Bayesian Nonparametric Models

David M. Blei
Columbia University

December 15, 2015

Introduction

¶ We have been looking at models that posit *latent structure* in high dimensional data. We use the posterior to uncover that structure.

¶ The two main types are mixtures (and variants, like mixed-membership) and factor models (like PCA, factor analysis, and others).

¶ A nagging concern for these methods is *model selection*—how do I choose the number of mixture components? the number of factors?

¶ This is a field unto itself. You have probably heard of cure-all techniques, like AIC, BIC, and DIC. It can be fitted, for example, with cross-validation or Bayes factors. Sometimes existing theory or problem constraints inform the number of components.

¶ *Bayesian nonparametric modeling* (BNP) provides an alternative solution. BNP methods have enjoyed a recent renaissance in machine learning.

¶ Loosely, BNP models posit an "infinite space" of latent structure where (in the generative process) the data only uses a finite part of it.

- BNP mixtures posit an infinite number of mixture components.
- BNP factor models posit an infinite number of latent factor loadings.

¶ The posterior thus expresses uncertainty about how much of the hidden structure is used and what specific form does it take?

- How many mixture components are expressed in the data? What are they?
- How many factors are there in the data? What are they?
- How many topics are there in this corpus of documents?

¶ OK, fine. But why the hype? Is it just yet another model selection technique? There are a few properties that distinguish BNP models

¶ First, the *predictive distribution* allows for the next data point—the data point to be predicted—to exhibit a previously unseen piece of the latent structure.

- Contrast this to classical model selection, for example, where the number of components is fitted and fixed.

¶ Second, BNP methods can be adapted to model uncertainty over complicated latent *structures*, like trees and graphs.

- Traditional model selection, which usually focuses on selecting a single integer, has difficulty with these kinds of problems.
- This is particularly relevant in hierarchical settings, where we might have multiple per-group model selection problems as well as global per-data problems.
- We can extend the idea of model selection—the inference of the structure of our latent space—to *depend* on other variables.
- And think again about the first point—these complicated latent structures can grow and change in the predictive distribution.

¶ Third (and related to above), casting the model selection & estimation parameters together in a single model means that one posterior inference algorithm does both the searching for a good model and filling in the details of that model.

- The algorithms don't waste time on poor settings of the model selection parameter.

¶ In summary, BNP models are flexible and powerful. They are more relevant now because the kinds of latent structure that we look for—and the kinds of data that we look at—are more complex.

¶ But they too are not a cure-all! BNP models need also to be checked. We avoid choosing the number of mixture components, but we make other choices that affect inference.

¶ A quick word on random measures—

- BNP models place a prior over the (infinite dimensional) space of distributions on an arbitrary space. (This is why they are called "nonparametric.")

- Imagine a Dirichlet—a random distribution over k elements—but for a distribution on any space (including continuous spaces). This is a *Dirichlet process*.
  - (Draw a draw from a Dirichlet.)
  - (Draw a random measure.)
  - (Draw a discrete random measure, and note the distinction.)

¶ The random distribution perspective is important for understanding properties of BNP models and for deriving some inference algorithms, like variational inference.

¶ First we will discuss random structures, partitions in particular, and connect those to mixture models. We will discuss random distributions later.

¶ (Talk about the roadmap here.)

The Chinese restaurant process

¶ More formally, BNP models give priors on structures that accommodate "infinite" sizes.

¶ When combined with data, the resulting posteriors give a distribution on structures that can grow with new observations.

¶ The simplest version is the *Chinese restaurant process*, a distribution on *partitions* that can be used as a prior over clusterings.

- The posterior gives a distribution over the clustering of a given data set.
- Different data will use different numbers of clusters.
- *New clusters* can spawn from new data.

¶ Imagine a restaurant with an infinite number off tables and imagine a sequence of customers entering the restaurant and sitting down. (Draw the picture.)

- The first customer enters and sits at the first table.
- The second customer enters and sits at the first table with probability $\frac{1}{1+\alpha}$, and the second table with probability $\frac{\alpha}{1+\alpha}$, where α is a positive real.
- When the nth customer enters the restaurant, she sits at each of the occupied tables with probability proportional to the number of previous customers sitting there, and at

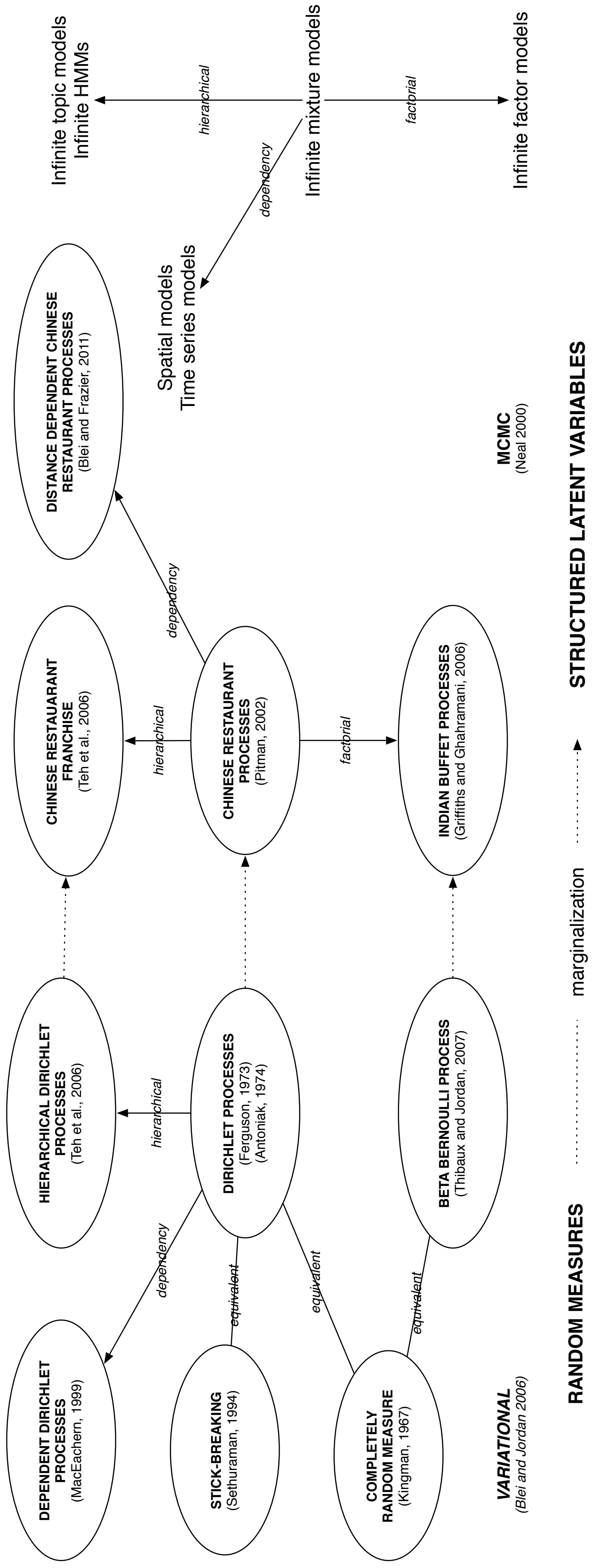

DEPENDENT DIRICHLET PROCESSES (MacEachern, 1999)
STICK-BREAKING (Sethuraman, 1994)
COMPLETELY RANDOM MEASURE (Kingman, 1967)
VARIATIONAL (Blei and Jordan 2006)
dependency
equivalent
equivalent
equivalent
HIERARCHICAL DIRICHLET PROCESSES (Teh et al., 2006)
hierarchical
DIRICHLET PROCESSES (Ferguson, 1973) (Antoniak, 1974)
BETA BERNOULLI PROCESS (Thibaux and Jordan, 2007)
RANDOM MEASURES
marginalization
STRUCTURED LATENT VARIABLES
CHINESE RESTAUARANT FRANCHISE (Teh et al., 2006)
hierarchical
CHINESE RESTAURANT PROCESSES (Pitman, 2002)
factorial
INDIAN BUFFET PROCESSES (Griffiths and Ghahramani, 2006)
dependency
DISTANCE DEPENDENT CHINESE RESTAURANT PROCESSES (Blei and Frazier, 2011)
MCMC (Neal 2000)
Spatial models
Time series models
dependency
Infinite topic models
Infinite HMMs
hierarchical
Infinite mixture models
factorial
Infinite factor models

Figure 1: Roadmap of BNP research

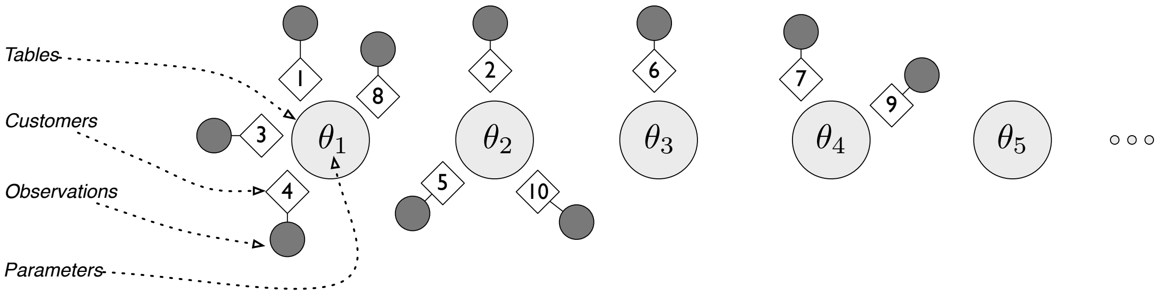


Figure 2: Example of a draw from a CRP and CRP mixture components.

the next unoccupied table with probability proportional to α.

- At any point in this process, the assignment of customers to tables defines a random partition.

¶ (Draw the figure here.)

¶ Let z_i be the table assignment of the nth customer. A draw from this distribution can be generated by sequentially assigning observations to classes with probability

$$P(z_i = k|\mathbf{z}_{1:i-1}) \propto \begin{cases} \frac{m_k}{i-1+\alpha} & \text{if } k \leq K_+ \text{ (i.e., } k \text{ is a previously occupied table)} \\ \frac{\alpha}{i-1+\alpha} & \text{otherwise (i.e., } k \text{ is the next unoccupied table)} \end{cases} \tag{1}$$

where m_k is the number of customers sitting at table k, and K_+ is the number of tables for which $m_k > 0$.

¶ The parameter α is called the *concentration parameter*. Intuitively, a larger value of α will produce more occupied tables (and fewer customers per table).

¶ The CRP exhibits an important invariance property: The cluster assignments under this distribution are *exchangeable*.

¶ This means that $p(\mathbf{z})$ is unchanged if the order of customers is shuffled (up to label changes). This may seem counter-intuitive at first, since the process in Eq. 1 is described sequentially.

¶ (Compute the joint for the figure. Swap one partition and recompute the joint.)

¶ The second important property is about the probability of new customers. The next customer sits at one of the previous tables with some probability. With other probability, she sits at a new table. The partition can expand.

The CRP and exchangeability

¶ We now prove exchangeability. Consider the joint distribution of a set of customer assignments $z_{1:n}$. It decomposes according to the chain rule,

$$p(z_1, z_2, \ldots, z_n) = p(z_1)p(z_2 \,|\, z_1)p(z_3 \,|\, z_1, z_2) \cdots p(z_n \,|\, z_1, z_2, \ldots, z_{n-1}), \qquad (2)$$

where each terms comes from Eq. 1.

¶ To show that this distribution is exchangeable, we will introduce some new notation.

- $K(z_{1:n})$ is the number of groups in which these assignments place the customers, which is a number between 1 and n.
- I_k is an ordered array of indices of customers assigned to the kth group.
- N_k is the number of customers assigned to that group (the cardinality of I_k).

¶ Now, examine the product of terms in Eq. 2 that correspond to the customers in group k. This product is

$$\frac{\alpha \cdot 1 \cdot 2 \cdots (N_k - 1)}{(I_{k,1} - 1 + \alpha)(I_{k,2} - 1 + \alpha) \cdots (I_{k,N_k} - 1 + \alpha)}. \qquad (3)$$

To see this, notice that the first customer in group k contributes probability $\frac{\alpha}{I_{k,1}-1+\alpha}$ because he is starting a new table; the second customer contributes probability $\frac{1}{I_{k,2}-1+\alpha}$ because he is sitting a table with one customer at it; the third customer contributes probability $\frac{2}{I_{k,3}-1+\alpha}$, and so on.

¶ The numerator of Eq. 3 can be more succinctly written as $\alpha(N_k - 1)!$

¶ With this expression, we now rewrite the joint distribution in Eq. 2 as a product over per-group terms,

$$p(z_{1:n}) = \prod_{k=1}^{K} \frac{\alpha(N_k - 1)!}{(I_{k,1} - 1 + \alpha)(I_{k,2} - 1 + \alpha) \cdots (I_{k,N_k} - 1 + \alpha)}. \qquad (4)$$

¶ Notice that the union of I_k across all groups k identifies each index once, because each customer is assigned to exactly one group. This simplifies the denominator and lets us write the joint as

$$p(z_{1:n}) = \frac{\alpha^K \prod_{k=1}^{K} (N_k - 1)!}{\prod_{i=1}^{n} (i - 1 + \alpha)}. \tag{5}$$

¶ Eq. 5 reveals that Eq. 2 is exchangeable. It only depends on the number of groups K and the size of each group N_k. The probability of a particular seating configuration $z_{1:N}$ does not depend on the order in which the customers arrived.

CRP mixtures

¶ The BNP clustering model uses the CRP in an infinite-capacity mixture model.

¶ Each table k is associated with a cluster and with a cluster parameter θ_k^*, drawn from a prior G_0.

¶ There are an infinite number of clusters, though a finite data set only exhibits a finite number of them.

¶ Each data point is a "customer," who sits at a table z_i and then draws its observed value from the distribution $F(y_i | \theta_{z_i})$.

¶ The concentration parameter α controls the prior expected number of clusters (i.e., occupied tables) K_+.

- This number grows logarithmically with the number of customers N,

  $$\mathbb{E}\left[[]\, K_+\right] = \alpha \log N \tag{6}$$

  for $\alpha < N/\log N$.
- If α is treated as unknown, we can put a prior over it.

¶ When we analyze data with a CRP, we form an approximation of the joint posterior over all latent variables and parameters.

¶ We can examine the likely partitioning of the data. This gives us a sense of how are data are grouped, and how many groups the CRP model chose to use.

¶ We can form predictions with the posterior predictive distribution. With a CRP mixture, the posterior predictive distribution is

$$P(y_{n+1}|\mathbf{y}_{1:n}) = \sum_{\mathbf{z}_{1:n+1}} \int_\theta P(y_{n+1}|z_{n+1},\theta)P(z_{n+1}|\mathbf{z}_{1:n})P(\mathbf{z}_{1:n},\theta|\mathbf{y}_{1:n})d\theta. \tag{7}$$

Since the CRP prior, $P(z_{n+1}|\mathbf{z}_{1:n})$, appears in the predictive distribution, the CRP mixture allows new data to possibly exhibit a previously unseen cluster.

Collapsed Gibbs sampling for CRP mixtures

¶ For concreteness, lets make a CRP mixture of Gaussians.

¶ The model is

1. Draw $\theta_k^* \sim N(0, \lambda)$ for $k \in \{1, \ldots\}$
2. For each data point $i \in \{1, \ldots, n\}$
   (a) Draw table $z_i \mid z_{1:(i-1)} \sim \text{CRP}(\alpha, z_{1:(i-1)})$.
   (b) Draw data $x_i \sim N(\theta_{z_i}^*, \sigma^2)$

¶ Repeatedly drawing from this model gives data sets with different numbers of clusters. Contrast to a traditional Gaussian mixture model, where K must be specified.

¶ When G_0 is conjugate to $p(x_n \mid \theta^*)$ (i.e., the model above) then we can use *collapsed Gibbs sampling*, only sampling the table assignments.

¶ We sample the ith table conditioned on the other tables and the data, $p(z_i \mid z_{-i}, x)$. (Recall this defines a Markov chain whose stationary distribution is the posterior.)

¶ The key to this is *exchangeability*. Because the table assignments are exchangeable, we can treat z_i as though it were the last table assignment sampled.

¶ Thus,

$$\begin{aligned} p(z_i = k \mid z_{-i}, x) &= p(z_i = k \mid z_{-i}) p(x \mid z) && (8) \\ &= p(z_i \mid z_{-i}) p(x[I_k(z_{-i})] \cup x_i \mid \lambda) \prod_{j \neq k} p(x[I_j(z_{-i})] \mid \lambda) && (9) \\ &\propto p(z_i \mid z_{-i}) \frac{p(x[I_k(z_{-i})] \cup x_i \mid \lambda)}{p(x[I_k(z_{-i}) \mid \lambda)} && (10) \end{aligned}$$

where $x[I_j(z_{-i})]$ denotes the observations that are assigned (in z_{-i}) to the jth table. Note that the denominator likelihood term does *not* include x_i; the numerator does. The values j, k run over the currently occupied tables and the next unoccupied table.

¶ The first term comes from the CRP. The second term is the usual likelihood integral that we can write down in conjugate-likelihood settings.

¶ Operationally, consider each cluster in z_{-i} and a new cluster.

- Compute the prior of z_i for that cluster.
- Put the ith data point in that cluster and compute the integrated likelihood.
- Multiply these two terms

Finally, normalize and sample. Note again, this allows z_i to start a new cluster.

¶ Exchangeability is key. If things weren't exchangeable then

$$p(z_i \mid z_{-i}, x) = p(z_i \mid z_{1:(i-1)}) p(z_{(i+1):n} \mid z_i) p(x \mid z) \tag{11}$$

and we would have to recompute partition probabilities in the second term for each possible assignment of z_i.

Dirichlet processes

¶ Next, we discuss *Dirichlet process (DP) mixture models*. We will reveal their connection to CRP mixtures in a little while.

¶ The Dirichlet process is a *distribution over distributions* (or, more formally, a random measure model).

¶ It has two parameters:

- The scaling factor α, which is a positive scalar.
- The base distribution G_0, which is an arbitrary distribution.

¶ We draw a random distribution G and then independently draw n values from it,

$$G \sim \mathrm{DP}(\alpha G_0) \tag{12}$$

$$\theta_i \sim G \quad i \in \{1 \ldots n\}. \tag{13}$$

¶ The parameter αG_0 is an *unnormalized measure*. We are construing it as a probability distribution G_0 multiplied by a $scalar$.

¶ Note how general this is. The distribution G is a distribution over whatever space G_0 is a distribution over. (But, of course, it is different from G_0.)

- If G_0 is Poisson, G is a random distribution over integers
- If G_0 is multivariate Gaussian, G is a random distribution over reals.
- If G_0 is Gamma, G is a random distribution over positive reals.

¶ G_0 might be a simple distribution in the exponential family, but G is not.

¶ Two theorems about the DP

- $\mathbb{E}\left[[]\, G\right] = G_0$
- Draws from $DP(\alpha G_0)$ are discrete. (Draw pictures.)

Definition of a DP based on finite dimensional distributions

¶ Consider a probability space Ω and a finite partition A, such that $\bigcup_{i=1}^{k} A_i = \Omega$.

- For concreteness, think of the real line or the real plane.
- A finite partition will include an infinite piece, i.e., the “rest” of the plane or line.

¶ Now consider the vector $\langle G(A_1), G(A_2), \ldots, G(A_k) \rangle$.

- $G(A_i) = \int_{\omega \in A_i} G(\omega)$
- a point on the $k - 1$ simplex because A is a partition of the entire probability space.
- a *random vector*, because G is random

¶ Assume for all k and all possible partitions A,

$$\langle G(A_1), G(A_2), \ldots, G(A_k)\rangle \sim \mathrm{Dir}(\alpha G_0(A_1), \alpha G_0(A_2), \ldots, \alpha G_0(A_k)). \tag{14}$$

These are called finite dimensional distributions (FDD). According to the Kolmogorov consistency theorem—and thanks to properties of these Dirichlets—there is a distribution of G. This is the Dirichlet process (Ferguson, 1973).

¶ A stochastic process is a collection of indexed random variables. In a DP, the index are the Borel sets of the probability space.

¶ This also shows you that $\mathbb{E}\left[[]\, G\right] = G_0$.

- Consider $E[G](A)$ for any set A. This is a probability.
- Put A in a partition with its complement A^c. By the FDD,

$$\begin{aligned} \mathbb{E}\left[[]\, G\right](A) &= \frac{\alpha G_0(A)}{\alpha G_0(A) + \alpha G_0(A^c)} && (15) \\ &= G_0(A) && (16) \end{aligned}$$

- Later, we'll want to write this as

$$\mathbb{E}\left[[]\, G\right] = \frac{\alpha G_0}{\int_\theta \alpha G_0(\theta)} \tag{17}$$

Conjugacy and the clustering effect

¶ Let's consider the posterior of G when we draw a single data point,

$$\begin{aligned} G &\sim \mathrm{DP}(\alpha G_0) && (18) \\ \theta_1 &\sim G && (19) \end{aligned}$$

¶ What is the distribution of $G \mid \theta_1$?

- Consider an arbitrary partition A.
- By definition the distribution of G applied to each set is $\mathrm{Dir}(\alpha G_0(A_1), \ldots, G_0(A_k))$.

- The posterior is given by the posterior Dirichlet

$$\langle G(A_1), \ldots, G(A_k) \rangle \mid \theta_1 \sim \mathrm{Dir}(\alpha G_0(A_1)+\delta_{\theta_1}(A_1), \ldots, \alpha G_0(A_k)+\delta_{\theta_1}(A_k)) \quad (20)$$

- The delta function equals 1 when $\theta_1 \in A_i$. This adds the count to the right component of the Dirichlet parameter.
- But, this is true for any partition. So

$$G \mid \theta_1 \sim \mathrm{DP}(\alpha G_0 + \delta_{\theta_1}). \quad (21)$$

¶ The DP posterior is itself a DP. It is "conjugate to itself." (Usually, conjugacy involves a prior likelihood pair. In this case, the prior generates the likelihood.)

¶ Suppose we observe $\theta_{1:n} \sim G$. What is the predictive distribution?

¶ The predictive distribution is the integral,

$$p(\theta \mid \theta_{1:n}) = \int_G G(\theta) p(G \mid \theta_{1:n}) dG \quad (22)$$

Note that we abuse notation a little. Here $G(\theta)$ is the density (applied to a point) while $G(A_i)$ denotes the integral of the density over the set A_i, i.e., $p(A_i \mid G)$. Above, we are marginalizing out the random distribution G.

¶ This is the expectation of G under its posterior, $G \mid \theta_{1:n}$,

$$\begin{aligned} \mathbb{E}\left[[] G \mid \theta_{1:n}\right] &= \frac{\alpha G_0 + \sum \delta_{\theta_i}}{\int \alpha G_0(\theta) + \sum \delta_{\theta_i}(\theta) d\theta} && (23) \\ &= \frac{\alpha G_0 + \sum \delta_{\theta_i}}{\alpha + n} && (24) \end{aligned}$$

¶ This is starting to look familiar. The distribution of the next θ is

$$\theta \mid \theta_{1:n} \sim \begin{cases} \text{A draw from } G_0 & \text{with probability } \frac{\alpha}{\alpha+n} \\ \theta_i & \text{with probability } \frac{1}{\alpha+n} \end{cases} . \quad (25)$$

¶ Suppose that $\theta_{1:n}$ exhibit K unique values, $\theta^*_{1:K}$. Then, the predictive distribution is

$$\theta \mid \theta_{1:n} \sim \begin{cases} \text{A draw from } G_0 & \text{with probability } \frac{\alpha}{\alpha+n} \\ \theta_k^* & \text{with probability } \frac{n_k}{\alpha+n} \end{cases}, \tag{26}$$

where n_k is the number of times we saw θ_k^* in $\theta_{1:n}$. Notice that which value it takes on (or if it takes on a new value) is given by the CRP distribution.

¶ Consider the DP model with G marginalized out,

$$p(\theta_1, \ldots, \theta_n \mid G_0, \alpha) = p(\theta_1 \mid G_0, \alpha) p(\theta_2 \mid \theta_1, G_0, \alpha) \cdots p(\theta_n \mid \theta_{1:(n-1)}, G_0, \alpha). \tag{27}$$

Each term is a predictive distribution. The cluster structure of $\theta_{1:n}$, i.e., which are equal to the same unique value, is given by a CRP with parameter α.

¶ This means we can draw from this joint distribution (again, with G marginalized out),

$$\begin{aligned} \theta_k^* &\sim G_0 \quad \text{for } k \in \{1, 2, 3, \ldots\} && (28) \\ z_i \mid z_{1:(i-1)} &\sim \text{CRP}(\alpha; z_{1:(i-1)}) && (29) \\ \theta_i &= \theta_{z_i}^* && (30) \end{aligned}$$

¶ The DP mixture (Antoniak, 1974) is

$$\begin{aligned} G &\sim \text{DP}(\alpha G_0) && (31) \\ \theta_i \mid G &\sim G && (32) \\ x_i &\sim f(\cdot \mid \theta_i). && (33) \end{aligned}$$

We now know this is equivalent to a CRP mixture. Note, this also shows exhangeability.

¶ (Draw the graphical model.)

¶ And it shows discreteness of the DP. The CRP reasoning tells us that a countably infinite set of points (called *atoms*) have finite probability. (Contrast to a density, like a Gaussian, where the probability of any particular point is zero.) This means that the distribution is discrete.

¶ Knowing these properties, we can think more about the parameters.

¶ What is the distribution of the atoms? It is G_0. From the CRP perspective, these occur when we sit at a new table.

¶ What affects their probabilities, i.e., how often we see them? It is α. If α is bigger than we revisit tables less often; if α is smaller than we revisit tables more often.

¶ Why? Think about the random measure again. When α is big—from our CRP knowledge—this means that most θ_i are drawn new from G_0. Thus, G will look more like G_0. When α is small, we see old values more often. This means that more values have higher probability (that usual, under G_0) and G looks less like G_0.

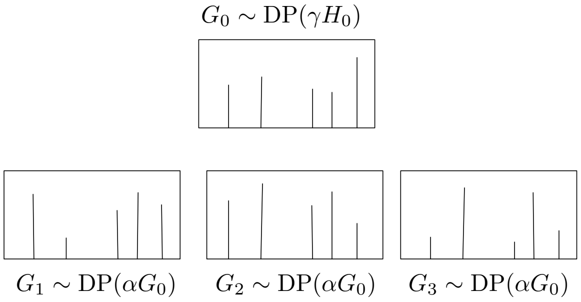
Hierarchical Dirichlet processes

¶ The hierarchical DP (HDP) is a distribution of grouped items,

$$\begin{aligned} G_0 &\sim \mathrm{DP}(\gamma H_0) && (34) \\ G_j \mid G_0 &\sim \mathrm{DP}(\alpha G_0) \quad j \in \{1, \ldots, m\} && (35) \\ \theta_{ij} \mid G_j &\sim G_j \quad i \in \{1, \ldots, n\} && (36) \end{aligned}$$

The parameters are a base measure H_0 and two scaling parameters α and γ.

¶ Here is a schematic for the distributions underlying an HDP



$G_0 \sim \mathrm{DP}(\gamma H_0)$
$G_1 \sim \mathrm{DP}(\alpha G_0)$
$G_2 \sim \mathrm{DP}(\alpha G_0)$
$G_3 \sim \mathrm{DP}(\alpha G_0)$

- At the top, the atoms of G_0 are drawn from H_0.
- In each group, the atoms are drawn from G_0.
- Since G_0 is from a DP, the same atoms are shared *across groups*.

¶ Now imagine this in a mixture setting. Suppose we have grouped data $x_j = \{x_{1j}, \ldots, x_{nj}\}$.

$$\begin{aligned}
G_0 &\sim \mathrm{DP}(\gamma H_0) && (37)\\
G_j \mid G_0 &\sim \mathrm{DP}(\alpha G_0) \quad j \in \{1, \ldots, m\} && (38)\\
\theta_{ij} &\sim G_j \quad i \in \{1, \ldots, n\} && (39)\\
x_{ij} \mid \theta_{ij} &\sim f(\cdot \mid \theta_{ij}) && (40)
\end{aligned}$$

- Each group draws from the same set of atoms, but with different proportions.
- Each observation in each group exhibits exhibits one the per-group atoms.

¶ Does this sound familiar? It is a *Bayesian nonparametric mixed membership model.* The number of components (i.e., the exhibited atoms of G_0) is unknown in advance.

¶ (Draw the graphical model.)

¶ Suppose H_0 is a Dirichlet over a vocabulary.

- The atoms of G_0 are distributions over words. Lets call them topics.
- The atoms of G_j are the same topics, but with different (per-group) proportions.
- Each observation (word) is drawn by choosing a topic from G_j and then drawing from its distribution. The distribution over topics changes from group to group.
- This is a Bayesian nonparametric topic model.

¶ We can integrate out the DPs in an HDP. This gives the "Chinese restaurant franchise."

¶ (Draw the picture of the CRF here.)

¶ Note that within each group, two tables can be assigned to the same atom. (Mathematically, they can be combined to one table.)

¶ There are several inference algorithms for the HDP. Chong Wang has Gibbs sampling code for infinite topic models. Gibbs sampling uses the CRF representation, integrating out the random distributions.

¶ The HDP is not restricted to two levels of hierarchy. Most applications only use two.

The stick-breaking representation

¶ We know that G is discrete. Therefore it can be represented as

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}, \tag{41}$$

where $\theta_k^* \sim G_0$. Intuitively, the proportions of each atom π_k should depend on the scaling parameter α. They do.

¶ Consider an infinite collection of Beta random variables $V_k \sim \text{Beta}(1, \alpha)$. Define

$$\pi_k = V_k \prod_{i=1}^{k-1} (1 - V_i). \tag{42}$$

This is like a unit-length stick where V_1 of the stick is broken off first, then V_2 of the stick is broken off of the rest of it, then V_3 of the stick is broken off of the rest of that, etc. Since the variables are Beta, there will always be a little stick left to break off.

¶ Draw a picture with this table:

Break off	Segment length	What's left
V_1	V_1	$(1 - V_1)$
V_2	$V_2(1 - V_1)$	$(1 - V_2)(1 - V_1)$
V_3	$V_3(1 - V_2)(1 - V_1)$	$(1 - V_3)(1 - V_2)(1 - V_1)$
etc.		

¶ Sethuraman (1994) showed that the $DP(\alpha G_0)$ is equal in distribution to

$$\begin{aligned} V_k &\sim \text{Beta}(1, \alpha) \quad k \in \{1, 2, 3, \ldots\} && (43) \\ \theta_k^* &\sim G_0 \quad k \in \{1, 2, 3, \ldots\} && (44) \\ \pi_k &= V_k \prod_{i=1}^{k-1} (1 - V_i) && (45) \\ G &= \sum \pi_k \delta_{\theta_k^*} && (46) \end{aligned}$$

¶ This is a *constructive definition of the DP*. It doesn't rely on the Kolmogorov consistency theorem and the random measure is not marginalized out.

¶ This is useful for variational inference in DP mixtures.

¶ (Go back to the roadmap and discuss what I didn't cover.)