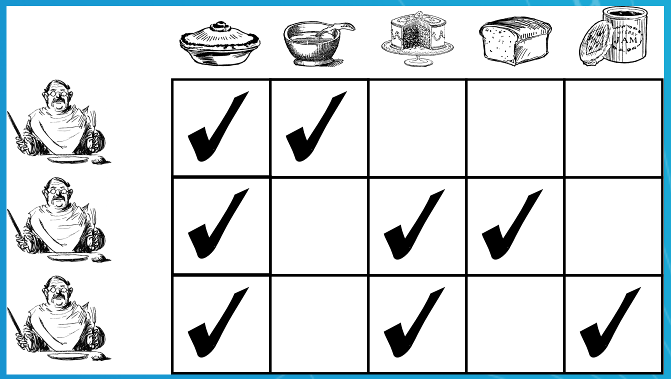
Carnegie Mellon University
School of Computer Science

Probabilistic Graphical Models

Bayesian Nonparametrics: Indian Buffet Process

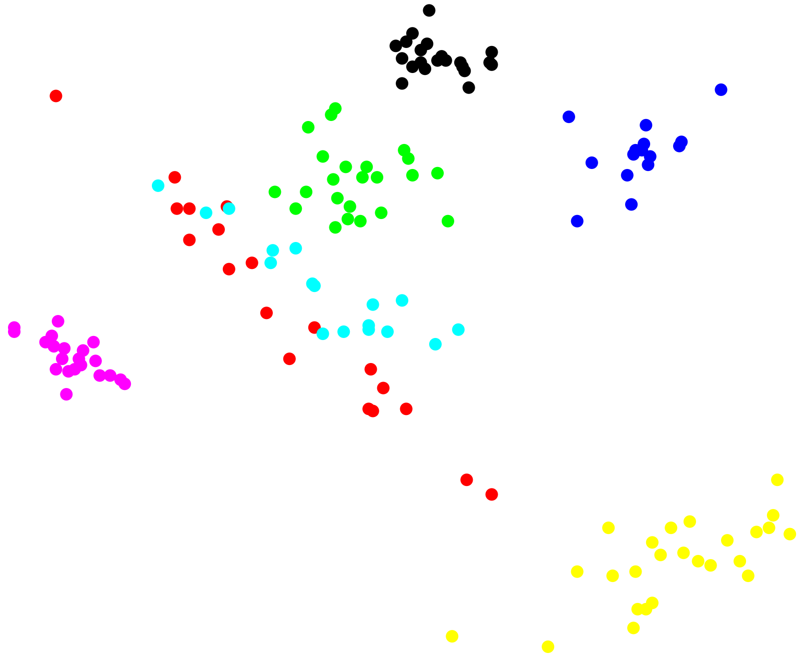
Eric Xing
Lecture 24, April 15, 2020

Reading: see class homepage



Recap of last lecture

- Dirichlet process: a distribution over discrete probability distributions with infinitely many atoms.
- Can be used to create a *nonparametric* version of a *finite mixture model*.



Recap of last lecture

- We can think of the Dirichlet process in a number of ways:
  - The **infinite limit** of a Dirichlet distribution.
  - A rich-gets-richer predictive distribution over the next data point (**Chinese restaurant process, Polya urn scheme**).
  - An iterative procedure for generating samples from the Dirichlet process – the **stick breaking representation**.

Limitations of a simple mixture model

- The Dirichlet distribution and the Dirichlet process are great if we want to cluster data into non-overlapping clusters.
- However, DP/Dirichlet mixture models cannot share features (i.e., cluster centroids, prototypes) between clusters.
- In many applications, data points exhibit properties of multiple latent features
  - Images contain multiple objects.
  - Actors in social networks belong to multiple social groups.
  - Movies contain aspects of multiple genres.

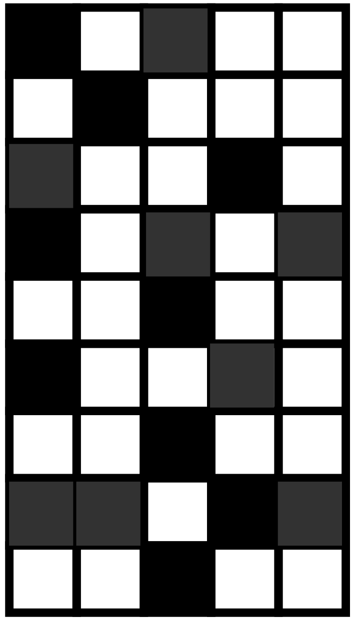
Latent variable models

- Latent variable models allow each data point to exhibit *multiple latent features*, to *varying degrees*.
- Example: Factor analysis

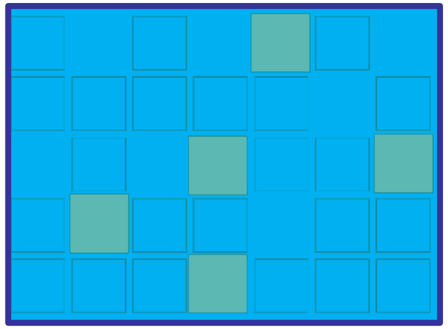
  $$\mathbf{X} = \mathbf{W}\mathbf{A}^{\top} + \varepsilon$$

  - Rows of **A** = latent features
  - Rows of **W** = data-point-specific weights for these features
  - ε = Gaussian noise.
- Example: LDA
  - Each document represented by a *mixture* of features.

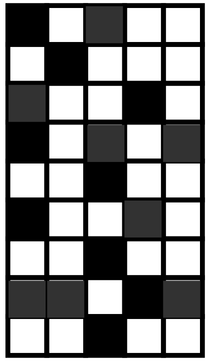
W

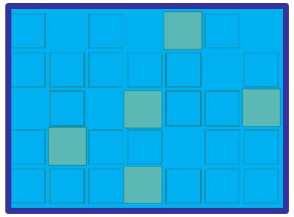


A



Infinite latent feature models





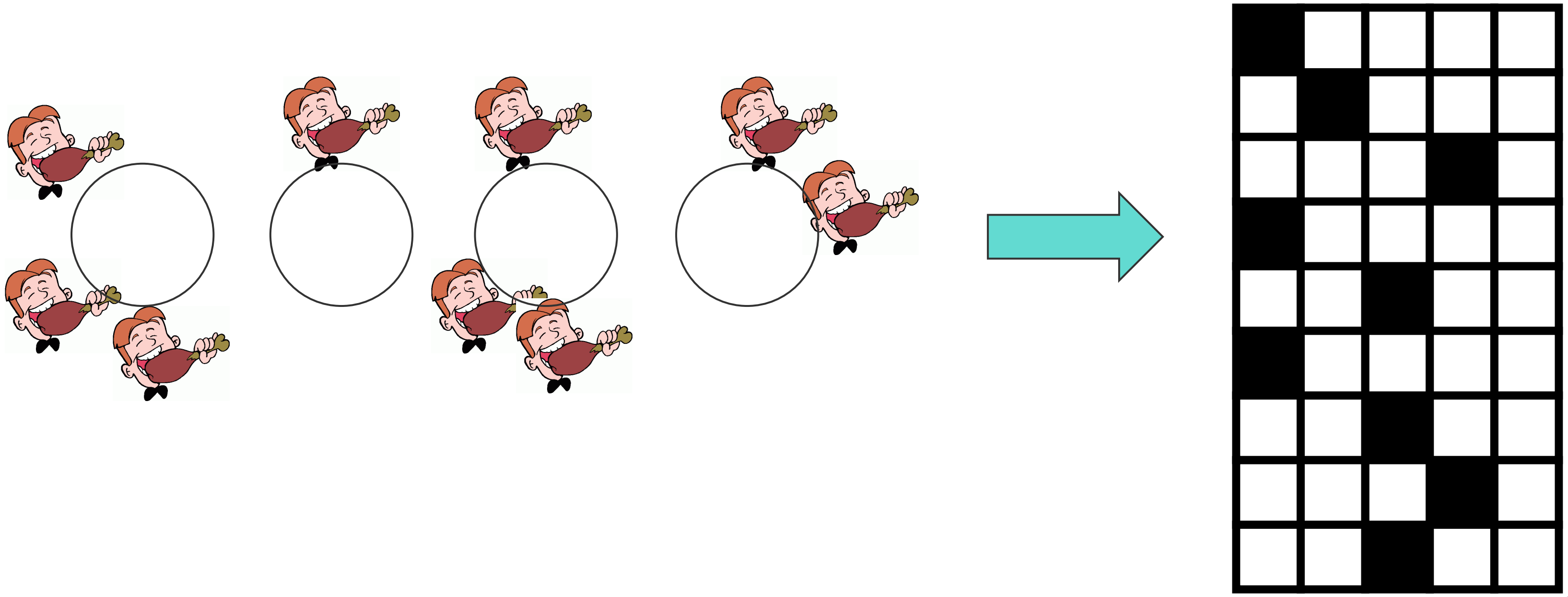
- Problem: How to choose the number of features?
- Example: Factor analysis

  $\mathbf{X} = \mathbf{W}\mathbf{A}^{\top} + \varepsilon$
- Each column of **W** (and row of **A**) corresponds to a feature.
- Question: Can we make the number of features *unbounded* a posteriori, as we did with the DP?
- Solution: allow *infinitely many* features a priori – i.e. let **W** (or **A**) have infinitely many columns (rows).
- Problem: We can't represent infinitely many features!
- Solution: make our infinitely large matrix *sparse*.

Griffiths and Ghaharamani, 2006

Recall the CRP: a distribution over indicator matrices

- Recall that the CRP gives us a distribution over *partitions* of our data.
- We can represent this as a distribution over *binary* (***indicator***) *matrices*, where each row (which is a "one-hot vector") corresponds to a data point, and each column to a cluster.



A sparse, finite latent variable model

- We want a *sparse* model – so let

$$\mathbf{X} = \mathbf{W}\mathbf{A}^T + \epsilon$$

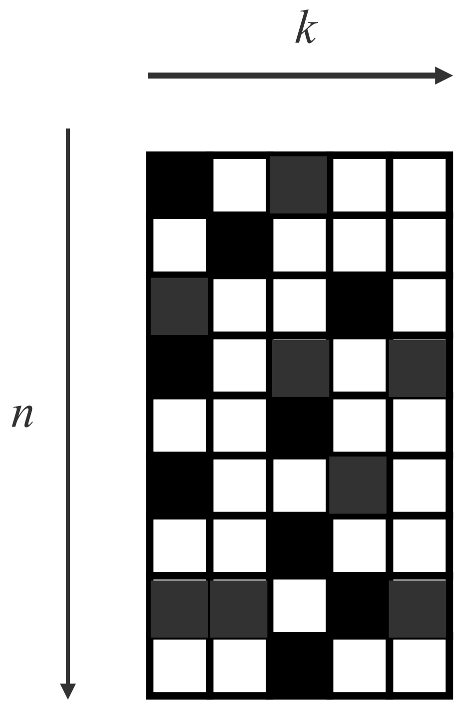
$$\mathbf{W} = \mathbf{Z} \odot \mathbf{V}$$

  for some sparse matrix **Z**.

- Place a *beta-Bernoulli prior* on **Z**:

$$\pi_k \sim \mathrm{Beta}\left(\frac{\alpha}{K}, 1\right), k = 1, \ldots, K$$

$$z_{nk} \sim \mathrm{Bernoulli}(\pi_k), n = 1, \ldots, N.$$



A sparse, finite latent variable model

- If we integrate out the π_k, the marginal probability of a matrix **Z** is:

$$p(\mathbf{Z}) = \prod_{k=1}^{K} \int \left(\prod_{n=1}^{N} p(z_{nk}|\pi_k) \right) p(\pi_k) d\pi_k$$

$$= \prod_{k=1}^{K} \frac{B(m_k + \alpha/K, N - m_k + 1)}{B(\alpha/K, 1)}$$

$$= \prod_{k=1}^{K} \frac{\alpha}{K} \frac{\Gamma(m_k + \alpha/K)\Gamma(N - m_k + 1)}{\Gamma(N + 1 + \alpha/K)}$$

where

$$m_k = \sum_{n=1}^{N} z_{nk}$$

- This is *exchangeable* (doesn't depend on the order of the rows or columns)

A sparse, finite latent variable model

- If we integrate out the π_k, the marginal probability of a matrix **Z** is:

$$p(\mathbf{Z}) = \prod_{k=1}^{K} \int \left(\prod_{n=1}^{N} p(z_{nk}|\pi_k) \right) p(\pi_k) d\pi_k$$

$$= \prod_{k=1}^{K} \frac{B(m_k + \alpha/K, N - m_k + 1)}{B(\alpha/K, 1)}$$

$$= \prod_{k=1}^{K} \frac{\alpha}{K} \frac{\Gamma(m_k + \alpha/K)\Gamma(N - m_k + 1)}{\Gamma(N + 1 + \alpha/K)}$$

where

$$m_k = \sum_{n=1}^{N} z_{nk}$$

- How is this sparse?

m_k! (N – m_k)! / (N + 1)!
= 1 * 2 * .. * (N – m_k) / (m_k + 1) * (m_k + 2) *… * (N + 1)
= (1 / (m_k + 1)) * (2 / (m_k + 2)) * … * ((N – m_k) / (N + 1))

An equivalence class of matrices

- We can naively take the infinite limit by taking K to infinity
- Because all the columns are equal in expectation, as K grows we are going to have more and more empty columns.
- We do not want to have to represent infinitely many empty columns!
- Define an *equivalence class* $[\mathbf{Z}]$ of matrices where the non-zero columns are all to the left of the empty columns.
- Let *lof(.)* be a function that maps binary matrices to *left-ordered* binary matrices – matrices ordered by the binary number made by their rows.

Left-ordered matrices

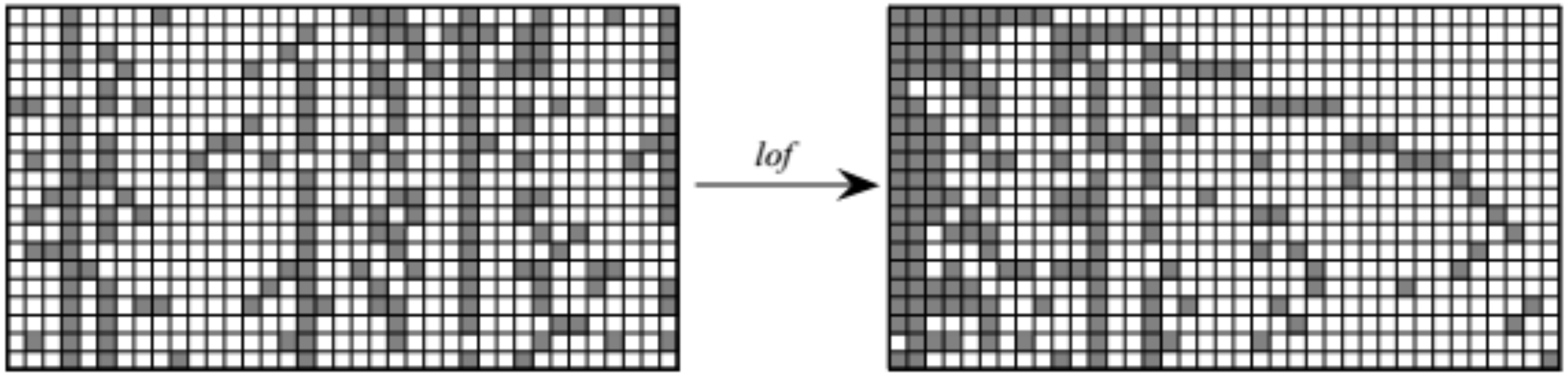
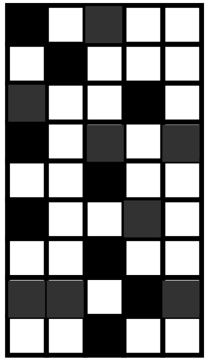


Figure 5: Binary matrices and the left-ordered form. The binary matrix on the left is transformed into the left-ordered binary matrix on the right by the function $lof(\cdot)$. This left-ordered matrix was generated from the exchangeable Indian buffet process with $\alpha = 10$. Empty columns are omitted from both matrices.

Image from Griffiths and Ghahramani, 2011

How big is the equivalence set?



- All matrices in the equivalence set [**Z**] are equiprobable (by exchangeability of the columns), so if we know the size of the equivalence set, we know its probability.
- Call the vector *(z_{1k},$z_{2,k}$,…,$z_{(n-1)k}$)* the *history* of feature *k* at data point *n* (a number represented in binary form).
- Let K_h be the number of features possessing history *h*, and let K_+ be the total number of features with non-zero history.
- The total number of lof-equivalent matrices in [**Z**] is

$$\binom{K}{K_0 \cdots K_{2^N-1}} = \frac{K!}{\prod_{n=0}^{2^N-1} K_n!}$$

Probability of an equivalence class of finite binary matrices.

- If we know the size of the equivalence class $[\mathbf{Z}]$, we can evaluate its probability:

$$
\begin{aligned}
p([\mathbf{Z}]) &= \sum_{\mathbf{Z}\in[\mathbf{Z}]} p(\mathbf{Z}) \\
&= \frac{K!}{\prod_{n=0}^{2^N-1} K_n!} \prod_{k=1}^{K} \frac{\alpha}{K} \frac{\Gamma(m_k+\alpha/K)\Gamma(N-m_k+1)}{\Gamma(N+1+\alpha/K)} \\
&= \frac{\alpha^{K_+}}{\prod_{n=1}^{2^N-1} K_n!} \frac{K!}{K_0!K^{K_+}} \left(\frac{N!}{\prod_{j=1}^{N} j+\alpha/K} \right)^K \\
&\quad \cdot \prod_{k=1}^{K_+} \frac{(N-m_k)! \prod_{j=1}^{m_k-1}(j+\alpha/K)}{N!}
\end{aligned}
$$

Taking the infinite limit

- We are now ready to take the limit of this finite model as K tends to infinity:

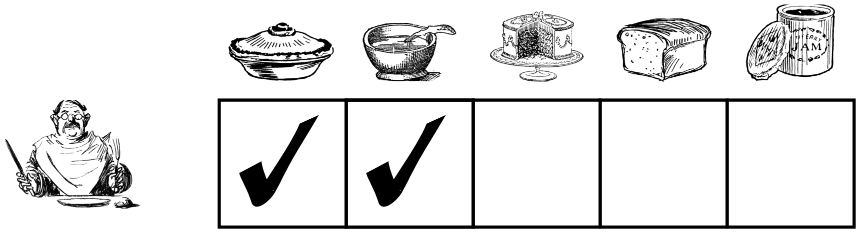
$$\frac{\alpha^{K_+}}{\prod_{n=1}^{2^N-1} K_n!} \frac{K!}{K_0! K^{K_+}} \left(\frac{N!}{\prod_{j=1}^{N} j + \frac{\alpha}{K}} \right)^K \prod_{k=1}^{K_+} \frac{(N-m_k)! \prod_{j=1}^{m_k-1} (j + \frac{\alpha}{K})}{N!}$$

$$\downarrow K \to \infty$$

$$\frac{\alpha^{K_+}}{\prod_{n=1}^{2^N-1} K_n!} \qquad 1 \qquad \exp\{-\alpha H_N\} \qquad \prod_{k=1}^{K_+} \frac{(N-m_k)!(m_k-1)!}{N!}$$

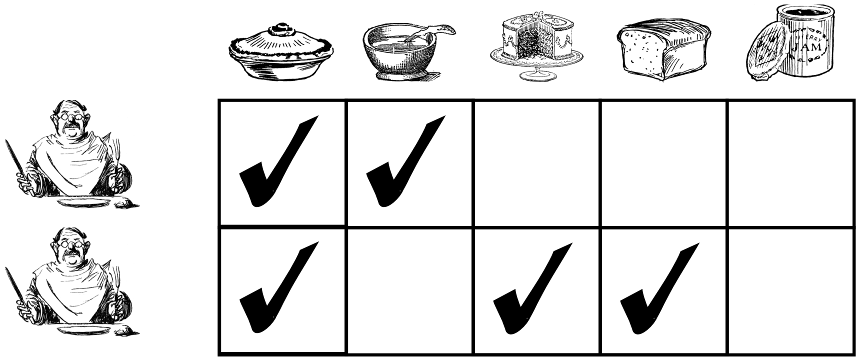
Predictive distribution: The Indian buffet process

- We can describe this model in terms of the following restaurant analogy.
  - A customer enters a restaurant with an infinitely large buffet
  - He helps himself to Poisson(α) dishes.



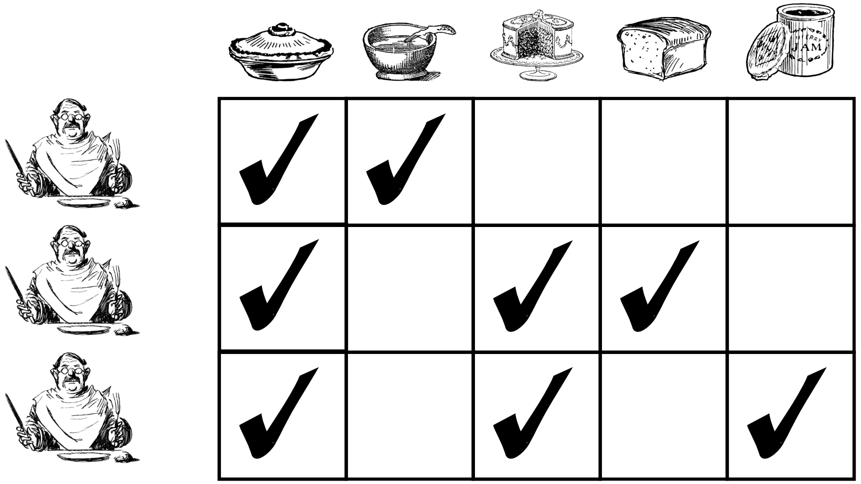
Predictive distribution: The Indian buffet process

- We can describe this model in terms of the following restaurant analogy.
  - A customer enters a restaurant with an infinitely large buffet
  - He helps himself to Poisson(α) dishes.
  - The n^{th} customer enters the restaurant
  - He helps himself to each previously chosen dish with probability m_k / n
  - He then tries Poisson(α/n) new dishes

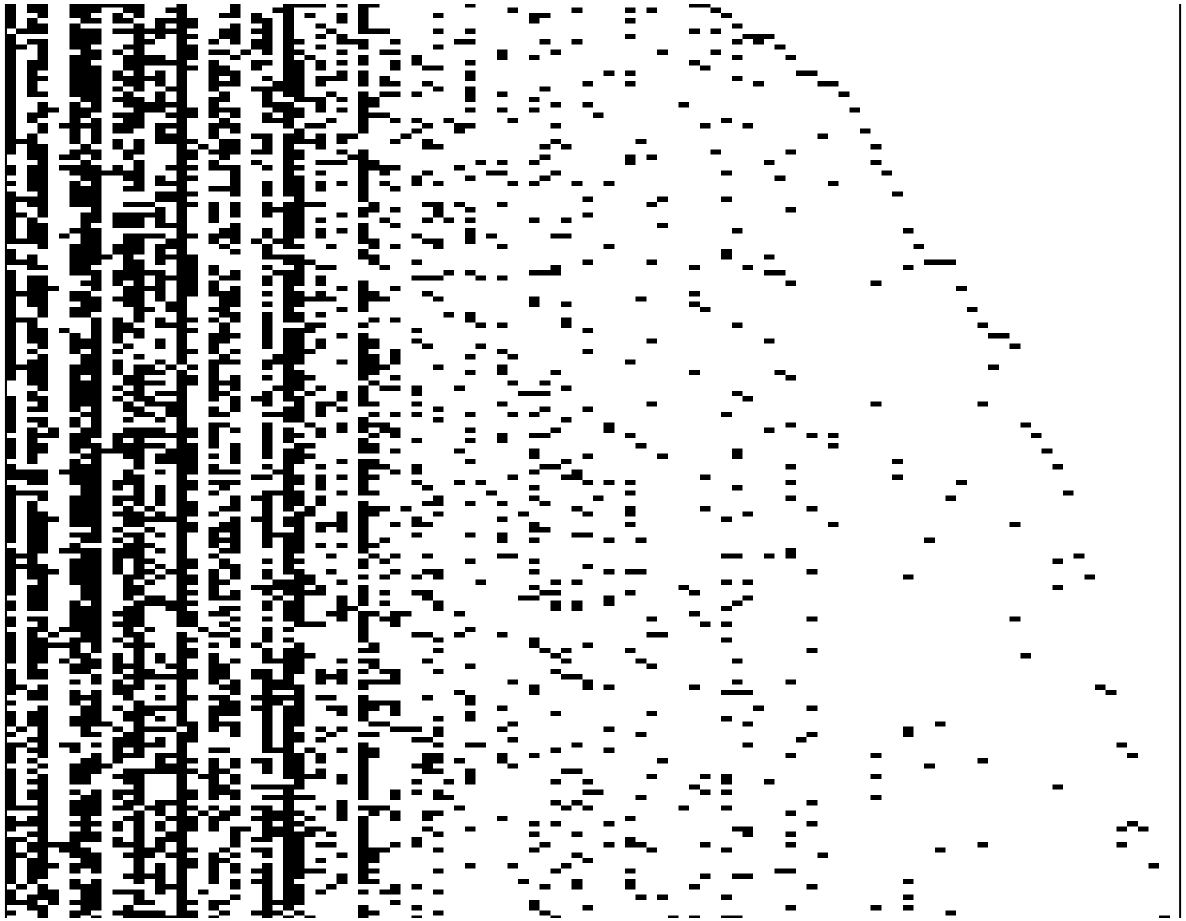


Predictive distribution: The Indian buffet process

- We can describe this model in terms of the following restaurant analogy.
  - A customer enters a restaurant with an infinitely large buffet
  - He helps himself to Poisson(α) dishes.
  - The n^{th} customer enters the restaurant
  - He helps himself to each previously chosen dish with probability m_k / n
  - He then tries Poisson(α/n) new dishes



Example



Proof that the IBP is lof-equivalent to the infinite beta-Bernoulli model

- What is the probability of a matrix **Z**?
- Let $K_1^{(n)}$ be the number of new features in the n^{th} row.

$$p(\mathbf{Z}) = \prod_{n=1}^{N} p(\mathbf{z}_n | \mathbf{z}_{1:(n-1)})$$

$$= \prod_{n=1}^{N} \text{Poisson}\left(K_1^{(n)} \middle| \frac{\alpha}{n}\right) \prod_{k=1}^{K_+} \left(\frac{\sum_{i=1}^{n-1} z_{ik}}{n}\right)^{z_{nk}} \left(\frac{n - \sum_{i=1}^{n-1} z_{ik}}{n}\right)^{1-z_{nk}}$$

$$= \prod_{n=1}^{N} \left(\frac{\alpha}{n}\right)^{K_1^{(n)}} \frac{1}{K_1^{(n)}!} e^{-\alpha/n} \prod_{k=1}^{K_+} \left(\frac{\sum_{i=1}^{n-1} z_{ik}}{n}\right)^{z_{nk}} \left(\frac{n - \sum_{i=1}^{n-1} z_{ik}}{n}\right)^{1-z_{nk}}$$

$$= \frac{\alpha^{K_+}}{\prod_{n=1}^{N} K_1^{(n)}!} \exp\{-\alpha H_N\} \prod_{k=1}^{K_+} \frac{N - m_k)!(m_k - 1)!}{N!}$$

- If we include the cardinality of [**Z**], this is the same as before

Properties of the IBP

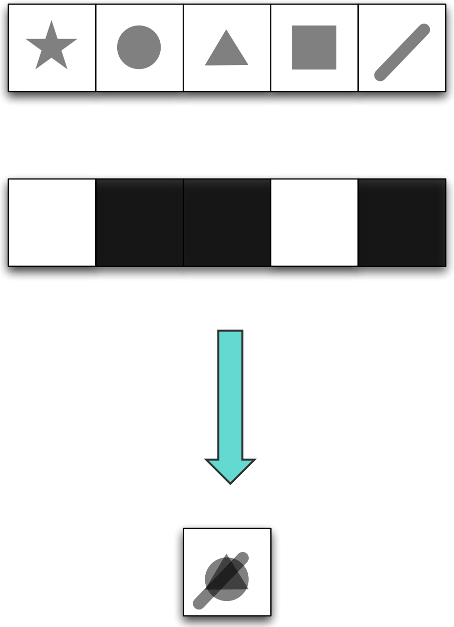
- "Rich get richer" property – "popular" dishes become more popular.
- The number of nonzero entries for each row is distributed according to Poisson(α) – due to exchangeability.
- Recall that if x_1~Poisson(α_1) and x_2~Poisson(α_2), then (x_1+x_2)~Poisson($\alpha_1+\alpha_2$)
  - The number of nonzero entries for the whole matrix is distributed according to Poisson($N\alpha$).
  - The number of non-empty columns is distributed according to Poisson(αH_N)

Building latent feature models using the IBP

- We can use the IBP to build latent feature models with an unbounded number of features.
- Let each column of the IBP correspond to one of an *infinite* number of features.
- Each row of the IBP selects a *finite subset* of these features.
- The **rich-get-richer** property of the IBP ensures features are shared between data points.
- We must pick a *likelihood model* that determines **what the features look like** and **how they are combined**.

A linear Gaussian model

- General form of latent factor model: $\mathbf{X} = \mathbf{W}\mathbf{A}^{\mathsf{T}} + \varepsilon$
- Simplest way to make an infinite factor model:
  - Sample $\mathbf{W} \sim \mathrm{IBP}(\alpha)$
  - Sample $\mathbf{a}_k \sim \mathcal{N}(\mathbf{0}, \sigma_a^2\mathbf{I})$
  - Sample $\varepsilon_{nk} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$



Griffiths and Ghahramani, 2006

Infinite factor analysis

- Problem with linear Gaussian model: Features are "all or nothing" due to the binary "loading matrix" W.
- Factor analysis: $\mathbf{X} = \mathbf{W}\mathbf{A}^T + \varepsilon$
  - Rows of **A** = latent features (Gaussian)
  - Rows of **W** = data-point-specific **weights** for these features (Gaussian)
  - ε = Gaussian noise.

- Write $\mathbf{W} = \mathbf{Z} \odot \mathbf{V}$
  - $\mathbf{Z} \sim \mathrm{IBP}(\alpha)$
  - $\mathbf{V} \sim \mathcal{N}(0,\sigma_V^2)$
  - $\mathbf{A} \sim \mathcal{N}(0,\sigma_A^2)$

Knowles and Ghahramani, 2007

A binary model for latent networks

- Motivation: Discovering latent causes for observed binary data
- Example:
  - Data points = patients
  - Observed features = presence/absence of symptoms
  - Goal: Identify biologically plausible “latent causes” – e.g. illnesses.
- Idea:
  - Each latent feature is associated with a set of symptoms
  - The more features a patient has that are associated with a given symptom, the more likely that patient is to exhibit the symptom.

Wood et al, 2006

A binary model for latent networks

- We can represent this in terms of a *Noisy-OR* model:

$$\mathbf{Z} \sim \mathrm{IBP}(\alpha)$$

$$y_{dk} \sim \mathrm{Bernoulli}(p)$$

d^{th} observed symptom,
k^{th} latent disease

$$p(x_{nd} = 1|\mathbf{Z}, \mathbf{Y}) = 1 - (1 - \lambda)^{\mathbf{z}_n \mathbf{y}_d^T}(1 - \epsilon)$$

- Intuition:
  - Each patient has a set of latent causes, as indicated by Z
  - Each latent cause (disease) k exhibit a symptom d with a Bernoulli rate
  - For each symptom,, we toss a coin with probability λ for each latent cause that is “on” for that patient and associated with that feature, plus an extra coin with probability ε.
  - If any of the coins land heads, we exhibit that feature.

Inference in the IBP

- Recall inference methods for the DP:
  - Gibbs sampler based on the exchangeable model.
  - Gibbs sampler based on the underlying Dirichlet distribution
  - Variational inference
  - Particle filter.

- We can construct analogous samplers for the IBP

Inference in the restaurant scheme

- Recall the exchangeability of the IBP means we can treat any data point as if it's our last.
- Let K_+ be the total number of used features, excluding the current data point.
- Let Θ be the set of parameters associated with the likelihood – eg the Gaussian matrix **A** in the linear Gaussian model
- The prior probability of choosing one of these features is m_k / N
- The posterior probability is proportional to

$$p(z_{nk} = 1 | \mathbf{x}_n, \mathbf{Z}_{-nk}, \Theta) \propto m_k f(\mathbf{x}_n | z_{nk} = 1, \mathbf{Z}_{-nk}, \Theta)$$
$$p(z_{nk} = 0 | \mathbf{x}_n, \mathbf{Z}_{-nk}, \Theta) \propto (N - m_k) f(\mathbf{x}_n | z_{nk} = 0, \mathbf{Z}_{-nk}, \Theta)$$

- In some cases, we can integrate out Θ, otherwise we must sample this.

Griffiths and Gharamani, 2006

Inference in the restaurant scheme

- In addition, we must propose adding new features.
- Metropolis Hastings method:
  - Let K^*_{old} be the number of features appearing only in the current data point.
  - Propose $K^*_{new} \sim \text{Poisson}(\alpha/N)$, and let $\mathbf{Z}^*$ be the matrix with K^*_{new} features appearing only in the current data point.
  - With probability

$$\min\left(1, \frac{f(\mathbf{x}_n|\mathbf{Z}^*, \Theta)}{f(\mathbf{x}_n|\mathbf{Z}, \Theta)}\right)$$

    accept the proposed matrix.

Beta processes and the IBP

- Recall the relationship between the Dirichlet process and the Chinese restaurant process:
  - The Dirichlet process is a prior on probability measures (distributions)
  - We can use this probability measure as cluster weights in a clustering model – cluster allocations are i.i.d. given this distribution.
  - If we integrate out the weights, we get an *exchangeable* distribution over partitions of the data – the **Chinese restaurant process**.
- De Finetti's theorem tells us that, if a distribution $X_1, X_2, \ldots$ is *exchangeable*, there **must** exist a measure conditioned on which $X_1, X_2, \ldots$ are i.i.d.

Beta processes and the IBP

- Recall the finite beta-Bernoulli model:

$$\pi_k \sim \mathrm{Beta}\left(\frac{\alpha}{K}, 1\right)$$

$$z_{nk} \sim \mathrm{Bernoulli}(\pi_k)$$

- The z_{nk} are i.i.d. given the π_k, but are exchangeable if we integrate out the π_k.
- The corresponding distribution for the IBP is the *infinite limit* of the beta random variables, as K tends to infinity.
- This distribution over discrete measures is called the **beta process**.
- Samples from the beta process have infinitely many atoms with masses between 0 and 1.

Thibaux and Jordan, 2007

Posterior distribution of the beta process

- **Question:** Can we obtain the posterior distribution of the column probabilities in closed form?
- **Answer:** Yes!
  - Recall that each atom of the beta process is the infinitesimal limit of a Beta$(\alpha/K, 1)$ random variable.
  - Our counts of observations for that atom are a Binomial(π_k, N) random variable.
  - We know the beta distribution is conjugate to the Binomial, so the posterior is the infinitesimal limit of a Beta$(\alpha/K+m_k, N+1-m_k)$ random variable.

Theorem: Let $X_1, X_2, \cdots, X_n$ be independent Bernoulli random variables, each with the same parameter p. Then the sum $X = X_1 + \cdots + X_n$ is a binomial random variable with parameters n and p.

A stick-breaking construction for the beta process

- We can construct the beta process using the following stick-breaking construction:
- Begin with a stick of unit length.
- For k=1,2,…
  - Sample a beta(α,1) random variable μ_k.
  - Break off a fraction μ_k of the stick. This is the k^{th} atom size.
  - Throw away *what's left* of the stick.
  - Recurse on the part of the stick that you broke off

$$\pi_k = \prod_{j=1}^{k} \mu_j \qquad \mu_j \sim \mathrm{Beta}(\alpha, 1)$$

- Note that, unlike the DP stick breaking construction, the atoms will *not* sum to one.

Teh et al, 2007

Inference in the stick-breaking construction

- We can also perform inference using the stick-breaking representation
  - Sample **Z|π,Θ**
  - Sample **π|Z**
- The posterior for atoms for which $m_k > 0$ is beta distributed.
- The atoms for which $m_k = 0$ can be sampled using the stick-breaking procedure.
- We can use a *slice sampler* to avoid representing all of the atoms, or using a fixed truncation level.

Teh et al, 2007

A two-parameter extension

- In the IBP, the parameter α governs both the *number of nonempty columns* and the *number of features per data point*.
- We might want to decouple these properties of our model.
- Reminder: We constructed the IBP as the limit of a finite beta-Bernoulli model where

$$\pi_k \sim \mathrm{Beta}\left(\frac{\alpha}{K}, 1\right)$$
$$z_{nk} \sim \mathrm{Bernoulli}(\pi_k)$$

- We can modify this to incorporate an extra parameter:

$$\pi_k \sim \mathrm{Beta}\left(\frac{\alpha\beta}{K}, \beta\right)$$
$$z_{nk} \sim \mathrm{Bernoulli}(\pi_k)$$

Sollich, 2005

A two-parameter extension

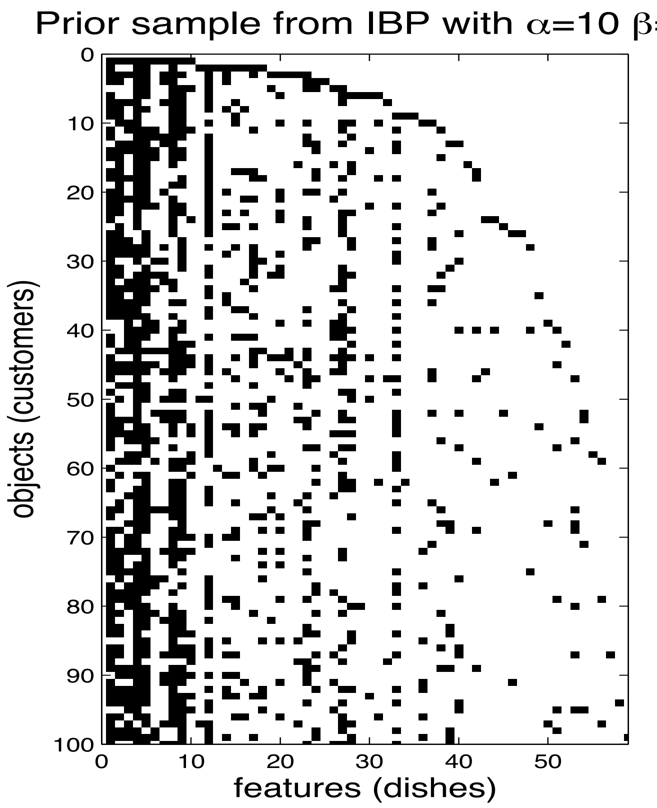
- Our restaurant scheme is now as follows:
  - A customer enters a restaurant with an infinitely large buffet
  - He helps himself to Poisson(α) dishes.
  - The n^{th} customer enters the restaurant
  - He helps himself to each dish with probability $m_k / (\beta+n-1)$
  - He then tries Poisson($\alpha\beta / (\beta+n-1)$) new dishes
- Note
  - The number of features per data point is still marginally Poisson(α).
  - The number of non-empty columns is now

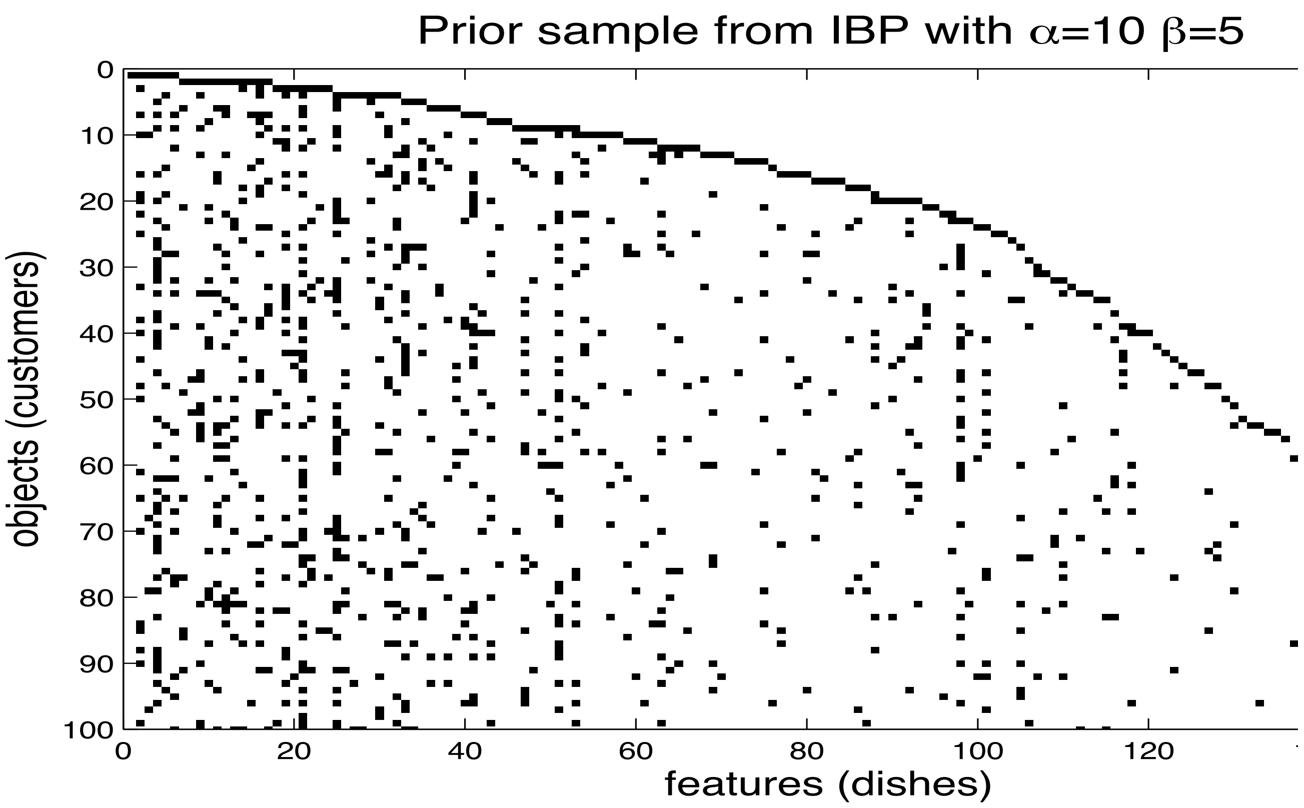
$$\text{Poisson}\left(\alpha \sum_{n=1}^{N} \frac{\beta}{\beta+n-1}\right)$$

  - We recover the IBP when $\beta = 1$.

Two parameter IBP: examples

ɔm IBP
2





ıes)

Image from Griffiths and Ghahramani, 2011

Other distributions over infinite, exchangeable matrices

- Recall the beta-Bernoulli process construction of the IBP.
- We start with a beta process – an infinite sequence of values between 0 and 1 that are distributed as the infinitesimal limit of the beta distribution.
- We combine this with a Bernoulli process, to get a binary matrix.
- If we integrate out the beta process, we get an exchangeable distribution over binary matrices.
- Integration is straightforward due to the beta-Bernoulli conjugacy.
- **Question:** Can we construct other infinite matrices in this way?

The infinite gamma-Poisson process

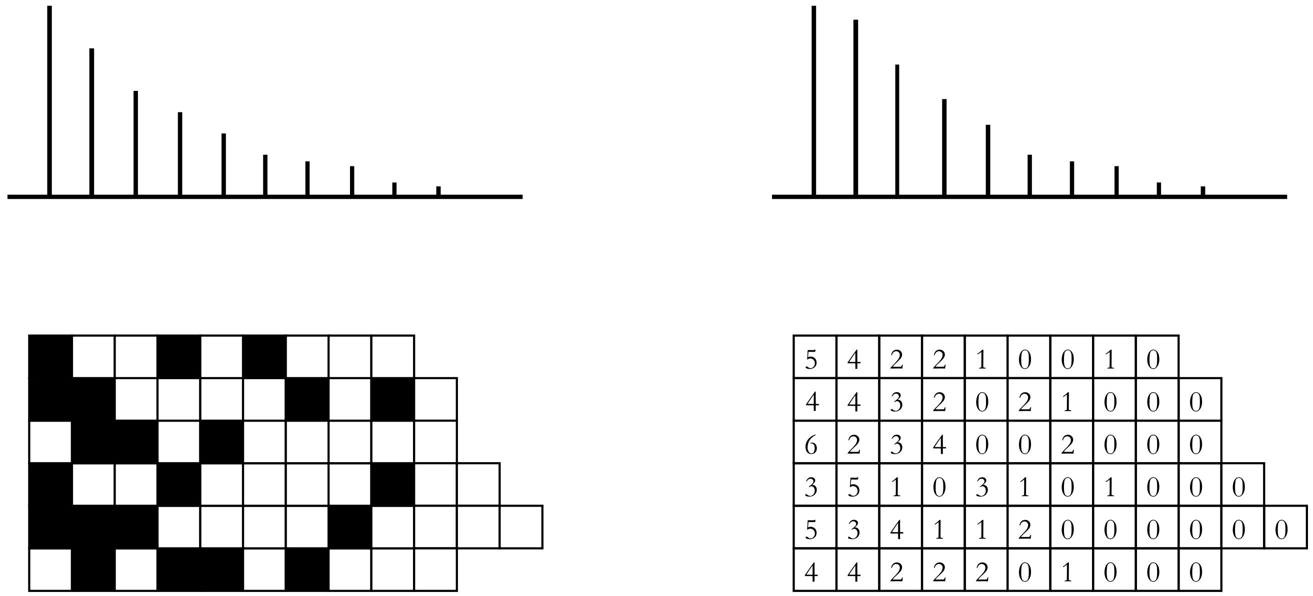
- The *gamma process* can be thought of as the infinitesimal limit of a sequence of gamma random variables.
- Alternatively,

$$\begin{aligned} \text{if } D &\sim \mathrm{DP}(\alpha, H) \\ \text{and } \gamma &\sim \mathrm{Gamma}(\alpha, 1) \\ \text{then } G = \gamma D &\sim \mathrm{GaP}(\alpha H) \end{aligned}$$

- The gamma distribution is conjugate to the Poisson distribution.

The infinite gamma-Poisson process

- We can associate each atom v_k of the gamma process with a column of a matrix (just like we did with the atoms of a beta process)
- We can generate entries for the matrix as $z_{nk} \sim \text{Poisson}(v_k)$



IBP infinite gamma-Poisson

Titsias, 2008

The infinite gamma-Poisson process

- Predictive distribution for the n^{th} row:
  - For each existing feature, sample a count z_{nk}~NegBinom$(m_k, n/(n+1))$

4	2	4	7	0	0	0	0	0
5	0	2	9	4	1	0	0	0
3	2	1	6	2	1	0	0	0
7	1	3	6	3	0	0	0	0

The infinite gamma-Poisson process

- Predictive distribution for the n^{th} row:
  - For each existing feature, sample a count z_{nk}~NegBinom$(m_k, n/(n+1))$

4	2	4	7	0	0	0	0	0
5	0	2	9	4	1	0	0	0
3	2	1	6	2	1	0	0	0
7	1	3	6	3	0	0	0	0
5								

The infinite gamma-Poisson process

- Predictive distribution for the n^{th} row:
  - For each existing feature, sample a count z_{nk}~NegBinom$(m_k, n/(n+1))$

4	2	4	7	0	0	0	0	0
5	0	2	9	4	1	0	0	0
3	2	1	6	2	1	0	0	0
7	1	3	6	3	0	0	0	0
5	0							

The infinite gamma-Poisson process

- Predictive distribution for the n^{th} row:
  - For each existing feature, sample a count z_{nk}~NegBinom$(m_k, n/(n+1))$

4	2	4	7	0	0	0	0	0
5	0	2	9	4	1	0	0	0
3	2	1	6	2	1	0	0	0
7	1	3	6	3	0	0	0	0
5	0	4	5	2	0			

The infinite gamma-Poisson process

- Predictive distribution for the n^{th} row:
  - For each existing feature, sample a count z_{nk}~NegBinom$(m_k, n/(n+1))$
  - Sample K^*_n~NegBinom$(\alpha, n/(n+1))$

4	2	4	7	0	0	0	0	0
5	0	2	9	4	1	0	0	0
3	2	1	6	2	1	0	0	0
7	1	3	6	3	0	0	0	0
5	0	4	5	2	0			

4

The infinite gamma-Poisson process

- Predictive distribution for the n^{th} row:
  - For each existing feature, sample a count z_{nk}~NegBinom$(m_k, n/(n+1))$.
  - Sample K^*_n~NegBinom$(\alpha, n/(n+1))$.
  - Partition K^*_n according to the CRP, and assign the resulting counts to new columns.

4	2	4	7	0	0	0	0	0
5	0	2	9	4	1	0	0	0
3	2	1	6	2	1	0	0	0
7	1	3	6	3	0	0	0	0
5	0	4	5	2	0	3	1	0