Hierarchical Dirichlet Processes

Presenters:
Micah Hodosh, Yizhou Sun

4/7/2010

Content

- Introduction and Motivation
- Dirichlet Processes
- Hierarchical Dirichlet Processes
  - Definition
  - Three Analogs
- Inference
  - Three Sampling Strategies

Introduction

- Hierarchical approach to model-based clustering of grouped data
- Find an unknown number of clusters to capture the structure of each group and allow for sharing among the groups
  - Documents with an arbitrary number of topics which are shared globably across the set of corpora.
- A Dirichlet Process will be used as a prior mixture components
- The DP will be extended to a HDP to allow for sharing clusters among related clustering problems

Motivation

- Interested in problems with observations organized into groups
- Let x_{ji} be the ith observation of group j = $\mathbf{x_j} = \{x_{j1}, x_{j2} \ldots\}$
- x_{ji} is exchangeable with any other element of $\mathbf{x_j}$
- For all j,k , $\mathbf{x_j}$ is exchangeable with $\mathbf{x_k}$

Motivation

- Assume each observation is drawn independently for a mixture model
  - Factor θ_{ji} is the mixture component associated with x_{ji}
- Let $F(\theta_{ji})$ be the distribution of x_{ji} given θ_{ji}
- Let G_j be the prior distribution of θ_{j1}, θ_{j2}... which are conditionally independent given G_j

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

$$x_{ji} \mid \theta_{ji} \sim F(\theta_{ji})$$

Content

- Introduction and Motivation
- Dirichlet Processes
- Hierarchical Dirichlet Processes
  - Definition
  - Three Analogs
- Inference
  - Three Sampling Strategies

The Dirichlet Process

- Let (Θ , β) be a measureable space,
  - Let G_0 be a probability measure on that space
  - Let $\mathbf{A} = (A_1, A_2 .., A_r)$ be a finite partition of that space
  - Let α_0 be a positive real number
- $G \sim DP(\alpha_0, G_0)$ is defined s.t. for all $\mathbf{A}$:

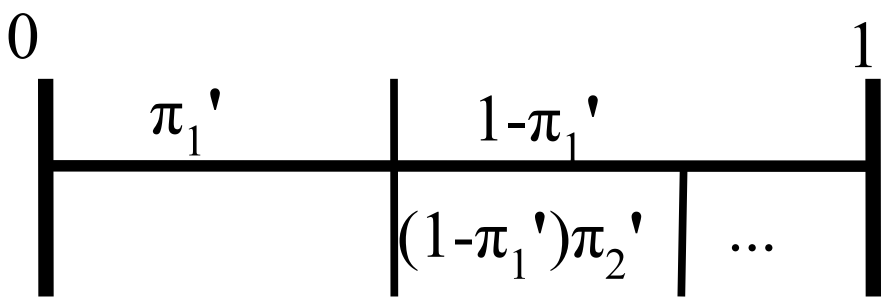
$$(G(A_1), \ldots, G(A_r)) \sim \mathrm{Dir}(\alpha_0 G_0(A_1), \ldots, \alpha_0 G_0(A_r))$$

Stick Breaking Construction

- The general idea is that the distribution G will be a weighted average of the distributions of a set of infinite random variables
- 2 infinite sets of i.i.d random variables
  - $\varphi_k \sim G_0$ – Samples from the initial probability measure
  - $\pi_k' \sim \text{Beta}(1, \alpha_0)$ – Defines the weights of these samples

Stick Breaking Construction

- $\pi_k' \sim \text{Beta}(1, \alpha_0)$
- Define π_k as $\pi_k = \pi_k' \prod_{l=1}^{k-1} (1 - \pi_l')$



$$\sum_{1}^{\infty} \pi_k = 1$$

Stick Breaking Construction

- $\pi_k \sim \text{GEM}(\alpha_0)$
- These π_k define the weight of drawing the value corresponding to φ_k.

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

Polya urn scheme/ CRP

- Let each θ_1, θ_2,.. be i.i.d. Random variables distributed according to G
- Consider the distribution of θ_i, given θ_1,...θ_{i-1}, integrating out G:

$$\theta_i \mid \theta_1, \ldots, \theta_{i-1}, \alpha_0, G_0 \sim \sum_{\ell=1}^{i-1} \frac{1}{i-1+\alpha_0} \delta_{\theta_\ell} + \frac{\alpha_0}{i-1+\alpha_0} G_0 \, .$$

Polya urn scheme

$$\theta_i \mid \theta_1, \ldots, \theta_{i-1}, \alpha_0, G_0 \sim \sum_{\ell=1}^{i-1} \frac{1}{i-1+\alpha_0} \delta_{\theta_\ell} + \frac{\alpha_0}{i-1+\alpha_0} G_0 \,.$$

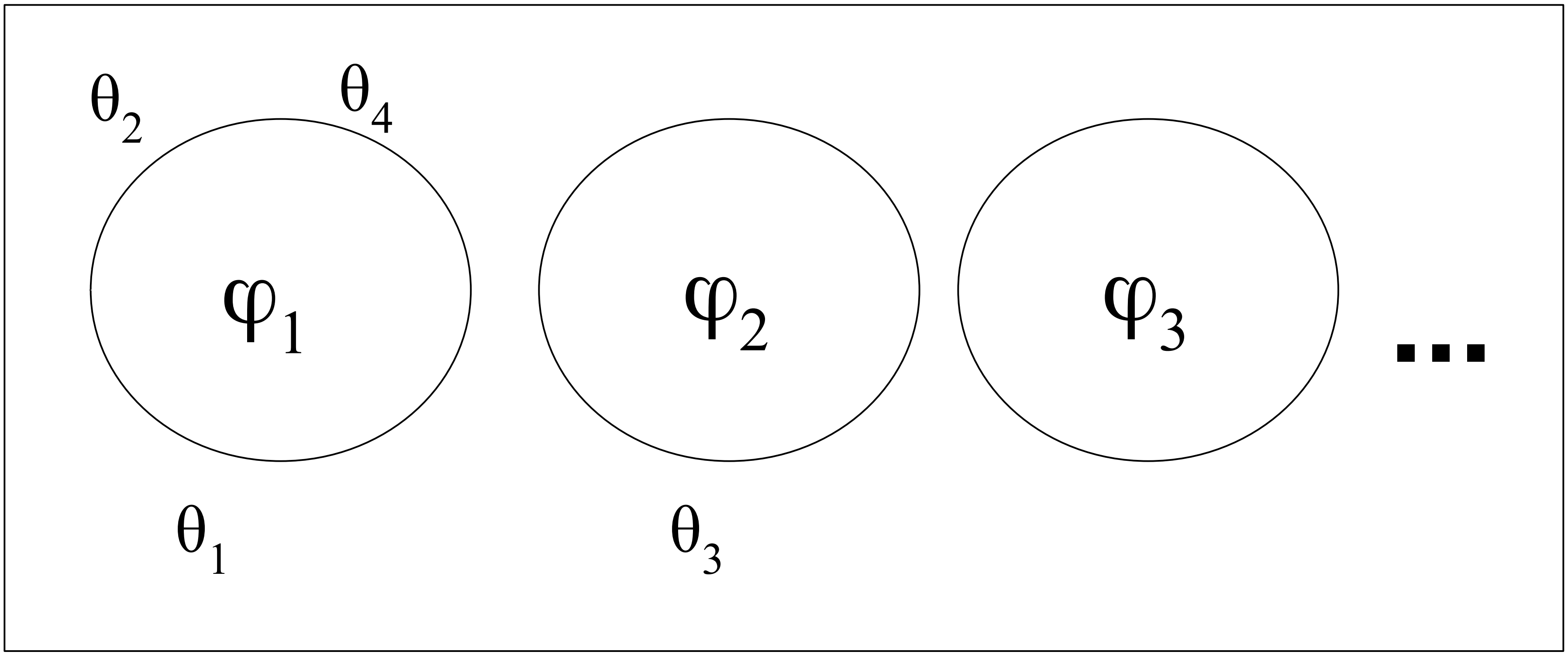
- Consider a simple urn model representation. Each sample is a ball of a certain color
- Balls are drawn equiprobably, and when a ball of color x is drawn, both that ball and a new ball of color x is returned to the urn
- With Probability proportional to α_0, a new atom is created from G_0,
  - A new ball of a new color is added to the urn

Polya urn scheme

- Let $\varphi_1 \ldots \varphi_K$ be the distinct values taken on by $\theta_1, \ldots \theta_{i-1}$,
- If m_k is the number of values of $\theta_1, \ldots \theta_{i-1}$, equal to φ_k:

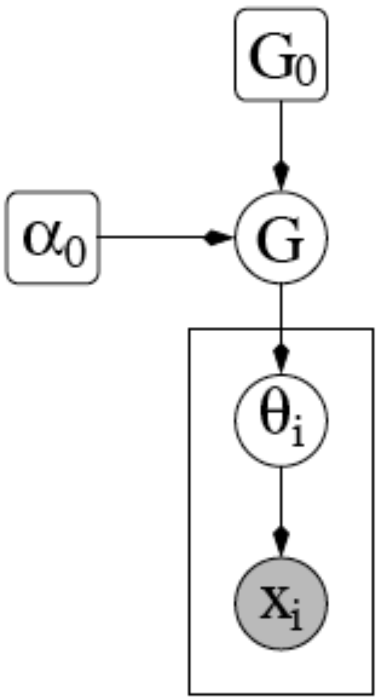
$$\theta_i \mid \theta_1, \ldots, \theta_{i-1}, \alpha_0, G_0 \ \sim \ \sum_{k=1}^{K} \frac{m_k}{i-1+\alpha_0} \delta_{\phi_k} + \frac{\alpha_0}{i-1+\alpha_0} G_0 \ .$$

Chinese restaurant process:



Dirichlet Process Mixture Model

- Dirichlet Process as nonparametric prior on the parameters of a mixture model:



$$\theta_i \mid G \sim G$$

$$x_i \mid \theta_i \sim F(\theta_i)$$

Dirichlet Process Mixture Model

- From the stick breaking representation:
  - θ_i will be the distribution represented by φ_k with probability π_k
- Let z_i be the indicator variable representing which φ_k θ_i is associated with:

$$\boldsymbol{\pi} \mid \alpha_0 \sim \mathrm{GEM}(\alpha_0)$$

$$\phi_k \mid G_0 \sim G_0$$

$$z_i \mid \boldsymbol{\pi} \sim \boldsymbol{\pi}$$

$$x_i \mid z_i, (\phi_k)_{k=1}^{\infty} \sim F(\phi_{z_i})$$

Infinite Limit of Finite Mixture Model

- Consider a multinomial on L mixture components with parameters $\boldsymbol{\pi} = (\pi_1, \ldots \pi_L)$
- Let $\boldsymbol{\pi}$ have a symmetric Dirichlet prior with hyperparameters $(\alpha_0/L, \ldots \alpha_0/L)$
- If x_i is drawn from a mixture component, z_i, according to the defined distribution:

$$\boldsymbol{\pi} \mid \alpha_0 \sim \mathrm{Dir}(\alpha_0/L, \ldots, \alpha_0/L) \qquad z_i \mid \boldsymbol{\pi} \sim \boldsymbol{\pi}$$

$$\phi_k \mid G_0 \sim G_0 \qquad x_i \mid z_i, (\phi_k)_{k=1}^{L} \sim F(\phi_{z_i})$$

Infinite Limit of Finite Mixture Model

$$\boldsymbol{\pi} \mid \alpha_0 \sim \mathrm{Dir}(\alpha_0/L, \ldots, \alpha_0/L) \qquad z_i \mid \boldsymbol{\pi} \sim \boldsymbol{\pi}$$

$$\phi_k \mid G_0 \sim G_0 \qquad x_i \mid z_i, (\phi_k)_{k=1}^{L} \sim F(\phi_{z_i})$$

- If $G^L = \sum_{k=1}^{L} \pi_k \delta_{\phi_k}$; L approaches ∞:

$$\int f(\theta)\, dG^L(\theta) \xrightarrow{\mathcal{D}} \int f(\theta)\, dG(\theta)$$

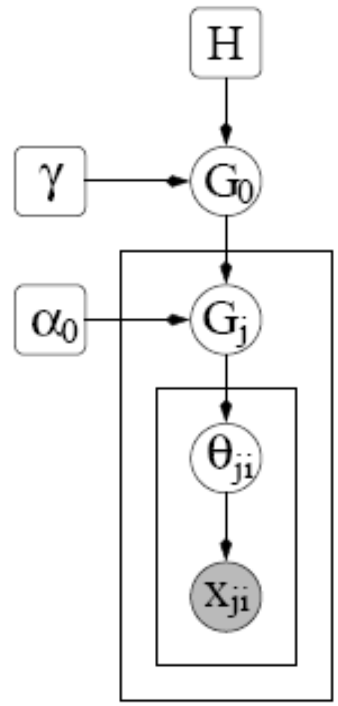
- The marginal distribution of x_1, x_2.... approaches that of a Dirichlet Process Mixture Model

Content

- Introduction and Motivation
- Dirichlet Processes
- Hierarchical Dirichlet Processes
  - Definition
  - Three Analogs
- Inference
  - Three Sampling Strategies

HDP Definition

- General idea
  - To model grouped data
    - Each group j <=> a Dirichlet process mixture model
    - Hierarchical prior to link these mixture models <=> hierarchical Dirichlet process
  - A hierarchical Dirichlet process is
    - A distribution over a set of random probability measures (   )



HDP Definition (Cont.)

- Formally, a hierarchical Dirichlet process defines
  - A set of random probability measures   , one for each group j
  - A global random probability measure
    - is a distributed as a Dirichlet process

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

is discrete!

    - are conditional independent given   , also follow DP

$$G_j \mid \alpha_0, G_0 \;\sim\; \mathrm{DP}(\alpha_0, G_0)$$

Hierarchical Dirichlet Process Mixture Model

- Hierarchical Dirichlet process as prior distribution over the factors for grouped data
- For each group j
  - Each observation corresponds to a factor
  - The factors are i.i.d random. variables distributed as

$$\theta_{ji} \mid G_j \sim G_j$$
$$x_{ji} \mid \theta_{ji} \sim F(\theta_{ji})$$

Some Notices

- HDP can be extended to more than two levels
  - The base measure *H* can be drawn from a DP, and so on and so forth
  - A tree can be formed
    - Each node is a DP
    - Children nodes are conditionally independent given their parent, which is a base measure
    - The atoms at a given node are shared among all its descendant nodes

Analog I: The stick-breaking construction

- Stick-breaking representation of

$$\phi_k \sim H$$

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}$$

$$\boldsymbol{\beta} = (\beta_k)_{k=1}^{\infty} \sim \mathrm{GEM}(\gamma)$$ i.e., $$\beta'_k \sim \mathrm{Beta}(1, \gamma) \qquad \beta_k = \beta'_k \prod_{l=1}^{k-1} (1 - \beta'_l)$$

- Stick-breaking representation of

$$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_k}$$

$$\boldsymbol{\pi}_j \sim \mathrm{DP}(\alpha_0, \boldsymbol{\beta})$$ i.e.,

$$\pi'_{jk} \sim \mathrm{Beta}\left(\alpha_0 \beta_k, \alpha_0 \left(1 - \sum_{l=1}^{k} \beta_l\right)\right)$$

$$\pi_{jk} = \pi'_{jk} \prod_{l=1}^{k-1} (1 - \pi'_{jl})$$

Equivalent representation using conditional distributions

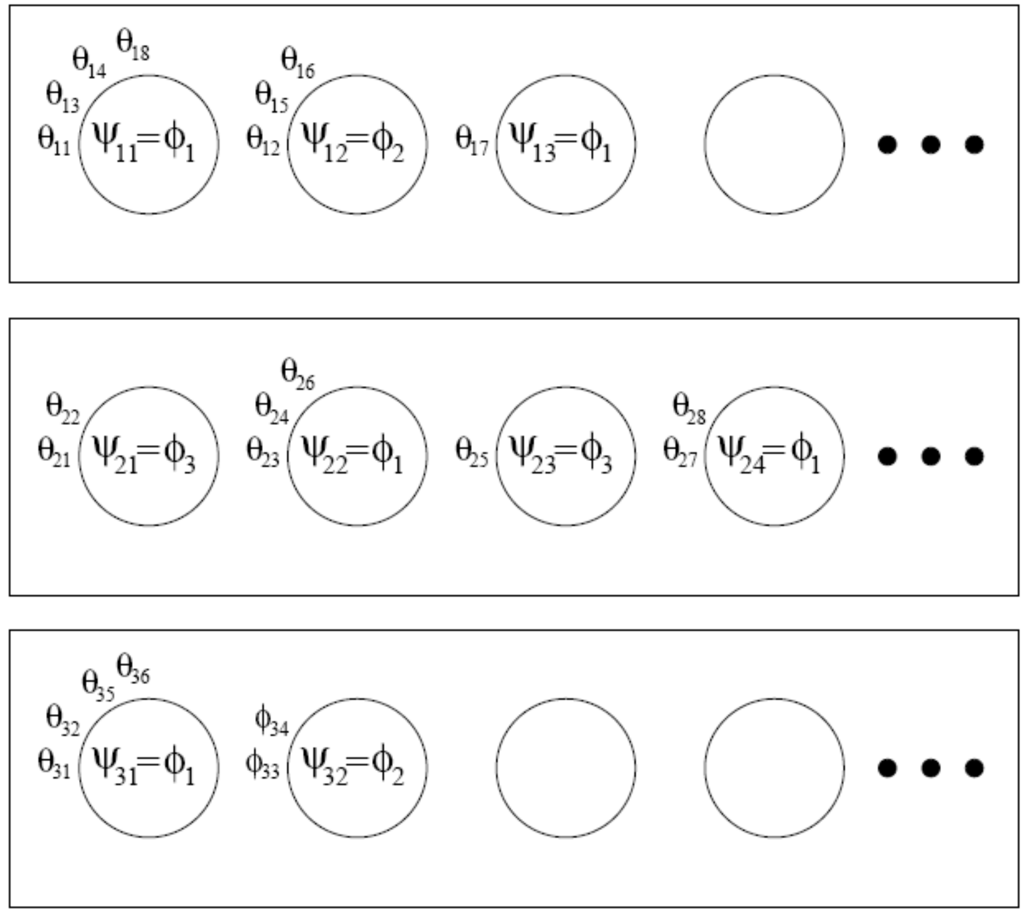
•

$$\begin{aligned}\boldsymbol{\beta} \mid \gamma &\sim \mathrm{GEM}(\gamma) \\ \boldsymbol{\pi}_j \mid \alpha_0, \boldsymbol{\beta} &\sim \mathrm{DP}(\alpha_0, \boldsymbol{\beta}) \\ \phi_k \mid H &\sim H\end{aligned}$$

$$\begin{aligned}z_{ji} \mid \boldsymbol{\pi}_j &\sim \boldsymbol{\pi}_j \\ x_{ji} \mid z_{ji}, (\phi_k)_{k=1}^{\infty} &\sim F(\phi_{z_{ji}})\end{aligned}$$

Analog II: the Chinese restaurant franchise

- General idea:
  - Allow multiple restaurants to share a common menu, which includes a set of dishes
  - A restaurant has infinite tables, each table has only one dish



Notations

-
  - The factor (dish) corresponding to
- $\phi_1, \cdots, \phi_K$
  - The factors (dishes) drawn from H
-
  - The dish chosen by table t in restaurant j
- : the index of associated with
- : the index of associated with

Conditional distributions

- Integrate out G_j (sampling table for customer)

$$\theta_{ji} \mid \theta_{j1}, \ldots, \theta_{j,i-1}, \alpha_0, G_0 \sim \sum_{t=1}^{m_{j\cdot}} \frac{n_{jt\cdot}}{i-1+\alpha_0} \delta_{\psi_{jt}} + \frac{\alpha_0}{i-1+\alpha_0} G_0$$

- Integrate out G_0 (sampling dish for table)

$$\psi_{jt} \mid \psi_{11}, \psi_{12}, \ldots, \psi_{21}, \ldots, \psi_{j\,t-1}, \gamma, H \sim \sum_{k=1}^{K} \frac{m_{\cdot k}}{m_{\cdot\cdot}+\gamma} \delta_{\phi_k} + \frac{\gamma}{m_{\cdot\cdot}+\gamma} H$$

Count notation: , number of customers in restaurant j, at table t, eating dish k
, number of tables in restaurant j, eating dish k

Analog III: The infinite limit of finite mixture models

- Two different finite models both yield HDPM
  - Global mixing proportions place a prior for group-specific mixing proportions

$$
\begin{aligned}
\boldsymbol{\beta} \mid \gamma &\sim \mathrm{Dir}(\gamma/L, \ldots, \gamma/L) \\
\boldsymbol{\pi}_j \mid \alpha_0, \boldsymbol{\beta} &\sim \mathrm{Dir}(\alpha_0 \boldsymbol{\beta}) \\
\phi_k \mid H &\sim H
\end{aligned}
$$

$$
\begin{aligned}
z_{ji} \mid \boldsymbol{\pi}_j &\sim \boldsymbol{\pi}_j \\
x_{ji} \mid z_{ji}, (\phi_k)_{k=1}^{L} &\sim F(\phi_{z_{ji}})
\end{aligned}
$$

As L goes infinity

– Each group choose a subset of T mixture components

$$
\begin{aligned}
\boldsymbol{\beta} \mid \gamma &\sim \mathrm{Dir}(\gamma/L, \ldots, \gamma/L) \\
\boldsymbol{\pi}_j \mid \alpha_0 &\sim \mathrm{Dir}(\alpha_0/T, \ldots, \alpha_0/T) \\
\phi_k \mid H &\sim H
\end{aligned}
\qquad
\begin{aligned}
k_{jt} \mid \boldsymbol{\beta} &\sim \boldsymbol{\beta} \\
t_{ji} \mid \boldsymbol{\pi}_j &\sim \boldsymbol{\pi}_j \\
x_{ji} \mid t_{ji}, (k_{jt})_{t=1}^{T}, (\phi_k)_{k=1}^{L} &\sim F(\phi_{k_{jt_{ji}}})
\end{aligned}
$$

As L, T go to infinity

Content

- Introduction and Motivation
- Dirichlet Processes
- Hierarchical Dirichlet Processes
  - Definition
  - Three Analogs
- Inference
  - Three Sampling Strategies

Introduction to three MCMC schemes

- Assumption: H is conjugate to F
  - A straightforward Gibbs sampler based on Chinese restaurant franchise
  - An augmented representation involving both the Chinese restaurant franchise and the posterior for G_0
  - A variation to scheme 2 with streamline bookkeeping

Conditional density of data under mixture component k

- For data , conditional density under component k given all data items except is:

$$f_k^{-x_{ji}}(x_{ji}) = \frac{\int f(x_{ji}|\phi_k) \prod_{j'i' \neq ji, z_{j'i'}=k} f(x_{j'i'}|\phi_k) h(\phi_k)\, d\phi_k}{\int \prod_{j'i' \neq ji, z_{j'i'}=k} f(x_{j'i'}|\phi_k) h(\phi_k)\, d\phi_k}$$

- For data set $\boldsymbol{x}_{jt}$, conditional density $f_k^{-\boldsymbol{x}_{jt}}(\boldsymbol{x}_{jt})$ is similarly defined

Scheme I: Posterior sampling in the Chinese restaurant franchise

- Sampling ***t*** and ***k***
  - Sampling ***t***

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

    - If is a new t, sampling the k corresponding to it

$$p(k_{jt^{\text{new}}} = k \mid \boldsymbol{t}, \boldsymbol{k}^{-jt^{\text{new}}}) \propto \begin{cases} m_{\cdot k} f_k^{-x_{ji}}(x_{ji}) & \text{if } k \text{ previously used,} \\ \gamma f_{k^{\text{new}}}^{-x_{ji}}(x_{ji}) & \text{if } k = k^{\text{new}}. \end{cases}$$

    - And

$$p(x_{ji} \mid \boldsymbol{t}^{-ji}, t_{ji} = t^{\text{new}}, \boldsymbol{k}) = \sum_{k=1}^{K} \frac{m_{\cdot k}}{m_{\cdot\cdot} + \gamma} f_k^{-x_{ji}}(x_{ji}) + \frac{\gamma}{m_{\cdot\cdot} + \gamma} f_{k^{\text{new}}}^{-x_{ji}}(x_{ji})$$

– Sampling ***k***

-

$$p(k_{jt} = k \mid \boldsymbol{t}, \boldsymbol{k}^{-jt}) \propto \begin{cases} m_{\cdot k}^{-jt} f_k^{-\boldsymbol{x}_{jt}}(\boldsymbol{x}_{jt}) & \text{if } k \text{ is previously used,} \\ \gamma f_{k^{\text{new}}}^{-\boldsymbol{x}_{jt}}(\boldsymbol{x}_{jt}) & \text{if } k = k^{\text{new}}. \end{cases}$$

Where is all the observations for table t in restaurant j

Scheme II: Posterior sampling with an augmented representation

- Posterior of G_0 given :

$$\mathrm{DP}(\gamma + m_{\cdot\cdot}, \frac{\gamma H + \sum_{k=1}^{K} m_{\cdot k}\delta_{\phi_k}}{\gamma + m_{\cdot\cdot}})$$

- An explicit construction for G_0 is given:

$$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K, \beta_u) \sim \mathrm{Dir}(m_{\cdot 1}, \ldots, m_{\cdot K}, \gamma) \qquad G_u \sim \mathrm{DP}(\gamma, H)$$

$$p(\phi_k \mid \boldsymbol{t}, \boldsymbol{k}) \propto h(\phi_k) \prod_{ji:k_{jt_{ji}}=k} f(x_{ji}|\phi_k) \qquad G_0 = \sum_{k=1}^{K} \beta_k \delta_{\phi_k} + \beta_u G_u$$

- Given a sample of G_0, posterior for each group is factorized and sampling in each group can be performed separately
- Sampling ***t*** and ***k***:
  - Almost the same as in Scheme I
    - Except using β_k, β_u to replace $m_{.k}, \gamma$
    - When a new component k_{new} is instantiated, draw $b \sim \mathrm{Beta}(1, \gamma)$, and set $\beta_{k^{\mathrm{new}}} = b\beta_u$ and $\beta_u^{\mathrm{new}} = (1-b)\beta_u$

– Sampling for

$$(\beta_1, \ldots, \beta_K, \beta_u) \mid \boldsymbol{t}, \boldsymbol{k} \;\sim\; \mathrm{Dir}(m_{\cdot 1}, \ldots, m_{\cdot K}, \gamma)$$

Scheme III: Posterior sampling by direct assignment

- Difference from Scheme I and II:
  - In I and II, data items are first assigned to some table t, and the tables are then assigned to some component k
  - In III, directly assign data items to component via variable , which is equivalent to
    - Tables are collapsed to numbers

- Sampling ***z***:

$$p(z_{ji} = k \mid \boldsymbol{z}^{-ji}, \boldsymbol{m}, \boldsymbol{\beta}) = \begin{cases} (n_{j\cdot k}^{-ji} + \alpha_0 \beta_k) f_k^{-x_{ji}}(x_{ji}) & \text{if } k \text{ previously used,} \\ \alpha_0 \beta_u f_{k^{\text{new}}}^{-x_{ji}}(x_{ji}) & \text{if } k = k^{\text{new}}. \end{cases}$$

- Sampling ***m***:

$$p(m_{jk} = m \mid \boldsymbol{z}, \boldsymbol{m}^{-jk}, \boldsymbol{\beta}) = \frac{\Gamma(\alpha_0 \beta_k)}{\Gamma(\alpha_0 \beta_k + n_{j\cdot k})} s(n_{j\cdot k}, m)(\alpha_0 \beta_k)^m$$

- Sampling

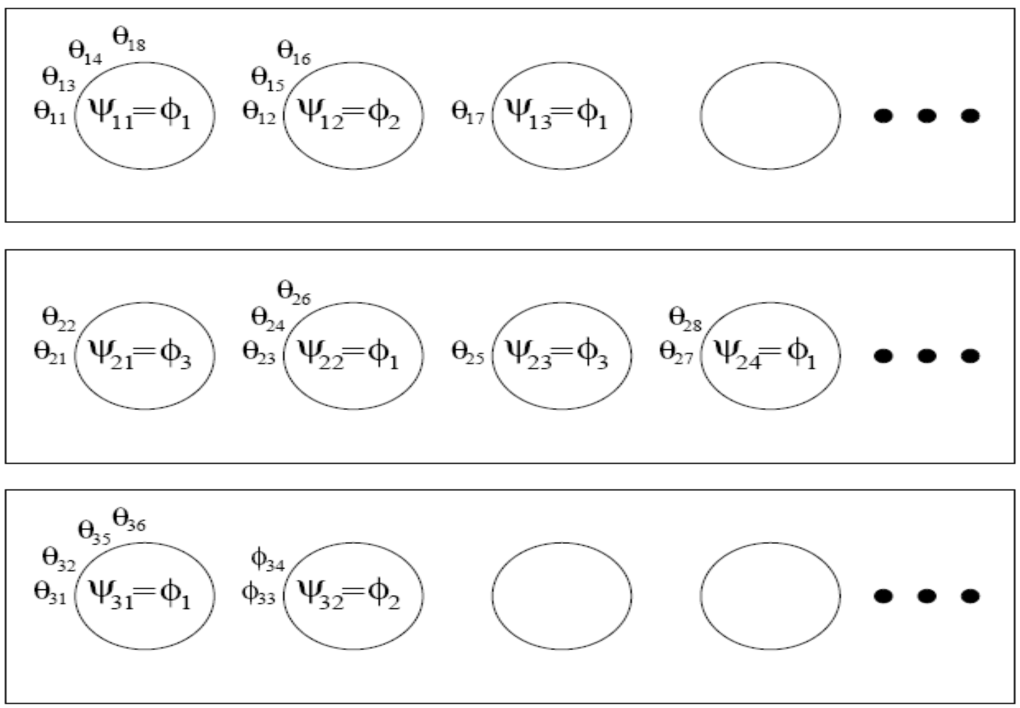
$$(\beta_1, \ldots, \beta_K, \beta_u) \mid \boldsymbol{t}, \boldsymbol{k} \ \sim \ \mathrm{Dir}(m_{\cdot 1}, \ldots, m_{\cdot K}, \gamma)$$

Comparison of Sampling Schemes

- In terms of ease of implementation
  - The direct assignment is better
- In terms of convergence speed
  - Direct assignment changes the component membership of data items one at a time
  - Scheme I and II, component membership of one table will change the membership of multiple data items at the same time, leading to better performance

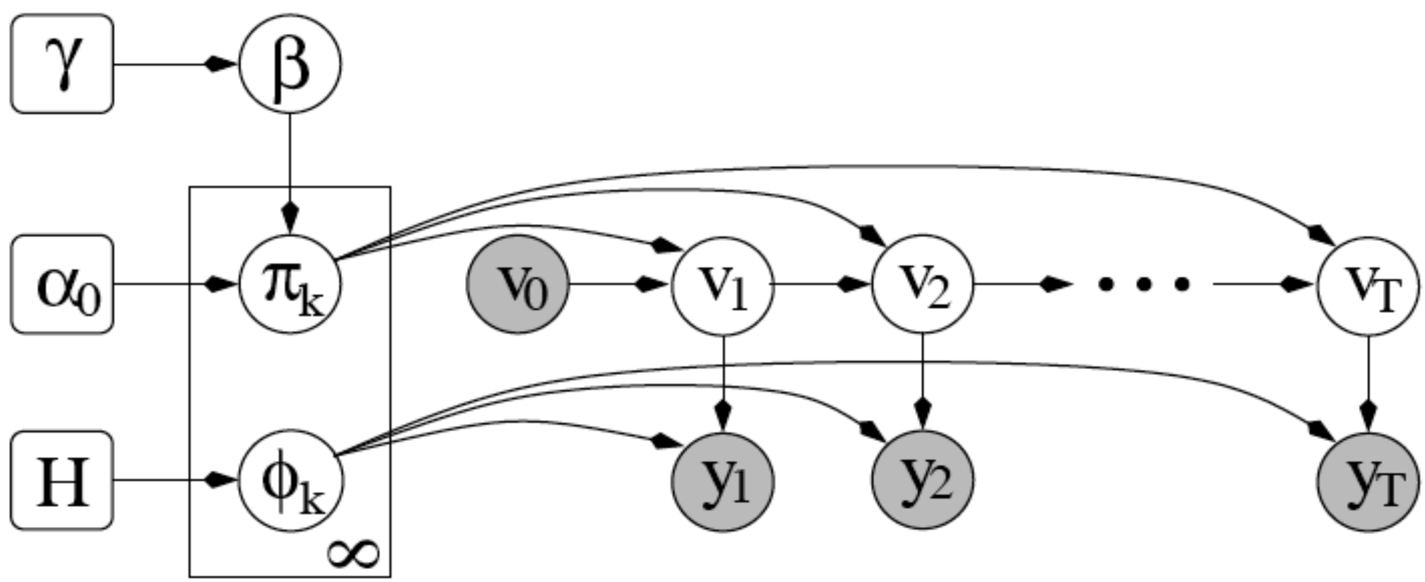
Applications

- Hierarchical DP extension of LDA
  - In CRF representation: dishes are topics, customers are the observed words



Applications

- HDP-HMM



References

- Yee Whye Teh et. al., Hierarchical Dirichlet Processes, 2006