
GURLS_{mk1}: A PFBS-based Implementation for Multiple Kernel Learning

(Jeremiah) Zhe Liu
Department of Biostatistics
Harvard University
Boston, MA 02115
zh1112@mail.harvard.edu

Contents

1	Hierarchical Dirichlet Process	1
1.1	Model	1
1.2	Inference	1
1.3	Application: Clustering Hierarchical Gaussian Data	2
2	HDP for Hidden Markov Model	2
A	Chinese Restaurant Franchise	3
B	References	4

1 Hierarchical Dirichlet Process

1.1 Model

Classic view:

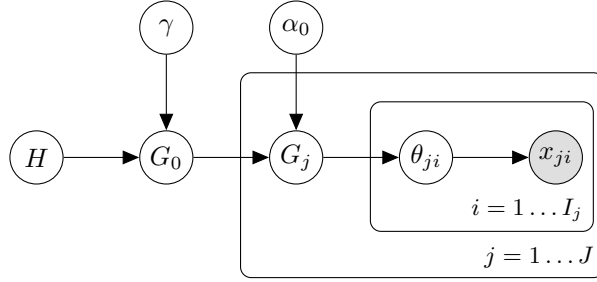
$$\begin{aligned} G_0 | \gamma, H &\sim DP(\gamma, H) \\ G_j | \alpha_0, G_0 &\sim DP(\alpha_0, G_0) \\ \theta_{ji} | G_j &\sim G_j \\ x_{ji} | \theta_{ji} &\sim F(\theta_{ji}) \end{aligned}$$

where $P \sim DP(\alpha, G)$ adopts the stick breaking representation w.p. 1:

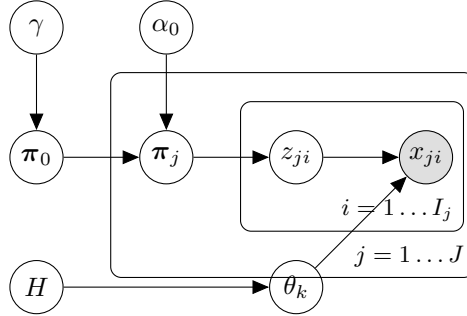
$$P = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k} \quad \text{where:} \quad \pi_k \sim GEM(\alpha), \quad \phi_k \sim G$$

Alternatively, one may describe the generative processes of π_k and θ_k separately as:

$$\begin{aligned} \pi_0 | \gamma &\sim GEM(\gamma) & \theta_k | H &\sim H \\ \pi_j | \alpha_0, \pi_0 &\sim DP(\alpha_0, \pi_0) \\ z_{ji} | \pi_j &\sim \pi_j \\ x_{ji} | z_{ji}, (\theta_k)_{k=1}^{\infty} &\sim F(\theta_{z_{ji}}) \end{aligned}$$



(a) Hierarchical Dirichlet Process



(b) Hierarchical Dirichlet Process

Figure 1: Hierarchical Dirichlet Process

1.2 Inference

Assuming conjugacy between H and F ¹ and holding (γ, α_0) fixed,

we now describe a simplified Gibbs approach to sample parameters (z_{ji}, m_{jk}, π_0) from the Chinese Restaurant Franchise (see Appendix A) representation of the posterior, where the parameter z_{ji} are referred to respectively as customer-specific dish assignment, m_{jk} as dish-specific table count, and π_0 as global dish distribution. This

¹so we can integrate out the mixture component parameters

particular method is referred to as "direct assignment" in Teh et al. [2006] since it circumvented the issue of bookkeeping for every t_{ij} (customer-specific table assignment) and k_{jt} (table-specific dish assignment) variables.

In each Gibbs iteration, denote $f_k^{-x_{ji}}(x_{ji}) = \frac{\int f(\mathbf{x}|\theta_k)h(\theta_k)d\theta_k}{\int f(\mathbf{x}_{-(ji)}|\theta_k)h(\theta_k)d\theta_k}$ the conditional distribution $x_{ji}|\mathbf{x}_{-(ji)}$ under $\theta = \theta_k$, and assume there are currently K dishes and T tables, we sample (z_{ji}, m_{jk}, π_0) iteratively as:

1. Sample $z_{ji} = k|\mathbf{z}_{-(ji)}, \mathbf{m}, \pi_0$ from the distribution:

$$z_{ji} = k|\mathbf{z}_{-(ji)}, \mathbf{m}, \pi_0 \propto \begin{cases} f_k^{-x_{ji}}(x_{ji}) * (n_{jk}^{-(ji)} + \alpha_0 \pi_{0,k}) & k \leq K \\ f_{K+1}^{-x_{ji}}(x_{ji}) * \alpha_0 \pi_{0,u} & k = K + 1 \end{cases}$$

2. Sample $m_{jk} = m|\mathbf{z}, \mathbf{m}_{-(jk)}, \pi_0$, by setting $m_{jk} = \sum_i I(t_{ji} = t_{new} | k_{jt_{new}} = k)$, we can sample t_{ji} from:

$$t_{ji} = t|k_{jt} = k, \mathbf{t}_{-(ji)}, \pi_0 \propto \begin{cases} n_{jt}^{-(ji)} & t \leq T \\ \alpha_0 \pi_{0,k} & t = T + 1 \end{cases}$$

and as in Fox [2009], sample $I(t_{ji} = t_{new} | k_{jt_{new}} = k)$ directly from:

$$\text{Bern}\left(\frac{\alpha_0 \pi_{0,k}}{n_{jk} + \alpha_0 \pi_{0,k}}\right)$$

3. Sample π_0 from distribution:

$$\pi_0 \sim \text{Dir}(m_1, \dots, m_K, \gamma)$$

Algorithm 1 HDP, Gibbs Sampler through Direct Assignment

```

1: procedure hdp_gibbs_ds( $\mathbf{K}, \mathbf{y}, (\tau, \mu, \sigma)$ )
2:    $\alpha^0 = \mathbf{0}$ 
3:   for  $p = 1$  to MAX_ITER do
4:      $\alpha_p^0 = (1 - \frac{\mu}{\sigma})\alpha^{p-1} - \frac{1}{\sigma n}(\mathbf{K}\alpha^{p-1} - \mathbf{y})$ 
5:      $\alpha^p = \mathbf{S}_{\frac{\tau}{\sigma}}(K, \alpha_p^0)$ 
6:   end for
7:   return  $f^{\text{MAX\_ITER}} = (\alpha^{\text{MAX\_ITER}})^T \mathbf{k}$ 
8: end procedure

```

1.3 Application: Clustering Hierarchical Gaussian Data

Consider mixture of Gaussian data $\mathbf{x} = \{\mathbf{x}_1, \dots, \mathbf{x}_K\}$ with $\mathbf{x}_k \stackrel{iid}{\sim} MVN(\boldsymbol{\theta}_{k,2 \times 1}, \mathbf{I}_{2 \times 2})$ with unknown mean $\boldsymbol{\theta}$. Assuming diffused Gaussian prior $\boldsymbol{\theta} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$, the form of likelihood F and base measure H are:

$$f(x_{ji}|\boldsymbol{\theta}_k) \propto \exp\left(-\frac{1}{2\sigma^2}(x_{ji} - \boldsymbol{\theta}_k)^T(x_{ji} - \boldsymbol{\theta}_k)\right)$$

$$h(\boldsymbol{\theta}_k) \propto \exp\left(-\frac{1}{2\sigma_0^2}\boldsymbol{\theta}_k^T \boldsymbol{\theta}_k\right)$$

Then $f_k^{-x_{ji}}(x_{ji})$ should be:

$$f_k^{-x_{ji}}(x_{ji}) \sim N\left(\frac{n_k^{-(ji)}\sigma_0^2}{n_k^{-(ji)}\sigma_0^2 + \sigma^2}\bar{\mathbf{x}}_k^{-(ji)}, \left(1 + \frac{\sigma_0^2}{n_k^{-(ji)}\sigma_0^2 + \sigma^2}\right)\mathbf{I}\right)$$

2 HDP for Hidden Markov Model

A Chinese Restaurant Franchise

A hierarchical analogy of Chinese Restaurant Process, the Chinese Restaurant Franchise offers a convenient scheme to sample from the posterior of cluster-specific θ 's in HDP. This process draw below analogy:

- H as the dish distribution for all possible dishes in the world, with the types of possible dishes being $(\theta_k)_{k=1}^{\infty}$.
- $G_0 \sim DP(\gamma, H)$ as the dish distribution for the franchise
- $G_j \sim DP(\alpha_0, G_0)$ as the dish distribution for restaurant j in the franchise
- $\psi_{jt} \sim G_0$ as the dish served at table t in restaurant j .
 $k_{jt} \sim \pi_0$ as the index of dish choice for this table.
- $\theta_{ji} \sim G_j$ as the dish will be enjoyed by customer i in restaurant j .
 $t_{ji} \sim \pi_j$ as the index of table choice for this customer.

Integrating over G_j , the sampling scheme for subject-specific dish $\theta_{ji} \sim G_j$ is:

$$\theta_{ji} | \boldsymbol{\theta}_{j(-i)}, \alpha_0, G_0 \sim \sum_{k=1}^K \frac{n_{jt.}}{\alpha_0 + n_{j..}} \delta_{\psi_{jt}} + \frac{\gamma}{n_{j..} + \gamma} G_0$$

Integrating over G_0 , the sampling scheme for table-specific dish $\psi_{jt} \sim G_0$ is:

$$\psi_{jk} | \Psi_{j(-k)}, \gamma, H \sim \sum_{k=1}^K \frac{m_{.k}}{\gamma + m_{..}} \delta_{\theta_k} + \frac{\gamma}{m_{..} + \gamma} H$$

B References

References

- Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006. ISSN 0162-1459. URL <http://www.jstor.org/stable/27639773>.
- Emily Beth Fox. *Bayesian nonparametric learning of complex dynamical phenomena*. Thesis, Massachusetts Institute of Technology, 2009. URL <http://dspace.mit.edu/handle/1721.1/55111>.