Bayesian Non Parametric and its Inference

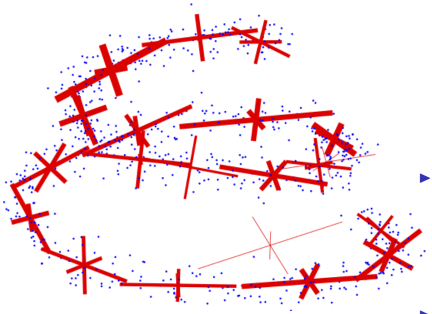
A/Prof Richard Yi Da Xu
Yida.Xu@uts.edu.au
Wechat: aubedata
https://github.com/roboticcam/machine-learning-notes

University of Technology Sydney (UTS)

February 18, 2018

Dirichlet Process: A diagrammatic representation

Rasmussen, Infinite Gaussian Mixture Model (1999):



- For a mixture model:
  Let $\mathbf{X} = x_1, \dots, x_N$:

$$P(\mathbf{X}|\theta_1, \dots \theta_K, w_1, \dots w_K) = \sum_{l=1}^{K} w_l f(\mathbf{X}|\theta_l)$$

$$\text{where } \sum_{l=1}^{K} w_l = 1$$

- If we allow K to also vary, what happens if you want to:

$$\underset{\theta_1, \dots \theta_K, w_1, \dots, w_K, K}{\arg\max} P(\mathbf{X}|\theta_1, \dots \theta_K, w_1, \dots w_K, K)?$$

- K = N for Gaussian case. Of course it's not desirable!

Dirichlet Process: Motivation

- For data $x_1, \ldots, x_N$, each x_i is associating with a parameter θ_i
- We need to a good prior for $\Pr(\theta_1 \ldots \theta_N)$:
- You also want K potentially be infinite
- A "clustering" property, controllable through a single parameter α
- Let's define it using Hierarchical prior, its marginal is:

$$p(\theta_1, \ldots \theta_n) = \int_G \Pr(\theta_1, \ldots, \theta_n | G) \mathbf{p}(\mathbf{G})$$

So, we are interested in the property of G:

- G needs to be **discrete** random distribution.
- Perhaps it should also some resemblence with some basic distribution H.

Dirichlet Process Definition

- We say G is a Dirichlet process, distributed with base distribution H and concentration parameter α:

$$G \sim DP(\alpha, H), \text{if}$$
$$(G(A1), ..., G(Ar)) \sim \text{Dir}(\alpha H(A1), ..., \alpha H(Ar))$$

- for every finite measurable partition $A_1, ..., Ar$ of Θ.
- What does this all mean? Let's visualise it!
- **note** $(A_1 \cup A_2 \cup \cdots \cup A_r) \subseteq \Omega$, this can be seen from the fact that:

$$\begin{aligned}
&(x_1, \ldots, x_k, \ldots, x_K) \sim \text{Dir}(\alpha_1, \ldots, \alpha_k, \ldots, \alpha_K) \\
\implies &\left(\frac{x_1}{1 - x_k}, \ldots, \frac{x_{k-1}}{1 - x_k}, \frac{x_{k+1}}{1 - x_k}, \ldots, \frac{x_K}{1 - x_k}\right) \sim \text{Dir}(\alpha_1, \alpha_{k-1}, \alpha_{k+1}, \alpha_K)
\end{aligned}$$

You need both the posterior and predictive distribution of Multinomial-Dirichlet:

Posterior

$$
\begin{aligned}
&P(p_1, \ldots, p_k | n_1, \ldots, n_k) \\
&\propto \underbrace{\frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k p_i^{\alpha_i - 1}}_{\text{Dir}(p_1,\ldots,p_k|\alpha_1,\ldots,\alpha_k)} \underbrace{\frac{n!}{n_1! \ldots n_k!} \prod_{i=1}^k p_i^{n_i}}_{\text{Mult}(n_1,\ldots,n_k|p_1,\ldots p_k)} \\
&\propto \prod_{i=1}^k p_i^{\alpha_i - 1} \prod_{i=1}^k p_i^{n_i} = \prod_{i=1}^k p_i^{\alpha_i - 1 + n_i} \\
&= \text{Dir}(p_1, \ldots p_k | \alpha_i + n_i, \ldots \alpha_k + n_k)
\end{aligned}
$$

Marginal

$$
\begin{aligned}
p(n_1, \ldots n_k) &= \int_{p_1,\ldots,p_k} P(p_1, \ldots, p_k, n_1, \ldots, n_k) \\
&= \frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\prod_{i=1}^k \Gamma(\alpha_i)} \frac{n!}{n_1! \ldots n_k!} \int_{p_1,\ldots,p_k} \prod_{i=1}^k p_i^{\alpha_i - 1 + n_i} \\
&= \frac{N!}{n_1! \ldots n_k!} \times \frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\prod_{i=1}^k \Gamma(\alpha_i)} \times \frac{\prod_{i=1}^k \Gamma(\alpha_i + n_i)}{\Gamma\left(N + \sum_{i=1}^k \alpha_i\right)}
\end{aligned}
$$

Expectation

- for any measurable set $A_i \in \Omega$: we have $\mathbb{E}[G(A_i)] = H(A_i)$, why?
- for a dirichlet distribution:

$$f(x_1, \ldots, x_K | \alpha_1, \ldots, \alpha_K) = \frac{\Gamma\left(\sum_{i=1}^K \alpha_i\right)}{\prod_{i=1}^K \Gamma(\alpha_i)} \prod_{i=1}^K x_i^{\alpha_i - 1}$$

- the expectation: $E[X_i] = \frac{\alpha_i}{\sum_k \alpha_k}$
- Therefore:

$$\mathbb{E}[G(A_i)] = \frac{\alpha H(A_i)}{\sum_i \alpha H(A_i)} = \frac{\alpha H(A_i)}{\alpha \sum_i H(A_i)} = H(A_i)$$

- note that the expectation is **independent of** α

Variances

- Variances for Dirichlet Distribution:

$$\mathbb{VAR}[X_i] = \frac{\alpha_i \left(\left(\sum_i^K \alpha_{i=1} \right) - \alpha_i \right)}{\left(\sum_i^K \alpha_{i=1} \right)^2 \left(\sum_i^K \alpha_{i=1} + 1 \right)}$$

- substitute $\alpha \to \alpha H(A_i)$:

$$\begin{aligned} \mathbb{VAR}(G(A_i)) &= \frac{\alpha H(A_i) \left(\alpha - \alpha H(A_i) \right)}{\alpha^2 \left(\alpha + 1 \right)} \\ &= \frac{H(A_i) \left(1 - H(A_i) \right)}{\left(\alpha + 1 \right)} \end{aligned}$$

- when $\alpha = 0$:

$$\mathbb{VAR}(G(A_i)_{\alpha=0}) = H(A_i)(1 - H(A_i))$$

Posterior

- from **multinomial-dirichlet conjugacy**, we have:

$$G' = G(A_1), \dots, G(A_r) | \theta_1, \dots, \theta_n \sim \text{Dir}(\alpha H(A_1) + n_1, \dots, \alpha H(A_k) + n_k)$$

- DP provides a conjugate family of priors over distributions that is **closed** under posterior updates given observations:

$$G' \sim \text{DP}\left(\alpha + n, \frac{\alpha H + \sum_{i=1}^n \delta_{\theta_i}}{\alpha + n}\right), \text{ or}$$

$$G' \sim \text{DP}\left(\alpha + n, \frac{\alpha}{\alpha + n} H + \frac{\sum_{i=1}^n \delta_{\theta_i}}{\alpha + n}\right)$$

- another way of specifying this is:

$$G_u \sim \text{DP}(\alpha, H) \qquad G' = \frac{1}{\alpha + n} \sum_{i=1}^n \delta_{\theta_i} + \frac{\alpha}{\alpha + n} G_u$$

In words: posterior of $\text{DP}(\alpha, H)$ is to **squash** $\text{DP}(\alpha, H)$ to a total mass of $\frac{\alpha}{\alpha+n}$ remaining mass was assigned to discrete points $\sum_{i=1}^n \delta_{\theta_i}$.

Predictive

- Let $P(\theta_{n+1} \in A | G) = G(A)$:

$$\begin{aligned} P(\theta_{n+1} \in A | \theta_1, \dots, \theta_n) &= \int_G P(\theta_{n+1} \in A | G) P(G | \theta_1, \dots, \theta_n) dG \\ &= \mathbb{E}(G(A) | \theta_1, \dots, \theta_n) \\ &= \mathbb{E}(G'(A)) \end{aligned}$$

- We know that:

$$\mathbb{E}(G(A)) = H(A) \implies \mathbb{E}(G'(A)) = \frac{\alpha}{\alpha + n} H(A) + \frac{\sum_{i=1}^{n} \delta_{\theta_i}}{\alpha + n}$$

Stick-Breaking construction

- $v_k \sim \text{Beta}(1, \alpha)$
- $\pi_k = v_k \prod_{l=1}^{k-1}(1 - v_l)$
- $\theta_k \sim H$
- $G_0 = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$

posterior sampling of π

- $v_k \sim \text{Beta}(1, \alpha)$
- $\pi_k = v_k \prod_{l=1}^{k-1}(1 - v_l)$
- given samples $\theta_1, \dots, \theta_N$ with k distinct values having $n_1, \dots, n_K$ counts

$$
\begin{aligned}
G' &= G(A_1), \dots, G(A_K) | \theta_1, \dots, \theta_n \\
&\sim \text{Dir}(\alpha H(A_1) + n_1, \dots, \alpha H(A_K) + n_k) \\
&\sim \text{Dir}\Big(\delta_{\theta_1 \in B_1} n_1, \dots, \delta_{\theta_K \in B_K} n_K, \alpha H\big(\Omega \setminus \{\text{d}B_1, \dots \text{d}B_K\}_{\|\text{d}B_k\| \to 0 \ \forall k}\big)
\end{aligned}
$$

$$
\implies (\pi_1, \dots, \pi_k, \pi_u) \sim \text{Dir}(n_1, n_2, \dots n_K, \alpha)
$$

- where π_u are all the probability mass assign to $\theta_{K+1}, \dots, \theta\infty$

Predictive

Let $\alpha_i = \frac{\alpha}{k}$: compute the density of i^{th} data belonging to existing component m.

$$
\begin{aligned}
\Pr(z_i = m|\mathbf{z}_{-1}) &= \int_{p_1,\dots,p_k} P(z_i = m|p_1,\dots,p_k)P(p_1,\dots,p_k|n_{1,-i},\dots,n_{k,-i}) \\
&= \frac{\int_{p_1,\dots,p_k} P(z_i = m|p_1,\dots,p_K)P(n_{1,-i},\dots,n_{k,-i}|p_1,\dots,p_K)P(p_1,\dots,p_K)}{P(n_{1,-i},\dots,n_{k,-i})} \\
&= \frac{\int_{p_1,\dots,p_k} P(z_i = m|p_1,\dots,p_K)P(n_{1,-i},\dots,n_{k,-i}|p_1,\dots,p_K)P(p_1,\dots,p_K)}{\int_{p_1,\dots,p_K} P(n_1^{-i},\dots,n_K^{-i}|p_1,\dots,p_K)P(p_1,\dots,p_K)} \\
&= \frac{\Gamma(\frac{\alpha}{k} + n_{m,-i} + 1)\prod_{l=1,l\neq m}^{k}\Gamma(\frac{\alpha}{k} + n_{l,-i})}{\Gamma(N+\alpha)} \times \frac{\Gamma(N-1+\alpha)}{\prod_{l=1}^{k}\Gamma(\frac{\alpha}{k} + n_{l,-1})} \\
&= \frac{\frac{\alpha}{k} + n_{m,-i}}{N+\alpha-1} \qquad \text{Let } k \to \infty = \frac{n_{m,-i}}{N+\alpha-1}
\end{aligned}
\tag{1}
$$

$\Pr(z_i = \text{new}) = \frac{\alpha}{N+\alpha-1}$.

DP Sampling algorithm

$$\Pr(z_i = m|\mathbf{z}_{-i}, \alpha) \propto \begin{cases} \frac{n_{m,-i}}{N+\alpha-1} & \text{for existing cluster } m \\ \frac{\alpha}{N+\alpha-1} & \text{for new cluster} \end{cases}$$

- **exercise** to write a Gibbs Sampling algorithm for above
- **homework** what is the joint density of $\Pr(z_1, \ldots z_N)$

- Using the following relations:

$$\psi(x+N) - \psi(x) = \sum_{k=0}^{N-1} \frac{1}{x+k}$$

- we know each i^{th} **new** person has $\frac{1}{\alpha+i}$ probability of occupying a new table:
- the probability of new table is **independent** of the existing seating arrangement:

$$\mathbb{E}(\text{\# of occupied tables}) = \sum_{k=0}^{N-1} \frac{\alpha}{\alpha+k} = \alpha\big(\psi(\alpha+N) - \psi(\alpha)\big)$$

$$\text{where } \psi(x) = \frac{d}{dx} \ln\big(\Gamma(x)\big) = \frac{\Gamma'(x)}{\Gamma(x)}$$

- **Homework** to also prove:

$$\mathbb{VAR}(\text{\# of occupied tables}) = \alpha\Big(\psi(\alpha+n) - \psi(\alpha)\Big) + \alpha^2\big(\psi'(\alpha+n) - \psi'(\alpha)\big)$$

probability of the number of tables

- number of times of sitting at **new** tables dictates k
- say if we are interested in $\Pr(k = 3)$: persons $\{1, 2, 4\}$ or $\{1, 6, 9\}$ can be the **first in a new table**
- what are the combinations (i.e, coefficient) for each k?

$$
\begin{aligned}
A_n(\alpha) &= \frac{(\overbrace{\alpha}^{\text{new}} + \overbrace{0}^{\text{old}})(\overbrace{\alpha}^{\text{new}} + \overbrace{1}^{\text{old}})\dots(\overbrace{\alpha}^{\text{new}} + \overbrace{n-1}^{\text{old}})}{\underbrace{(\alpha+0)(\alpha+1)(\alpha+2)(\alpha+3)\dots}_{\text{same}}} \\
&= \frac{\left[{n \atop 1}\right]\alpha + \left[{n \atop 2}\right]\alpha^2 + \dots \left[{n \atop n}\right]\alpha^n}{(\alpha+0)(\alpha+1)(\alpha+2)(\alpha+3)\dots} \\
&= \frac{\Gamma(\alpha)}{\Gamma(\alpha+n)}\left(\left[{n \atop 1}\right]\alpha + \left[{n \atop 2}\right]\alpha^2 + \dots \left[{n \atop n}\right]\alpha^n\right)
\end{aligned}
$$

- remove the denominator (which is a constant), we have $\Pr(\# = k) \propto \left[{n \atop k}\right]\alpha^k$
- $\left[{n \atop k}\right]$ is called **stirling number of the first kind**

- in binomial expansion:

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$$

$$(y=1) \implies (x+1)^n = \sum_{k=0}^{n} \binom{n}{k} x^k \qquad \text{there is no } y$$

- However, instead of $(x+1)^n$:

$$(x+0)(x+1)(x+2)\dots(x+n) = \sum_{k=0}^{n} \begin{bmatrix} n \\ k \end{bmatrix} x^k$$

- $\begin{bmatrix} n \\ k \end{bmatrix}$ is called **stirling number of the first kind**

Slice sampling for Dirichlet Process

- an infinite mixture density (e.g. Gaussian) can be written as:

$$f_{\pi,\theta}(y) = \sum_{j=1}^{\infty} \pi_{j=1} \mathcal{N}(y|\theta_j) \qquad \text{where } \theta = (\mu, \sigma^2)$$

- adding slice variable u:

$$f_{\pi,\theta}(y, u) = \sum_{j=1}^{\infty} \mathbf{1}(u < \pi_j) \mathcal{N}(y|\theta_j)$$

- to ensure **marginal is invariant**:

$$\begin{aligned}
\int f_{\pi,\theta}(y, u) \mathrm{d}u &= \int_0^{\pi_j} \sum_{j=1} \mathbf{1}(u < \pi_j) \mathcal{N}(y|\theta_j) \mathrm{d}u \\
&= \sum_{j=1}^{\infty} \mathcal{N}(y|\theta_j) \int_0^{\pi_j} \mathbf{1}(u < \pi_j) \mathrm{d}u \\
&= \sum_{j=1}^{\infty} \mathcal{N}(y|\theta_j) \times \pi_j \\
&= f_{\pi,\theta}(y)
\end{aligned}$$

- note this is in the **absence of latent variable** z_i (later slides)

Slice variable u

finite model: $$P(y|\pi,\theta) = \frac{1}{K}\sum_{j\in\{1\ldots K\}} \mathcal{N}(y|\theta_j)$$

infinite model: $$P(y|\pi,\theta,u) \equiv f_{\pi,\theta}(y|u) = \underbrace{\frac{1}{\#\{A_\pi(u)\}}}_{f_\pi(u)} \sum_{j\in A_\pi(u)} \mathcal{N}(y|\theta_j) = \frac{1}{f_\pi(u)} \sum_{j\in A_\pi(u)} \mathcal{N}(y|\theta_j)$$

- $f_\pi(u)$ is a **a random integer**

$$\begin{aligned} f_\pi(u) &= \sum_{j=0}^{\infty} \mathbf{1}(u < \pi_j) \\ &= \sum_{j=0}^{\infty} \pi_j \mathcal{U}(u|0,\pi_j) \qquad \text{where } \mathcal{U}(u|0,\pi_j) = \begin{cases} \frac{1}{\pi_j}, & u < \pi_j \\ 0, & u > \pi_j \end{cases} \end{aligned}$$

Latent variable z_i

- latent variable z identify the component which y is to be taken:

$$f_{\pi,\theta}(u, z, y) = \mathcal{N}(y|\theta_z)\mathbf{1}(z \in A(u))$$

- for example, $u_6 = 0.15$ and $A(u_6) = \{2, 4, 5, 6\}, k_6 = 4 \in A(u_6) \implies \pi_4 > 0.15$
- If there are n samples, complete data likelihood:

$$\mathcal{L}_{\pi,\theta}(\{y_i, u_i, z_i\}_{i=1}^n) = \prod_{i=1}^n \mathcal{N}(y_i|\theta_{z_i})\mathbf{1}(u_i < \pi_{z_i})$$

1. $u_i \sim U(0, \pi_{z_i})$

2. $f(\theta_j | \cdots) \propto H(\theta_j) \prod_{z_i = j} \mathcal{N}(y_i | \theta_j)$
   If there are no $z_i = j$, then $f(\theta_j | \cdots) = H(\theta_j)$

3. $f(v | \cdots) \propto \pi(v) \prod_{i=1}^{n} \mathbf{1}(\pi_{z_i} > u_i)$

$$f(v | \cdots) \propto \pi(v) \prod_{i=1}^{n} \mathbf{1}(\pi_{z_i} > u_i) = \pi(v) \prod_{i=1}^{n} \mathbf{1}\Big(\underbrace{v_{z_i} \prod_{l < z_i} (1 - v_l)}_{\pi_{z_i}} > u_i \Big)$$

$$= \underbrace{\pi(v)}_{\text{beta}(1,\alpha)} \prod_{i=1}^{n} \mathbf{1} \underbrace{\Big(v_{z_i} \prod_{l < z_i} (1 - v_l) > u_i \Big)}_{\gamma_j < v_j < \beta_j}$$

- the above only applies when $j \leq z^*$, where z^* is the maximum of $\{z_1, \ldots, z_n\}$
- for γ_j and β_j must be a function of u_i and α
- for $j > z^*$, $f(v_j | \cdots) = \text{beta}(1, \alpha)$

lower bound γ_j

$$f(v|\cdots) = \underbrace{\pi(v)}_{\text{beta}(1,\alpha)} \prod_{i=1}^{n} \mathbf{1}\left(\underbrace{v_{k_i} \prod_{l<z_i} (1 - v_l) > u_i}_{\gamma_j < v_j < \beta_j} \right)$$

- **lower bound** means how **low** you can reduce v_j to
- **reduce** $v_j \implies$ **reduce** π_j
- therefore, one needs to ensure all: $\{\pi_{z_i=j}\} > u_i$:

$$\begin{aligned}
v_{z_i} \prod_{l<z_i} (1 - v_l) &> \max_{\{i:z_i=j\}} (u_i) \\
\implies v_{z_i} &> \frac{\max_{\{z_i=j\}} (u_i)}{\prod_{l<z_i} (1 - v_l)} \\
\implies v_{z_i} &> \underbrace{\max_{\{z_i=j\}} \left(\frac{u_i}{\prod_{l<z_i} (1 - v_l)} \right)}_{\gamma_j}
\end{aligned}$$

- $\pi_{j+1}, \pi_{j+2}, \ldots$ will **increase**: there is more to share now - but not affected by lower bound
- $\pi_1, \ldots, \pi_{j-1}$ will **not** be affected

upper bound β_j

$$f(v|\cdots) = \underbrace{\pi(v)}_{\text{beta}(1,\alpha)} \prod_{i=1}^{n} \mathbf{1}\underbrace{\left(v_{z_i} \prod_{l<z_i} (1-v_l) > u_i \right)}_{\gamma_j < v_j < \beta_j}$$

- **increase** $v_j \implies$ **increase** $\pi_j \implies$ **reduce** $\pi_{j+1}, \pi_{j+2}, \dots$
- therefore, one needs to ensure all: $\{\pi_{k_i > j}\} > u_i$
- as an **illustrative example**, we let ($j = 3$) and a particular ($z_i = 5$):

$$\begin{aligned}
&\pi_{z_i=5} > u_i \\
&\qquad \implies (1-v_1)(1-v_2)\mathbf{(1-v_3)}(1-v_4)v_5 > u_i \\
&\qquad \implies (1-v_1)(1-v_2)(1-v_4)v_5 - \mathbf{v_3}(1-v_1)(1-v_2)(1-v_4)v_5 > u_i \\
&\qquad \implies v_3(1-v_1)(1-v_2)(1-v_4)v_5 < (1-v_1)(1-v_2)(1-v_4)v_5 - u_i \\
&\qquad \implies v_3 < 1 - \frac{u_i}{(1-v_1)(1-v_2)(1-v_4)v_5}
\end{aligned}$$

- however, one needs to ensure v_3 (or v_j in general) satisfies: $\{\forall\ z_i > j\}$, write it generally:

$$\begin{aligned}
& v_j < \min_{\{z_i > j\}} \left(1 - \frac{u_i}{v_{z_j} \prod_{l<z_i, l\neq j} (1-v_l)} \right) \\
\implies & v_j < \underbrace{1 - \max_{\{z_i > j\}} \left(\frac{u_i}{v_{z_j} \prod_{l<z_i, l\neq j} (1-v_l)} \right)}_{\beta_j}
\end{aligned}$$

- $\pi_1, \dots, \pi_{j-1}$ and π_j will **not** be affected

sampling via inverse CDF of v_j

- We can define the **truncated** CDF distribuiton of v:

$$F(v) = \frac{1}{C}\int_{\gamma_j}^{v} f(v|\cdots)\mathrm{d}v$$
$$= \frac{\int_0^v \mathrm{beta}(v|1,\alpha)\mathbf{1}(\gamma_j < v < \beta_j)\mathrm{d}v}{\int_0^1 \mathrm{beta}(v|1,\alpha)\mathbf{1}(\gamma_j < v < \beta_j)\mathrm{d}v} = \frac{\int_{\gamma_j}^{v} \mathrm{beta}(v|1,\alpha)\mathrm{d}v}{\int_{\gamma_j}^{\beta_j} \mathrm{beta}(v|1,\alpha)\mathrm{d}v}$$

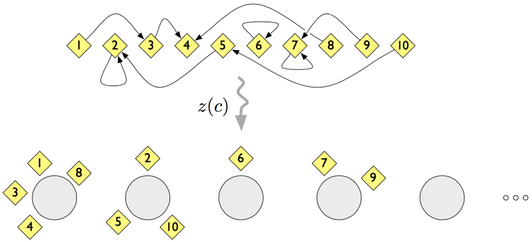
- looking at the property of beta distribution:

$$\int_{\gamma_j}^{v_j} \mathrm{beta}(v|1,\alpha)\mathrm{d}v = \int_{\gamma_j}^{v_j} \frac{\Gamma(1+\alpha)}{\Gamma(1)\Gamma(\alpha)} v^{1-1}(1-v)^{\alpha-1}\mathrm{d}v$$
$$= \alpha\int_{\gamma_j}^{v_j}(1-v)^{\alpha-1}\mathrm{d}v$$
$$= (1-\gamma_j)^{\alpha} - (1-v_j)^{\alpha}$$

- So, we can prove that:

$$F(v_j) = \frac{(1-\gamma_j)^{\alpha} - (1-v_j)^{\alpha}}{(1-\gamma_j)^{\alpha} - (1-\beta_j)^{\alpha}}$$

- this is where **inverse CDF** becomes useful



- instead of sample class variable for nodes, it samples links:

$$\Pr(c_i = j | D, \alpha) \propto \begin{cases} f(d_{ij}) & \text{if } j \neq i \\ \alpha & \text{else} \end{cases}$$

- MATLAB code download:
`http://www-staff.it.uts.edu.au/~ydxu/software1.htm`

Some Extensions to DP

- **Hierarchical Dirichlet Process (HDP)**
- HDP-Hidden Marko Model
- Indian Buffet Process

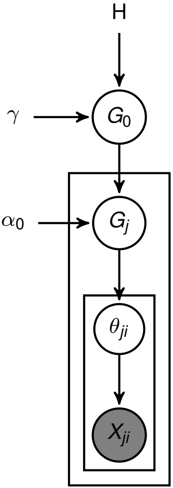
Hierarchical Dirichlet Process (HDP)

Generative model

$$
\begin{aligned}
G_0 &\sim \mathrm{DP}(\gamma, H) \\
G_j &\sim \mathrm{DP}(\alpha_0, G_0) \\
\theta_{ji} &\sim G_j \\
X_{ji} &\sim F(x|\theta_{ij})
\end{aligned}
$$

- Drawing $G_0 \sim \mathrm{DP}(.)$ can be done using stick breaking process, i.e., $\sim \mathrm{Beta}(1, \gamma)$.
- What about stick breaking construction for G_j?
- Certainly, it's NOT $\sim \mathrm{Beta}(1, \alpha_0)$

Graphical model



Generative model

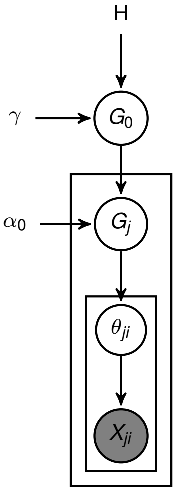
$$\boldsymbol{\beta} \sim \text{GEM}(\gamma) \quad G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}$$

$$\boldsymbol{\pi}_j \sim \text{DP}(\alpha_0, \boldsymbol{\beta}) \qquad G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_k}$$

$$z_{ji} \sim \pi_j \qquad \phi_k \sim H \qquad X_{ji} \sim F(x|\phi_{z_{ji}})$$

- Using $\boldsymbol{\beta}$ as a base, discrete distribution define on range $\{0 \ldots \infty\}$.

Graphical model



New Stick breaking for π_{jk} using β

- Dirichlet Process:

$$v_k \sim \text{Beta}(1, \alpha) \qquad \pi_k = v_k \prod_{l=1}^{k-1} (1 - v_l)$$

$$\theta_k \sim H \qquad G_0 = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$$

- Hierarchical Dirichlet Process:

$$v_{jk} = \frac{\pi_k}{1 - \sum_{l=1}^{k-1} \pi_l} \sim \text{Beta}\left(\alpha\beta_k, 1 - \sum_{l=1}^{k} \beta_l\right) \qquad \pi_{jk} = v_{jk} \prod_{l=1}^{k-1} (1 - v_{jl})$$

- In DP, each v_k is distributed iid from Beta(1α)
- In HDP, each v_{jk} is distributed independently, but having different distribution

Suppose $\boldsymbol{\beta}|\gamma \sim \text{GEM}(\gamma)$ and $\pi|\alpha, \beta \sim \text{DP}(\alpha, \boldsymbol{\beta})$. Notice that the support is $\{1, \dots, k, \dots, \infty\}$:

$$
\begin{aligned}
&(G_j(A_1), \dots, G_j(A_r)) \sim \text{Dir}\left(\alpha G_0(A_1), \dots, \alpha G_0(A_r)\right) \\
\implies & \left(\sum_{k \in K_1} u_k, \dots, \sum_{k \in K_r} u_k\right) \sim \text{Dir}\left(\alpha \sum_{k \in K_1} \beta_k, \dots, \alpha \sum_{k \in K_r} \beta_k\right) \\
\implies & \left(\sum_{l=1}^{k-1} u_l, u_k, \sum_{l=k+1}^{\infty} u_l\right) \sim \text{Dir}\left(\alpha \sum_{l \in 1}^{k-1} \beta_l, \alpha\beta_k, \sum_{l=k+1}^{\infty} \beta_l\right) \\
\implies & \left(\frac{u_k}{1 - \sum_{l=1}^{k-1} u_l}, \frac{\sum_{l=k+1}^{\infty} u_l}{1 - \sum_{l=1}^{k-1} u_l}\right) \sim \text{Dir}\left(\alpha\beta_k, \sum_{l=k+1}^{\infty} \beta_l\right) \qquad \textbf{exercise} \text{ prove this} \\
\implies & \left(\frac{u_k}{1 - \sum_{l=1}^{k-1} u_l}, \frac{\sum_{l=k+1}^{\infty} u_l}{1 - \sum_{l=1}^{k-1} u_l}\right) \sim \text{Dir}\left(\alpha\beta_k, 1 - \sum_{l=1}^{k} \beta_l\right) \\
\implies & \left(v = \frac{u_k}{1 - \sum_{l=1}^{k-1} u_l}\right) \sim \text{Beta}\left(\alpha\beta_k, 1 - \sum_{l=1}^{k} \beta_l\right)
\end{aligned}
$$

Additional proof (1)

$$\left(\sum_{l=1}^{k-1} u_l, u_k, \sum_{l=k+1}^{\infty} u_l\right) \sim \text{Dir}\left(\alpha \sum_{l\in 1}^{k-1} \beta_l, \alpha\beta_k, \sum_{l=k+1}^{\infty} \beta_l\right)$$

$$\implies \left(\frac{u_k}{1-\sum_{l=1}^{k-1} u_l}, \frac{\sum_{l=k+1}^{\infty} u_l}{1-\sum_{l=1}^{k-1} u_l}\right) \sim \text{Dir}\left(\alpha\beta_k, \sum_{l=k+1}^{\infty} \beta_l\right)$$

Additional proof (2)

Let $g_i \sim \text{Gamma}(\alpha_i, 1)$ for $i = 1, \ldots, n$:

$$\left(\frac{g_1}{\sum_{i=1}^n g_i}, \ldots, \frac{g_n}{\sum_{i=1}^n g_i} \right) \sim \text{DIR}(\alpha_1, \alpha_2, \ldots \alpha_n)$$

The following is also true:

$$\left(\frac{g_2}{\sum_{i=2}^n g_i}, \ldots, \frac{g_n}{\sum_{i=2}^n g_i} \right) \sim \text{Dirichlet}(\alpha_2, \ldots \alpha_n)$$

Look at a particular term:

$$\frac{g_j}{\sum_{i=2}^n g_i} = \frac{\frac{g_j}{\sum_{i=1}^n g_i}}{\frac{\sum_{i=2}^n g_i}{\sum_{i=1}^n g_i}} = \frac{\pi_j}{\frac{(\sum_{i=1}^n g_i) - g_1}{\sum_{i=1}^n g_i}} = \frac{\pi_j}{1 - \pi_1}$$

So we can write:

$$\left(\frac{\pi_2}{1 - \pi_1}, \ldots, \frac{\pi_n}{1 - \pi_1} \right) \sim \text{Dirichlet}(\alpha_2, \ldots \alpha_n)$$

Sampling for HDP: notation using restaurant franchise

- x_{ji}: i^{th} customer at the j^{th} restaurant.
- N customers at each restaurant j.
- each customer x_{ji} associates a table index $t_{ji} \in \{1, \dots T\}$, $T << N$.
- each table t_{ji} associates with a dish number $k_{jt} \in \{1, \dots, K\}$, $K << T$.
- a **shorthand** notation $z_{ji} = k_{jt_{ji}}$: customer x_{ji} has table number t_{ji} which serve dish k_{jt}
- $m_{.}$ is the count of all dish served.

Sampling t

- the equation is:

$$p(t_{ji} = t|\mathbf{t}^{-ji}, \mathbf{k}, x_{ji}) \propto \begin{cases} n_{jt.}^{-ji} f_{k_{ji}}^{\mathbf{x}-ji}(x_{ji}) & \text{IF } t \text{ is previously used} \\ \alpha_0 p(x_{ji}|\mathbf{t}^{-ji}, t_{ji} = t^{\text{new}}, \mathbf{k}) & \text{IF } t = t^{\text{new}} \end{cases}$$

- when t_{ji} is a **new table**, x_{ji} should associate a new dish k.
- just like $f(x|k^{\text{new}}) = \int_\phi f(x|\phi)h(\phi)\mathrm{d}\phi$, we also need to **integrate** out possible values of $k_{jt^{\text{new}}}$:
- However, this dish may be an existing or a **new** one in the entire franchise.

$$p(x_{ji}|\mathbf{x}^{-ji}, t_{jt} = t^{\text{new}}, \mathbf{k}) = \underbrace{\sum_{k=1}^{K} \frac{m_{\cdot k}}{m_{\cdot\cdot} + \gamma} f_k^{\mathbf{x}-ji}(x_{ji})}_{\textbf{part 1}} + \underbrace{\frac{\gamma}{m_{\cdot\cdot} + \gamma} f_{k^{\text{new}}}^{\mathbf{x}-ji}(x_{ji})}_{\textbf{part 2}}$$

1. **part 1**: $k_{jt_{ji}}$ is an **existing** dish in the franchise
2. **part 2**: $k_{jt_{ji}}$ is a **new** dish in the franchise

- **exercise** what is **after** a customer sits in a **new** table?

Sampling k

- this is to decide dish for all customers of the same **table** k_{jt}:

$$p(k_{jt} = k | \mathbf{k}^{-jt}, \mathbf{t}, \mathbf{x}_{jt}) \propto \begin{cases} m_{\cdot k}^{-jt} f_{\mathbf{x}_{jt}}^{\mathbf{x}_{-jt}}(\mathbf{x}_{jt}) & \text{IF } k \text{ is previously used} \\ \gamma f_{k^{\text{new}}}^{\mathbf{x}_{-jt}}(\mathbf{x}_{jt}) & \text{IF } k = k^{\text{new}} \end{cases}$$

  where $\mathbf{x}_{-jt}$ is every customer of the same table t, and x_{ji} is a single customer
- there is also a single person version:

$$p(k_{jt^{\text{new}}} = k | \mathbf{k}^{-ji}, \mathbf{t}, \mathbf{x}_{jt}) \propto \begin{cases} m_{\cdot k}^{-ji} f_{\mathbf{x}_{jt}}^{\mathbf{x}_{-jt}}(\mathbf{x}_{jt}) & \text{IF } k \text{ is previously used} \\ \gamma f_{k^{\text{new}}}^{\mathbf{x}_{-jt}}(\mathbf{x}_{jt}) & \text{IF } k = k^{\text{new}} \end{cases}$$

  exercise think about when you may use this version?

Likelihood function $f_{\mathbf{k}}^{\mathbf{x}_{-ji}}(x_{ji})$

- the likelihood function for $z_{ji} = k$, i.e., sitting on **existing** table

$$
\begin{aligned}
f_{\mathbf{k}}^{\mathbf{x}_{-ji}}(x_{ji}) &= p(x_{ji}|\mathbf{x}_{-ji}, z_{jt} = \mathbf{k}, \mathbf{z}^{-ji}) \\
&= \int_{\phi_k} p(x_{ji}|\phi_k) p(\phi_k|\mathbf{x}_{-ji} = k) \mathrm{d}\phi_k \\
&= \int_{\phi_k} p(x_{ji}|\phi_k) p(\mathbf{x}_{-ji} = k|\phi_k) p(\phi_k) \mathrm{d}\phi_k \\
&\propto \int_{\phi_k} f(x_{ji}|\phi_k) \prod_{j' \neq j, i' \neq i, z_{j'i'} = k} f(x_{j'i'}|\phi_k) h(\phi_k) \mathrm{d}\phi_k \\
&= \frac{\int_{\phi_k} f(x_{ji}|\phi_k) \prod_{j' \neq j, i' \neq i, z_{j'i'} = k} f(x_{j'i'}|\phi_k) h(\phi_k) \mathrm{d}\phi_k}{p(\mathbf{x}_{-ji}, z_{jt} = k, \mathbf{z}^{-ji})} \\
&= \frac{\int_{\phi_k} f(x_{ji}|\phi_k) \prod_{j' \neq j, i' \neq i, z_{j'i'} = k} f(x_{j'i'}|\phi_k) h(\phi_k) d\phi_k}{\int_{\phi_k} \prod_{j' \neq j, i' \neq i, z_{j'i'} = k} f(x_{j'i'}|\phi_k) h(\phi_k) \mathrm{d}\phi_k}
\end{aligned}
$$

- the likelihood function for $z_{ji} = \text{new}$, i.e., sitting on **new** table:

$$
\begin{aligned}
f_{\mathbf{k}\text{new}}^{\mathbf{x}_{-ji}}(x_{ji}) &= p(x_{ji}|\mathbf{x}_{-ji}, z_{jt} = \text{new}, \mathbf{z}^{-ji}) \\
&= \int_{\phi} p(x_{ji}|\phi) p(\phi) \mathrm{d}\phi
\end{aligned}
$$

Sampling G_0 explicitly

- in previous sampling scheme, all groups are coupled since G_0 is integrated out.
- this is just like the DP case: $z_i | \mathbf{z}_{-1}$
- alternative sampling scheme is to have explicit $G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}$
- allow posterior conditioned on G_0 factorizes across groups.

Sampling G_0 explicitly (2)

- given $(\mathbf{t}, \mathbf{k})$, we can draw G_0 by noting:
  - $G_0 \sim \text{DP}(\gamma, H)$
  - $\psi_{jt} \sim G_0$ for each table t
- this is just the posterior of DP we saw earlier:
  $G' = G(A_1), \ldots, G(A_r) | \theta_1, \ldots, \theta_n \sim \text{Dir}(\alpha H(A_1) + n_1, \ldots, \alpha H(A_k) + n_k)$

$$G_0 | \mathbf{t}, \mathbf{k}, \gamma, H, \{\psi_{jt}\} = \text{DP}\left(\gamma + m_{\cdot\cdot}, \frac{\gamma H + \sum_{k=1}^{K} m_{\cdot k} \delta_{\phi_k}}{\gamma + m_{\cdot\cdot}}\right)$$

- posterior of G_0 constructed from different elements:

$$\begin{aligned}
\boldsymbol{\beta} &= (\beta_1, \ldots, \beta_K, \beta_u) \sim \text{Dir}(m_{\cdot 1}, \ldots, m_{\cdot K}, \gamma) \\
p(\phi_k | \mathbf{t}, \mathbf{k}) &\propto h(\phi_k) \prod_{ji: z_{ji} = k} f(x_{ji} | \phi_k) \\
G_u &\sim \text{DP}(\gamma, H) \\
G_0 &= \sum_{k=1}^{K} \beta_k \delta_{\phi_k} + \beta_u G_u
\end{aligned}$$

- when **new** component is instantiated:
  1. $b \sim \text{Beta}(1, \gamma)$
  2. $K \leftarrow K + 1$
  3. $\beta_K = b \beta_u$
  4. $\beta_u \leftarrow (1 - b) \beta_u$

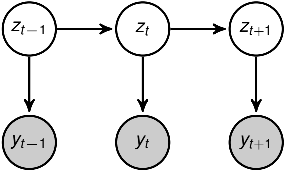
Some Extensions to DP

- Hierarchical Dirichlet Process (HDP)
- **HDP-Hidden Marko Model**
- Indian Buffet Process

Traditional HMM

Under normal HMM, you have a transition matrix A, let the j^{th} row of A to be π_i, then:

$$
A = \begin{bmatrix} \pi_1 \\ \pi_2 \\ \dots \\ \pi_K \end{bmatrix} = \begin{bmatrix} p(z_{t+1} = 1 | z_t = 1) & p(z_{t+1} = 2 | z_t = 1) & \dots & p(z_{t+1} = K | z_t = 1) \\ p(z_{t+1} = 1 | z_t = 2) & p(z_{t+1} = 2 | z_t = 2) & \dots & p(z_{t+1} = K | z_t = 2) \\ \dots & \dots & \dots & \dots \\ p(z_{t+1} = 1 | z_t = K) & p(z_{t+1} = 2 | z_t = K) & \dots & p(z_{t+1} = K | z_t = K) \end{bmatrix}
$$



To obtain the current latent state, we need to sample $z_t \sim \text{Mult}(\pi_{z_{t-1}})$.

HDP-HMM

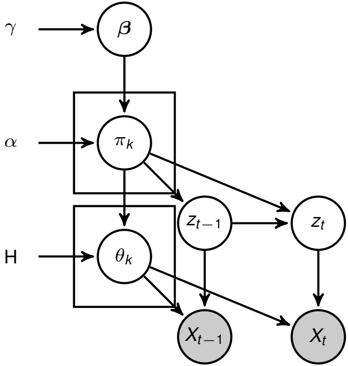
- Same idea has been extended to non-parametric bayes,
- Allow π_j to have infinite many components.
- Matrix A has size $\infty \times \infty$. But the "recovered" number of states are finite, so you only "jumping around" in the upper-left corner of matrix A.

$$\begin{bmatrix} p(z_{t+1} = 1 | z_t = 1) & p(z_{t+1} = 2 | z_t = 1) & \dots & p(z_{t+1} = \infty | z_t = 1) \\ p(z_{t+1} = 1 | z_t = 2) & p(z_{t+1} = 2 | z_t = 2) & \dots & p(z_{t+1} = \infty | z_t = 2) \\ \dots & \dots & \dots & \dots \\ p(z_{t+1} = 1 | z_t = \infty) & p(z_{t+1} = 2 | z_t = \infty) & \dots & p(z_{t+1} = \infty | z_t = \infty) \end{bmatrix}$$

Generative model

$$\boldsymbol{\beta} \sim \text{GEM}(\gamma)$$
$$\boldsymbol{\pi}_j \sim \text{DP}\,(\alpha, \boldsymbol{\beta})$$
$$z_t \sim \text{Mult}(\pi_{z_{t-1}})$$
$$\theta_k \sim H$$
$$X_t \sim F(x|\theta_{z_t})$$

Graphical model



HMM conditional

$$z_1 = 1 \to z_2 = 3 \to z_3 = 2 \to \boxed{z_4 = k} \to z_5 = 1 \to z_6 = 2 \to z_7 = 1 \to z_8 = 3 \to z_9 = 2 \to z_{10} = 2$$

- let $t-1 = 3$, $\mathbf{t} = \mathbf{4}$, $t+1 = 5$
- n_{ij} is the number of transitions from state i to j **excluding** time steps $t-1$ and t:

$$\begin{array}{llll}
n_{1,1} = 0 & n_{1,2} = 1 & n_{1,3} = 2 & \mathbf{n}_{1,:} = 3 \\
n_{2,1} = 1 & n_{2,2} = 1 & n_{2,3} = 0 & \mathbf{n}_{2,:} = 2 \\
n_{3,1} = 1 & n_{3,2} = 2 & n_{3,3} = 0 & \mathbf{n}_{3,:} = 3 \\
\mathbf{n}_{:,1} = 2 & \mathbf{n}_{:,2} = 4 & \mathbf{n}_{:,3} = 2 &
\end{array}$$

- $\mathbf{n}_{:,k}$ is the number of transitions **INTO** state k
- $\mathbf{n}_{k,:}$ is the number of transitions **FROM** state k

$$\Pr(z_t = k|\mathbf{z}_{-t}) \propto \Pr\left(\{z_t = k|z_{t-1} = \mathbf{2}\}_{t=2:T}\right) \Pr\left(\{z_{t+1} = \mathbf{1}|z_t = k\}_{t=1:T-1}\right)$$

$$\Pr(z_t = 1|\mathbf{z}_{-t}) \propto \Pr\left(\{z_t = 1|z_{t-1} = \mathbf{2}\}_{t=2:T}\right) \Pr\left(\{z_{t+1} = \mathbf{1}|z_t = 1\}_{t=1:T-1}\right)$$

$$= \frac{n_{2,1}}{\mathbf{n}_{:,1}} \frac{n_{1,1}}{\mathbf{n}_{1,:}}$$

$$\Pr(z_t = 2|\mathbf{z}_{-t}) \propto \Pr\left(\{z_t = 2|z_{t-1} = \mathbf{2}\}_{t=2:T}\right) \Pr\left(\{z_{t+1} = \mathbf{1}|z_t = 2\}_{t=1:T-1}\right)$$

$$= \frac{n_{2,2}}{\mathbf{n}_{:,2}} \frac{n_{2,1}}{\mathbf{n}_{2,:} + 1}$$

exercise why denominator increase by 1? What happens when $z_{t+1} = z_t$

$$\Pr(z_t = 3|\mathbf{z}_{-t}) \propto \Pr\left(\{z_t = 3|z_{t-1} = \mathbf{2}\}_{t=2:T}\right) \Pr\left(\{z_{t+1} = \mathbf{1}|z_t = 3\}_{t=1:T-1}\right)$$

$$= \frac{n_{2,3}}{\mathbf{n}_{:,3}} \frac{n_{3,1}}{\mathbf{n}_{3,:}}$$

The probability $\Pr(z_t | z_{t-1}, \boldsymbol{\beta}, \mathbf{Y}, \alpha, H)$ without slice variables

$$\Pr(z_t | z_{t-1}, \boldsymbol{\beta}, \mathbf{Y}, \alpha, H) \propto p(y_t | z_t, \mathbf{z}_{-t}, \mathbf{y}_{-t}, H) \underbrace{\Pr(z_t | \mathbf{z}_{-t}, \boldsymbol{\beta}, \alpha)}$$

$$\Pr(z_t = k | \mathbf{z}_{-t}, \boldsymbol{\beta}, \alpha) \propto \begin{cases} \left(\frac{n_{z_{t-1},k} + \alpha \beta_k}{\mathbf{n}_{:,k} + \alpha} \right) \left(\frac{n_{k,z_{t+1}} + \alpha \beta_{z_{t+1}}}{\mathbf{n}_{k,:} + \alpha} \right) & \text{if} \quad k \leq K, k \neq z_{t-1} \\ \left(\frac{n_{z_{t-1},k} + \alpha \beta_k}{\mathbf{n}_{:,k} + \alpha} \right) \left(\frac{n_{k,z_{t+1}} + 1 + \alpha \beta_{z_{t+1}}}{\mathbf{n}_{k,:} + 1 + \alpha} \right) & \text{if} \quad k = z_{t-1} = z_{t+1} \\ \left(\frac{n_{z_{t-1},k} + \alpha \beta_k}{\mathbf{n}_{:,k} + \alpha} \right) \left(\frac{n_{k,z_{t+1}} + \alpha \beta_{z_{t+1}}}{\mathbf{n}_{k,:} + 1 + \alpha} \right) & \text{if} \quad k = z_{t-1} \neq z_{t+1} \\ \alpha \beta_k \beta_{z_{t+1}} & \text{if} \quad k = K + 1 \end{cases}$$

- note that the DP sampling $\Pr(z_t = k | \mathbf{z}_{-t}, \alpha) \propto \begin{cases} \frac{n_k + \alpha}{\cancel{\mathbf{n} + \alpha}} & \text{if existing} \\ \frac{\alpha}{\cancel{\mathbf{n} + \alpha}} & \text{if new} \end{cases}$ does not apply in HDP-HMM, as $\mathbf{n}$ is not constant.
- also when $k = \text{new}$, $\mathbf{n}_{k,:} = \mathbf{n}_{:,k} = n_{z_{t-1},k} = n_{k,z_{t+1}} = 0$
- in DP sampling $\mathbf{n} > 0$ and remain constant.

Slice variables $u_1, \dots u_T$

- Introduce auxiliary variables $u_1, \dots u_t$:

$$u_t \sim \mathrm{U}(0, \pi_{z_{t-1}, z_t}) \implies p(u_t | \mathbf{z}, \boldsymbol{\pi}) = p(u_t | z_{t-1}, z_t, \boldsymbol{\pi})$$

- Another way of writing it:

$$p(u_t | z_{t-1}, z_t, \boldsymbol{\pi}) = \frac{\mathbb{I}\left(0 < u_t < \pi_{z_{t-1}, z_t}\right)}{\pi_{z_{t-1}, z_t}}$$

$$\begin{aligned}
p(z_t | y_{1:t}, u_{1:t}) &\propto p(z_t, u_t, y_t | y_{1:t-1}, u_{1:t-1}) \\
&= \sum_{z_{t-1}} p(z_t, u_t, y_t, z_{t-1} | y_{1:t-1}, u_{1:t-1}) \\
&= \sum_{z_{t-1}} p(y_t | z_t) \underbrace{p(u_t | z_t, z_{t-1})} p(z_t | z_{t-1}) p(z_{t-1} | y_{1:t-1}, u_{1:t-1}) \\
&= p(y_t | z_t) \sum_{z_{t-1}} \underbrace{\frac{\mathbb{I}\left(0 < u_t < \pi_{z_{t-1}, z_t}\right)}{\pi_{z_{t-1}, z_t}}} p(z_t | z_{t-1}) p(z_{t-1} | y_{1:t-1}, u_{1:t-1}) \\
&= p(y_t | z_t) \sum_{z_{t-1}} \mathbb{I}\left(u_t < \pi_{z_{t-1}, z_t}\right) p(z_{t-1} | y_{1:t-1}, u_{1:t-1})
\end{aligned}$$

Slice variables $u_1, \dots u_T$ (2)

- **forward pass**:

$$
\begin{aligned}
\Pr(z_t|y_{1:t}, u_{1:t}) &\propto \Pr(z_t, u_t, y_t|y_{1:t-1}, u_{1:t-1}) \\
&= \Pr(y_t|z_t) \sum_{z_{t-1}} \mathbb{I}\left(u_t < \pi_{z_{t-1},z_t}\right) \Pr(z_{t-1}|y_{1:t-1}, u_{1:t-1}) \\
&= \Pr(y_t|z_t) \sum_{\{z_{t-1}\} u_t < \pi_{z_{t-1},z_t}} \Pr(z_{t-1}|y_{1:t-1}, u_{1:t-1})
\end{aligned}
$$

  u_t truncates the above summation to **finitely many** z_{t-1}s that satisfy both constraints:

  1. $u_t < \pi_{z_{t-1},z_t}$
  2. $\Pr(z_{t-1}|y_{1:t-1}, u_{1:t-1}) > 0$

- To sample the whole trajectory $z_{1:t}$:

  1. Sample $\mathbf{z_T} \sim \Pr(z_T|y_{1:T}, u_{1:T})$ - which is used in the "likelihood" function for z_{T-1}:
  2. then, perform a **backward pass**, where we sample:

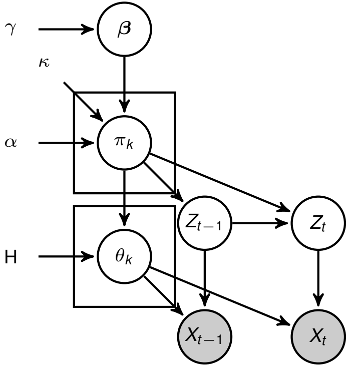
$$
z_t|z_{t+1} : \Pr(z_t|z_{t+1}, y_{1:T}, u_{1:T}) \propto \Pr(\mathbf{z_{t+1}}|z_t, u_{t+1}) \Pr(z_t|y_{1:t}, u_{1:t})
$$

Sticky HDP - HMM

Generative model

$$
\begin{aligned}
\boldsymbol{\beta} &\sim \text{GEM}(\gamma) \\
\boldsymbol{\pi}_j &\sim \text{DP}\left(\alpha + \kappa, \frac{\alpha\boldsymbol{\beta} + \kappa\delta_j}{\alpha + \kappa}\right) \\
z_t &\sim \text{Mult}(\pi_{z_{t-1}}) \\
\theta_k &\sim H \\
X_t &\sim F(x|\theta_{z_t})
\end{aligned}
$$

Graphical model



Some Extensions to DP

- Hierarchical Dirichlet Process (HDP)
- HDP-Hidden Marko Model
- **Indian Buffet Process**

Indian Buffet Process: Its relationship with DP

DP

- $\Pr(z_1 \dots z_N)$, where $z_i \in (1 \dots K)$ indicate category.
- You also want K potentially be infinite
- A "clustering" property, controllable through a single parameter α
- Can also be thought as a special $N \times K$ Z matrix, where there is only one "1" in each row.

IBP

- More general than DP: z_i can take multiple values $\in (1, \dots K)$
- This is equivelently of saying that, z_i is a binary vector of K elements.
- Given N such data, we have a binary matrix of size $N \times K$
- A "clustering" property, controllable through a single parameter α, a column with more 1, results it to have more 1s.

The big Z matrix

An example of Z matrix:

1	0	1	1	0	...	1
0	1	0	0	0	...	0
...	...	...	...	...	...	0
1	1	0	0	0	...	0

For each column: $Pr(z_{ik} = 1) \sim \text{Ber}(\mu_k)$ independently.
Each $u_k \sim \text{Beta}\left(\frac{\alpha}{k}, 1\right)$ is also distributed independently.
The marginal distribution:

Multinomial-Dirichlet

$$P(p_1, \dots, p_k | n_1, \dots, n_k)$$

$$\propto \underbrace{\frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} p_i^{\alpha_i - 1}}_{\text{Dir}(p_1,\dots,p_k|\alpha_1,\dots,\alpha_k)} \underbrace{\frac{n!}{n_1! \dots n_k!} \prod_{i=1}^{k} p_i^{n_i}}_{\text{Mult}(n_1,\dots,n_k|p_1,\dots p_k)}$$

$$\propto \prod_{i=1}^{k} p_i^{\alpha_i - 1} \prod_{i=1}^{k} p_i^{n_i} = \prod_{i=1}^{k} p_i^{\alpha_i - 1 + n_i}$$

$$= \text{Dir}(p_1, \dots p_k | \alpha_i + n_i, \dots \alpha_k + n_k)$$

Bernoulli-Binomial

$$P(p | n_1 = m)$$

$$\propto \underbrace{\frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha - 1}(1 - p)^{\beta - 1}}_{\text{Beta}(p|\alpha,\beta)} \underbrace{\frac{N!}{m!(N - m)!} p^k (1 - p)^{N - k}}_{\text{Binomial}(n_1,n_2|p)}$$

$$\propto p^{\alpha - 1}(1 - p)^{\beta - 1} p^k (1 - p)^{N - k}$$

$$= p^{\alpha - 1 + k}(1 - p)^{\beta - 1 + N - k}$$

$$= \text{Beta}(p | \alpha_i + k, \beta + N - k)$$

Bernoulli- Beta vs Multinomial-Dirichlet: Marginal

Multinomial-Dirichlet

$$\int_{p_1,\dots,p_k} P(p_1,\dots,p_k,n_1,\dots,n_k)$$

$$= \frac{N!}{n_1!\dots n_k!} \frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\prod_{i=1}^k \Gamma(\alpha_i)} \frac{\prod_{i=1}^k \Gamma(\alpha_i + n_i)}{\Gamma\left(N + \sum_{i=1}^k \alpha_i\right)}$$

Bernoulli-Beta

$$\int_p P(p, n_1, n_2)$$

$$= \frac{N!}{k!(N-k)!} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+k)\Gamma(\beta+N-k)}{\Gamma(N+\alpha+\beta)}$$

Bernoulli-Beta Predictivie

$\mu_k \sim \text{Beta}\left(\frac{\alpha}{k}, 1\right)$ $\qquad\qquad \Pr(z_{ik} = 1) \sim \text{Ber}(\mu_k)$.
$n_{k,-i}$ is the number of 1s of k^{th} column, above row i.
Let $\alpha_i = \frac{\alpha}{k}$: compute the density of i^{th} data belonging to existing component m.

$$
\begin{aligned}
\Pr(z_{ik} = 1 | \mathbf{z}_{-i,k}) &= \int_p \Pr(z_{ik} = 1|p) P(p|\underbrace{n_{-i,k}}_{n_1}, \underbrace{i - 1 - n_{-i,k}}_{n_2}) \\
&= \frac{\int_p \Pr(z_{ik} = 1|p) \Pr(n_1, n_2|p) P(p)}{\Pr(n_1, n_2)} = \frac{\int_p \Pr(z_{ik} = 1|p) \Pr(n_1, n_2|p) P(p)}{\int_p \Pr(n_{-i,k}, i - 1 - n_{-i,k}|p) P(p)} \\
&= \frac{\Gamma(\frac{\alpha}{k} + n_{-i,k} + 1)\Gamma(1 + i - 1 - n_{-i,k})}{\Gamma\left(i + \frac{\alpha}{k} + 1\right)} \frac{\Gamma\left(i - 1 + \frac{\alpha}{k} + 1\right)}{\Gamma(\frac{\alpha}{k} + n_{-i,k})\Gamma(1 + i - 1 - n_{-i,k})} = \frac{\frac{\alpha}{k} + n_{-i,k}}{i + \frac{\alpha}{k}}
\end{aligned}
$$

One more factor: relationship between Binomial and Poisson

$$\exp(x) = \lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n$$

Let $\lambda = np$:

$$\begin{aligned}
\text{Binomial}(x|n,p) &= \binom{n}{x}p^x(1-p)^{n-x} = \frac{n!}{x!(n-x)!}\frac{\lambda}{n}^x(1-\frac{\lambda}{n})^{n-x} \\
&= \underbrace{\frac{\lambda^x}{x!}}_{\text{constant}}\underbrace{\frac{n!}{(n-x)!}\frac{1}{n^x}}\left(1-\frac{\lambda}{n}\right)^{n-x} \\
&= \frac{\lambda^x}{x!}\frac{\overbrace{n(n-1),\ldots(n-x+1)}^{n\text{ terms}}}{n^x}\left(1-\frac{\lambda}{n}\right)^{n-x} \\
&= \frac{\lambda^x}{x!}\frac{n}{n}\frac{n-1}{n}\cdots\frac{n-x+1}{n}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-x} \\
&= \frac{\lambda^x}{x!}1\left(1-\frac{1}{n}\right)\cdots\left(1-\frac{x+1}{n}\right)\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-x}
\end{aligned}$$

$$\lim_{n\to\infty}\text{Binomial}(x|n,p) = \lim_{n\to\infty}\binom{n}{x}p^x(1-p)^{n-x}$$
$$= \frac{\lambda^x}{x!}\lim_{n\to\infty}\left(1-\frac{1}{n}\right)\cdots\lim_{n\to\infty}\left(1-\frac{x+1}{n}\right)\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{-x} = \frac{\lambda^x}{x!}\exp(-\lambda)$$

Taking limit $k \to \infty$

$$\lim_{k\to\infty} \Pr(z_{ik}) = \lim_{k\to\infty} \frac{\frac{\alpha}{k} + n_{-i,k}}{i + \frac{\alpha}{k}} = \frac{n_{-i,k}}{i}$$

$$\lim_{n\to\infty} \text{Binomial}(\frac{\lambda}{n}, n) = \text{Poisson}(\lambda)$$

$$\text{Let } k \to \infty : \qquad = \frac{n_{-i,k}}{i}$$

For "new" dishes, i.e., $n_{-i,k} = 0$, then, $\Pr(z_{ik} = 1) = \text{Bernoulli}\left(\frac{\frac{\alpha}{K}}{i+\frac{\alpha}{K}}\right)$

i.e., how many new dishes across all columns would be: $\text{Binomial}\left(\frac{\frac{\alpha}{K}}{i+\frac{\alpha}{K}}, K\right)$

Since $\frac{\frac{\alpha}{k}}{i+\frac{\alpha}{k}} \times k = \frac{\alpha}{i+\frac{\alpha}{k}}$, we have:

$$\lim_{K\to\infty} \text{Binomial}\left(\frac{\frac{\alpha}{K}}{i + \frac{\alpha}{K}}, K\right) = \text{Poisson}\left(\frac{\alpha}{i}\right)$$

So, how many K^+ columns there are?

Let $n_i \sim \text{Poisson}\left(\frac{\alpha}{i}\right)$ $\left(\sum_{i=1}^{N} n_i\right) \sim \text{Poisson}\left(\sum_{i=1}^{N} \frac{\alpha}{i}\right)$

An motivational example of IBP: Factor Analysis

What is Factor Analysis? There are $N = 1000$ students, each having ($p = 10$) scores. Therefore:

$$\begin{bmatrix} y_{11} & y_{12} & \dots & y_{1N} \\ y_{21} & y_{22} & \dots & y_{2N} \\ \dots & \dots & \dots & \dots \\ y_{p1} & y_{p2} & \dots & y_{pN} \end{bmatrix} = \begin{bmatrix} g_{11} & \dots & g_{1k} \\ g_{21} & \dots & g_{2k} \\ \dots & \dots & \dots \\ g_{p1} & \dots & g_{pk} \end{bmatrix} \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1N} \\ \dots & \dots & \dots & \dots \\ x_{k1} & x_{k2} & \dots & x_{kN} \end{bmatrix} + \mathbf{E}$$

$$\mathbf{E} = \begin{bmatrix} e_{11} & e_{12} & \dots & e_{1N} \\ e_{21} & e_{22} & \dots & e_{2N} \\ \dots & \dots & \dots & \dots \\ e_{p1} & e_{p2} & \dots & e_{pN} \end{bmatrix} \text{ and } k << p$$

Or in a matrix form: $\mathbf{Y} = \mathbf{GX} + \mathbf{E}$.

Factor analysis cont.

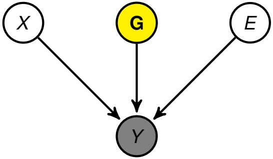
What this means is that a person's i's raw mark is interpretted as:

$$\begin{bmatrix} y_{1i} \\ y_{2i} \\ \dots \\ y_{pi} \end{bmatrix} = x_{1i} \begin{bmatrix} g_{11} \\ g_{21} \\ \dots \\ g_{p1} \end{bmatrix} + x_{2i} \begin{bmatrix} g_{11} \\ g_{21} \\ \dots \\ g_{p1} \end{bmatrix} + \dots x_{ki} \begin{bmatrix} g_{1k} \\ g_{2k} \\ \dots \\ g_{pk} \end{bmatrix} + \begin{bmatrix} e_{1i} \\ e_{2i} \\ \dots \\ e_{pi} \end{bmatrix}$$

- Given a set of k loading factors (vectors) each with dimension p: $\{\mathbf{g}_{:,i}\}_{i=1}^{k}$, the $x_{:,i}$ can be thought as the latent linear weights.
- Of course, you are only given data matrix Y, one has to infer the latent structure. **G**, **X** and **E**. Ths is not as silly as it seems, as DoF is much reduced.

The Bayesian Treatment:

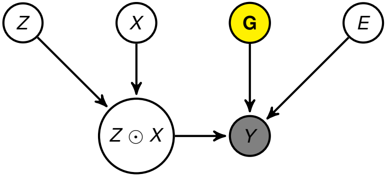
$$e_i \sim \mathcal{N}(0, \sigma_e^2 \mathbf{I}) \qquad \sigma_e^2 \sim \mathcal{IG}(a, b)$$
$$g_k \sim \mathcal{N}(0, \sigma_G^2) \qquad \sigma_G^2 \sim \mathcal{IG}(c, d)$$
$$x_{ki} \sim \mathcal{N}(0, 1) \qquad y_i = \mathbf{G} x_i + e_i$$



- Knowles, d and Ghahramani, Z, Innite Sparse Factor Analysis
- K should known beforehand. What about making K a variable?
- Although $[x_{1,i}, \dots x_{k,i}]^T$ has a reduced dimension, it can still cause “overfitting”.
- We need to introcuce variable number of latent factors K, at the same time, have **sparsity**!

How?

$$e_i \sim \mathcal{N}(0, \sigma_e^2\mathbf{I}) \qquad \sigma_e^2 \sim \mathcal{IG}(a, b)$$
$$g_k \sim \mathcal{N}(0, \sigma_G^2) \qquad \sigma_G^2 \sim \mathcal{IG}(c, d)$$
$$Z \sim \mathcal{IBP}(\alpha) \qquad \alpha \sim \mathcal{G}(e, f)$$
$$x_{ki} \sim \mathcal{N}(0, 1) \qquad y_i = \mathbf{G}(x_i \odot z_i) + e_i$$



- What about if there are two sets of data matrix **Y** and $\mathbf{Y}^{'}$, each having different number of entries. They share the same loading vectors **G**, but with different level of **sparsities**.

$$
\begin{aligned}
e_i &\sim \mathcal{N}(0, \sigma_e^2 \mathbf{I}) & \sigma_e^2 &\sim \mathcal{IG}(a, b) \\
g_k &\sim \mathcal{N}(0, \sigma_G^2) & \sigma_G^2 &\sim \mathcal{IG}(c, d) \\
Z &\sim \mathcal{IBP}(\alpha) & \alpha &\sim \mathcal{G}(e, f) \\
x_{ki} &\sim \mathcal{N}(0, 1) & y_i &= \mathbf{G}(x_i \odot z_i) + e_i
\end{aligned}
$$

$$
\begin{aligned}
e'_i &\sim \mathcal{N}(0, \sigma_e^2 \mathbf{I}) & \sigma_e^2 &\sim \mathcal{IG}(a', b') \\
g_k &\sim \mathcal{N}(0, \sigma_G^2) & \sigma_G^2 &\sim \mathcal{IG}(c, d) \\
Z' &\sim \mathcal{IBP}(\alpha') & \alpha' &\sim \mathcal{G}(e', f') \\
x'_{ki} &\sim \mathcal{N}(0, 1) & y'_i &= \mathbf{G}(x'_i \odot z'_i) + e'_i
\end{aligned}
$$