电子科技大学
University of Electronic Science and Technology of China

Dirichlet Process

Zhongjing Yu

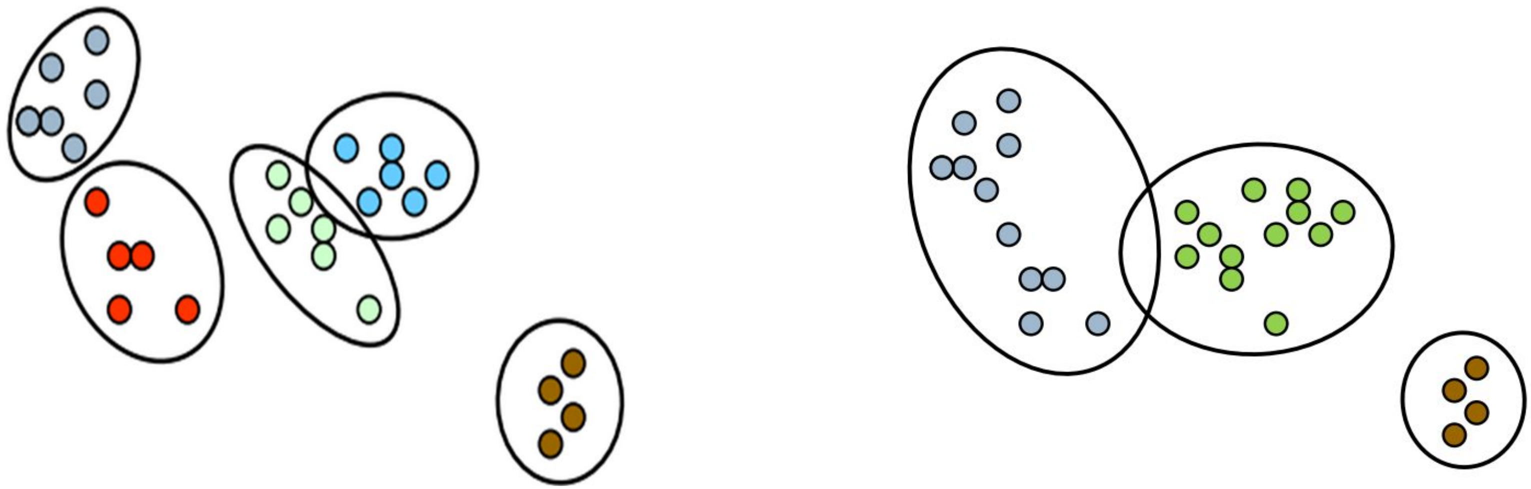
Data Mining Lab, Big Data Research Center, UESTC
Email：junmshao@uestc.edu.cn
http://staff.uestc.edu.cn/shaojunming

Outline

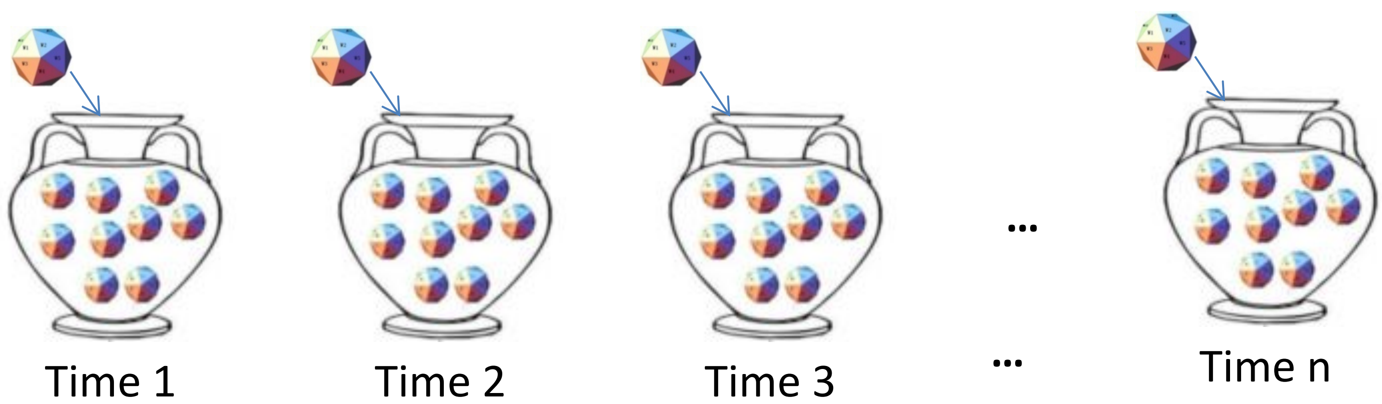
- Motivation of Dirichlet Process
- Dirichlet Distribution
- Dirichlet Process
- The Dirichlet Process, the Chinese Restaurant Process and other representations
- Application of Chinese Restaurant Process

Motivation of Dirichlet Process

- The problem of identifying the number of Clusters.

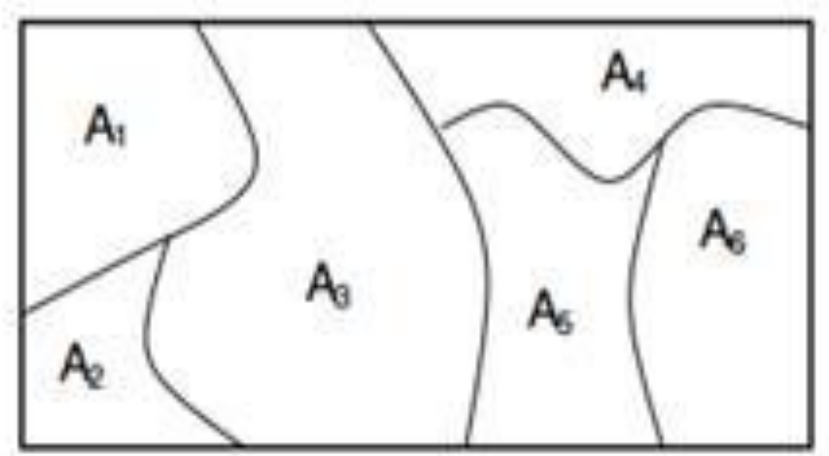


- The number of clusters is expected to change as we add more observations over time



Motivation of Dirichlet Process

- Dirichlet Process : a family of **non-parametric** Bayesian models(in a sense, **infinite number of parameters**).



7 parameters vectors

Dirichlet Process Mixture Models perform **clustering**.

Feature : Don’t require to define the **number of clusters** .

Adapt the number of active clusters over time

Unsupervised.

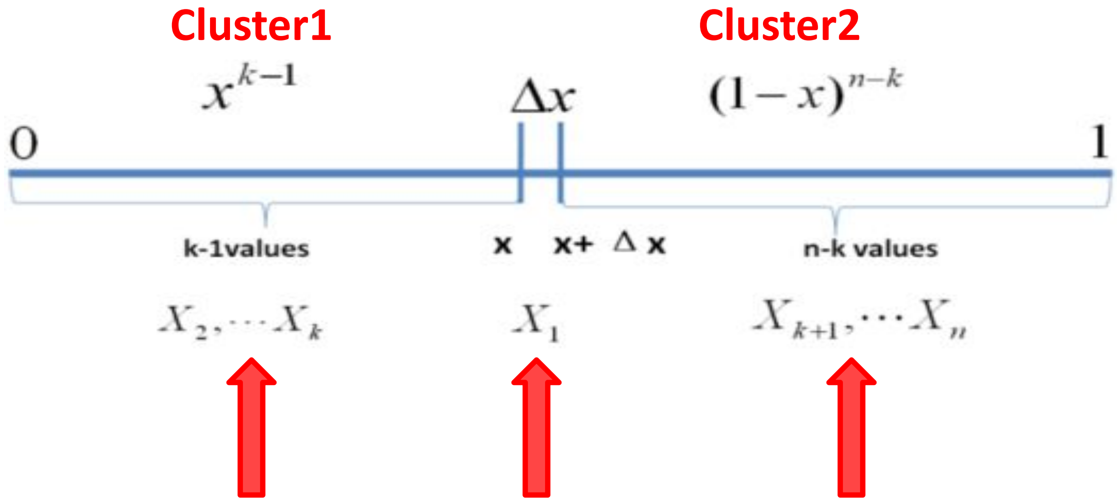
Dirichlet Distribution

- **Game (Bata distribution):**
-

Random figures x1,x2,…x10 ~Uniform(0,1)
Guess the 7th large number. How do you guess (error less 0.01)?

1: $X_1, X_2, \cdots, X_n \overset{\text{iid}}{\sim} Uniform(0,1)$,
2: 把这n 个随机变量排序后得到顺序统计量$X_{(1)}, X_{(2)} \cdots, X_{(n)}$,
3: 问$X_{(k)}$ 的分布是什么

Dirichlet Distribution



$$E=\{X_i \in [0,x)(i=2,\ldots k), X_1 \in [x,x+\Delta x], X_j \in (x+\Delta x,1](j=k+1,\ldots n)\}$$

$$P(E) = x^{k-1}(1-x-\Delta x)^{n-k}\Delta x = x^{k-1}(1-x)^{n-k}\Delta x + o(\Delta x)$$

$$P(x<X_{(k)}<x+\Delta x) = n\binom{n-1}{k-1}x^{k-1}(1-x)^{n-k}\Delta x + o(\Delta x) \Rightarrow f(x)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$$

Return game, choose the peak of f(x),where $\alpha=7, \beta=4$

where $\alpha=k, \beta=n-k+1$

Dirichlet Distribution

数据挖掘实验室
Data Mining Lab

Game(Beta - Bernoulli)

Generate random figure $x_1,x_2,x_3,\ldots x_n$ ~ Uniform(0,1), **guess** p^{th} large number, **Given** $y_1,y_2,\ldots y_m$ ~ Uniform(0,1), $y_1,\ldots y_{m1}$ less p and $y_{m1+1}, y_{m1+2},\ldots y_m$ bigger p. **How do you guess** $(p+m_1)^{th}$?

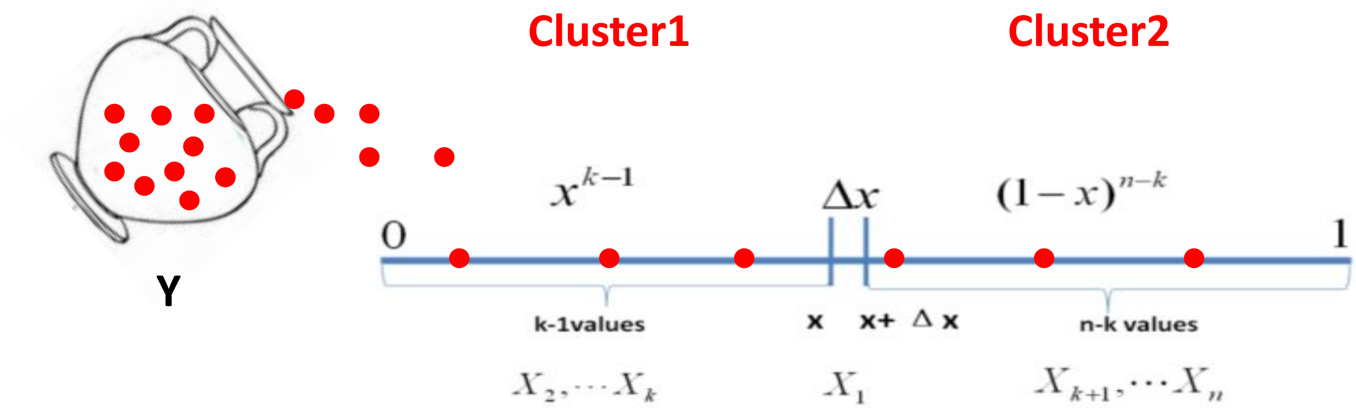
1: $X_1, X_2, \cdots, X_n \overset{\text{iid}}{\sim} Uniform(0,1)$，排序后对应的顺序统计量$X_{(1)}, X_{(2)} \cdots, X_{(n)}$，我们要猜测$p = X_{(k)}$；

2: $Y_1, Y_2, \cdots, Y_m \overset{\text{iid}}{\sim} Uniform(0,1)$, Y_i中有m_1个比p小，m_2个比p大；

3: 问$P(p|Y_1, Y_2, \cdots, Y_m)$ 的分布是什么。

Dirichlet Distribution

- **Game(Beta-Bernoulli)**



Step 1: $p = X_{(k)}$是我们要猜测的参数，推导出p的分布为 $f(p) = Beta(p|k, n-k+1)$称为p的**先验分布**。

Step 2 : 数据 Y_i中有m_1 个比p小，m_2 个比p大，Y_i 相当于是做了m 次贝努利实验，所以 m_1服从二项分布 $B(m, p)$；

Step 3 : 在给定了来自数据提供的(m_1 ,m_2)的知识后，p的后验分布变为

$$f(p|m_1, m_2) = Beta(p|k+m_1, n-k+1+m_2)$$

Dirichlet Distribution

$$Beta(p \mid k, n-k+1) + BernouCount(m_1, m_2) = Beta(p \mid k+m_1, n-k+1+m_2)$$

$$\alpha = k, \beta = n-k+1$$

α, β : **physical count**

- Especially, $Beta(p \mid 1,1) + BernouCount(\alpha-1, \beta-1) = Beta(p \mid \alpha, \beta)$

$Beta(p \mid 1,1)$ is *Uniform(0,1)*

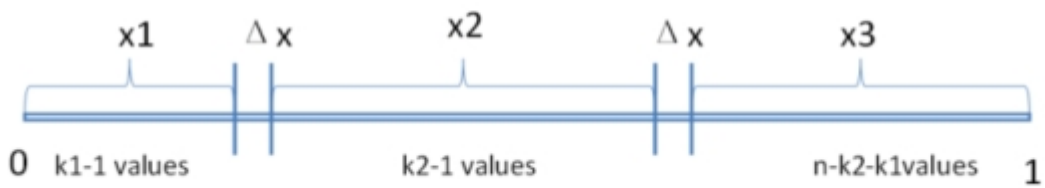
Dirichlet Distribution

Game (Dirichlet):

1: $X_1, X_2, \cdots, X_n \overset{\text{iid}}{\sim} Uniform(0,1)$,

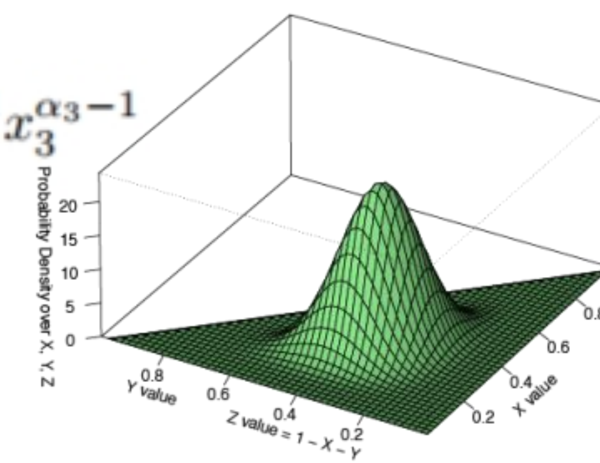
2: 排序后对应的顺序统计量$X_{(1)}, X_{(2)} \cdots, X_{(n)}$,

3: 问$(X_{(k_1)}, X_{(k_1+k_2)})$的联合分布是什么；



$$P(X_{(k_1)} \in (x_1, x_1+\Delta x), X_{(k_1+k_2)} \in (x_2, x_2+\Delta x)) = n(n-1)\binom{n-2}{k_1-1, k_2-1} x_1^{k_1-1} x_2^{k_2-1} x_2^{n-k_1-k_2} (\Delta x)^2$$

$$f(x_1, x_2, x_3) = \frac{\Gamma(\alpha_1+\alpha_2+\alpha_3)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)} x_1^{\alpha_1-1} x_2^{\alpha_2-1} x_3^{\alpha_3-1}$$



Dirichlet Distribution

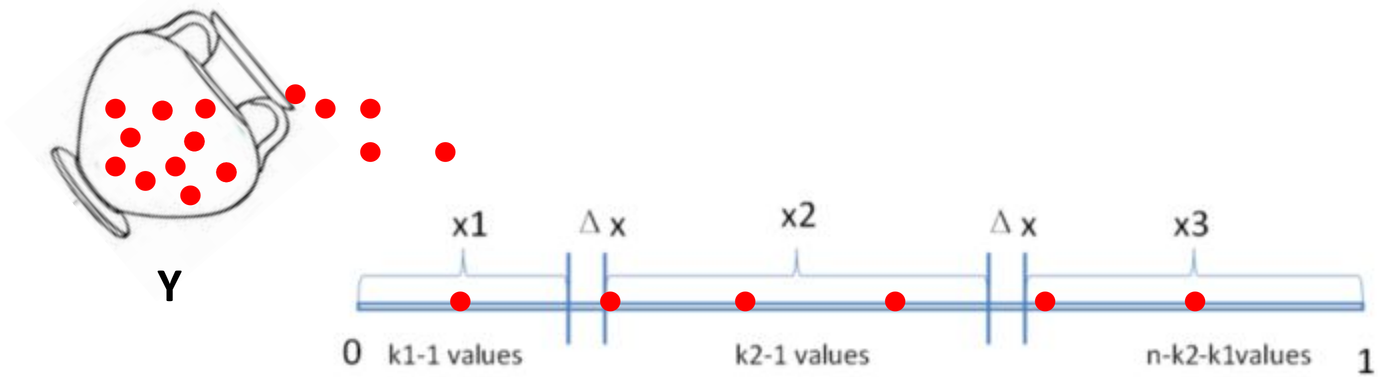
- **Dirichlet-Multi**

1: $X_1, X_2, \cdots, X_n \overset{\text{iid}}{\sim} Uniform(0,1)$，排序后对应的顺序统计量$X_{(1)}, X_{(2)} \cdots, X_{(n)}$

2: 令$p_1 = X_{(k_1)}, p_2 = X_{(k_1+k_2)}, p_3 = 1 - p_1 - p_2$(加上$p_3$是为了数学表达简洁对称)，我们要猜测$\vec{p} = (p_1, p_2, p_3)$；

3: $Y_1, Y_2, \cdots, Y_m \overset{\text{iid}}{\sim} Uniform(0,1)$，$Y_i$中落到$[0, p_1), [p_1, p_2), [p_2, 1]$ 三个区间的个数分别为m_1, m_2, m_3，$m = m_1 + m_2 + m3$；

4: 问后验分布$P(\vec{p}|Y_1, Y_2, \cdots, Y_m)$ 的分布是什么。

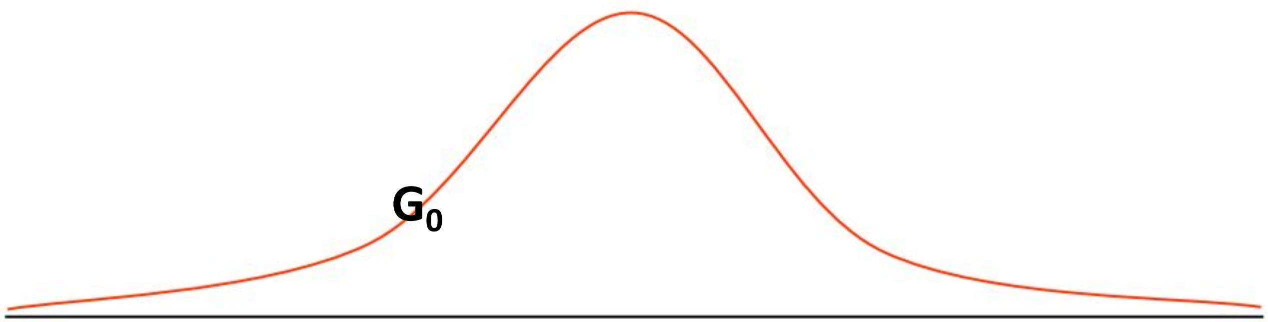


$$Dir(\vec{p}|\vec{k}) + MultCount(\vec{m}) = Dir(\vec{p}|\vec{k} + \vec{m})$$

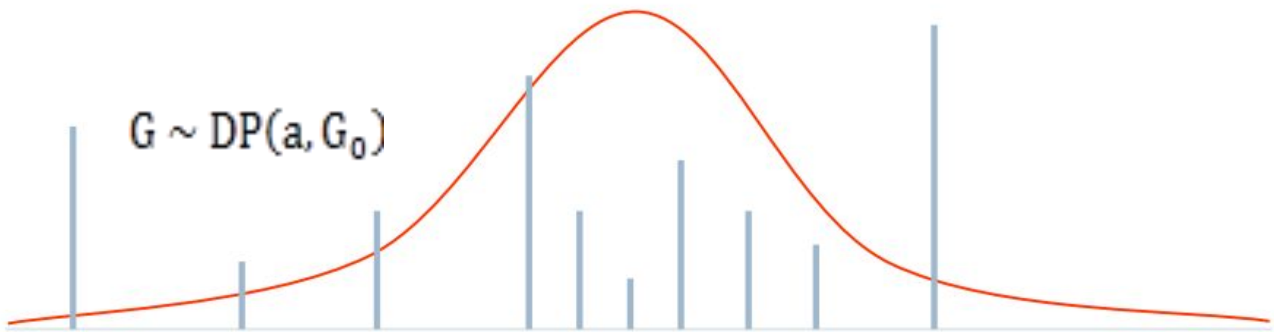
How to apply in Dirichlet Process? Introduce next part

Dirichlet Process

- **Dirichlet Process Induction**:
- Given $x_1, x_2, x_3, \ldots, x_n$. Corresponding feature $\Theta_1, \Theta_2, \Theta_3, \ldots, \Theta_n$.
- Want to divide into k classes.



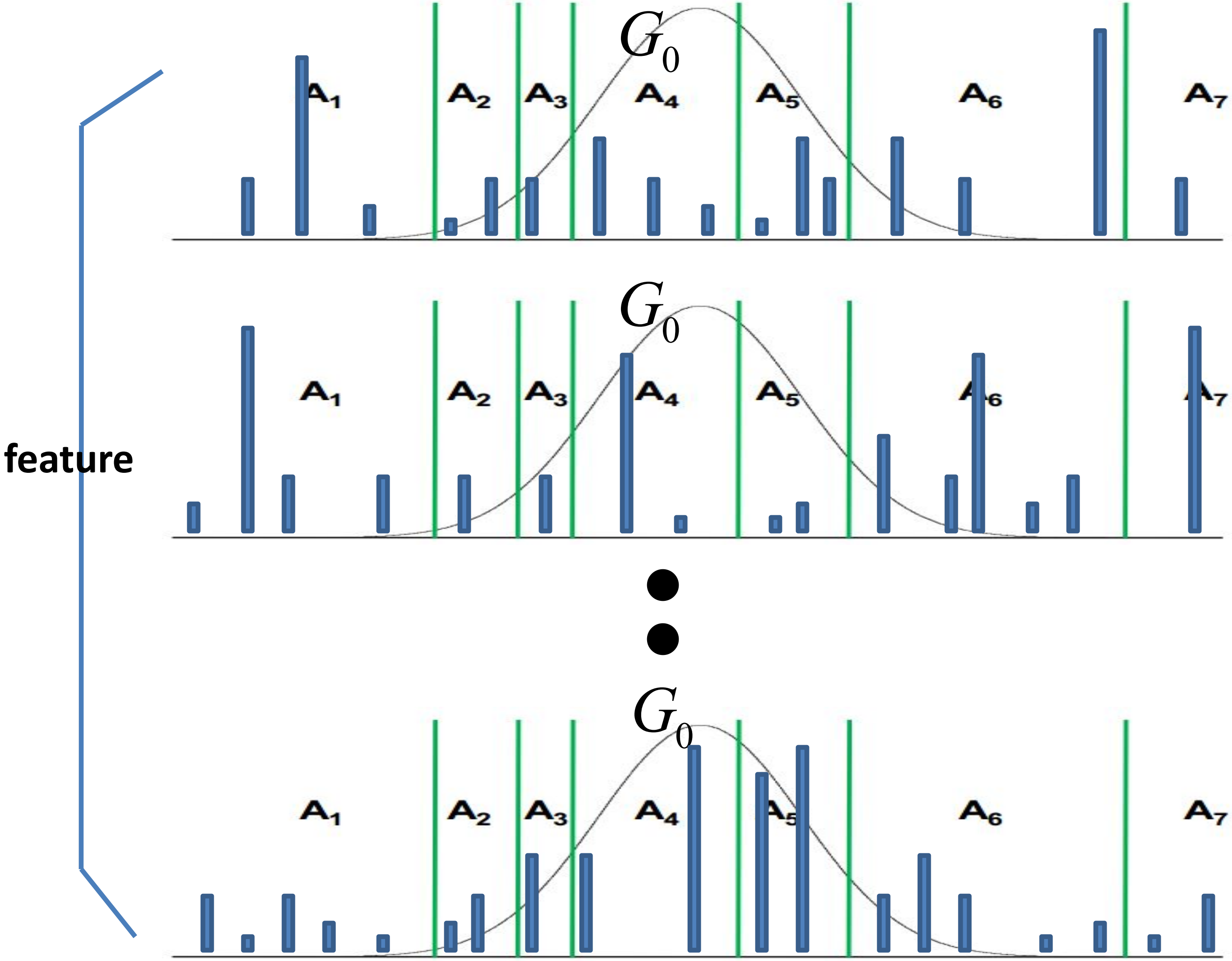
$P(\Theta_i = \Theta_j) = 0$, if G_0 is continuous. ➔ discretization G_0



Notice: even if $G_0(\bullet)$ is a continuous, the distributions drawn from the Dirichlet Process are almost surely **discrete (G).** G **made up** of a countable infinite **number of point masses. (How to comprehend G ?)**

Dirichlet Process

数据挖掘实验室
Data Mining Lab

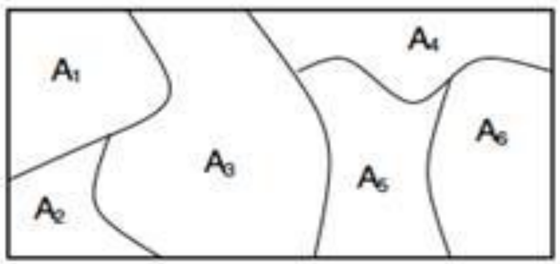


$$(G(A_1),\ldots,G(A_n)) \sim Dirichlet(aG_0(A_1),\ldots,aG_0(A_n))$$

Dirichlet Process

- Dirichlet Process :
- A family of stochastic processes

  A probability distribution whose domain is **itself a set of probability distribution.**



Dirichlet Process： a **probability distribution** over “**probability distribution** over Θ space” .

$$(G(A_1),\ldots,G(A_n)) \sim Dirichlet(aG_0(A_1),\ldots,aG_0(A_n))$$

Sum of each area object to Dirichlet distribution

Where G is ,a random probability measure, a function of **subsets** of space Θ to [0,1]

$G_0(\bullet)$ is a base distribution and the excepted distributions.

a is a strength value.

Dirichlet Process

$$(G(A_1),\ldots,G(A_n)) \sim Dirichlet(aG_0(A_1),\ldots,aG_0(A_n))$$

$$E(G(A)) = G_0(A)$$

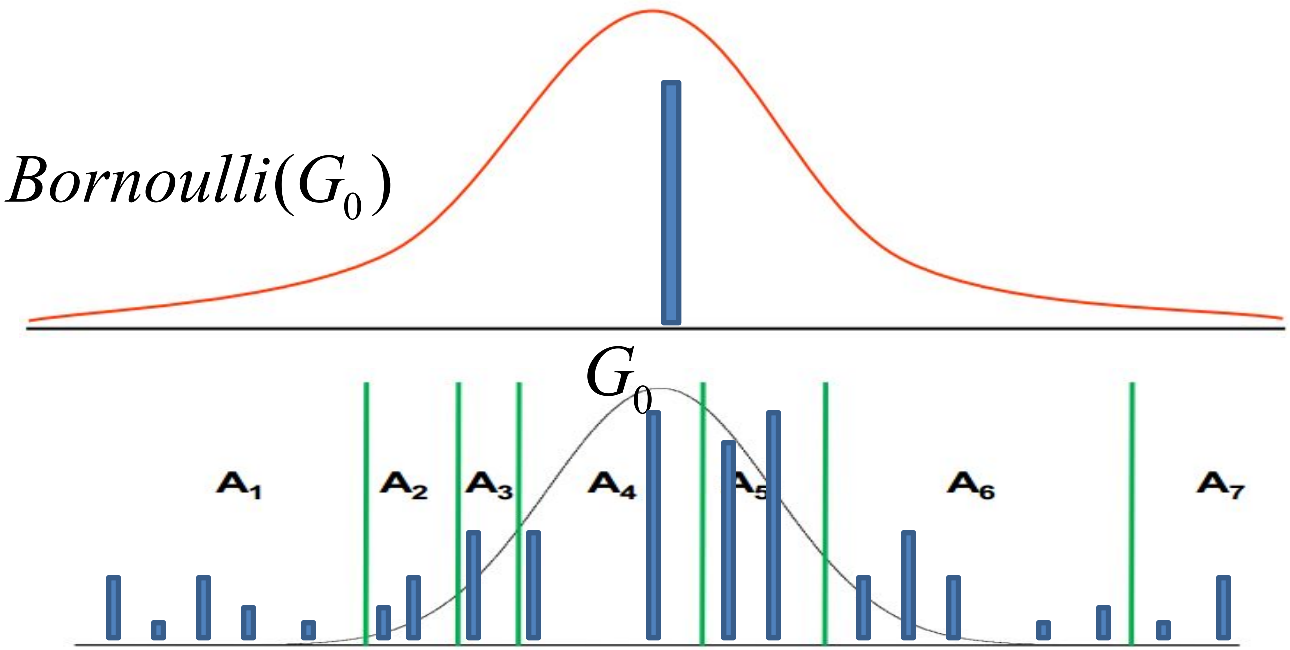
$$V(G(A)) = \frac{G_0(A)(1-G_0(A))}{a+1}$$

$$a \to 0$$

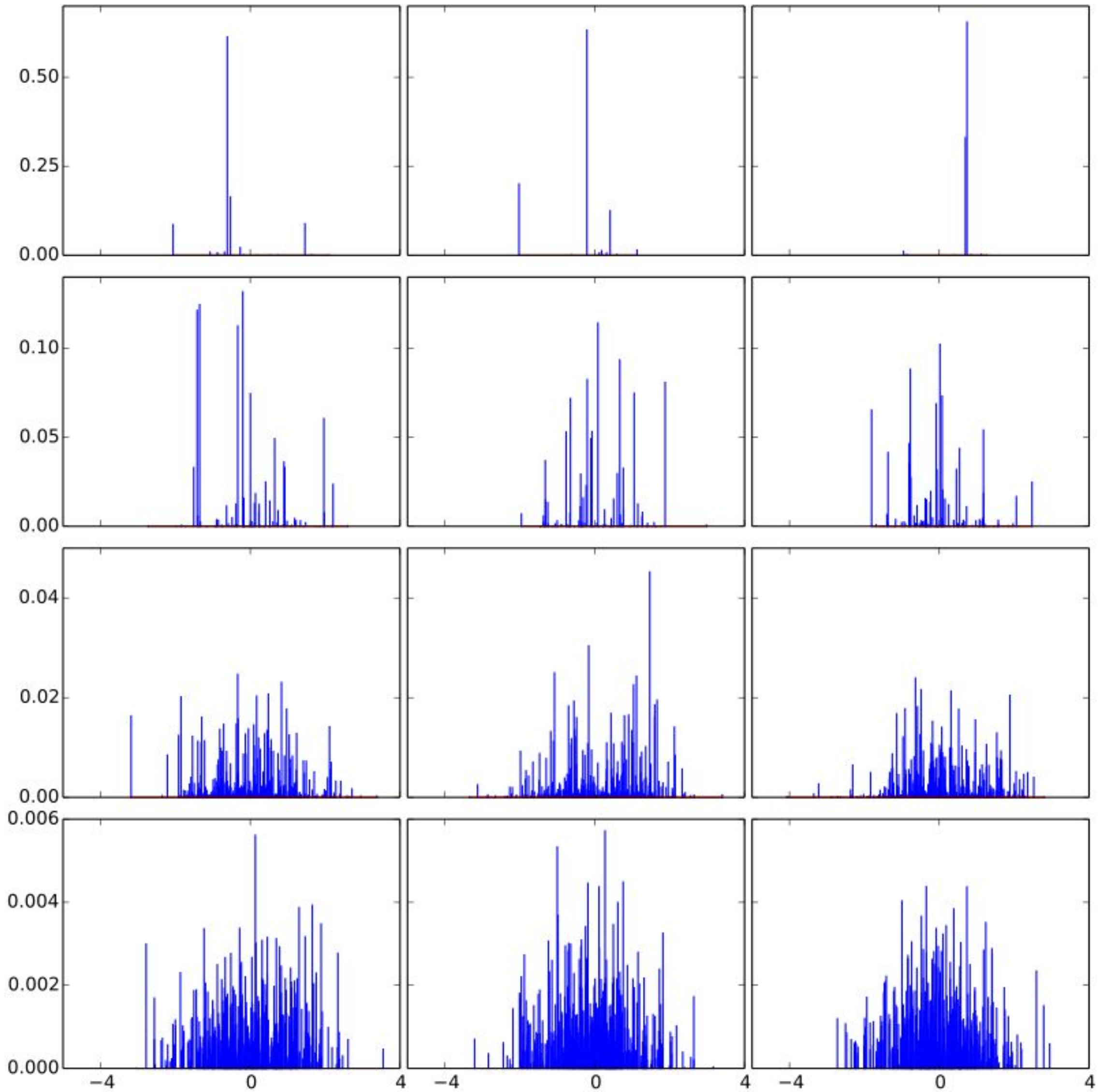
$$V(G) = G_0(1-G_0) \sim Bornoulli(G_0)$$

$$a \to \infty$$

$$V(G) = \frac{G_0(1-G_0)}{a+1}$$



Dirichlet Process



Draws from the Dirichlet process DP(N(0,1), alpha). Each row uses a different alpha: 1, 10, 100 and 1000. A row contains 3 repetitions of the same experiment

Dirichlet Process

数据挖掘实验室 Data Mining Lab

Graph model:

- A random probability measure G : $G \sim DP(a, G_0)$
- Samples : $\boldsymbol{\theta_1, \ldots, \theta_n \sim G}$ **(**discrete**)**

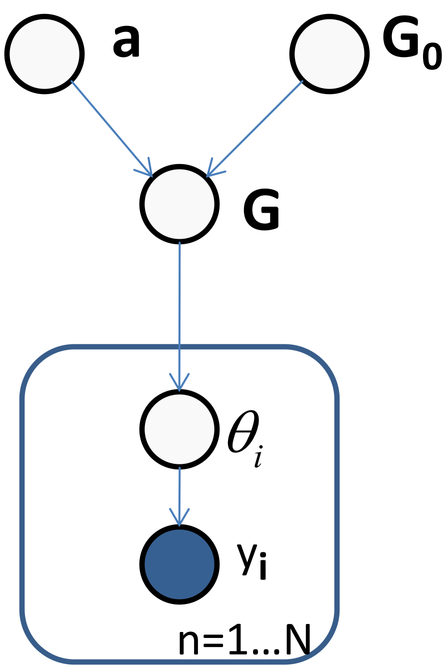
represent : $G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}$

$$\delta_{\theta_k^*} = 1 \quad \text{, if} \quad \theta_k^* = \theta_i \quad ;$$

$$\delta_{\theta_k^*} = 0 \quad \text{, if} \quad \theta_k^* \neq \theta_i$$

where π_k is a weight of samples.

θ_i remarks that feature of y_i



Dirichlet Process Mixture Model

- **Prior:**

$$(G(A_1),...,G(A_n)) \sim Dirichlet(aG_0(A_1),...,aG_0(A_n))$$

$$<=> (p_1, p_2, p_3, ..., p_k) \sim Dir(a_1, a_2, a_3, ..., a_k)$$

- **Likelihood:**

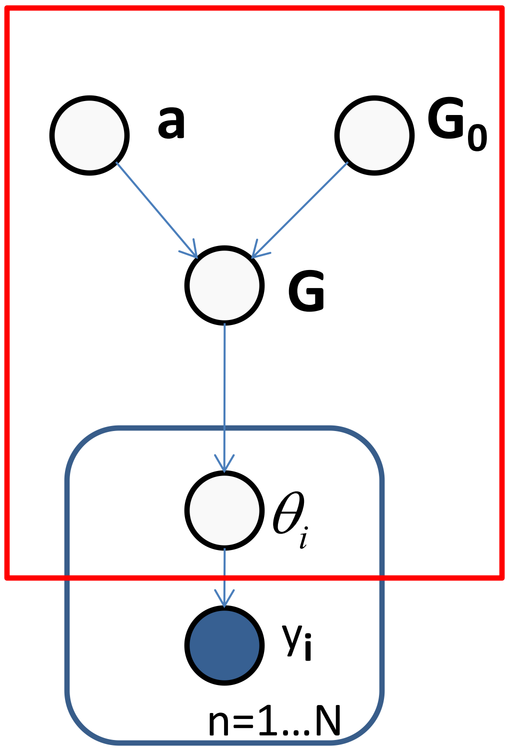
$$\theta_i \sim G(A_i)$$

$$<=> (\theta_1, \theta_2, \theta_3, ..., \theta_k) \sim Multi(p_1, p_2, p_3, ..., p_k)$$

- **Posterior:**

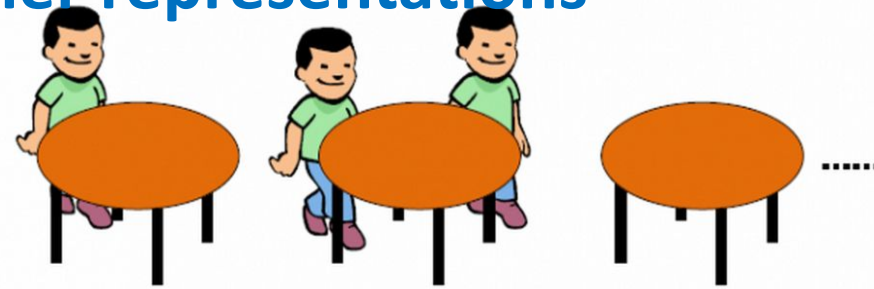
$$likelihood + Porior \propto \prod_{i=1}^{k} p_i^{a_i + n_i - 1}$$

$$= Dir(a_1 + n_1, a_2 + n_2, a_3 + n_3, ..., a_k + n_k)$$



However, we is not likely to consider the order of samples.

The Dirichlet Process, the Chinese Restaurant Process and other representations



$P(z_i = c \mid c_{-i}) =$	1	0	0
	$\frac{1}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	0
	$\frac{1}{2+\alpha}$	$\frac{1}{2+\alpha}$	$\frac{\alpha}{2+\alpha}$
	$\frac{1}{3+\alpha}$	$\frac{2}{3+\alpha}$	$\frac{\alpha}{3+\alpha}$

$$P(X_1, \ldots, X_N) = \int P(G) \prod_{n=1}^{N} P(X_n|G) dG$$

$$X_n|X_1, \ldots, X_{n-1} = \begin{cases} X_i & \text{with probability } \frac{1}{n-1+\alpha} \\ \text{new draw from } G_0 & \text{with probability } \frac{\alpha}{n-1+\alpha} \end{cases}$$

Conclusion : joint probability is same! Exchangeable

Rich get richer

The Dirichlet Process, the Chinese Restaurant Process and other representations

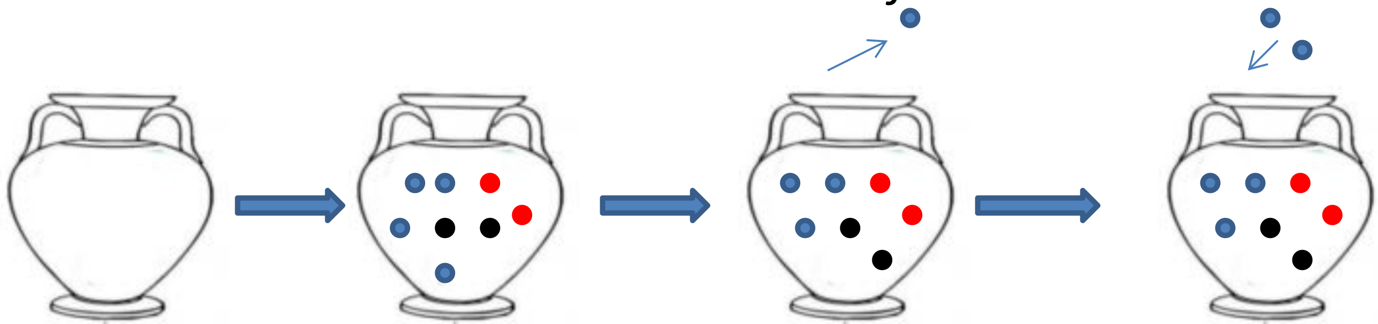
The Pólya urn scheme :

The algorithm:

step 1 : need an **observation** urn, draw a ball from urn (a **non-transparent**)

step 2 : observation is **black**, generate a **new (non-black) color** uniformly, label a new ball this color, drop the new ball into urn.

step 3 : we draw a random ball from the urn, we ***observe*** *its color, we* ***place it back*** *to the urn and we* ***add*** *an additional ball of the same color in the urn.*



Representation : a sequence of $\theta_1,\theta_2,\ldots$ with conditional probabilities

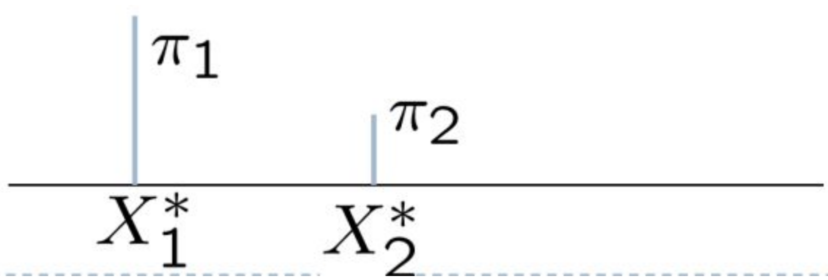
$\theta_n|\theta_{1:n-1} \sim \frac{aG_0+\sum_{i=1}^{n-1}\delta_{\theta_i}}{a+n-1}$ G_0 : distribution over colors. θ_n : the color of the ball

The Dirichlet Process, the Chinese Restaurant Process and other representations

- **The Stick-breaking construction :**
- 一尺之棰，日取其半，万世不竭。------庄子
- We assume that we have a stick **of length 1**, we **break it at position β_1** and we assign π_1 equal to the **length of the part of the stick** that we broke. We **repeat** the same process to obtain π_2, π_3,… etc; due to the way that this scheme is defined we can continue doing it **infinite** times.
- For each $\beta_{i,}$ choose a **θi,** corresponding to a cluster, and then pick out π_i **.** Similarly , stick is divided to some clusters.
- $\beta_1, \beta_2, \ldots, \beta_i, \ldots$ ~Beta(1α)

$$\pi_i = \beta_i \prod_{j=1}^{j=i-1} (1 - \pi_j)$$

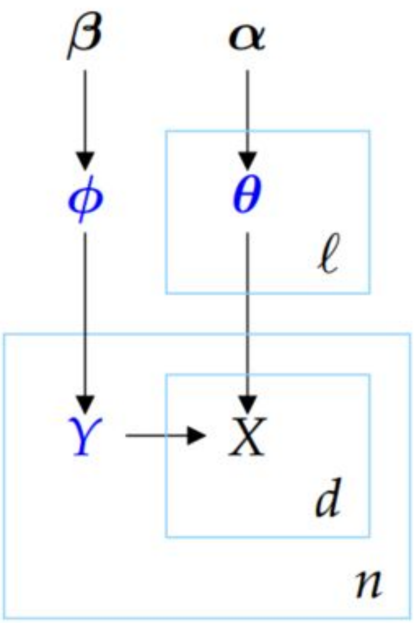
$$G = \sum_{i=1}^{\infty} \pi_i \delta_{X_i^*}$$



Application of Chinese Restaurant Process

数据挖掘实验室 Data Mining Lab

- **Topic modeling**
- **Customers** correspond to **documents** $\boldsymbol{X_i}$ in topic model.
- **Tables** corrspond to hidden **classes** $\boldsymbol{k}$ in topic model



- Data consists of "documents" $\boldsymbol{X}_i$
- Each $\boldsymbol{X}_i$ is a sequence of "words" $X_{i,j}$
- Initialize by *randomly* assign each document $\boldsymbol{X}_i$ to a topic Y_i
- Repeat the following:
  - Replace ϕ with a sample from a Dirichlet with parameters $\boldsymbol{\beta} + \boldsymbol{N}(\boldsymbol{Y})$
  - For each topic k, replace $\boldsymbol{\theta}_k$ with a sample from a Dirichlet with parameters $\boldsymbol{\alpha} + \sum_{i:Y_i=k} \boldsymbol{N}(\boldsymbol{X}_i))$
  - For each document i, replace Y_i with a sample from

$$\mathrm{P}(Y_i = k | \phi, \boldsymbol{\theta}, \boldsymbol{X}_i) \propto \phi_k \prod_{j=1}^{m} \theta_{k,j}^{N_j(\boldsymbol{X}_i)}$$

end

thanks