SDS 383C: Statistical Modeling I, Fall 2021
Homework 6, Due Dec 09, 12:00 Noon

Instructor: Abhra Sarkar (abhra.sarkar@utexas.edu)
Teaching Assistant: Angela Ting (angelating@utexas.edu)
Department of Statistics and Data Sciences
The University of Texas at Austin
2317 Speedway D9800, Austin, TX 78712-1823, USA

All homework must be submitted typed-in as a single pdf file. Name the file "firstname-lastname-SDS383C-HW-6.pdf" Submit this file without compression such as zip or rar. Figures accompanying the solutions must be presented close to the actual solution. Computer codes will be rarely evaluated but must still be submitted separately from the main file. Codes must be commented properly and should run easily on other machines. Precise, concise, clear, innovative solutions may be rewarded with bonus points. Explain your answer with logic reasoning and/or mathematical proofs. Organize your solutions in the same order as they were presented. If you can solve a problem using multiple techniques, present only your best solution.

1. (10 points) Consider the 'WoodPewee.csv' dataset provided with the HW set. The dataset descsribes the morning song of a north American songbird recorded as a sequence of syllables.

   Fit a first order Markov model to this data set. Specifically, find out the MLEs of the transition probability matrix and the stationary distribution.

2. (20 points) The 'faithful' dataset from package 'datasets' in R gives eruption and waiting times of the old faithful geyser in Yellowstone national park (export this dataset from R if you are using a different programming language).

   Using Gibbs sampling, fit a Bayesian HMM with location-scale mixture of normals as emission densities to the waiting times

$$p(z_t = k \mid z_{t-1} = j) = \pi_{j,k}$$
$$p(y_t \mid z_t = k) = \text{Normal}(y_t \mid \mu_k, \sigma_k^2)$$

   with $K = 2, 3, 4, 5$ components. Use the priors

$$\boldsymbol{\pi}_j \sim \text{Dir}(1/K, \ldots, 1/K), \quad \mu_k \sim \text{Normal}(\mu_0, \sigma_0^2), \quad \sigma_k^2 \sim \text{Inv-Ga}\,(a_0, b_0)$$

   with appropriately chosen prior hyper-parameters. Provide a brief general description of your algorithm, explicitly listing the full conditionals. Summarize your results by showing the posterior means superimposed over a plot of the data points, and the estimated one-step ahead predictive density (posterior mean and 90% point-wise credible intervals) super imposed over a histogram of the data.

3. **(Bonus Problem)** (40 points) Consider the 'coriel1.csv' dataset supplied with the HW.

   Using Gibbs sampling, fit a Bayesian HMM with location mixture of normals as emission densities to this dataset first (a) using the entire data set with $T = 2271$ data points, and

then (b) using the first $T = 2270$ data points, excluding the last data point. Specifically, fit the following model

$$p(z_t = k \mid z_{t-1} = j) = \pi_{j,k}$$
$$p(y_t \mid z_t = k) = \text{Normal}(y_t \mid \mu_k, \sigma^2)$$

with $K = 3$ components. Use the priors

$$\boldsymbol{\pi}_j \sim \text{Dir}(1/K, \ldots, 1/K),$$
$$\mu_1 \sim \text{Normal}(-0.5, 1/6), \quad \mu_2 \sim \text{Normal}(0, 10^{-6}), \quad \mu_3 \sim \text{Normal}(0.5, 1/6),$$
$$\sigma^2 \sim \text{Inv-Ga}\,(1, 1)\,.$$

Provide a brief general description of your algorithm, explicitly listing the full conditionals. In each case, summarize your results by showing the posterior means superimposed over a plot of the data points, and the estimated one-step ahead predictive density (posterior mean and 90% point-wise credible intervals) super imposed over a histogram of the data.