STATS270/370 - Bayesian Statistics

Fall 2022, Final Project

Due date: Dec 8, 2022 at 11:59 pm

Note: You may use any programming language of your choice, but you cannot use any package designed specifically for Markov Chain Monte Carlo sampling algorithms. You may use existing routines for generating random variables or running optimization problems.

Suppose $(y_1, y_2, ..., y_n)$ are gene expression measurements for two genes on n samples, where $y_i = (y_{i1}, y_{i2})$ represent the two gene expressions for sample i. There are four different groups of samples: each group-1 sample contains cells of cell type A (say adipose cells), each group 2 sample contains cells of cell type B (say blood vessel cells), each group 3 sample contain cells from tissue type C which is a 50%-50% mixture of type A and type B cells, each group 4 sample contain cells from tissue type D which is also a mixture of type A and type B cells but the mixing proportion τ is unknown. The group labels for all samples are known and are denoted by $(t_1, t_2, ..., t_n)$. For example, $t_9 = 2$ if sample 9 is in group 2.

We assume that for each gene the mean expression (but not the variance) depends on the cell type, and that, given the cell type, the expressions of different genes are independently normally distributed. Specifically,

- $Y_i \sim N(\mu, \sigma^2\mathbf{I})$ if sample i is from group 1 (μ is a 2-vector, $\mathbf{I}$ is the 2 by 2 identity matrix)
- $Y_i \sim N(\gamma, \sigma^2\mathbf{I})$ if sample i is from group 2
- $Y_i \sim N(0.5\mu + 0.5\gamma, \sigma^2\mathbf{I})$ if sample i is from group 3
- $Y_i \sim N(\tau\mu + (1-\tau)\gamma, \sigma^2\mathbf{I})$ if sample i is from group 4

This model has a 6-dimensional parameter $\theta = (\sigma^2, \tau, \mu_1, \mu_2, \gamma_1, \gamma_2)$.

In the data set for this project, we have n=24 samples with 4 samples in each of groups 1 and 2, and 8 samples in each of groups 3 and 4 respectively. You are asked to implement and compare the following approaches to generate samples from the posterior distribution $p(\theta|y_1, ..., y_n)$, where our prior for τ will be Uniform on $[0, 1]$, our prior for μ, γ will be the (improper) uniform prior on the real line, and our prior for σ^2 will be proportional to $1/\sigma^2$.

1. Metropolis-Hasting
2. Hamiltonian Monte Carlo
3. Gibbs sampling
4. Importance sampling (hint: try 2-stage importance sampling)

Since there is considerable freedom in the design of each algorithm, you should discuss and justify your design choices. For example, how do you choose your proposal distributions in importance sampling and in Metropolis-Hasting? and why do you partition the parameters into a certain way if you use the blocked version of Gibbs sampling? Finally, please also discuss their scalabilities when the sample sizes increase.