Statistical Machine Learning

Lecture 07
Dirichlet Process (DP)
Chinese Restaurant Process (CRP)
Indian Buffet Process (IBP)

Spring 2021

Sharif University of Technology

Recall: Dirichlet Process

Dirichlet Process Mixtures

The Dirichlet Process (DP)
A distribution on countably infinite discrete probability measures. Sampling yields a ***Polya urn****.*

Chinese Restaurant Process (CRP)
The distribution on partitions induced by a DP prior

Stick-Breaking
An explicit construction for the weights in DP realizations

Infinite Mixture Models
As an infinite limit of finite mixtures with Dirichlet weight priors

Dirichlet Processes: Big Picture

There are many ways to derive the Dirichlet Process:

- Dirichlet distribution
- Urn model
- Chinese restaurant process
- Stick breaking
- Gamma process

Recall

Exchangable Random Variables:

consider a sequence $X_1, X_2, X_3, \ldots$

If joint prob. dist. $p(X_1, X_2, \ldots)$ does not change when the location of sequence ~~locat~~ changes.

$\Rightarrow$ sequence is exchangable

X_1, X_2, X_3, X_4, X_5 $\quad$ X_3, X_5, X_1, X_2, X_4

exchangable

both have same joint dist.

Recall

de Finetti's theorem

exchangeability $\xleftrightarrow{\text{relate}}$ independence

If an identically distributed sequence is independent then the seq. is exchangeable.

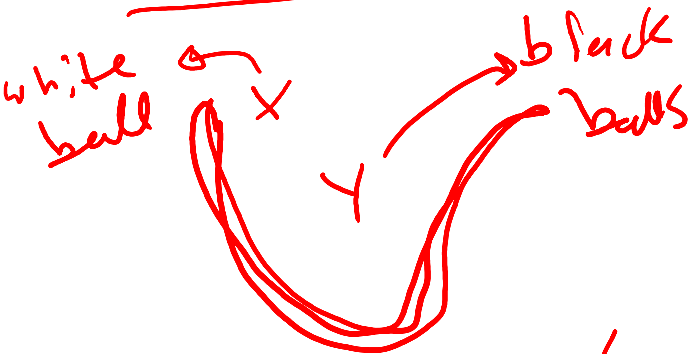
$$X_1, X_2, X_3, X_4, X_5 \quad \text{\& independent}$$

$$N(\mu, \delta^2) \quad N(\mu, \delta^2) \quad \cdots \quad N(\mu, \delta^2)$$

$$\Longrightarrow X_1, X_2, \ldots X_5 \quad \text{exchangable}$$

Recall

Polya urn



rich gets richer

1. one ball drawn randomly look at its color
2. return the ball to the urn
3. add additional ball with same color to the urn
4. repeat

Recall

Dirichlet-multinomial dist.

assume K different color

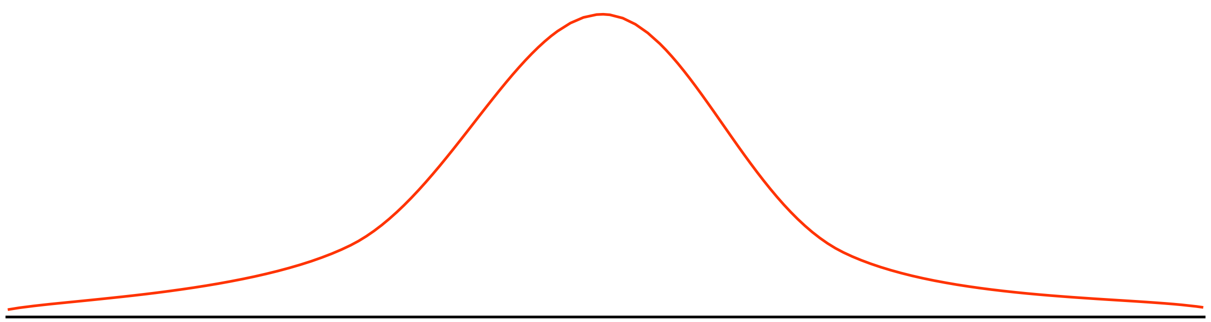
The dist. over the number of balls of each color given n draws

K=2

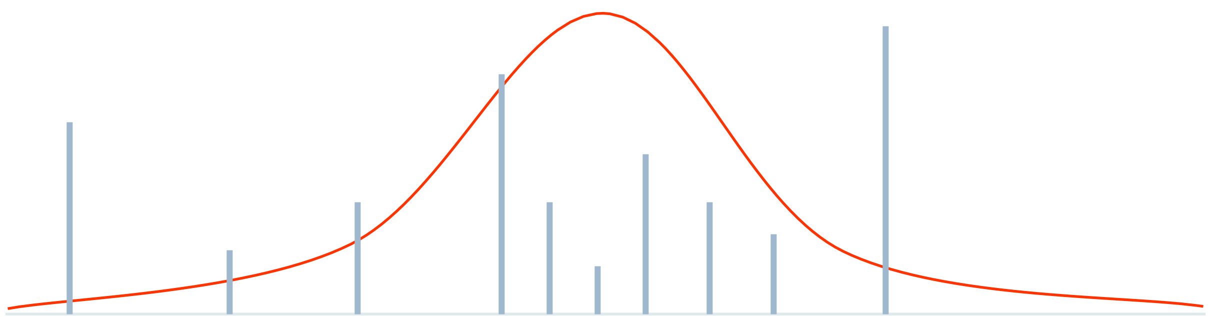
Beta distribution

Dirichlet Process

- Consider Gaussian G_0

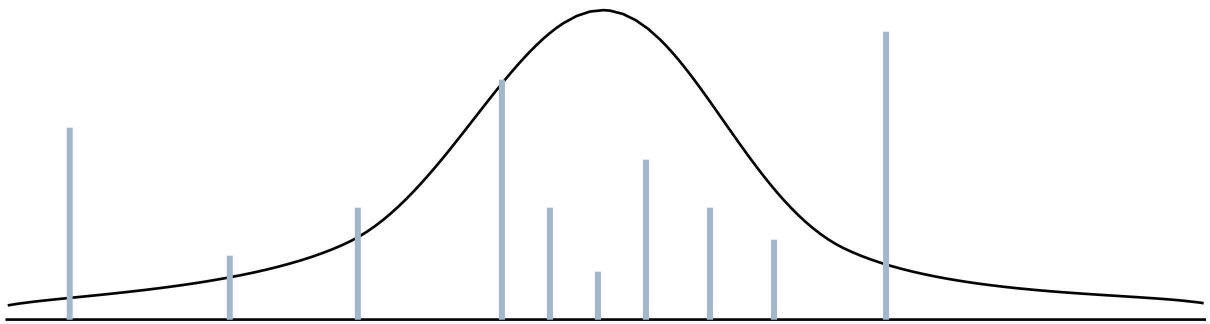


- $G \sim DP(\alpha, G_0)$



Dirichlet Process

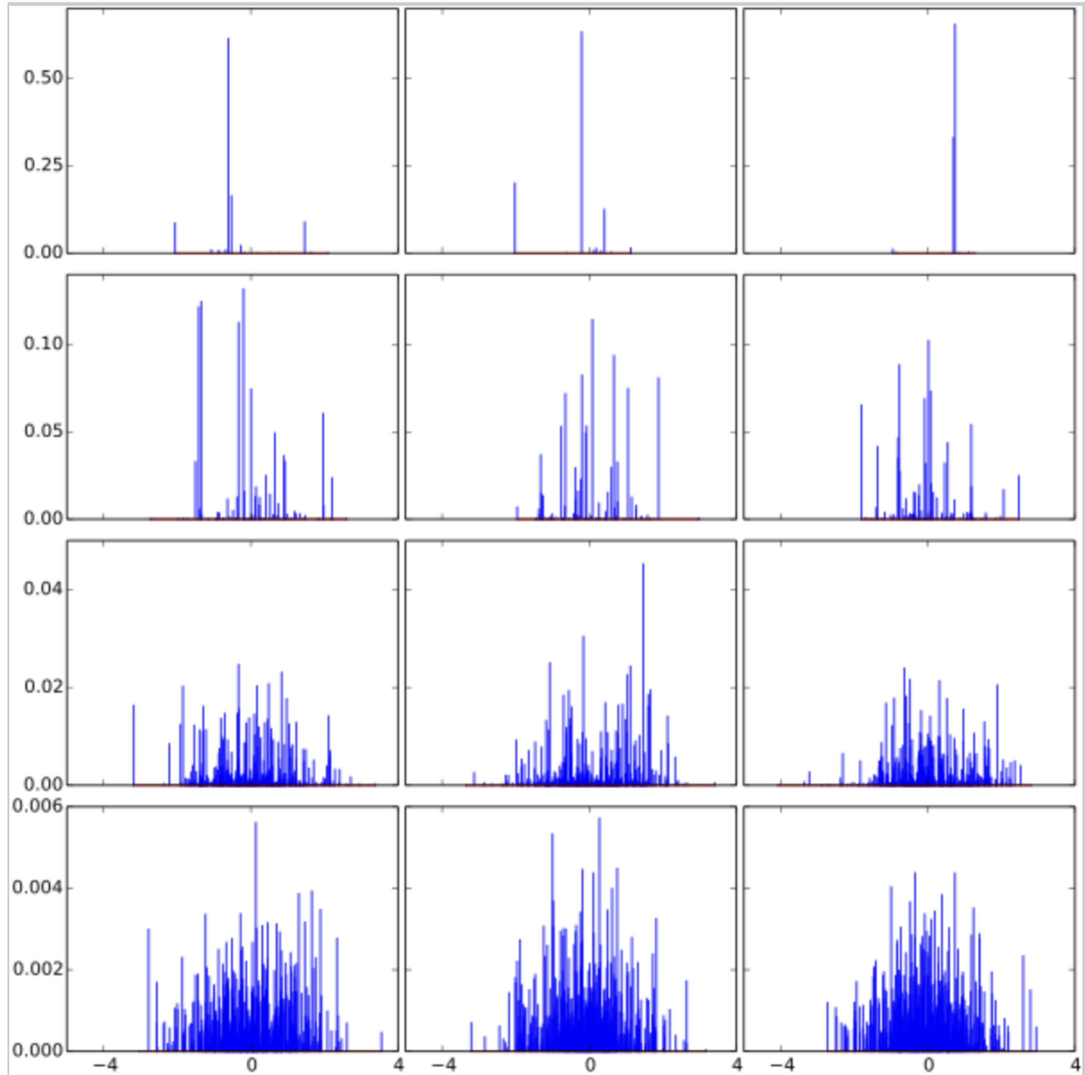
- $G \sim DP(\alpha, G_0)$



- 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.
  - Therefore, there is always a non-zero probability of two samples colliding

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.

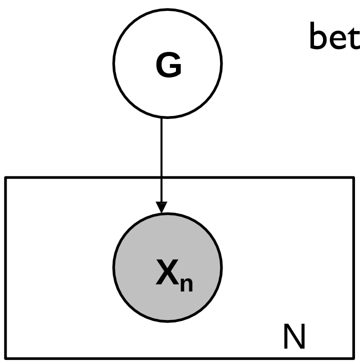


Dirichlet Process

Samples from a Dirichlet Process

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

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



Marginalizing out G introduces dependencies between the X_n variables

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

Dirichlet Process

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:

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

Dirichlet Process

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.

Dirichlet Process

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:

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

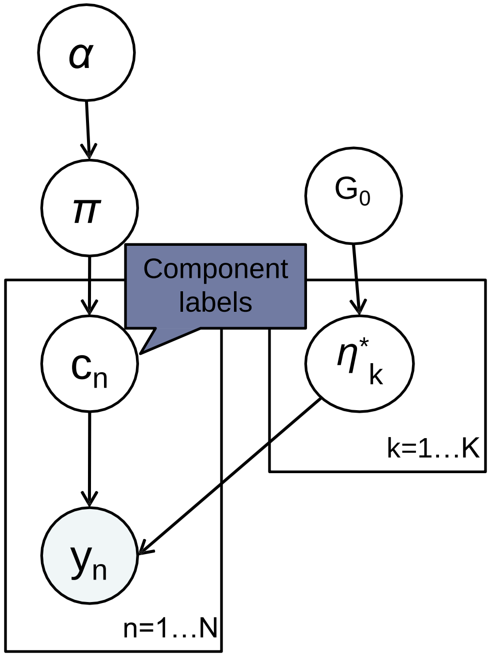
Can rewrite as:

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

Dirichlet Process

Finite Mixture Models

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



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

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

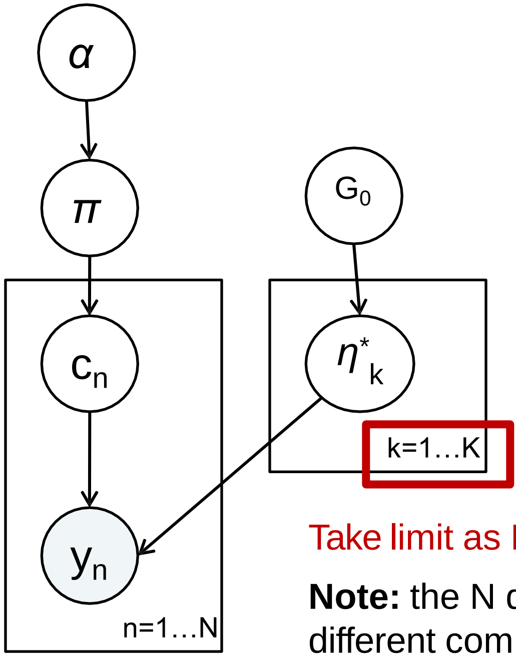
$\eta_k \sim G_0$

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

Dirichlet Process

Infinite Mixture Models

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



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

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

$\eta_k \sim G_0$

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

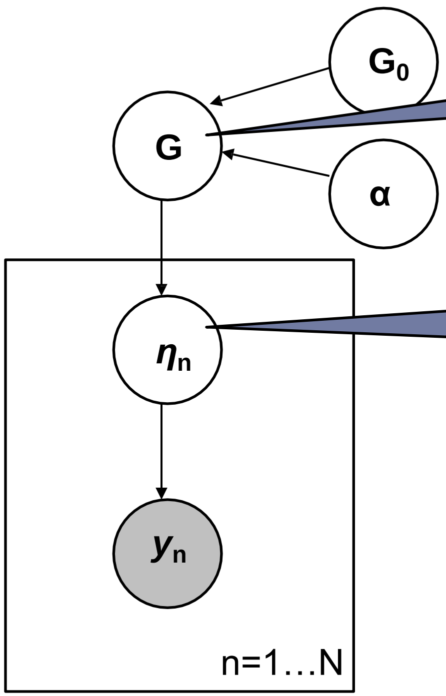
Take limit as K goes to ∞

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

[Rasmussen 2000]

Dirichlet Process

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

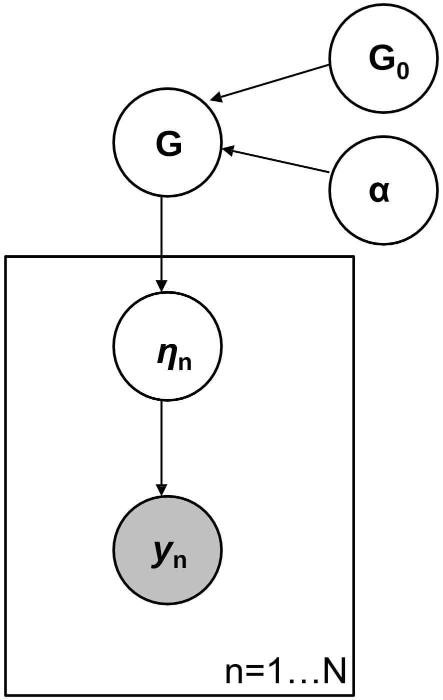
unique values for η_n = # mixture components

Dirichlet Process

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]



Dirichlet Process

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:

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

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

[Lafferty and Wasserman]

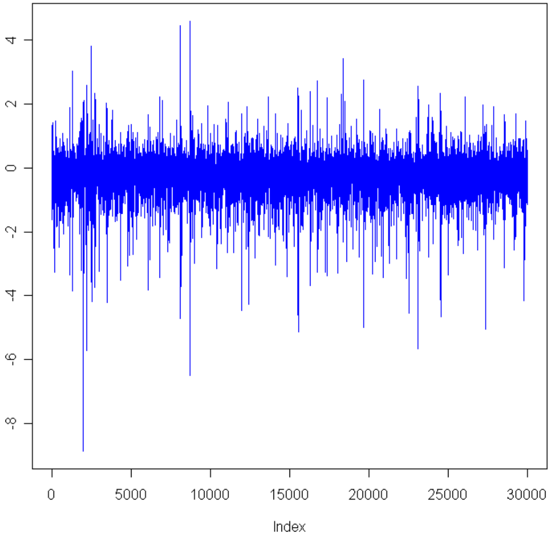
Dirichlet Process

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

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



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

Dirichlet Distribution & Dirichlet Process

- 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.

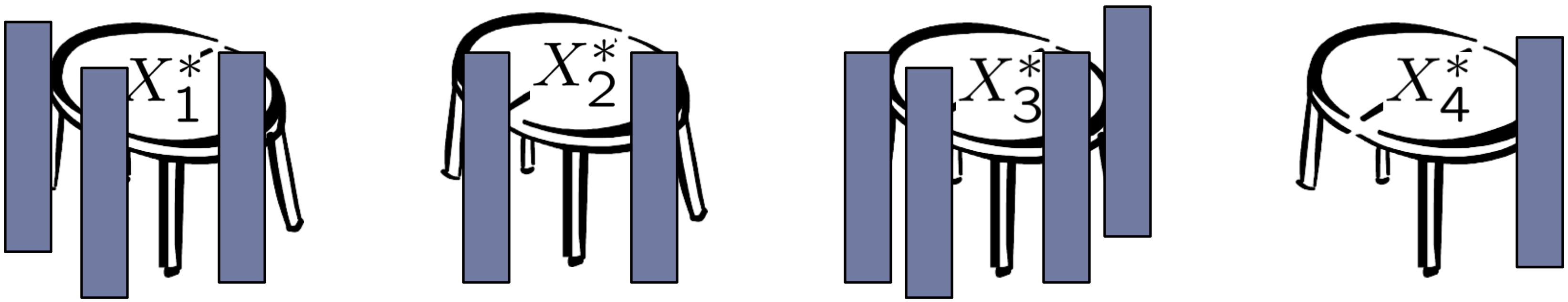
Chinese Restaurant Process (CRP)

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

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

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

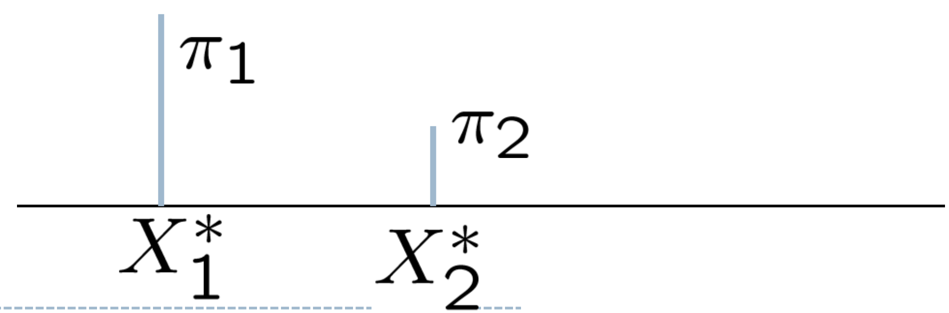
$$X_1^*, X_2^*, \ldots, X_i^*, \ldots \sim G_0$$

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

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

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. **$\pi_2 = v_2(1 - v_1)$**

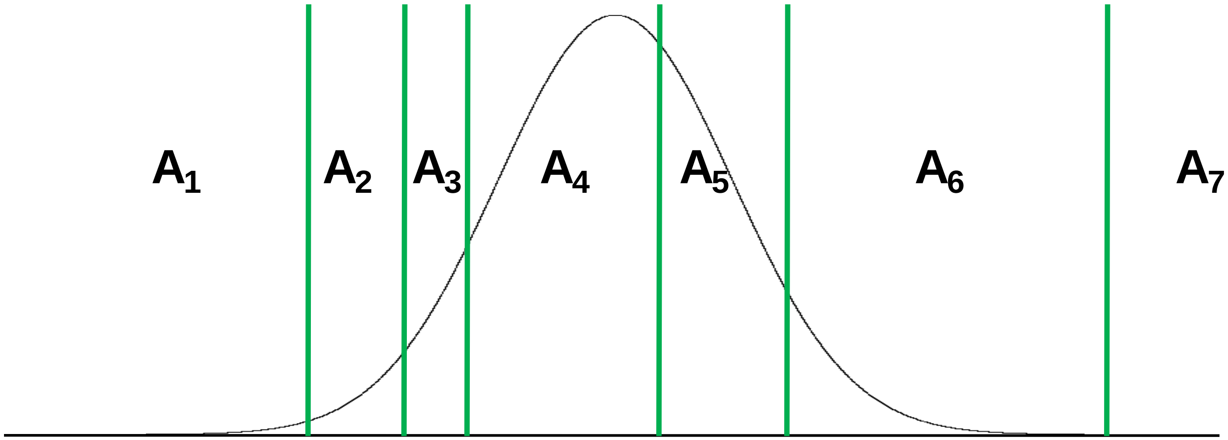
…



Formal Definition (not constructive)

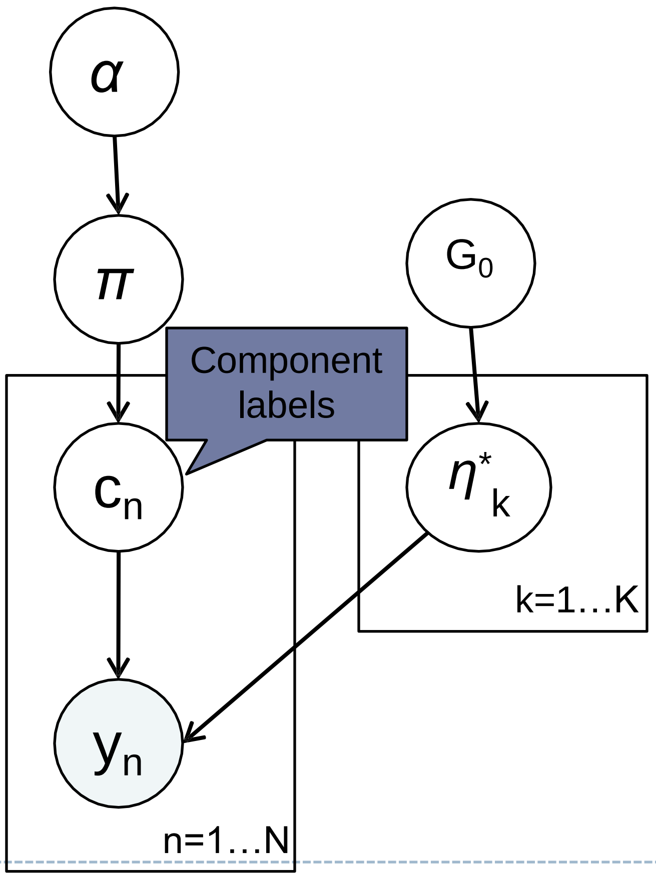
- Let α be a positive, real-valued scalar
- Let G_0 be a probability distribution over support set A
- If G ~ DP(α, G_0), then for any finite set of partitions $A_1 \cup A_2 \cup \ldots \cup A_k$ of A:

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



Finite Mixture Models

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



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

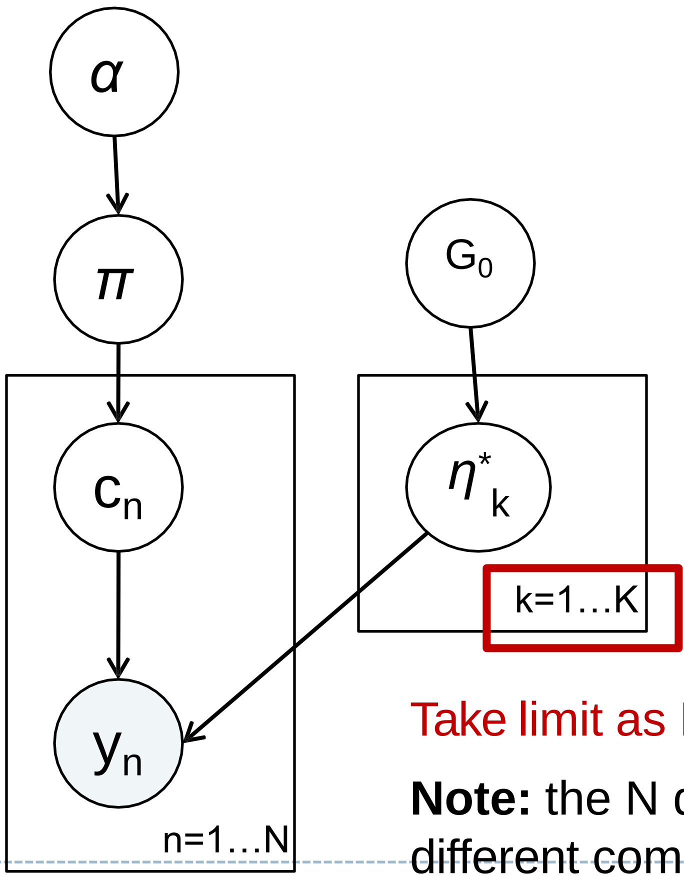
$c_n \sim \text{Multinomial}(\pi)$

$\eta_k \sim G_0$

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

Infinite Mixture Models

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



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

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

$$\eta_k \sim G_0$$

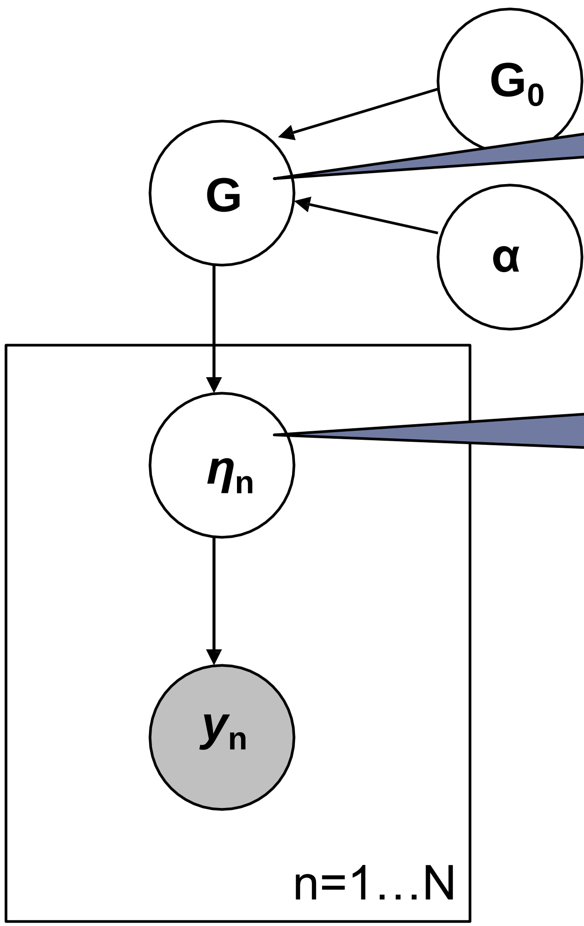
$$y_n \mid c_n, \eta_1, \ldots \eta_K \sim F(\,\cdot \mid \eta_{c_n})$$

Take limit as K goes to ∞

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

[Rasmussen 2000]

Dirichlet Process Mixture



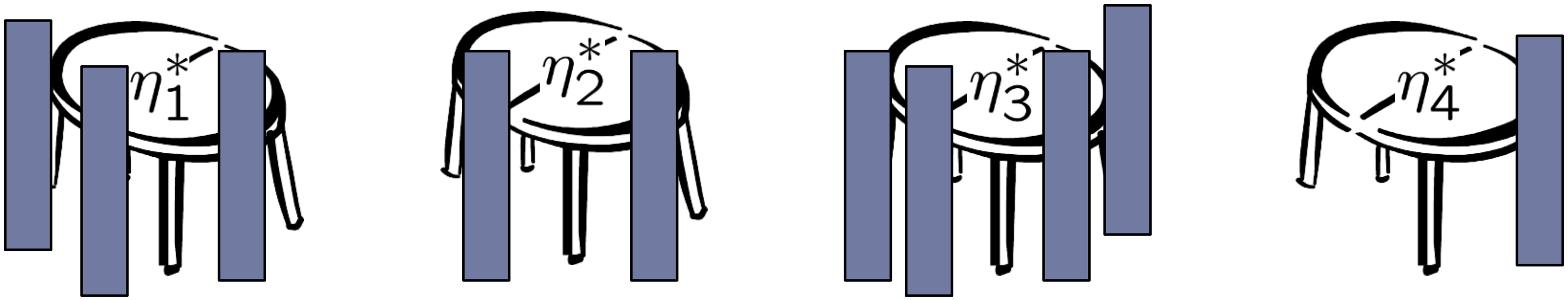
countably infinite number of point masses

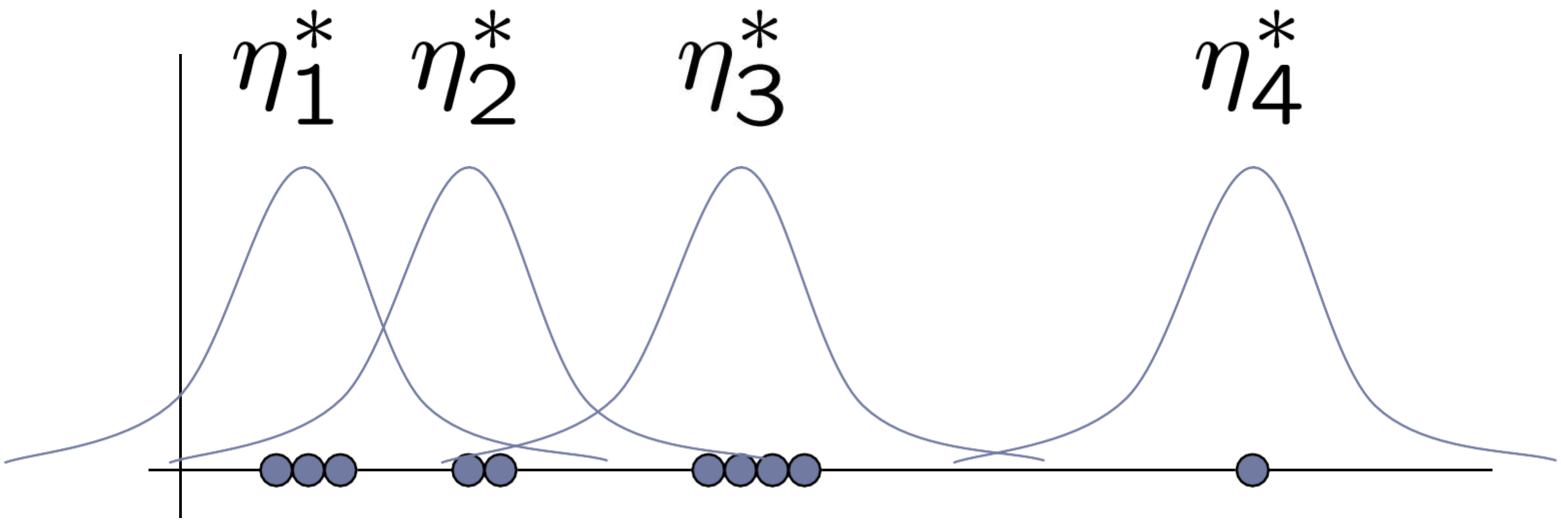
draw N times from G to get parameters for different mixture components

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



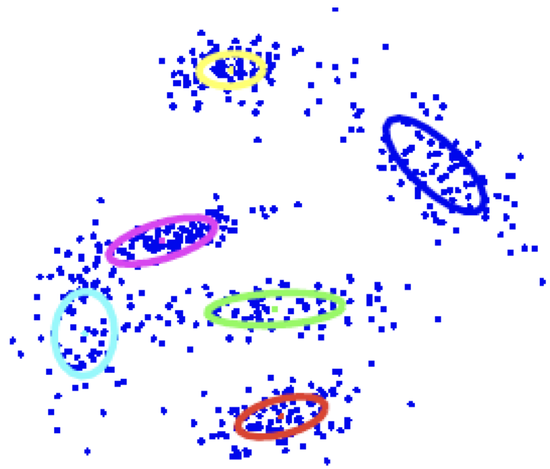


Indian Buffet Process (IBP)

(clustering with Non-Parametric Bayesian Models)

Clustering with Non-Parametric Bayesian Models

Assume: each data point belongs to a cluster:



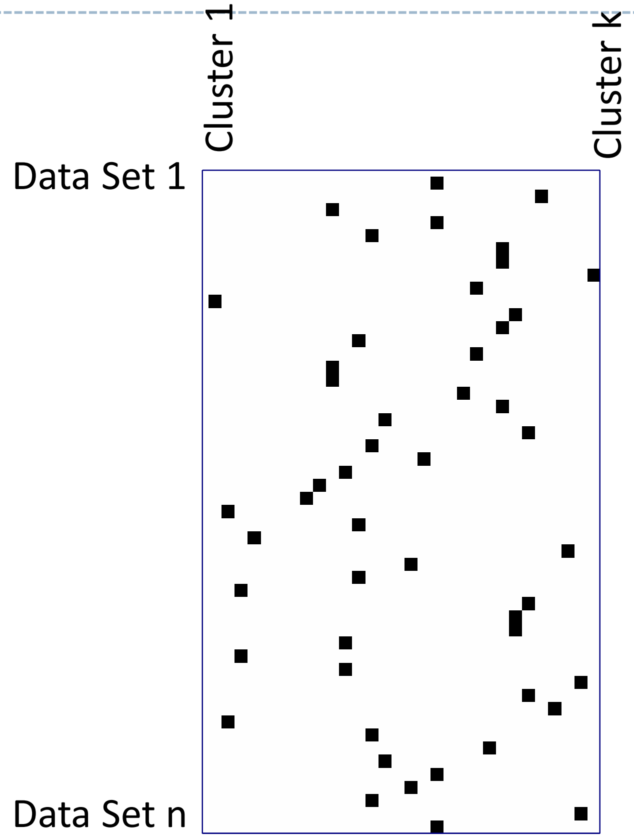
Goals:

- 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

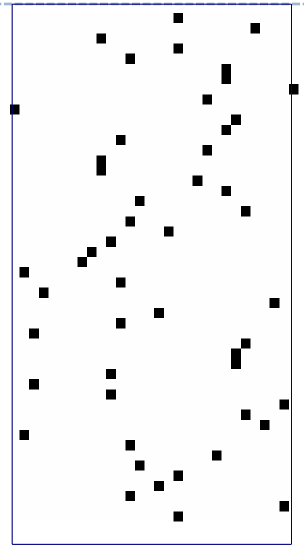
A Bayesian Nonparametric Approach: Dirichlet process mixtures, with countably infinitely many components

A binary matrix representation of data for clustering



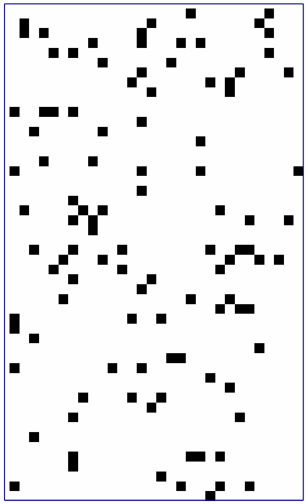
- Rows are data points
- Columns are clusters

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.

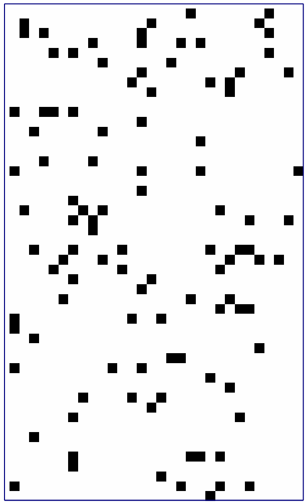
Consider more general distributions on binary matrices



- 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



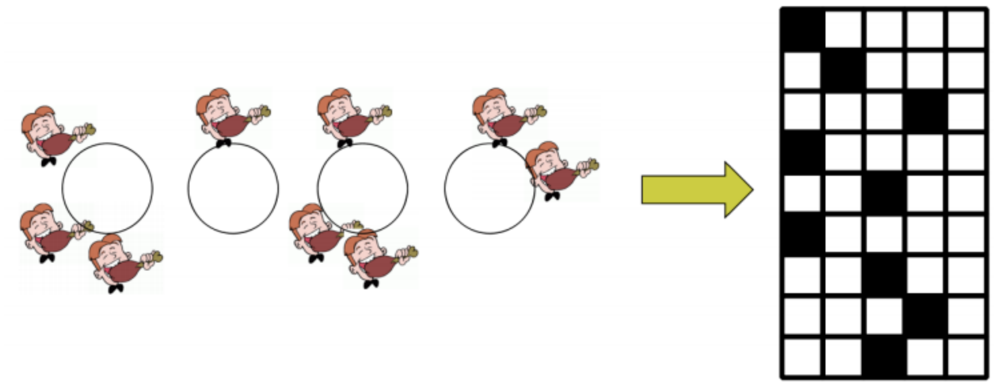
Therefore:

- 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

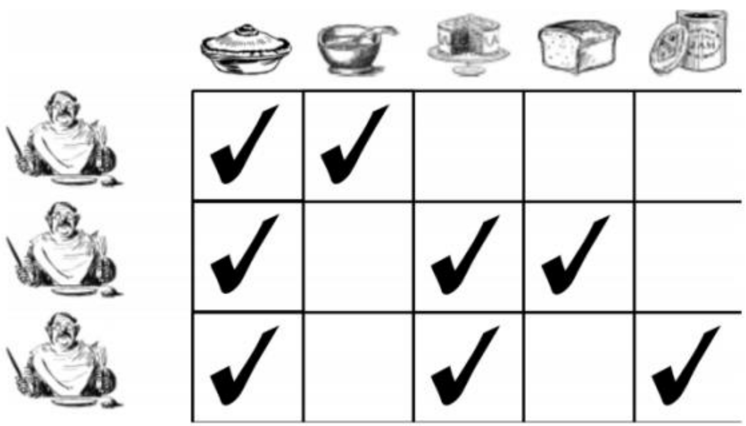
- Many statistical models can be utilized to model data in terms of hidden or latent variables.
- Clustering algorithms (using mixture models) represent data in terms of which cluster each data point belongs to.
- **Issues:**
- Consider modelling people's movie preferences:
  - A movie might be described 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**.

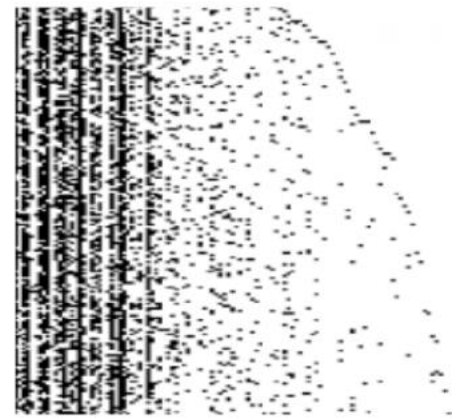
Recall CRP



- Rows are data points
- Columns are clusters
- Rows add up to 1
- Each Data belongs to only 1 cluster

In Summary: Indian Buffet Process





- Rows are data points
- Columns are clusters
- Rows may add up to more than 1
- Each Data may belongs to more than 1 cluster

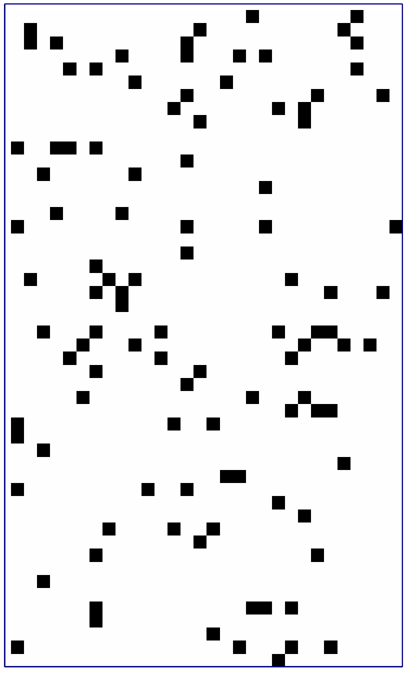
Solution: Finite to infinite binary matrices

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 $K \to \infty$ limit, the matrix is expected to have a finite number of non-zero entries.