COS 597C: Bayesian Nonparametrics

Lecturer: David Blei — Lecture # 7
Scribe: Min Sun, Mike Wawrzoniak — November 5, 2007

Grouped Data

Let's consider grouped data, which is drawn from a mixture model, and share mixture components between groups. Although the data are separated into groups, there are a underlying links between data so that the data "shares statistical strength". This setting is consistent with the setting of "transfer learning", where learning the statistic in one group enhance the learning in another group.

An example of grouped data is from the field of information retrieval (IR) of modeling of relationships among sets of documents. A document can be commonly viewed as a collection of words. It is also common to view the words in a document as arising from a number of latent clusters or "topics, where a topic is generally modeled as a multinomial probability distribution on words from some basic vocabulary. Finally, topics are usually shared among the documents in the corpus. As a second example, consider a set of images, each image can also be view as a collection of pixels, patches, or features. It's also common to view each of them as a drawn from a number of hidden topics. Finally, topics are shared among the set of images as well.

Hierarchical Dirichlet Process

Data generated from Hierarchical Dirichlet Process (HDP) mixture models exactly satisfy the grouped data characteristic described above. From the graphical model of HDP mixture models shown in 1(a), we know HDP is built on multiple DPs. By adding one more level of DP over G_0, HDP enables data in groups to share countable infinite cluster identities and to exhibit unique cluster propositions. But why is the second level of DP important to guaranty sharing clusters among groups? Let's consider a naive multiple DPs mixture models without the second level of DP (in figure 1(b)). The model models each group as DP mixture and each DP shares the same concentration parameter α and base measure G_0. However, this model doesn't allow groups to share cluster identities when G_0 is not a discrete distribution. By simply adding a second level of DP over G_0 with concentration parameter γ and base measure H , HDP guaranties the discreteness of G_0. Therefore, HDP mixture models yields exactly the grouped data characteristic.

Chinese restaurant franchise(CRF)

In this section we describe an analog of the Chinese restaurant process (CRP) for hierarchical Dirichlet processes which we refer to as the Chinese restaurant franchise. Consider each restaurant serves a group of costumers and all restaurants of the franchise share the same menu of dishes. This is exactly the same setting as grouped data where costumers having the same dishes across restaurants is the analog of data sharing the same cluster identities across groups. In another way, we can imagine HDP as two level of CRP as in figure 2. At the first level, the table assignment θ_{ji} [1] of costumer i in restaurant j is drawn from

[1] Notice each integer i around tables of group j in Figure 2(a) represent θ_{ji}

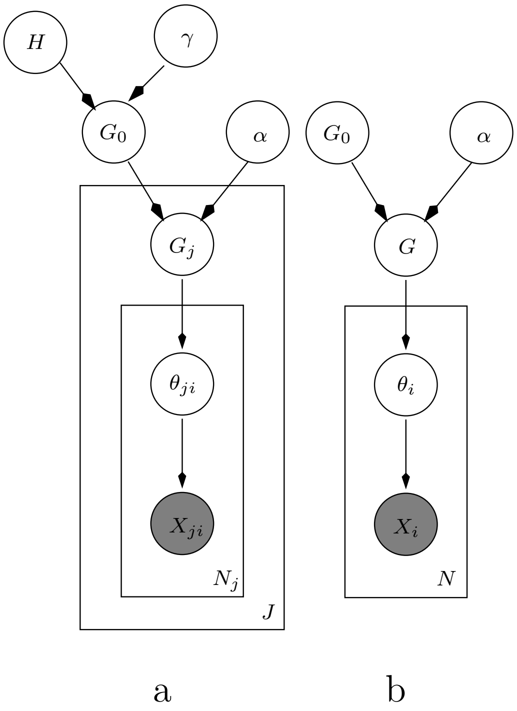

Figure 1: Graphical model representation of an HDP mixture models(a), and an naive multiple DPs mixture models (b).

CRP with parameter α and G_j. Since G_j is drawn from $DP(\gamma, H)$ in the HDP model, the dishes θ^*_{jt} of restaurant j and table t can be model as drawn from the second level CRP with parameter γ and H. And θ^{**} is a unique dish/identity in second level CRP.

Notice that, in CRF model, dishes in different tables might be the same in each group. For example, in figure 2(b), θ^*_{23} and θ^*_{21} share the same dish/identity θ^{**}_1. This is different from CRP model when different tables represent different cluster identity. However, in CRF, the probability of the next costumer having dish d (no matter in what table) is the same as if there is a big table merging all the customers having dish d in different tables. Therefore, dish index works as the cluster identity in each group of CRF model, compares to table index works as cluster identity in CRP model. The subtlety further makes HDP have two different stick-breaking constructions which will be discussed in the following section.

The stick-breaking construction

The HDP construction can also be represented in the stick-breaking interpretations as

$$\begin{aligned}
\beta &\sim GEM(\gamma) \\
\theta^{**}_i &\sim H \\
G_0 &= \sum_{i=1}^{\infty} \beta_i \delta(\theta^{**}_i) \\
\pi_j &\sim GEM(\alpha) \\
\theta^*_{ji} &\sim G_0
\end{aligned}$$

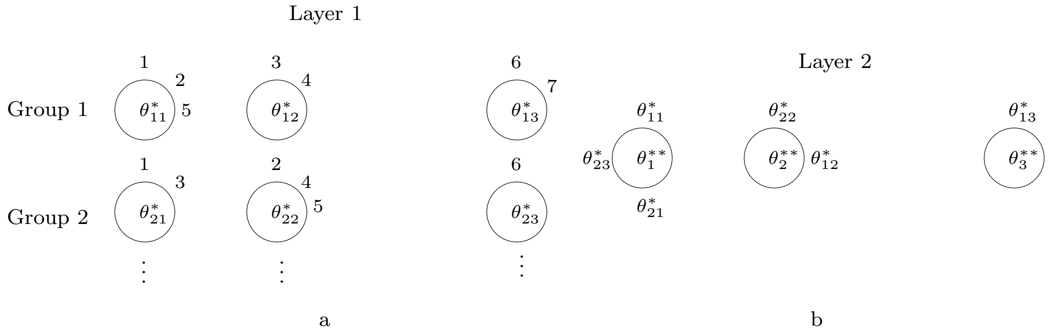

Figure 2: Chinese restaurant franchise representation of HDP. Each group as a 1st level CRP in (a). 2nd level CRP in (b).

$$G_j \quad = \quad \sum_{i=1}^{\infty} \pi_{ji} \delta(\theta^*_{ji}) \ . \tag{1}$$

Because G_0 has support at the points θ^{**}_i, G_j has support at these points as well, and therefore can also be written as

$$G_j = \sum_{i=1}^{\infty} \omega_{ji} \delta(\theta^{**}_i) \ . \tag{2}$$

Since $G_j \sim \mathrm{DP}(\alpha, G_0)$, then for a measurable partition $(A_1, \ldots, A_r)$ of Θ

$$(G_j(A_1), \ldots, G_j(A_r)) \sim \mathrm{Dir}(\alpha G_0(A_1), \ldots, \alpha G_0(A_r)) \tag{3}$$

Therefore, if for $l = 1, \ldots, r$ let $I_l = \{i : \theta^{**}_i \in A_l\}$

$$\left(\sum_{i \in I_1} \omega_{ji}, \ldots, \sum_{i \in I_r} \omega_{ji} \right) \sim \mathrm{Dir}\left(\alpha \sum_{i \in I_1} \beta_i, \ldots, \alpha \sum_{i \in I_r} \beta_i \right) \tag{4}$$

Hence, $\omega_j \sim \mathrm{DP}(\alpha, \beta)$.

The random probability measure ω_j is also produced with the stick-breaking construction

$$\begin{aligned} \omega'_{ji} \quad &\sim \quad \mathrm{beta}\left(\alpha\beta_i, \alpha \left(1 - \sum_{l=1}^{i} \beta_l \right) \right) \\ \omega_{ji} \quad &= \quad \omega'_{ji} \prod_{l=1}^{i-1} (1 - \omega'_{jl}) \ , \end{aligned} \tag{5}$$

and also by

$$\omega_{ji} \sim \mathrm{beta}(\alpha\beta_i, \alpha(1 - \beta_i)) \ . \tag{6}$$

To arrive at eq. 6 and eq. 5, note that for partition $(1, \ldots, i-1, i, i+1, i+2, \ldots)$ by eq. 4

$$\left(\sum_{l=1}^{i-1} \omega_{jl}, \omega_{ji}, \sum_{l=i+1}^{\infty} \omega_{jl} \right) \sim \mathrm{Dir}\left(\alpha \sum_{l=1}^{i-1} \beta_l, \alpha\beta_i, \alpha \sum_{l=i+1}^{\infty} \beta_l \right) \tag{7}$$

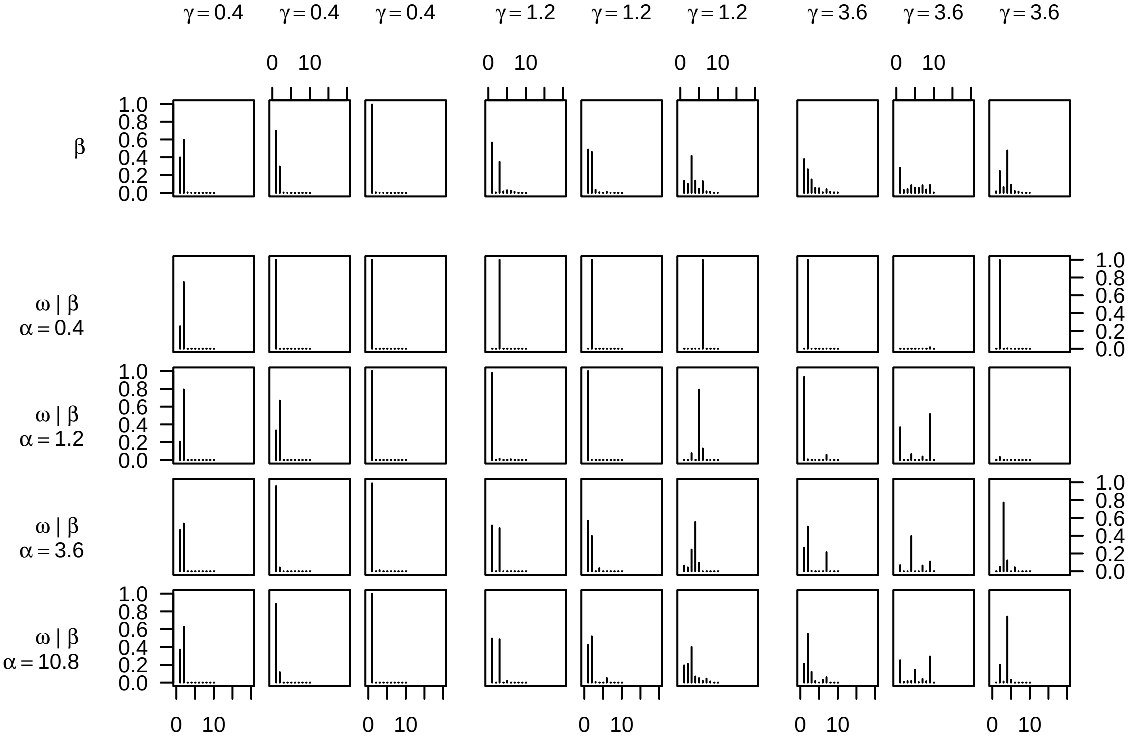

Figure 3: Random draw for β and γ. The top row shows 9 draws for β for γ of $0.4, 1.2$ and 3.6. Below each draw of β, 4 draws for ω are show given the draw of β.

Eq. 6 follows by standard properties of Dirichlet distribution. Also by standard properties of Dirichlet distribution,

$$\left(\frac{\omega_{ji}}{1-\sum_{l=1}^{i-1}\omega_{jl}}, \frac{\sum_{l=i+1}^{\infty}\omega_{jl}}{1-\sum_{l=1}^{i-1}\omega_{jl}}\right) \sim \text{Dir}\left(\alpha\beta_i, \alpha\sum_{l=i+1}^{\infty}\beta_l\right) . \tag{8}$$

Then defining,

$$\omega'_{ji} = \frac{\omega_{ji}}{1-\sum_{l=1}^{i-1}\omega_{jl}} , \tag{9}$$

and therefore,

$$\omega_{ji} = \omega'_{ji}\prod_{l=1}^{i-1}(1-\omega'_{jl}) . \tag{10}$$

Together with

$$1-\sum_{l=1}^{i}\beta_l = \sum_{l=i+1}^{\infty}\beta_l \tag{11}$$

arrive at eq. 5.

The concentration parameters γ, α and the baseline probability measure H are the hyperparameters of an HDP. For small values of γ, the mass is concentrated on a few atoms of H, as the value of γ increases, mass shifts away to be more spread out. This can be observed on the top row of figure 3 with draws for different values of γ. Similarly, for small

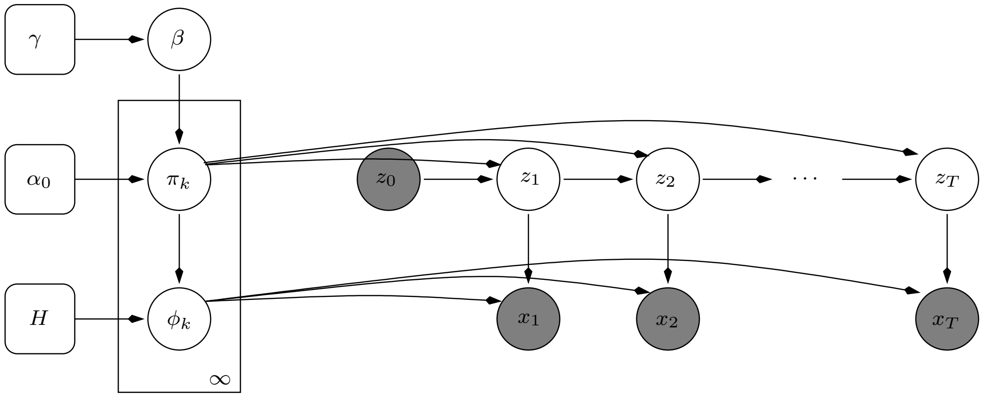

Figure 4: A Graphical representation of the HDP-HMM model.

values of α, mass is concentrated on few atoms of G_0, and as it increases the mass is less concentrated and more spread out. Since the distribution of G_0 is govern by the parameter γ, the parameter α can be interpreted as a refinement of the concentration on H set by γ. This can be seen in figure 3 with draws of ω for different values of α given a particular draw of β.

To formulate the HDP mixture model in the stick-breaking representation, let θ_{ji} be the factors corresponding to a single observation x_{ij}, and let :

$$\begin{aligned} \theta_{ji} &\sim G_j \\ x_{ji} &\sim F(\theta_{ji}) \end{aligned}$$

where $F(\theta_{ji})$ denotes the distribution of the observation x_{ji}.

Hierarchical Dirichlet Process Hidden Markov Model (HDP-HMM)

One application of the HDP model is to construct a variant of the Hidden Markov Model (HMM) that is not restricted to a fixed number of states. In the HMM model, a sequence of multinomial state variables (z_1, z_2, $\ldots$, z_T) is linked through a state transition matrix. Each row of the state transition matrix describes the mixing proportions of the choice of following state value (z_{t+1}) for a particular current state z_t. In he HMM model, the number of states is fixed, and therefore the transition matrix is fixed in size, in the HDP-HMM model, the number of states is unbounded, and so is the size of matrix, see figure 5. The observations $(x_1, x_2, \ldots, x_T)$ are independent conditional on z_t. Figure 4 shows a graphical representation of the HDP-HMM model.

The model can be describe as follows :

$$\begin{aligned} \beta|\gamma &\sim GEM(\gamma) \\ \pi_k|\alpha_0, \beta &\sim DP(\alpha_0, \beta) \\ \phi_k|H &\sim H \end{aligned} \tag{12}$$

for all k. The state transition distribution at time step t,

$$z_{t+1}|z_t, (\pi_k)_{k=1}^{\infty} \sim \pi_{z_t} \tag{13}$$

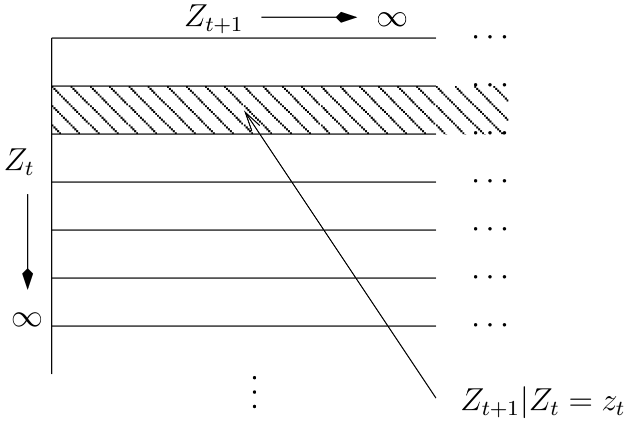

Figure 5: A view of the state transition matrix $P(Z_{t+1}|Z_t)$ of the HDP-HMM. A row of the transition matrix indexed by the value z_t represents the distribution for Z_t, the state at the next time step. In contrast to the classical HMM, in HDP-HMM the number of states is unbounded.

and the observation distribution at time step t,

$$x_t|z_t, (\phi_k)_{k=1}^{\infty} \quad \sim \quad F(\phi_{z_t}) \tag{14}$$

.

The resulting HDP-HMM is a strongly connected automata with countably infinite number of states. However its strength lies in the fact that the base distribution for all of the state transition distributions is shared. Because of this, all of the transition distributions can be viewed as refinements of a common transition distribution, resulting in models that favor fewer common states.

The model parameters are γ, α_0, and H. Intuitively, the parameter γ influences how concentrated is the distribution over states in the model. The parameter α_0 influences the tendency of each transition distribution to be focused on reaching a few other states. The parameter H is the distribution over priors for the observation distributions.