Biostat 285, HW 2

Problem 1

Simulation of Dirichlet process prior realizations

1. Consider a DP (α, G_0) prior over the space of distributions (equivalently c.d.f.s) G on $\mathbb{R}$, with $G_0 = \text{N}(0,1)$.

   Use Sethuraman's constructive definition to generate (multiple) prior realizations from the $\text{DP}(\alpha, \text{N}(0,1))$ for α fixed at these values $\alpha = 0.1, 0.5, 1, 10$.

   Read: Using finite Truncations to approximate an infinite process.

   Ishwaran and Zarepour (2000) and Ishwaran and James (2001, 2002) proposed to approximate DP (and DPM) models by truncating Sethuraman's stick-breaking representation of the DP.
   The fractions v_h are truncated after H terms by setting $v_H = 1$, leaving

   $$G(\cdot) = \sum_{h=1}^{H} w_h \delta_{m_h}(\cdot)$$

   with $w_h = v_h \prod_{\ell<h}(1 - v_\ell)$ with $v_h \sim \text{Be}(1, \alpha), h = 1, \ldots, H-1$ and $v_H = 1$
   The prior on the point masses remains unchanged, i.e., $m_h \overset{\text{i.i.d.}}{\sim} G_0, h = 1, \ldots, H$.
   We write $G \sim \text{DP}_H(MG_0)$ and DPM_H for a DPM model when a DP_H replaces a DP prior, $f_G(y) = \sum_{h=1}^{H} w_h f_{m_h}(y)$
   Thus you can approximate the Sethuraman's construction by considering a truncation of the infinite sum to K terms, with K large. For example, consider $K = 20$ in the following.

   In addition to the prior c.d.f. realizations, for each value of α, obtain the prior distribution for the mean functional $\mu(G) = \int x \; G(dx)$ and for the variance functional $\sigma^2(G) = \int x^2 \, G(dx) - \left\{\int x \, G(dx)\right\}^2$. Note that, since G_0 has finite first and second moments, both of the random variables $\mu(G)$ and $\sigma^2(G)$ take finite values almost surely; see Section 4 in Ferguson (1973).

   Comment on the results.

2. Now download the following data from:

   http://138.68.227.229/ mguindani/teaching/bio285/DPdata.csv

   Assume the following model for the data

   $$X_i | G \overset{i.i.d.}{\sim} G \quad \text{with} \quad G \sim DP(\alpha G_0)$$

   with $G_0 = \text{N}(0,1)$.

Antoniak (1974) has shown that in a DP, if M indicates the number of *non-empty* components in the infinite stick-breaking sum, the prior expectation $E(M) = \alpha \log((\alpha+n)/\alpha) \to \alpha \log(n)$ for $n \to \infty$. Compute the prior expectation implied by the α values as fixed above. Indicate if they are reasonable based on the prior realizations obtained under (1).

3. Compute the probability $p(X_i = X_j)$ when $i \neq j$ (co-clustering probability).

4. Obtain posterior samples for the unknown distribution $G|X_1, \ldots, X_n$ using the same settings of the DP parameters discussed above. Plot the cdf of the posterior samples, the cdf of the centering distribution G_0 and the cdf of the empirical distribution for each value of α, in a single graph. Comment on the results.

5. Now consider $G_0 = Poisson(3)$. Does your answer under (3) change? Motivate your answer.

Problem 2

The Dirichlet process prior to model the old faithful data

Propose and discuss the use of a DP mixture model to fit the old faithful data in R. Use the package BNPmix (see link) to get posterior density estimates and posterior inference on the possible number of clusters in the datasets. Do sensitivity analysis for various choices of the precision parameter and the loss functions used for data-driven clustering.

Problem 3

An alternative finite approximation of a DP is given by the following model:

$$\begin{aligned} X_1, X_2, \ldots | \pi, \theta &\overset{\text{iid}}{\sim} G_K \\ G_K &= \sum_{i=1}^{K} \pi_i \delta_{\theta_i} \\ \theta_1, \ldots, \theta_K &\overset{\text{iid}}{\sim} H \\ \pi_1, \ldots, \pi_K &\sim \text{Dirichlet}(\beta/K, \ldots, \beta/K) \end{aligned} \tag{1}$$

Let $H = N(0, 1)$ as in Problem 1.

1. Consider the sequence of distributions $G_k(\cdot)$ and the total mass parameter $G_K(\mathfrak{X}) = |\alpha_K|$. Try to justify why it is true that G_K converges to a $DP(\alpha, H)$ (hint, see slide 54).

2. Compute the predictive distribution

$$p(X_n | X_1, \ldots, X_{n-1})$$

based on model (1)

3. Argue that the predictive distributions of $(X_i)_{i \geq 1}$ converge to those of a Blackwell-MacQueen urn as $K \to \infty$.