Dirichlet Processes
A gentle tutorial

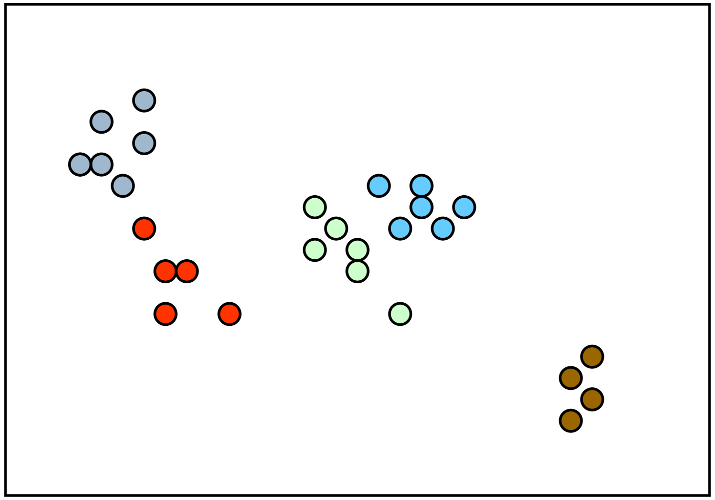
SELECT Lab Meeting

October 14, 2008

Khalid El-Arini

Motivation

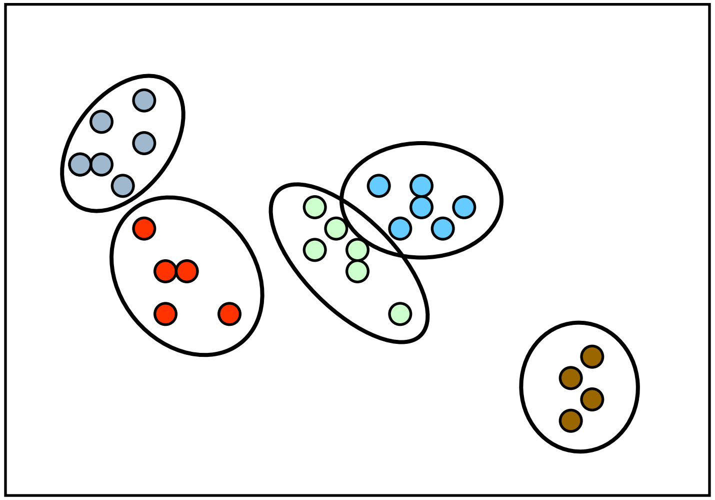
- We are given a data set, and are told that it was generated from a mixture of Gaussian distributions.



- Unfortunately, no one has any idea *how many* Gaussians produced the data.

Motivation

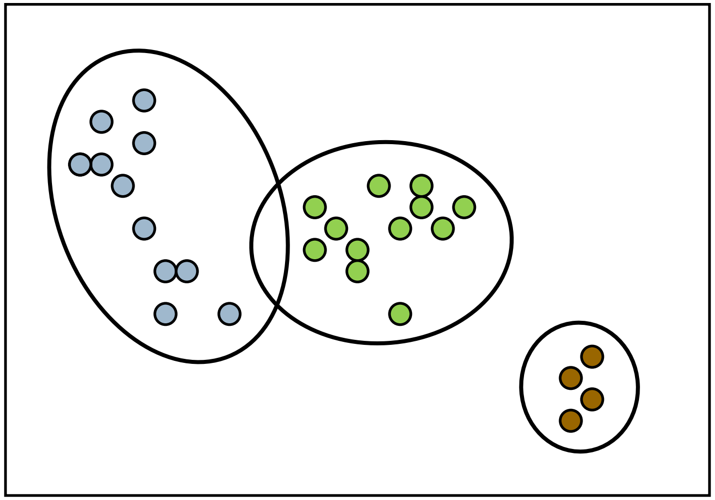
- We are given a data set, and are told that it was generated from a mixture of Gaussian distributions.



- Unfortunately, no one has any idea *how many* Gaussians produced the data.

Motivation

- We are given a data set, and are told that it was generated from a mixture of Gaussian distributions.



- Unfortunately, no one has any idea *how many* Gaussians produced the data.

What to do?

- We can guess the number of clusters, run Expectation Maximization (EM) for Gaussian Mixture Models, look at the results, and then try again…
- We can run hierarchical agglomerative clustering, and cut the tree at a visually appealing level…

- We want to cluster the data in a statistically principled manner, without resorting to hacks.

Other motivating examples

- Brain Imaging: Model an unknown number of spatial activation patterns in fMRI images [Kim and Smyth, NIPS 2006]
- Topic Modeling: Model an unknown number of topics across several corpora of documents [Teh et al. 2006]
- …

Overview

- Dirichlet distribution, and Dirichlet Process introduction
- Dirichlet Processes from different perspectives
  - Samples from a Dirichlet Process
  - Chinese Restaurant Process representation
  - Stick Breaking
  - Formal Definition
- Dirichlet Process Mixtures
- Inference

The Dirichlet Distribution

- Let $\boldsymbol{\Theta} = \{\theta_1, \theta_2, \ldots, \theta_m\}$
- We write:

$$\boldsymbol{\Theta} \sim \mathsf{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_m)$$

$$P(\theta_1, \theta_2, \ldots, \theta_m) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{m} \theta_k^{\alpha_k - 1}$$

- Samples from the distribution lie in the *m-1* dimensional probability simplex

The Dirichlet Distribution

- Let $\boldsymbol{\Theta} = \{\theta_1, \theta_2, \ldots, \theta_m\}$
- We write:

$$\boldsymbol{\Theta} \sim \mathsf{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_m)$$

- Distribution over possible parameter vectors for a multinomial distribution, and is the conjugate prior for the multinomial.
- Beta distribution is the special case of a Dirichlet for 2 dimensions.
- Thus, it is in fact a "distribution over distributions."

Dirichlet Process

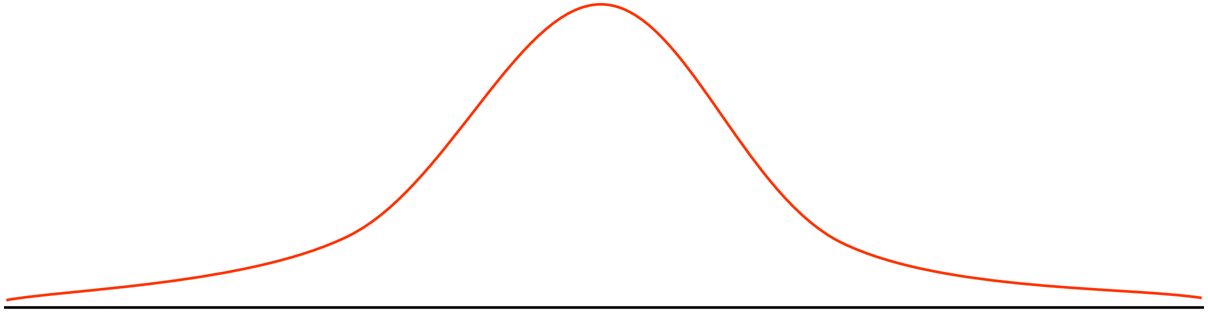
- A *Dirichlet Process* is also a distribution over distributions.
- Let G be Dirichlet Process distributed:

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

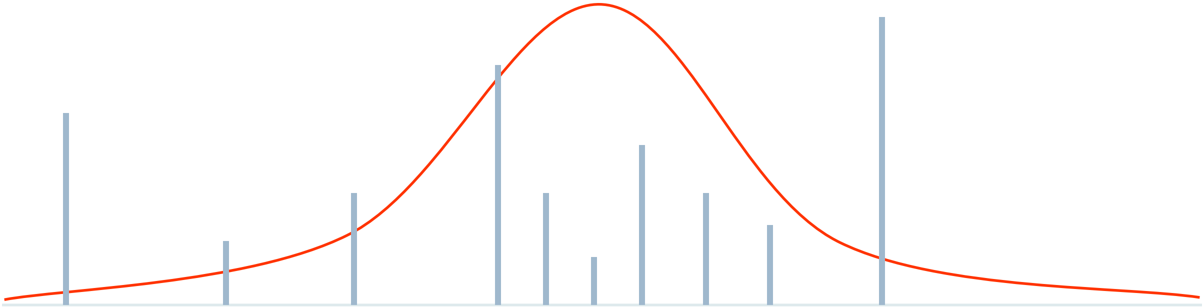
  - G_0 is a base distribution
  - α is a positive scaling parameter
- G is a random probability measure that has the same support as G_0

Dirichlet Process

- Consider Gaussian G_0

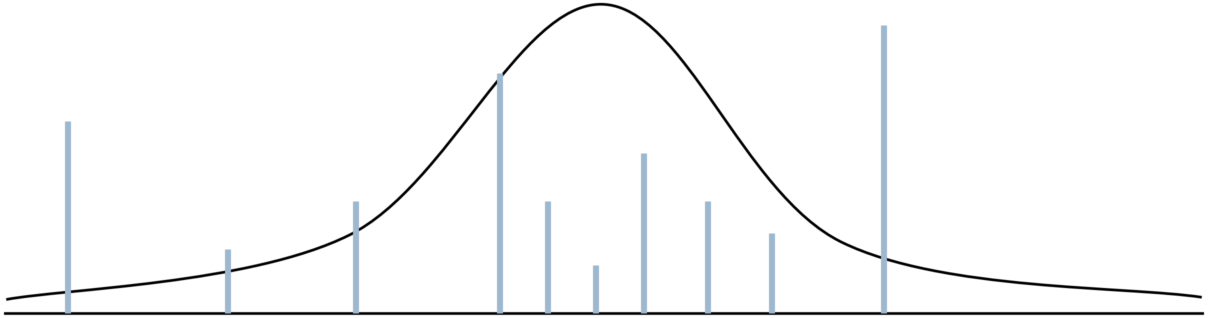


- $G \sim \mathrm{DP}(\alpha, G_0)$



Dirichlet Process

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



- G_0 is continuous, so the probability that any two samples are equal is precisely zero.
- However, G is a discrete distribution, made up of a countably infinite number of point masses [Blackwell]
  - Therefore, there is always a non-zero probability of two samples colliding

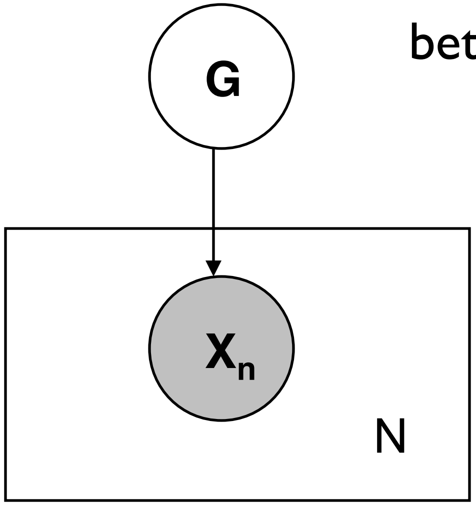
Overview

- Dirichlet distribution, and Dirichlet Process introduction
- Dirichlet Processes from different perspectives
  - Samples from a Dirichlet Process
  - Chinese Restaurant Process representation
  - Stick Breaking
  - Formal Definition
- Dirichlet Process Mixtures
- Inference

Samples from a Dirichlet Process

G ~ DP(α, G_0)

X_n | G ~ G for n = {1, …, N} (iid given G)



Marginalizing out G introduces dependencies between the X_n variables

$$P(X_1, \ldots, X_N) = \int P(G) \prod_{n=1}^{N} P(X_n|G) dG$$

Samples from a Dirichlet Process

$$P(X_1, \ldots, X_N) = \int P(G) \prod_{n=1}^{N} P(X_n|G) dG$$

Assume we view these variables in a specific order, and are interested in the behavior of X_n given the previous *n* - 1 observations.

$$X_n | X_1, \ldots, X_{n-1} = \begin{cases} X_i & \text{with probability } \frac{1}{n-1+\alpha} \\ \text{new draw from } G_0 & \text{with probability } \frac{\alpha}{n-1+\alpha} \end{cases}$$

Let there be *K* unique values for the variables:

$$X_k^* \text{ for } k \in \{1, \ldots, K\}$$

Samples from a Dirichlet Process

$$X_n | X_1, \ldots, X_{n-1} = \begin{cases} X_i & \text{with probability } \frac{1}{n-1+\alpha} \\ \text{new draw from } G_0 & \text{with probability } \frac{\alpha}{n-1+\alpha} \end{cases}$$

$$P(X_1, \ldots, X_N) = P(X_1)P(X_2|X_1)\ldots P(X_N|X_1, \ldots, X_{N-1})$$

Chain rule

$$= \frac{\alpha^K \prod_{k=1}^K (\text{num}(X_k^*) - 1)!}{\alpha(1+\alpha)\ldots(N-1+\alpha)} \prod_{k=1}^K G_0(X_k^*)$$

P(partition)

P(draws)

Notice that the above formulation of the joint distribution does not depend on the order we consider the variables.

Samples from a Dirichlet Process

$$X_n|X_1,\ldots,X_{n-1} = \begin{cases} X_i & \text{with probability } \frac{1}{n-1+\alpha} \\ \text{new draw from } G_0 & \text{with probability } \frac{\alpha}{n-1+\alpha} \end{cases}$$

Let there be *K* unique values for the variables:

$$X_k^* \text{ for } k \in \{1,\ldots,K\}$$

Can rewrite as:

$$X_n|X_1,\ldots,X_{n-1} = \begin{cases} X_k^* & \text{with probability } \frac{\text{num}_{n-1}(X_k^*)}{n-1+\alpha} \\ \text{new draw from } G_0 & \text{with probability } \frac{\alpha}{n-1+\alpha} \end{cases}$$

Blackwell-MacQueen Urn Scheme

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

- Assume that G_0 is a distribution over colors, and that each X_n represents the color of a single ball placed in the urn.
- Start with an empty urn.
- On step *n*:
  - With probability proportional to α, draw $X_n \sim G_0$, and add a ball of that color to the urn.
  - With probability proportional to $n - 1$ (i.e., the number of balls currently in the urn), pick a ball at random from the urn. Record its color as X_n, and return the ball into the urn, along with a new one of the same color.

[Blackwell and Macqueen, 1973]

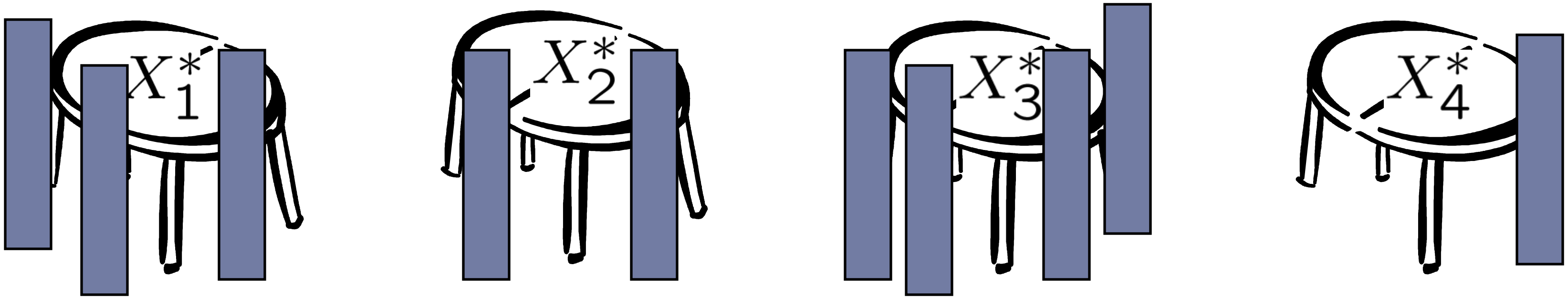
Chinese Restaurant Process

$$X_n|X_1,\ldots,X_{n-1} = \begin{cases} X_k^* & \text{with probability } \frac{\text{num}_{n-1}(X_k^*)}{n-1+\alpha} \\ \text{new draw from } G_0 & \text{with probability } \frac{\alpha}{n-1+\alpha} \end{cases}$$

Consider a restaurant with infinitely many tables, where the X_n's represent the patrons of the restaurant. From the above conditional probability distribution, we can see that a customer is more likely to sit at a table if there are already many people sitting there. However, with probability proportional to α, the customer will sit at a new table.

Also known as the "clustering effect," and can be seen in the setting of social clubs. [Aldous]

Chinese Restaurant Process



Stick Breaking

- So far, we've just mentioned properties of a distribution G drawn from a Dirichlet Process
- In 1994, Sethuraman developed a constructive way of forming G, known as "stick breaking"

Stick Breaking

$$V_1, V_2, \ldots, V_i, \ldots \sim \text{Beta}(1, \alpha)$$

$$f(V_i = v_i|\alpha) = \alpha(1 - v_i)^{\alpha-1}$$

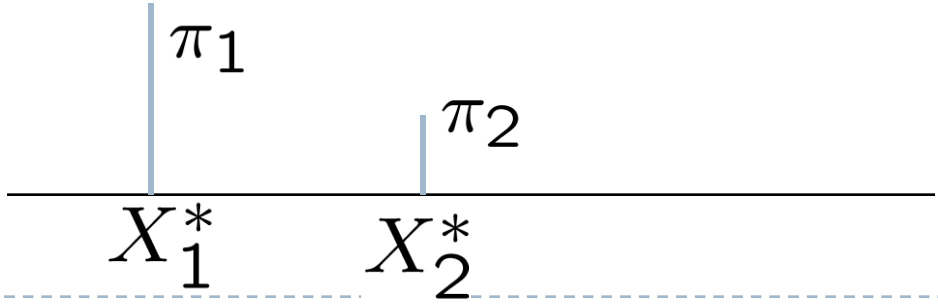
$$X_1^*, X_2^*, \ldots, X_i^*, \ldots \sim G_0$$

$$\pi_i(\mathbf{v}) = v_i \prod_{j=1}^{i-1} (1 - v_j)$$

$$G = \sum_{i=1}^{\infty} \pi_i(\mathbf{v}) \delta_{X_i^*}$$

1. Draw X_1* from G_0
2. Draw v_1 from Beta(1, α)
3. $\pi_1 = v_1$
4. Draw X_2* from G_0
5. Draw v_2 from Beta(1, α)
6. $\pi_2 = v_2(1 – v_1)$

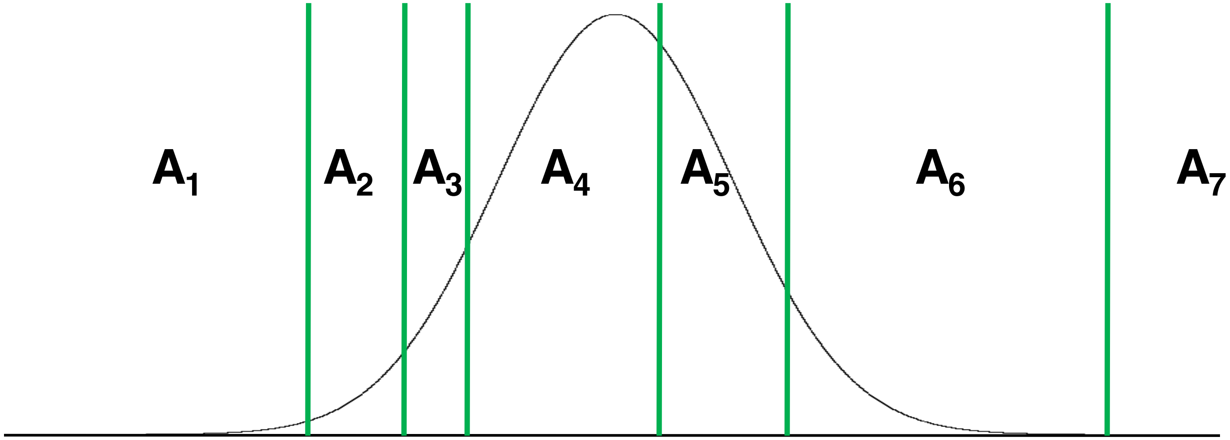
…



Formal Definition (not constructive)

- Let α be a positive, real-valued scalar
- Let G_0 be a non-atomic probability distribution over support set A
- If G ~ DP(α, G_0), then for any finite set of partitions $A_1 \cup A_2 \cup \ldots \cup A_k$ of A:

$$(G(A_1), \ldots, G(A_k)) \sim \mathrm{Dirichlet}(\alpha G_0(A_1), \ldots, \alpha G_0(A_k))$$

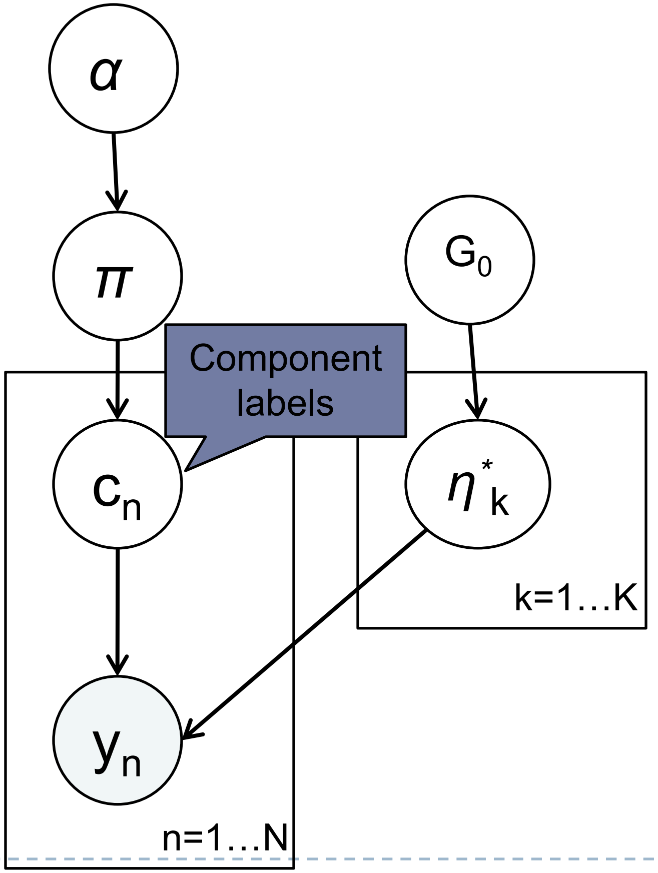


Overview

- Dirichlet distribution, and Dirichlet Process introduction
- Dirichlet Processes from different perspectives
  - Samples from a Dirichlet Process
  - Chinese Restaurant Process representation
  - Stick Breaking
  - Formal Definition
- Dirichlet Process Mixtures
- Inference

Finite Mixture Models

- A finite mixture model assumes that the data come from a mixture of a finite number of distributions.



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

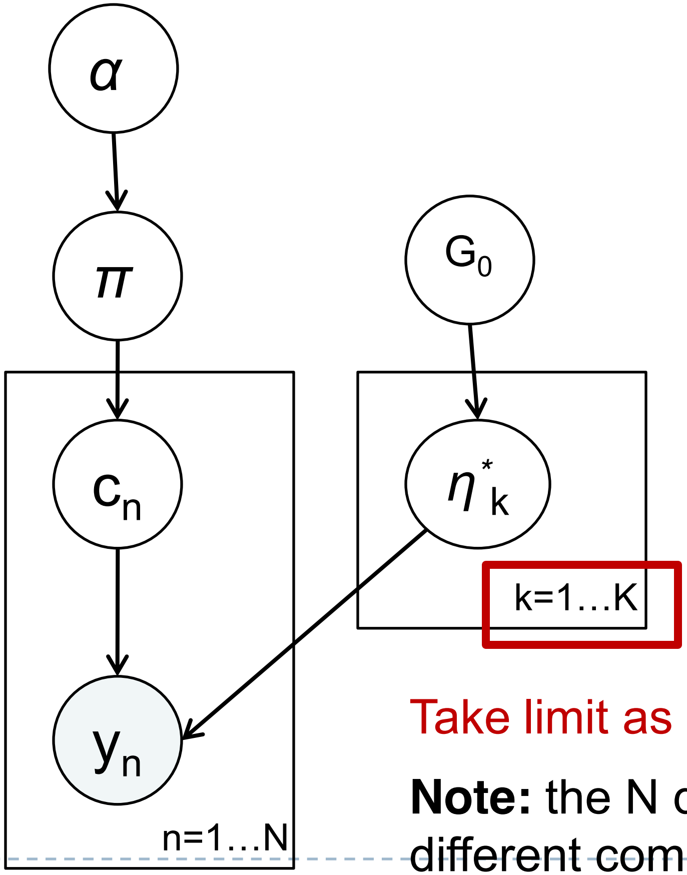
$c_n \sim \text{Multinomial}(\pi)$

$\eta_k \sim G_0$

$y_n \mid c_n, \eta_1, \ldots \eta_K \sim F(\cdot \mid \eta_{c_n})$

Infinite Mixture Models

- An infinite mixture model assumes that the data come from a mixture of an *infinite* number of distributions



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

$c_n \sim \text{Multinomial}(\pi)$

$\eta_k \sim G_0$

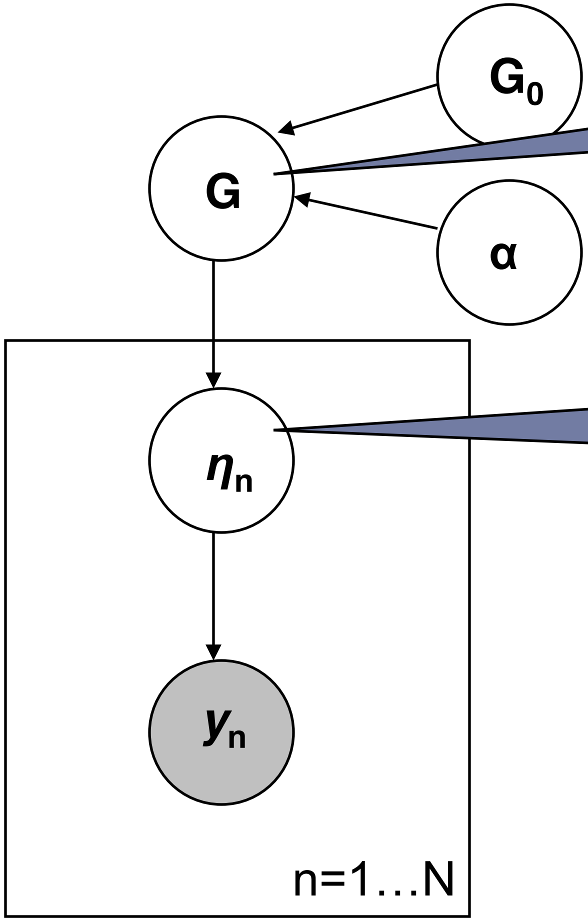
$y_n \mid c_n, \eta_1, \ldots \eta_K \sim F(\,\cdot \mid \eta_{c_n})$

Take limit as K goes to ∞

Note: the N data points still come from at most N different components

[Rasmussen 2000]

Dirichlet Process Mixture



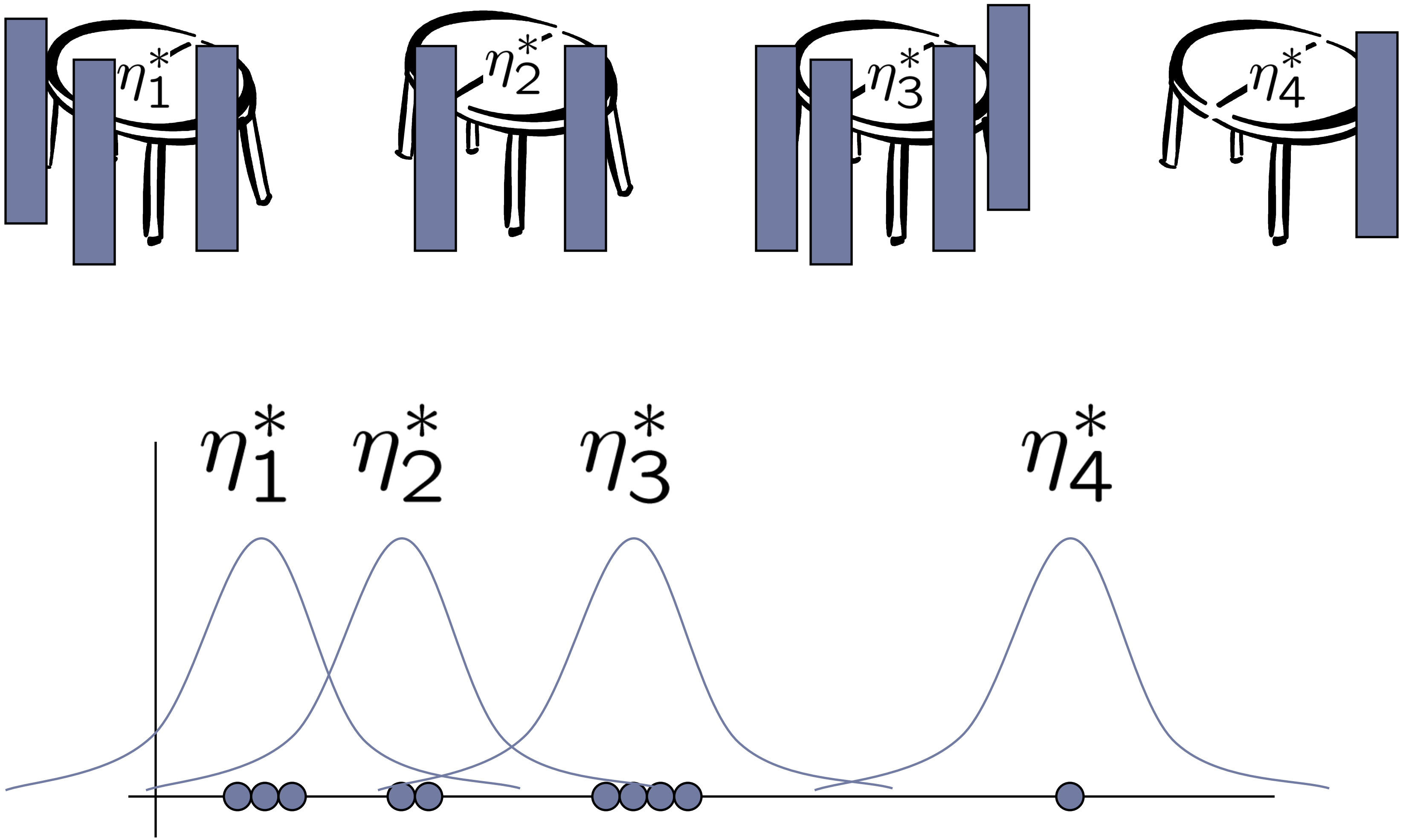
countably infinite number of point masses

draw N times from G to get parameters for different mixture components

If η_n were drawn from, e.g., a Gaussian, no two values would be the same, but since they are drawn from a Dirichlet Process-distributed distribution, we expect a clustering of the η_n

# unique values for η_n = # mixture components

CRP Mixture



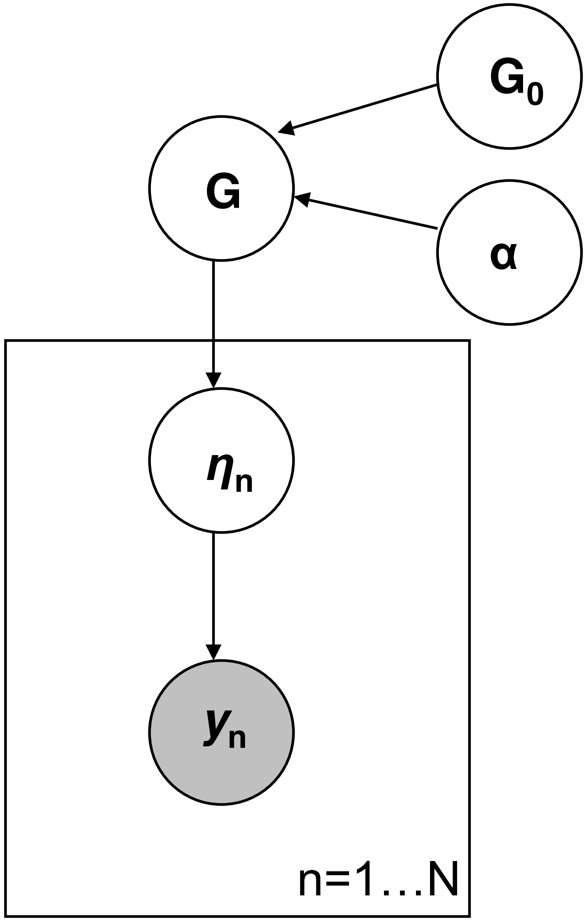
Overview

- Dirichlet distribution, and Dirichlet Process introduction
- Dirichlet Processes from different perspectives
  - Samples from a Dirichlet Process
  - Chinese Restaurant Process representation
  - Stick Breaking
  - Formal Definition
- Dirichlet Process Mixtures
- Inference

Inference for Dirichlet Process Mixtures

- Expectation Maximization (EM) is generally used for inference in a mixture model, but G is nonparametric, making EM difficult

- Markov Chain Monte Carlo techniques [Neal 2000]
- Variational Inference [Blei and Jordan 2006]



Aside: Monte Carlo Methods

[Basic Integration]

- We want to compute the integral,

$$I = \int h(x) f(x) dx$$

where f(x) is a probability density function.

- In other words, we want E_f [h(x)].
- We can approximate this as:

$$\hat{I} = \frac{1}{N} \sum_{i=1}^{N} h(X_i)$$

where X_1, X_2, …, X_N are sampled from f.

- By the law of large numbers,

$$\hat{I} \xrightarrow{p} I$$

[Lafferty and Wasserman]

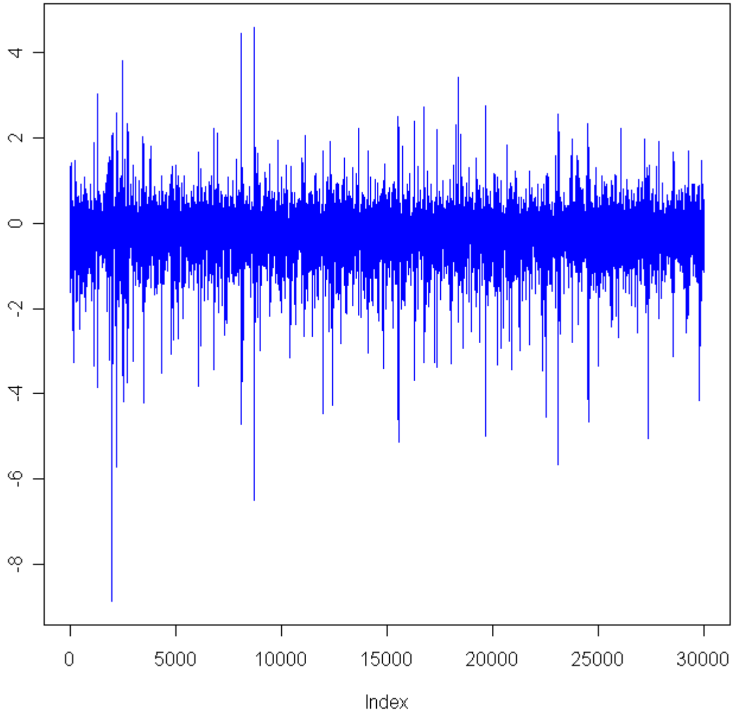
Aside: Monte Carlo Methods

[What if we don't know how to sample from f?]

- Importance Sampling
- Markov Chain Monte Carlo (MCMC)
  - Goal is to generate a Markov chain $X_1, X_2, \ldots,$ whose stationary distribution is f.
  - If so, then

$$\frac{1}{N}\sum_{i=1}^{N} h(X_i) \xrightarrow{p} I$$

(under certain conditions)



Aside: Monte Carlo Methods

[MCMC I: Metropolis-Hastings Algorithm]

Goal: Generate a Markov chain with stationary distribution f(x)

Initialization:

- Let q(y | x) be an arbitrary distribution that we know how to sample from. We call q the **ion**.
- Arbitrarily choose X

A common choice is N(x, b²) for b > 0

Assume we have generated $X_0, X_1, \ldots, X_i$. T
X_{i+1}:

- Generate a proposal value Y ~ q(y|X_i)
- Evaluate r ≡ r(X_i, Y) where:

$$r(x,y) = \min\left\{\boxed{\frac{f(y)}{f(x)}\frac{q(x|y)}{q(y|x)}},\ 1\right\}$$

If q is symmetric, simplifies to: $\frac{f(y)}{f(x)}$

- Set:

$$X_{i+1} = \begin{cases} Y & \text{with probability r} \\ X_i & \text{with probability 1-r} \end{cases}$$

[Lafferty and Wasserman]

Aside: Monte Carlo Methods

[MCMC II: Gibbs Sampling]

Goal: Generate a Markov chain with stationary distribution f(x, y) [Easily extendable to higher dimensions.]

Assumption:

- We know how to sample from the conditional distributions $f_{X|Y}(x \mid y)$ and $f_{Y|X}(y \mid x)$

If not, then we run one iteration of Metropolis-Hastings each time we need to sample from a conditional.

Initialization:

- Arbitrarily choose X_0, Y_0.

Assume we have generated (X_0,Y_0), …, (X_i,Y_i). To generate (X_{i+1},Y_{i+1}):

- $X_{i+1} \sim f_{X|Y}(x \mid Y_i)$
- $Y_{i+1} \sim f_{Y|X}(y \mid X_{i+1})$

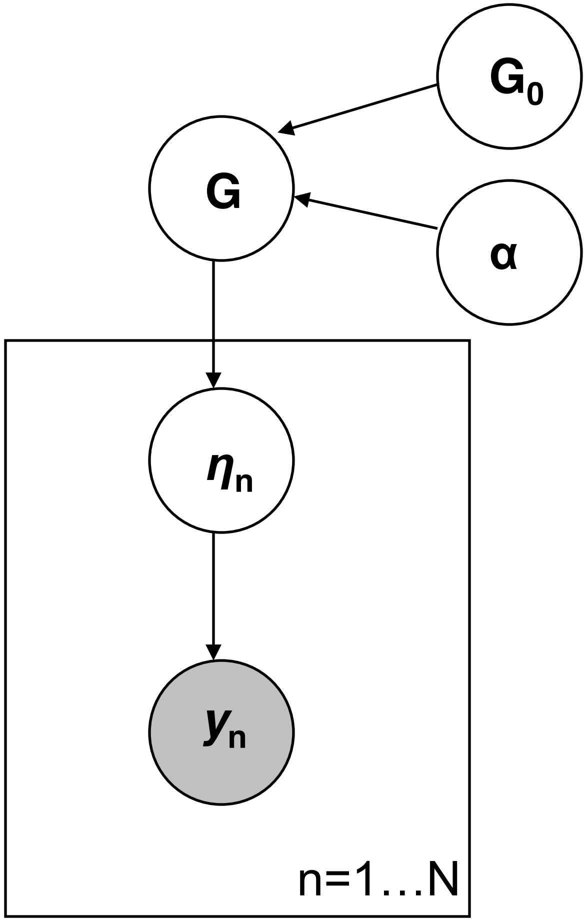
[Lafferty and Wasserman]

MCMC for Dirichlet Process Mixtures
[Overview]

- We would like to sample from the posterior distribution:

  $$P(\eta_1,\ldots,\eta_N \mid y_1,\ldots y_N)$$

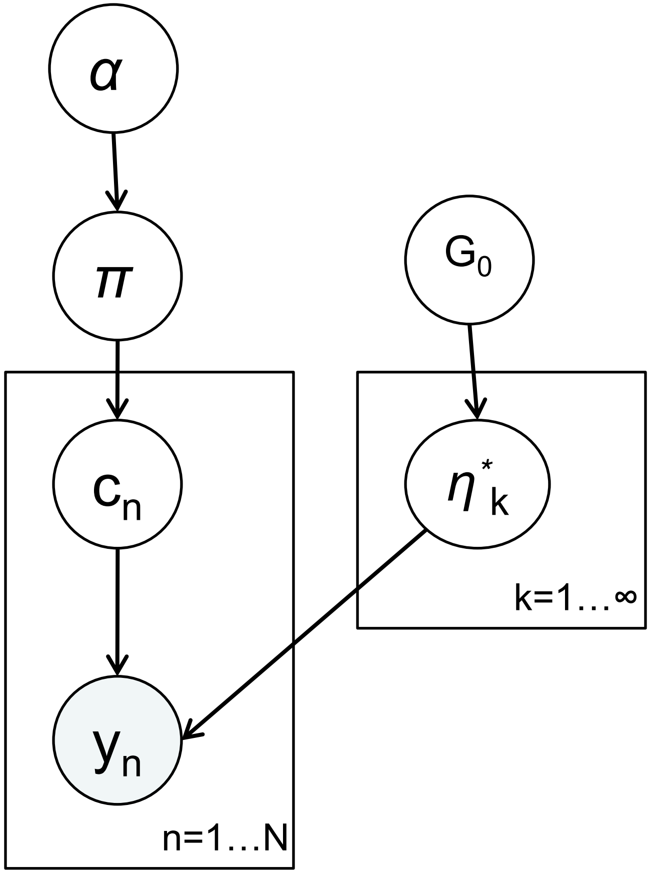
- If we could, we would be able to determine:
  - how many distinct components are likely contributing to our data.
  - what the parameters are for each component.
- [Neal 2000] is an excellent resource describing several MCMC algorithms for solving this problem.
  - We will briefly take a look at two of them.



MCMC for Dirichlet Process Mixtures

[Infinite Mixture Model representation]

- MCMC algorithms that are based on the infinite mixture model representation of Dirichlet Process Mixtures are found to be simpler to implement and converge faster than those based on the direct representation.
- Thus, rather than sampling for $\eta_1,\ldots,\eta_N$ directly, we will instead sample for the component indicators $c_1, \ldots, c_N$, as well as the component parameters $\eta^*_{c,}$ for all c in $\{c_1, \ldots, c_N\}$



[Neal 2000]

MCMC for Dirichlet Process Mixtures

[Gibbs Sampling with Conjugate Priors]

Assume current state of Markov chain consists of $c_1, \ldots, c_N$, as well as the component parameters $\eta^*_{c,}$ for all c in $\{c_1, \ldots, c_N\}$.

To generate the next sample:

1. For i = 1,…,N:
   - If c_i is currently a singleton, remove $\eta^*_{c_i}$ from the state.
   - Draw a new value for c_i from the conditional distribution:

$$P(c_i = c | c_{-i}, y_i, \eta^*) \propto \begin{cases} \dfrac{N_{-i,c}}{N-1+\alpha} F(y_i, \eta^*_c) & \text{for existing c} \\ \dfrac{\alpha}{N-1+\alpha} \int F(y_i, \eta^*) dG_0(\eta^*) & \text{for new c} \end{cases}$$

   - If the new c_i is not associated with any other observation, draw a value for $\eta^*_{c_i}$ from:

$$P(\eta^* | y_i) \propto F(y_i, \eta^*) G_0(\eta^*)$$

[Neal 2000, Algorithm 2]

MCMC for Dirichlet Process Mixtures

[Gibbs Sampling with Conjugate Priors]

2. For all c in $\{c_1, \ldots, c_N\}$:
   - Draw a new value for η^*_c from the posterior distribution based on the prior G_0 and all the data points currently associated with component c:

$$\begin{aligned} P(\eta^* | \mathbf{y_c}) &\propto \prod_{i:c_i=c} P(y_i|\eta^*)P(\eta^*) \\ &= \prod_{i:c_i=c} F(y_i, \eta^*)G_0(\eta^*) \end{aligned}$$

This algorithm breaks down when G_0 is not a conjugate prior.

[Neal 2000, Algorithm 2]

MCMC for Dirichlet Process Mixtures

[Gibbs Sampling with Auxiliary Parameters]

- Recall from the Gibbs sampling overview: if we do not know how to sample from the conditional distributions, we can interleave one or more Metropolis-Hastings steps.
  - We can apply this technique when G_0 is not a conjugate prior, but it can lead to convergence issues [Neal 2000, Algorithms 5-7]
  - Instead, we will use **auxiliary parameters**.
- Previously, the state of our Markov chain consisted of $c_1, \ldots, c_N$, as well as component parameters $\eta^*_{c,}$ for all c in $\{c_1, \ldots, c_N\}$.
- When updating c_i, we either:
  - choose an existing component c from c_{-i} (i.e., all c_j such that $j \neq i$).
  - choose a brand new component.
    - In the previous algorithm, this involved integrating with respect to G_0, which is difficult in the non-conjugate case.

[Neal 2000, Algorithm 8]

MCC for Dirichlet Process Mixtures

[Gibbs Sampling with Auxiliary Parameters]

- When updating c_i, we either:
  - choose an existing component c from c_{-i} (i.e., all c_j such that $j \neq i$).
  - choose a brand new component.
- Let K_{-i} be the number of distinct components c in c_{-i}.
- WLOG, let these components c_{-i} have values in $\{1, \ldots, K_{-i}\}$.
- Instead of integrating over G_0, we will add m auxiliary parameters, each corresponding to a new component independently drawn from G_0 :

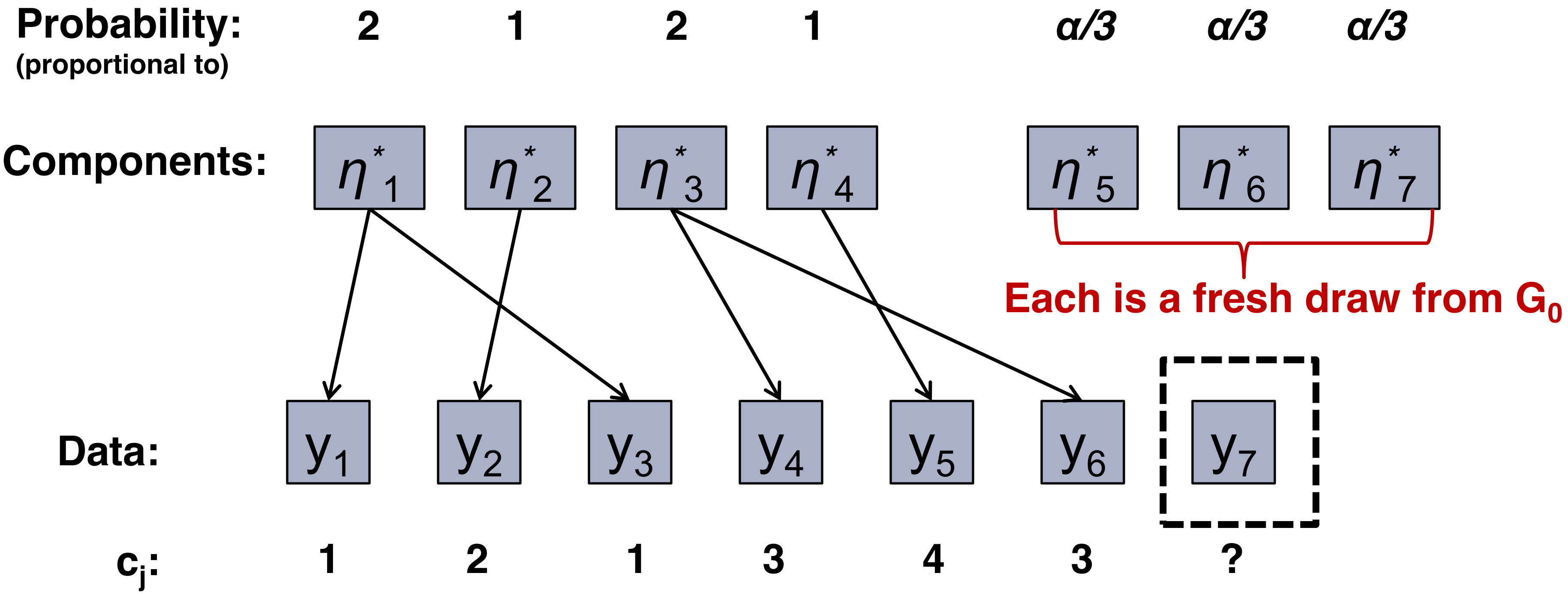
$$[\eta^*_{K_{-i}+1}, \ldots, \eta^*_{K_{-i}+m}]$$

- Recall that the probability of selecting a new component is proportional to α.
  - Here, we divide α equally among the m auxiliary components.

[Neal 2000, Algorithm 8]

MCMC for Dirichlet Process Mixtures

[Gibbs Sampling with Auxiliary Parameters]



Probability: **(proportional to)** 2 1 2 1 $\alpha/3$ $\alpha/3$ $\alpha/3$

Components: η^*_1 η^*_2 η^*_3 η^*_4 η^*_5 η^*_6 η^*_7

Each is a fresh draw from G_0

Data: y_1 y_2 y_3 y_4 y_5 y_6 y_7

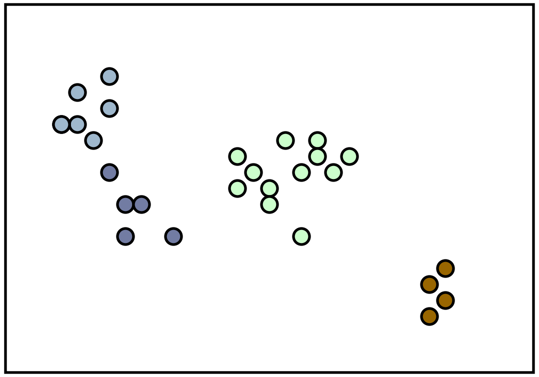
c_j: 1 2 1 3 4 3 ?

- This takes care of sampling for c_i in the non-conjugate case.
- A Metropolis-Hastings step can be used to sample η^*_c.
  - See Neal's paper for more details.

[Neal 2000, Algorithm 8]

Conclusion

- We now have a statistically principled mechanism for solving our original problem.



- This was intended as a general and fairly high level overview of Dirichlet Processes.
  - Topics left out include Hierarchical Dirichlet Processes, variational inference for Dirichlet Processes, and many more.
  - Teh's MLSS '07 tutorial provides a much deeper and more detailed take on DPs—highly recommended!

Acknowledgments

- Much thanks goes to David Blei.
- Some material for this presentation was inspired by slides from Teg Grenager, Zoubin Ghahramani, and Yee Whye Teh

References

David Blackwell and James B. MacQueen. "Ferguson Distributions via Polya Urn Schemes." *Annals of Statistics* 1(2), 1973, 353-355.

David M. Blei and Michael I. Jordan. "Variational inference for Dirichlet process mixtures." *Bayesian Analysis* 1(1), 2006.

Thomas S. Ferguson. "A Bayesian Analysis of Some Nonparametric Problems" *Annals of Statistics* 1(2), 1973, 209-230.

Zoubin Gharamani. "Non-parametric Bayesian Methods." UAI Tutorial July 2005.

Teg Grenager. "Chinese Restaurants and Stick Breaking: An Introduction to the Dirichlet Process"

R.M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9:249-265, 2000.

C.E. Rasmussen. The Infinite Gaussian Mixture Model. *Advances in Neural Information Processing Systems* 12, 554-560. (Eds.) Solla, S. A., T. K. Leen and K. R. Müller, MIT Press (2000).

Y.W. Teh. "Dirichlet Processes." Machine Learning Summer School 2007 Tutorial and Practical Course.

Y.W. Teh, M.I. Jordan, M.J. Beal and D.M. Blei. "Hierarchical Dirichlet Processes." *J. American Statistical Association* 101(476):1566-1581, 2006.