UCLA Biostatistics 285: Homework 2

Instructor: Dr. Michele Guindani

Soumyakanti Pan, e-mail: span18@ucla.edu

May 19, 2023

1 Problem 1

1.1 Part 1

The function `Sethu_jump` generates the jumps given a truncation option and a α. The `generate_DPH` uses a jump function (here we use the `Sethu_jump`) and takes input a base measure and its parameters along with α, truncation parameter K and number of samples to be generated. The final output is realizations of DP$(\alpha, \mathcal{N}(0,1))$ approximated by finite truncation with 20 terms as described in Ishwaran and Zarepour (2002).

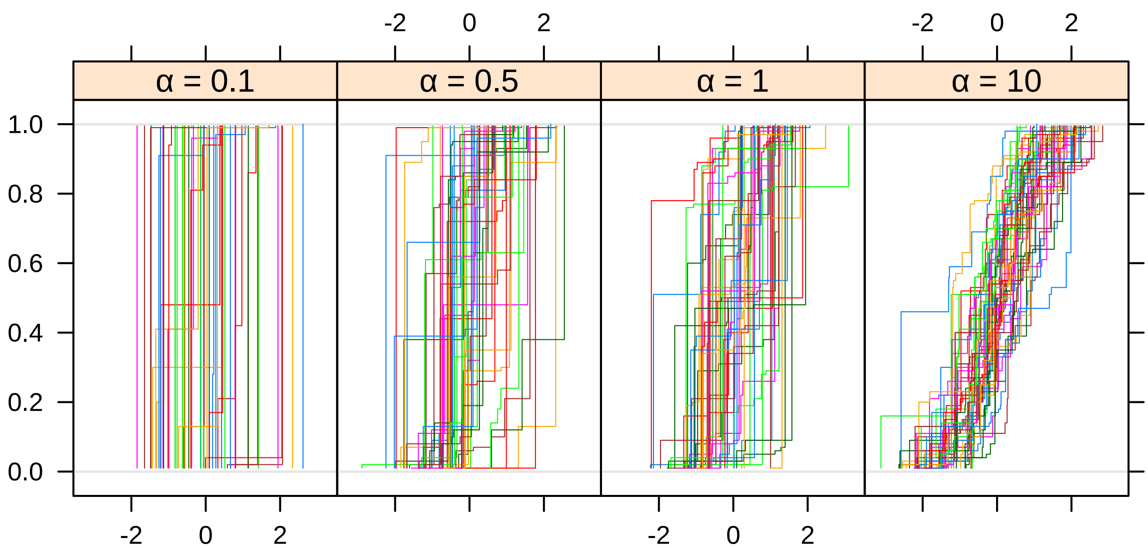

Figure 1: Prior c.d.f realizations of Dirichlet process with different base measures.

We can also get Monte Carlo estimates of the mean functional $\mu(G)$ and the variance functional $\sigma^2(G)$ from the prior realizations of G, drawn by assuming a truncation upto $K = 20$ terms.

We see in Figure 2, that naturally the mean functional $\mu(G)$ is centered around 0 since the base measure is centered around 0 but higher the value of α, more is the concentration of the mean functional around 0. In other words, the mean functional has higher kurtosis for higher α. On the other hand, the variance functional indicates that lower value of α indicates the lower dispersion of atoms in G, whereas, higher value of α indicates higher dispersion of atoms in G.

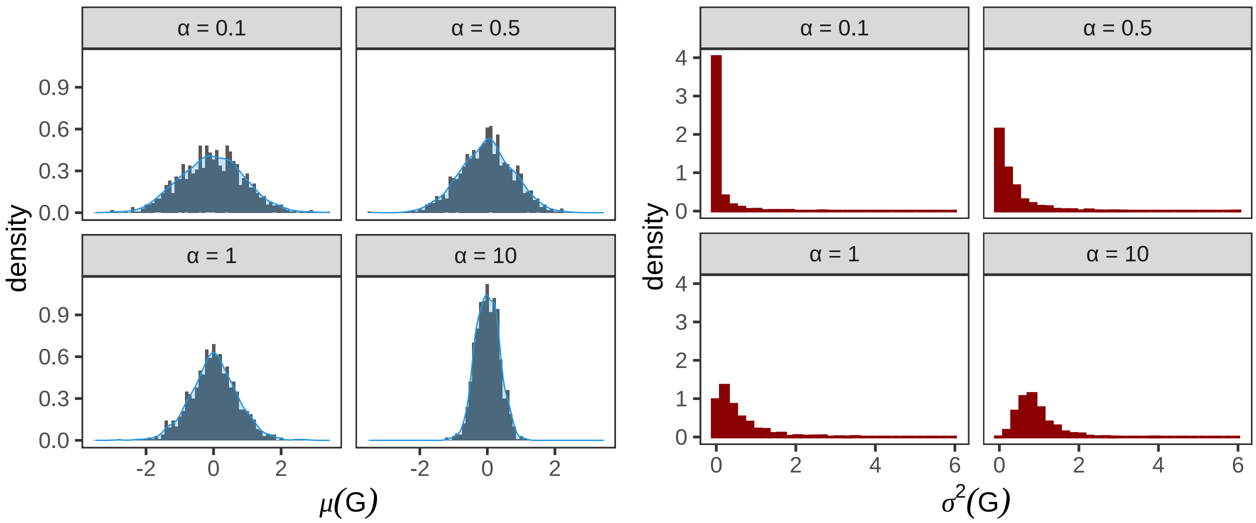

Figure 2: Prior distributions of mean and variance functionals from 1000 prior samples

1.2 Part 2

Following Weak Law of large Numbers (WLLN), we can estimate the expected number of nonempty clusters $E(M)$ by the mean number of unique atoms from each realization of c.d.f. sampled from the prior DP$(\alpha, \mathcal{N}(0,1))$ with different values of α. Here, we have considered $\alpha = 0.1, 0.5, 1, 10$. Note that the sampled prior realizations are generated from a approximate prior achieved by truncation the infinite mixture to a mixture of 20 atoms. The red line in Figure 3 denotes the theoretical expected number of nonempty clusters as shown by Antoniak (1974) given by $E(M) = \alpha \log((\alpha + n)/\alpha)$. We see for smaller values of α, the approximation may have given reasonable estimates of number of non-empty clusters but I suspect that for a large value of α, the truncation approximation might not be reasonable.

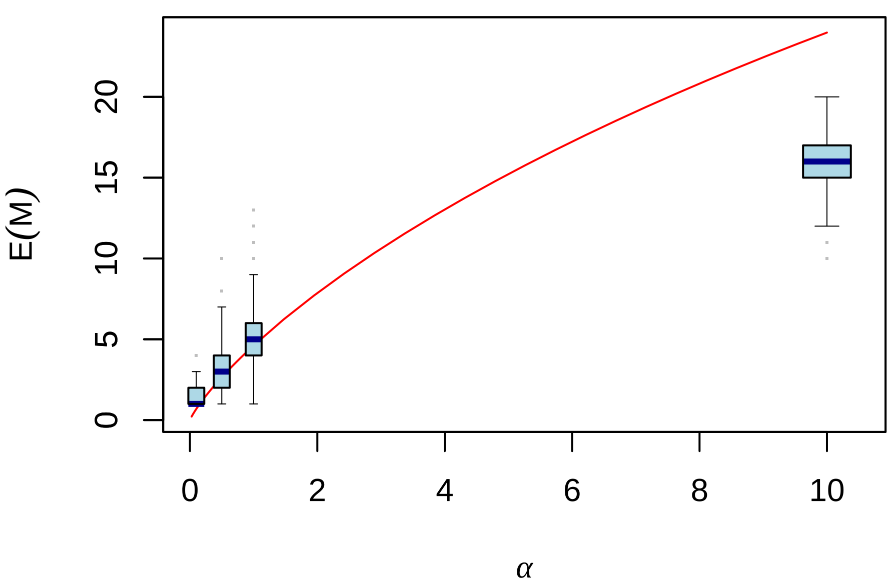

Figure 3: Comparison of theoretical expected number of non-empty clusters and estimated with finite truncation approximation based on 1000 samples

1.3 Part 3

As we know that, we can construct a Dirichlet process on $(\mathfrak{X}, \mathfrak{S}, P)$ from a Pólya sequence, invoking de Finetti's theorem, assuming $i < j$, we can write the distribution of the j-th observation as

$$X_j \mid X_{j-1}, \ldots, X_i, \ldots, X_1 \sim \frac{\alpha}{\alpha + j - 1} G_0 + \frac{1}{\alpha + j - 1} \sum_{k=1}^{j-1} \delta_{X_k}.$$

Hence, for any event $A \in \mathfrak{S}$, we can calculate the probability $P(X_j \in A)$ as follows

$$P(X_j \in A) = \frac{\alpha}{\alpha + j - 1} G_0(A) + \frac{1}{\alpha + j - 1} \sum_{k=1}^{j-1} \delta_{X_k}(A).$$

Considering $A = \mathbf{1}(X_i)$ denoting the event that the observed value is equal to X_i, we can calculate the above quantity following from $G_0(A) = 0$ as G_0 is non-atomic and

$$P(X_j = X_i) = \frac{1}{\alpha + j - 1} \sum_{k=1}^{j-1} \delta_{X_k}(A) = \frac{n_i}{\alpha + j - 1}$$

where n_i is the number of occurrence of X_i among the samples $X_1, X_2, \ldots, X_{j-1}$.

1.4 Part 4

We know from the conjugacy of a Dirichlet process, if we have the data generation model as $X_1, X_2, \ldots, X_n \overset{\text{ind}}{\sim} G$ with $G \sim \text{DP}(\alpha G_0), G_0 = \mathcal{N}(0, 1)$, then we can write the posterior as again a Dirichket process with a different base measure.

$$G \mid X_1, X_2, \ldots, X_n \sim \text{DP}(\alpha + n, \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta_{X_i} + \frac{\alpha}{\alpha + n} G_0)$$

Hence, from Figure 4, we see that with increase in α, we see more posterior samples (in gray) tending towards the prior base measure (given in red dashed line) and for lesser values of α, we have more attenuation towards the empirical cdf of the given data (drawn in solid black line).

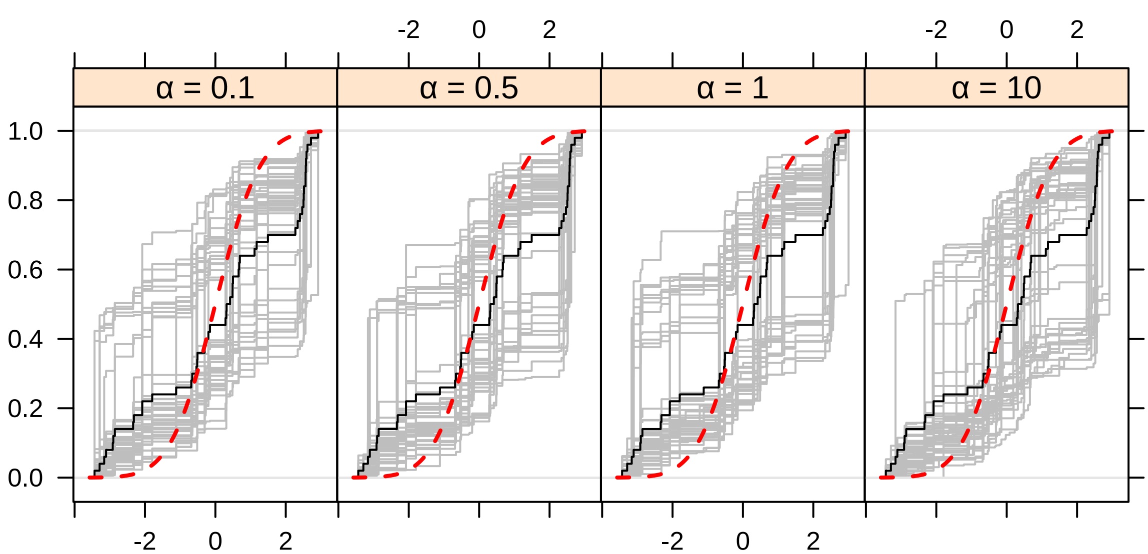

Figure 4: Posterior samples (gray) of the generative random measure compared with the empirical c.d.f. (black) and the original base measure of the DP prior (red).

1.5 Part 5

Now if the prior base measure is atomic, $G_0 = \text{Poisson}(3)$, $G_0(A)$ will no longer be 0 and hence we will have

$$P(X_j = X_i) = G_0(A) + \frac{n_i}{\alpha + j - 1} = \frac{e^{-3}3^{X_i}}{\Gamma(X_i + 1)} + \frac{n_i}{\alpha + j - 1}$$

where n_i is the number of occurrence of X_i among the samples $X_1, X_2, \ldots, X_{j-1}$.

2 Problem 2

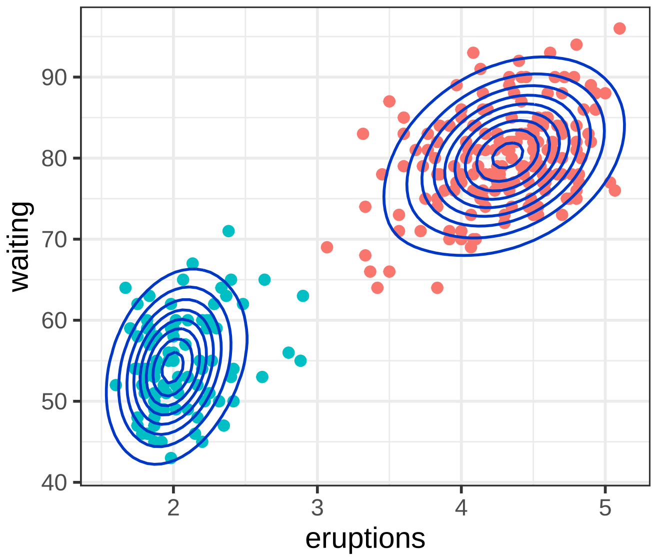

Figure 5: Posterior density

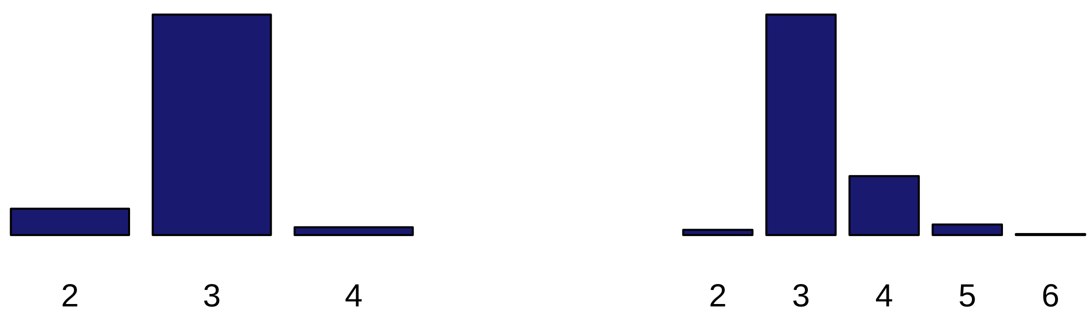

Figure 6: Posterior samples of number of clusters with different values of strength in PY prior

3 Problem 3

The model is a Dirichlet-multinomial process given as follows.

$$X_1, X_2, \ldots \overset{\text{ind}}{\sim} G_K, G_K = \sum_{k=1}^{K} \pi_k \delta_{\theta_k}$$
$$\theta_1, \ldots, \theta_K \overset{\text{iid}}{\sim} H$$
$$\pi_1, \ldots, \pi_K \sim \text{Dir}(\beta/K, \ldots, \beta/K)$$

The base measure H is $\mathcal{N}(0, 1)$.

3.1 Part 1

Here, G_K is a Dirichlet-multinomial process of order K, and from part (ii) of Theorem 4.19 of Ghoshal, van der Vaart (2017), which says that if for a Dirichlet-multinomial process with parameters $(\alpha_{1,K}, \ldots, \alpha_{K,K})$ with $\max_{1 \le k \le K} \alpha_{k,K}/\alpha_{.K} \to 0$ where $\alpha_{.K} = \sum_{k=1}^K \alpha_{k,K}$, then if $\alpha_{.K} \to M$, then $\int \psi\, dG_K \to \int \psi\, dG$, where $G \sim DP(M\bar{G})$, for any $\psi \in \mathbb{L}^1(G)$.

In our case, $\alpha_{.K} = \sum_{k=1}^K \alpha_{k,K} = \beta$ is constant for all K and hence trivially the condition holds. Hence, we can argue the usage of Dirichlet-multinomial process as a approximation of the infinite dimensional DP prior. Moreover, we also know that for sequence of finite measures $\bar{\alpha}_m \to \bar{\alpha}$, if $|\alpha_m| \to M$ finite, then $\mathrm{DP}(\alpha_m) \to \mathrm{DP}(\alpha)$.

3.2 Part 2

Suppose $s_{i,k}$ denotes the indicator variable $\mathbf{1}(X_i = \theta_k)$ for $i = 1, \ldots, n+1$ and $k = 1, \ldots, K$. Then the posterior predictive distribution can be given as follows.

$$
\begin{aligned}
p(X_{n+1} \mid X_1, \ldots, X_n) &= \int\int p(X_{n+1} \mid \theta, \pi) p(\pi, \theta \mid X_1, \ldots, X_n) d\pi d\theta \\
&\propto \int\int p(X_{n+1} \mid \theta, \pi) p(X_1, \ldots, X_n \mid \pi, \theta) p(\pi) p(\theta) d\pi d\theta \\
&= \int \prod_{k=1}^K \pi_k^{s_{n+1,k}} p(\theta) \int \prod_{k=1}^K \pi_k^{n_k + \frac{\beta}{K} - 1} d\pi d\theta \\
&= \int p(\theta) \int \prod_{k=1}^K \pi_k^{s_{n+1,k} + n_k + \frac{\beta}{K} - 1} d\pi d\theta \\
&\propto \int p(\theta) \prod_{k=1}^K \Gamma(s_{n+1,k} + n_k + \frac{\beta}{K}) d\theta \\
&= \prod_{k=1}^K \int \exp\{-\frac{1}{2}\theta_k^2\} \Gamma\left(s_{n+1,k} + n_k + \frac{\beta}{K}\right) d\theta
\end{aligned}
$$

Matching with the theorems relating to analogous MDP models in Ishwaran and Zarepour (2002), I guess that the predictive posterior can be given by

$$
\sum_{k=1}^K \frac{n_k + \frac{\beta}{K}}{n+\beta} \delta_{X_k^*} + \beta\left(1 - \frac{K}{n}\right) G_{n-K}^{(n)}
$$

where,

$$
G_{n-K}^{(n)} = \sum_{k=1}^{N-k} p_k^* \delta_{\theta_k}
$$

where p_k^* are jumps (could not find the correct value, but can be given as a stick-breaking construction) and θ_k are iid from $\mathcal{N}(0,1)$.

3.3 Part 3

Clearly, as $K \to \infty$, we have the jumps $\frac{n_k + \beta/K}{n+\beta} \to \frac{n_k}{n+\beta}$ which aligns with the predictive posterior update of Pólya urn scheme.

References

Antoniak, Charles E. 1974. "Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems." *The Annals of Statistics* 2 (6): 1152–74. http://www.jstor.org/stable/2958336.

Ishwaran, Hemant, and Mahmoud Zarepour. 2002. "Exact and Approximate Sum Representations for the Dirichlet Process." *Canadian Journal of Statistics* 30 (2): 269–83. https://doi.org/https://doi.org/10.2307/3315951.