Statistical Machine Learning

Dirichlet Process (DP)
Chinese Restaurant Process (CRP)
Indian Buffet Process (IBP)

Spring 2020

Dirichlet Process (DP)

Dirichlet Process (DP): What’s DP good for?

- A good Bayesian method for fitting a mixture model with an unknown number of clusters

- Because it’s Bayesian, can build Hierarchies Dirichlet Process (HDP) and integrate with other random variables in a principled way

What is the difference between
Dirichlet Distribution &
Dirichlet Process

The Dirichlet Distribution

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

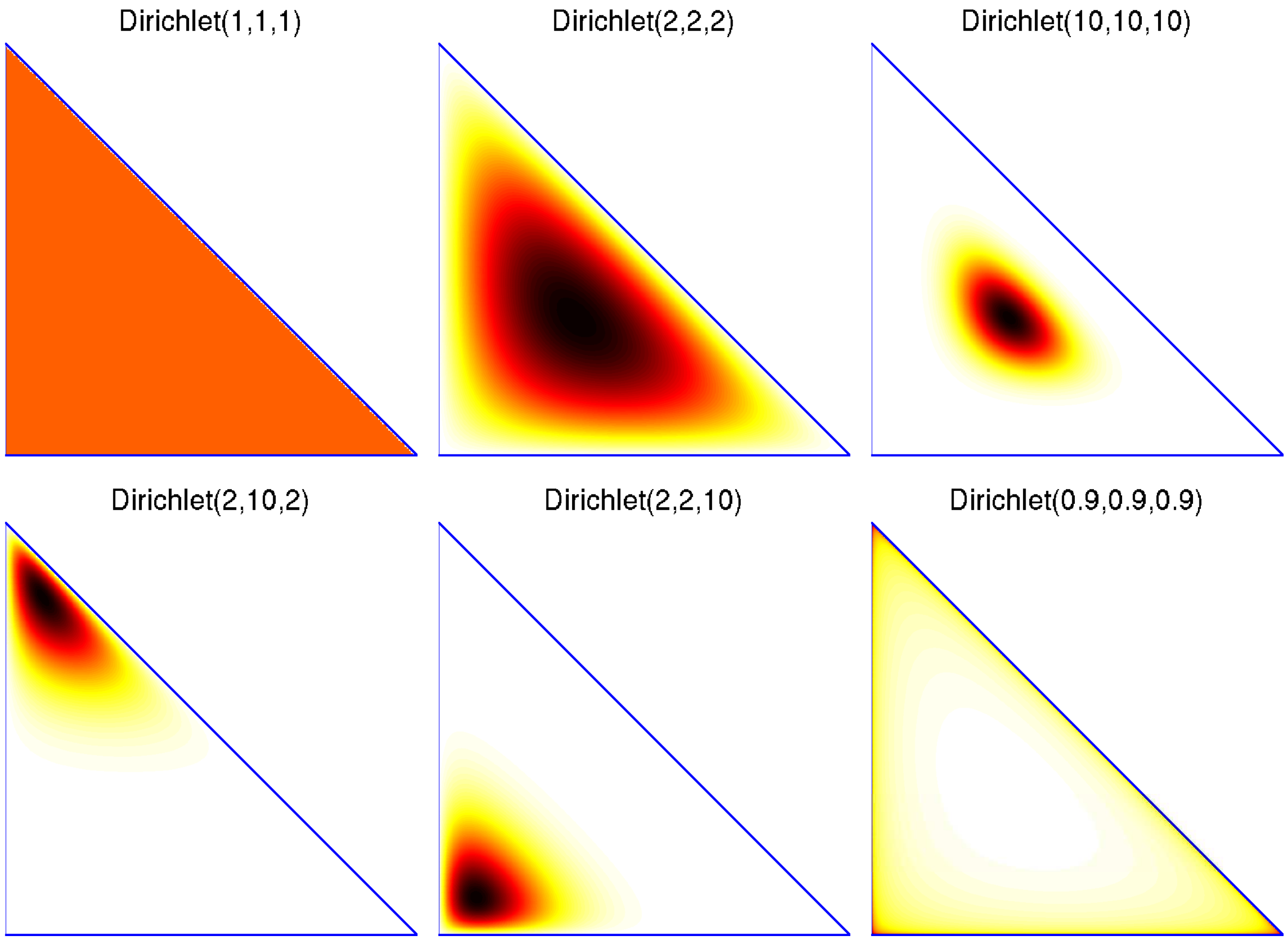
  Then:

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

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

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

Dirichlet Distributions



Dirichlet(1,1,1)

Dirichlet(2,2,2)

Dirichlet(10,10,10)

Dirichlet(2,10,2)

Dirichlet(2,2,10)

Dirichlet(0.9,0.9,0.9)

Examples of Dirichlet distributions over **p** = (p_1, p_2, p_3) which can be plotted in 2D

The Dirichlet Distribution

- Dirichlet distribution, is a distribution over possible parameter vectors for a multinomial distribution, and is the conjugate prior for the multinomial (categorical) distribution.

- Beta distribution is the special case of a Dirichlet for 2 dimensions.

- Dirichlet distribution is in fact a distribution over distributions.

- The infinite-dimensional generalization of the Dirichlet distribution is the Dirichlet Process (DP).

The Dirichlet Process

- Dirichlet processes are a family of stochastic processes whose realizations are probability distributions.

- A Dirichlet process is a probability distribution whose range is itself a set of probability distributions (how likely it is that the random variables are distributed according to one or another particular distribution).

- The Dirichlet process is specified by a base distribution G_0 and a positive real number α called the concentration (scaling) parameter.

The Dirichlet Process

- The base distribution is the expected value of the process; the Dirichlet process draws distributions "around" the base distribution the way a normal distribution draws real numbers around its mean.

- Even if the base distribution is continuous, the distributions drawn from DP are discrete.

- The scaling parameter specifies how strong this discretization is:
    - As α goes to 0, the realizations are all concentrated at a single value
    - As α goes to infinity, the realizations become continuous
    - Between the two extremes the realizations are discrete distributions

The Dirichlet Process

In summary:

- A Dirichlet Process a distribution over distributions.
- Let G be a Dirichlet Process:

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

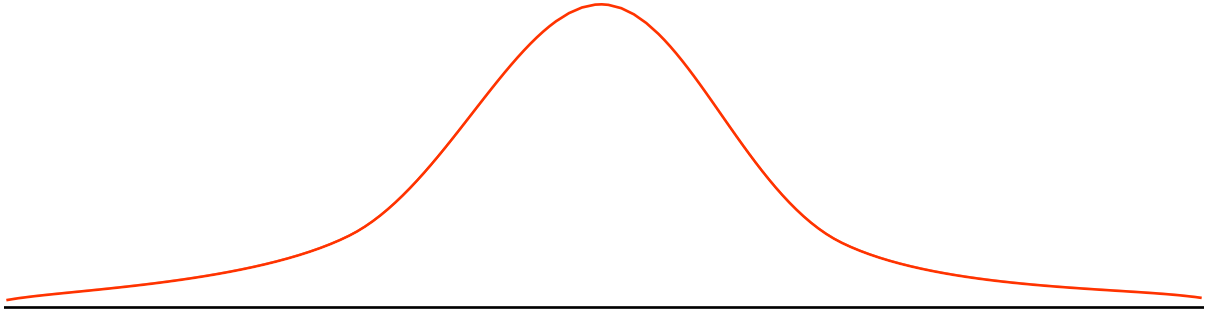
  - G_0 is a base distribution

  - α is a positive scaling parameter

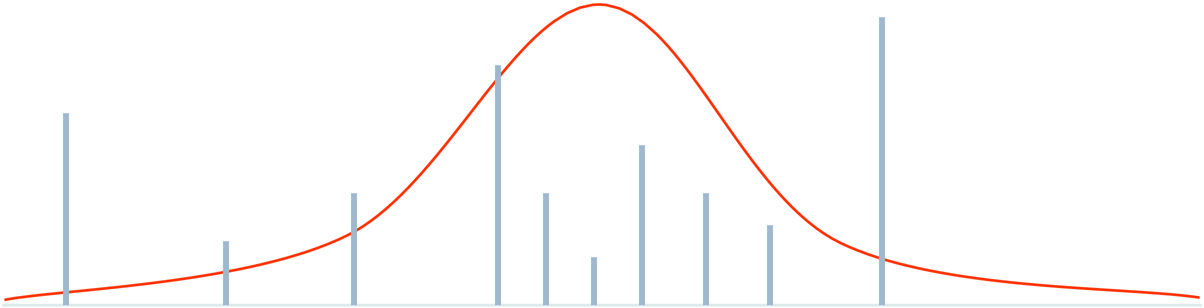
- G is a random probability measure that has the same support as G_0
- Dirichlet process is the conjugate prior for infinite, nonparametric discrete distributions.
- An important application of DP is as a prior probability distribution in infinite mixture models.

The Dirichlet Process

- Consider Gaussian G_0

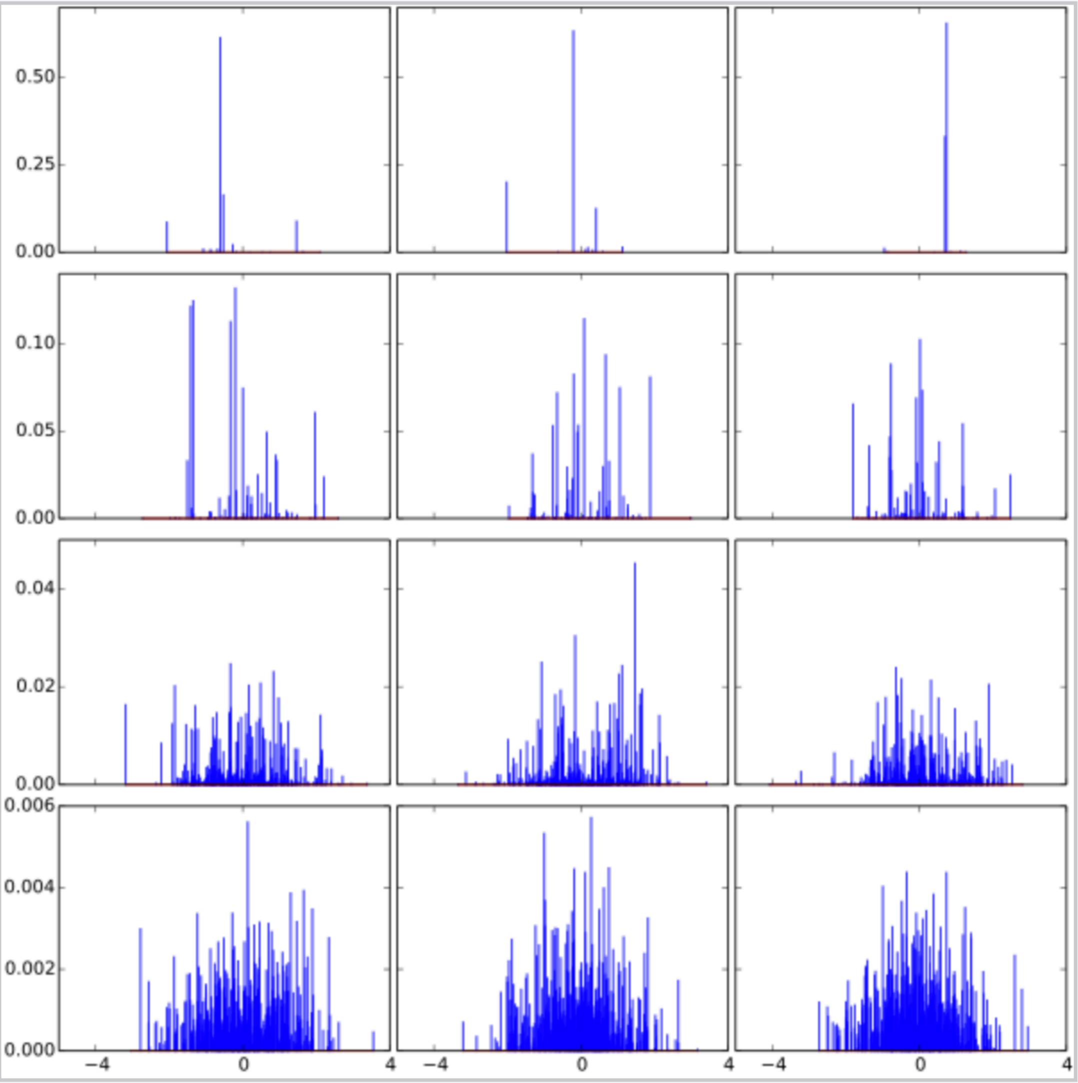


- $G \sim DP(\alpha, G_0)$



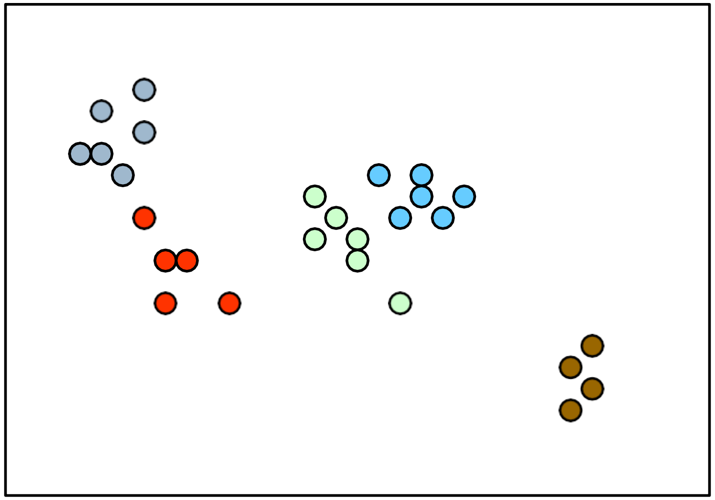
The Dirichlet Process

- Samples from the Dirichlet process D(N(0,1), α) Top to bottom α is 1, 10, 100, and 1000.
- Each row contains three repetitions of the same experiment.
- Samples from a Dirichlet process are discrete distributions.



DP Application

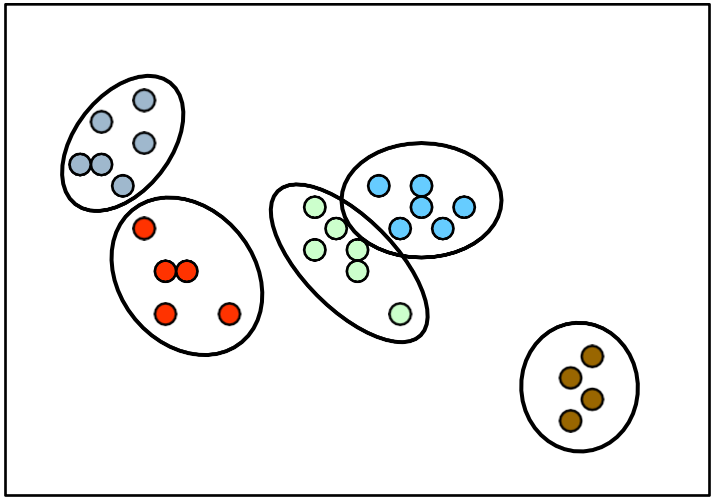
- We are given a data set, and are told that it was generated from a mixture of Gaussian distributions.



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

DP Application

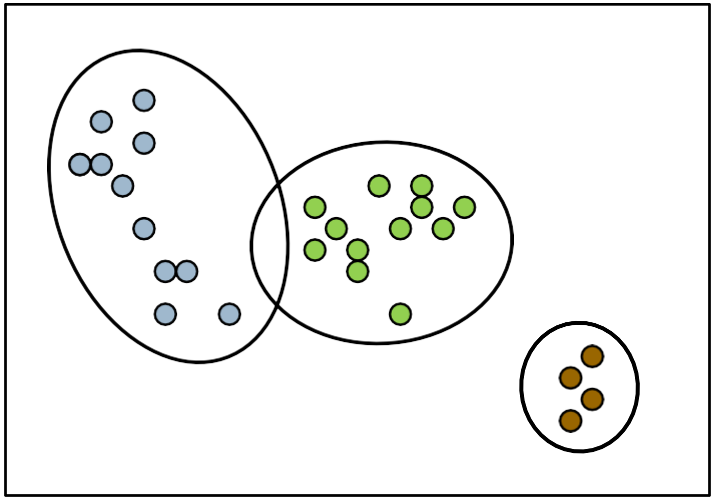
- We are given a data set, and are told that it was generated from a mixture of Gaussian distributions.



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

DP Application

- 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…

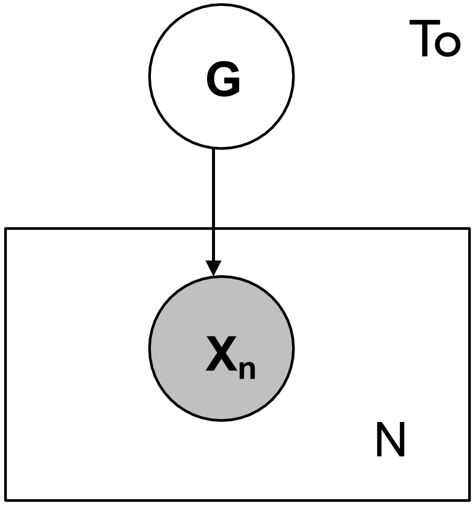
- We want to cluster the data in a statistically principled manner, without resorting to adhoc solutions.

Samples from a Dirichlet Process

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

$X_n \mid G \sim G$ for n = {1,…,N}   (iid given G)

Marginalizing out G
To obtain 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:

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

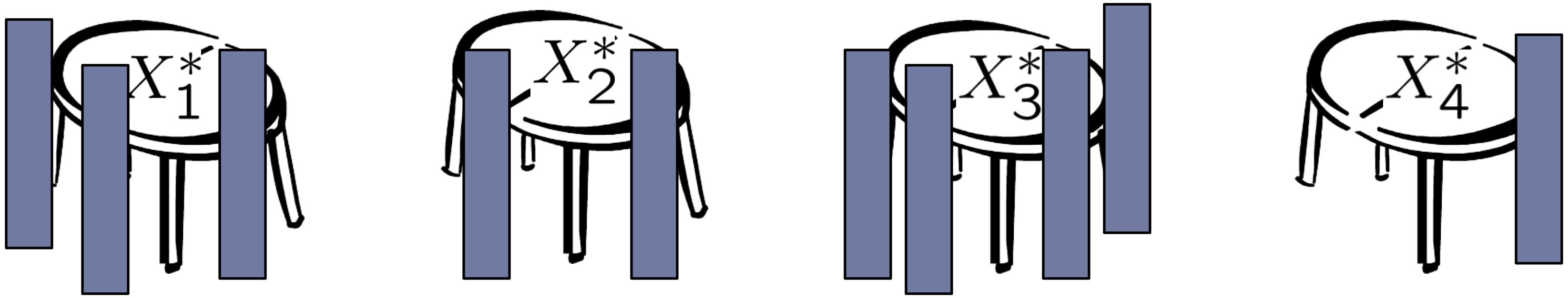
Chinese Restaurant Process (CRP)

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.

Chinese Restaurant Process



Stick Breaking

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

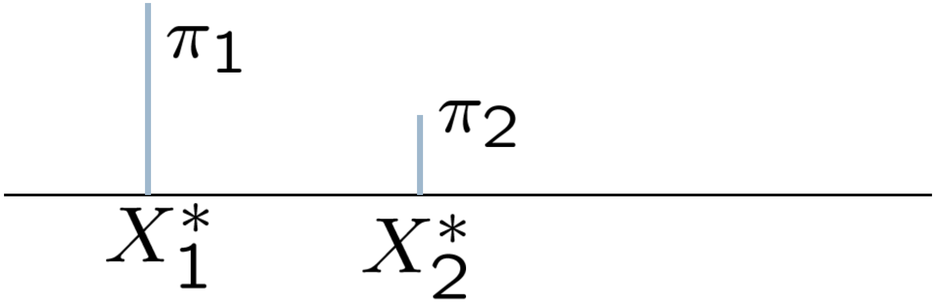
$$f(V_i = v_i|\alpha) = \alpha(1 - v_i)^{\alpha - 1}$$

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

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

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

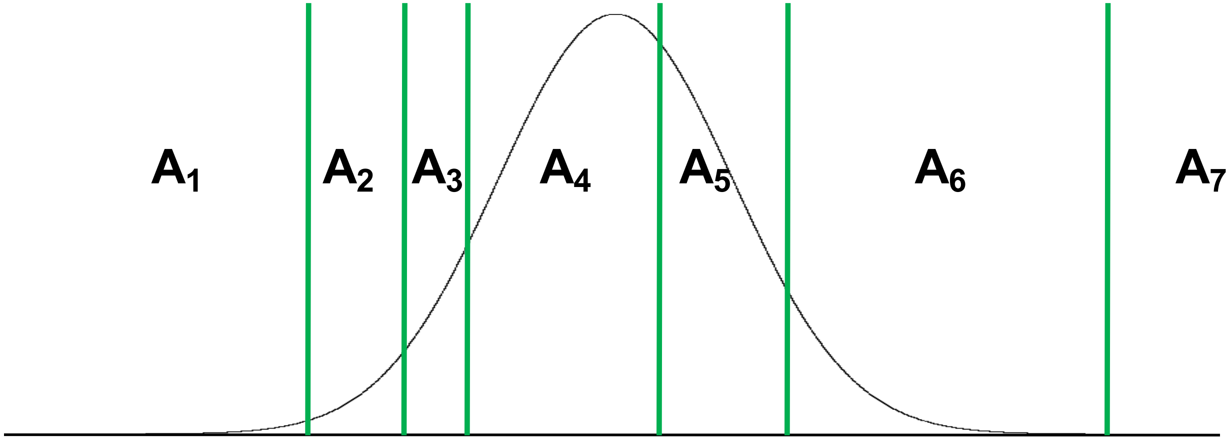
1. Draw X_1* from G_0
2.Draw v_1 from Beta(1, α)
3. $\pi_1 = v_1$
4. Draw X_2* from G_0
5.Draw v_2 from Beta(1, α)
6. $\pi_2 = v_2(1 – v_1)$
…



Formal Definition (not constructive)

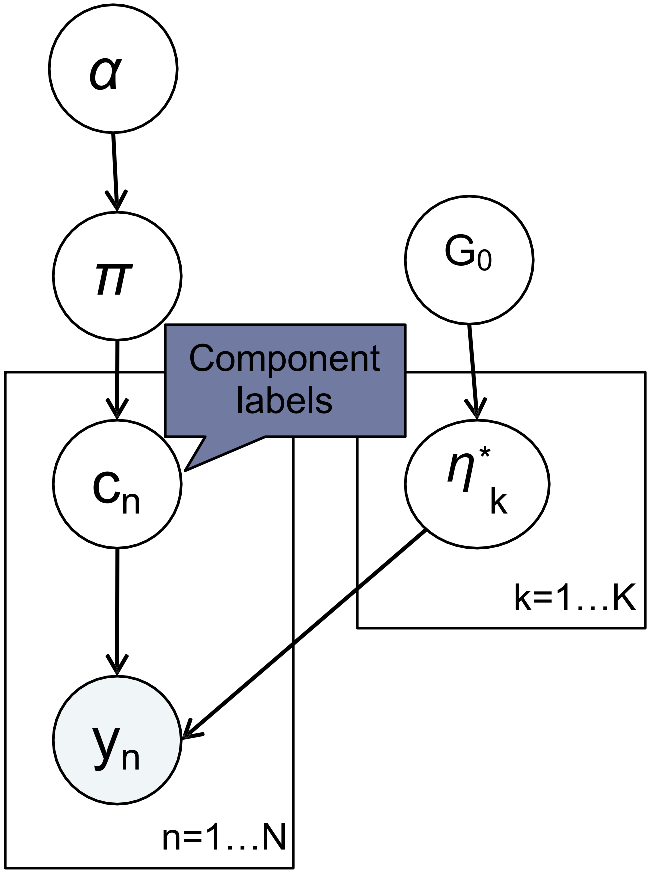
- Let α be a positive, real-valued scalar
- Let G_0 be a probability distribution over support setA
- If G ~ DP(α, G_0),then for any finite set of partitions $A_1 \cup A_2 \cup \ldots \cup A_k$ ofA:

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



Finite Mixture Models

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



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

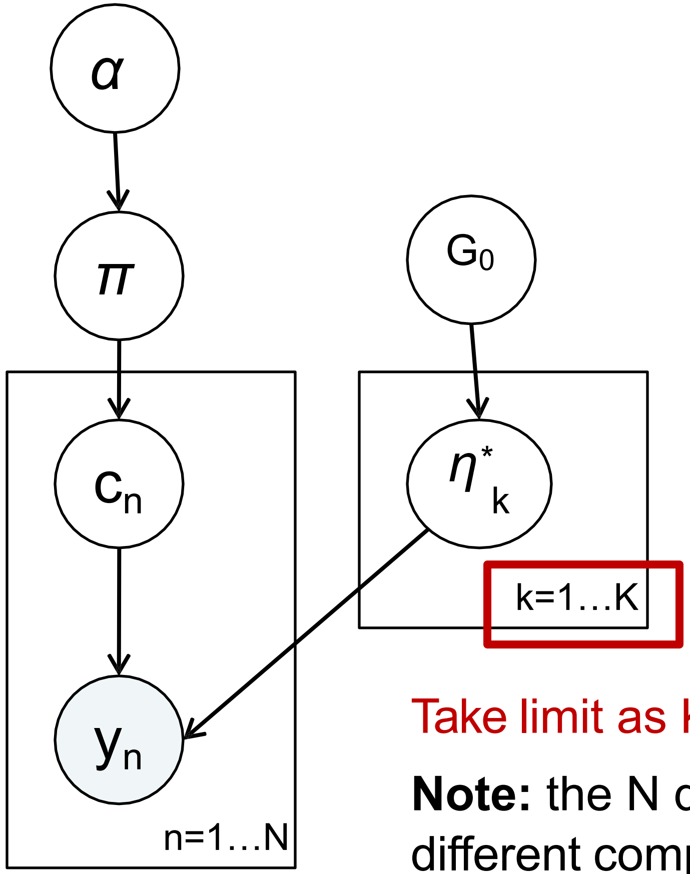
$c_n \sim \text{Multinomial}(\pi)$

$\eta_k \sim G_0$

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

Infinite Mixture Models

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



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

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

$\eta_k \sim G_0$

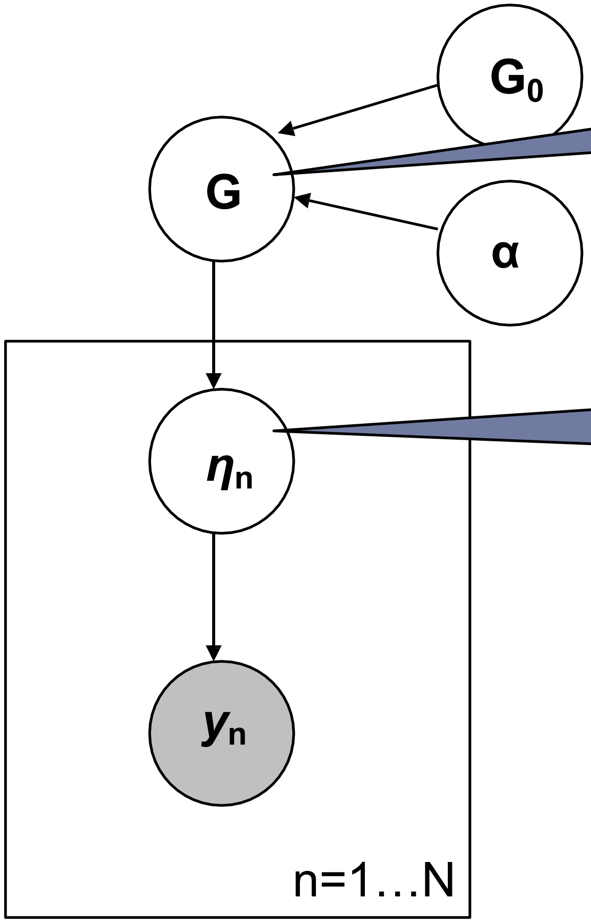
$y_n \mid c_n, \eta_1, \dots \eta_K \sim F(\cdot \mid \eta_{c_n})$

Take limit as K goes to ∞

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

[Rasmussen 2000]

Dirichlet Process Mixture



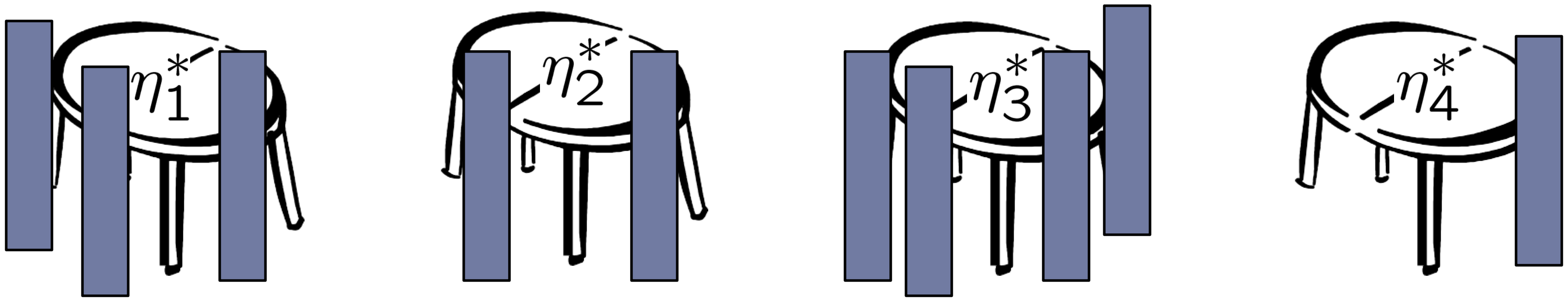
countably infinite number of point masses

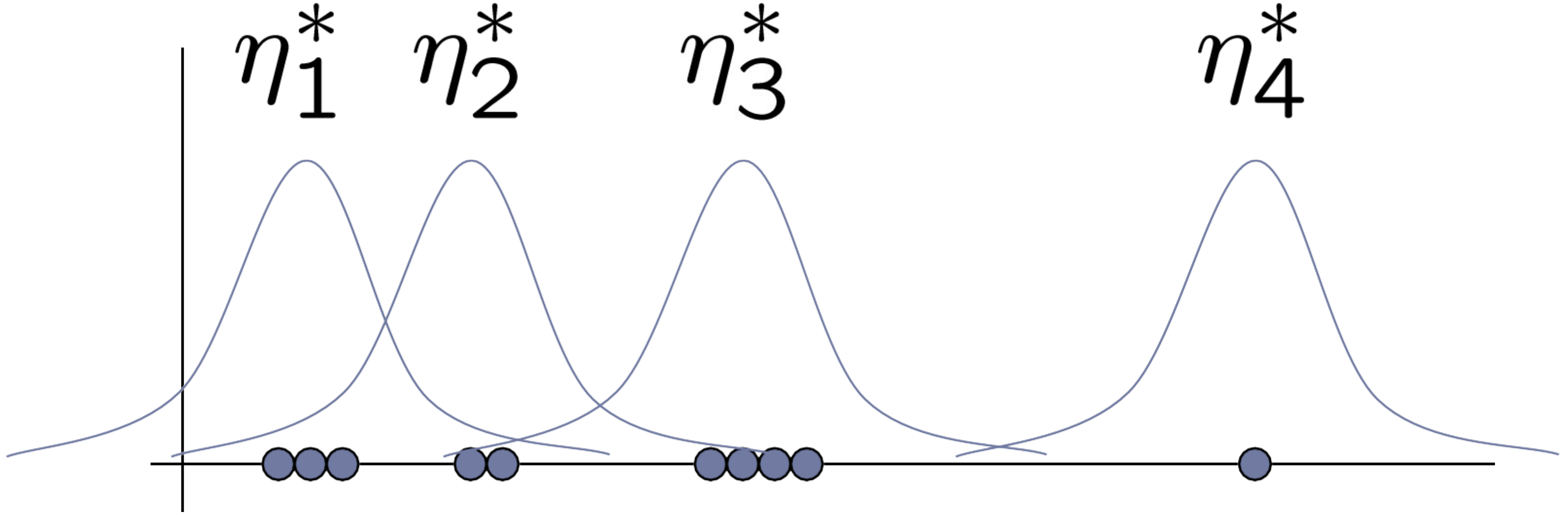
draw N times from G to get parameters for different mixture components

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

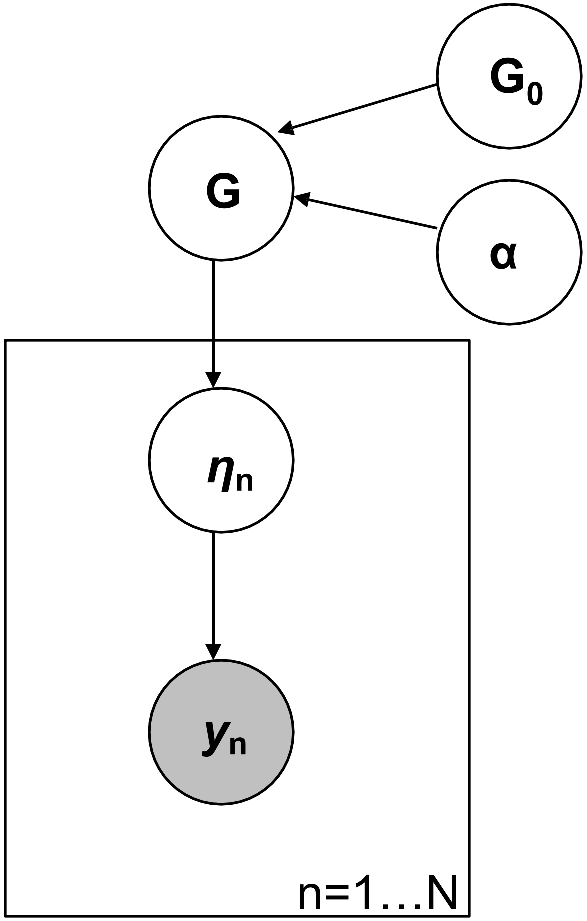




Inference for Dirichlet Process Mixtures

- 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 $X_1, X_2, \ldots, X_N$ are sampled from f.
- By the law of large numbers, $\hat{I} \xrightarrow{p} I$

[Lafferty and Wasserman]

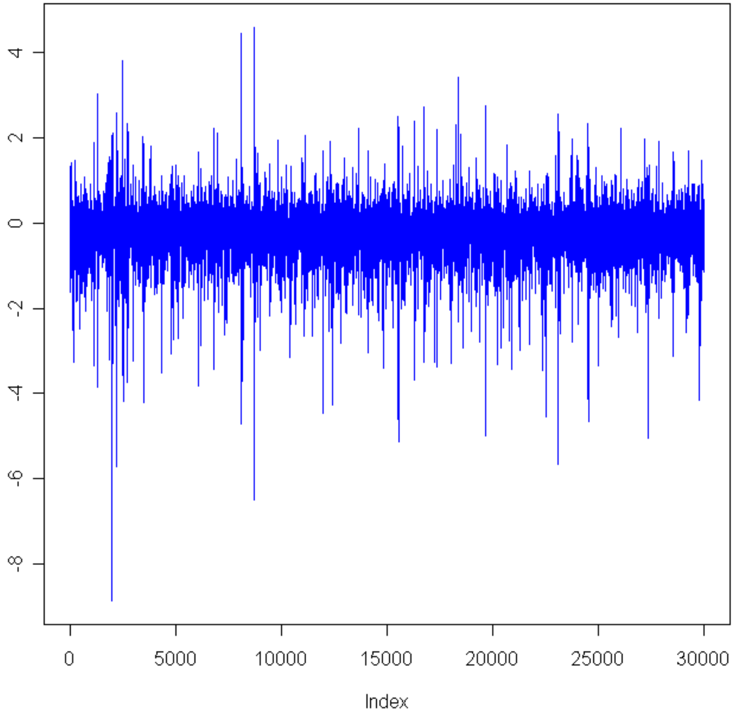
Aside: Monte Carlo Methods

[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)

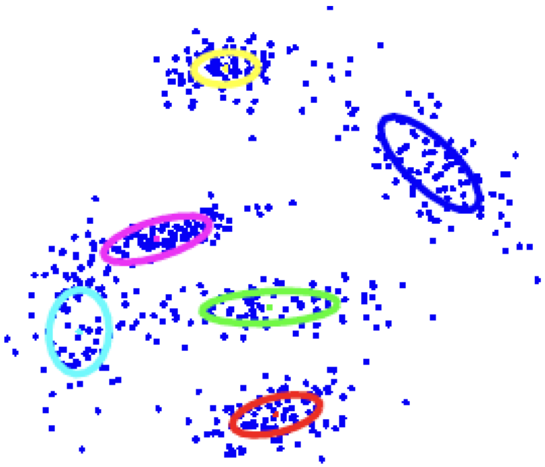


Indian Buffet Process (IBP)

(clustering with Non-Parametric Bayesian Models)

Clustering with Non-Parametric Bayesian Models

Assume: each data point belongs to a cluster:



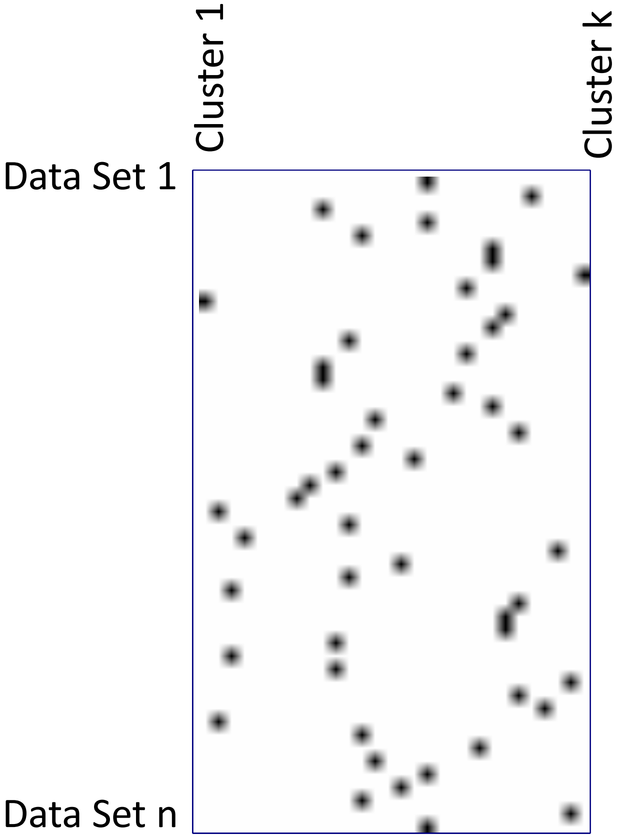
Goals:

- to model the distribution of data;
- to partition data into groups;
- to infer the number of groups

A Classical Approach: mixture modelling with finitely many components

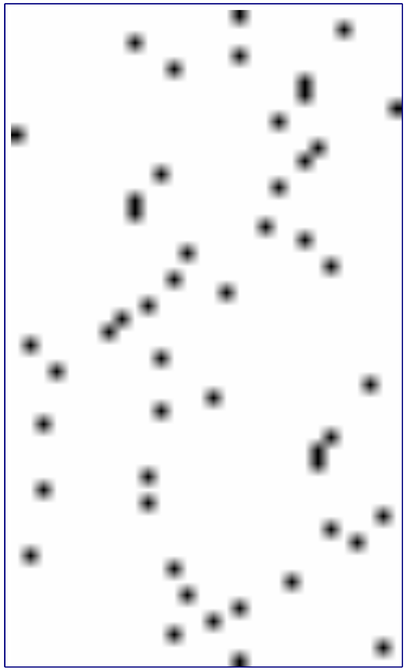
A Bayesian Nonparametric Approach: Dirichlet process mixtures, with countably infinitely many components

A binary matrix representation of data for clustering



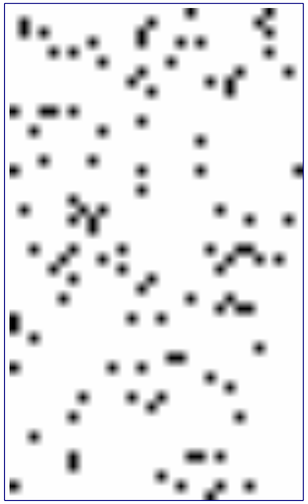
- Rows are data points
- Columns are clusters

A binary matrix representation of data for clustering



- Each data point is assigned to one and only one cluster ➔ rows sum to one.
- Parametric Model: Finite mixture models: number of columns is finite
- Non-Parametric Model: Dirichlet Process Mixtures (DPM): number of columns is countably infinite
- Note: Chinese Restaurant Process (CRP) is the distribution on partitions of the data by a DPM. Thus, we can think of the CRP as a distribution on such binary matrices.

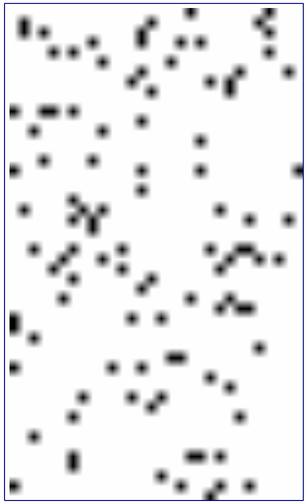
Consider more general distributions on binary matrices



- Rows are data points
- Columns are latent features

- We can think of **infinite** binary matrices where each data point can now have *multiple* features → the rows can sum to more than one.

Consider more general distributions on binary matrices



Therefore:

- there are multiple overlapping clusters
- each data point can belong to several clusters, simultaneously.
- If there are K latent features, then there are 2^K possible settings of the binary latent features for each data point.

Why Considering more general distributions on binary matrices

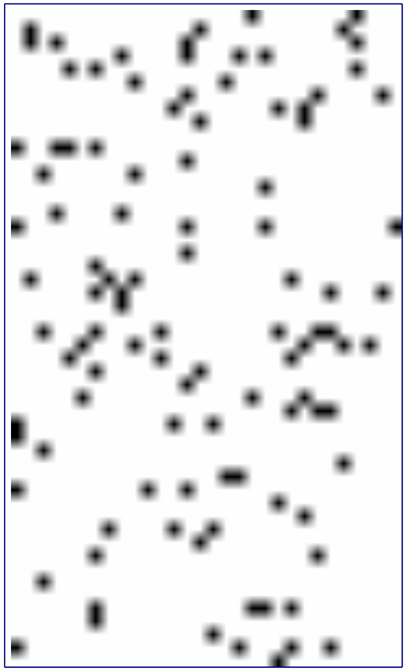
- Many statistical models can be utilized to model data in terms of hidden or latent variables.
- Clustering algorithms (using mixture models) represent data in terms of which cluster each data point belongs to.
- **Issues:**
- Consider modelling people's movie preferences:
    - A movie might be described using features such as "is science fiction", "has Charlton Heston", "was made in the US", "was made in 1970s", "has apes in it"... these features may be unobserved (**latent**).
- The number of potential latent features for describing a movie (or person, news story, image, gene, speech waveform, etc) is **unlimited**.

Solution: From finite to infinite binary matrices

Assume:
$z_{nk} = 1$ means object n has feature k

$z_{nk} \sim \text{Bernoulli}(\theta_k)$

$\theta_k \sim \text{Beta}(\alpha/K, 1)$



- Note that $P(z_{nk} = 1|\alpha) = E(\theta_k) = \frac{\alpha/K}{\alpha/K+1}$ and as K grows larger the matrix gets sparser.
- If $\mathbf{Z}$ is $N \times K$, the expected number of nonzero entries is $N\alpha/(1 + \alpha/K) < N\alpha$.
- Even in the $K \rightarrow \infty$ limit, the matrix is expected to have a finite number of non-zero entries.

Solution: From finite to infinite binary matrices

If we integrate out θ:

$$\begin{aligned} P(\mathbf{Z}|\alpha) &= \int P(\mathbf{Z}|\theta)P(\theta|\alpha)d\theta \\ &= \prod_k \frac{\Gamma(m_k+\frac{\alpha}{K})\Gamma(N-m_k+1)}{\Gamma(\frac{\alpha}{K})} \frac{\Gamma(1+\frac{\alpha}{K})}{\Gamma(N+1+\frac{\alpha}{K})} \end{aligned}$$

The conditional feature assignments are:

$$\begin{aligned} P(z_{nk}=1|\mathbf{z}_{-n,k}) &= \int_0^1 P(z_{nk}|\theta_k)p(\theta_k|\mathbf{z}_{-n,k})\ d\theta_k \\ &= \frac{m_{-n,k}+\frac{\alpha}{K}}{N+\frac{\alpha}{K}} \end{aligned}$$

where $\mathbf{z}_{-n,k}$ is the set of assignments of all objects, not including n, for feature k, and $m_{-n,k}$ is the number of objects having feature k, not including n. We can take limit as $K \rightarrow \infty$.

Solution: From finite to infinite binary matrices

Problem: the probability for any particular matrix goes to zero as $K \to \infty$:

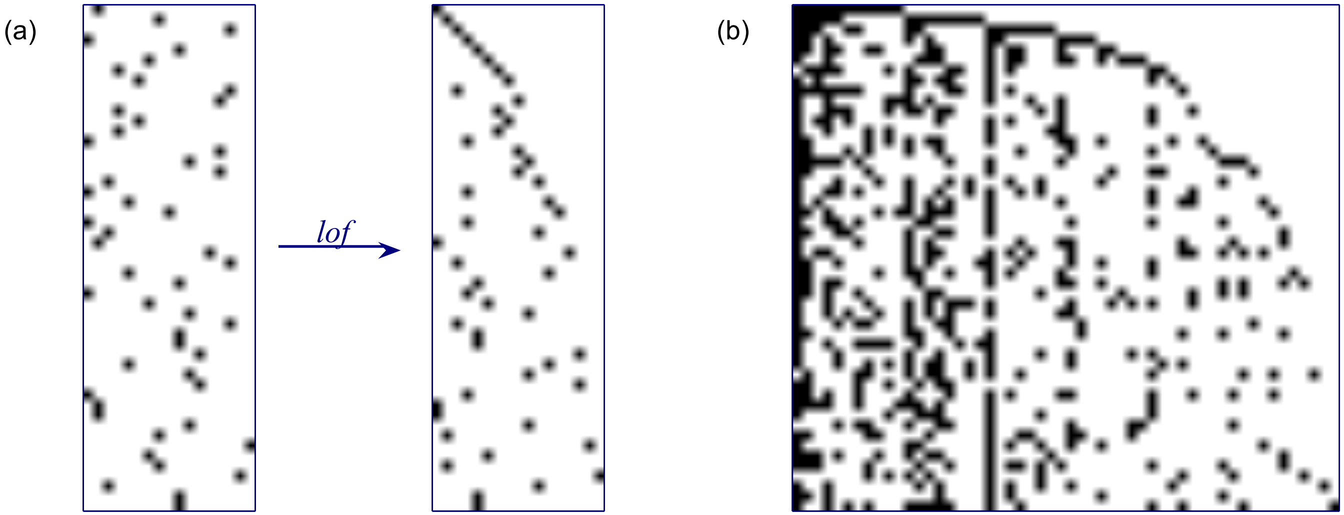
$$\lim_{K\to\infty} P(\mathbf{Z}|\alpha) \quad = \quad 0$$

However, if we consider equivalence classes of matrices in left-ordered form (lof) obtained by reordering the columns: $[\mathbf{Z}] = lof(\mathbf{Z})$:

$$\lim_{K\to\infty} P([\mathbf{Z}]|\alpha) = \exp\{-\alpha H_N\}\frac{\alpha^{K_+}}{\prod_{h>0} K_h!}\prod_{k\leq K_+}\frac{(N-m_k)!(m_k-1)!}{N!}$$

- K_+ is the number of features assigned (i.e. non-zero columns).
- H_N is the N th harmonic number.
- K_h are the number of features with history h.
- This distribution is **infinitely exchangeable**, i.e. it is not affected by the ordering on objects. This is important for its use as a prior in settings where the objects have no natural ordering.

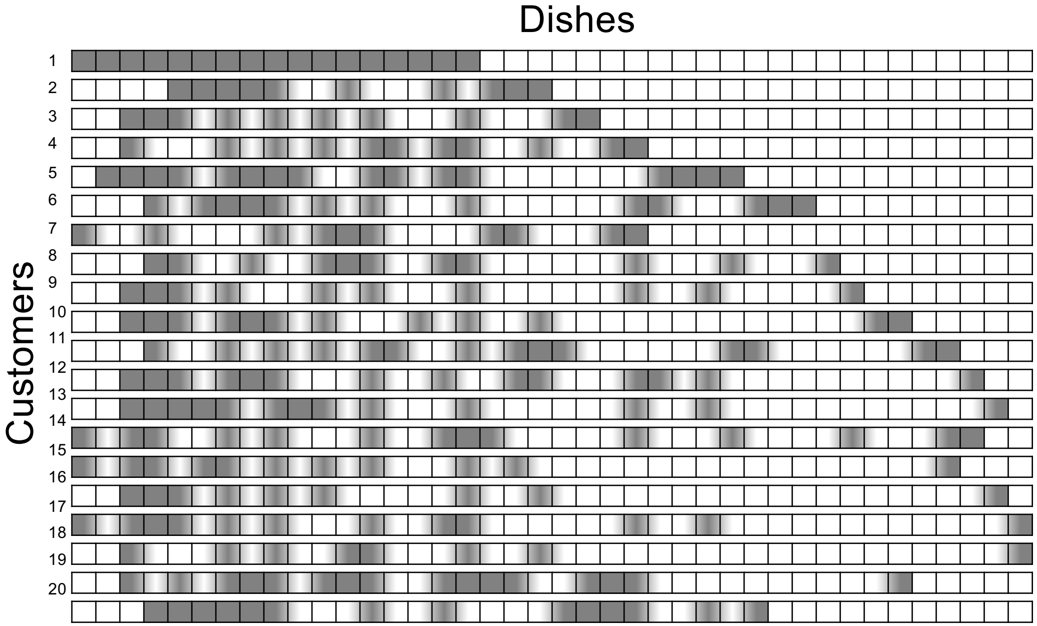
Binary matrices in left-ordered form



(a) The matrix on the left is transformed into the matrix on the right by the function *lof* (). The resulting left-ordered matrix was generated from a Chinese restaurant process (CRP) with *α* = 10.

(b) A left-ordered feature matrix. This matrix was generated from the prior on infinite binary matrices with *α* = 10.

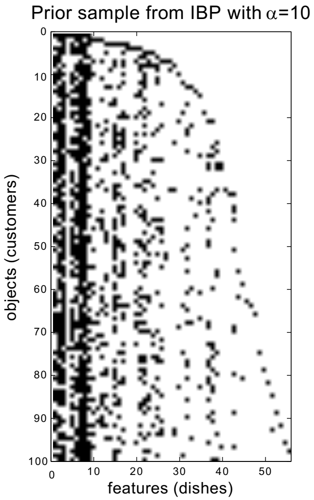
The Indian buffet process (IBP)

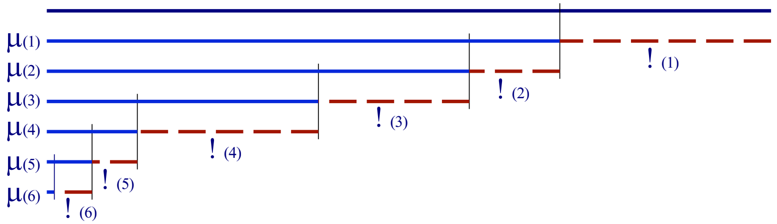


- First customer starts at the left of the buffet, and takes a serving from each dish, stopping after a Poisson(α) number of dishes as her plate becomes overburdened.
- The *n*th customer moves along the buffet, sampling dishes in proportion to their popularity, serving himself with probability m_k / n, and trying a Poisson(α/n) number of new dishes.
- The customer-dish matrix is our feature matrix, **Z**.

Properties of the Indian buffet process

$$P([\mathbf{Z}]|\alpha) = \exp\{-\alpha H_N\} \frac{\alpha^{K_+}}{\prod_{h>0} K_h!} \prod_{k \leq K_+} \frac{(N-m_k)!(m_k-1)!}{N!}$$





Stick-breaking construction for the DP and IBP. The black stick at top has length 1. At each iteration the vertical black line represents the break point. The brown dotted stick on the right is the weight obtained for the DP, while the blue stick on the left is the weight obtained for the IBP.

Shown in (Griffiths and Ghahramani, 2005):

- It is infinitely exchangeable.
- The number of ones in each row is Poisson(α)
- The expected total number of ones is αN .
- The number of nonzero columns grows as $O(\alpha \log N)$.

Additional properties:

- Has a stick-breaking representation (Teh, 2007)
- Can be interpreted using a Beta-Bernoulli process (Thibaux, 2007)

Modelling Data

Latent variable model: let $\mathbf{X}$ be the $N \times D$ matrix of observed data, and $\mathbf{Z}$ be the $N \times K$ matrix of binary latent features

$$P(\mathbf{X}, \mathbf{Z}|\alpha) = P(\mathbf{X}|\mathbf{Z})P(\mathbf{Z}|\alpha)$$

By combining the IBP with different likelihood functions we can get different kinds of models:

- Models for graph structures
- Models for protein complexes
- Models for overlapping clusters
- Models for choice behaviour
- Models for users in collaborative filtering
- Sparse latent factor models

Posterior Inference in IBPs

$$P(Z, \alpha|X) \propto P(X|Z)P(Z|\alpha)P(\alpha)$$

Gibbs sampling: $P(z_{nk} = 1|\mathbf{Z}_{-(nk)}, \mathbf{X}, \alpha) \propto P(z_{nk} = 1|\mathbf{Z}_{-(nk)}, \alpha)P(\mathbf{X}|\mathbf{Z})$

- If $m_{-n,k} > 0$, $P(z_{nk} = 1|\mathbf{z}_{-n,k}) = \frac{m_{-n,k}}{N}$
- For infinitely many k such that $m_{-n,k} = 0$: Metropolis steps with truncation* to sample from the number of new features for each object.
- If α has a Gamma prior then the posterior is also Gamma → Gibbs sample.

Conjugate sampler: assumes that P (**X**|**Z**) can be computed.

Non-conjugate sampler: $P(\mathbf{X}|\mathbf{Z}) = \int P(\mathbf{X}|\mathbf{Z}, \theta)P(\theta)d\theta$ cannot be computed, requires sampling latent θ as well (c.f. (Neal 2000) non-conjugate DPM samplers).

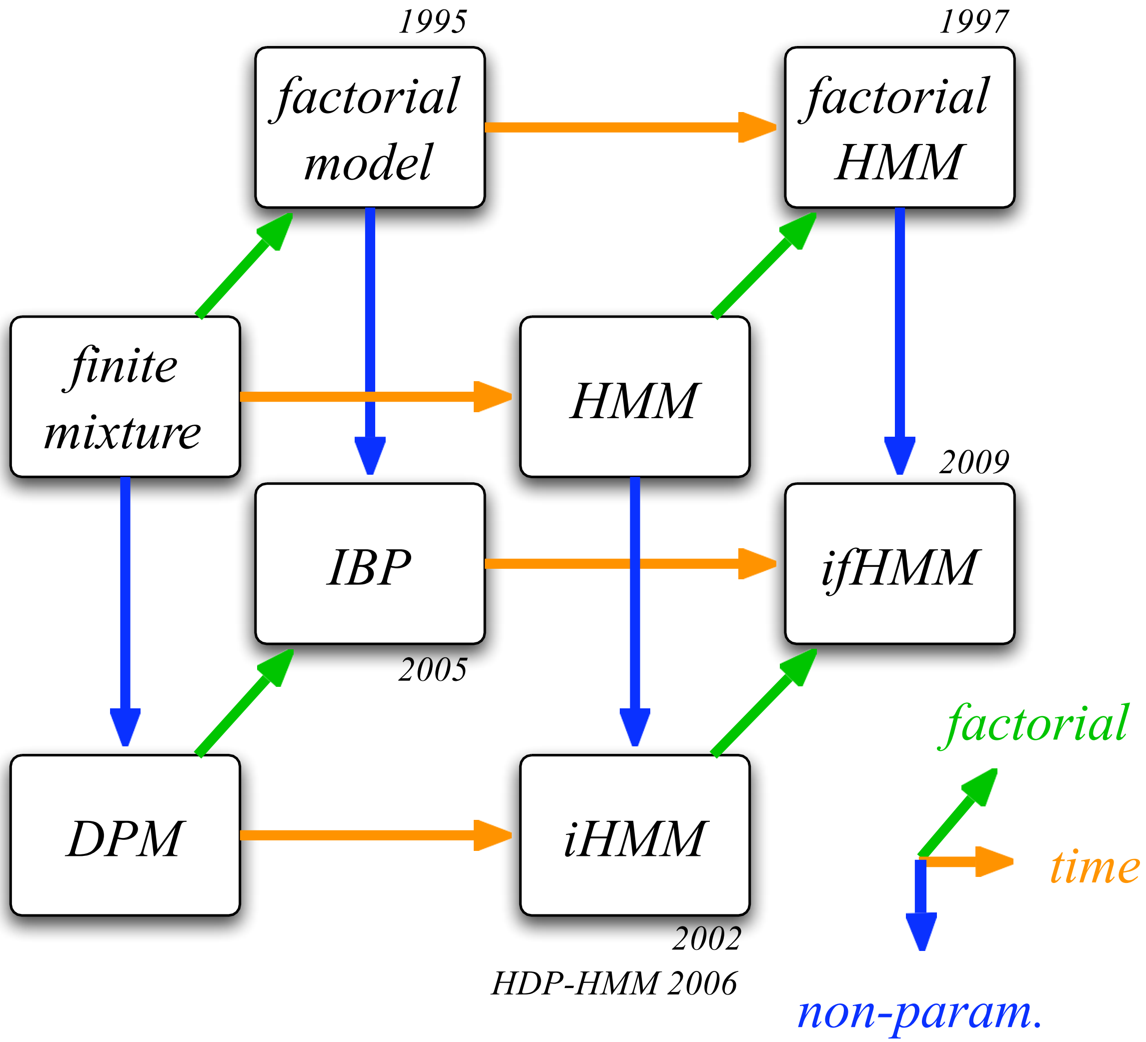
***Slice sampler:** non-conjugate case, is not approximate, and has an adaptive truncation level using a stick-breaking construction of the IBP (Teh, et al, 2007).

Particle Filter: (Wood & Griffiths, 2007).

Accelerated Gibbs Sampling: maintaining a probability distribution over some of the variables (Doshi-Velez & Ghahramani, 2009).

Variational inference: (Doshi-Velez, Miller, van Gael, & Teh, 2009).

Summary


1995
factorial model
1997
factorial HMM
finite mixture
HMM
2009
IBP
ifHMM
2005
DPM
iHMM
2002
HDP-HMM 2006
factorial
time
non-param.