Dirichlet Process

DaeYoung Lim and SeoYoon Cho

2016.05.16 (Mon)

Contents

Dirichlet-Multinomial Model

- Prior and Model

$$
\begin{aligned}
Y_i|p &\overset{ind}{\sim} \text{Multinomial}(p), \qquad i = 1, \cdots, n \\
p = (p_1, \cdots, p_k) &\sim \text{Dirichlet}(\alpha_1, \cdots, \alpha_k) \\
\pi(p) &= \frac{\Gamma(\sum \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k p_i^{\alpha_i - 1}
\end{aligned}
$$

- Posterior

$$
\text{Dirichlet}(\alpha_1 + \sum I(Y_i = 1), \cdots, \alpha_k + \sum I(Y_i = k))
$$

- Prior **has full support** and is **conjugate**, thus leading to an easy update

DD to DP

- Dirichlet Distribution

$$G \sim \text{Dir}(\alpha), \quad \theta_i | G \sim G$$

  - For any measurable set A, partition $\mathbf{R}$ into A and A^c
  - Dirichlet-Multinomial (Beta-Binomial) model on A and A^c

$$P(\theta_i \in A) = \frac{\alpha(A)}{\alpha(\mathbf{R})}$$

- Dirichlet Process

$$G \sim \text{DP}(MG_0), \quad \theta_i | G \sim G$$

  - Baseline probability distribution G_0, mass(precision) parameter M, MG_0: base measure of DP

Ferguson (1973)

- Probability Space : (Θ, A, G)
- **arbitrary** partition $\{A_1, \cdots, A_k\}$ of Θ
- $G \sim DP(M, G_0)$ if

$$(G(A_1), \cdots, G(A_k)) \sim DP(MG_0(A_1), \cdots, MG_0(A_k))$$

- well-defined infinite dimensional model $p(G)$ because of **Kolmogorov's consistency conditions**

  (guarantees suitably consistent collection of finite-dimensional distributions define a stochastic process.)

Kolmogorov's Consistency Conditions

- Let T: interval, $n \in \mathbf{N}$. Then for each $k \in \mathbf{N}$ and finite sequence of times $t_1, \cdots, t_k \in T$, let $v_{t_1 \cdots t_k}$ be a probability measure on $(\mathbf{R}^n)^k$

1. $\forall$ permutations π of $\{1, \cdots, k\}$, measurable sets $F_i \subseteq \mathbf{R}^n$,
$$\mathbf{v_{t_{\pi(1)} \cdots t_{\pi(k)}}(F_{\pi(1)} \times \cdots \times F_{\pi(k)}) = v_{t_1 \cdots t_k}(F_1 \times \cdots \times F_k)}$$
2. For $\forall$ measurable sets $F_i \subseteq \mathbf{R}^n, m \in \mathbf{N}$,
$$v_{t_1 \cdots t_k}(F_1 \times \cdots \times F_k) = v_{t_1 \cdots t_k t_{k+1} \cdots t_{k+m}}(F_1 \times \cdots \times F_k \times \mathbf{R}^n \times \cdots \times \mathbf{R}^n)$$

- When the two conditions satisfied, $\exists$ a probability space (Ω, F, P) and a stochastic process $X : T \times \Omega \to \mathbf{R}^n$ s.t.
$$v_{t_1 \cdots t_k}(F_1 \times \cdots \times F_k) = P(X_{t_1} \in F_1, \cdots, X_{t_k} \in F_k)$$
for all $t_i \in T,\ k \in \mathbf{N}$, and measurable sets $F_i \subseteq \mathbf{R}^n$
- X has $v_{t_1 \cdots t_k}$ as its finite-dimensional distribution relative to times $t_1, \cdots, t_k$

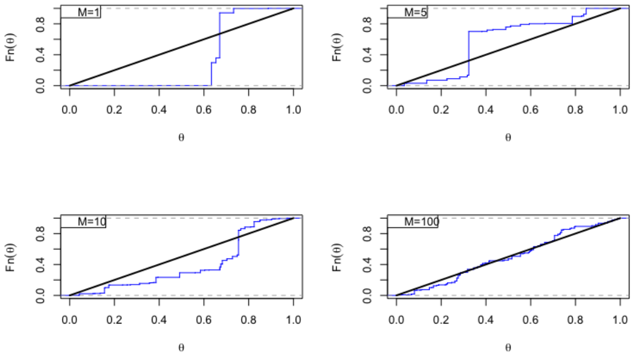
Sethuraman (1994)

- Stick-Breaking Construction
  - $\delta_\theta(\cdot)$: point mass at θ
  - if $(\tilde{\theta}_h) \stackrel{iid}{\sim} G_0$
  - $v_h \stackrel{iid}{\sim}$ Beta$(1, M)$
  - $w_h = v_h \prod_{k<h}\{1 - v_k\}$

$$G(\cdot) = \sum_{h=1}^{\infty} w_h \delta_{\tilde{\theta}_h}(\cdot) \ \sim DP(M, G_0) \text{ prior}$$

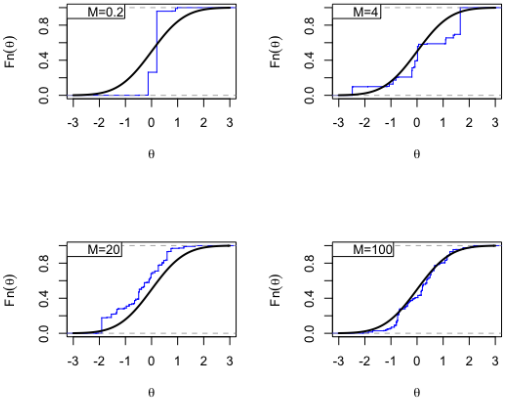
Simulation

- $G_0 =$ Unif$(0,1)$, 1000 samples from $G(\cdot)$



Simulation

- $G_0 = \mathrm{N}(0, 1)$, 1000 samples from $G(\cdot)$



Simulation Interpretation

- $M \uparrow$: reduce the variability
- $M \downarrow$: small number of weights concentrate most of the probability mass

Predictive Probability Function

- $\theta_i | G \overset{iid}{\sim} G$ where $G \sim DP(M, G_0)$.
- k_n: number of unique values among $\{\theta_1, \cdots, \theta_n\}$
- $\{\theta_1^*, \cdots, \theta_{k_n}^*\}$ be these unique values
- n_{n_j}: number of draws among $\{\theta_1, \cdots, \theta_n\}$ that are equal to θ_j^*

$$p(\theta_{n+1} | \theta_n, \cdots, \theta_1) \propto \sum_{j=1}^{k_n} n_{n_j} \delta_{\theta_j^*} + MG_0$$

Blackwell and MacQueen (1973)

$$p(\theta_{n+1}|\theta_n, \cdots, \theta_1) \propto \sum_{j=1}^{k_n} n_{n_j} \delta_{\theta_j^*} + MG_0$$

- new $\theta_i = \theta_j^*$ with probability $\propto n_{n_j}$

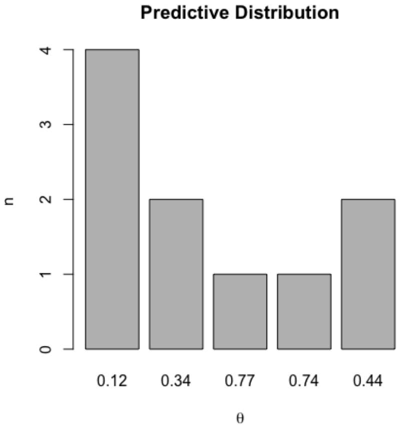
  or
- new θ_i sampled from G_0 with probability $\propto M$.
- After integrating G, observations are **exchangeable**, have identical marginal distribution G_0 but are not independent.

Polya Urn (Chinese Restaurant Process)

- Urn has initially M **black** and **one colored** ball (color is randomly selected according to G_0)
- If a **colored** ball is drawn, return it along with another ball of the **same color** to the urn
- If a **black** ball is drawn, return it along with a ball of a **new color** randomly selected according to G_0

Simulation

- $G_0 = \text{Unif}(0,1)$, $M = 5$, 10 predictive samples



Normalized Random Measure with Independent Increments (NRMI)

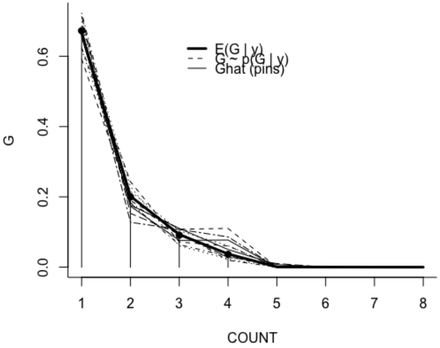
- Dirichlet from Gamma

$$\begin{aligned} y_1, \cdots, y_k &\overset{iid}{\sim} \text{Gamma}(\alpha_i, 1) \\ x_i &= \frac{y_i}{\sum_{i=1}^k y_i} \\ \Rightarrow (x_1, \cdots, x_k) &\overset{iid}{\sim} \text{Dir}(\alpha_1, \cdots, \alpha_k) \end{aligned}$$

- $\mu(A) \sim$ Gamma$(MG_0(A), 1)$ for any $A \subset \Theta$
- $G(\cdot) \equiv \frac{\mu(\cdot)}{\mu(\Theta)} \sim DP(M, G_0)$

Simulation

- $y_i \sim G$ with prior $G \sim DP(M, G_0)$
- $M = 1$ and $G_0 = Poi^+(2)$
- 10 posterior draws



Large Weak Support

- Under mild conditions, any distribution with the same support as G_0 can be well approximated weakly by a DP random probability measure
- Let $supp(Q) \subset supp(G_0)$. For any finite number of measurable sets $A_1, \cdots, A_k$ and $\epsilon > 0$,

$$\pi\{|G(A_i) - Q(A_i)| < \epsilon, \text{ for } i = 1, \cdots, k\} > 0$$

Ferguson's Definition

- Random Variable $G(A)$ for any $A \subset \Theta$

$$G(A) \sim \text{Beta}\left(MG_0(A), M(1 - G_0(A))\right)$$

$$E[G(A)] = G_0(A), \qquad Var[G(A)] = \frac{G_0(A)(1 - G_0(A))}{M + 1}$$

- G_0 : expected shape of G
- M : controls the variability of the realizations around G_0
- $E(w_h) = \frac{1}{M+1}\left(\frac{M}{M+1}\right)^{h-1} \quad \because v_h \overset{iid}{\sim} \text{Beta}(1, M)$

Conjugacy

- $\theta_1, \cdots, \theta_n$ iid
- $\theta_i|G \sim G$ and $G \sim DP(M, G_0)$
- Posterior

$$\Rightarrow G|\theta_1, \cdots, \theta_n \sim \text{DP}\left(M+n, \frac{MG_0 + \sum \delta_{\theta_i}}{M+n}\right)$$

- Posterior Mean

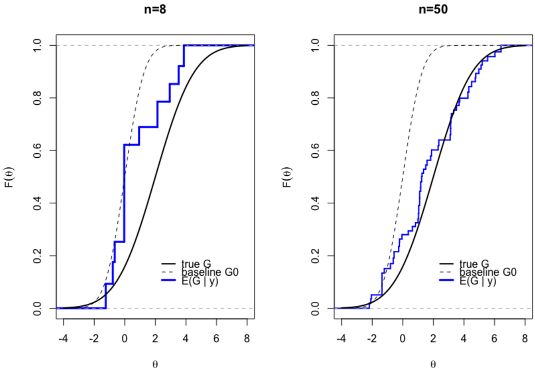
$$E(G|\theta_1, \cdots, \theta_n) = \frac{M}{M+n}G_0 + \frac{n}{M+n}\frac{\sum_{i=1}^n \delta_{\theta_i}}{n}$$

- Consistency

  since empirical cdf is consistent if iid, for some true distribution G_T, as $n \to \infty$, $G(A)|\theta_1, \cdots, \theta_n \xrightarrow{p} G_T(A)$ for any measurable set A

Simulation (Consistency)

- true $G \sim N(2, 2^2)$
- baseline $G_0 \sim N(0, 1)$
- $M = 5$



Dirichlet Process Mixtures

- Motivation
  - Discrete nature of DP
  - However, unknown distribution may be **continuous**
  - For some hierarchical models, DP prior may lead to inconsistent estimator when the true distribution is continuous.
- Mitigation
  - Add a convolution with a continuous kernel to G
- DPM

$$y_1, \cdots, y_n \sim F(y_i) = \int P(y_i|\theta)\, G(d\theta), \quad G \sim \text{DP}(M, G_0)$$

where $p(y_i|\theta)$ is a parametric distribution indexed by a finite dimensional parameter θ.

Stick-Breaking Construction

$$y_i|(w_h),(\tilde{\theta}_h) \sim \sum_{h=1}^{\infty} w_h P(y_i|\tilde{\theta}_h) = F(y_i)$$

where $w_h = v_h \prod_{\ell<h}(1-v_\ell),\ v_h \sim \text{Beta}(1,M),\ \tilde{\theta}_h \sim G_0$

- **countable** mixtures with an **infinite** number of components
- support on a large classes of distributions

Clustering

DPM induces clustering among the observations, with M controlling the a priori expected number clusters in the sample.

- $M \to 0$: model reduces to a single component mixture (fully parametric model)

$$y_i \overset{iid}{\sim} p(y|\theta), \quad \theta \sim G_0$$

- $M \to \infty$: each observation assigned its own singleton cluster

$$y_i \overset{iid}{\sim} \int p(y_i|\theta) G_0(d\theta)$$

Hierarchical Model

- latent random effects θ_i

$$y_i|\theta_i \sim p(y_i|\theta_i), \quad \theta_i|G \sim G, \quad G \sim \text{DP}(M, G_0)$$

- highlighting the nature of clusters generated by ties among the θ_i

$$y_i|\theta_i \sim p(y_i|\theta_i), \quad (\theta_1, \cdots, \theta_n) \sim p(\theta_1, \cdots, \theta_n)$$

Cluster Indicator variable

- cluster indicator variables (s_i) s.t. $\theta_i = \theta^*_{s_i}$

$$y_i | s_i, (\theta^*_j) \sim p(y_i | \theta^*_{s_i}), \quad \theta^*_j \sim G_0,$$

$$p(s_1, \cdots, s_n) = \frac{\Gamma(M)}{\Gamma(M+n)} M^k \prod_{j=1}^{k} \Gamma(n_j)$$

where k: number of distinct values among $s_1, \cdots, s_n$ and $n_j = \sum_i I(s_i = j)$

  - prior distribution on all possible partitions of the data into at most n groups
  - for any finite sample size n, at most n distinct $\tilde{\theta}$ are sampled as θ^*_j

Posterior Simulation for DPM Models

$$y_i|\theta_i \sim p(y_i|\theta_i), \quad \theta_i|G \sim G(\theta_i), \quad G \sim \text{DP}(M, G_0)$$

- kernel $p(y_i|\theta_i)$
- unknown mixing measure $G \sim$ DP prior

Collapsed Gibbs Samplers (Conjugate models)

- Species Sampling Model

$$\theta_n|\theta_{n-1},\cdots,\theta_1 \sim \sum_{j=1}^{k_n-1} \frac{n_{n-1,j}}{M+n-1}\delta_{\theta_j^*} + \frac{M}{M+n-1}G_0$$

where $n_{n-1,j}$: number of θ_i equal to θ_j^*

  - **exchangeable**: full conditional prior distribution for any θ_i given θ_{-i}

Full Conditional Posterior Distribution for θ_i

$$
\begin{aligned}
\theta_i|\theta_{-i}, y \quad &\propto \quad \sum_{j=1}^{k^-} n_j^- p(y_i|\theta_j^{*-})\delta_{\theta_j^{*-}} + Mp(y_i|\theta_i)G_0(\theta_i) \\
&= \quad \sum_{j=1}^{k^-} \{n_j^- p(y_i|\theta_j^{*-})\}\delta_{\theta_j^{*-}} + \\
&\qquad \left\{ M \int p(y_i|\theta_i)\, dG_0(\theta_i) \right\} p(\theta_i|y_i, G_0)
\end{aligned}
$$

where $^-$: the appropriate quantity with θ_i excluded.

- $p(\theta_i|y_i, G_0) = \frac{p(y_i|\theta_i)\, dG_0(\theta_i)}{\int p(y_i|\theta_i)\, dG_0(\theta_i)}$: posterior on θ_i in a singleton cluster
- $\int p(y_i|\theta_i)\, dG_0(\theta_i)$: prior marginal distribution for y_i under G_0

Gibbs Sampler for θ_i

- sample θ_i equal to one of the unique θ_j^*'s with probability $\propto \; n_j^- p(y_i|\theta_j^*)$
  or
- sample from the posterior distribution based solely on y_i with probability $\propto \; M \int p(y_i|\theta_i)\, dG_0(\theta_i)$

- when the mixture components are well separated, slow mixing
- faster mixing by including an additional transition probability
- more efficient sampler by first sampling indicators from $p(s_i|s^-, y)$ sequentially and then sampling each θ_j^* from $p(\theta_j^*|y, s)$

$p(s_i|s^-, y)$

- hierarchical model

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

and

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

$p(s_i | s^-, y)$

- $\mathbf{y}_j^{*-} = (y_\ell;\ s_\ell = j$ and $\ell \neq i)$: obs in the jth cluster w/o y_i
- Remove θ_j^{*-} from the conditioning set by integrating with respect to $p(\theta_j^{*-} | s^-, y) = p(\theta_j^{*-} | y_j^{*-})$

$$p(s_i = j | s^-, y) \propto \begin{cases} n_j^- \int p(y_i | \theta_j^{*-})\, dp(\theta_j^{*-} | y_j^{*-}) & j \leq k^- \\ M \int p(y_i | \theta_i)\, dG(\theta_i) & j = k^- + 1 \end{cases}$$

- Full conditional posterior for θ_j^*

$$p(\theta_j^* | s, y) \propto G_0(\theta_j^*) \prod_{\{i: s_i = j\}} p(y_i | \theta_j^*)$$

- When $G_0(\theta)$ is conjugate to $p(y_i | \theta)$, all of $\int p(y_i | \theta_j^{*-})\, dp(\theta_j^{*-} | y_j^{*-}),\ \int p(y_i | \theta_i)\, dG_0,\ p(\theta_j^* | s, y)$ are usually available in closed form and implementation of the algorithm is straightforward.