Dirichlet process note

Daeyoung Lim*
Department of Statistics
Korea University

May 17, 2016

1 Dirichlet Distribution

As Dirichlet process is the infinite-dimensional generalization of the Dirichlet distribution, we should first take a look at what Dirichlet distribution is and what properties it possesses. Afterwards, it will be quite straightforward to make the stochastic process have those properties as we construct it. The Dirichlet distribution of dimension D is a continuous probability measure on Δ_D having the density function

$$p(\boldsymbol{\pi}|\beta_1, \ldots, \beta_D) = \frac{\Gamma\left(\sum_i \beta_i\right)}{\prod_i \Gamma(\beta_i)} \prod_{i=1}^{D} \pi_i^{\beta_i - 1} \tag{1}$$

where the parameters $\beta_i \geq 0, \forall i$. In preparation of the generalization to Dirichlet process, let us reparameterize the distribution as

$$p(\boldsymbol{\pi}|\alpha g_{01}, \ldots, \alpha g_{0D}) = \frac{\Gamma(\alpha)}{\prod_i \Gamma(\alpha g_{0i})} \prod_{i=1}^{D} \pi_i^{\alpha g_{0i} - 1} \tag{2}$$

where $\alpha = \sum_i \beta_i$ and $g_{0i} = \beta_i / \left(\sum_i \beta_i\right)$. We will hereafter denote the distribution by $\boldsymbol{\pi} \sim \operatorname{Dir}(\alpha g_0)$; the mean and variance of such a reparameterized Dirichlet random variable are

$$\mathbb{E}[\pi_i] = g_{0i}, \operatorname{Var}[\pi_i] = \frac{g_{0i}(1 - g_{0i})}{\alpha + 1}. \tag{3}$$

2 How to interpret Dirichlet distribution

It is a well-known fact that the sum of all elements in a Dirichlet random vector is unity. Therefore, it is not difficult to admit that somehow a Dirichlet random vector is a realization of some sort of discrete distribution with a finite number of possible values. In other words, we also refer to a Dirichlet distribution as a distribution on a probability simplex, which is basically the same thing.

$$X_{n+1}|X_1, \ldots, X_n \sim \sum_{i=1}^{n} \frac{1}{\alpha(\mathcal{X}) + n} \delta_{X_i} + \frac{1}{\alpha(\mathcal{X}) + n} \alpha \tag{4}$$

3 DPM: Rao-Blackwellized MCMC

c_i is the cluster to which y_i belongs. Fast MCMC algorithm integrates out θ and only constructs a chain for the categorical variable c.

- If $c = c_j$ for some $j \neq i$:

$$\mathrm{P}(c_i = c \mid c_{-i}, y_i) = b \frac{n_{-i,c}}{n - 1 + M} \int F(y_i, \theta) \, dH_{-i,c}(\theta). \tag{5}$$

*Prof. Taeryon Choi

- Otherwise,

$$\mathrm{P}\left(c_i \neq c_j \text{ for all } j \neq i \mid c_{-i}, y_i\right) = b\frac{M}{n-1+M}\int F\left(y_i, \theta\right)\, dG_0\left(\theta\right). \tag{6}$$

Now let $G_0 \equiv \mathrm{N}\left(\mu_0, \sigma_0^2\right)$ and $F \equiv \mathrm{N}\left(\theta, \sigma^2\right)$.

$$p\left(\theta \mid \left\{y_j \mid j \neq i, c_j = c\right\}, G_0\right) \equiv \mathrm{N}\left(\frac{\sigma_0^2 \sum_{c_j=c} y_j + \mu_0\sigma^2}{n_c\sigma_0^2 + \sigma^2}, \frac{\sigma^2\sigma_0^2}{n_c\sigma_0^2 + \sigma^2}\right) \tag{7}$$

$$\int F\left(y_i, \theta\right)\, dH_{-i,c}\left(\theta\right) \equiv \mathrm{N}\left(y_i \middle| \mu_\theta, \sigma_\theta^2 + \sigma_y^2\right) \tag{8}$$

where the mean of (7) is μ_θ in (8) and the same for σ_θ^2. σ_y^2 in (8) is simply σ^2 but the subscript was attached to make the distinction clear.

4 DPM: Latent variable MCMC

We have marginalized out θ in the previous section, which is also called *Rao-Blackwellization*. We examine another form of MCMC which also has a chain with respect to θ along with the categorical variable c.

- If $c = c_j$ for some $j \neq i$,

$$\mathrm{P}\left(c_i = c \mid c_{-i}, y_i, \theta\right) = b\frac{n_{-i,c}}{n-1+M}F\left(y_i, \theta_c\right). \tag{9}$$

- Otherwise,

$$\mathrm{P}\left(c_i \neq c_j \text{ for all } j \neq i \mid c_{-i}, y_i, \theta\right) = b\frac{M}{n-1+M}\int F\left(y_i, \theta\right)\, dG_0\left(\theta\right). \tag{10}$$

In Neal's paper, the Gibbs sampler construction is summerized as follows:

- For $i = 1, \ldots, n$: If the present value of c_i is associated with no other observation (i.e., $n_{-i,c_i} = 0$), remove θ_{c_i} from the state. Draw a new value for c_i from $c_i \mid c_{-i}, y_i, \theta$ as defined above. If the new c_i is not associated with any other observation, draw a value for θ_{c_i} from H_i and add it to the state. H_i is the posterior distribution based on the prior G_0 and the single observation y_i.
- For all $c \in \{c_1, \ldots, c_n\}$: Draw a new value from $\theta_c \mid$ all y_i for which $c_i = c$— that is, from the posterior distribution based on the prior G_0 and all the data points currently associated with latent class c.

4.1 Gaussian mixtures

Let's play with an actual example. Recall

$$y_i|\theta_i \sim \mathrm{N}\left(\theta_i, \sigma^2\right) \tag{11}$$
$$\theta_i|G \sim G \tag{12}$$
$$G \sim \mathrm{DP}\left(M, G_0\right) \tag{13}$$
$$G_0 \equiv \mathrm{N}\left(\mu_0, \sigma_0^2\right). \tag{14}$$

And

$$p\left(s_i = j|s^-, \theta^{*-}, y\right) \propto \begin{cases} n_j^- p\left(y_i|\theta_j^{*-}\right) & j = 1, \ldots k^- \\ M\int p\left(y_i|\theta_i\right)\, dG_0\left(\theta_i\right) & j = k^- + 1 \end{cases} \tag{15}$$

$$p\left(\theta_i|s_i = j, s^-, \theta^{*-}, y\right) = \begin{cases} \delta_{\theta_j^{*-}} & j = 1, \ldots, k^- \\ p\left(\theta_i|y_i, G_0\right) & j = k^- + 1 \end{cases}. \tag{16}$$

$$
\begin{aligned}
\int p\left(y_i|\theta_i\right)\, dG_0\left(\theta_i\right) &= \frac{1}{2\pi\sqrt{\sigma^2\sigma_0^2}}\int \exp\left\{-\frac{1}{2\sigma^2}\left(y_i-\theta_i\right)^2-\frac{1}{2\sigma_0^2}\left(\theta_i-\mu_0\right)^2\right\} d\theta_i \qquad (17)\\
&= \frac{1}{2\pi\sqrt{\sigma^2\sigma_0^2}}\int \exp\left\{-\frac{\left(\sigma_0^2+\sigma^2\right)\theta_i^2-2\left(y_i\sigma_0^2+\mu_0\sigma^2\right)+y_i^2\sigma_0^2+\mu_0^2\sigma^2}{2\sigma^2\sigma_0^2}\right\} d\theta_i \qquad (18)\\
&= \frac{1}{2\pi\sqrt{\sigma_0^2\sigma^2}}\int \exp\left\{-\frac{\sigma_0^2+\sigma^2}{2\sigma^2\sigma_0^2}\left(\theta_i^2-2\frac{y_i\sigma_0^2+\mu_0\sigma^2}{\sigma_0^2+\sigma^2}\theta_i\right)-\frac{y_i^2\sigma_0^2+\mu_0^2\sigma^2}{2\sigma^2\sigma_0^2}\right\} d\theta_i \qquad (19)\\
&= \frac{1}{2\pi\sqrt{\sigma_0^2\sigma^2}}\exp\left\{-\frac{y_i^2\sigma_0^2+\mu_0^2\sigma^2}{2\sigma^2\sigma_0^2}+\frac{\left(y_i\sigma_0^2+\mu_0\sigma^2\right)^2}{2\sigma^2\sigma_0^2\left(\sigma_0^2+\sigma^2\right)}\right\}\int \exp\left\{-\frac{\sigma_0^2+\sigma^2}{2\sigma^2\sigma_0^2}\left(\theta_i-\frac{y_i\sigma_0^2+\mu_0\sigma^2}{\sigma_0^2+\sigma^2}\right)^2 d\theta_i\right\} \qquad (20)\\
&= \frac{1}{2\pi\left(\sigma_0^2+\sigma^2\right)}\exp\left\{-\frac{\left(y_i-\mu_0\right)^2}{2\left(\sigma_0^2+\sigma^2\right)}\right\} \qquad (21)\\
&\equiv \mathrm{N}\left(y_i|\mu_0,\sigma^2+\sigma_0^2\right) \qquad (22)
\end{aligned}
$$

And then for the posterior,

$$
\begin{aligned}
p\left(\theta_i|y_i,G_0\right) &= \frac{p\left(y_i|\theta_i\right)\, dG_0\left(\theta_i\right)}{\int p\left(y_i|\theta_i\right)\, dG_0\left(\theta_i\right)} \qquad (23)\\
&= \frac{\mathrm{N}\left(y_i|\theta_i,\sigma^2\right)\mathrm{N}\left(\theta_i|\mu_0,\sigma_0^2\right)}{\mathrm{N}\left(y_i|\mu_0,\sigma^2+\sigma_0^2\right)} \qquad (24)\\
&\sim \mathrm{N}\left(\frac{y_i\sigma_0^2+\mu_0\sigma^2}{\sigma_0^2+\sigma^2},\frac{\sigma^2\sigma_0^2}{\sigma^2+\sigma_0^2}\right) \qquad (25)
\end{aligned}
$$