Some Applications of Dirichlet Processes

Sepehr Akhavan

May 15, 2014

Dirichlet Distributions

- A **Dirichlet Distribution** is a continuous multivariate probability distribution over a K-dimensional probability simplex where:

$$\Delta_K = \{(\pi_1, \pi_2, \ldots, \pi_K) : \pi_j \geq 0, \sum_{j=1}^{K} \pi_j = 1\}$$

- If $(\pi_1, \pi_2, \ldots, \pi_K)$ are Dirichlet distributed, then the density is of the form:

$$(\pi_1, \pi_2, \ldots, \pi_K) \sim Dirichlet(\alpha_1, \alpha_2, \ldots, \alpha_K) \text{ where } \alpha_j > 0$$

- The density function for this Dirichlet distribution is then of the form:

$$P(\pi_1, \pi_2, \ldots, \pi_K) \propto \prod_{j=1}^{K} {\pi_j}^{\alpha_j - 1}$$

Dirichlet Distributions - Agglomerative Property

- Based on this property we can combine some elements of the probability vector $\vec{\pi}$ and get a new Dirichlet distribution (remember condition C in Ferguson!)
- Suppose:

$$(\pi_1, \pi_2, \ldots, \pi_K) \sim Dirichlet(\alpha_1, \alpha_2, \ldots, \alpha_K)$$

Based on this property for the vector $(\pi_1 + \pi_2, \pi_3, \ldots, \pi_K)$ we can write:

$$(\pi_1 + \pi_2, \pi_3, \ldots, \pi_K) \sim Dirichlet(\alpha_1 + \alpha_2, \alpha_3, \ldots, \alpha_K)$$

Dirichlet Distributions - Decimative Property

- Decimative property is the opposite of the agglomerative property. Consider the probability vector $\vec{\pi}$ that is distributed as Dirichlet:

$$(\pi_1, \pi_2, \ldots, \pi_K) \sim \mathit{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_K)$$

- Now Suppose we want to break π_1 randomly into two pieces. Consider another Dirichlet distribution of the form:

$$(\tau_1, \tau_2) \sim \mathit{Dirichlet}(\alpha_1\beta_1, \alpha_1\beta_2) \text{ where } \beta_1 + \beta_2 = 1$$

- Based on decimative property of Dirichlet distributions, we can then conclude:

$$(\pi_1\tau_1, \pi_2\tau_2, \pi_3, \ldots, \pi_K) \sim \mathit{Dirichlet}(\alpha_1\beta_1, \alpha_1\beta_2, \alpha_3, \ldots, \alpha_K)$$

"Infinite-Dimension" Dirichlet Distribution

- Using the Decimative property of Dirichlet Distributions, we can add "dimensions" to our probability vector $\vec{\pi}$ as:

$$1 \sim Dirichlet(\alpha)$$

$$(\pi_1, \pi_2) \sim Dirichlet(\alpha/2, \alpha/2) \text{ where: } \pi_1 + \pi_2 = 1$$

$$(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}) \sim Dirichlet(\alpha/4, \alpha/4, \alpha/4, \alpha/4) \text{ where: } \pi_{i1} + \pi_{i2} = \pi_i$$

$$\vdots$$

- and we can do it on and on.
- in the process above, at each step we divide each π into two piece (based on a Beta distribution) —> Stick Breaking ?!
- Claim: A **Dirichlet Process (DP)** is "infinitely decimated" Dirichlet distribution.

"Infinite-Dimension" Dirichlet Distribution - Demo

- Nice Demo by Yee Whye Teh (Fork it on Github: https://github.com/probml)
- We already know realizations of DP are discrete almost surely from Sethuraman construction.
- The Demo also "visually" shows why realizations of DP are discrete almost surely.

Dirichlet Processes

- A Dirichlet Process is distribution over probability measures such that marginals on finite partitions are distributed as Dirichlet.
- How do we know such a Distribution exsits? (next slide!)
- Consider $G \sim DP(\alpha, G_0)$. Then for any finite partition of our sample space $\mathcal{X}$ that is of the form $(A_1, \ldots, A_K)$, we have:

$$(G(A_1), \ldots, G(A_K)) \sim Dirichlet(\alpha G_0(A_1), \ldots, \alpha G_0(A_K))$$

- The first two moments of DP for any measurable subset of $\mathcal{X}$ like A is:
  1. $E(G(A)) = G_0(A)$
  2. $Var(G(A)) = \frac{G_0(A)(1-G_0(A))}{\alpha+1}$

Why DP Exists?

- Kolmogorov Consistency Theorem [Ferguson 1973]
- de Finitti's Theorem [Blackwell and MacQueen 1973, Aldous 1985]
- Stick-breaking Construction [Sethuraman 1994]

Marginal and Posterior of DP

- We can show if:

$$\theta | G \sim G$$

$$G \sim DP(\alpha, G_0)$$

Then it implies:

$$\theta \sim G_0$$

$$G|\theta \sim DP(\alpha + 1, \frac{\alpha}{\alpha + 1} G_0 + \frac{1}{\alpha + 1} \delta_\theta)$$

Representations of Dirichlet Processes

- **Blackwell-MacQueen Urn Scheme:**

$$\theta_n|\theta_{1:n-1} \sim \frac{\alpha}{\alpha+n-1}G_0 + \frac{1}{\alpha+n-1}\sum_{i=1}^{n-1}\delta_{\theta_i}$$

- **Chinese Restaurent Process:**

$$P(\text{customer n sat at table K}|...) = \begin{cases} \frac{n_k}{n-1+\alpha} & \text{one of current tables} \\ \frac{\alpha}{n-1+\alpha} & \text{new table} \end{cases}$$

- **Stick-Breaking Construction - Sethuraman:**

$$\pi_k = \beta_k \prod_{i=1}^{k-1}(1-\beta_i); \text{ where: } \beta_k \sim Beta(1,\alpha) \text{ , } \theta_k^* \sim G_0$$

Then G can be written as:

$$G = \sum_{k=1}^{\infty}\pi_k\delta_{\theta_k^*}$$

Application

Here we focus on two applications of Dirichlet processes:

- Density Estimation
- Semi-parametric Modelling

Density Estimation

- A typical parametric density estimation is as follows:

$$\text{Observed Data: } X = \{x_1, x_2, \ldots, x_n\}$$

$$\mathit{Model} : X_i|w \sim F(.|w) \text{ ,F is a parametric distribution}$$

- In Bayesian non-parametric density estimation with Dirichlet Processes we directly put prior on F without any explicit assumption. In other words, our model is:

$$\text{Observed Data: } X = \{x_1, x_2, \ldots, x_n\}$$

$$X_i|F \sim F \text{ , } F \sim DP(\alpha, G_0)$$

Model above does not work due to discreteness of DP!

Density Estimation

- In order to be able to estimate continuous distributions, we can convolve DP with a smooth distribution. This means instead of setting a DP prior on the F distribution, we put a DP prior on the distribution of the **parameters** of a smooth distribution. In other words:

$$X_i|\theta_i \sim^{ind} F(.|\theta_i)$$

$$\theta_i \sim^{iid} G$$

$$G \sim DP(\alpha, G_0)$$

This will lead to $X_i|G \sim F_x$ where:

$$F_x(.) = \int F(.|\theta)dG(\theta) = \sum_{k=1}^{\infty} \pi_k F(.|\theta_k^*)$$

This model is called Dirichlet Process Mixture.

Density Estimation

- Let's consider the predictive density of $f(X_{n+1}|X_1, X_2, \ldots, X_n)$ explained in class one more time:

$$f(X_{n+1}|X_1, \ldots, X_n) = \int (1) * (2) * (3) * d\theta_1 d\theta_2 \ldots d\theta_{n+1}$$

where:

- (1): $f(X_{n+1}|\theta_{n+1})$
- (2): $f(\theta_{n+1}|\theta_1, \ldots, \theta_n, X_1, \ldots, X_n) = f(\theta_{n+1}|\theta_1, \ldots, \theta_n)$
- (3): $f(\theta_1, \ldots, \theta_n|X_1, \ldots, X_n)$ you may call it posterior of DPM!

Difficult to sample 3 !

Semiparametric Models

- Consider a mixed effect model of the form:

$$Y_{ij} = \beta^T X_{ij} + b_i^T Z_{ij} + \epsilon_{ij}$$

- We would like to be able to interpret regression coefficients, β, so we have parameteric assumption for them.
- Model might include other parts which we would like to be as flexible as possible (here ϵ and b_i)
- Intead of having restrictive parametric assumptions on ϵ and b_i (usually they are assumed to be Normally distributed in parameteric setting), we can relax any distributional assumptions by putting a DP prior on them as follows:

$$\epsilon_{ij} \sim F \ ; \ F \sim DP$$

or

$$b_i \sim G \ ; \ G \sim DP$$

Semiparametric Models

- Again, **sampling from the Posterior distribution of a Dirichlet process mixture model** is a challenge!

A Dirichlet Process Mixture Model

- Let's consider $y_1, \ldots, y_n$ being independently drawn from some unknown distribution.
- We can model that unknown distribution as a mixture of distributions of the form $F(.|\theta)$, with θ coming from a mixing distribution, G.
- We put a DP prior on G

$$y_i|\theta_i \sim^{ind} F(.|\theta_i)$$

$$\theta_i|G \sim^{iid} G$$

$$G \sim DP(G_0, \alpha)$$

Goal: Predictive Density

- Our goal is to get predictive density of future Y_{n+1} given the observed data $y_1, \ldots, y_n$.
- Predictive density is of the form:

$$f(Y_{n+1}|y_1, \ldots, y_n) = \int (1) * (2) * (3) * d\theta_1 d\theta_2 \ldots d\theta_{n+1}$$

where:

- (1): $f(Y_{n+1}|\theta_{n+1})$
- (2): $f(\theta_{n+1}|\theta_1, \ldots, \theta_n, y_1, \ldots, y_n) = f(\theta_{n+1}|\theta_1, \ldots, \theta_n)$
- (3): $f(\theta_1, \ldots, \theta_n|y_1, \ldots, y_n)$

We use MCMC methods to do a numerical approximation to the predictive density

Posterior Sampling of a DPM

- Goal is to be able to sample from: $f(\theta_1, \ldots, \theta_n | y_1, \ldots, y_n)$
- we know:
  $p(\theta_1, \ldots, \theta_n | y_1, \ldots, y_n) \propto f(Y_1, \ldots, Y_n | \theta_1, \ldots, \theta_n) p(\theta_1, \ldots, \theta_n)$
- as showed in class, we can repeatedly draw values for each θ_i from it's conditional distribution given the data and other θ's:

$$p(\theta_i | \theta_{(-i)}, Y_1, \ldots, Y_i, \ldots, Y_n) \propto f(Y_i | \theta_i) p(\theta_i | \theta_{(-i)})$$

- $p(\theta_i | \theta_{(-i)}) \sim \frac{\alpha}{n-1+\alpha} G_0 + \frac{1}{n-1+\alpha} \sum_{j \neq i} \delta_{\theta_j}(\theta_i)$ via polya urn scheme
- Combining with likelihood, we get the following conditional distribution:

$$\theta_i | \theta_{(-i)}, y_i \sim r_i H_i + \sum_{j \neq i} q_{i,j} \delta(\theta_j)$$

Posterior Sampling of a DPM

- Combining with likelihood, we get the following conditional distribution:

$$\theta_i|\theta_{(-i)}, y_i \sim r_i H_i + \sum_{j \neq i} q_{i,j}\delta(\theta_j)$$

where:

H_i : posterior distn for θ with G_0 (prior) and likelihood with signle y_i

$$q_{i,j} = bF(y_i, \theta_j)$$

$$r_i = b\alpha \int F(y_i, \theta) dG_0(\theta)$$

$$\text{b is such that: } \sum_{j \neq i} q_{i,j} + r_i = 1$$

Algorithm 1 (Conjugate)

- **when:** G_0 is a conjugate prior for F.
- **How:**
  - state of the Markov chain consists of $(\theta_1, \ldots, \theta_n)$
  - For $i = 1, \ldots, n$: Draw a new value from $\theta_i | \theta_{(-i)}, y_i$
  - Remixing is recommended for faster convergence
- **comment:** Convergence to the posterior is slow (inefficient sampling!)
- Often times there are groups of observations associated with the same θ with high probability. Since the algorithm can't change the θ value for more than one observation, we get the so-called "sticky-cluster" problem.

Demo - Model

- Consider a mixed effect model with a simple Random Intercept, where:

$$\vec{Y_i} = b_0^i + \beta_1 * \vec{T_i} + \vec{e_i}$$

  - m_i: number of measurements for subject i
  - $\vec{Y_i}$: a vector of length m_i of Albumin measures
  - $\vec{T_i}$: a vector of time for subject i
  - β_1: a common covariate for all subjects
  - $\vec{e_i} \sim N_{m_i}(\vec{0}, \Sigma = sigma_\epsilon^2 * diag(m_i))$

Demo - Data Simulation

- nSub = 30
- b_0^{true}: -5 or 0 or 5 - each 10
- $\beta_1 = 1$
- $\sigma_\epsilon^2 = 0.2$
- m_i = Integers from 5-10
- **Priors:**
  - $b^i{}_0 | G \sim G$ where$G \sim DP(\alpha = 1.5, G_0 = N(\mu_0 = 0, \sigma_0 = 15))$
  - $\beta_1 \sim N(\mu_{\beta_1} = 0, \sigma_{\beta_1} = 2)$

Demo - Results

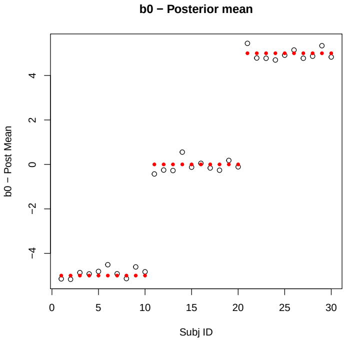

Figure: Posterior mean of b0's (black) v. True Values (red)

Demo - Results

	True Value	Posterior Mean (Beta1)	95% CR (Beta1)	acceptance rate
Beta1	1	1	(0.9918, 1.0093)	0.4895

Table: Random Intercept Demo - Results

Algorithm 2

- Consider a finite mixture model with K components as follows:

$$Y_i|c_i, \vec{\theta^*} \sim F(\theta^*{}_{c_i})$$

$$c_i|\vec{p} \sim Discrete(p_1, \ldots, p_K)$$

$$\theta^*_c \sim G_0$$

$$\vec{p} \sim Dirichlet(\alpha/K, \ldots, \alpha/K)$$

- Corresponding to each Y_i, there is a latent class indicator c_i. It works as an index.
- $\vec{\theta^*}$ is a vector of K different θ values.
- **Claim: $\sum_{i=1}^{K} p_i F(.|\theta^*_i)$ converges to DPM as K goes to inifinity, so the model above is an approximation to DPM!** GoToAlg5

Algorithm 2

- By integrating over the mixing proportions, **p**, in our finite mixture model, we can write:

$$P(c_i = c | c_1, \ldots, c_{i-1}) = \frac{n_{i,c} + \alpha/K}{i - 1 + \alpha}$$

where $n_{i,c} = \sum_{j<i} \delta_{c_j}(c)$

- Now, if we let $K \to \infty$, the conditional probability above (prior for c_i) reaches the following limits:

$$P(c_i = c | c_1, \ldots, c_{i-1}) \to \frac{n_{i,c}}{i - 1 + \alpha}$$

$$P(c_i \neq c_j \text{ for all } j < i | c_1, \ldots, c_{i-1}) \to \frac{\alpha}{i - 1 + \alpha}$$

Algorithm 2

- In a finite setting, the conditional probabilities for c_i is:

$$P(c_i = c | c_{(-i)}, y_i, \vec{\theta^*}) = bF(y_i, \theta^*_c) \frac{n_{-i,c} + \alpha/K}{n - 1 + \alpha}$$

- as $k \to \infty$, $\vec{\theta^*}$ will go to infinite dimension. However, we can do Gibbs sampling on only θ^*'s that are currently associated with at least one observation. So we can write:

If $c = c_j$ for some $j \neq i$: $P(c_i = c | c_{(-i)}, y_i, \vec{\theta^*}) = b \frac{n_{-i,c}}{n - 1 + \alpha} F(y_i, \theta^*_c)$

$$P(c_i \neq c_j \text{ for all } j \neq i | c_{(-i)}, y_i, \vec{\theta^*}) = b \frac{\alpha}{n - 1 + \alpha} \int F(y_i, \theta^*) dG_0(\theta^*)$$

Algorithm 2 - Conjugate

- **Algorithm 2:** Let the state of the Markov chain consist of $\vec{c} = (c_1, \ldots, c_n)$ and $\vec{\theta^*} = (\theta^*{}_c : c \in \{c_1, \ldots, c_n\})$
- For $i = 1, \ldots, n$: Using the formula on last page, draw a value for c_i. c_i is either one of the exisiting ones or if not, draw a new $\theta^*_{c_i}$ from H_i (posterior with prior G_0 and a likelihood based on y_i only).
- do remixing for observations with the same c_i.
- Easy when G_0 is a conjugate prior !

Algorithm 3 - Conjugate

- In algorithm 2 when we have a conjugate G_0, we can analytically integrate over ${\theta^*}_c$.
- In that case, the state of the Markov chain will contain only the indeces c_i's. We then get:

$$\text{If } c = c_j \text{ for some } j \neq i :$$

$$P(c_i = c | c_{(-i)}, y_i, \vec{\theta^*}) = b \frac{n_{-i,c}}{n-1+\alpha} \int F(y_i, \theta^*) dH_{-i,c}(\theta^*)$$

$$P(c_i \neq c_j \text{ for all } j \neq i | c_{(-i)}, y_i, \vec{\theta^*}) = b \frac{\alpha}{n-1+\alpha} \int F(y_i, \theta^*) dG_0(\theta^*)$$

When G_0 is Non-Conjugate

- Algorithms 1 to 3 cannot easily be applied to models where G_0 is not a conjugate prior.
- Perhaps Metropolis-Hasting algorithm is the simplest way to handle non-conjugate priors.
- One idea is to use MH to update c_i's where the conditional prior of c_i's used as the proposal distribution.

Metropolis-Hasting – Review

- Suppose we want to sample for X where X is distributed $\pi(X)$.
- Consider $g(X^*|X)$ as a proposal distribution that proposes a new state (X^*) given our current state X.
- We accept the proposed state X^* with the acceptance probability:

$$a(X^*|X) = min\left[1, \frac{g(X|X^*)}{g(X^*|X)}\frac{\pi(X^*)}{\pi(X)}\right]$$

Algorithm 5 - Non-Conjugate

- We earlier showed in our finite mixture model that the conditional prior for c_i's is:

$$P(c_i = c | c_{(-i)}) = \frac{n_{-i,c} + \alpha/K}{n - 1 + \alpha}$$

  Where $n_{-i,c}$ is the number of $c_j = c$ for $j \neq i$

- Considering the probability above as our proposal distribution (symmetrix), we can compute our acceptance probability as:

Finite Mixture Model

$$a(c_i^*, c_i) = min\left[1, \frac{F(y_i, \theta^*_{c_i^*})}{F(y_i, \theta^*_{c_i})}\right]$$

Algorithm 5 - Non-Conjugate

- Analogous to our Finite mixture model, our conditional prior on c_i's for a DPM model is:

$$\text{If } c = c_j \text{for some j: } P(c_i = c | c_{(-i)}) = \frac{n_{-i,c}}{n - 1 + \alpha}$$

$$P(c_i \neq c_j \text{ for all } j | c_{(-i)}) = \frac{\alpha}{n - 1 + \alpha}$$

- we can use the probability above as our propsal distribution.
- In each step, we may do several MH update.

Algorithm 5 - Non-Conjugate

- For $i = 1, \ldots, n$, repeat the following update of c_i, R times:
- Draw a candidate c_i^* from the condtional prior of c_i
- if $c_i^* \notin c1, \ldots, c_n$, sample a value for $\theta^*_{c_i^*}$ from G_0 and accept the new value of c_i^* and it's corresponding θ^* with probability $a(c_i^*, c_i)$.
- do remixing

Algorithm 7 - Non-Conjugate

- The MH method in Algorithm 5 is more likely to consider changing c_i to a component associated with many observations than a component associated with few observations.
- Creating a new component is proportional to α. In general we know that the probability of making a new component depends on α but in our MH case and by considering that in practice α is usually small (around 1), the issue is such a change might not even be considered in this algorithm.
- A new algorithm with a desire to create a new component more often might be more efficient. To do so, we need to modify our proposal distribution.

Algorithm 7 - Non-Conjugate

- State of the Markov chain: $\vec{c} = (c_1, \ldots, c_n)$ and $vec\theta^* = (\theta_c^* : c \in \{c_1, \ldots, c_n\})$
- For $i = 1, \ldots, n$, update c_i as follows:
- If c_i is not a singleton ($c_i = c_j$ for some $j \neq i$), let c_i^* be a newly created component with a $\theta^*_{c_i^*}$ drawn from G_0. Accept this new c_i^* with probability:

$$a(c_i^*, c_i) = min\left[1, \frac{\alpha}{n-1} \frac{F(y_i, \theta^*_{c_i^*})}{F(y_i, \theta^*_{c_i})}\right]$$

- Otherwise, if c_i is a singleton, draw c_i^* from $c_{(-i)}$ with probability $\frac{n_{-i,c}}{n-1}$ for $c_i^* = c$ and accept the new c_i^* with probability:

$$a(c_i^*, c_i) = min\left[1, \frac{\alpha}{n-1} \frac{F(y_i, \theta^*_{c_i^*})}{F(y_i, \theta^*_{c_i})}\right]$$

Algorithm 7 - Non-Conjugate

- For $i = 1, \ldots, n$: If c_i is not a singleton, choose a new value for c_i from $\{c_1, \ldots, c_n\}$ using the following probability:

$$P(c_i = c | c_{(-i)}, y_i, \vec{\theta^*}, c_i \in c_1, \ldots, c_n) = b \frac{n_{-i,c}}{n-1} F(y_i, \theta_c^*)$$

- do remixng.

Algorithm 8 - Non-Conjugate

- Algorithm 8 handles models with non-conjugate priors by applying Gibbs sampling to an extended state with some auxiliary parameters.
- **Idea:**
  - Suppose we are interested in sampling for X from the distribution π_x.
  - We can sample from π_x by sampling from π_{xy} with the marginal distribution of π_x.
- Now consider a Markov chain with the permanent state of X and with some auxiliary variables introduced temprorarily during an update of the following form:
  1. Draw a value for y from $\pi_{Y|X}$
  2. Perform some update of (x,y) that leaves π_{xy} invariant.
  3. Discard y and only keep x value.
- Claim: As long as π_x is the marginal distribution of π_{xy}, this update leaves x invariant.

Algorithm 8 - Non-Conjugate

- **Permanent state of the Markov Chain:** includes $\vec{c}$ and $\vec{\theta^*}$ as in algorithm 2.
- when c_i is updated, temporary auxiliary variables are introduced.
- Temporary auxiliary variables represent possible values for the parameters that are not currently associated with any observation.
- We update c_i by Gibbs sampling and from a pool of current c_j's and the temporary auxiliary parameters. c_j for other parameters ($j \neq i$) is in the set $\{1, \ldots, k^-\}$ where k^- is the number of distinct $c_j, j \neq i$.

Algorithm 8 - Non-Conjugate

- The conditional prior distribution for c_i given other c_j and our auxiliary variables (m of them) is:
  - choose one of the existing $c \in \{1, \ldots, k^-\}$ with probability $\frac{n_{-i,c}}{n-1+\alpha}$ $n_{-i,c}$: frequency of in $c_j, j \neq i$
  - or choose an auxiliary variable with prob $\frac{\alpha}{n-1+\alpha}$ that is equally distributed.

Algorithm 8 - Non-Conjugate

- **State of the Markov chain:** $\vec{c} = c(c_1, \ldots, c_n)$ and $\vec{\theta^*} = (\theta^*_c : c \in \{c_1, \ldots, c_n\})$
- For $i = 1, \ldots, n$ k^- is the number of distinct c_j for $j \neq i$ and define $h = k^- + m$.
- Label these c_j's with values in $\{1, \ldots, k^-\}$
- if $c_i = c_j$ for some $j \neq i$ then draw m independent samples from G_0 for the auxiliary variables.
- if $c_i \neq c_j$ for some $j \neq i$ then draw m - 1 independent samples from G_0 for the auxiliary variables and use c_i as one of the auxiliary variables.
- Now we have a pool of h different c_i values and their corresponding $\theta*$ values.

Algorithm 8 - Non-Conjugate

- We then draw a value for c_i from $\{1, \ldots, h\}$:

$$P(c_i = c | c_{(-i)}, y_i, \theta_1^*, \ldots, \theta_n^*) =$$

  - for $1 \leq c \leq k^-$:

$$b \frac{n_{-i,c}}{n-1+\alpha} F(y_i, \theta_c^*)$$

  - for $k^- < c \leq h$:

$$b \frac{\alpha/m}{n-1+\alpha} F(y_i, \theta_c^*)$$

  Where $n_{-i,c}$ is the number of $c_j = c$ for $j \neq i$

  - Throw away all θ_c^* that are not associated with any subject.
- Do remixing !

References

- Neal, Radford M. "Markov chain sampling methods for Dirichlet process mixture models." Journal of computational and graphical statistics 9.2 (2000): 249-265.
- Antoniak, Charles E. "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems." The annals of statistics 2.6 (1974): 1152-1174.
- Ferguson, Thomas S. "A Bayesian analysis of some nonparametric problems." The annals of statistics (1973): 209-230.
- Sethuraman, Jayaram. A constructive definition of Dirichlet priors. No. FSU-TR-M-843. FLORIDA STATE UNIV TALLAHASSEE DEPT OF STATISTICS, 1991.
- Teh, Yee Whye. "Dirichlet processes: Tutorial and practical course." Machine Learning Summer School (2007).