Dr. Nonparametric Bayes

Or: How I Learned to Stop Worrying
and Love the Dirichlet Process

Kurt Miller
CS 294: Practical Machine Learning
November 19, 2009

Today we will discuss Nonparametric Bayesian methods.

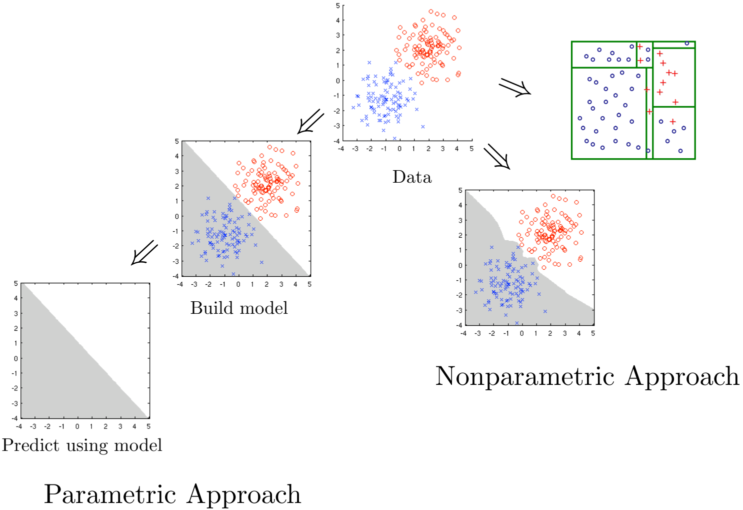
Today we will discuss Nonparametric Bayesian methods.

"Nonparametric Bayesian methods"?
What does that mean?

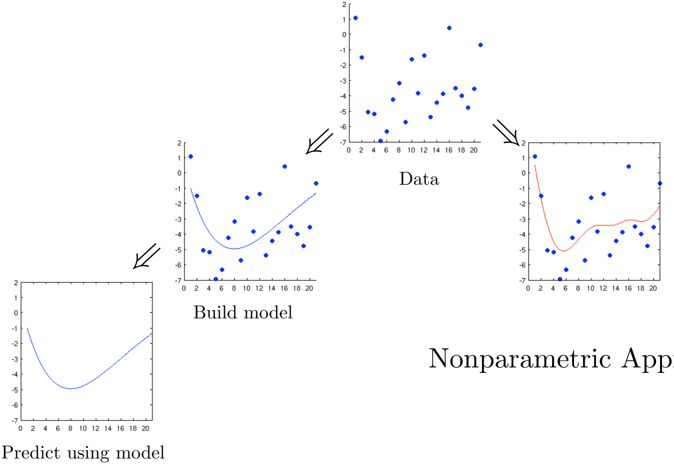
Nonparametric

Nonparametric: Does NOT mean there are no parameters.

Example: Classification



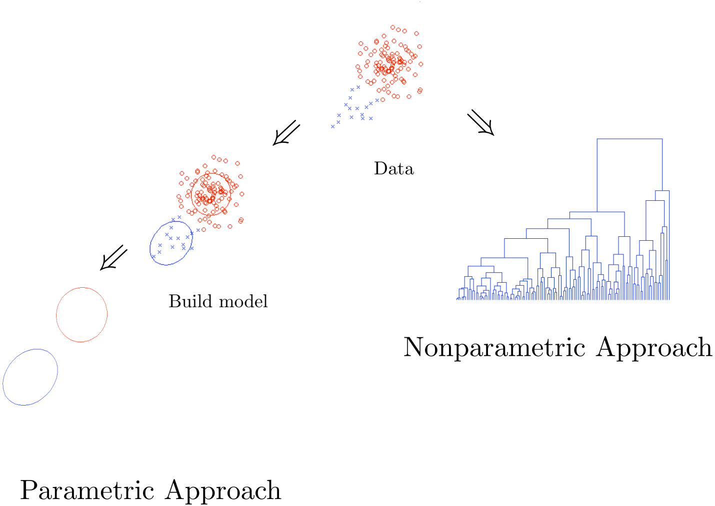
Example: Regression



Nonparametric Approach

Parametric Approach

Example: Clustering



So now we know what nonparametric means, but what does Bayesian mean?

Statistics: Bayesian Basics (Slide from tutorial lecture)

- The Bayesian approach treats statistical problems by maintaining probability distributions over possible parameter values.
- That is, we treat the parameters themselves as random variables having distributions:
  1. We have some beliefs about our parameter values θ before we see any data. These beliefs are encoded in the *prior distribution* $P(\theta)$.
  2. Treating the parameters θ as random variables, we can write the likelihood of the data X as a conditional probability: $P(X|\theta)$.
  3. We would like to update our beliefs about θ based on the data by obtaining $P(\theta|X)$, the *posterior distribution*. *Solution*: by Bayes' theorem,

     $$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

     where

     $$P(X) = \int P(X|\theta)P(\theta)d\theta$$

Why Be Bayesian?

You can take a course on this question.

Why Be Bayesian?

You can take a course on this question. One answer:

Infinite Exchangeability: $\forall n \; p(x_1, \ldots, x_n) = p(x_{\sigma(1)}, \ldots, x_{\sigma(n)})$

Why Be Bayesian?

You can take a course on this question. One answer:

Infinite Exchangeability: $\forall n \; p(x_1, \ldots, x_n) = p(x_{\sigma(1)}, \ldots, x_{\sigma(n)})$

De Finetti's Theorem (1955): If $(x_1, x_2, \ldots)$ are *infinitely exchangeable*, then $\forall n$

$$p(x_1, \ldots, x_n) = \int \left(\prod_{i=1}^{n} p(x_i|\theta) \right) dP(\theta)$$

for some random variable θ.

Simple Example

Task: Toss a (potentially biased) coin N times. Compute θ, the probability of heads.

Suppose we observe: {T, H, H, T}. What do we think θ is?

Simple Example

Task: Toss a (potentially biased) coin N times. Compute θ, the probability of heads.

Suppose we observe: $\{$T, H, H, T$\}$. What do we think θ is? The maximum likelihood estimate is $\theta = 1/2$. Seems reasonable.

Simple Example

Task: Toss a (potentially biased) coin N times. Compute θ, the probability of heads.

Suppose we observe: $\{$T, H, H, T$\}$. What do we think θ is? The maximum likelihood estimate is $\theta = 1/2$. Seems reasonable.

Now suppose we observe: $\{$H, H, H, H$\}$. What do we think θ is?

Simple Example

Task: Toss a (potentially biased) coin N times. Compute θ, the probability of heads.

Suppose we observe: $\{$T, H, H, T$\}$. What do we think θ is? The maximum likelihood estimate is $\theta = 1/2$. Seems reasonable.

Now suppose we observe: $\{$H, H, H, H$\}$. What do we think θ is? The maximum likelihood estimate is $\theta = 1$. Seem reasonable?

Simple Example

Task: Toss a (potentially biased) coin N times. Compute θ, the probability of heads.

Suppose we observe: $\{$T, H, H, T$\}$. What do we think θ is? The maximum likelihood estimate is $\theta = 1/2$. Seems reasonable.

Now suppose we observe: $\{$H, H, H, H$\}$. What do we think θ is? The maximum likelihood estimate is $\theta = 1$. Seem reasonable?

Not really. Why?

Simple Example

When we observe $\{$H, H, H, H$\}$, why does $\theta = 1$ seem unreasonable?

Simple Example

When we observe {H, H, H, H}, why does $\theta = 1$ seem unreasonable?

Prior knowledge! We believe coins generally have $\theta \approx 1/2$. How to encode this? By using a Beta *prior on* θ.

Bayesian Approach to Estimating θ

Place a Beta(a, b) prior on θ. This prior has the form

$$p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}.$$

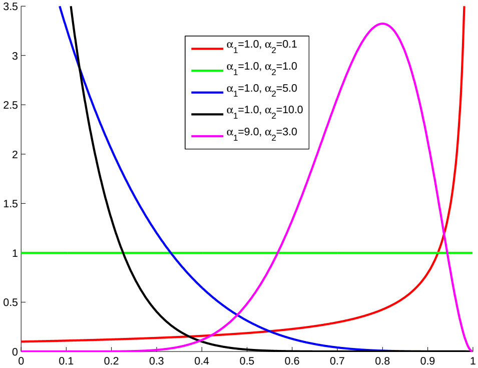
What does this distribution look like?

Bayesian Approach to Estimating θ

Place a Beta(a, b) prior on θ. This prior has the form

$$p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}.$$

What does this distribution look like?



Bayesian Approach to Estimating θ

After observing X, a sequence with n heads and m tails, the posterior on θ is:

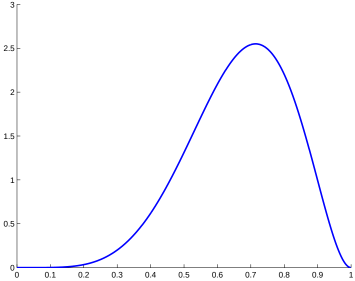
$$\begin{aligned} p(\theta|X) &\propto p(X|\theta)p(\theta) \\ &\propto \theta^{a+n-1}(1-\theta)^{b+m-1} \\ &\sim \mathrm{Beta}(a+n, b+m). \end{aligned}$$

Bayesian Approach to Estimating θ

After observing X, a sequence with n heads and m tails, the posterior on θ is:

$$\begin{aligned} p(\theta|X) &\propto p(X|\theta)p(\theta) \\ &\propto \theta^{a+n-1}(1-\theta)^{b+m-1} \\ &\sim \mathrm{Beta}(a+n, b+m). \end{aligned}$$

If $a = b = 1$ and we observe 5 heads and 2 tails, Beta$(6, 3)$ looks like



Nonparametric Bayesian Methods

Now we know what nonparametric and Bayesian mean. What should we expect from nonparametric Bayesian methods?

Nonparametric Bayesian Methods

Now we know what nonparametric and Bayesian mean. What should we expect from nonparametric Bayesian methods?

- Complexity of our model should be allowed to grow as we get more data.

Nonparametric Bayesian Methods

Now we know what nonparametric and Bayesian mean. What should we expect from nonparametric Bayesian methods?

- Complexity of our model should be allowed to grow as we get more data.
- Place a prior on an unbounded number of parameters.

Nonparametric Bayesian Methods overview

- Dirichlet Process/Chinese Restaurant Process
  Latent class models - often used in the clustering context
- Beta Process/Indian Buffet Process
  Latent feature models
- Gaussian Process (No culinary metaphor - oh well)
  Regression

Today we focus on the Dirichlet Process!

Today's topic: The Dirichlet Process

A nonparametric approach to clustering. It can be used in *any* probabilistic model for clustering.

Before diving into the details, we first introduce several key ideas.

Key ideas to be discussed today

- A parametric Bayesian approach to clustering
  - Defining the model
  - Markov Chain Monte Carlo (MCMC) inference
- A nonparametric approach to clustering
  - Defining the model - The Dirichlet Process!
  - MCMC inference
- Extensions

Key ideas to be discussed today

- A parametric Bayesian approach to clustering
    - Defining the model
    - Markov Chain Monte Carlo (MCMC) inference
- A nonparametric approach to clustering
    - Defining the model - The Dirichlet Process!
    - MCMC inference
- Extensions

A Bayesian Approach to Clustering

We must specify two things:

- The likelihood term (how data is affected by the parameters):

$$p(X|\theta)$$

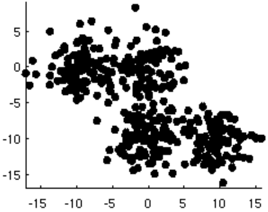
- The prior (the prior distirubution on the parameters):

$$p(\theta)$$

We will slowly develop what these are in the Bayesian clustering context.

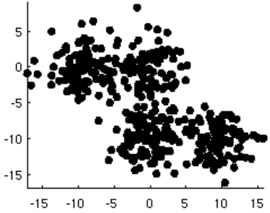
Motivating example: Clustering

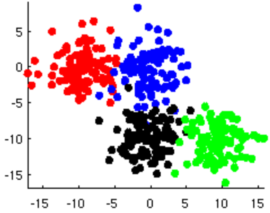
How many clusters?

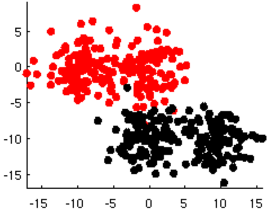


Motivating example: Clustering

How many clusters?







Clustering – A Parametric Approach

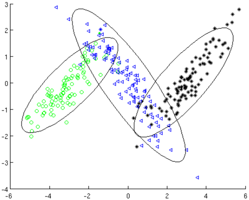
Frequentist approach: Gaussian Mixture Models with K mixtures

Distribution over classes: $\pi = (\pi_1, \ldots, \pi_K)$

Each cluster has a mean and covariance: $\phi_i = (\mu_i, \Sigma_i)$

Then

$$p(x|\pi, \phi) = \sum_{k=1}^{K} \pi_k p(x|\phi_k)$$



Use Expectation Maximization (EM) to maximize the likelihood of the data with respect to (π, ϕ).

Clustering – A Parametric Approach

Frequentist approach: Gaussian Mixture Models with K mixtures

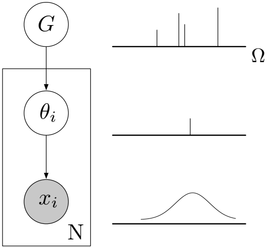
Alternate definition:

$$G = \sum_{k=1}^{K} \pi_k \delta_{\phi_k}$$

where δ_{ϕ_k} is an *atom* at ϕ_k.

Then

$$\begin{aligned} \theta_i &\sim G \\ x_i &\sim p(x|\theta_i) \end{aligned}$$



Clustering – A Parametric Approach

Bayesian approach: Bayesian Gaussian Mixture Models with K mixtures

Distribution over classes: $\pi = (\pi_1, \ldots, \pi_K)$

$$\pi \sim \mathrm{Dirichlet}(\alpha/K, \ldots, \alpha/K)$$

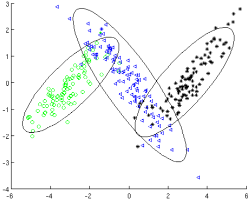
(We'll review the Dirichlet Distribution in a several slides.)

Each cluster has a mean and covariance: $\phi_k = (\mu_k, \Sigma_k)$

$$(\mu_k, \Sigma_k) \sim \text{Normal-Inverse-Wishart}(\nu)$$

We still have

$$p(x|\pi, \phi) = \sum_{k=1}^{K} \pi_k p(x|\phi_k)$$

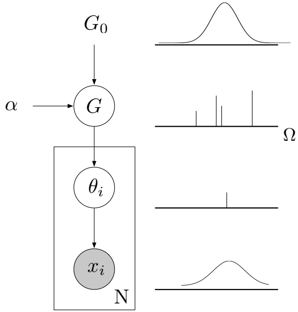


Clustering – A Parametric Approach

Bayesian approach: Bayesian Gaussian Mixture Models with K mixtures

G is now a *random* measure.

$$
\begin{aligned}
\phi_k &\sim G_0 \\
\pi &\sim \mathrm{Dirichlet}(\alpha/K, \ldots, \alpha/K) \\
G &= \sum_{i=1}^{K} \pi_k \delta_{\phi_k} \\
\theta_i &\sim G \\
x_i &\sim p(x|\theta_i)
\end{aligned}
$$



The Dirichlet Distribution

We had

$$\pi \sim \mathrm{Dirichlet}(\alpha_1, \ldots, \alpha_K)$$

The Dirichlet density is defined as

$$p(\pi|\alpha) = \frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\prod_{k=1}^K \Gamma(\alpha_k)} \pi_1^{\alpha_1 - 1} \pi_2^{\alpha_2 - 1} \cdots \pi_K^{\alpha_K - 1}$$

where $\pi_K = 1 - \sum_{k=1}^{K-1} \pi_k$.

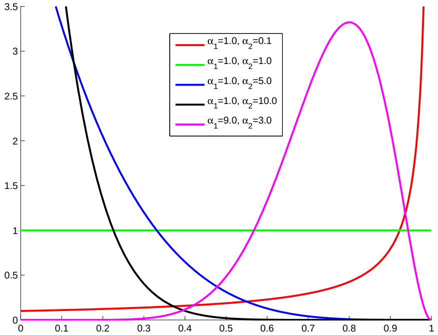
The expectations of π are

$$E(\pi_i) = \frac{\alpha_i}{\sum_{i=1}^K \alpha_i}$$

The Beta Distribution

A special case of the Dirichlet distribution is the Beta distribution for when $K = 2$.

$$p(\pi|\alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \pi^{\alpha_1 - 1}(1 - \pi)^{\alpha_2 - 1}$$

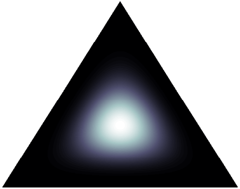


The Dirichlet Distribution

In three dimensions:

$$p(\pi|\alpha_1, \alpha_2, \alpha_3) = \frac{\Gamma(\alpha_1 + \alpha_2 + \alpha_3)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)} \pi_1^{\alpha_1 - 1} \pi_2^{\alpha_2 - 1} (1 - \pi_1 - \pi_2)^{\alpha_3 - 1}$$

$\alpha = (2, 2, 2)$



$\alpha = (5, 5, 5)$

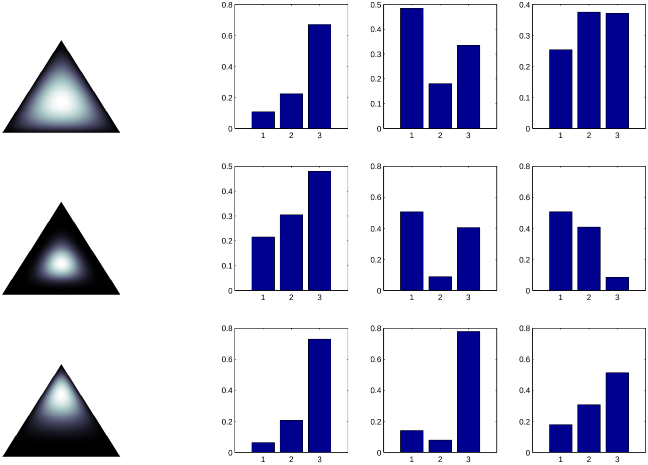
$\alpha = (2, 2, 25)$

Draws from the Dirichlet Distribution

$\alpha = (2, 2, 2)$

$\alpha = (5, 5, 5)$

$\alpha = (2, 2, 5)$



Key Property of the Dirichlet Distribution

The Aggregation Property: If

$$(\pi_1, \ldots, \pi_i, \pi_{i+1}, \ldots, \pi_K) \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_i, \alpha_{i+1}, \ldots, \alpha_K)$$

then

$$(\pi_1, \ldots, \pi_i + \pi_{i+1}, \ldots, \pi_K) \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_i + \alpha_{i+1}, \ldots, \alpha_K)$$

Key Property of the Dirichlet Distribution

The Aggregation Property: If

$$(\pi_1, \ldots, \pi_i, \pi_{i+1}, \ldots, \pi_K) \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_i, \alpha_{i+1}, \ldots, \alpha_K)$$

then

$$(\pi_1, \ldots, \pi_i + \pi_{i+1}, \ldots, \pi_K) \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_i + \alpha_{i+1}, \ldots, \alpha_K)$$

This is also valid for any aggregation:

$$\left(\pi_1 + \pi_2, \sum_{k=3^K} \pi_k\right) \sim \mathrm{Beta}\left(\alpha_1 + \alpha_2, \sum_{k=3}^{K} \alpha_k\right)$$

Multinomial-Dirichlet Conjugacy

Let $Z \sim \text{Multinomial}(\pi)$ and $\pi \sim \text{Dir}(\alpha)$.

Posterior:

$$\begin{aligned} p(\pi|z) &\propto p(z|\pi)p(\pi) \\ &= (\pi_1^{z_1} \cdots \pi_K^{z_K})(\pi_1^{\alpha_1 - 1} \cdots \pi_K^{\alpha_K - 1}) \\ &= (\pi_1^{z_1 + \alpha_1 - 1} \cdots \pi_K^{z_K + \alpha_K - 1}) \end{aligned}$$

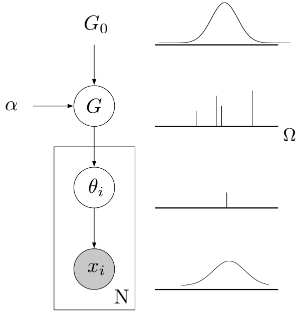
which is $\text{Dir}(\alpha + z)$.

Clustering – A Parametric Approach

Bayesian approach: Bayesian Gaussian Mixture Models with K mixtures

G is now a *random* measure.

$$
\begin{aligned}
\phi_k &\sim G_0 \\
\pi &\sim \mathrm{Dirichlet}(\alpha/K, \ldots, \alpha/K) \\
G &= \sum_{i=1}^{K} \pi_k \delta_{\phi_k} \\
\theta_i &\sim G \\
x_i &\sim p(x|\theta_i)
\end{aligned}
$$



Bayesian Mixture Models

We no longer want just the maximum likelihood parameters, we want the full posterior:

$$p(\pi, \phi | X) \propto p(X | \pi, \phi) p(\pi, \phi)$$

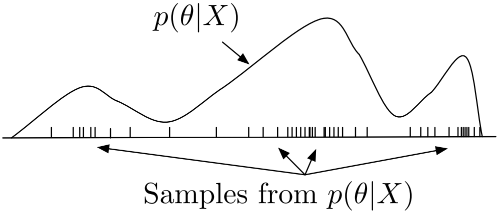
Unfortunately, this is not analytically tractable.

Two main approaches to approximate inference:

- Markov Chain Monte Carlo (MCMC) methods
- Variational approximations

Monte Carlo Methods

Suppose we wish to reason about $p(\theta|X)$, but we cannot compute this distribution exactly. If instead, we can sample $\theta \sim p(\theta|X)$, what can we do?



This is the idea behind *Monte Carlo* methods.

Markov Chain Monte Carlo (MCMC)

We do not have access to an oracle that will give use samples $\theta \sim p(\theta|X)$. How do we get these samples?

Markov Chain Monte Carlo (MCMC) methods have been developed to solve this problem.

We focus on *Gibbs sampling*, a special case of the *Metropolis-Hastings algorithm*.

Gibbs sampling

An MCMC technique

Assume θ consists of several parameters $\theta = (\theta_1, \ldots, \theta_m)$. In the finite mixture model, $\theta = (\pi, \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K)$.

Then do

- Initialize $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_m^{(0)})$ at time step 0.
- For $t = 1, 2, \ldots$, draw $\theta^{(t)}$ given $\theta^{(t-1)}$ in such a way that eventually $\theta^{(t)}$ are samples from $p(\theta|X)$.

Gibbs sampling

An MCMC technique

In Gibbs sampling, we only need to be able to sample

$$\theta_i^{(t)} \sim p(\theta_i | \theta_1^{(t)}, \ldots, \theta_{i-1}^{(t)}, \theta_{i+1}^{(t-1)}, \ldots, \theta_m^{(t-1)}, X).$$

If we repeat this for any model we discuss today, theory tells us that eventually we get samples $\theta^{(t)}$ from $p(\theta|X)$.

Gibbs sampling

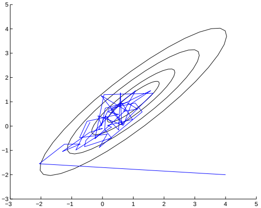
An MCMC technique

In Gibbs sampling, we only need to be able to sample

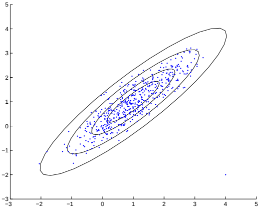
$$\theta_i^{(t)} \sim p(\theta_i|\theta_1^{(t)},\ldots,\theta_{i-1}^{(t)},\theta_{i+1}^{(t-1)},\ldots,\theta_m^{(t-1)},X).$$

If we repeat this for any model we discuss today, theory tells us that eventually we get samples $\theta^{(t)}$ from $p(\theta|X)$.

Example: $\theta = (\theta_1, \theta_2)$ and $p(\theta) \sim \mathcal{N}(\mu, \Sigma)$.



First 50 samples

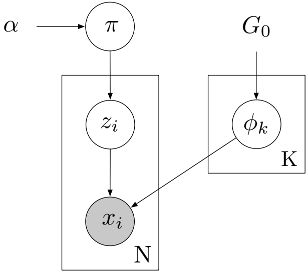


First 500 samples

Bayesian Mixture Models - MCMC inference

Introduce "membership" indicators z_i where $z_i \sim \mathrm{Multinomial}(\pi)$ indicates which cluster the i^{th} data point belongs to.

$$p(\pi, Z, \phi | X) \propto p(X | Z, \phi) p(Z | \pi) p(\pi, \phi)$$



Gibbs sampling for the Bayesian Mixture Model

Randomly initialize Z, π, ϕ. Repeat until we have enough samples:

1. Sample each z_i from

$$z_i | Z_{-i}, \pi, \phi, X \propto \sum_{k=1}^{K} \pi_k p(x_i|\phi_k) \mathbb{1}_{\{z_i = k\}}$$

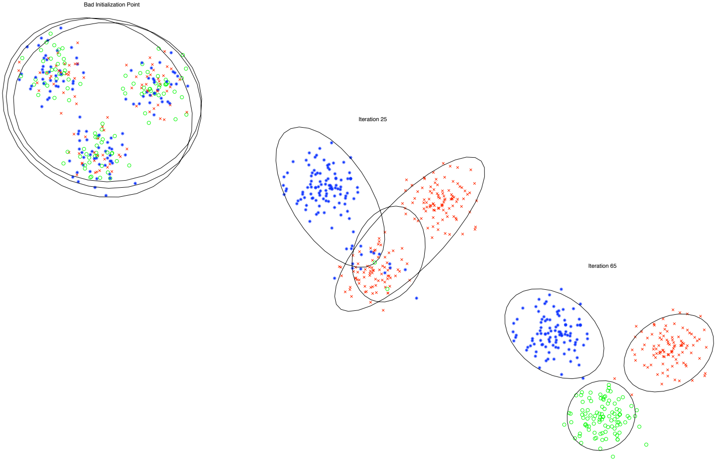
2. Sample each π from

$$\pi | Z, \phi, X \sim \mathrm{Dir}(n_1 + \alpha/K, \ldots, n_K + \alpha/K)$$

   where n_i is the number of points assigned to cluster i.

3. Sample each ϕ_k from the NIW posterior based on Z and X.

MCMC in Action



[Matlab demo]

Collapsed Gibbs Sampler

Idea for an improvement: we can marginalize out some variables due to conjugacy, so do not need to sample it. This is called a *collapsed sampler*. Here marginalize out π.

Randomly initialize Z, ϕ. Repeat:

1. Sample each z_i from

$$z_i | Z_{-i}, \phi, X \propto \sum_{k=1}^{K} (n_k + \alpha/K) p(x_i|\phi_k) \mathbb{1}_{\{z_i = k\}}$$

2. Sample each ϕ_k from the NIW posterior based on Z and X.

Note about the likelihood term

For easy visualization, we used a Gaussian mixture model.

You should use the appropriate likelihood model for your application!

Summary: Parametric Bayesian clustering

- First specify the likelihood - application specific.
- Next specify a prior on all parameters.
- Exact posterior inference is intractable. Can use a Gibbs sampler for approximate inference.

5 minute break

How to Choose K?

Generic model selection: cross-validation, AIC, BIC, MDL, etc.

Can place of parametric prior on K.

How to Choose K?

Generic model selection: cross-validation, AIC, BIC, MDL, etc.

Can place of parametric prior on K.

What if we just let $K \to \infty$ in our parametric model?

Thought Experiment

Let $K \to \infty$.

$$\begin{aligned}
\phi_k &\sim G_0 \\
\pi &\sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K) \\
G &= \sum_{i=1}^{K} \pi_k \delta_{\phi_k} \\
\theta_i &\sim G \\
x_i &\sim p(x|\theta_i)
\end{aligned}$$

Thought Experiment: Collapsed Gibbs Sampler

Randomly initialize Z, ϕ. Repeat:

1. Sample each z_i from

$$
\begin{aligned}
z_i | Z_{-i}, \phi, X \quad &\propto \quad \sum_{k=1}^{K} (n_k + \alpha/K) p(x_i|\phi_k) \mathbb{1}_{\{z_i = k\}} \\
&\to \quad \sum_{k=1}^{K} n_k p(x_i|\phi_k) \mathbb{1}_{\{z_i = k\}}
\end{aligned}
$$

   Note that $n_k = 0$ for empty clusters.

2. Sample each ϕ_k based on Z and X.

Thought Experiment: Collapsed Gibbs Sampler

What about empty clusters? Lump all empty clusters together. Let K^+ be the number of occupied clusters. Then the posterior probability of sitting at *any* empty cluster is:

$$\begin{aligned} z_i | Z_{-i}, \phi, X &\propto \alpha/K \times (K - K^+) f(x_i | G_0) \\ &\rightarrow \alpha f(x_i | G_0) \end{aligned}$$

for $f(x_i|G_0) = \int p(x|\phi) dG_0(\phi)$.

Key ideas to be discussed today

- A parametric Bayesian approach to clustering
  - Defining the model
  - Markov Chain Monte Carlo (MCMC) inference
- A nonparametric approach to clustering
  - Defining the model - The Dirichlet Process!
  - MCMC inference
- Extensions

A Nonparametric Bayesian Approach to Clustering

We must again specify two things:

- The likelihood term (how data is affected by the parameters):

$$p(X|\theta)$$

  Identical to the parametric case.

- The prior (the prior distirubution on the parameters):

$$p(\theta)$$

  The Dirichlet Process!

Exact posterior inference is still intractable. But we have already derived the Gibbs update equations!

What is the Dirichlet Process?

Image from http://www.nature.com/nsmb/journal/v7/n6/fig_tab/nsb0600_443_F1.html

What is the Dirichlet Process?

The Dirichlet Process

A flexible, nonparametric prior over an infinite number of clusters/classes as well as the parameters for those classes.

Parameters for the Dirichlet Process

- α - The concentration parameter.
- G_0 - The base measure. A prior distribution for the cluster specific parameters.

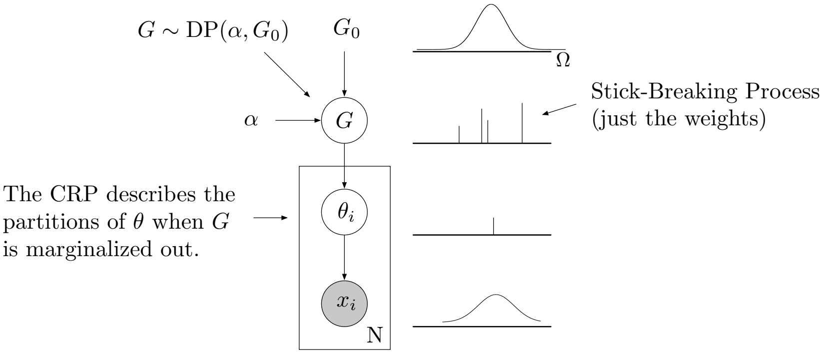
The Dirichlet Process (DP) is a *distribution over distributions*. We write

$$G \sim DP(\alpha, G_0)$$

to indicate G is a distribution drawn from the DP.

It will become clearer in a bit what α and G_0 are.

The DP, CRP, and Stick-Breaking Process

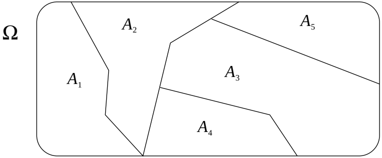


The Dirichlet Process

Definition: Let G_0 be a probability measure on the measurable space (Ω, B) and $\alpha \in \mathbb{R}^+$.

The *Dirichlet Process* $DP(\alpha, G_0)$ is the distribution on probability measures G such that for any finite partition $(A_1, \ldots, A_m)$ of Ω,

$$(G(A_1), \ldots, G(A_m)) \sim \mathrm{Dir}(\alpha G_0(A_1), \ldots, \alpha G_0(A_m)).$$



(Ferguson, '73)

Mathematical Properties of the Dirichlet Process

Suppose we sample

- $G \sim DP(\alpha, G_0)$
- $\theta_1 \sim G$

What is the posterior distribution of G given θ_1?

Mathematical Properties of the Dirichlet Process

Suppose we sample

- $G \sim DP(\alpha, G_0)$
- $\theta_1 \sim G$

What is the posterior distribution of G given θ_1?

$$G|\theta_1 \sim \mathrm{DP}\left(\alpha+1, \frac{\alpha}{\alpha+1}G_0 + \frac{1}{\alpha+1}\delta_{\theta_1}\right)$$

More generally

$$G|\theta_1,\ldots,\theta_n \sim \mathrm{DP}\left(\alpha+n, \frac{\alpha}{\alpha+n}G_0 + \frac{1}{\alpha+n}\sum_{i=1}^{n}\delta_{\theta_i}\right)$$

Mathematical Properties of the Dirichlet Process

With probability 1, a sample $G \sim DP(\alpha, G_0)$ is of the form

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$$

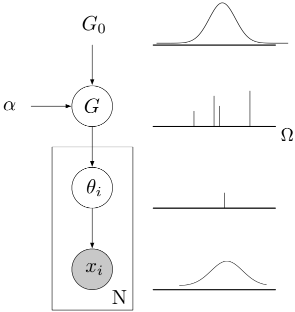
(Sethuraman, '94)

The Dirichlet Process and Clustering

Draw $G \sim DP(\alpha, G_0)$ to get

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$$

Use this in a mixture model:



The Stick-Breaking Process

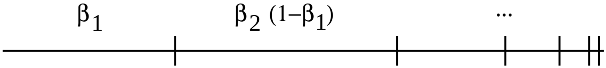
- Define an infinite sequence of Beta random variables:

$$\beta_k \sim \mathrm{Beta}(1, \alpha) \qquad\qquad k = 1, 2, \ldots$$

- And then define an infinite sequence of mixing proportions as:

$$\begin{aligned} \pi_1 &= \beta_1 \\ \pi_k &= \beta_k \prod_{l=1}^{k-1} (1 - \beta_l) \qquad\qquad k = 2, 3, \ldots \end{aligned}$$

- This can be viewed as breaking off portions of a stick:



- When π are drawn this way, we can write $\pi \sim \mathrm{GEM}(\alpha)$.

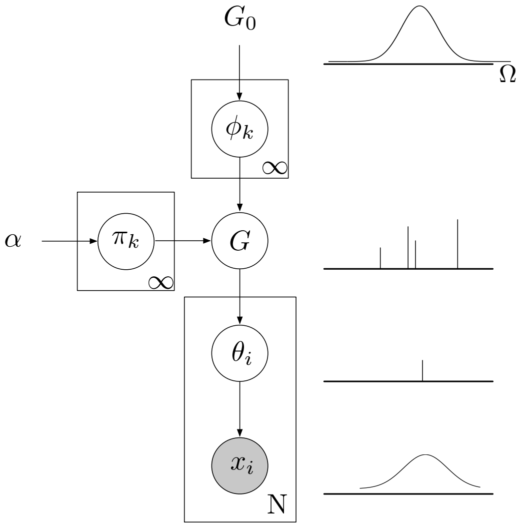
The Stick-Breaking Process

- We now have an explicit formula for each π_k:
  $\pi_k = \beta_k \prod_{l=1}^{k-1} (1-\beta_l)$
- We can also easily see that $\sum_{k=1}^{\infty} \pi_k = 1 \;\; (\mathrm{wp1})$:

$$\begin{aligned} 1 - \sum_{k=1}^{K} \pi_k &= 1 - \beta_1 - \beta_2(1-\beta_1) - \beta_3(1-\beta_1)(1-\beta_2) - \cdots \\ &= (1-\beta_1)(1-\beta_2-\beta_3(1-\beta_2)-\cdots) \\ &= \prod_{k=1}^{K}(1-\beta_k) \\ &\rightarrow 0 \qquad\qquad (\mathrm{wp1} \text{ as } K \rightarrow \infty) \end{aligned}$$

- So now $G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$ has a clean definition as a random measure

The Stick-Breaking Process

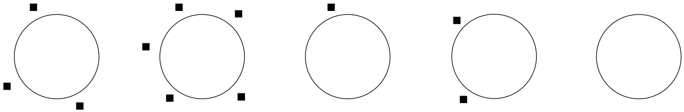


The Chinese Restaurant Process (CRP)

- A random process in which n customers sit down in a Chinese restaurant with an infinite number of tables
  - first customer sits at the first table
  - mth subsequent customer sits at a table drawn from the following distribution:

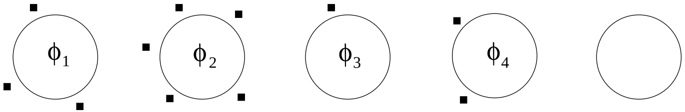
$$\begin{aligned} P(\text{previously occupied table } i|\mathcal{F}_{m-1}) &\propto n_i \\ P(\text{the next unoccupied table}|\mathcal{F}_{m-1}) &\propto \alpha \end{aligned}$$

    where n_i is the number of customers currently at table i and where $\mathcal{F}_{m-1}$ denotes the state of the restaurant after $m-1$ customers have been seated



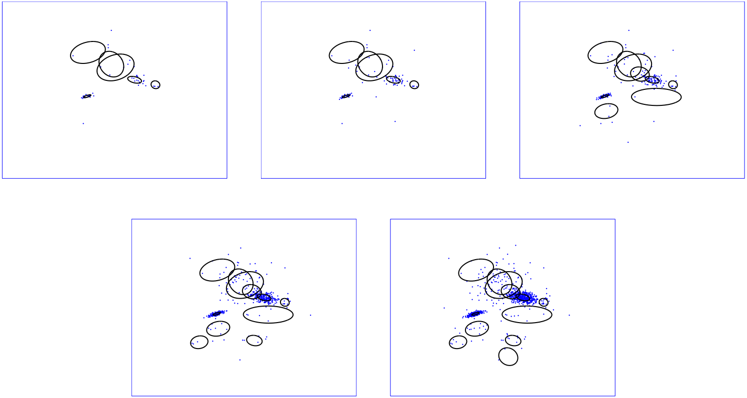
The CRP and Clustering

- Data points are customers; tables are clusters
  - the CRP defines a prior distribution on the partitioning of the data and on the number of tables
- This prior can be completed with:
  - a likelihood—e.g., associate a parameterized probability distribution with each table
  - a prior for the parameters—the first customer to sit at table k chooses the parameter vector for that table (ϕ_k) from the prior



- So we now have a distribution—or can obtain one—for any quantity that we might care about in the clustering setting

The CRP Prior, Gaussian Likelihood, Conjugate Prior



$$\phi_k = (\mu_k, \Sigma_k) \sim N(a, b) \otimes IW(\alpha, \beta)$$

$$x_i \sim N(\phi_k) \qquad \text{for a data point } i \text{ sitting at table } k$$

The CRP and the DP

OK, so we've seen how the CRP relates to clustering. How does it relate to the DP?

The CRP and the DP

OK, so we've seen how the CRP relates to clustering. How does it relate to the DP?

Important fact: The CRP is *exchangeable*.

Remember De Finetti's Theorem: If $(x_1, x_2, \ldots)$ are *infinitely exchangeable*, then $\forall n$

$$p(x_1, \ldots, x_n) = \int \left(\prod_{i=1}^{n} p(x_i | G) \right) dP(G)$$

for some random variable G.

The CRP and the DP

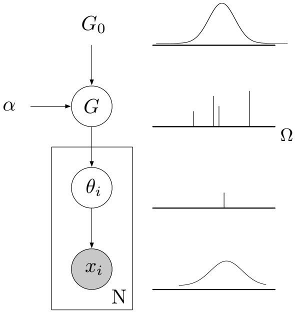
The *Dirichlet Process* is the *De Finetti mixing distribution* for the *CRP*.

The CRP and the DP

The *Dirichlet Process* is the *De Finetti mixing distribution* for the *CRP*.

That means, when we integrate out G, we get the CRP.

$$p(\theta_1, \ldots, \theta_n) = \int \prod_{i=1}^{n} p(\theta_i | G) dP(G)$$

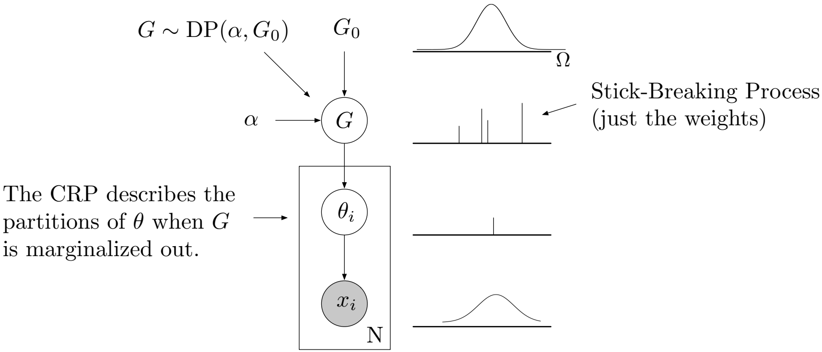


The CRP and the DP

The *Dirichlet Process* is the *De Finetti mixing distribution* for the *CRP*.

In English, this means that if the DP is the prior on G, then the CRP defines how points are assigned to clusters when we integrate out G.

The DP, CRP, and Stick-Breaking Process Summary



Inference for the DP - Gibbs sampler

We introduce the indicators z_i and use the CRP representation.

Randomly initialize Z, ϕ. Repeat:

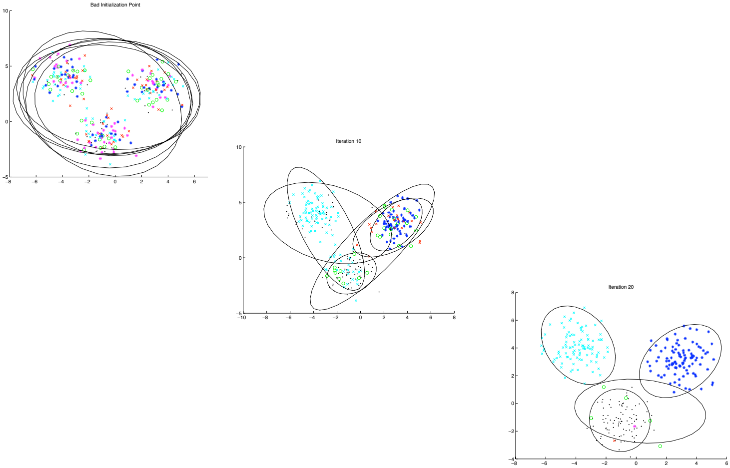
1. Sample each z_i from

$$z_i | Z_{-i}, \phi, X \propto \sum_{k=1}^{K} n_k p(x_i|\phi_k) \mathbb{1}_{\{z_i = k\}} + \alpha f(x_i|G_0) \mathbb{1}_{\{z_i = K+1\}}$$

2. Sample each ϕ_k based on Z and X only for occupied clusters.

This is the sampler we saw earlier, but now with some theoretical basis.

MCMC in Action for the DP



[Matlab demo]

Improvements to the MCMC algorithm

- Collapse out the ϕ_k if conjugate model.
- Split-merge algorithms.

Summary: Nonparametric Bayesian clustering

- First specify the likelihood - application specific.
- Next specify a prior on all parameters - the Dirichlet Process!
- Exact posterior inference is intractable. Can use a Gibbs sampler for approximate inference. This is based on the CRP representation.

Key ideas to be discussed today

- A parametric Bayesian approach to clustering
  - Defining the model
  - Markov Chain Monte Carlo (MCMC) inference
- A nonparametric approach to clustering
  - Defining the model - The Dirichlet Process!
  - MCMC inference
- Extensions

Hierarchical Bayesian Models

Original Bayesian idea

View parameters as random variables - place a prior on them.

Hierarchical Bayesian Models

Original Bayesian idea

View parameters as random variables - place a prior on them.

"Problem"?

Often the priors themselves need parameters.

Hierarchical Bayesian Models

Original Bayesian idea

View parameters as random variables - place a prior on them.

"Problem"?

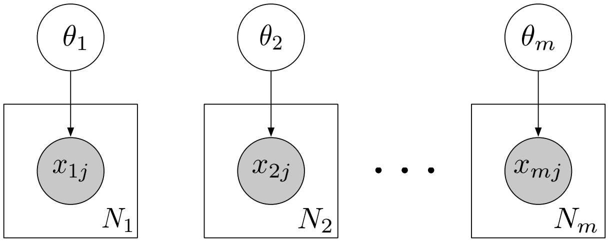
Often the priors themselves need parameters.

Solution

Place a prior on these parameters!

Multiple Learning Problems

Example: $x_i \sim \mathcal{N}(\theta_i, \sigma^2)$ in m different groups.



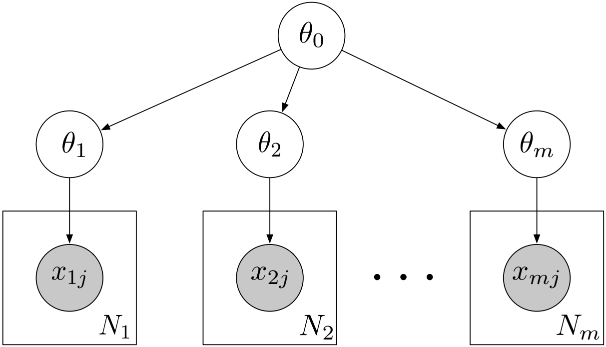
How to estimate θ_i for each group?

Multiple Learning Problems

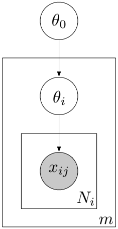
Example: $x_i \sim \mathcal{N}(\theta_i, \sigma^2)$ in m different groups.

Treat θ_is as random variables sampled from a common prior

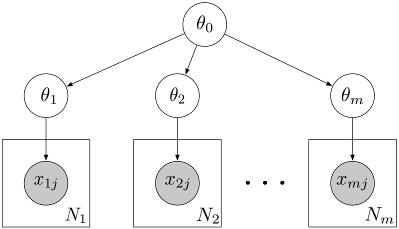
$$\theta_i \sim \mathcal{N}(\theta_0, \sigma_0^2)$$



Recall Plate Notation:

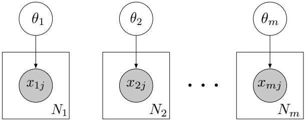


is equivalent to



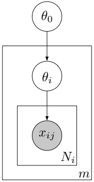
Let's Be Bold!

Independent estimation



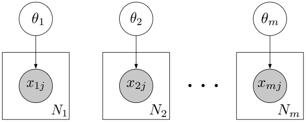
$\Rightarrow$

Hierarchical Bayesian



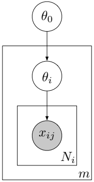
Let's Be Bold!

Independent estimation

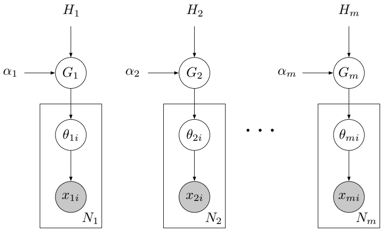


$\Rightarrow$

Hierarchical Bayesian



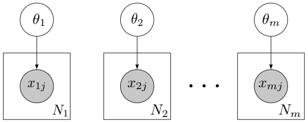
What do we do if we have DPs for multiple related datasets?



$\Rightarrow$

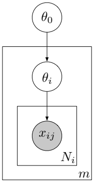
Let's Be Bold!

Independent estimation

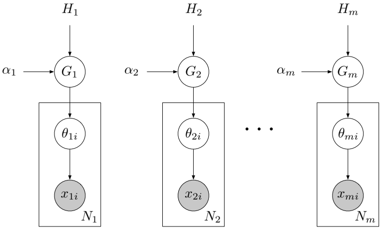


$\Rightarrow$

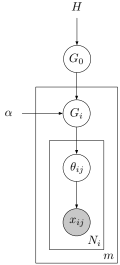
Hierarchical Bayesian



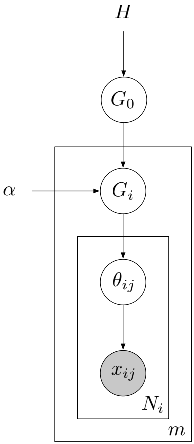
What do we do if we have DPs for multiple related datasets?



$\Rightarrow$



Attempt 1



What kind of distribution do we use for G_0? H?

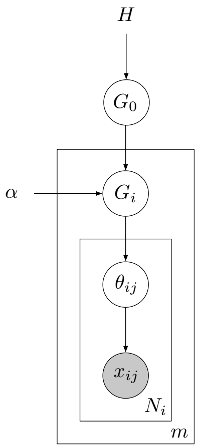
Suppose θ_{ij} are mean parameters for a Gaussian where

$$G_i \sim \mathrm{DP}(\alpha, G_0)$$

and G_0 is a Gaussian with unknown mean?

$$G_0 = \mathcal{N}(\theta_0, \sigma_0^2)$$

Attempt 1



What kind of distribution do we use for G_0? H?

Suppose θ_{ij} are mean parameters for a Gaussian where

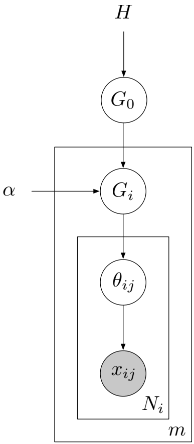
$$G_i \sim \mathrm{DP}(\alpha, G_0)$$

and G_0 is a Gaussian with unknown mean?

$$G_0 = \mathcal{N}(\theta_0, \sigma_0^2)$$

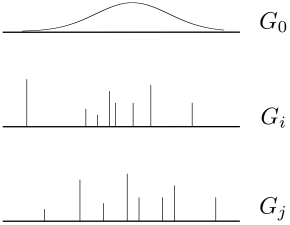
This does NOT work! Why?

Attempt 1

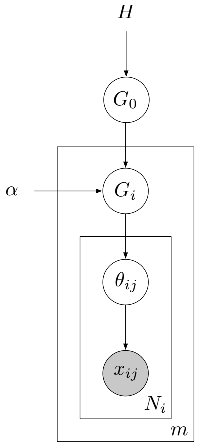


The problem: If G_0 is continuous, then with probability ONE, G_i and G_j will share ZERO atoms.

$\Rightarrow$ This means NO clustering!

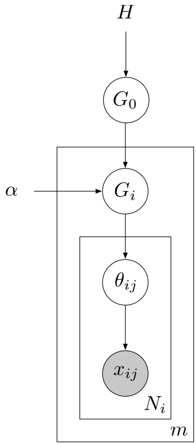


Attempt 2



So G_0 must be discrete. What discrete prior can we use on G_0?

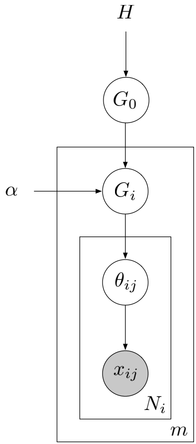
Attempt 2



So G_0 must be discrete. What discrete prior can we use on G_0?

How about a parametric prior?

Attempt 2



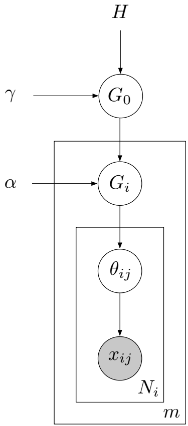
So G_0 must be discrete. What discrete prior can we use on G_0?

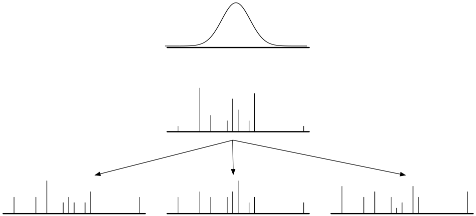
How about a parametric prior?

Gee, if only we had a nonparametric prior on discrete measures...

The Hierarchical Dirichlet Process

Solution:





$$
\begin{aligned}
G_0 &\sim \mathrm{DP}(\gamma, H) \\
G_i &\sim \mathrm{DP}(\alpha, G_0) \\
\theta_{ij} | G_i &\sim G_i \\
x_{ij} | \theta_{ij} &\sim p(x_{ij} | \theta_{ij})
\end{aligned}
$$

(Teh, Jordan, Beal, Blei, 2004)

G_0 vs. G_i

Since

$$\begin{aligned} G_0 &\sim \mathrm{DP}(\gamma, H) \\ G_i &\sim \mathrm{DP}(\alpha, G_0) \end{aligned}$$

we know

$$\begin{aligned} G_0 &= \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k} \\ G_i &= \sum_{k=1}^{\infty} \pi_{ik} \delta_{\phi_k} \end{aligned}$$

G_0 vs. G_i

Since

$$\begin{aligned} G_0 &\sim \mathrm{DP}(\gamma, H) \\ G_i &\sim \mathrm{DP}(\alpha, G_0) \end{aligned}$$

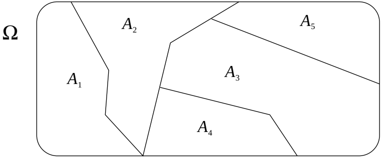
we know

$$\begin{aligned} G_0 &= \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k} \\ G_i &= \sum_{k=1}^{\infty} \pi_{ik} \delta_{\phi_k} \end{aligned}$$

What is the relationship between π_k and π_{ik}?

Relationship between π_k and π_{jk}

Let $(A_1, \ldots, A_m)$ be a partition of Ω.

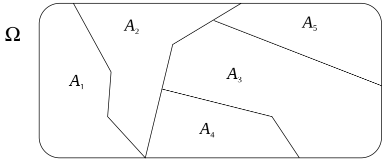


By properties of the DP

$$(G_i(A_1), \ldots, G_i(A_m)) \quad \sim \quad \mathrm{Dir}(\alpha G_0(A_1), \ldots, \alpha G_0(A_m))$$

Relationship between π_k and π_{jk}

Let $(A_1, \ldots, A_m)$ be a partition of Ω.



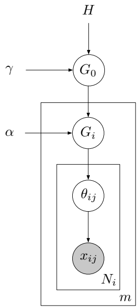
By properties of the DP

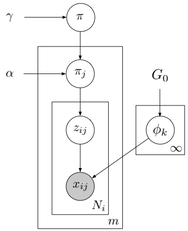
$$\begin{aligned} (G_i(A_1), \ldots, G_i(A_m)) &\sim \mathrm{Dir}(\alpha G_0(A_1), \ldots, \alpha G_0(A_m)) \\ \Rightarrow \left(\sum_{k \in K_1} \pi_{ik}, \ldots, \sum_{k \in K_m} \pi_{ik} \right) &\sim \mathrm{Dir}\left(\alpha \sum_{k \in K_1} \pi_k, \ldots, \alpha \sum_{k \in K_m} \pi_k \right) \end{aligned}$$

Stick-Breaking Construction for the HDP

$$
\begin{aligned}
G_0 &\sim \mathrm{DP}(\gamma, H) \\
G_i &\sim \mathrm{DP}(\alpha, G_0) \\
\theta_{ij} | G_i &\sim G_i \\
x_{ij} | \theta_{ij} &\sim p(x_{ij} | \theta_{ij})
\end{aligned}
$$

$$
\begin{aligned}
\pi &\sim \mathrm{GEM}(\gamma) \\
\pi_i &\sim \mathrm{DP}(\alpha, \pi) \\
\phi_k &\sim H \\
z_{ij} &\sim \pi_i \\
x_{ij} &\sim p(x_{ij} | \phi_{z_{ij}})
\end{aligned}
$$





Stick-Breaking Construction for the HDP

Remember:

$$\left(\sum_{k \in K_1} \pi_{ik}, \ldots, \sum_{k \in K_m} \pi_{ik} \right) \sim \mathrm{Dir}\left(\alpha \sum_{k \in K_1} \pi_k, \ldots, \alpha \sum_{k \in K_m} \pi_k \right)$$

Explicit relationship between π and π_i:

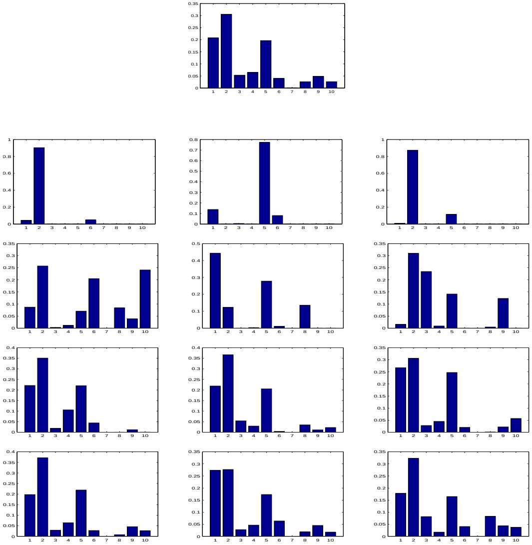
$$\begin{aligned}
\beta_k &\sim \mathrm{Beta}(1, \gamma) \\
\pi_k &= \beta_k \prod_{j=1}^{k-1} (1 - \beta_j) \\
\beta_{ik} &\sim \mathrm{Beta}\left(\alpha \pi_k, \alpha \left(1 - \sum_{j=1}^{k} \pi_j \right) \right) \\
\pi_{ik} &= \beta_{ik} \prod_{j=1}^{k-1} (1 - \beta_{ij})
\end{aligned}$$

The Effect of α

$$\pi \sim \mathrm{GEM}(\gamma),\ \pi_i \sim \mathrm{DP}(\alpha, \pi)$$

π: $\gamma = 2$

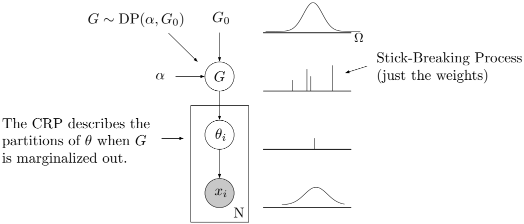
π_i: $\alpha = 1$, $\alpha = 5$, $\alpha = 20$, $\alpha = 100$



The Hierarchical Dirichlet Process

For the DP, we had:

- Mathematical definition
- Stick-breaking construction
- Chinese restaurant process

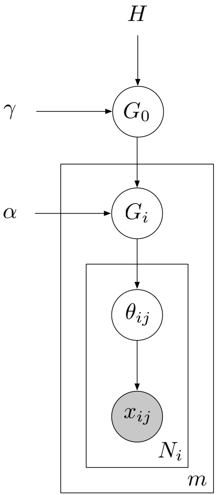


For the HDP, we have

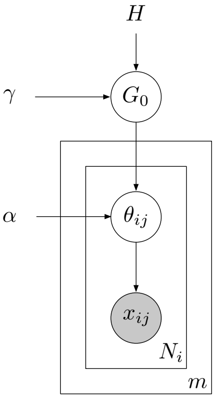
- Mathematical definition
- Stick-breaking construction
- ?

The Chinese Restaurant Franchise (CRF) - Step 1

First integrate out the G_i.



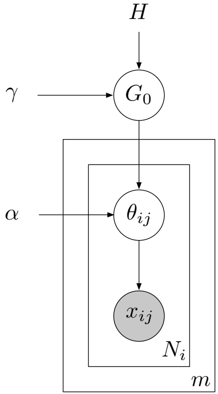
$\Rightarrow$

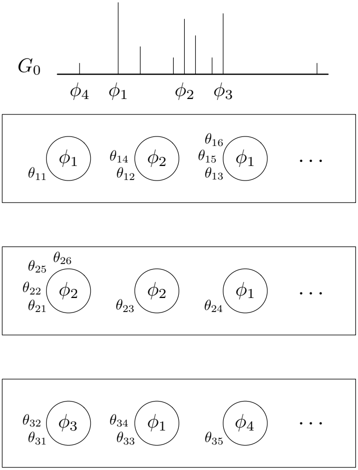


The Chinese Restaurant Franchise (CRF) - Step 1

What is the generative process when we integrate out G_i?

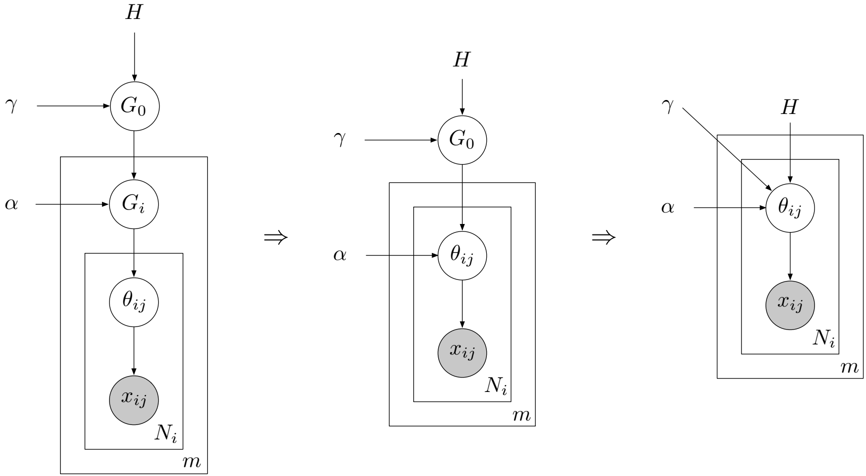
1. Draw global $G_0 = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$.
2. Each group acts like a separate CRP.



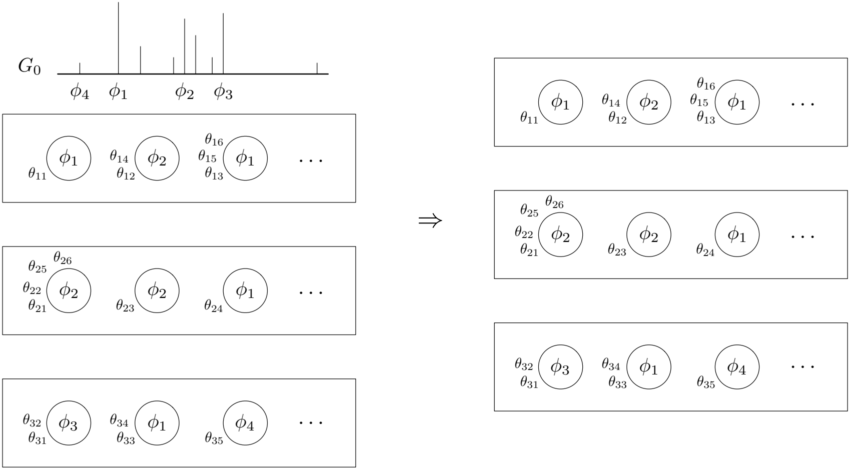


The Chinese Restaurant Franchise (CRF)

First integrate out the G_i, then integrate out G_0



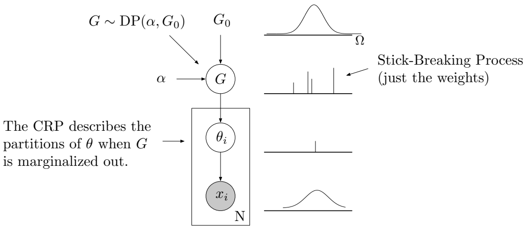
Chinese Restaurant Franchise (CRF)



The Hierarchical Dirichlet Process

For the DP, we had:

- Mathematical definition
- Stick-breaking construction
- Chinese restaurant process



For the HDP, we have

- Mathematical definition
- Stick-breaking construction
- Chinese restaurant franchise process

Inference

Same classes of algorithms used for the DP:

- MCMC
  - CRF representation
  - Stick-breaking representation
- Variational

We will not go into these.

Application of the HDP - Infinite Hidden Markov Model

Finite Hidden Markov Models (HMMs):

- m states $s_1, \ldots, s_m$
- s_i has parameter ϕ_i with emission distribution

$$y \sim p(y|\phi_i)$$

- $m \times m$ transition matrix

	s_1	s_2	$\cdots$	s_m
s_1	π_{11}	π_{12}	$\cdots$	π_{1m}
s_2	π_{21}	π_{22}	$\cdots$	π_{2m}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
s_m	π_{m1}	π_{m2}	$\cdots$	π_{mm}

How do we let $m \to \infty$?

Application of the HDP - Infinite Hidden Markov Model

How do we let $m \to \infty$?

Think a bit outside the traditional clustering context.

Let each state s_i corresponds to a group.

$$
\begin{aligned}
\pi|\gamma &\sim \mathrm{GEM}(\gamma) \\
\pi_i|\alpha, \pi &\sim \mathrm{DP}(\alpha, \pi) \\
\phi_k|H &\sim H \\
x_t|x_{t-1}, (\pi_i)_{i=1}^{\infty} &\sim \pi_{x_{t-1}} \\
y_t|x_t, (\pi_i)_{i=1}^{\infty} &\sim p(y_t|\phi_{x_t})
\end{aligned}
$$

Questions?

Great set of references for the Machine Learning community:
http://npbayes.wikidot.com/references
Includes both the "classics" as well as modern applications.