The Dirichlet Process

COS 424/524, SML 302: Fundamentals of Machine Learning

Professor Engelhardt

COS424/524, SML302

Lecture 20

Unsupervised learning: where we are

We have been discussing unsupervised learning using latent variable models and dimension reduction.

One problem that repeatedly comes up is how to pick the size K of the lower dimensional subspace for

- clustering: number of clusters
- matrix factorization: number of factors
- latent Dirichlet allocation: number of topics

Today we will discuss the Dirichlet process, which has the effect of letting K be a random variable that grows with respect to the data.

Bayesian nonparametrics

Dirichlet processes (DPs) are a class of Bayesian nonparametric distributions.

Nonparametric (in the Bayesian context) means that the number of parameters grows with the number of data points n.

Nonparametric unfortunately refers to classes of models that have an infinite dimensional parameter space in the prior.

These models only use a finite number of parameters to model a finite number of samples; the number of parameters grows with the data.

(In parametric mixture models, the number of parameters remains constant; the number of latent variables grows with the data.)

Bayesian nonparametrics

What does it mean to have an infinite dimensional parameter space?

The number of model parameters grows with the data n.

- in density estimation: the PDF supports the set of all densities
- in regression: the PDF supports the set of all continuous functions on the real line

Bayesian nonparametrics: two examples

In this lecture and the next, we will learn about two of these Bayesian nonparametric distributions

- Dirichlet process (clustering): in clustering, adapts the number of clusters to the data
- Gaussian process (regression): covariate structure grows with the sample size

Why Bayesian nonparametrics?

One theme of this course is that, as data analysts, we want to select and adapt our model to data to avoid over- or under-fitting the data.

- Clustering: setting the number of clusters
- Hidden Markov models: selecting the number of states
- Factor model: selecting the number of factors
- Sparse regression: selecting the number of included predictors
- Nonlinear regression: selecting the complexity of the function

Bayesian nonparametrics formalizes this process using explicit distributions.

Nonparametrics methods

We have already seen a number of nonparametric methods in this class

- Support vector machines: with Gaussian kernel, Gram matrix—and, by the representer theorem, the complexity of the decision boundary—grows with the number of samples
- K-nearest neighbors: complexity of the space grows with the samples
- Kernel density estimation: estimate a density by summing over a small Gaussian distribution centered at each sample

Today we are going to discuss **Bayesian nonparametric models**, and the Dirichlet process in particular.

Dirichlet process: motivation

Applications of DP

- Email clustering: sometimes a type of email comes in that the spam filter has not seen before (e.g., Twitter notices, library events);
- Scientific publications: sometimes a "new" scientific sub discipline will arise (e.g., LDA; SVM, deep learning)
- Collaborative filtering: in recommendation systems, occasionally a new subpopulation of users will join (e.g., Facebook in Brazil, Quentin Tarantino fans)
- Astrophysics: we want to cluster each galaxy by its velocity, assuming a small number of velocities and Gaussian noise.
- Genomics: we want to find the set of ancestral populations for a collection of genomic samples.

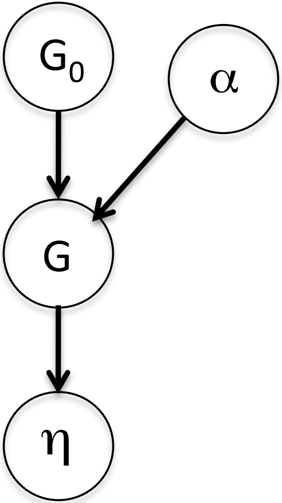
Dirichlet process (DP)

The Dirichlet process is a distribution on the data partition, where the number of partitions is unknown a priori (and, in the prior, infinite).

Several models we have seen where we will benefit from having unknown number of latent components are:

- Clustering
- Latent factor models
- Latent Dirichlet allocation (LDA)

The Dirichlet process



The Dirichlet process is a distribution on distributions.

Let *base distribution* G_0 be a probability measure on a probability space.

Motivated by the example of the Gaussian mixture model, we will choose G_0 to be a Gaussian.

Let *concentration parameter* α be a nonnegative real number.

Dirichlet process, formally

We say that a distribution G is distributed according to a Dirichlet process whose parameters are the base distribution G_0 and the *concentration parameter* or scale α.

Given any partition of the probability space $B_1, B_2, ..., B_K$, we define the prior, for continuous variable η:

$$(G(\eta \in B_1), G(\eta \in B_2), \ldots, G(\eta \in B_K)) \\ \sim Dir(\alpha G_0(B_1), \alpha G_0(B_2), \ldots, \alpha G_0(B_K)).$$

- $(G(B_1), G(B_2), ..., G(B_K))$ is a vector whose entries are each greater than 0 and sum to 1
- each entry $G(B_k)$ represents the probability of partition B_k

In clustering, each partition will correspond to a specific cluster mean, and the proportion of samples in that cluster is $G(\eta \in B_k)$.

Dirichlet process, generative process

The posterior distribution of a DP has the following property. After the first sample η_1 is drawn we have:

$$G \mid \eta_1, \alpha, G_0 \sim DP(\alpha, G_0 + \delta_{\eta_1}),$$

where $\delta(\cdot)$ is the dirac delta function. Rewritten with respect to the Dirichlet distribution:

$$\begin{aligned}&(G(B_1), G(B_2), ..., G(B_k))\\ &\qquad \sim Dir(\alpha \cdot G_0(B_1), \alpha \cdot G_0(B_2), ..., \alpha \cdot G_0(B_i) + 1, ..., \alpha \cdot G_0(B_k))\end{aligned}$$

where sample η_1 represents partition B_i.

Dirichlet process, generative process

We draw the $(n+1)st$ sample as :

$$G \mid \eta_{1:n}, \alpha, G_0 \sim Dir(\alpha \cdot G_0(B_1) + n_1, \alpha \cdot G_0(B_2) + n_2, \ldots, \alpha \cdot G_0(B_K) + n_K)$$

where n_i is the number of samples representing partition B_i, and $n_1 + ...n_K = n$.

We can write this as:

$$G \mid \eta_{1:n}, \alpha, G_0 \sim DP(\alpha, G_0 + \sum_{i=1}^{n} \delta_{\eta_i})$$

The sample obtained in the $(n+1)$st draw, η_{n+1}, is either one of the previous η_i values or it is drawn from G_0.

The probability of drawing η_i representing partition k will grow as more samples are drawn from that partition.

Dirichlet process, generative model

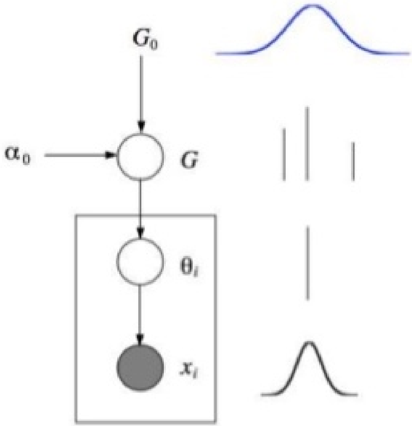
The Dirichlet process is generated as:

- draw η_1 from G_0
- draw $\eta_2 | \eta_1, G_0$
- ...
- draw $\eta_n | \eta_{1:(n-1)}, G_0$.

I find this representation not all that informative.

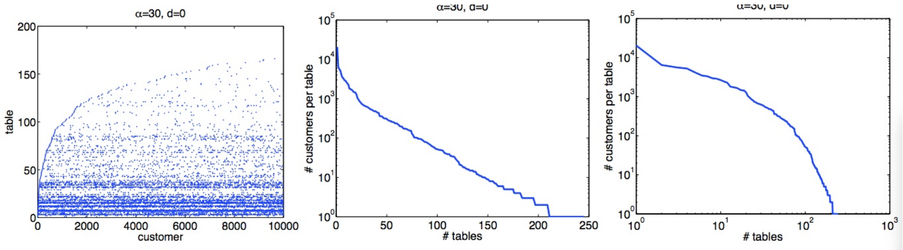
A DP discretizes a continuous distribution

For this representation of a DP, this is the picture I like [Jordan 2005].



DP: how many tables do we find? [YWT 2006]

When we consider drawing samples from a DP, how many tables are there?



Dirichlet process alternative representations

There are two other representations of the Dirichlet process that are more informative, both in terms of intuition and parameter estimation:

- *Chinese restaurant process*: marginal probability of the distribution over the partitions
- *Stick breaking process*: constructive definition of the DP

Chinese Restaurant Process (CRP), intuition

Imagine a Chinese restaurant with an infinite number of tables in a line.

- The first customer sits down at the first table.
- The second customer sits at table 1 with probability $\frac{1}{1+\alpha}$ and table 2 with probability $\frac{\alpha}{1+\alpha}$
- ...
- The $n+1$st customer sits at table k with probability $\frac{n_k}{n+\alpha}$, and an empty table with probability $\frac{\alpha}{n+\alpha}$

Generalization of the Chinese Restaurant Process

Alternatively, for n customers and concentration parameter α:

- $p(n+1\text{st customer sits at an occupied table } k \mid \text{previous } n \text{ customers}) \propto n_k$,
- $p(n+1\text{st customer sits at an unoccupied table } \mid \text{previous } n \text{ customers}) \propto \alpha$,

- the probability of sitting at table k is proportional to the number of people at that table
- the probability of sitting at an unoccupied table is proportional to the concentration parameter α

- The number of occupied tables grows roughly at $O(\log n)$

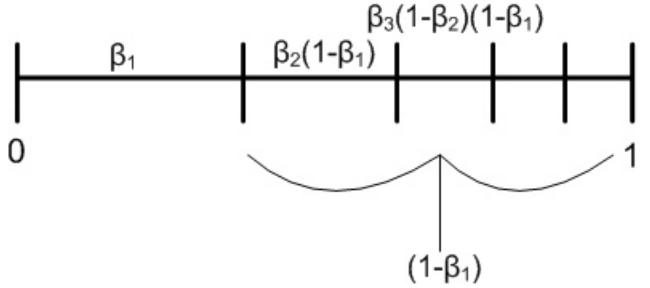
Stick Breaking Process (SBP)

A stick breaking process is a constructive definition of a Dirichlet process.

Start with $\beta_k \sim Beta(1, \alpha)$, where α is our concentration parameter.

Use the independent draws from the beta distribution to partition the $(0, 1)$ line (our *stick*).

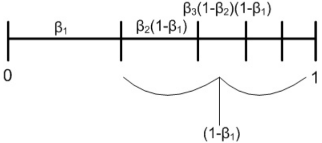
In particular, we have $\pi_1 = \beta_1$ and $\pi_k = \beta_k \prod_{\ell=1}^{k}(1 - \beta_\ell)$ for $k = 2, 3, \ldots$



Stick breaking process

At the kth draw from the stick breaking process,

- the remaining part of the stick is $\prod_{\ell=1}^{K}(1-\beta_\ell)$
- break off β_k proportion of the remaining stick.

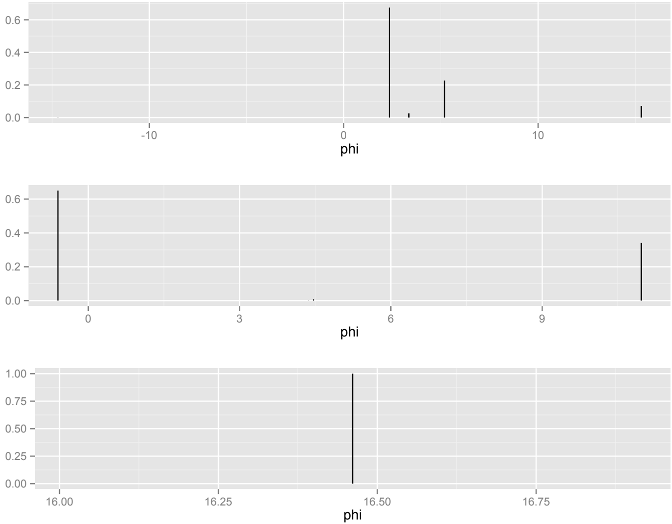


Since $\beta_1 + \beta_1^c = 1$, we know that $\sum_{k=1}^{\infty} \pi_k = 1$.

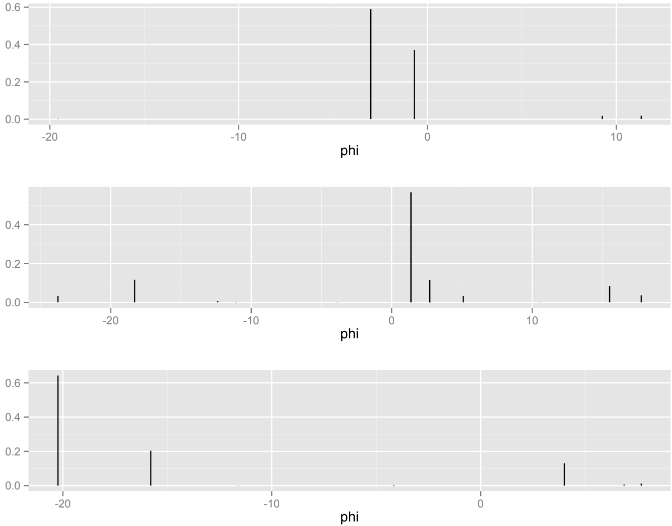
Randomly draw $\eta_k \sim G_0$ and assign to kth stick partition. This constructively defines the DP:

$$
\begin{aligned}
G &\sim DP(\alpha, G_0) \\
\eta_i &\sim G
\end{aligned}
$$

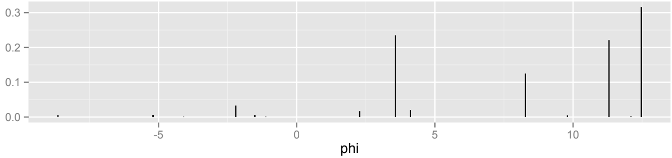
Samples from the stick breaking process $\alpha = 0.5$

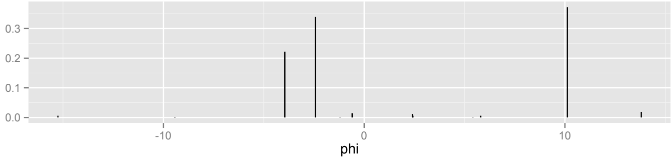


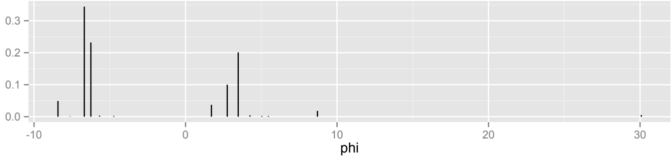
Samples from the stick breaking process $\alpha = 1$



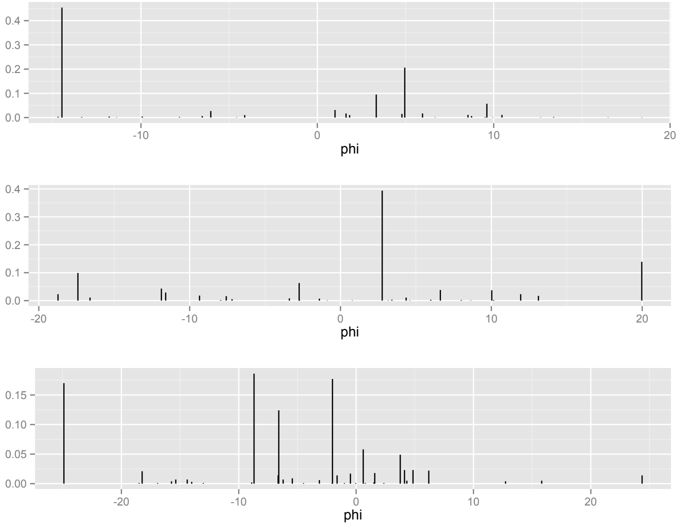
Samples from the stick breaking process $\alpha = 2$



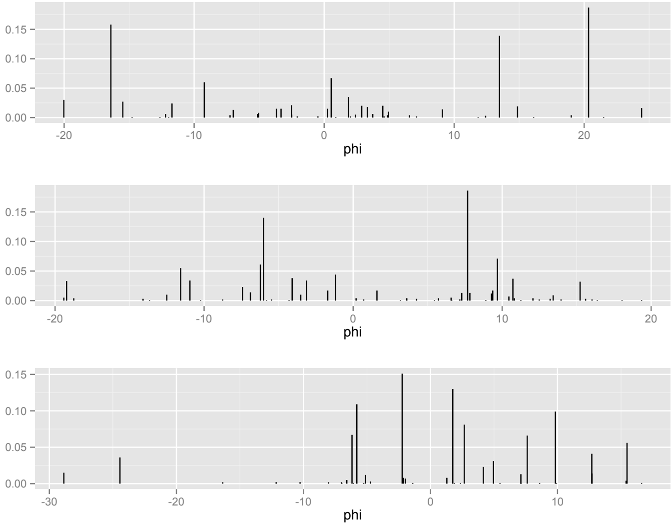




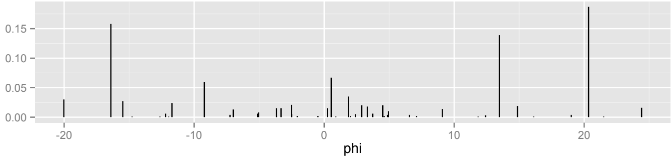
Samples from the stick breaking process $\alpha = 5$

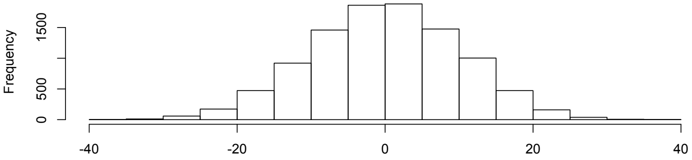


Samples from the stick breaking process $\alpha = 10$



Samples from SBP versus base distribution G_0





Consider the interpretation of a DP as a formal way to discretize a continuous distribution.

How is the density of the base distribution reflected in a DP sample?

Dirichlet process mixture model

Let's now show how we can use a DP to define an infinite Gaussian mixture model.

Finite mixture models define a density function of the form:

$$p(x) = \prod_{k=1}^{K} \pi_k p(x|\theta_k),$$

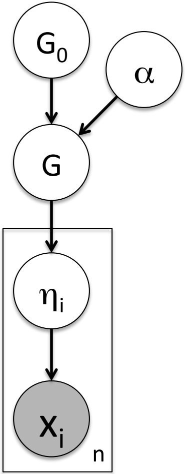
where π_k are mixing proportions and θ_k are parameters for component k.

We can write the density as an integral:

$$p(x) = \int_{\theta} p(x \mid \theta) G(\theta) d\theta,$$

where $G = \sum_{k=1}^{K} \pi_k \delta(\theta_k)$ is a discrete mixing distribution.

Dirichlet process mixture model

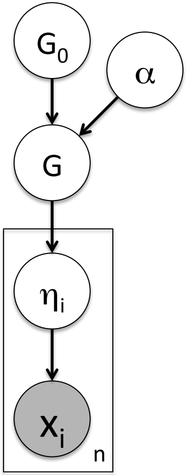


DP mixtures instead use infinite discrete mixing distributions:

$$G = \sum_{k=1}^{\infty} \pi_k \delta(\theta_k)$$

This gives rise to mixture models with an infinite possible number of components

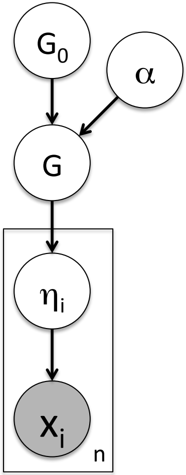
Dirichlet process mixture model



We need to specify a prior over the mixing distribution G

When we use a Dirichlet process (DP), the resulting mixture model is called a DP mixture model

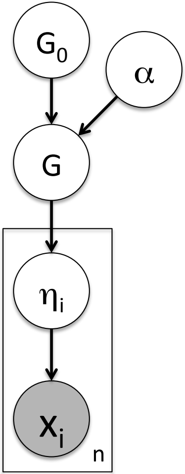
Dirichlet process mixture model



For finite samples, only a finite (but varying) number of components will be used to model the data: each data item is associated with exactly one component but each component can be associated with multiple data items.

Model fitting in a DPMM estimates both the number of components to use and the parameters of those components.

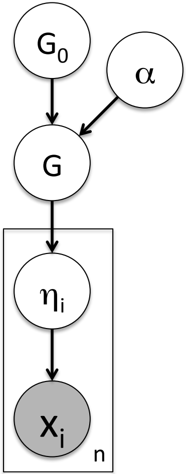
Dirichlet process mixture model: generative model



The generative model for a Dirichlet process Gaussian mixture model:

$$
\begin{aligned}
G &\sim DP(\alpha, G_0) \\
\eta_i &\sim G \\
x_i &\sim p(x_i \mid \eta_i) = \mathcal{N}(x_i | \mu_i = \eta_i)
\end{aligned}
$$

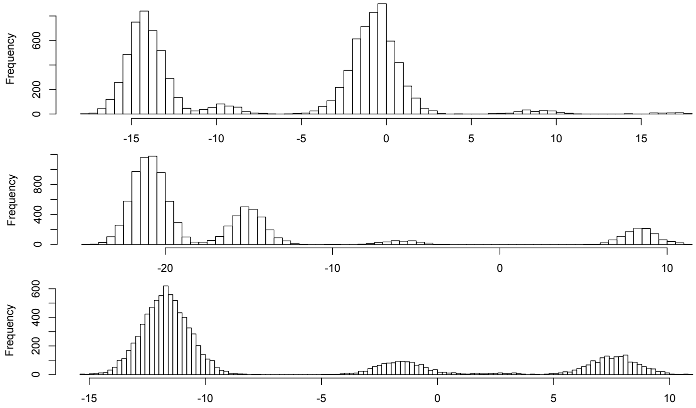
DPMM: stick breaking representation



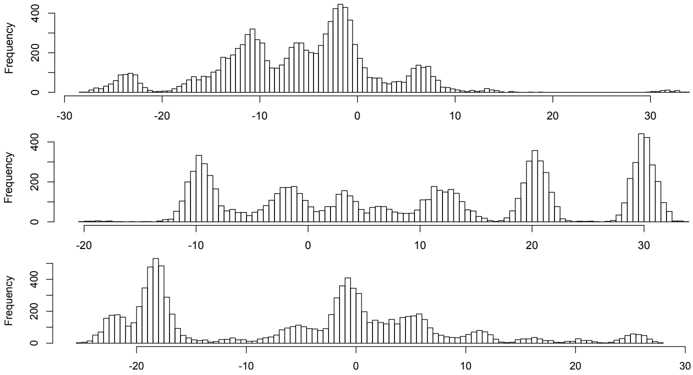
The stick breaking representation for a Dirichlet process Gaussian mixture model:

$$
\begin{aligned}
\beta_k &\sim Beta(1, \alpha) \\
\pi_k &= \beta_k \prod_{\ell=1}^{K} (1 - \beta_\ell) \\
\eta_k &\sim G_0 \\
z_i &\sim Mult(\pi) \\
x_i &\sim \mathcal{N}(\eta_{z_i}).
\end{aligned}
$$

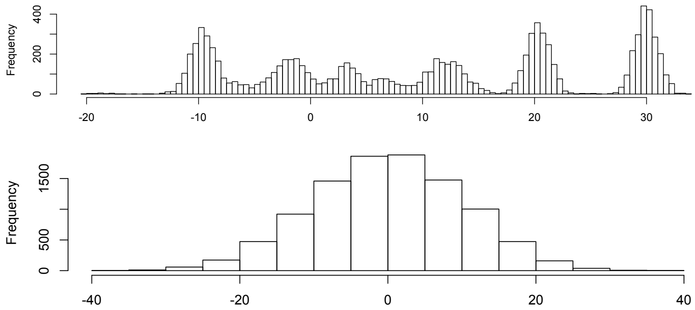
Samples from the DPGMM $\alpha = 1$



Samples from the DPGMM $\alpha = 10$

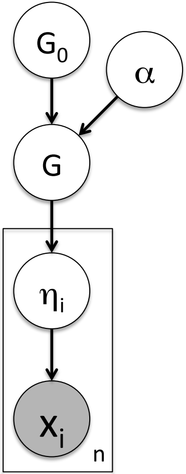


Samples from DPGMM versus base distribution G_0



How is the density of the base distribution reflected in a DPMM sample?

Dirichlet process mixture model for text

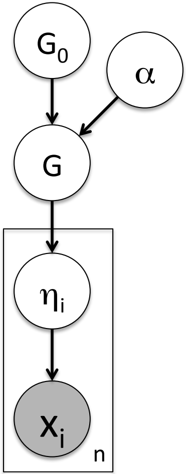


This is a generic mixture model now, and we are not constrained by Gaussian distributions of our mixture components.

What if our observations x_i are now are bag-of-words representation of a document i? What distribution is $p(x_i \mid \eta_i)$?

$$
\begin{aligned}
G &\sim DP(\alpha, G_0) \\
\eta_i &\sim G \\
x_i &\sim p(x_i \mid \eta_i)
\end{aligned}
$$

Dirichlet process mixture model for text

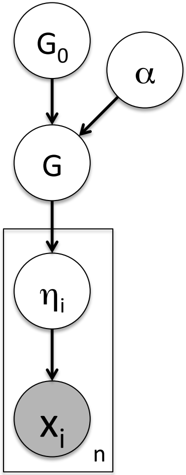


Let's model the bag-of-words for document i as a draw from a multinomial distribution.

What should our base distribution G_0 be to make this model as simple as possible?

$$
\begin{aligned}
G &\sim DP(\alpha, G_0) \\
\eta_i &\sim G \\
x_i &\sim Mult(x_i \mid \eta_i)
\end{aligned}
$$

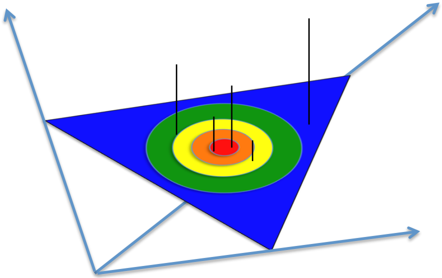
Dirichlet process mixture model for text



The conjugate prior for a multinomial is a Dirichlet.

Here, G_0 is a Dirichlet distribution on the V-dimensional simplex where V is the size of the vocabulary.

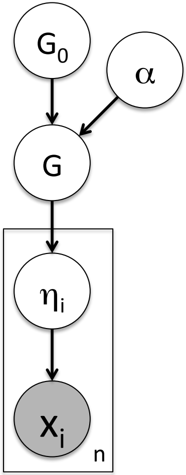
Dirichlet process mixture model for text



When G_0 is a Dirichlet distribution on the V-dimensional simplex, then $G \sim DP(\alpha, G_0)$ is a discretized distribution on the V dimensional simplex.

For visualization purposes, $V = 3$.

Dirichlet process mixture model for text

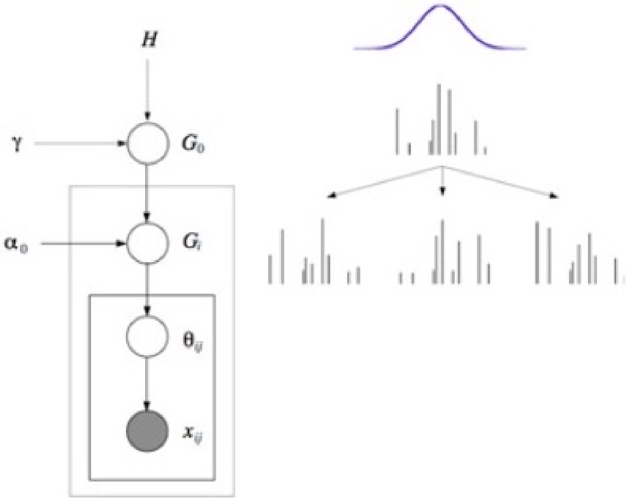


This model has the feel of a topic model with an infinite number of possible topics.

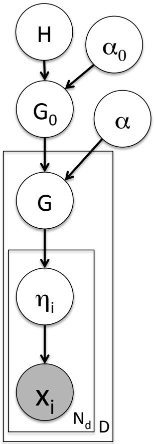
But it is not quite right. Why is this model not an appropriate model for topics in a collection of documents?

Hierarchical Dirichlet process mixture model [Jordan 2005]

Returning to the Gaussian base distribution for clarity:



Hierarchical Dirichlet process mixture model for text



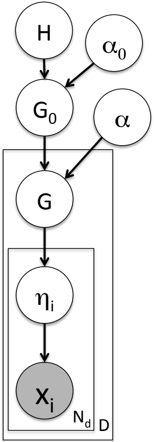
We will let the base distribution be a Dirichlet process with a Dirichlet base distribution H.

Now G_0 discretizes the continuous Dirichlet distribution, allowing documents to share specific topics (where topic is a distribution on words, or a point on the simplex)

Then G specifies the set of topics and the topic proportions for a specific document.

And η_i selects a specific topic and corresponding word distribution for word x_i.

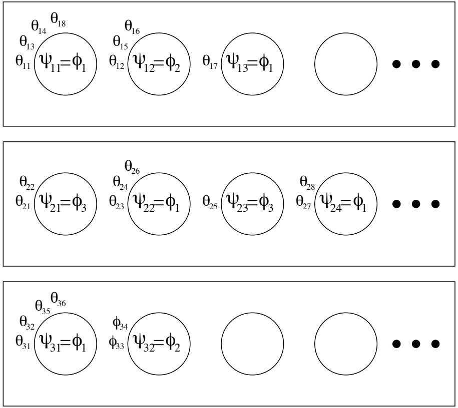
Hierarchical Dirichlet process mixture model for text



The HDP model is specified as follows:

$$
\begin{aligned}
G_0 &\sim DP(\alpha_0, H) \\
G_i &\sim DP(\alpha, G_0) \\
\eta_i &\sim G_i \\
x_i &\sim Mult(x_i \mid \eta_i)
\end{aligned}
$$

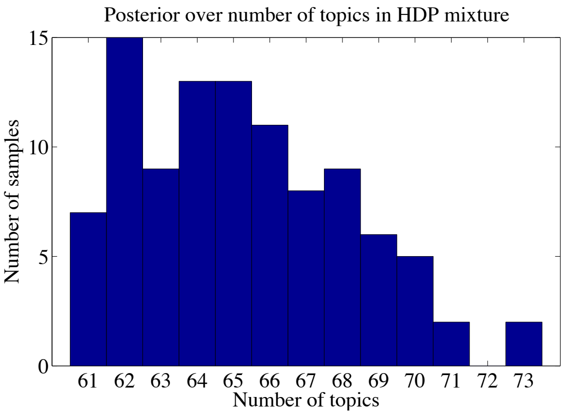
Hierarchical Dirichlet process: Chinese restaurant franchise



The restaurant metaphor used to explain the HDP is the "Chinese restaurant franchise"

Figure from [Teh et al. 2006]

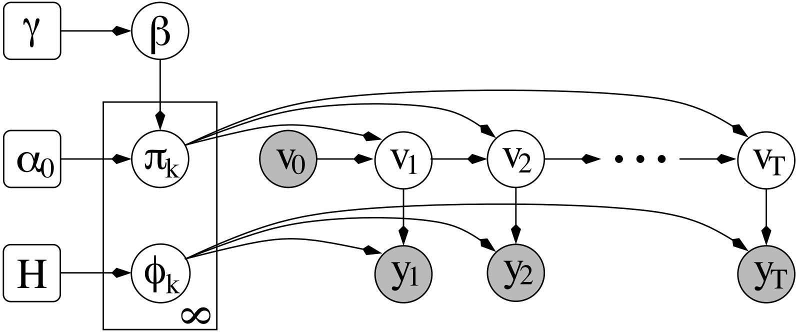
Hierarchical Dirichlet process: posterior over topics



Uses the corpus of nematode biology abstracts, fitting an HDP.

Figure from [Teh et al. 2006]

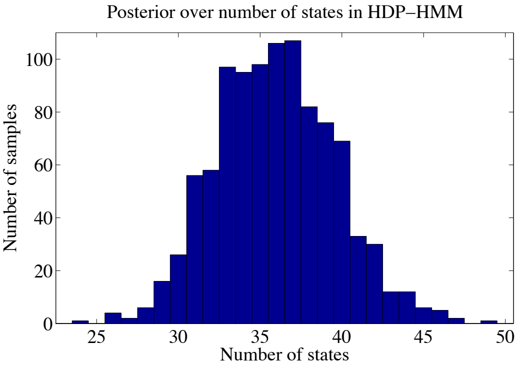
HDP-HMM



HDP-HMM models sequential data with possibly infinite number of latent states.

Figure from [Teh et al. 2006]

HDP-HMM: posterior distribution on latent states



HDP-HMM to predict next character string in *Alice in Wonderland*; posterior distribution over number of latent states.

Figure from [Teh et al. 2006]

How to estimate parameters in DP models?

As with LDA, EM is difficult in (infinite) latent variable models

- MCMC: Gibbs sampling, collapsed Gibbs sampling
- Variational approaches: mean field, collapsed variational
- Stochastic variational inference (Hoffman et al. 2013)
- variational approaches often faster
- sampling approaches give you an estimate of the full posterior distribution (which may or may not be interpretable!), including the number of latent clusters

DP assumptions and cautions

- The assumptions and cautions are identical for all of the mixture models and topic models we have discussed
- Additional caution 1: the parameter estimates may not be robust to α setting (might want to estimate this parameter too)
- Additional caution 2: avoid interpreting the estimated number of components K as *truth*. It is a draw from an (often very flat) posterior distribution.

Extensions to the Dirichlet process

- Dirichlet process regression
- Dirichlet process generalized linear models
- Dirichlet process factor analysis
- Spatial models with Dirichlet processes
- Network analysis and stochastic block models
- Anywhere a latent variable model exists

History of the Dirichlet process

- Polya Urn scheme (Blackwell & MacQueen 1973)
- DP mixture model (Antoniak 1974)
- Stick breaking process (Sethuraman 1994)
- MCMC sampling for DP mixtures (Escobar & West 1994)
- Connections between DPs and other distributions on partitions (Pitman 2001 summer school notes)
- Hierarchical Dirichlet process (Teh et al. 2006)

Additional Resources

- MLAPA: Chapter 25
- (reading) Orbanz & Teh 2010. *Bayesian Nonparametric Models*
- (reading) Rasmussen 1999. *The Infinite Gaussian Mixture Model*
- (video) Michael Jordan *Dirichlet Processes, Chinese Restaurant Processes and All That*
- (video) Yee Whye Teh *Dirichlet Processes: Tutorial and Practical Course*
- (video) Tom Griffiths – Inferring Structure from Data
- Metacademy; *Dirichlet Process*
- Metacademy: *Chinese Restaurant Process*