Inference Methods for Nonparametric Bayesian Models

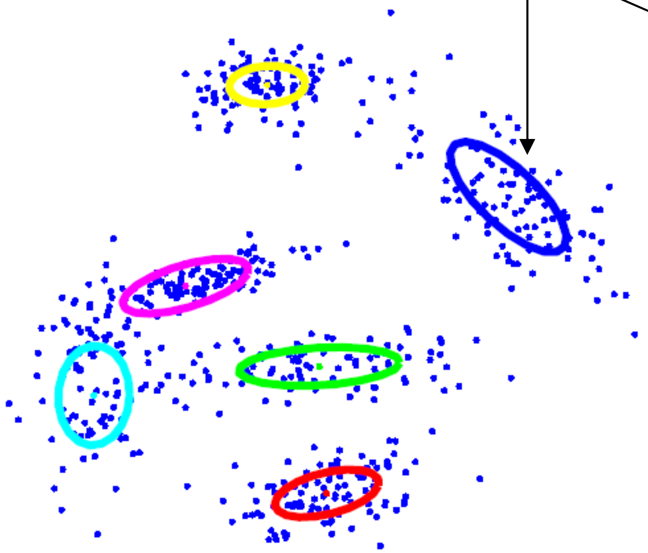
Tutorial: Workshop on Bayesian Inference at ISM

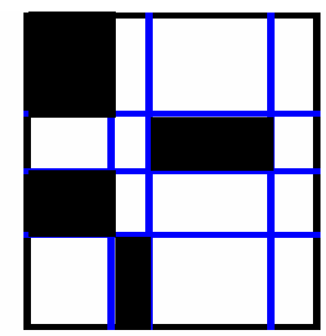
Aug. 21, 2007

NTT Communication Science Laboratories
Kyoto, Japan

Naonori Ueda

In mixture modeling (or latent variable modeling) each sample is assumed to be generated from some unit component model.





Stochastic block model

Gaussian mixture model

Given observed data, we try to estimate latent component models by optimally partitioning data.

Posterior Inference

We (Bayesian) try to maximize posterior over both data partitioning and parameters given observed data

$$p(Z,\Theta \mid D) = \frac{P(D \mid Z,\Theta)P(Z,\Theta)}{\sum_{Z} P(D \mid Z,\Theta)P(Z,\Theta)} \longrightarrow \underset{Z\ \Theta}{\text{Max}}$$

$D = \{x_{1:n}\}$: observed data $\Theta = \{\theta_{(k)}\}$: a set of parameters

$Z = \{z_{1:n}\}$: latent variable $z_i = k$ means that sample *i* belongs to unit cluster *k*

If possible, we want to maximize the marginalized posterior since we want to find clusters:

$$P(Z \mid D) \longrightarrow \underset{Z}{\text{Max}}$$

Important Problems

(1) How define the prior over the partition $P(Z)$

➡ Dirichlet processes (DPs) are available!

(2) How to compute the posterior

$$p(Z,\Theta \mid D) = \frac{P(D \mid Z,\Theta)P(Z,\Theta)}{\sum_{Z} P(D \mid Z,\Theta)P(Z,\Theta)}$$

intractable!

➡ Several methods are proposed

This tutorial explains these two topics

Contents

(1) Brief review of Dirichlet Process Mixture (DPM)

(2) Inference Methods for DPM

- Variational Bayes
- Gibbs sampling

(3) Inference Methods for Hierarchical DPM (HDPM)

- Gibbs sampling

(4) Others

Dirichlet Process (definition)

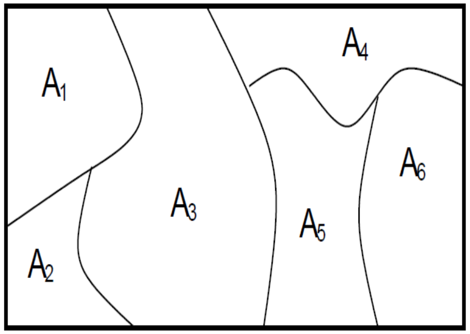
Dirichlet Process (Ferguson, 1973)

$$G = \mathrm{DP}(\gamma, G_0) \quad \text{if and only if}$$

$$G = (P(A_1), \ldots, P(A_K))$$

$$\sim \mathrm{Dirichlet}(G; \gamma G_0(A_1), \ldots, \gamma G_0(A_K))$$

finite partition



$$\{A_i\}_{i=1}^{K}$$

A DP is a distribution over probability

γ : concentration parameter (inverse-variance of DP)

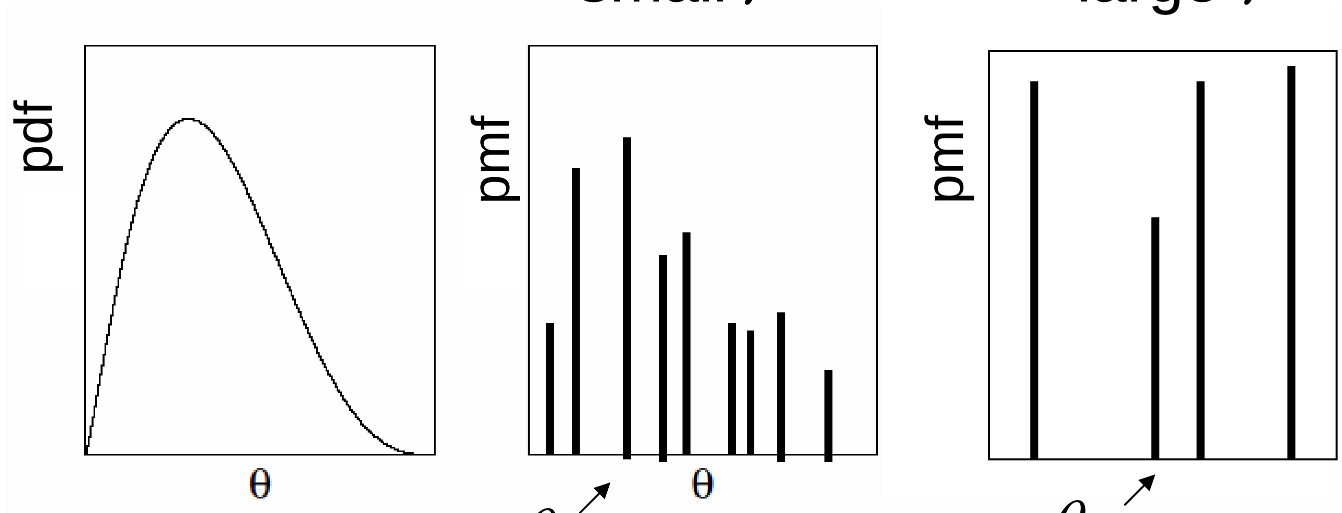
G_0 : base distribution (base measure)

Intuitively, DP can be regarded as an extension of Dirichlet distribution in a continuous domain

What does $G(\theta)$ look like?

$G(\theta) = DP(\gamma, G_0)$

$G_0(\theta)$ small γ large γ



θ $\theta_{(k)}$ $\theta_{(k)}$

clustering effect

Smaller (larger) γ : larger (smaller) number of $\theta_{(k)}$ will appear.

$$\gamma \to \infty \quad G(\theta) \to G_0(\theta)$$

Note: Even if $G_0(\theta)$ is a continuous distribution, $G(\theta)$ becomes a discrete distribution with prob. 1.

Existence of DP

How to construct a DP

(1) Hoppe’s urn model (Polya’s urn model)

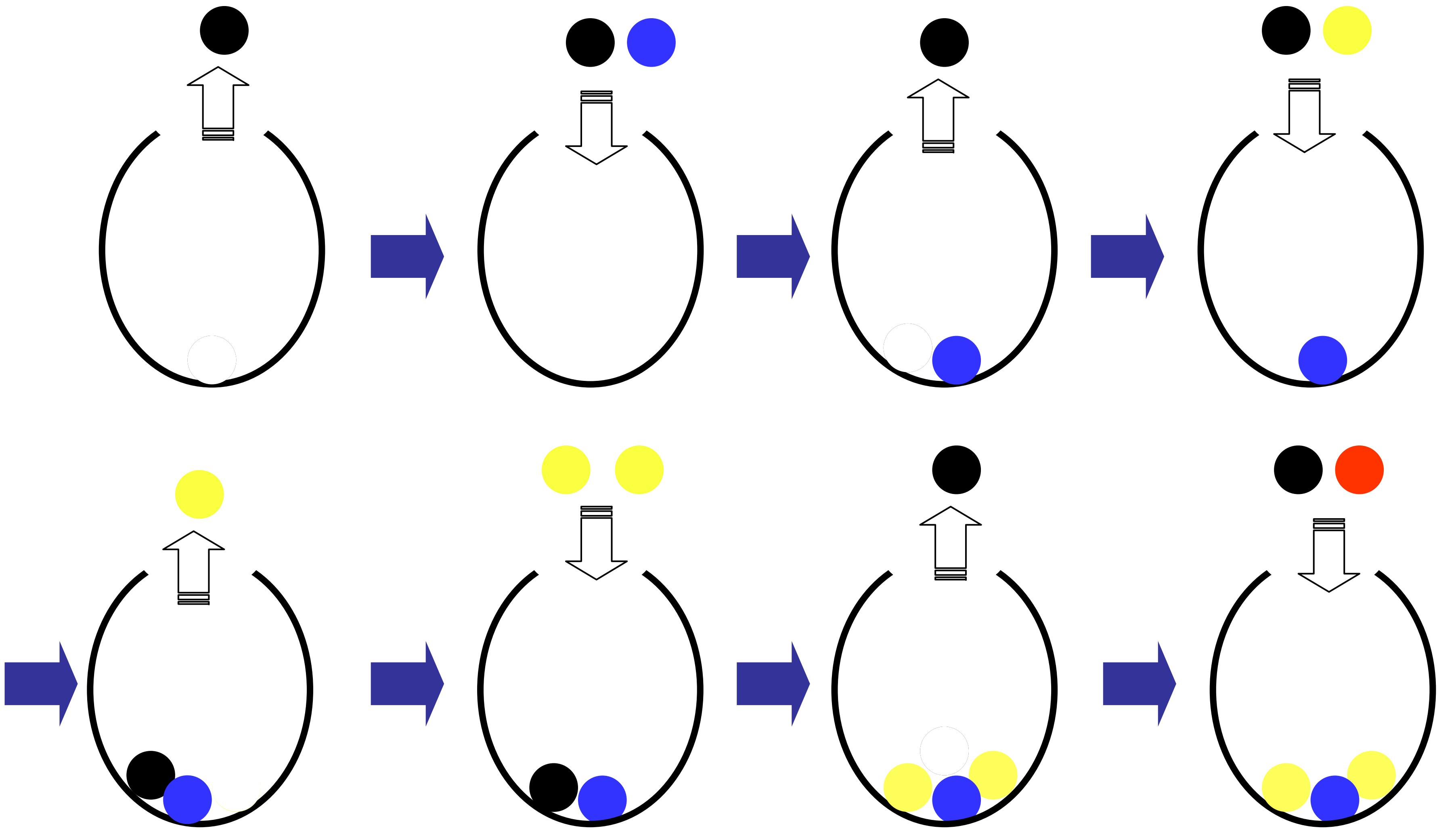
(2) Chinese restaurant process (CRP)

(3) Stick-breaking process

Hoppe's urn model

If the black ball is drawn, insert another color ball with the black ball

If a color ball is drawn, insert the same color ball with the color ball



Hoppe's urn model (cont'd)

By assigning the weight 1 to a color ball and $\gamma(>0)$ to a black ball, we can compute the probability of a draw as follows:

Let $z_i \in \{1,\ldots,K\}$ be a color index,

Then, the conditional probability that the *i*th ball is color *k* given $z_{1:i-1} = \{z_1,\ldots,z_{i-1}\}$

$$P(z_i = k \mid z_{1:i-1}, \gamma) = \begin{cases} \dfrac{m_k}{i-1+\gamma} & \text{if } k \text{ is an old color} \\ \dfrac{\gamma}{i-1+\gamma} & \text{if } k \text{ is a new color} \end{cases}$$

Note: $\sum_{k=1}^{K} m_k = i-1$

m_k: # of balls with color *k* given $z_{1:i-1}$

K: # of existing colors given $z_{1:i-1}$

Probability of a partition

weight 1 to a color ball and weight $\gamma (> 0)$ to a black ball

$$P(R,Y,Y,R,R,G) = \frac{\gamma}{\gamma} \times \frac{\gamma}{1+\gamma} \times \frac{1}{2+\gamma} \times \frac{1}{3+\gamma} \times \frac{2}{4+\gamma} \times \frac{\gamma}{5+\gamma} = \frac{\gamma^3 2!}{AF(\gamma, 6)}$$

Note: Ascending factorial: $AF(a,n) = a(a+1)\cdots(a+n-1)$

$$P(G,Y,Y,R,R,R) = \frac{\gamma}{\gamma} \times \frac{\gamma}{1+\gamma} \times \frac{1}{2+\gamma} \times \frac{\gamma}{3+\gamma} \times \frac{1}{4+\gamma} \times \frac{2}{5+\gamma} = \frac{\gamma^3 2!}{AF(\gamma, 6)}$$

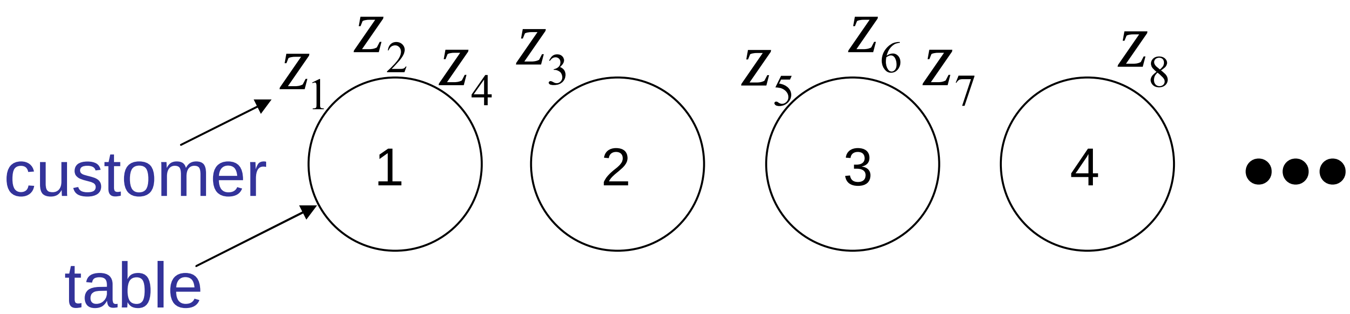
➡ The order does not affect the probability.
An exchangeable process.

In general,

$$P(Z) = \frac{\gamma^K \prod_{k=1}^{K} (m_k - 1)!}{AF(\gamma, n)}$$

Chinese Restaurant Process (Aldous, 1985)

A customer prefers to sit at the crowed table



The probability that the *i*th customer will sit at the *k*th table is:

$$P(z_i = k \mid z_{1:i-1}, \gamma) = \begin{cases} \dfrac{m_k}{i-1+\gamma} & \text{if } k \text{ is a old table} \\ \dfrac{\gamma}{i-1+\gamma} & \text{if } k \text{ is a new table} \end{cases}$$

(# of customers at table *k* → m_k)

This is the same as Hoppe's urn scheme!

DP construction theorem

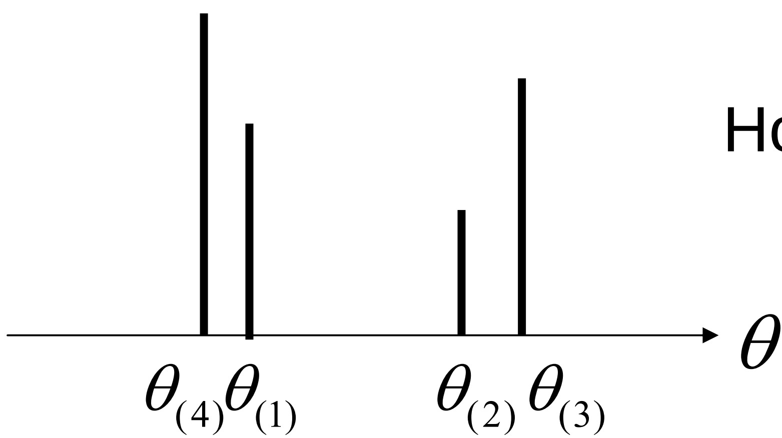
(Sethuraman, 1994)

Theorem

$G \sim \mathrm{DP}(\alpha, G_0)$ can be represented by

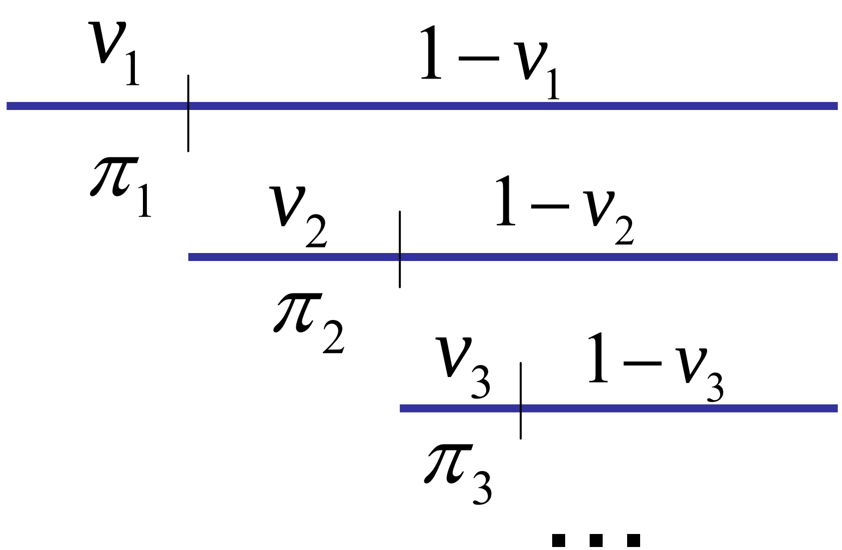
$$G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_{(k)}}(\theta)$$

where $\pi_k \geq 0, \quad \sum_{k=1}^{\infty} \pi_k = 1 \qquad \theta_{(k)} \sim G_0(\theta)$



How do we get this ?

Stick-breaking process (SBP)



stick whose length is 1

break the stick with the ratio

$$v_j : 1 - v_j$$

$$\therefore \pi_k = v_k \prod_{j=1}^{k-1} \left(1 - v_j\right)$$

$$v_j \sim \mathrm{Beta}\left(v;\ 1, \gamma\right)$$ Beta distribution

larger γ $\longrightarrow$ smaller v_j $\longrightarrow$ larger # of components

Other SBP

Beta Two-parameter process (Ishwaran & Zarepour, 2000)

$$v_k \sim \mathrm{Beta}(v; a, b)$$

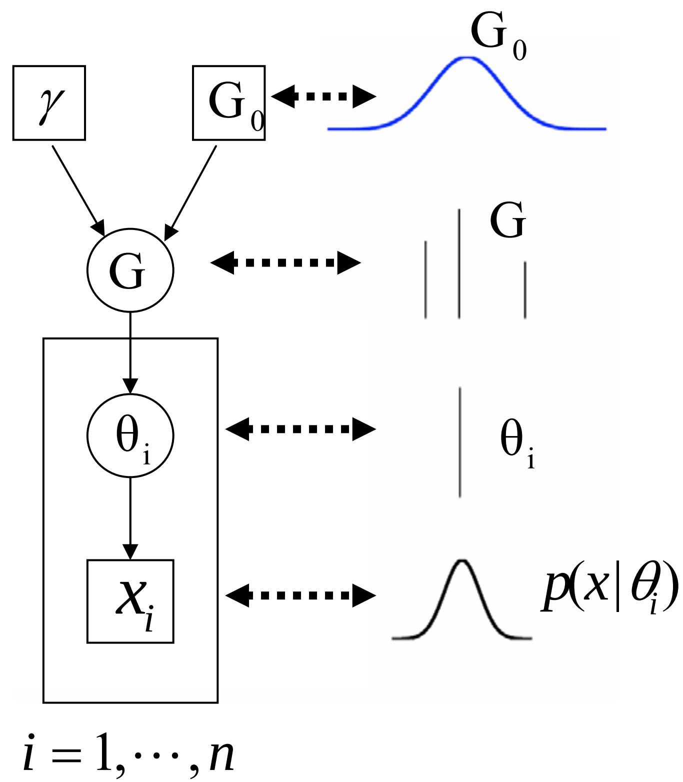
Pitman-Yor process (two-parameter Poisson-Dirichlet, Pitman & Yor, 1997)

$$v_k \sim \mathrm{Beta}(v; 1-a, b+ka)$$

Dirichlet Process Mixture (DPM) Models

In DPM, a DP is used as a parameter generation process.

$$G \sim \mathrm{DP}(\gamma, \mathrm{G}_0)$$
$$\theta_i \mid G \sim G, \ \text{for } i = 1, \ldots, n$$
$$x_i \sim p(x \mid \theta_i), \ \text{for } i = 1, \ldots, n$$



Usually G_0 is set to a prior conjugate to p(x|.)

Note: Due to the clustering effect of DP, some of data can share the same parameter $\theta_{(k)}$.

mixing weights

By integrating out G, we have

$$P(\theta_i \mid \theta_{1:i-1}) = \frac{\gamma}{i-1+\gamma} G_0(\theta_i) + \frac{1}{i-1+\gamma} \sum_{k=1}^{K} m_k \delta_{\theta_{(k)}}(\theta_i)$$

DPM construction without G

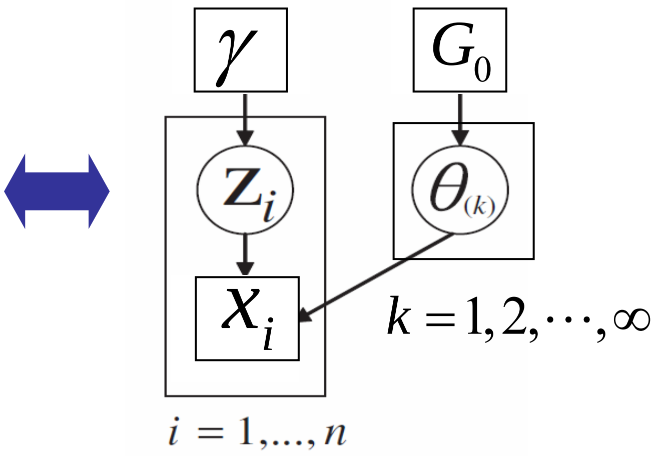
CRP

$$Z \sim \text{CRP}(\gamma)$$

$$\theta_{(k)} \mid G_0 \sim G_0$$

$$x_i \sim p(x \mid \theta_{(z_i)}), \ \text{for } i = 1, \ldots, n$$

The *k*th table corresponds to $\theta_{(k)}$



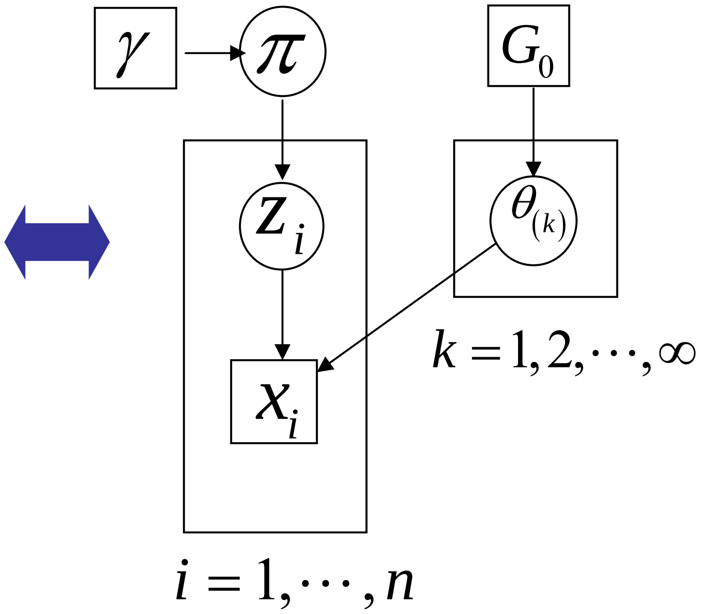
SBP

$$\pi \sim \text{Stick}(\gamma) \leftarrow \quad v_j \sim \text{Beta}(1, \gamma), \quad \pi_k = v_k \prod_{j=1}^{k-1} (1 - v_j)$$

$$\theta_{(k)} \mid G_0 \sim G_0$$

$$z_i \mid \pi \sim \text{Discrete}(z_i; \pi)$$

$$x_i \sim p(x \mid \theta_{(z_i)}), \ \text{for } i = 1, \ldots, n$$



Inference Methods

(1) SBP-DPM

- Variational method

(2) CRP-DPM

- Gibbs sampling and Collapsed version

Variational Bayes

We consider the log-evidence:

latent variable set

$$\mathcal{L}(D) = \log p(D) = \log \sum_{Z} \int p(D, Z, \theta) d\theta$$

$$= \log \sum_{Z} \int q(Z, \theta) \frac{p(D, Z, \theta)}{q(Z, \theta)} d\theta$$

$$\geq \sum_{Z} \int q(Z, \theta) \log \frac{p(D, Z, \theta)}{q(Z, \theta)} d\theta \triangleq F[q]$$

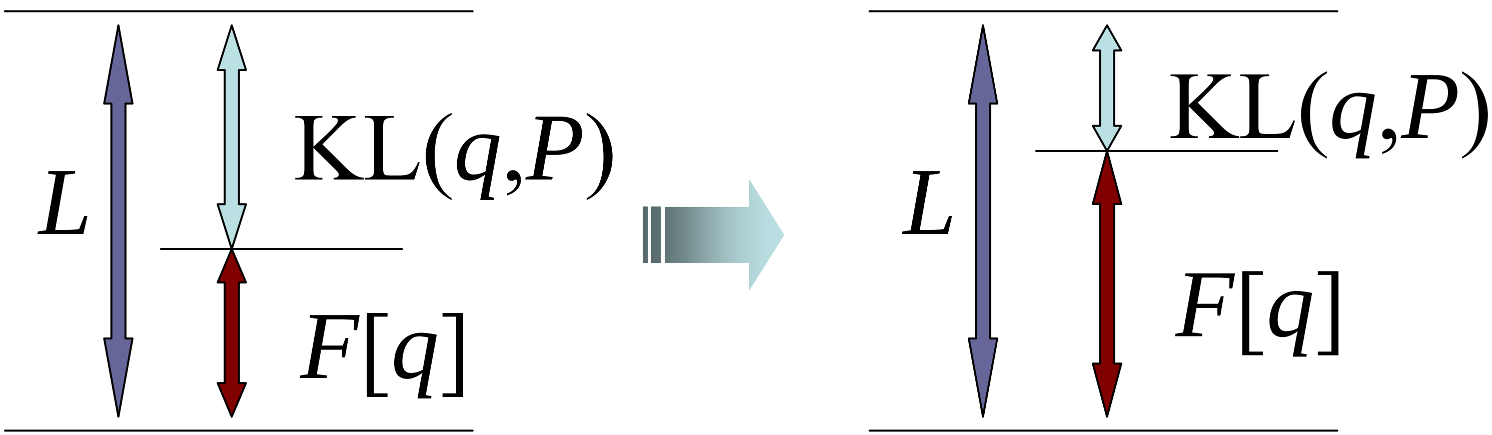
Jensen' inequality $\log \mathrm{E}\{f(x)\} \geq \mathrm{E}\{\log f(x)\}$

$F[q]$ gives a lower bound of the log-evidence

Important Relationship between *L* and *F*

$$L(D) = F[q] + \mathrm{KL}(q, P(Z, \theta \mid D))$$

log-evidence   variational posterior   true posterior



Maxmizing F[q] w.r.t. q = Minimizing KL

Optimizing Variational Posteriors

Z,θ are assumed to be independent.

➡ $q(Z,\theta) \cong q(Z)q(\theta)$

Then

$$F[q] = \left\langle \log \frac{p(D,Z \mid \theta)}{q(Z)} \right\rangle_{q(Z)q(\theta)} + \left\langle \log \frac{p(\theta)}{q(\theta)} \right\rangle_{q(\theta)}$$

Note: $< f(x) >= \int f(x)p(x)dx$

➡ By using the variational calculus, we can obtain optimal $q(Z), q(\theta)$

Solution

$$\text{Max} \quad F[q] \quad \text{w.r.t. } q(Z)$$

$$q(Z) = \frac{1}{C_Z} \exp\left\{ \left\langle \log p(D, Z \mid \Theta) \right\rangle_{q(\Theta)} \right\}$$

normalizing constant

$$\text{Max } F[q] \text{ w.r.t. } q(\Theta)$$

$$q(\Theta) = \frac{1}{C_\Theta} p(\Theta) \exp\left\{ \left\langle \log p(D, Z \mid \Theta) \right\rangle_{q(Z)} \right\}$$

These are mutually dependent, so we iteratively estimate each of them.

Variational Bayes (VB)EM algorithm

Step 1. Initialization. Set $t \leftarrow 0$

Step 2. Repeat EM-steps until convergence

VB-Estep:

$$q(Z)^{(t+1)} = \frac{1}{C_Z} \exp\left\langle \log p(D, Z \mid \theta) \right\rangle_{q(\theta)^{(t)}}$$

VB-Mstep:

$$q(\Theta)^{(t+1)} = \frac{1}{C_\Theta} p(\Theta) \exp\left\langle \log p(D, Z \mid \Theta) \right\rangle_{q(Z)^{(t+1)}}$$

Set $t \leftarrow t+1$

Modification to SBP-DPM

(Blei, et al., 2004)

(1) Introduction of DP prior over *Z*

In parametric Bayes, prior over Z is not specified.

$$P(z_i = k \mid V) = v_k \prod_{j=1}^{k-1} (1 - v_j)$$ ← SBP

Note: $I(f) = 1(0)$

f is true (false)

$$P(z_i \mid V) = \prod_{k=1}^{\infty} \left(v_k \prod_{j=1}^{k-1} (1 - v_j)\right)^{I(z_i = k)}$$

$$= \prod_{k=1}^{\infty} v_k^{I(z_i = k)} (1 - v_k)^{I(z_i > k)}$$

(2) Truncation

$$P(z_i \mid V) = \prod_{k=1}^{T} v_k^{I(z_i = k)} (1 - v_k)^{I(z_i > k)}$$

To make the computation feasible, the no. of components is truncated.

Optimal Posteriors

$$q(Z) = \frac{1}{C_Z} \exp\left\{ \left\langle \log P(Z|V) \right\rangle_{q(V)} + \left\langle \log p(D, Z|\Theta) \right\rangle_{q(\Theta)} \right\}$$

$$q(\Theta) = \frac{1}{C_\Theta} p(\Theta) \exp\left\{ \left\langle \log p(D|Z, \Theta) \right\rangle_{q(Z)} \right\}$$

$$q(V) = \frac{1}{C_V} \exp\left\{ \left\langle \log p(Z|V) \right\rangle_{q(Z)} + \left\langle \log p(V|\gamma) \right\rangle_{q(\gamma)} \right\}$$

$$q(\gamma) = \frac{1}{C_\gamma} p(\gamma) \exp\left\{ \left\langle \log p(V|\gamma) \right\rangle_{q(V)} \right\}$$

These are iteratively estimated in the same manner as the conventional VB-EM.

The introduction of prior over *Z* requires the extra computations. But, they are straightforwardly computed.

Gibbs sampling

Our goal it to sample θ with a distribution $p(\theta)$ using Markov chain.

Suppose $\theta = \left(\theta_1, \ldots, \theta_d\right)$

We just sample one variable with the remaining variables fixed.

$$\text{for } t = 1, 2, \ldots$$

$$\text{for } i = 1, \ldots, d$$

$$\theta_i^{(t+1)} \sim p(\theta_i \mid \theta_{-i}^{(t)})$$

where $\theta_{-i}^{(t)} = \left(\underbrace{\theta_1^{(t+1)}, \ldots, \theta_{i-1}^{(t+1)}}, \underbrace{\theta_{i+1}^{(t)}, \ldots, \theta_d^{(t)}}\right)$

values at (t+1)-step values at t-step

Naive Gibbs sampling for DPM

To compute the conditional distribution given the other variables, we make use of the exchangeability and treat θ_i as the last variable being sampled.

Noting that

$$P(\theta_i = \theta \mid \theta_{-i}) = \frac{\gamma}{n-1+\gamma} G_0 + \frac{1}{n-1+\gamma} \sum_{\substack{j=1 \\ j \neq i}}^{n} \delta_{\theta_j}(\theta)$$

We have

$$P(\theta_i \mid \theta_{-i}, x_i) = \frac{p(x_i \mid \theta_i) P(\theta_i \mid \theta_{-i})}{\int p(x_i \mid \theta_i) P(\theta_i \mid \theta_{-i}) d\theta_i}$$

This method is inefficient for large *n*

We never use this!

Smart Method using CRP-DPM

Rather than sampling θ_i , we sample z_i,conditioned on $z_{-i} = \{z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n\}$

Then $$P(z_i = k \mid z_{-i}, x_i, \Theta) = \frac{P(z_i = k, x_i \mid z_{-i}, \Theta)}{P(x_i \mid z_{-i}, \Theta)}$$

$$\propto P(z_i = k \mid z_{-i})\, p(x_i \mid z_i = k, \Theta)$$

According to the exchangeability, we can regard z_i as the last customer to arrive after the other customers are seated.

CRP $$P(z_i = k \mid z_{-i}) = \begin{cases} \dfrac{m_{-i,k}}{n-1+\gamma} & \text{if } k \text{ is old} \\ \dfrac{\gamma}{n-1+\gamma} & \text{if } k \text{ is new} \end{cases}$$

where $m_{-i,k} = \#\{j;\ z_j = k \text{ for all } j \neq i\}$

Moreover,

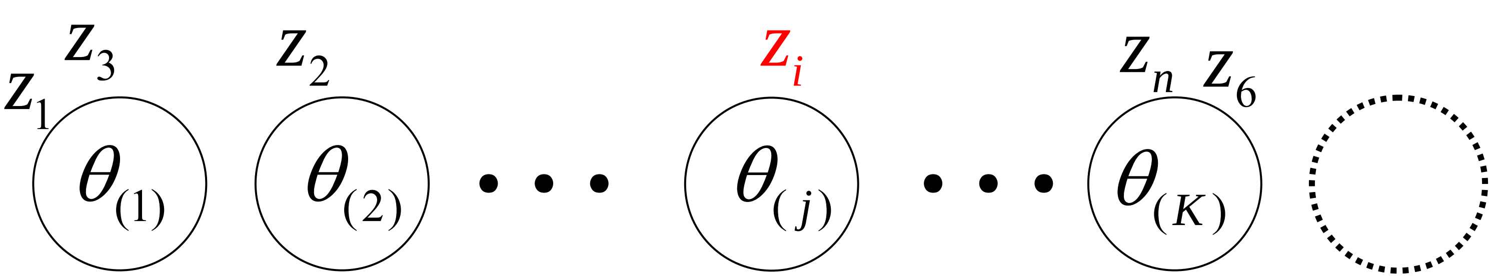
$$P(x_i \mid z_i = k, \Theta) = \begin{cases} p(x_i \mid \theta_{(k)}) & \text{if } k \text{ is old} \\ \int p(x_i \mid \theta) G_0(\theta) d\theta & \text{if } k \text{ is new} \end{cases}$$

$$\text{or } p(x_i \mid \theta_{(k^{new})}) \quad \theta_{(k^{new})} \sim G_0 \quad \text{if } k \text{ is new}$$

Finally, we have

For all $k \in \{z_1, \ldots, z_n\}$

$$P(z_i = k \mid z_{-i}, x_i, \Theta) \propto \begin{cases} m_{-i,k} p(x_i \mid \theta_{(k)}) & \text{if } k \text{ is old} \\ \gamma \int p(x_i \mid \theta) G_0(\theta) d\theta & \text{if } k \text{ is new} \end{cases}$$



Gibbs sampling for DPM (West et al., 1994)

Let the state of the Markov chain consist of $z_1, \ldots, z_n$ and $\Theta = \{\theta_{(k)};\ k \in \{z_1, \ldots, z_n\}\}$. Repeatedly sample as follows:

[1] If z_i is associated with no other z_{-i}, then remove $\theta_{(z_i)}$ from Θ.

Draw a new value for z_i from $P(z_i \mid z_{-i}, x_i, \Theta)$

If the new z_i is not associated with z_{-i}, then draw a value for $\theta_{(z_i)}$ from $P(\theta_i \mid x_i)$ and add it to Θ.

Here $P(\theta_i \mid x_i) = \dfrac{p(x_i \mid \theta_i) G_0(\theta_i)}{\int p(x_i \mid \theta) G_0(\theta) d\theta}$

[2] For $k \in \{z_1, \ldots, z_n\}$: Draw a new value from

$$P(\theta_{(k)} \mid \{x_s\} \text{ such that } z_s = k) = \frac{\prod_{s:z_s=k} p(x_s \mid \theta_{(k)}) G_0(\theta_{(k)})}{\int \prod_{s:z_s=k} p(x_s \mid \theta) G_0(\theta) d\theta}$$

Marginalized version (MacEachern, 1994)

Eliminating Θ from the previous algorithm

$$P(z_i = k \mid z_{-i}, x_i, x_{-i}) \propto \begin{cases} m_{-i,k} \int p(x_i \mid \theta_{(k)}) H_{-i,k} \, d\theta_{(k)} & \text{if } k \text{ is old} \\ \gamma \int p(x_i \mid \theta) G_0(\theta) \, d\theta & \text{if } k \text{ is new} \end{cases}$$

H is the posterior prob. of $\theta_{(k)}$, conditioned on $z_{s \neq i}$ that fall into cluster *k*.

$$H_{-i,k} = p(\theta_{(k)} \mid x_{-i}, z_{s \neq i} = k) = \frac{\prod_{s: z_s = k, s \neq i} p(x_s \mid \theta_{(k)}) G_0(\theta_{(k)})}{\int \prod_{s: z_s = k, s \neq i} p(x_s \mid \theta) G_0(\theta) \, d\theta}$$

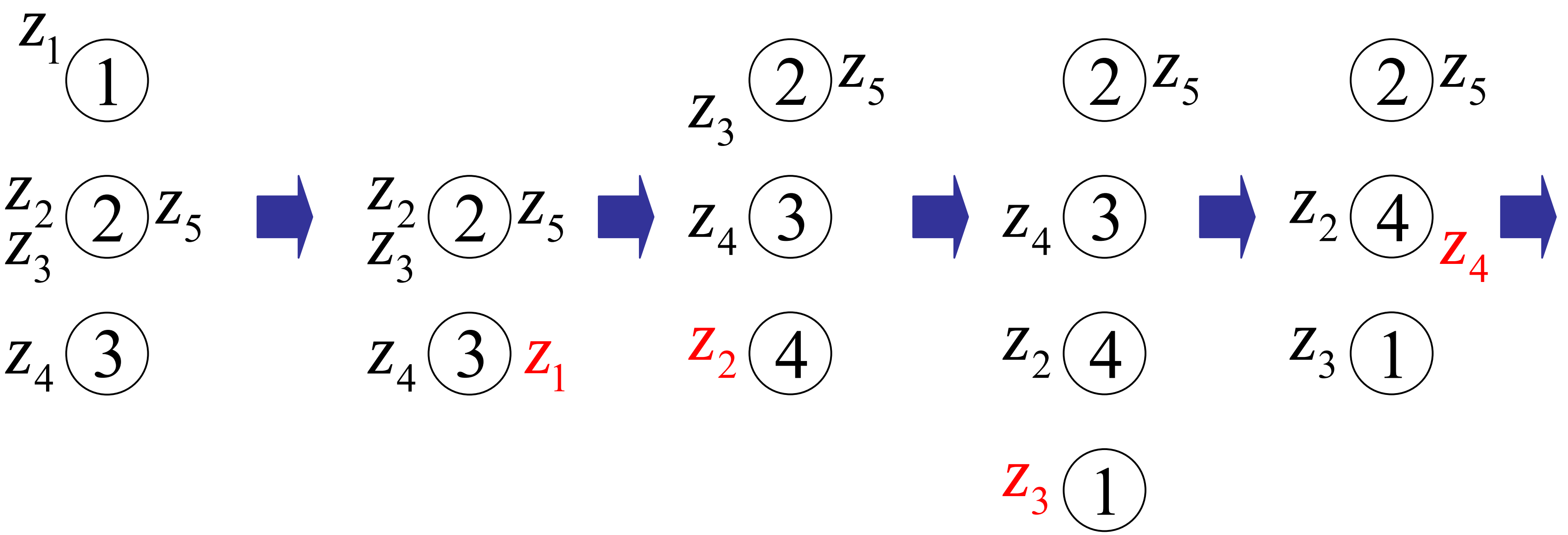
Modified algorithm: Let the state of the Markov chain consist of $z_1, \ldots, z_n$ Repeatedly sample as follows:

For *i* =1,…,*n*: Draw z_i from $P(z_i \mid z_{-i}, x_i, x_{-i})$

$$P(z_i = k \mid z_{-i}, x_i, x_{-i}) \propto \begin{cases} m_{-i,k}\int p(x_i \mid \theta_{(k)}) H_{-i,k}\, d\theta_{(k)} = a_{ik} & \text{if } k \text{ is old} \\ \gamma \int p(x_i \mid \theta) G_0(\theta)\, d\theta = b_i & \text{if } k \text{ is new} \end{cases}$$

constant

$n = 5$ case



We just want to cluster data. Table ID is not important.

Hierarchical Dirichlet Process (HDP)

(Teh, et al., 2004)

Assume we have *J* groups data

Each group has clusters, and these clusters are shared between groups. Such model can be realized by HDP.

$$G_0 \sim \mathrm{DP}(\gamma, \mathrm{H})$$

$$\text{for } j = 1, \ldots, J$$

$$G_j \sim \mathrm{DP}(\gamma, \mathrm{G}_0)$$

$$\text{for } i = 1, \ldots, n_j$$

$$\phi_{ji} \mid G_j \sim G_j$$

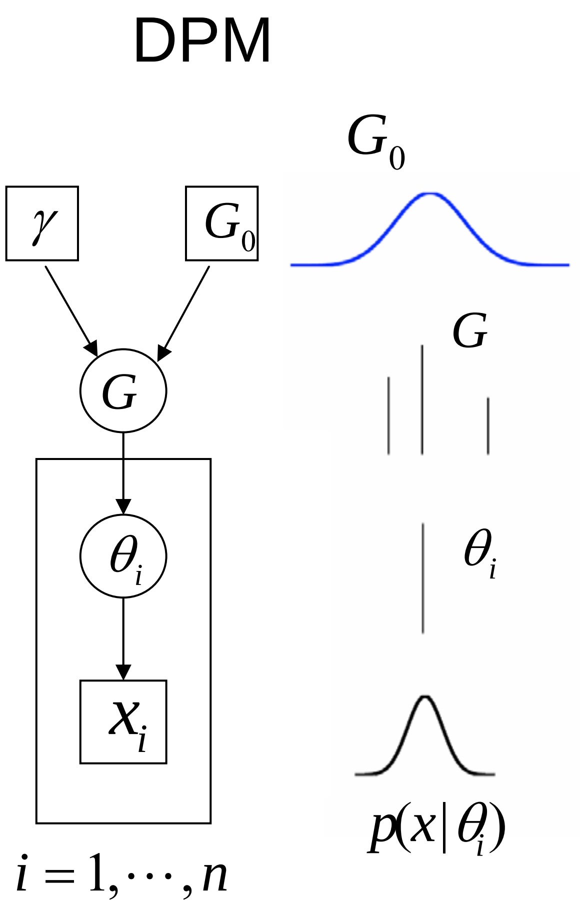
$$x_{ji} \sim p(x_{ji} \mid \phi_{ji})$$

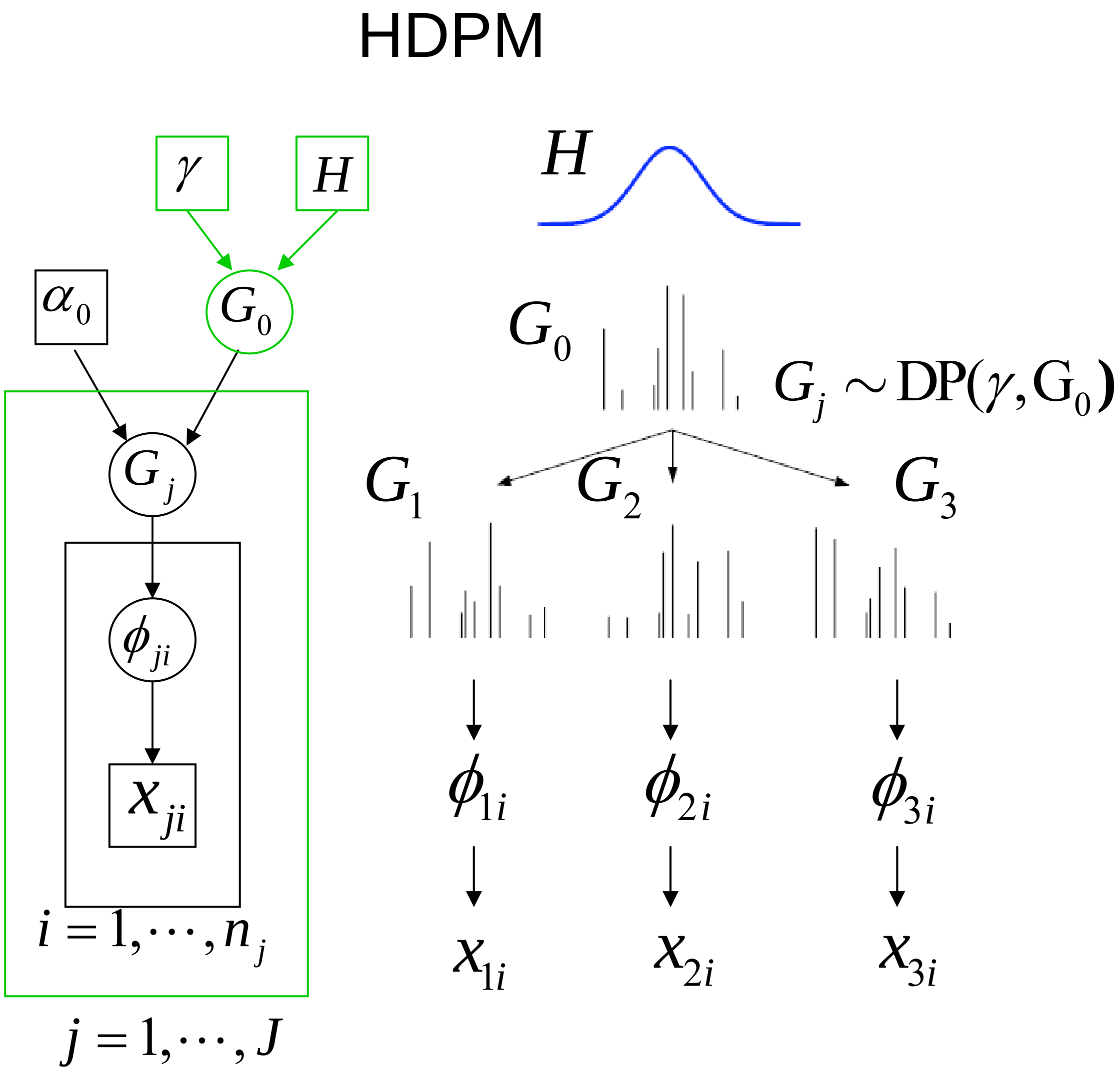
Note:

$\{G_j\}$ are conditionally independent given G_0

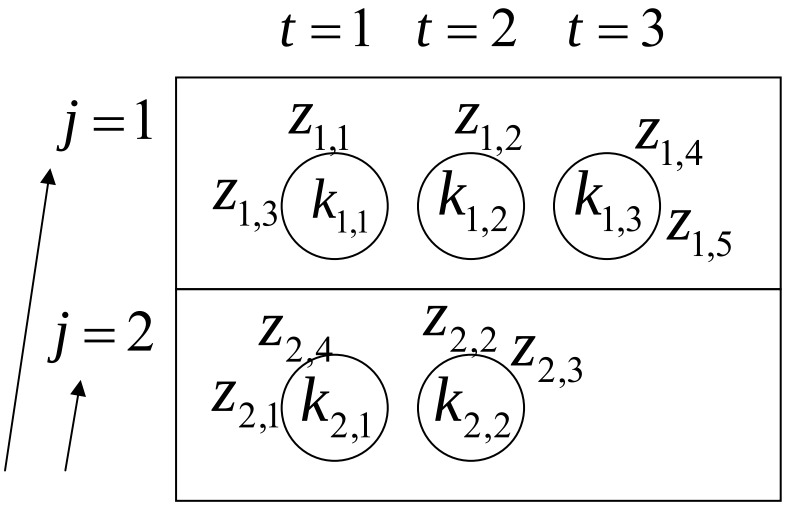
Note: $\phi_{ji} \in \{\theta_{(1)}, \ldots, \theta_{(K)}\}$

Parameters can be shared not only within groups, but also between groups





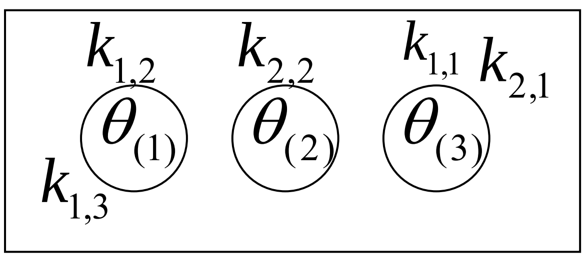
Chinese Restaurant Franchise



restaurant index

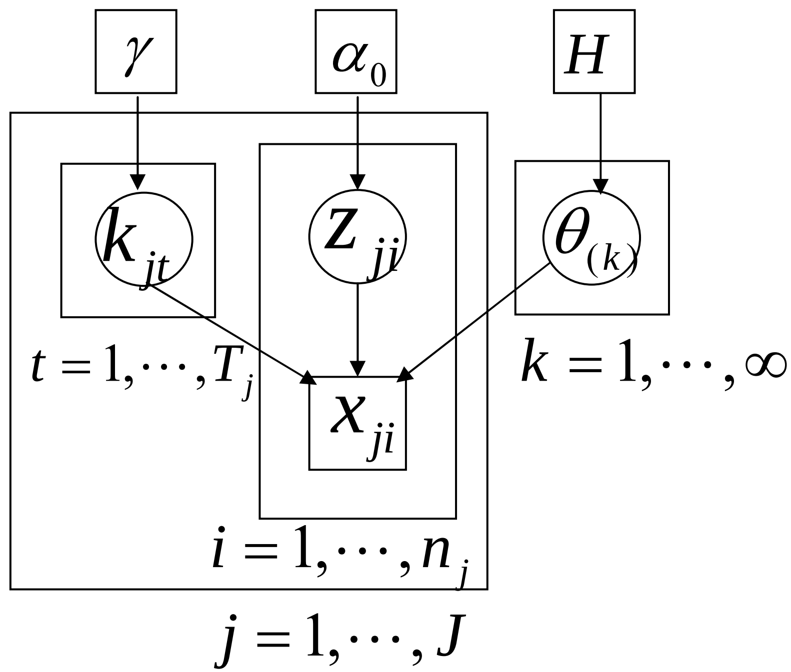
$z_{ji} \in \{1,..,T_j\}$ table index

$k_{jt} \in \{1,..,K\}$ dish index

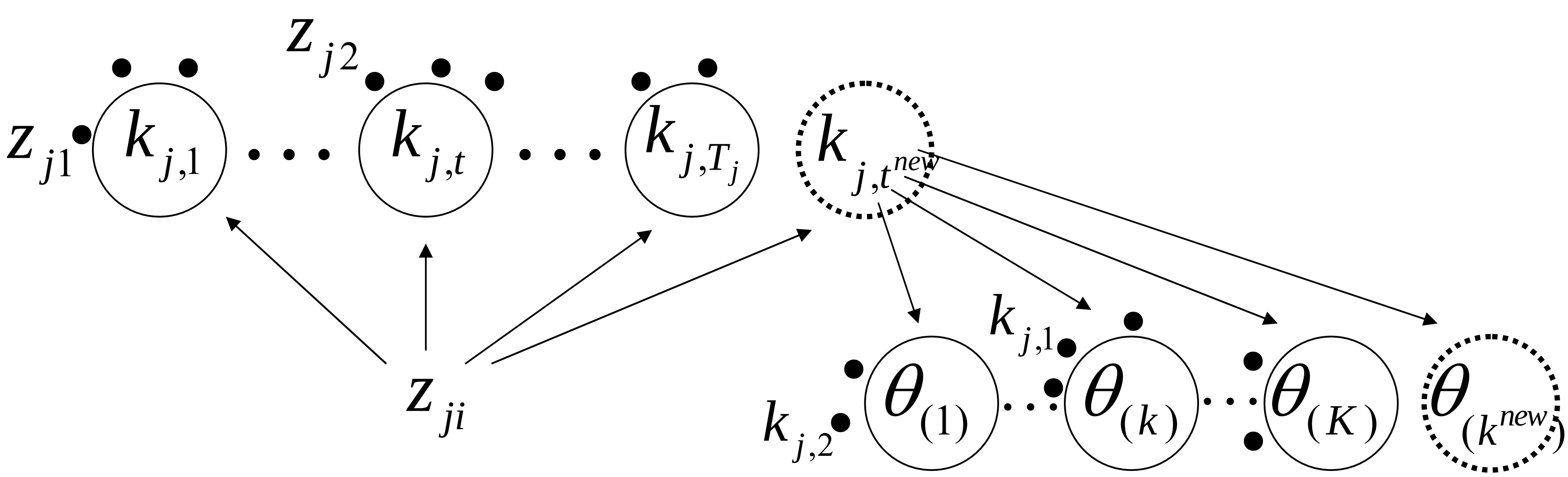


ex) If $k_{1,1}=3$, then $x_{1,1}, x_{1,3} \sim p(x \mid \theta_{(3)})$

HDPM



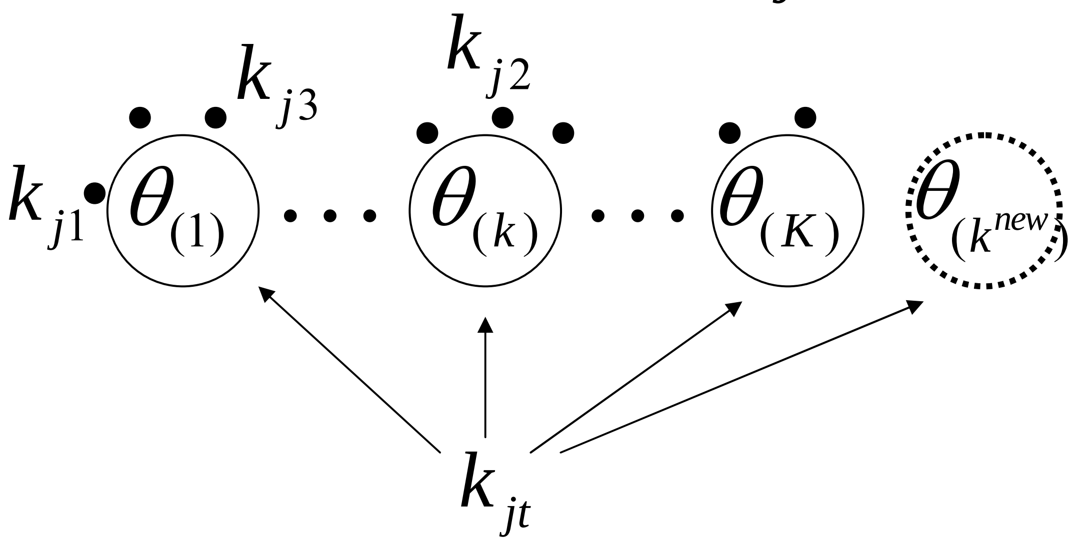
Sampling z_{ji}



$$P(z_{ji}=t \mid z^{-ji}, \boldsymbol{k}, \Theta, x_{ji}) \propto P(z_{ji}=t \mid z^{-ji})\, p(x_{ji} \mid z_{ji}=t, \Theta)$$

$$\propto \begin{cases} n_{jt}^{-i}\, p(x_{ji} \mid \theta_{(k_{jt})}) & \text{if } t \text{ is old} \\ \alpha_0\, p(x_{ji} \mid \theta_{(k_{jt})}) & \text{if } t \text{ is new} \end{cases}$$

Sampling k_{jt}



$$P(k_{jt}=k \mid k^{\text{-}jt},\mathbf{z},\Theta,X) \propto P(k_{jt}=k \mid k^{-ji})\, p(X \mid k_{jt}=k,\mathbf{z},\Theta)$$

$$\propto \begin{cases} m_k^{-t} \prod_{s:\, z_{js}=t} p(x_{js} \mid \theta_{(k)}) & \text{if } k \text{ is old} \\ \gamma \prod_{s:\, z_{js}=t} p(x_{js} \mid \theta_{(k)}) & \text{if } k \text{ is new} \end{cases}$$

Likelihood when setting $k_{jt}=k$

Sampling $\theta_{(k)}$

$$P(\theta_{(k)} \mid \Theta^{-k}, \boldsymbol{z}, \boldsymbol{k}, X) \propto P(\theta_{(k)}) P(X \mid \theta_{(k)}, \boldsymbol{k}, \boldsymbol{z})$$

$$= H(\theta_{(k)}) \prod_{j=1}^{J} \prod_{i=1}^{n_j} p(x_{ji} \mid \theta_{(k)})^{\overbrace{I(z_{ji} = t \,\&\, k_{jt} = k)}^{}_{I(k_{jz_{ji}} = k)}}$$

$$= H(\theta_{(k)}) \prod_{ji:k_{jz_{ji}}=k}^{J} p(x_{ji} \mid \theta_{(k)})$$

Likelihood associated with $\theta_{(k)}$

Recent Efficient Methods

Particle filters for mixture models with an unknown number of components (Fearnhead, 2004)

sequential sampler

Accelerated Variational DPM (Kurihara et al., 2006)

Incorporating the *K*d-trees to VB

Fast search for DPM (Daume, 2007)

A-star search

A Permutation-Augmented Sampler (Liang et al., 2007)

global move sampler

Selected References

Original paper of DP：

- Ferguson T. S., “A Bayesian analysis of some nonparametric problems,” Annals of Statistics, vol.1, pp. 209-230, 1973.

DPM-VB：

- Blei, David M. and Michael I. Jordan. “Variational inference for Dirichlet process mixtures.” Bayesian Analysis 1(1), 2004.

MCMC for DPM：

- Neal, R. M. “Markov chain sampling methods for Dirichlet process mixture models", Journal of Computational and Graphical Statistics, vol. 9, pp. 249-265:

HDP：

- Teh, Y. W., Jordan, M. I, Beal, M. J. and Blei, D. M., “Hierarchical Dirichlet processes,” Tech. report 653, Dept. of statistics, Univ. of California, Berkeley, 2004.

Tutorial：

- http://www.cs.berkeley.edu/~jordan/nips-tutorial05.ps
  Michael Jordan's NIPS 2005 tutorial: Nonparametric Bayesian Methods:
  Dirichlet Processes, Chinese Restaurant Processes and All That.