

# Infinite latent feature models and the Indian Buffet Process

**Tom Griffiths**

Cognitive and Linguistic Sciences

Brown University

Joint work with **Zoubin Ghahramani**

# Beyond latent classes

- Unsupervised learning often uses latent classes
- Many domains require richer representations
  - membership in multiple latent classes
  - more generally, latent features
- How do we choose the dimensionality of representations?
  - problem of *model selection*

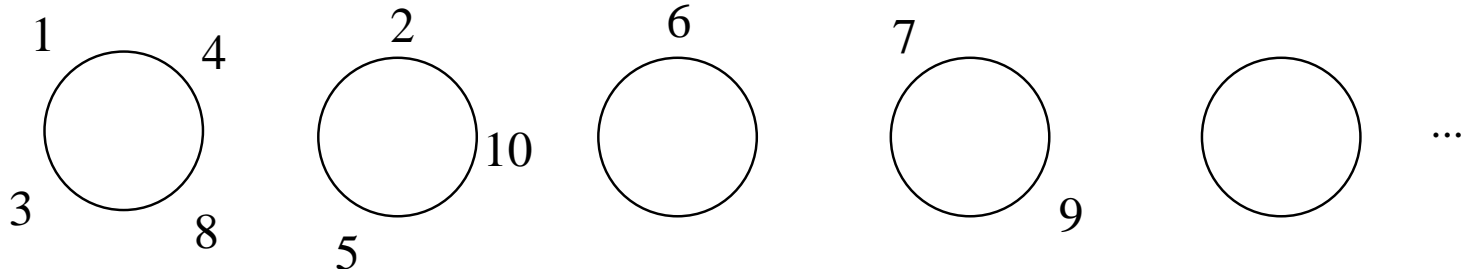
# Perspectives on model selection

- Compare multiple models of different dimensionality
  - Bayes factors, cross-validation, etc.
  - (usually) false assumption of fixed dimension
  - hard to apply to large model spaces
- Define a single model of unbounded dimensionality
  - posterior on dimensionality via posterior on parameters
  - allows dimensionality to grow with new data
  - pursued in nonparametric Bayesian density estimation (e.g., Antoniak, 1974; Escobar & West, 1995)

# Latent class models

- Associate each datapoint  $\mathbf{x}_i$  with a latent class  $z_i$ 
  - data matrix  $\mathbf{X} = [\mathbf{x}_1 \dots \mathbf{x}_N]^T$
  - class vector  $\mathbf{z} = [z_1 \dots z_N]^T$
- Model defined by
  - prior on class assignments  $P(\mathbf{z})$
  - likelihood  $p(\mathbf{X}|\mathbf{z})$
- How do we choose the number of classes?
- Nonparametric Bayes: define  $P(\mathbf{z})$  to allow infinitely many classes, of which a finite subset are used

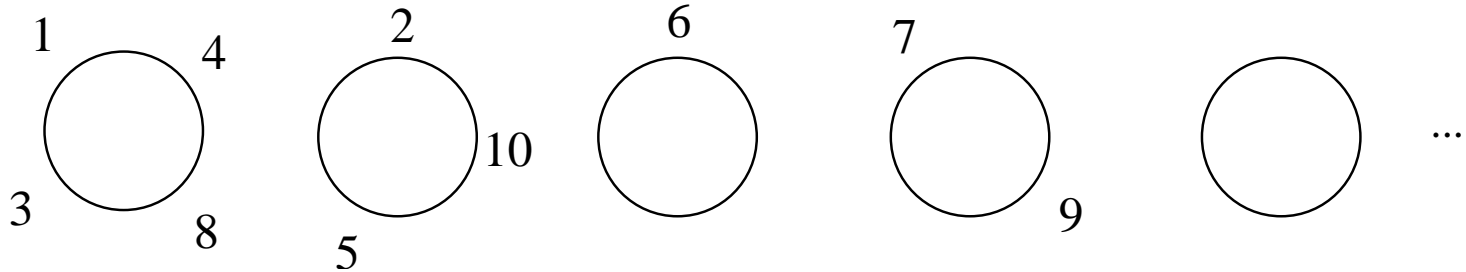
# Chinese restaurant process (CRP)



- Chinese restaurant with infinitely many infinite tables
- $N$  customers sit down
  - the first customer sits at the first table
  - the  $i$ th customer chooses a table at random

$$P(\text{occupied table } k | \text{previous customers}) = \frac{m_k}{\alpha + i - 1}$$
$$P(\text{next unoccupied table} | \text{previous customers}) = \frac{\alpha}{\alpha + i - 1}$$

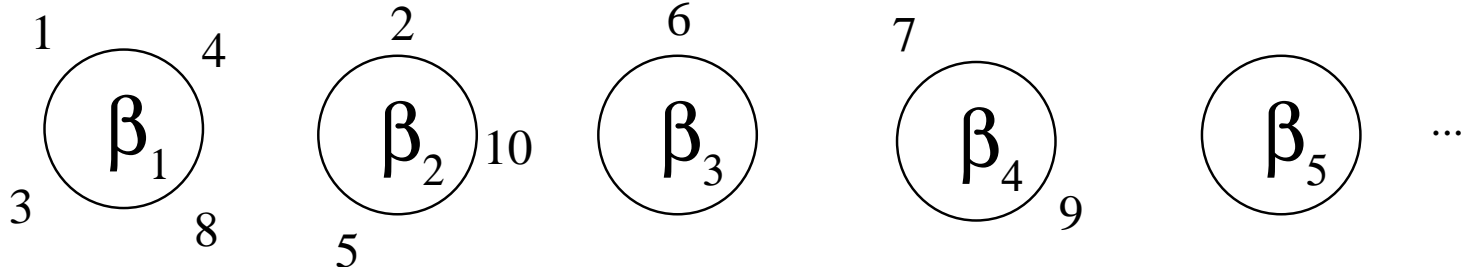
# Chinese restaurant process (CRP)



- Defines a distribution over partitions
- e.g., (1 3 4 8) (2 5 10) (6) (7 9)
- Customers are exchangeable (Aldous, 1985; Pitman, 2002)

$$P(\mathbf{z}) = \alpha^{K_+} \left( \prod_{k=1}^{K_+} (m_k - 1)! \right) \frac{\Gamma(\alpha)}{\Gamma(N + \alpha)}$$

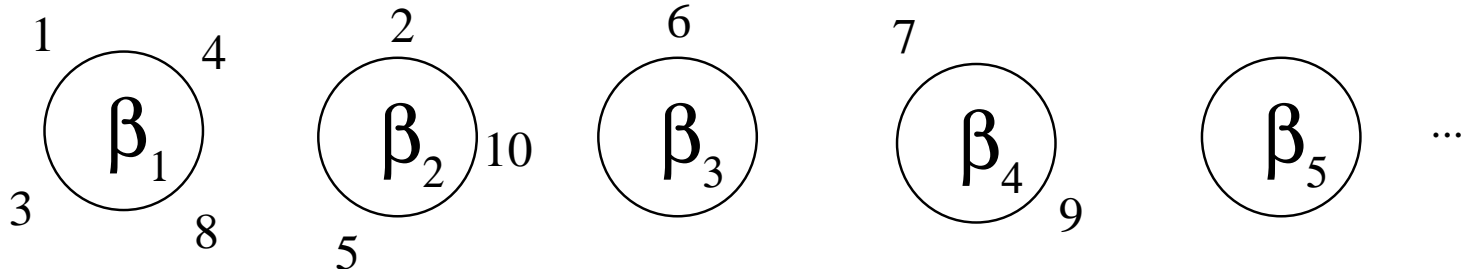
# CRP and mixture modeling



- Each table  $k$ 
  - corresponds to a mixture component
  - associated with a parameter  $\beta_k$  drawn from a prior
- e.g., Gaussian CRP mixture model:

$$\begin{aligned} \mathbf{z} &\sim \text{CRP}(\alpha) \\ x_i | z_i, \beta &\sim \text{Gaussian}(\beta_{z_i}, \sigma_X) \\ \beta_k &\sim \text{Gaussian}(0, \sigma_\beta) \end{aligned}$$

# CRP and mixture modeling



- Given data  $\mathbf{x}$ , posterior on  $\mathbf{z}$  gives
  - # of classes (# of occupied tables)
  - which data are assigned to each class
  - parameter for each class,  $p(\beta_k | \text{data assigned to table } k)$
- Posterior inference via Gibbs sampling (e.g., Neal, 1998)

# Gibbs sampling

- Sequentially sample class assignments (for each  $i$ )

$$P(z_i | \mathbf{x}, \mathbf{z}_{-i}) \propto p(x_i | \mathbf{x}_{-i}, \mathbf{z}) P(z_i | \mathbf{z}_{-i})$$

- CRP provides  $P(z_i | \mathbf{z}_{-i})$

$$P(z_i = \text{occupied class } k | \mathbf{z}_{-i}) = \frac{m_{k,-i}}{\alpha + N - 1}$$

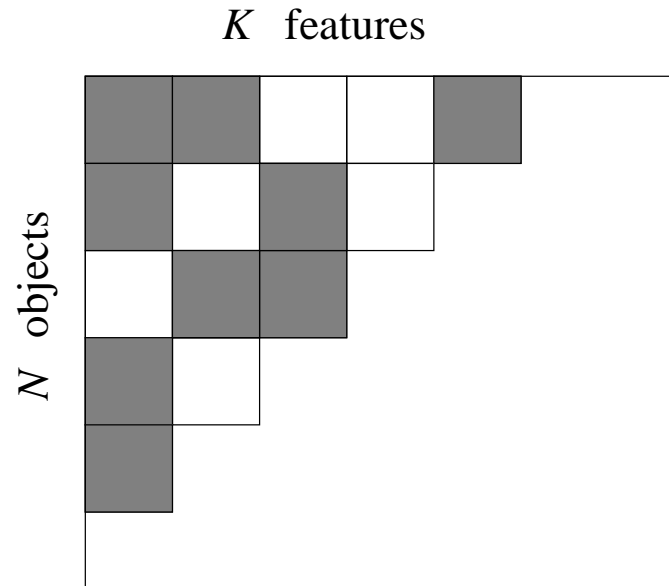
$$P(z_i = \text{new class} | \mathbf{z}_{-i}) = \frac{\alpha}{\alpha + N - 1}$$

- Allows datapoints to come from new classes
- Also split-merge algorithms (Jain & Neal, 2000; Dahl, 2003)

# Beyond latent classes

- The CRP allows number of classes to be inferred
  - a prior on class assignments of unbounded dimension
  - distribution over partitions
- Can we apply a parallel strategy with other representations?
- Infinite latent feature models
  - a prior on feature assignments of unbounded dimension
  - distribution over binary matrices

# Different feature representations



- Binary features

# Different feature representations

$K$  features

	1	3	0	0	4	
$N$ objects	5	0	3	0		
	0	1	4			
	2	0				
	5					

- Binary features
- Factorial features

# Different feature representations

$K$  features

	0.9	1.4	0	0	-0.3
$N$ objects	-3.2	0	0.9	0	
	0	0.2	-2.8		
	1.8	0			
	-0.1				

- Binary features
- Factorial features
- Continuous features

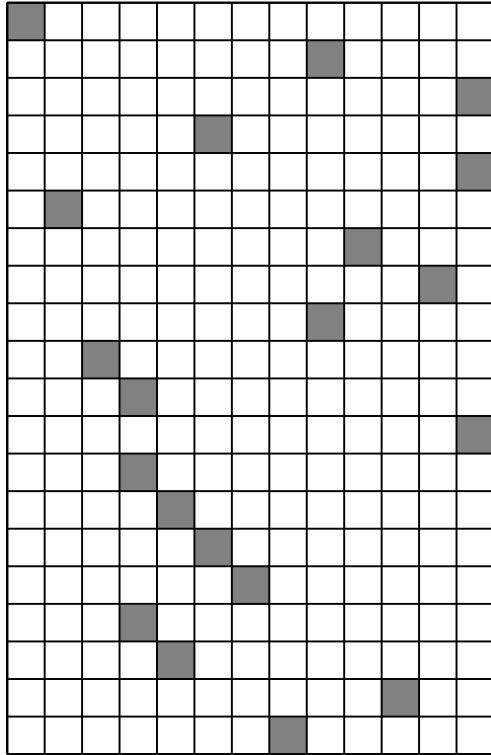
# Latent feature models

- Associate each datapoint  $\mathbf{x}_i$  with a latent feature vector  $\mathbf{z}_i$ 
  - data matrix  $\mathbf{X} = [\mathbf{x}_1 \dots \mathbf{x}_N]^T$
  - feature matrix  $\mathbf{Z} = [\mathbf{z}_1 \dots \mathbf{z}_N]^T$
- Model defined by
  - prior on feature assignments  $P(\mathbf{Z})$
  - likelihood  $p(\mathbf{X}|\mathbf{Z})$
- How do we choose the number of features?
- Nonparametric Bayes: define  $P(\mathbf{Z})$  to allow infinitely many features, of which a finite subset are used

# Priors on binary matrices

- Start with priors on  $N \times K$  matrices, take  $K \rightarrow \infty$
- Two cases:
  - “class matrices”: one 1 per row
  - “feature matrices”: general binary matrices
- Two priors:
  - the Chinese restaurant process
  - the Indian buffet process

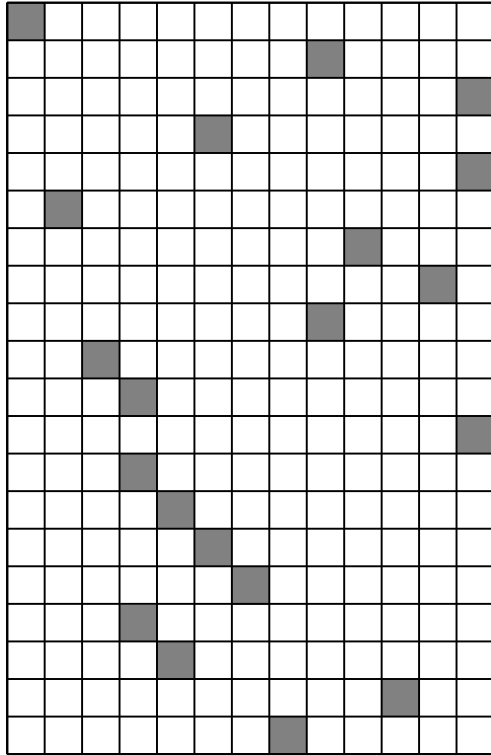
# Class matrices



$$\mathbf{z}_i | \theta \sim \text{Discrete}(\theta)$$

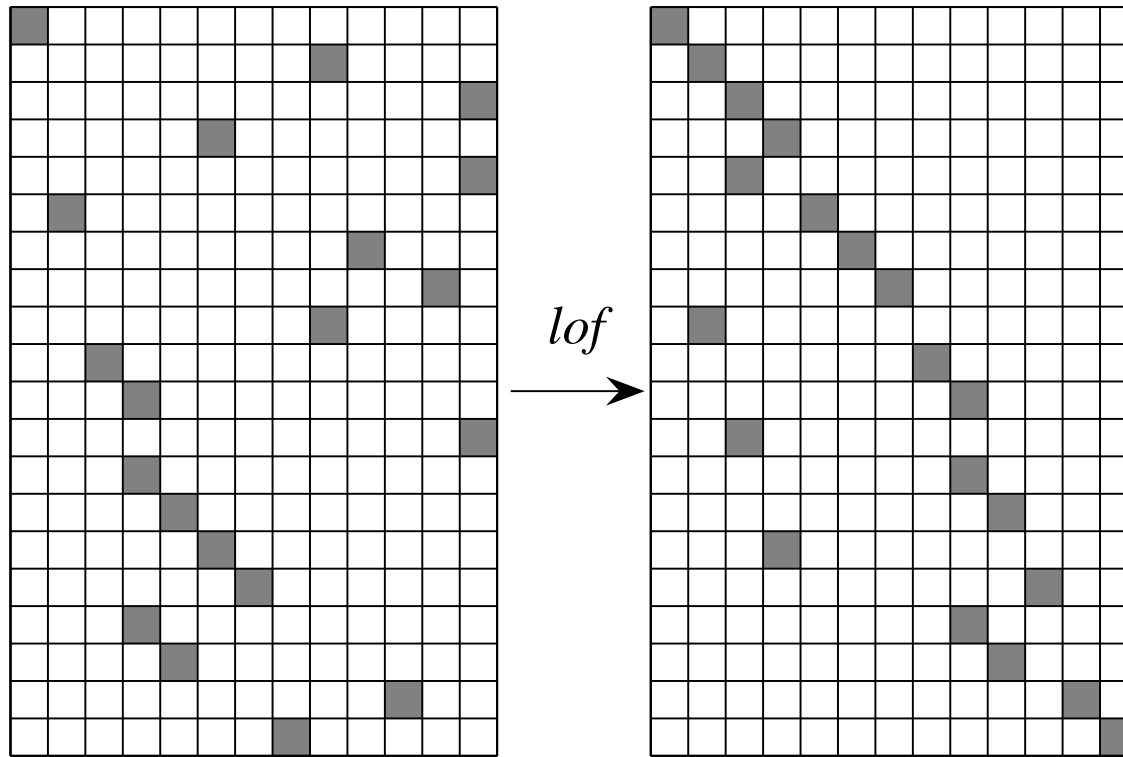
$$\theta \sim \text{Dirichlet}\left(\frac{\alpha}{K}\right)$$

# Class matrices



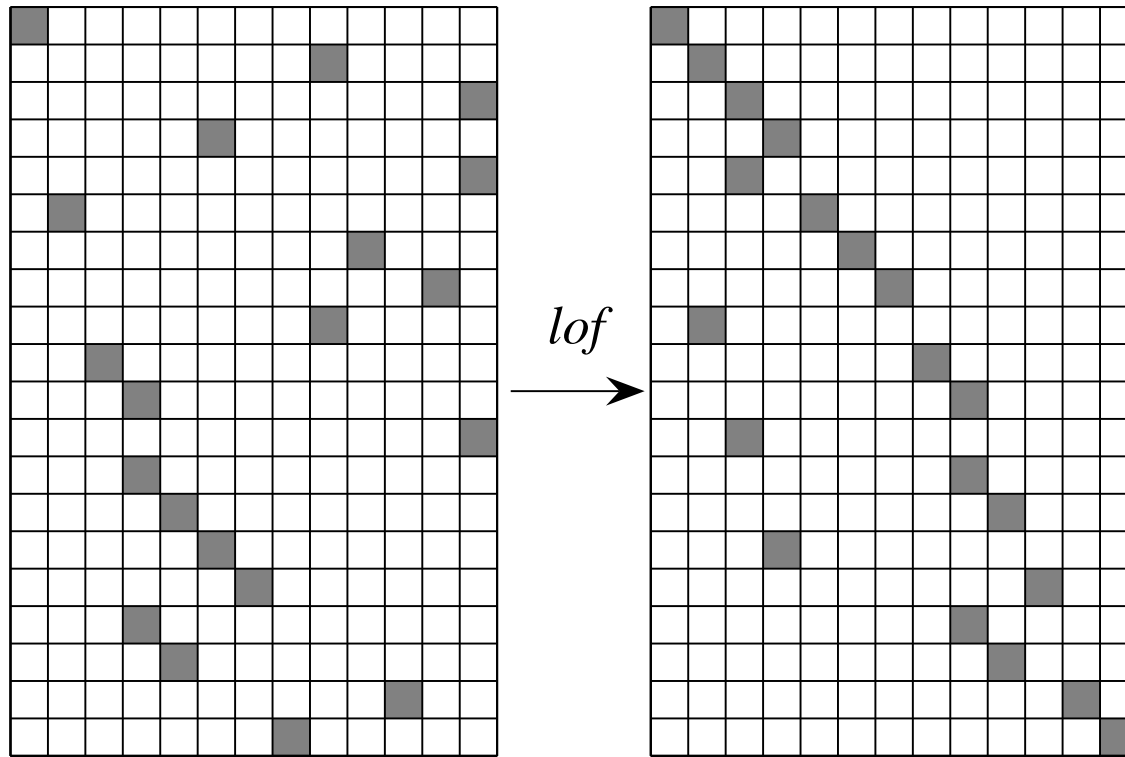
$$P(\mathbf{Z}) = \int_{\Delta} \prod_{i=1}^N P(\mathbf{z}_i | \theta) p(\theta) d\theta$$

# Left-ordered form



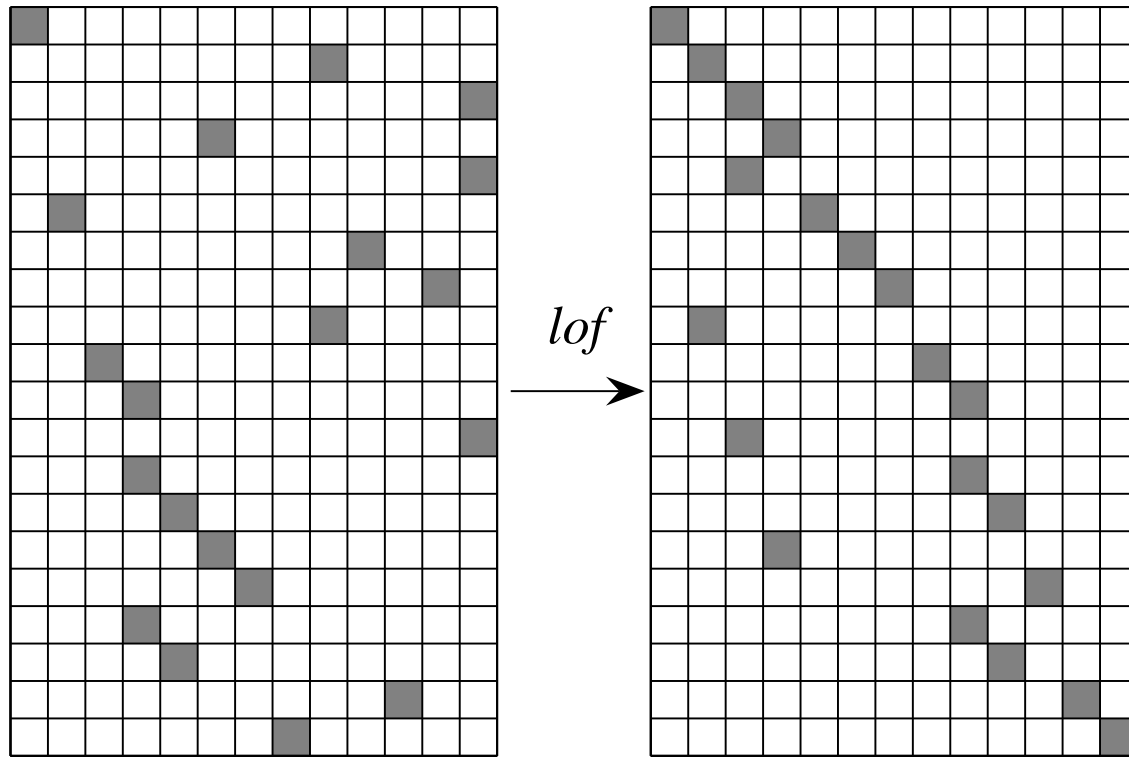
- History  $h$  of each class: binary column vector
- *lof* orders columns by values of binary histories

# *lof* equivalence classes



- $\mathbf{X}$  and  $\mathbf{Y}$  are *lof* equivalent iff  $lof(\mathbf{X}) = lof(\mathbf{Y})$
- Class matrices: *lof* equivalence classes are partitions

# *lof* equivalence classes



$$\lim_{k \rightarrow \infty} P([\mathbf{Z}]) = \alpha^{K_+} \left( \prod_{k=1}^{K_+} (m_k - 1)! \right) \frac{\Gamma(\alpha)}{\Gamma(N + \alpha)}$$

(see also Green & Richardson, 2001; Neal, 1992)

# Feature matrices

- For general binary matrices

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

$$\theta_k \sim \text{Beta}\left(\frac{\alpha}{K}, 1\right)$$

# Feature matrices

- For general binary matrices

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

$$\theta_k \sim \text{Beta}\left(\frac{\alpha}{K}, 1\right)$$

- For a finite matrix  $\mathbf{Z}$

$$P(\mathbf{Z}) = \int_0^1 \cdots \int_0^1 P(\mathbf{Z}|\theta_1, \dots, \theta_k) \prod_{k=1}^K p(\theta_k) d\theta_k$$

# Feature matrices

- For general binary matrices

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

$$\theta_k \sim \text{Beta}\left(\frac{\alpha}{K}, 1\right)$$

- For a finite matrix  $\mathbf{Z}$

$$P(\mathbf{Z}) = \int_0^1 \cdots \int_0^1 P(\mathbf{Z}|\theta_1, \dots, \theta_k) \prod_{k=1}^K p(\theta_k) d\theta_k$$

- Taking the limit as  $K \rightarrow \infty \dots$

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

# Indian buffet process (IBP)

- Indian restaurant with infinitely many infinite dishes
- $N$  customers serve themselves
  - the first customer samples  $\text{Poisson}(\alpha)$  dishes
  - the  $i$ th customer
    - samples a previously sampled dish with probability  $\frac{m_k}{i+1}$
    - then samples  $\text{Poisson}(\frac{\alpha}{i})$  new dishes

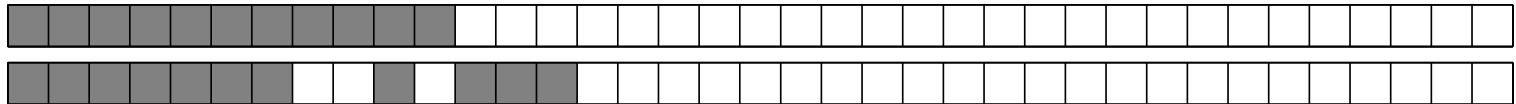
# Indian buffet process (IBP)

- Indian restaurant with infinitely many infinite dishes
- $N$  customers serve themselves
  - the first customer samples  $\text{Poisson}(\alpha)$  dishes
  - the  $i$ th customer
    - samples a previously sampled dish with probability  $\frac{m_k}{i+1}$
    - then samples  $\text{Poisson}(\frac{\alpha}{i})$  new dishes



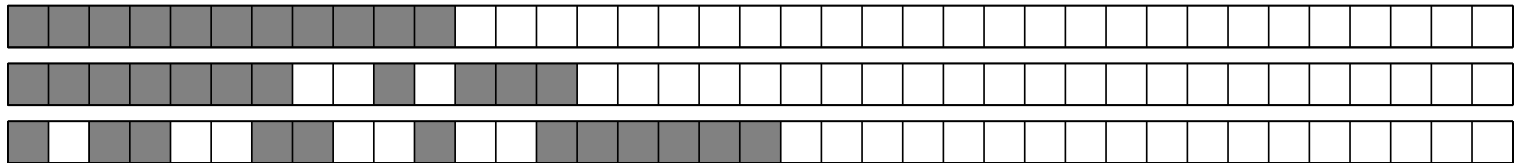
# Indian buffet process (IBP)

- Indian restaurant with infinitely many infinite dishes
- $N$  customers serve themselves
  - the first customer samples  $\text{Poisson}(\alpha)$  dishes
  - the  $i$ th customer
    - samples a previously sampled dish with probability  $\frac{m_k}{i+1}$
    - then samples  $\text{Poisson}(\frac{\alpha}{i})$  new dishes



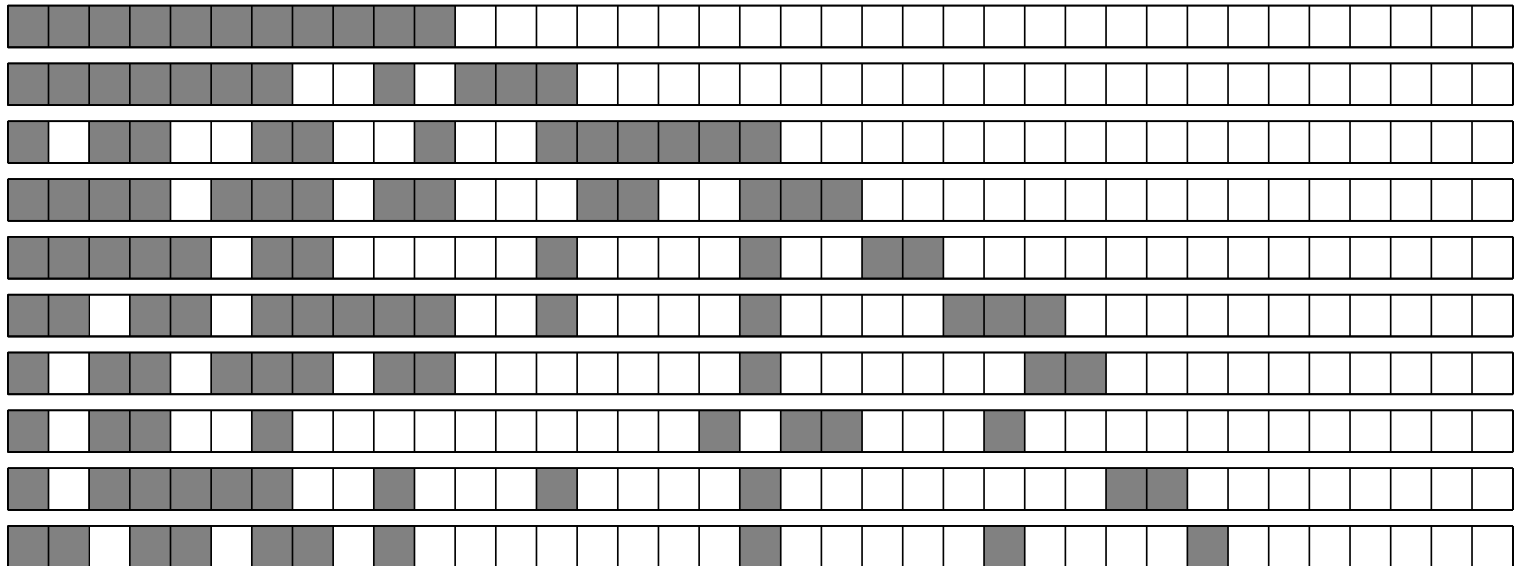
# Indian buffet process (IBP)

- Indian restaurant with infinitely many infinite dishes
- $N$  customers serve themselves
  - the first customer samples  $\text{Poisson}(\alpha)$  dishes
  - the  $i$ th customer
    - samples a previously sampled dish with probability  $\frac{m_k}{i+1}$
    - then samples  $\text{Poisson}(\frac{\alpha}{i})$  new dishes



# Indian buffet process (IBP)

- Indian restaurant with infinitely many infinite dishes
- $N$  customers serve themselves
  - the first customer samples  $\text{Poisson}(\alpha)$  dishes
  - the  $i$ th customer
    - samples a previously sampled dish with probability  $\frac{m_k}{i+1}$
    - then samples  $\text{Poisson}(\frac{\alpha}{i})$  new dishes



# Properties of the IBP

- Customers are exchangeable
- Total number of dishes  $K_+ \sim \text{Poisson}(\alpha \sum_{i=1}^N \frac{1}{i})$ 
  - i.e.  $K_+ \rightarrow \infty$  as  $N \rightarrow \infty$ , as with the CRP
- Number of dishes sampled by each customer  $\sim \text{Poisson}(\alpha)$ 
  - sparsity is coupled to dimension
- Expected number of non-zero entries in  $\mathbf{Z}$  is  $N\alpha$

# A linear-Gaussian model

- Likelihood  $P(\mathbf{X}|\mathbf{Z})$  specified by

$$\mathbf{x}_i \sim \text{Gaussian}(\mathbf{z}_i\mathbf{A}, \sigma_X\mathbf{I})$$

$$\mathbf{A} \sim \text{Gaussian}(\mathbf{0}, \sigma_A\mathbf{I})$$

- For  $\mathbf{Z} \sim \text{CRP}(\alpha)$ , spherical Gaussian mixture model
- For  $\mathbf{Z} \sim \text{IBP}(\alpha)$ , binary latent factor model
- Compute posterior distribution  $P(\mathbf{Z}|\mathbf{X})$

# Gibbs sampling

- Sequentially sample feature assignments

$$P(z_{ik} | \mathbf{X}, \mathbf{z}_{(-i)k}) \propto p(\mathbf{x}_i | \mathbf{X}_{-i}, \mathbf{Z}) P(z_{ik} | \mathbf{z}_{(-i)k})$$

- IBP provides  $P(z_{ik} | \mathbf{X}, \mathbf{z}_{(-i)k})$

$$P(z_{ik} | \mathbf{z}_{(-i)k}) = \frac{m_{k,-i}}{N}$$

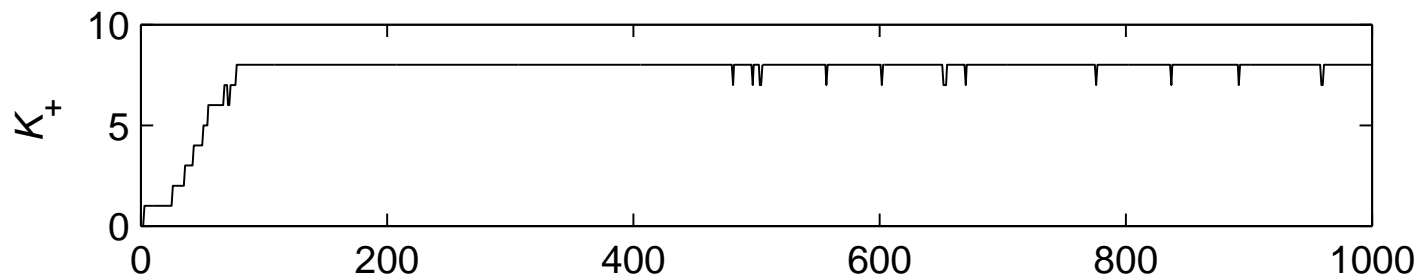
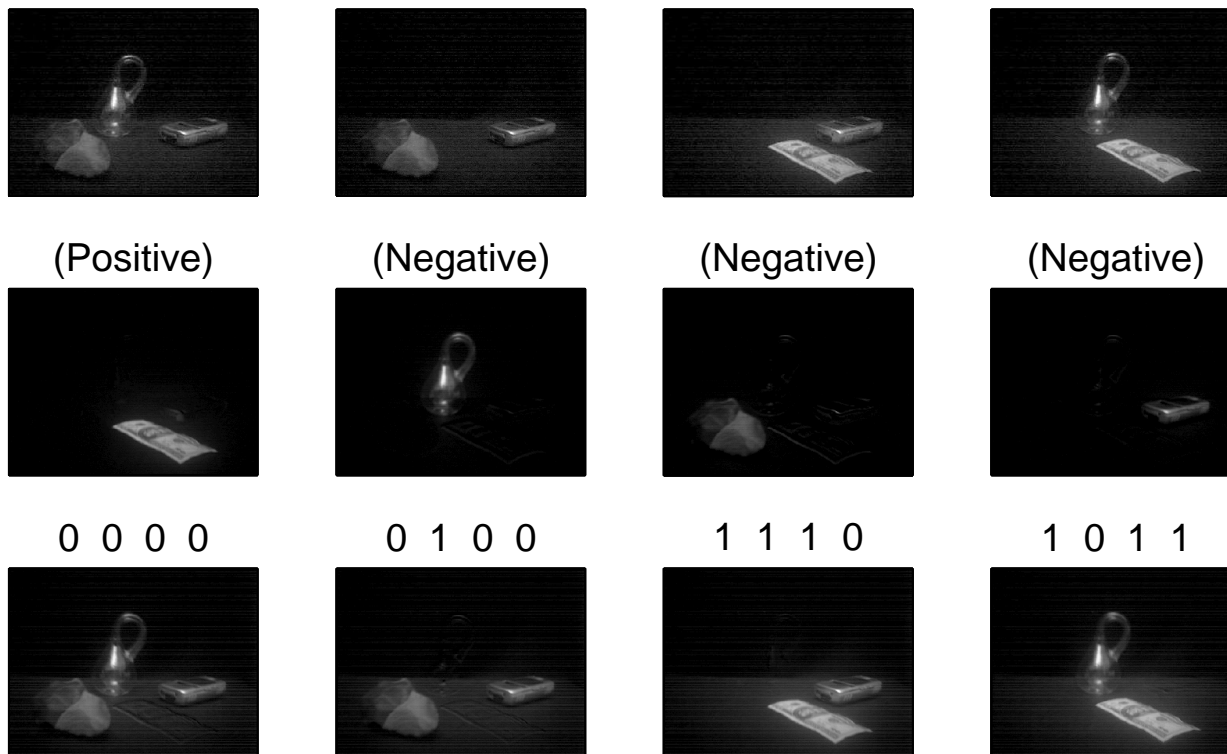
- Draw  $\text{Poisson}(\frac{\alpha}{N})$  new features

# Coding for the presence of objects

- Photographs of everyday objects taken with a webcam
- 100 images, each  $320 \times 240$  pixels
- Each image contained from 1 to 4 (fixed position) objects



# Coding for the presence of objects



# Conclusion

- Strategy for model selection from nonparametric Bayes: prior over combinatorial structures of variable dimension

Structure	Distribution	Models
partitions	CRP	infinite mixture models
binary matrices	IBP	infinite binary latent factors
$(\times \mathbb{Z})$		infinite CVQ
$(\times \mathbb{R})$		infinite sparse PCA

- Other exchangeable distributions?