Final Report: HDP Model

Stats Group

Department of Statistics, The University of Chicago

2023/05/25

1. Literature Review
    - Slice Sampler for Mixture Models (Walker, 2007; Kalli et al., 2011)
    - HDP Model (Teh et al., 2004)

2. Repeat and Research
    - Repeat: Chinese Restaurant Franchise Sampling Method
    - HDP Model with Latent Indicator Variables
    - Slice Sampling for HDP

1 Literature Review

Slice Sampler for Mixture Models (Walker, 2007; Kalli et al., 2011)

HDP Model (Teh et al., 2004)

2 Repeat and Research

1. Literature Review
   - Slice Sampler for Mixture Models (Walker, 2007; Kalli et al., 2011)
   - HDP Model (Teh et al., 2004)
2. Repeat and Research

Sampling the Dirichlet Mixture Model with Slices, Walker, 2007

- In this article, Walker introduce a new method for sampling the mixture of Dirichlet process (MDP) model, a complex task due to the countable infiniteness of the discrete masses from the random distribution functions chosen from the Dirichlet process prior.
- Ishwaran and James (2001) recently devised a Gibbs sampling scheme that incorporates more general stick-breaking priors, serving as a direct extension of Escobar's (1988) methodology.

The Dirichlet Process Model

The Mixed Dirichlet Process (MDP) model is founded on the concept of creating continuous random distribution functions, as initially proposed by Lo in 1984. The random distribution function selected from a Dirichlet process is almost certain to be discrete, as posited by Blackwell in 1973. Therefore, we turn our attention to the random density function.

$$f_p(y) = \int N(y|\theta)dP(\theta)$$

$N(y|\theta)$ denotes a conditional density function, which will typically be a normal distribution with parameters θ. Therefore, in the normal case $\theta = (\mu, \sigma^2)$

The Dirichlet Process Model

Given the form for P, we can write

$$f_{w,\theta}(y) = \sum_j w_j N(y|\theta_j)$$

Via Gibbs sampling ideas, our attempt to estimate the model is to introduce a latent variable μ such that the joint density of (y,u) given (ω, θ) is given by

$$f_{\omega,\theta}(y,u) = \sum_j \mathbf{1}(u < \omega_j) N(y|\theta_j)$$

The joint distribution exists and we can write

$$f_{\omega,\theta}(y,u) = \sum_{j=1}^{\infty} \omega U(u|0,\omega_j) N(y|\theta_j)$$

The Dirichlet Process Model

We can get the marginal density for u is given by

$$f_\omega(u) = \sum_{j=1}^{\infty} \omega_j U(u|0,\omega_j) = \sum_{j=1}^{\infty} \mathbf{1}(u < \omega_j)$$

let $A_\omega(u) = j : \omega_j > u$ then we can get

$$f_{\omega,\theta}(y,u) = \sum_{j \in A_\omega(u)} N(y|\theta_j)$$

It is clear that $A_\omega(u)$ is a finite set for all u>0. The conditional density of y given u is

$$f_{\omega,\theta}(y|u) = \frac{1}{f_\omega(u)} \sum_{j \in A_\omega(u)} N(y|\theta_j)$$

The Dirichlet Process Model

Here, we move from an infinite sum to a finite sum, which makes a lot of difference when sampling involved.
Thus, given u, we have a finite mixture model with equal to $1/f_\omega(u)$.
We can introduce a further indicator latent variable which will identify the component of the mixture from which y is to be taken.
Joint density:

$$f_{\omega,\theta}(y, \delta = k, u) = N(y|\theta_k)\mathbf{1}(k \in A(u))$$

The complete data likelihood based on a sample of size n:

$$l_{\omega,\theta}(y_i, u_i, \delta_i = k_i{}_{i=1}^{n}) = \prod_{i=1}^{n} N(y_i|\theta_{k_i})\mathbf{1}(u_i < \omega_{k_i})$$

Slice Sampling Mixture Models, Kalli, 2011

- In this article, Kalli et.al (2011) proposed a modified slice sampler called slice efficient sampler for inference problems of Dirichlet process mixture models. While preserving the property of reducing infinite number of components to finite number, slice efficient sampler is able to control the updating process of auxiliary variables so that the MCMC sampling process is smooth.
- The authors extended the idea of slice sampler to normalized weights priors for mixture models and performed two numerical experiments to demonstrate the effeteness of slice sampler.

Slice Sampling Mixture Models, Kalli, 2011

Slice efficient sampler: by introducing ζ_i, $i = 1, 2, 3, 4, ...$, a positive decreasing deterministic sequence and d, a latent variable indicating which finite number of components can produce the observed data y. d together with u, where u here is the auxiliary variable, reduce the infinite number of mixture of components to finite components so that MCMC algorithms are smooth. The construction is followed:

$$f_{v,\mu,\sigma^2}(y, u, d) = \zeta_d^{-1} \mathbf{1}(u < \zeta_d) w_d N(y; u_d, \sigma_d^2)$$

The joint posterior likelihood is then:

$$\prod_{i=1}^{n} \zeta_{d_i}^{-1} \mathbf{1}(u < \zeta_{d_i}) w_{d_i} N(y; u_{d_i}, \sigma_{d_i}^2)$$

Slice Sampling Mixture Models, Kalli, 2011

Notice, here ζ_i is decreasing and is positive, so $\mathbf{1}(u < \zeta_i)$ happens with less chance, while $\frac{w_i}{\zeta_i}$ gets bigger so gives more weights to the Gaussian kernel. Thus, the sampling algorithm is followed:

$$1.\pi\left(\mu_j, \sigma_j^2 \mid \cdots\right) \propto p_0\left(\mu_j, \sigma_j^2\right) \prod_{d_i=j} \mathrm{N}\left(y_i; \mu_j, \sigma_j^2\right)$$

$$2.\pi\left(v_j\right) \propto \mathrm{Be}\left(v_j; a_j, b_j\right), a_j = 1 + \sum_{i=1}^{n} \mathbf{1}\left(d_i = j\right), b_j = M + \sum_{i=1}^{n} \mathbf{1}\left(d_i > j\right)$$

$$3.\pi\left(u_i \mid \cdots\right) \propto \mathbf{1}\left(0 < u_i < \xi_{d_i}\right)$$

$$4.\mathrm{P}\left(d_i = k \mid \cdots\right) \propto \mathbf{1}\left(k : \xi_k > u_i\right) w_k / \xi_k \ \mathrm{N}\left(y_i; \mu_k, \sigma_k^2\right)$$

Slice Sampling Mixture Models, Kalli, 2011

Mixtures based on normalized weights prior

- Consider $f(y) = \sum_{j=1}^{\infty} w_j K(y; \phi_j)$, $w_j = \lambda_j / \Lambda, \Lambda = \sum_{j=1}^{\infty} \lambda_j$ and $\phi_1, \phi_2, \phi_3, \ldots$ are iid. $\Lambda_m = \sum_{j=m+1}^{\infty} \lambda_j$ and $\lambda_j \sim \pi_j(\lambda_j)$
- π_j as the authors suggested can be gamma distribution and inverse Gaussian distribution

Slice Sampling Mixture Models, Kalli, 2011

Slice sampler and normalized weights prior

- joint density:

$$f(y, v, u, d) = \exp(-v\Lambda)\mathbf{1}\left(u < \xi_d\right)\lambda_d/\xi_d K\left(y; \phi_d\right)$$

- let $v_1, v_2, ..., v_n$ be $v = \sum_{i=1}^{n} v_i$, the likelihood is then

$$v^{n-1}\exp(-v\Lambda)\prod_{i=1}^{n}\mathbf{1}\left(u_i < \xi_{d_i}\right)\lambda_{d_i}/\xi_{d_i} K\left(y_i; \phi_{d_i}\right)$$

Then, the authors provided two algorithms to perform sampling:
(1) dependent slice-efficient sampler and $\xi_j = \lambda_j$
(2) independent slice-efficient sampler and ξ_j, λ_j independent

Slice Sampling Mixture Models, Kalli, 2011

- Random hazard function has same posterior as normalized mixture models using slice efficient sampler
- The authors compared the efficiency of slice sampler and retrospective sampler, pointing out that slice sampler can save much pre-runnung work and so easier to work.

1. Literature Review
   - Slice Sampler for Mixture Models (Walker, 2007; Kalli et al., 2011)
   - HDP Model (Teh et al., 2004)

2. Repeat and Research

Main Idea

- The paper discusses the representation of hierarchical Dirichlet processes using a **stick-breaking process** and introduces a generalization of the Chinese restaurant process called the **"Chinese restaurant franchise."** These representations facilitate the understanding and implementation of hierarchical Dirichlet processes.

Dirichlet Process

- Let (Θ, B) be a measurable space, with G_0 a probability measure on the space. Let α_0 be a positive real number. For any finite partition $\{A_1, ..., A_k\}$ of the sample space, consider the random vector $(G(A_1), ..., G(A_k))$ where $G()$ is a random distribution function. If

$$(G(A_1), ..., G(A_k)) \sim Dirichlet(\alpha_0, G_0(A_1), ..., G_0(A_k))$$

where G_0 is also a fixed distribution function, then $G \sim DP(\alpha_0, G_0)$ is a Dirichlet process.

Dirichlet process

- We have the model

$$\theta_i | G \sim G$$
$$x_i | \theta_i \sim F(\theta_i)$$

  where $F(\theta_i)$ denotes the distribution of the observation x_i given θ_i. The factors θ_i are conditionally independent given G, and the observation x_i is conditionally independent of the other observations given the factor θ_i.
- When G is distributed according to a Dirichlet process, this model is referred to as a Dirichlet process mixture model.

Hierarchical Dirichlet processes

- We have the global random probability measure G_0,

$$G_0|\gamma, H \sim DP(\gamma, H)$$

- We have conditional random measure G_j follow,

$$G_j|\alpha_0, G_0 \sim DP(\alpha_0, G_0)$$

- In this model we have three hyper-parameters: baseline probability measure H, concentration parameters γ, α_0

Hierarchical Dirichlet processes

- For now, we can have HDP

$$
\begin{aligned}
\theta_{ji} | G_j &\sim G_j \\
G_j | \alpha_0, G_0 &\sim DP(\alpha_0, G_0) \\
G_0 | \gamma, H &\sim DP(\gamma, H) \\
x_{ji} | \theta_{ji} &\sim F(\theta_{ji}), \text{ for each } j \text{ and } i
\end{aligned}
$$

Hierarchical Dirichlet processes

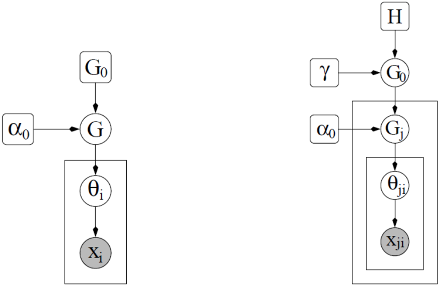


Figure 1: (Left) A representation of a Dirichlet process mixture model as a graphical model. (Right) A hierarchical Dirichlet process mixture model. In the graphical model formalism, each node in the graph is associated with a random variable, where shading denotes an observed variable. Rectangles denote replication of the model within the rectangle. Sometimes the number of replicates is given in the bottom right corner of the rectangle.

Figure 1:

Stick-breaking Representation

- Given that the global measure G_0 is distributed as a Dirichlet process, it can be expressed using a stick-breaking representation:

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k} \quad (16)$$

where $\phi_k \sim H$ independently and $\beta = (\beta_k)_{k=1}^{\infty} \sim \mathrm{GEM}(\gamma)$ are mutually independent. Since G_0 has support at the points $\phi = (\phi_k)_{k=1}^{\infty}$, each G_j necessarily has support at these points as well, and can thus be written as:

$$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_k} \quad (17)$$

Stick-breaking Representation

- Let $\pi_j = (\pi_{jk})_{k=1}^{\infty}$. Note that the weights π_j are independent given β (since the G_j are independent given G_0). We now describe how the weights π_j are related to the global weights β.
- Let $(A_1, ..., A_r)$ be a measurable partition, and $K_l = \{k : \phi_k \in A_l\}, l = 1, ..., r$ is a finite partition of the positive integers. We have

$$
\begin{aligned}
&(G_j(A_1), ..., G_j(A_r)) \sim Dir(\alpha_0 G_0(A_1), ..., \alpha_0 G_0(A_r)) \\
\Rightarrow &\Big(\sum_{k \in K_1} \pi_{jk}, ..., \sum_{k \in K_r} \pi_{jk}\Big) \sim Dir\Big(\alpha_0 \sum_{k \in K_1} \beta_k, ..., \alpha_0 \sum_{k \in K_r} \beta_k\Big) \quad (18)
\end{aligned}
$$

Stick-breaking Representation

- Then try to derive the explicit relationship of β, π_j.

$$\beta'_k \sim Beta(1, \gamma), \quad \beta_k = \beta'_k \Pi_{l=1}^{k-1}(1 - \beta'_l) \quad (20)$$

- It can be showed that the following stick-breaking construction produces a random probability measure $\pi_j \sim DP(\alpha_0, \beta)$:

$$\pi'_{jk} \sim Beta(\alpha_0 \beta_k, \alpha_0(1 - \sum_{l=1}^{k} \beta_l)), \quad \pi_{jk} = \pi'_{jk} \Pi_{l=1}^{k-1}(1 - \pi'_{jl}) \quad (21)$$

Stick-breaking Representation

- To derive (21), first notice that for a partition $(\{1,...,k-1\},\{k\},\{k+1,k+2,...\})$, (18) gives

$$(\sum_{l=1}^{k-1}\pi_{jl},\pi_{jk},\sum_{l=k+1}^{\infty}\pi_{jl}) \sim Dir(\alpha_0\sum_{l=1}^{k-1}\beta_l,\alpha_0\beta_k,\alpha_0\sum_{l=k+1}^{\infty}\beta_l) \quad (22)$$

- Removing the first element, and using standard properties of the Dirichlet distribution that support should be added to 1(Re-normalization Property). Dirichlet Distribution follows the renormalization property. Let $\pi_1,...,\pi_K \sim Dir(\alpha_1,...,\alpha_K)$,

$$\frac{(\pi_2,...,\pi_K)}{\sum_{k=2}^{K}\pi_k} \sim Dir(\alpha_2,...,\alpha_K)$$

Stick-breaking Representation

- So we have

$$\frac{1}{1-\sum_{l=1}^{k-1}\pi_{jl}}(\pi_{jk}, \sum_{l=k+1}^{\infty}\pi_{jl}) \sim Dir(\alpha_0\beta_k, \alpha_0\sum_{l=k+1}^{\infty}\beta_l) \quad (23)$$

- Finally, define $\pi'_{jk} = \frac{\pi_{jk}}{1-\sum_{l=1}^{k-1}\pi_{jl}}$ and notice that $1-\sum_{l=1}^{k}\beta_l = \sum_{l=k+1}^{\infty}\beta_l$ to obtain (21)

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

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

Chinese Restaurant Franchise (CRF) Process

- Different from original Chinese Restaurant Process, We have a restaurant franchise with a shared menu across the restaurants.
- At each table of each restaurant one dish is ordered from the menu by the first customer who sits there, and it is shared among all customers who sit at that table. Multiple tables in multiple restaurants can serve the same dish.

Chinese Restaurant Franchise (CRF) Process

Process assumes $j = 1, ..., J$ Chinese restaurants:

- with infinite tables in each restaurant;
- all share the same menu;
- 1st customer to a new table orders a plate for the table;
- multiple tables in multiple restaurants can serve the same plate.

Each customer i entering a restaurant j does the following:

- Chooses a table to sit proportion to the number of customers already at that table.
- If sits in a new table, then order a plate from the menu proportion to the popularity of the plate across all restaurants.

1. Literature Review

2. Repeat and Research
   - Repeat: Chinese Restaurant Franchise Sampling Method
   - HDP Model with Latent Indicator Variables
   - Slice Sampling for HDP

1. Literature Review
2. Repeat and Research
   - Repeat: Chinese Restaurant Franchise Sampling Method
   - HDP Model with Latent Indicator Variables
   - Slice Sampling for HDP

The Chinese Restaurant Franchise

- The Chinese restaurant process is extended to allow multiple restaurants to share a set of dishes.

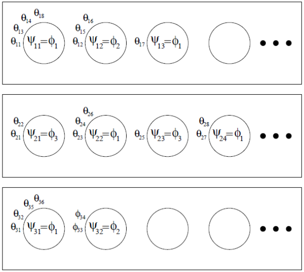


Figure 2: A depiction of a Chinese restaurant franchise. Each restaurant is represented by a rectangle. Customers (θ_{ji}'s) are seated at tables (circles) in the restaurants. At each table a dish is served. The dish is served from a global menu (ϕ_k), whereas the parameter ψ_{jt} is a table-specific indicator that serves to index items on the global menu. The customer θ_{ji} sits at the table to which it has been assigned in (24).

The Chinese Restaurant Franchise Notations

- θ_{ji}: customers; ϕ_k: a global menu
- x_{ji}: the observed data, each one is a draw from a distribution $F(\theta_{ji})$
- ψ_{jt}: the dish served at table t in restaurant j
- t_{ji}: the index of the ψ_{jt} associated with θ_{ji}
- k_{jt}: the index of ϕ_k associated with ψ_{jt}
- We also maintain counts of customers and counts of tables
  n_{jtk}: the number of customers in restaurant j at table t eating dish k
  m_{jk}: the number of tables in restaurant j serving dish k

The Chinese Restaurant Franchise

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

$$\psi_{jt} \mid \psi_{11}, \psi_{12}, \cdots, \psi_{21}, \cdots, \psi_{j,t-1}, \gamma, H \sim \sum_{k=1}^{K} \frac{m_{\cdot k}}{m_{\cdot\cdot} + \gamma} \delta_{\phi_k} + \frac{\gamma}{m_{\cdot\cdot} + \gamma} H \tag{2}$$

For each j and i, first sample θ_{ji} using (1). If a new sample from G_0 is needed, we use (2) to obtain a new sample ψ_{jt} and set $\theta_{ji} = \psi_{jt}$

Posterior Sampling in the Chinese Restaurant Franchise

- Rather than dealing with the θ_{ji}'s and ψ_{jt}'s directly, we shall sample their index variables t_{ji} and k_{jt} instead
- Motivation: the θ_{ji}'s and ψ_{jt}'s can be reconstructed from these index variables and the ϕ_k's
- Advantage: It makes the MCMC sampling scheme more efficient
- Notice that the t_{ji} and k_{jt} inherit the exchangeability properties of the θ_{ji} and ψ_{jt}, allowing us to adapt the conditional distributions in (1) and (2) to be expressed in terms of t_{ji} and k_{jt}

Sampling t

Obtain the conditional posterior for t_{ji} by combining the conditional prior distribution for t_{ji} with the likelihood of generating x_{ji}

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

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

Sampling t

- If the sampled value of t_{ji} is t^{new},

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

- If some table t becomes unoccupied as a result of updating t_{ji}, i.e., $n_{jt\cdot} = 0$, then the probability that this table will be reoccupied in the future will be zero. Hence, we delete the corresponding k_{jt} from the data structure.

Sampling k

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

Implementing the CRF Sampling Method

- Given hyperparameters: $\alpha_0 = \gamma = 1$
- Priors: $\phi_k \mid H \sim H = N(0, \frac{1}{\tau_\phi^2})$, $x_{ji} \mid \theta_{ji} \sim F = N(\theta_{ji}, \frac{1}{\tau_x^2})$ where $\tau_\phi = \tau_x = 1$ (predetermined, tuned afterwards)

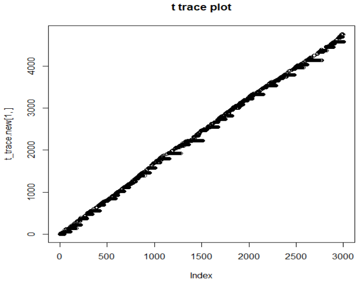
Implementing the CRF Sampling Method

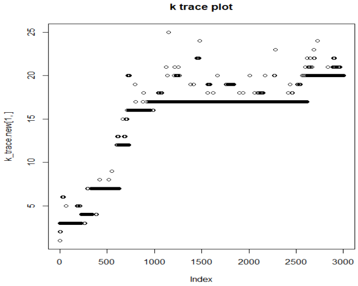
Derive $f_k^{-x_{ji}}(x_{ji})$:

$$f_k^{-x_{ji}}(x_{ji}) = \frac{\tau_x}{\sqrt{2\pi}} \tau_\pi \sqrt{\frac{\tau_x^2 n_{\cdot\cdot k}^{-ij} + \tau_\phi^2}{\tau_x^2 (n_{\cdot\cdot k}^{-ij} + 1) + \tau_\phi^2}} e^{-\frac{1}{2}[x_{ji}^2 + s_1 - s_2]}$$

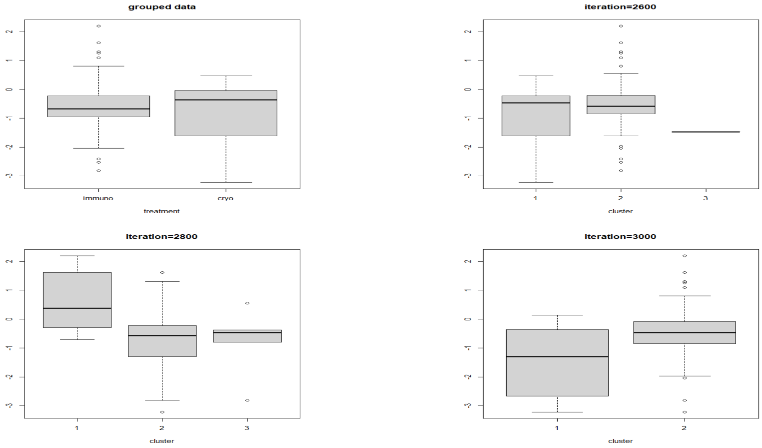
where $s_1 = \frac{(\tau_x^2 \sum_{j'i' \neq ji, z_{j'i'}=k} x_{j'i'})^2}{\tau_x^2 n_{\cdot\cdot k}^{-ji} + \tau_\phi^2}$ and $s_2 = \frac{(\tau_x^2 \sum_{z_{j'i'}=k} x_{j'i'})^2}{\tau_x^2 n_{\cdot\cdot k}^{-ji} + \tau_\phi^2}$

Result: Trace Plots of t_{ji} and k_{ji}

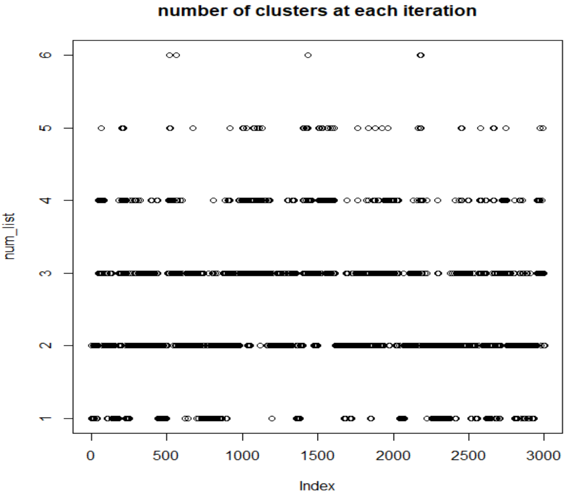




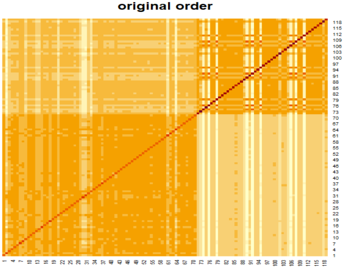
Issue: Clustering is not Stable

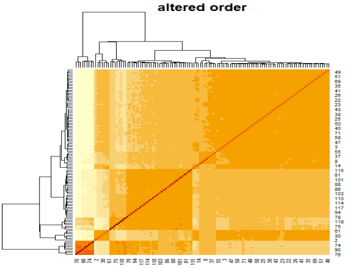


Result: Number of Clusters in Each Iteration



Result: Simularity Matrix





1. Literature Review
2. Repeat and Research
    - Repeat: Chinese Restaurant Franchise Sampling Method
    - HDP Model with Latent Indicator Variables
    - Slice Sampling for HDP

HDP Model with Latent Indicator Variables (1)

Our model setting is

$$y_{ij} \mid \theta_{ij} \sim F(\theta_{ij}) \qquad \theta_{ij} \mid G_j \sim G_j$$

$$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_k} \qquad \pi_{jk} = \pi'_{jk} \prod_{l=1}^{k-1} (1 - \pi_{jl})$$

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

$$\beta'_k \sim \mathrm{Beta}(1, \gamma)$$

$$\phi_k \sim H \quad \text{(Prior for location } \phi_k)$$

HDP Model with Latent Indicator Variables (2)

- Now,

$$F(\cdot) = N(\cdot, \cdot) \qquad \theta_{ij} = \{\mu_{ij}, \sigma^2_{ij}\}$$
$$\phi_k = \{\mu_k, \sigma^2_k\}$$

- Then, introduce latent indicator variable z_{ij} indicating which component the observation ij belongs to. In other words,

$$\theta_{ji} = \phi_{(z_{ji}=k')}$$

HDP Model with Latent Indicator Variables (3)

Hence, by using z_{ij}, we can write the model as

$$y_{ij} \mid z_{ij}, \{\mu_k\}_{k=1}^{\infty}, \{\sigma_k^2\}_{k=1}^{\infty} \sim N(\mu_{z_{ij}}, \sigma_{z_{ij}})$$

$$z_{ij} \mid \{\pi_{jk}\}_{k=1}^{\infty} \sim \{\pi_{jk}\}_{k=1}^{\infty}$$

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

$$\beta_k \sim \beta'_k \prod_{l=1}^{k-1}(1-\beta'_l) \quad \beta'_k \sim \mathrm{Beta}(1,\gamma)$$

$$\begin{cases} \mu_k \mid H_\mu \sim H_\mu \\ \sigma_k^2 \mid H_{\sigma^2} \sim H_{\sigma^2} \end{cases}$$

A Change of Expression

In the initial expression, we have

$$\boldsymbol{\beta}|\gamma_0 \sim \mathrm{GEM}(\gamma_0)$$

$$\boldsymbol{\pi}_j|\alpha_0, \boldsymbol{\beta} \sim \mathrm{DP}(\alpha_0, \boldsymbol{\beta})$$

$$z_{ij}|\boldsymbol{\pi}_j \sim \boldsymbol{\pi}_j$$

However, this expression is not suitable for doing slice sampling, as $\boldsymbol{\pi}$ is heavily based on $\boldsymbol{\beta}$. Note that $\boldsymbol{\pi}$ itself is a DP.

$$\boldsymbol{\beta}|\gamma \sim \mathrm{GEM}(\gamma_0)$$

$$\gamma_j|\alpha_0 \sim \mathrm{GEM}(\alpha_0), \gamma_j = (\gamma_{jt})$$

$$k_{jt}|\boldsymbol{\beta} \sim \boldsymbol{\beta}, t = 1, 2, 3, \ldots$$

$$\boldsymbol{\pi}_j = \sum_{t=1}^{\infty} \gamma_{jt}\delta_{k_{jt}}$$

$$z_{ji}|\boldsymbol{\pi}_j \sim \boldsymbol{\pi}_j.$$

A Change of Expression Cont.

This can also be written as the following.

$$\boldsymbol{\beta}|\gamma \sim \text{GEM}(\gamma_0)$$

$$\gamma_j|\alpha_0 \sim \text{GEM}(\alpha_0), \gamma_j = (\gamma_{jt})$$

$$k_{jt}|\boldsymbol{\beta} \sim \boldsymbol{\beta}, t = 1, 2, 3, ...$$

$$t_{ij}|\boldsymbol{\gamma}_j \sim \boldsymbol{\gamma}_j, i = 1, 2, ..., n_j$$

$$z_{ji}|\boldsymbol{t_j}, \boldsymbol{k}_j = k_{i,t_{ij}}.$$

We can get the joint density of $\boldsymbol{t}, \boldsymbol{k}, \boldsymbol{\gamma}, \boldsymbol{\beta}$, and use this for sampling. (Coding is still in progress...)

Thanks!