Slice sampling in nested IBP

Jinsan Yang, Jinseok Nam and Byoung-Tak Zhang
Biointelligence Laboratory
School of Computer Science and Engineering
Seoul National University
Seoul 151-742, Korea
jsyang@bi.snu.ac.kr

Abstract

We develop a nonparametric Bayesian method that explores the infinite space of latent features and finds the best subset in the sense of posterior probability. When the data appear in several groups, there should be different measures reflecting the differences between the groups. We formalize this as a nested Indian buffet process (nIBP) by assuming different measures according to the specific types where corresponding set of groups belong to. For efficiently running the Gibbs sampling, slice sampling method is applied to our model with proper choice of stick components. Our contributions here are two fold. First we extended the Indian buffet process model by imposing nested structure above the groups of data sets and hence providing a hierarchically clustering method for objects having multiple features in nonparametric approach without imposing class numbers or feature numbers. We followed similar formulation as in nDP model by extending the IBP model instead of the DP model (Rodriguez, Dunson & Gelfand, 2008). Secondly in the computation of the stick breaking expressions, we applied the slice sampling method used in IBP model for properly expanding the stick components (Teh, Gorur & Ghahramani, 2007).

1 Introduction

In the analysis of the observations characterized by a certain set of latent features, the Indian buffet process (IBP) method is widely applied (Griffiths & Ghahramani, 2005). In IBP, the number of classes or features is assumed potentially unbounded and the observed objects reflect a subset of these features (Rasmusen & Ghahrahmani, 2001). Instead of using a fixed set of latent features, each observed object is modeled very generally using unbounded number of latent features selected based on the beta process (Thibaux & Jordan 2007). Hierarchical Dirichlet process (HBP) models assume multiple groups of data with observed objects in each group. In HDP, each group follows a Dirichlet process with common measure and this common measure is again comes from a Dirichlet process with a base measure (Teh, Jordan, Beal & Blei, 2006). Therefore, different groups of data share clusters from a common set of clusters.

Recently, several approaches have been proposed to combine feature selections and data clusterings together. DP-IBP and IBP-IBP models (Doshi-Velez, & Ghahramani, 2009) assign different feature sets to each data cluster by assuming single (DP-IBP) or multiple (IBP-IBP) cluster memberships to each data observation. These models consider data clusterings and feature selections for each data together and provide models for the correlated features.

In the nested Dirichlet process (Rodriguez, Dunson & Gelfand, 2008), two stage clustering is considered, where clustering separates distributions of each data group as well as observed data in each nested group. The nested Dirichlet process looks similar to the nested Chinese restaurant process (Blei, Griffiths, Jordan & Tenenbaum, 2004) in selecting clusters hierarchically in each level but in nCRP, individual data point is provided without group memberships and the selection is repeated to

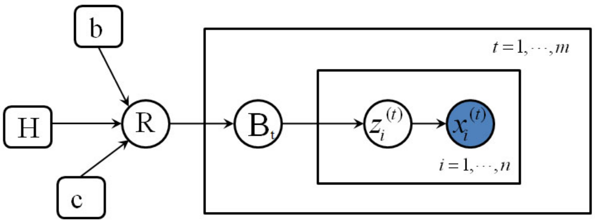

Figure 1: Graphical representation of the nested IBP model. Nodes represent the fixed hyperparameters (rounded cornered squares), variables (circles) and an observable (shaded circle). Rectangles denote the replications with the replication number on the corner. Arrows between the nodes represent the dependency relations.

make a path over the tree structure for each observed data.

We consider the case when there are multiple groups of data and classifying each group depends on the different set of latent features used to characterize each data group. When there are several groups of observed objects, we consider the clustering of the groups as well as the finding of latent features for each observed objects nested in each group. Therefore, the measure for modeling the set of observed objects are sampled from a Dirichlet process and the set of latent features are selected nested in that measure.

2 Nested IBP model

2.1 Nonparametric Bayesian Models

In nonparametric Bayesian models, the model complexity is unbounded and the prior over the underlying distribution is not limited by the parametric family but includes the space of all distributions. The Chinese restaurant process (CRP) and the Indian buffet process (IBP) are two of the most popular constructive processes in nonparametric Bayesian models relating to Dirichlet process and beta process, respectively (Teh, Jordan, Beal & Blei, 2006; Griffiths & Ghahramani, 2005).

$G \sim DP(\alpha G_0)$ is a Dirichlet process where α is a concentration parameter, G_0 is a base measure over some probability space Θ and G is a random measure over any partition $A_1, A_2, \cdots, A_K$ of Θ with distributional property, $(G(A_1), \cdots, G(A_K)) \sim dir(\alpha G_0(A_1), \cdots, \alpha G_0(A_K))$. Since G is a distribution over Θ, the parameters could be sampled from G.

The Indian buffet process is a constructive sampling process for generating sets of feature selections based on the beta process (Griffiths & Ghahramani, 2005; Thibaux & Jordan, 2007) formulated as: $Z \sim BeP(B)$, $B \sim BP(cB_0)$ where c is a concentration parameter, B_0 is a base measure over latent feature space Ω and Z is a binary row of feature selections sampled from a Bernoulli process with parameters depending on B. B is a random measure with distributional property, $B(A_k) \sim beta(cB_0(A_k), c(1 - B_0(A_k)))$ independently for any $A_k \subset \Omega$.

Stick breaking construction (Sethuraman, 1994) provides an explicit method to construct a sampled measure: For G sampled from $DP(\alpha G_0)$, $G = \sum_{k=1}^{\infty} \pi_k \delta(\theta_k^*)$ where $\delta(\theta_k^*)$ is a point measure concentrated on a representing element θ_k^* for the k^{th} partition of Θ, $\beta_k \sim Beta(1, \alpha)$, $\pi_k = \beta_k \prod_{i=1}^{k-1}(1 - \beta_i)$ and $\theta_k^* \sim G_0$ for each k. Stick breaking construction is very useful in computing the likelihoods in the case of nested DP model due to the expressive representation of arbitrary distributions (Rodriguez, Dunson & Gelfand, 2008).

2.2 Nested Indian buffet process

When the data are observed in several groups, in addition to the finding of latent feature sets in each group, it is also important to categorize each group of objects according to their properties. For example, when there are sets of images and we want to know which image sets can be characterized based on the same sets of latent features in common. We can consider to categorize the groups of images according to their characteristics. Nested Indian buffet process is a model for clustering the general classes as well as finding the set of latent features for each observed object. Nested IBP model should not be confused with the DP-IBP model (Doshi-Velez, & Ghahramani, 2009) where assuming feature correlations, each object is associated with a finite number of categories and each category is associated with a finite number of features.

Denoting $z_i^{(t)}$ for a vector of binary values for selecting latent features for the i^{th} object in the t^{th} group, we can formulate the nested IBP for $i = 1, \cdots, n_t$ and $t = 1, \cdots, m$ as follows:

$$z_i^{(t)} | B_t \sim BeP(B_t) \tag{1}$$

$$B_t \sim R \sim DP(b\tau) \quad (\tau \equiv BP(cH)) \tag{2}$$

In (1), each element of binary vector $z_i^{(t)}$ is sampled from a Bernoulli distribution with success probability determined by a measure B_t over parameter space. In (2), the measure B_t for generating features among the t^{th} group is distributed according to R which is also a measure over $\mathcal{C} = \{B\}$ ($\mathcal{C}$ is a collection of measures like B_j). τ is a measure for generating R with positive concentration parameter b and set to be a beta process with base measure H with a positive concentration parameter c. If B_j^* is a representing member of the j^{th} category in $\mathcal{C}$, it can be sampled from τ in the stick breaking representation. We will show detailed representations in the next section.

3 Stick breaking construction in nested IBP by slice sampling

3.1 Stick breaking construction

Applications of stick breaking method in the defining relations (1) and (2) can provide expressions of R as well as B_j^* , a representing element in the j^{th}category of $\mathcal{C}$ for $j = 1, 2, \cdots$. Denoting $\delta(x)$ as a point mass concentrated to a point x:

$$R = \sum_{j=1}^{\infty} \pi_j \delta(B_j^*), \pi_j = \beta_j \prod_{i=1}^{j-1} (1 - \beta_i) \tag{3}$$

$$\beta_i \sim beta(1, a), \quad B_j^* \sim BP(cH) \tag{4}$$

$$B_j^* = \sum_{k=1}^{\infty} \mu_k^{(j)} \delta(\varphi_k^{(j)}) \tag{5}$$

$$\mu_k^{(j)} = \nu_k^{(j)} \mu_{k-1}^{(j)} = \prod_{i=1}^{k} \nu_i^{(j)} \tag{6}$$

$$\nu_i^{(j)} \sim beta(c, 1) \tag{7}$$

$$\varphi_k^{(j)} \sim H \tag{8}$$

where H is a discrete uniform base measure, $\mu_{l(1)} > \mu_{l(2)} > \cdots > \mu_{l(k)}$ are rearrangements of weights $\mu_{l1}, \mu_{l2}, \cdots, \mu_{lk}$ in the l^{th} category and θ_{lk}^* is the k^{th} feature of the l^{th} category (Teh, Gorur & Ghahramani, 2007). Since there are infinite terms of weights in each representation, the exact procedure is infeasible and truncation is an easy implementing method using the approximation model instead of the full model. But by introducing the slice sampling (Kalli, Griffin, & Walker, 2011), it is possible to avoid infinite number of weights and still take advantages of the full model. In the following section, we will explain the use of slice sampling method in our model.

3.2 Slice sampling

With the development of Markov chain Monte Carlo methods in statistics, the mixture of Dirichlet process (MDP) is one of the most popular models in Bayesian nonparametrics. But the sampled distribution from MDP has countably infinite number of masses that make inferences intractable in many cases. To cope these difficulties, there has been various approaches of approximations ... (restrospective sampler, truncated model,...)

Slice sampling (Damien et al 1999) is a trick in sampling process by introducing certain set of latent variables that can change the sampling of the infinite part of the random distribution to be conditionally finite with easy sampling process.

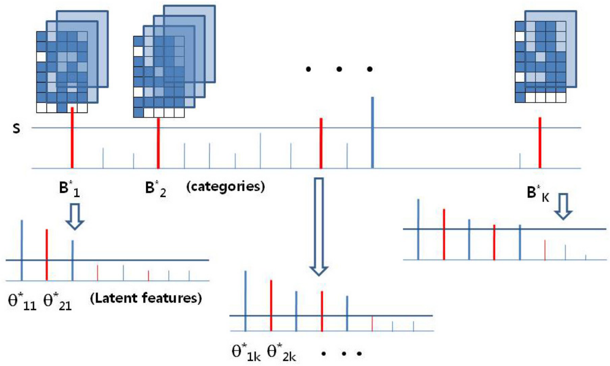

Figure 2: In nested IBP model, each data set is allocated to the constructed stick components taller than the slice value and also each data entry is taking latent features among the feature components with larger weights than the given slice values enabling one to consider only finitely many components out of infinitely many components.

The basic result of slice sampling (Damien et al 1999) states that if f is the required conditional distribution for a random variable v in Gibbs sampler, then it is possible to introduce a latent variable u with f being marginal density and extend the Gibbs sampler to include the additional full conditional of u. Let $f(v)$ be a posterior distribution and $g(v)$ be a prior distribution for a random variable v. Suppose

$$f(v) \propto g(v) \prod_{i=1}^{n} l_i(v) \tag{9}$$

where $l_i : v \to l_i(v)$ are nonnegative invertible functions (possibly but not necessarily densities for the i^{th} observations) so that when $l_i(v) > u_i$ for some positive u_i, it is always possible to find the corresponding v.
By introducing a set of latent variables $u = (u_1, \cdots, u_n)$, $(u_i \in (0, \infty); i = 1, \cdots, n)$ to the posterior density f,

$$f(v, u_1, \cdots, u_n) \propto g(v) \prod_{i=1}^{n} \mathbf{1}\{u_i < l_i(v)\} \tag{10}$$

a Gibbs sampler for generating random variables from f can be constructed with full conditionals for u_i

$$p(u_i|\cdots) = U(0, l_i(v)) \quad i = 1, \cdots, n \tag{11}$$

and a full conditional for v,

$$p(v|\cdots) \propto g(v)\mathbf{1}(v \in \Upsilon_u) \tag{12}$$

where $\Upsilon_u = \{v | l_i(v) > u_i, \text{ for } i = 1, \cdots, n\}$.

3.3 Slice sampling process of nested IBP

The state variables of Gibbs sampler for nested IBP model (from (1) and (2)) are

$$\{(l_i), (\pi_j), (z_i^{(j)}), (\mu_{(k)}^{(j)}); i = 1, \cdots, n, k = 1, 2, \cdots; j = 1, 2, \cdots\}. \tag{13}$$

Denoting $\Pi = \{\mu^{(j)} | j = 1, 2, \cdots\}$ for the collection of stick sets, the full conditionals of $l_i = j$ are given by

$$\begin{aligned} & p(j|\pi, \Pi, z_i^{(j)}, Z_{-i}^{(j)}, x_i, X_{-i}^{(j)},) \\ \propto\ & p(j|\pi) p(z_i^{(j)}|\Pi, j, Z_{-i}^{(j)}) p(x_i, X_{-i}^{(j)}|z_i^{(j)}, Z_{-i}^{(j)}) \\ \propto\ & \pi_j p(z_i^{(j)}|\mu^{(j)}, Z_{-i}^{(j)}) p(x_i, X_{-i}^{(j)}|z_i^{(j)}, Z_{-i}^{(j)}). \end{aligned} \tag{14}$$

Note that $p(z_i^{(j)}|\mu^{(j)}, Z_{-i}^{(j)})$ in (14) represents the likelihood of a binary indicator $z_i^{(j)}$ of the i^{th} data

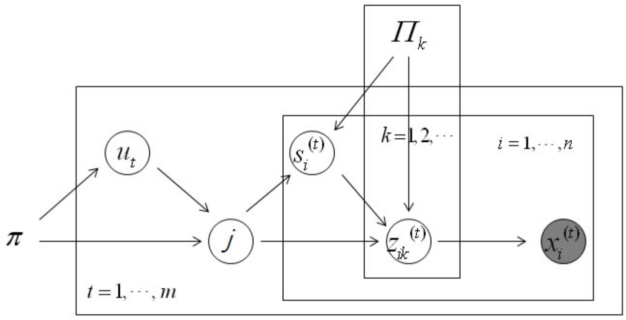

Figure 3: graph for nIBP

with respect to the j^{th} data set type. Since there are infinite number of weights in π and Π, direct computation is infeasible and we introduce latent variables u_i and $s_i^{(j)}$ to truncate weights for the i^{th} data x_i with regard to the j^{th} data set type. We introduce a latent variable $s_i^{(j)}$ and assume the conditional distribution of $(z_i^{(j)}, s_i^{(j)})$ given others as

$$p(z_i^{(j)}, s_i^{(j)}|\cdots) \propto p(z_i^{(j)}|\mu^{(j)}, x_i, X_{-i}^{(j)}, Z_{-i}^{(j)}) \prod_{\substack{k=1,2,\cdots \\ i=1,\cdots,n_j}} \mathbf{1}(s_i^{(j)} < [\mu_{(k)}^{(j)}]^{z_{ik}^{(j)}}). \tag{15}$$

Let j_i^* be the minimum index satisfying the relation:

$$1 - \sum_{j=1}^{j_i^*} \pi_j = \sum_{j=j_i^*+1}^{\infty} \pi_j < u_i. \tag{16}$$

Then the set $\{\pi_j | j = 1, 2, \cdots, j_i^*\}$ satisfying (16) would be large enough for the selection of l_i in (14). The probability distribution of a set of weights $(\pi_1, \cdots, \pi_J)$ generated by (3) is a generalized Dirichlet distribution as in the case of DP. Therefore, the posterior stick breaking construction process becomes (ref)

$$\pi_j = \beta_j \prod_{i=1}^{j-1}(1 - \beta_i) \quad j = 1, 2, \cdots, J-1 \tag{17}$$

$$\beta_i \sim beta(a_i^*, b_i^*) \quad i = 1, \cdots, j-1 \tag{18}$$

$$a_j^* = 1 + \sum_{i=1}^{n} \mathbf{1}(l_i = j), b_j^* = a + \sum_{i=1}^{n} \mathbf{1}(l_i > j). \tag{19}$$

For each $j = 1, \cdots, j^*$ and for each i with $z_{ik}^{(j)}$ for $k = 1, \cdots, K_{i^\dagger}{}^{(j)}$, the full conditional becomes

$$\begin{aligned} & p(j|u_i, \pi, \Pi, s_i^{(j)}, z_i^{(j)}, Z_{-i}^{(j)}, x_i, X_{-i}^{(j)}) \\ \propto & \frac{1}{\mu_{i^{**}}^{(j)}} p(z_i^{(j)}|\mu^{(j)}, Z_{-i}^{(j)}) p(x_i, X_{-i}^{(j)}|z_i^{(j)}, Z_{-i}^{(j)}). \end{aligned} \tag{20}$$

Now the state variables of Gibbs sampler for nested IBP model (from (1) and (2)) are

$$\{(l_i, u_i), (\pi_j), (z_i^{(j)}, s_i^{(j)}), (\mu_{(k)}^{(j)}); i = 1, \cdots, n_j; j = 1, 2, \cdots; k = 1, 2, \cdots\} \tag{21}$$

where full conditionals of $l_i, u_i, z_i^{(j)}$ and $s_i^{(j)}$ are given by (20), (22), (24) and (23).

$$p(u_i|\cdots) \propto \mathbf{1}(u_i < \pi_j). \tag{22}$$

$$s_i^{(j)}|\mu^{(j)}, Z^{(j)} \sim Unif(0, \mu_{i^*}^{(j)}) \tag{23}$$

$$p(z_i^{(j)}|s_i^{(j)}, \mu^{(j)}, Z_{-i}^{(j)}, x_i, X_{-i}) \propto p(z_i^{(j)}|\mu^{(j)}, Z_{-i}^{(j)}, x_i, X_{-i}) \frac{1}{\mu_{i^{**}}^{(j)}} \mathbf{1}(z_i^{(j)} \in B_{s_i}^{(j)}) \tag{24}$$

For each component $z_{ik}^{(j)}$ in (24) for $k = 1, \cdots, {K_{i^\dagger}}^{(j)}$, the full conditional becomes

$$\begin{aligned} &p(z_{ik}^{(j)}|s_i^{(j)}, \mu^{(j)}, Z_{-ik}^{(j)}, x_i, X_{-i}^{(j)}) \\ &\propto p(z_{ik}^{(j)}|\mu^{(j)}, Z_{-ik}^{(j)}, x_i, X_{-i}^{(j)}) \frac{1}{\mu_{i^{***}}^{(j)}}. \end{aligned} \tag{25}$$

In terms of the feature weight and likelihood, (25) becomes

$$\begin{aligned} p(z_{ik}^{(j)} = 1|s_i^{(j)}, \mu^{(j)}, Z_{-ik}^{(j)}, x_i, X_{-i}^{(j)}) & \\ \propto p(z_{ik}^{(j)} = 1|\mu^{(j)}, Z_{-ik}^{(j)}, x_i, X_{-i}^{(j)}) \frac{1}{\mu_{i^{***}}^{(j)}} & \\ \propto \frac{\mu_{(k)}^{(j)}}{\mu_{i^{***}}^{(j)}} h(x_i, X_{-i}^{(j)}|Z_{-ik}^{(j)}, z_{ik}^{(j)} = 1). & \end{aligned} \tag{26}$$

From the basic properties of order statistics, the conditional probability of $\mu_{(k+1)}^{(j)}$ conditional on $\mu_{(1:k)}^{(j)}$ when $\mu_k^{(j)}$ has a prior distribution $beta(\frac{c^{(t)}}{K}, 1)$ for $k = 1, \cdots, K$ is given by (ref),

$$p(\mu_{(k+1)}^{(j)}|\mu_{(1:k)}^{(j)}) = c^{(j)}[\mu_{(k+1)}^{(j)}]^{c^{(j)}-1}[\mu_{(k)}^{(j)}]^{-c^{(j)}}\mathbf{1}(0 < \mu_{(k+1)}^{(j)} < \mu_{(k)}^{(j)}). \tag{27}$$

Setting $\nu_k^{(j)} = \frac{\mu_{(k)}^{(j)}}{\mu_{(k-1)}^{(j)}}$, each weight $\mu_{(k)}^{(j)}$ can be obtained by stick breaking construction from $\nu_k^{(j)} \sim beta(c^{(j)}, 1)$, for $k = 1, 2, \cdots$ (ref). The full conditional of component weights $\mu_{(k)}^{(j)}$ is given as,

$$p(\mu_{(k)}^{(j)}|\mu_{(-k)}^{(j)}, Z^{(j)}) \propto \mu_{(k)}^{m_{\cdot,k}^{(j)}-1}(1 - \mu_{(k)})^{n^{(j)}-m_{\cdot,k}^{(j)}}\mathbf{1}(\mu_{(k+1)}^{(j)} < \mu_{(k)}^{(j)} < \mu_{(k-1)}^{(j)}) \tag{28}$$

and the $\mu_{(k)}^{(j)}$ for the new feature $k^{(j)}$ is distributed as

$$p(\mu_{(k)}^{(j)}|\mu_{(1:k-1)}^{(j)}, Z_{\cdot,\geq k}^{(j)} = 0) \propto [\mu_{(k)}^{(j)}]^{c^{(j)}-1}(1 - \mu_{(k)}^{(j)})^{n_j} \exp\left(c^{(j)} \sum_{i=1}^{n_j} \frac{(1 - \mu_{(k)}^{(j)})^i}{i}\right)\mathbf{1}(0 < \mu_{(k)}^{(j)} < \mu_{(k-1)}^{(j)}). \tag{29}$$

4 Experimental results

5 Conclusions and discussions

We proposed a model for analyzing large scales data sets using both clustering and feature allocation methods with effective sampling method. We will test our model with various image data for further properties and possibilities.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MEST) (No. 2011-0016483), the Industrial Strategic Technology Development Program (10035348) funded by the Korean government (MKE), and the BK21-IT Program. The ICT at Seoul National University provides research facilities for this study.

References

[1] Antoniak, C. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. The Annals of Statistics, 2, 1152-1174.

[2] Blei, D. M., Griffiths, T. L., Jordan, M. I., and Tenenbaum, J. B. (2004). Hierarchical topic models and the nested Chinese restaurant process. In Advances in Neural Information Processing Systems, volume 16, Cambridge, MA. MIT Press.

[3] Doshi-Velez, F. & Ghahramani, Z. (2009) Correlated non-parametric latent feature models In: UAI 2009 (Conference on Uncertainty in Artificial Intelligence), 18-21 June 2009, Montreal, Quebec.

[4] Escobar, M. D., & West, M. (1995). Bayesian density estimation and inference using mixtures. Journal of the American Statistical Association, 90, 577-588.

[5] Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. In M. Rizvi, J. Rustagi, & D. Siegmund (Eds.), Recent advances in statistics (p. 287-302). New York: Academic Press.

[6] Ishwaran, H., & James., L.F. (2001). Gibbs sampling methods for stick-breaking priors. Journal of the American Statistical Association, 96(453):161-173.

[7] Kalli, M., Griffin, J.E., & Walker, S.G. (2011). Slice sampling mixture models. Journal of Statistics and Computing, Vol. 21, Issue 1.

[8] Griffiths, T., & Ghahramani, Z. (2005). Infinite latent feature models and the Indian buffet process. TR 2005-001, Gatsby Comp. Neuroscience Unit.

[9] McLachlan, G. J. and Peel, D. (2000). Finite Mixture Models, Wiley.

[10] Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. Journal of Computational and Graphical Statistics, 9, 249-265.

[11] Neal, R. M. (2003). Slice sampling, Ann. Statist. Volume 31, Number 3, 705-767.

[12] Paisley, J., Zaas, A., Woods, C. W., Ginsburg, G. S., & Carlin, L. (2010). A Stick-Breaking Construction of the Beta Process. ICML2010, Haifa, Israel.

[13] Rodriguez, A., Dunson, D. B., Gelfand, A. E. (2008). The nested Dirichlet process. Journal of the American Statistical Association, Vol. 103, No. 483, pp. 1131-1154.

[14] Sethuraman, J. (1994). A constructive definition of Dirichlet priors. Statistica Sinica, 4:639-650.

[15] Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. Journal of the American Statistical Association, 101(476):1566.1581.

[16] Teh, Y.W.,Gorur, D.,Ghahramani, Z. (2007). Stick-breaking Construction for the Indian Buffet Process. Proceedings of the International Conference on Artificial Intelligence and Statistics, Vol. 11.

[17] Thibaux, R. & Jordan, M. I. (2007). Hierarchical beta processes and the Indian buffet process. In Proceedings of the International Workshop on Artificial Intelligence and Statistics, volume 11.

Overview of slice sampling algorithm for the nested IBP

Setting parameters $\alpha, a, \sigma_x, \sigma_A$,
Initialize $Z^{(j)}, J^{\dagger}, {K^{\dagger}}^{(j)}, \pi, \mu^{(j)}$

(Iterative process)
for $j = 1, \cdots, J^{\dagger}$
  update π_j of types
end
for $i = 1, \cdots, N$
  expand type number using slice value u ($j = 1, \cdots, J^{\dagger}$)
  for $j = 1, \cdots, J^{\dagger}$
    expand $\mu_{(\cdot)}^{(j)}$ in each type
    update $Z_{i,\cdot}^{(j)}$ in each type
  end
  choose type for input data x_i
  sample u
end
for $j = 1, \cdots, J^{\dagger}$
  for $k = 1, \cdots, {K^{\dagger}}^{(j)}$
    update ${\mu_{(k)}}^{(j)}$ in each type
  end
end