

# Exercises for module 11

## A mixture model and slice sampling

### Exercise 1: Artificial mixture data

1. Read in the dataset `simmix.csv` from the website. It is a `data.frame` which contains 500 observations of a variable `x` which was artificially generated.
2. For  $k = 1, 2, \dots$ , let  $\lambda = (\lambda_1, \dots, \lambda_k)$  and  $\mu = (\mu_1, \dots, \mu_k)$  and consider a  $k$ -component normal mixture density

$$\pi(y_i|\lambda, \theta) = \sum_{j=1}^k \lambda_j \pi_j(y_i|\mu_j)$$

where  $\lambda \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_k)$ ,  $\pi_j(y_i|\mu_j) \sim \mathcal{N}(\mu_j, 1)$  and  $\mu_j \sim \mathcal{N}(\mu_{j,0}, \tau_{j,0})$ . For any given values of  $k$  and  $\alpha_j, \mu_{j,0}, \tau_{j,0}$  with  $j = 1, \dots, k$ , write a code for a Gibbs sampler which simulates from the posterior density  $\pi(\lambda, \mu, z|y)$ , where using the notation from the lecture,  $z = (z_1, \dots, z_{500})$  is the vector of dummy variables.

3. With  $k = 4$ , discuss
  - how you would specify the values of  $\alpha_j, \mu_{j,0}, \tau_{j,0}$  with  $j = 1, 2, 3, 4$ ,
  - results obtained by a Bayesian analysis using the Gibbs sampler.
4. There is no simple way of telling what the “correct” number of mixture components is. One suggestion is to assume a maximum number of components  $H$  and let  $k = H$  and  $\alpha_1 = \dots = \alpha_k = 1/H$ .
  - For instance, then one may study the posterior distribution of the means  $(\mu_1, \dots, \mu_k)$ ; what would you conclude if some of the means tend to be close to each other?
  - Apply the approach for the 500 simulated data points when  $k = H = 4$ .