$.

(c) Suppose $K \geq 2$. For each y_i, we introduce a latent indicator $z_i \in \{1, \ldots, K\}$; and assume that

$$\begin{array}{rcl} [y_i \mid z_i = k, \mu_k, \sigma_k^2\ ] & \overset{\text{i.i.d.}}{\sim} & \mathrm{N}(\mu_k, \sigma_k^2) \\ \Pr(z_i = k) & = & \lambda_k \end{array} \qquad (2)$$

i. Show that (1) and (2) are equivalent.

ii. Design and implement an MCMC algorithm to simulate the joint posterior distribution of $\boldsymbol{\mu}$, $\boldsymbol{\sigma}^{-2}$ and $\boldsymbol{\lambda}$.

(d) Perform a simulation study to compare the three algorithms that you developed in Part 1, Part 2 and Part 3, when $K = 2$; and compare the two MCMC algorithms when $K = 10$. For the comparison, we mainly focus on the computational time and the accuracy of parameter estimations.

3. Consider a linear regression model.

$$\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \sigma^2 \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}),$$

where $\mathbf{y} = (y_1, \ldots, y_n)^T$ is an $n \times 1$ vector and $\mathbf{X} = (x_{ij})$ is an $n \times p$ matrix. The parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ and σ^2 are of our primary interests. We consider the following prior specifications

$$\begin{array}{rcl} [\beta_j \mid \sigma^2, z_j = k\ ] & \overset{\text{i.i.d.}}{\sim} & \mathrm{N}(0, \sigma^2 \tau_k^2), \quad \text{for } k = 0, 1, \\ \Pr(z_j = 1) & = & \pi, \\ \sigma^{-2} & \sim & \mathrm{G}(\alpha_1, \alpha_2), \end{array}$$

where τ_0^2, τ_1^2, α_1, α_2 and π are pre-determined.

(a) Design and implement a Gibbs sampler to simulate marginal posterior distribution of $\boldsymbol{\beta}$ and σ^2.

(b) Perform simulation studies to evaluate the performance of the proposed algorithm. In particular, simulate data with the following settings: $\mathbf{X}$ are generated from the multivariate normal distribution with zero mean and compound symmetric covariance variance $0.75\mathbf{I}_p + 0.25\mathbf{1}_p\mathbf{1}_p^T$. The true parameters are set as follows

$$\boldsymbol{\beta} = (0.6, 1.2, 1.8, 2.4, 3.0, \underbrace{0, 0, \ldots, 0}_{p-5})^T, \qquad \sigma^2 = 1$$

The suggested hyper parameter specifications are

$$\tau_0^2 = \frac{1}{10n}, \quad \tau_1^2 = \log(n), \qquad \pi = 0.5$$

Please consider three different cases $(n, p) = (100, 100), (100, 200)$ and $(200, 500)$.