Priors for Nonparametric Bayes with Structured Representations

Tom Griffiths

Brain and Cognitive Sciences

MIT

Identifying dimensionality

- Common questions raised by high-dimensional data:
  - how many clusters/classes?
  - which tree?
  - how many dimensions?

- These are problems of *model selection*

Perspectives on model selection

- Compare multiple models of different dimensionality
  - Bayes factors, cross-validation, etc.
  - 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)

Outline

- Latent classes
  - distribution on partitions
- Latent paths
  - distribution on trees
- Latent features
  - distribution on binary matrices

Mixture models

- Associate each datapoint x_i with a latent class z_i

$$p(x_i) = \sum_{k=1}^{K} p(x_i|z_i = k)p(z_i)$$

Mixture models

- Associate each datapoint x_i with a latent class z_i

$$p(x_i) = \sum_{k=1}^{K} p(x_i|z_i = k)p(z_i)$$

- e.g., Gaussian mixture model:

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

Mixture models

- Associate each datapoint x_i with a latent class z_i

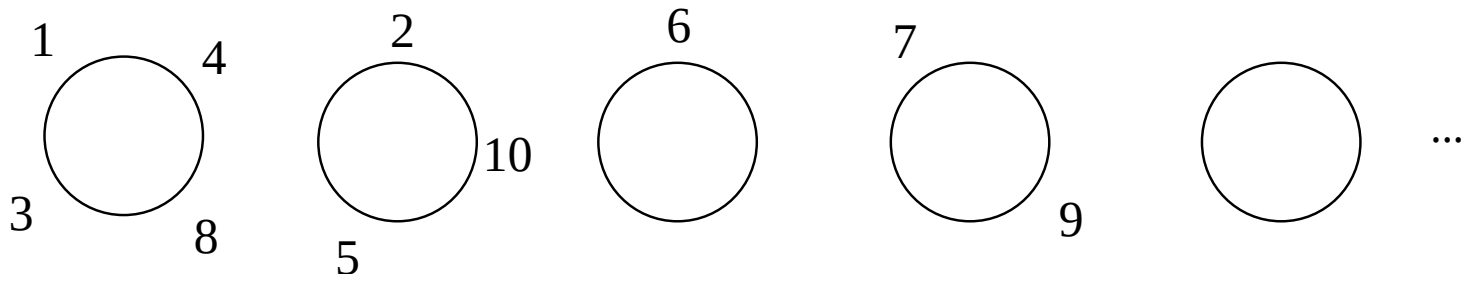
$$p(x_i) = \sum_{k=1}^{K} p(x_i|z_i = k)p(z_i)$$

- e.g., Gaussian mixture model:

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

- How do we choose K?

Chinese restaurant process (CRP)

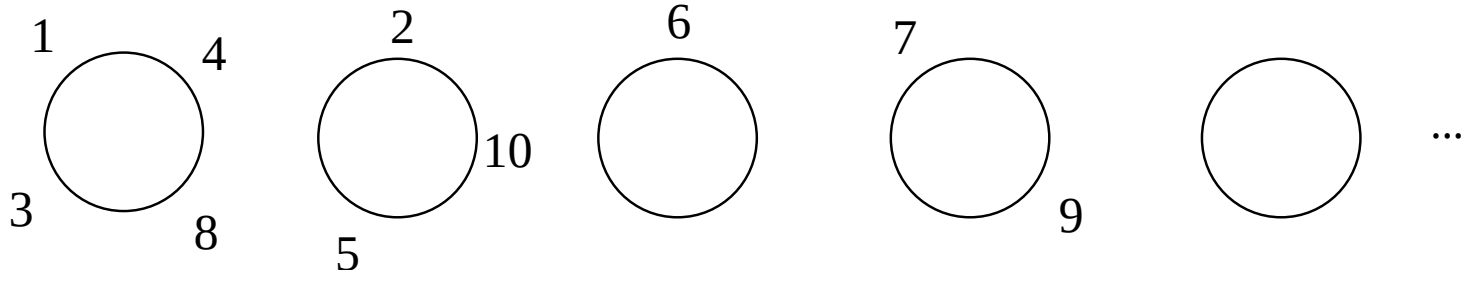


- Chinese restaurant with infinitely many infinite tables
- N customers sit down
  - the first customer sits at the first table
  - the ith customer chooses a table at random

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

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

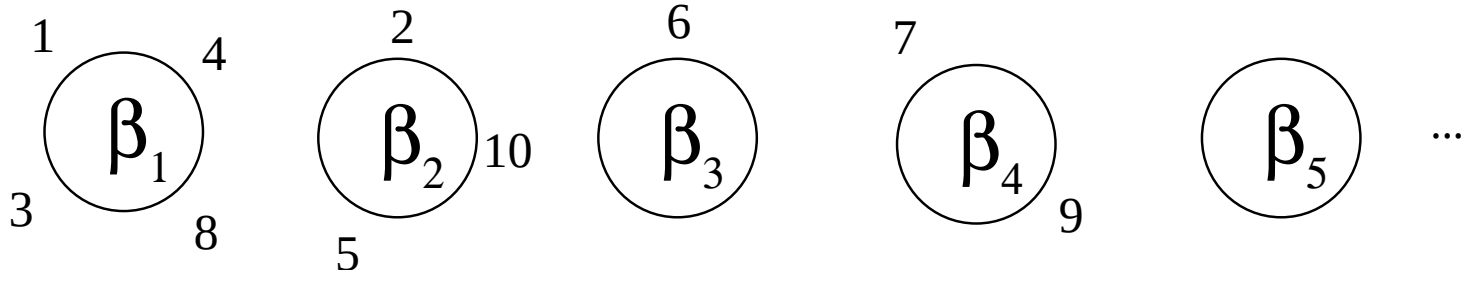
Chinese restaurant process (CRP)



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

$$p(\text{partition}) = \alpha^{K_+} \left(\prod_{k=1}^{K_+} (m_k - 1)! \right) \frac{\Gamma(\alpha)}{\Gamma(N + \alpha)}$$

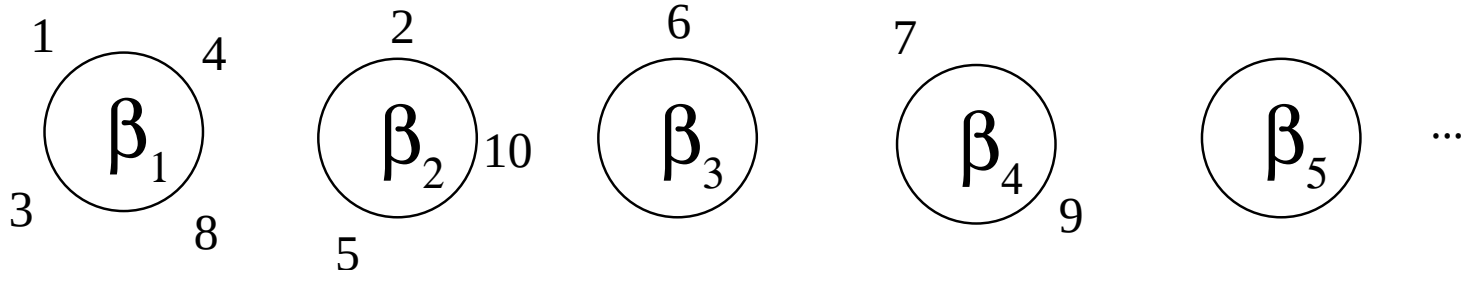
CRP and mixture modeling



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

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

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

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

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

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

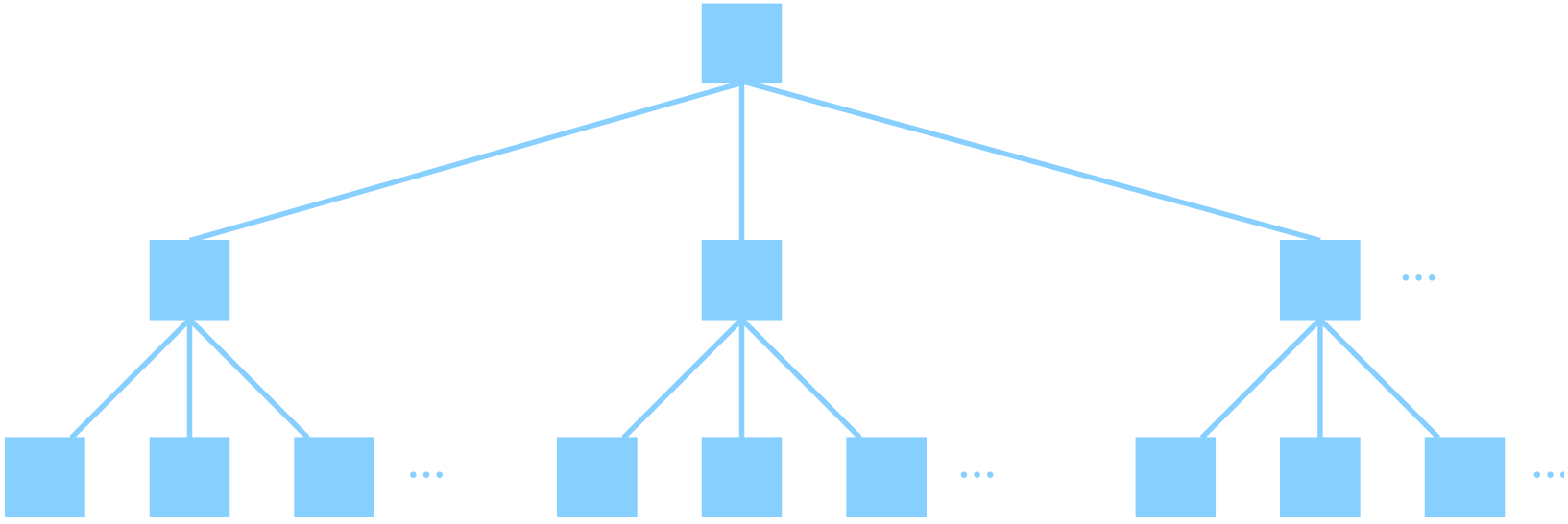
$$P(z_i = \text{new class}|\mathbf{z}_{-i}) \quad = \quad \frac{\alpha}{\alpha + N - 1}$$

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

Beyond the CRP

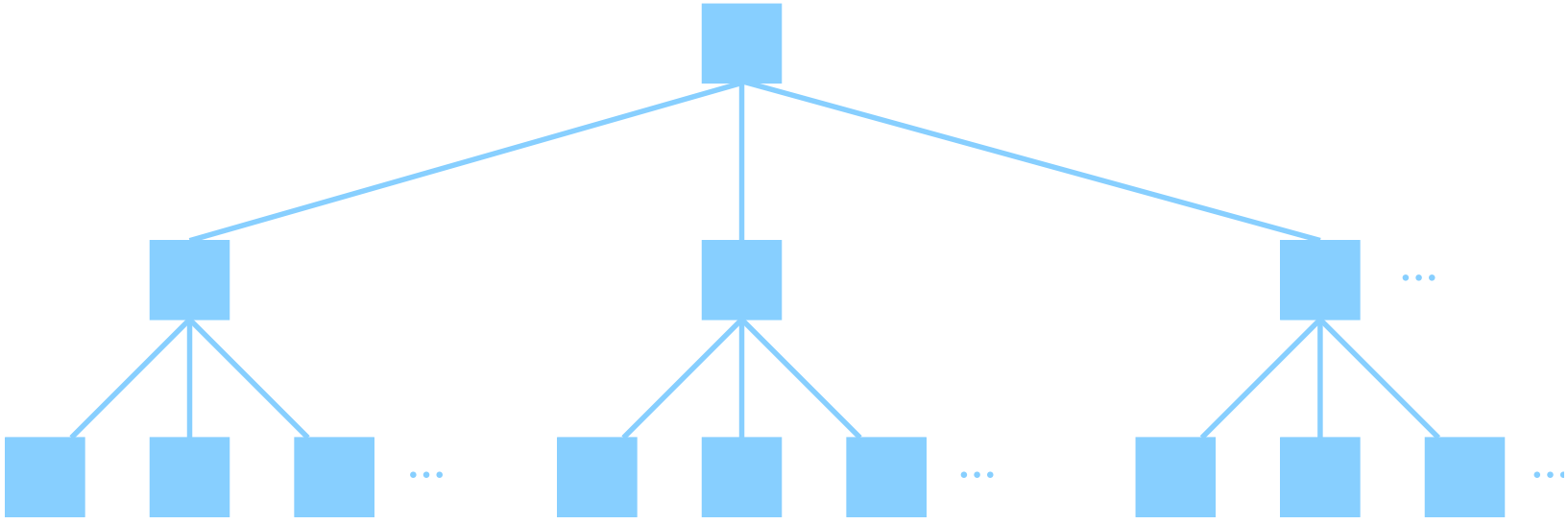
- The CRP allows number of classes to be inferred
- But. . .
  - testing multiple models still feasible for mixtures
  - many kinds of data require other representations
- Can we apply a parallel strategy with other structures?
  - trees (Blei, Griffiths, Jordan, & Tenenbaum, 2004)
  - binary matrices (Griffiths & Ghahramani, in prep)

Trees



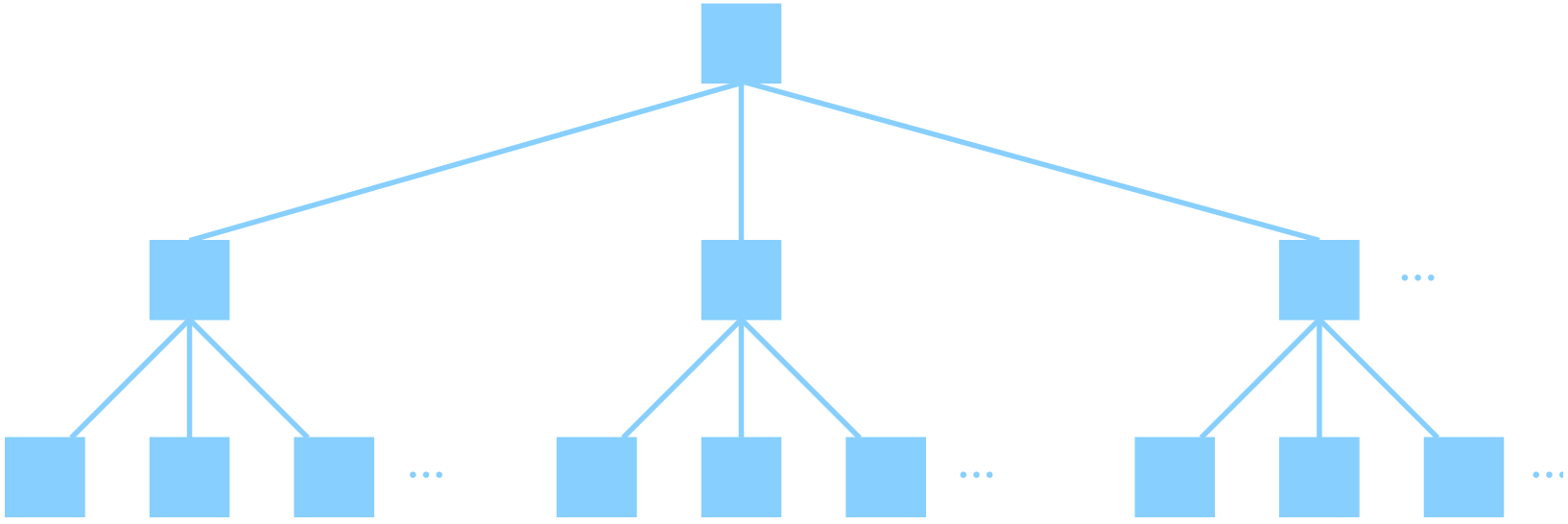
- One representation of trees is as nested partitions

Trees



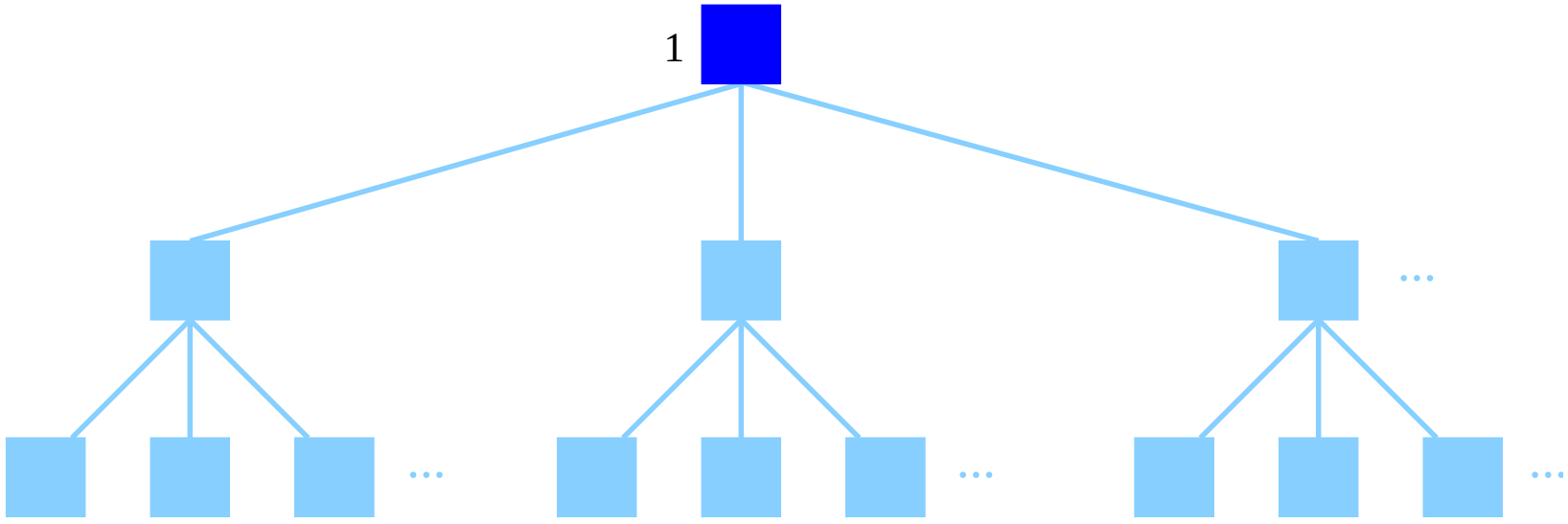
- One representation of trees is as nested partitions
- Suggests method for defining prior on trees:
  the *nested* Chinese restaurant process

Nested Chinese restaurant process



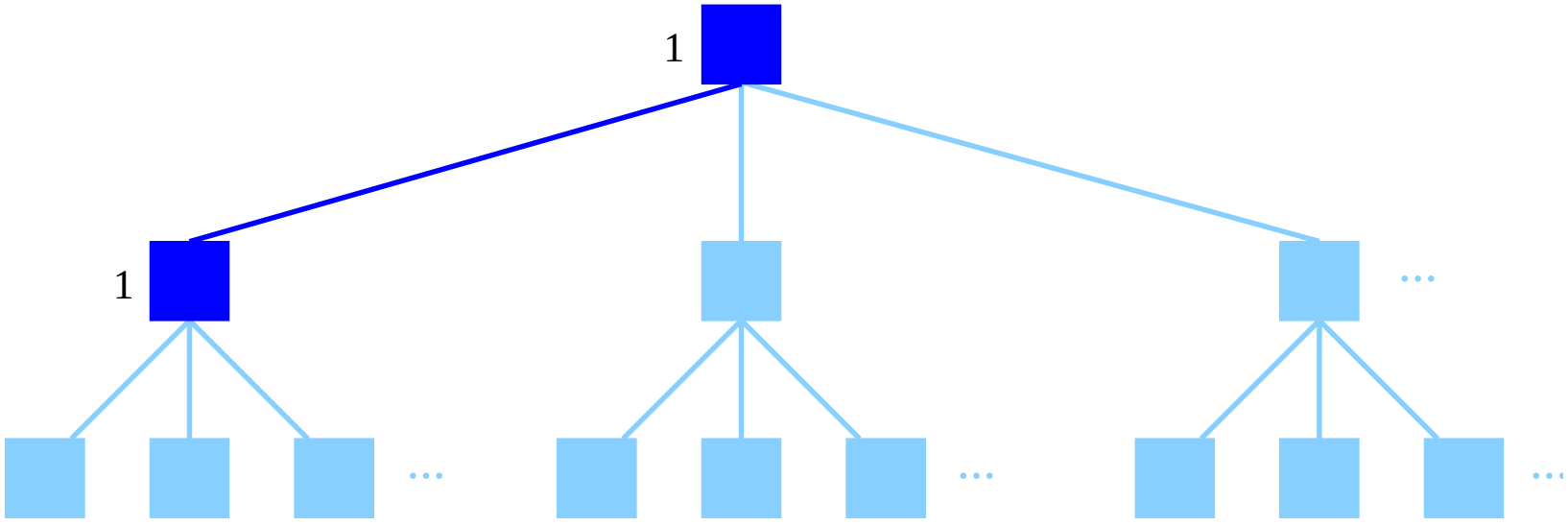
- Infinite number of Chinese restaurants in a city:
  - one restaurant is the root. On each of its infinite tables is a card with the name of another restaurant
  - on each of the tables in those restaurants are cards that refer to other restaurants, and this structure repeats
- Restaurants are organized into an infinitely branching tree

Nested Chinese restaurant process



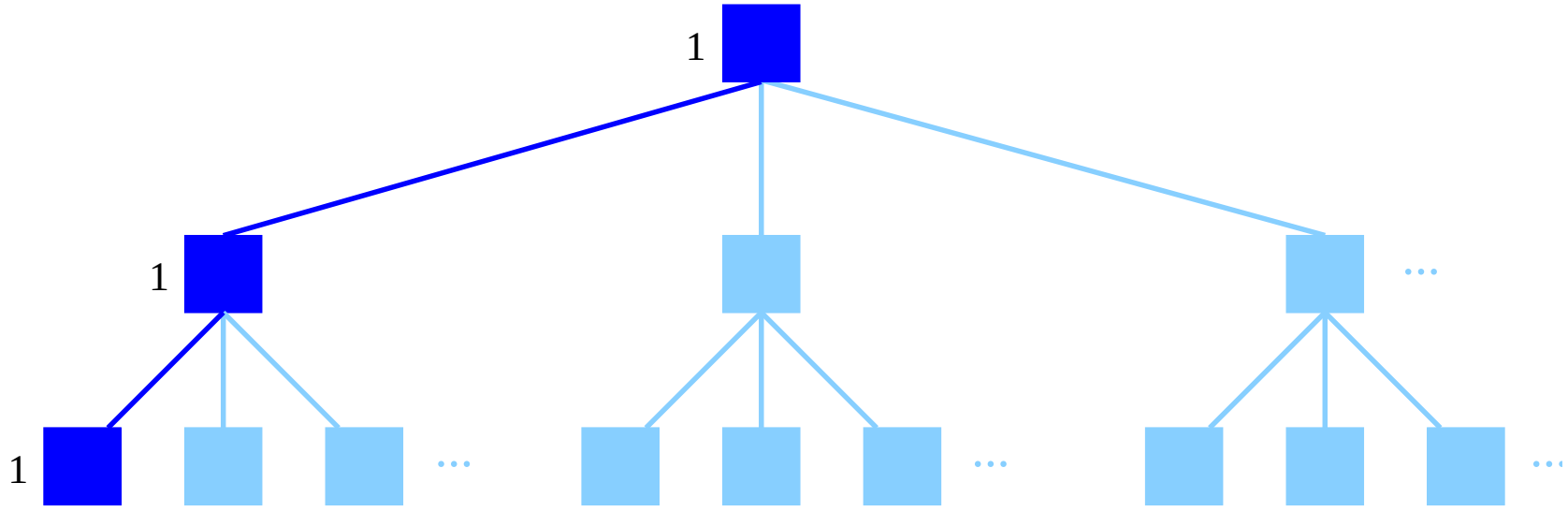
- A tourist arrives in the city for a culinary vacation
  - on the first evening, he enters the root restaurant and chooses a table, taking the card on that table
  - on the second evening, he goes to the restaurant identified on the card and chooses another table
  - he repeats this process for L days

Nested Chinese restaurant process



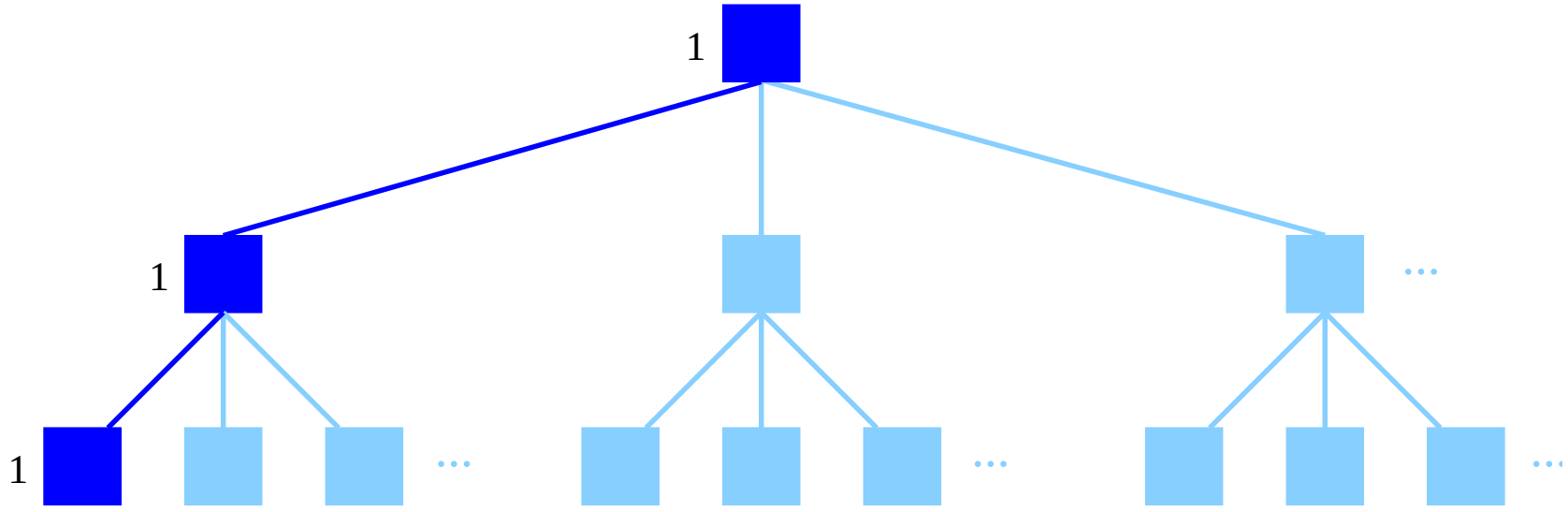
- A tourist arrives in the city for a culinary vacation
  - on the first evening, he enters the root restaurant and chooses a table, taking the card on that table
  - on the second evening, he goes to the restaurant identified on the card and chooses another table
  - he repeats this process for L days

Nested Chinese restaurant process



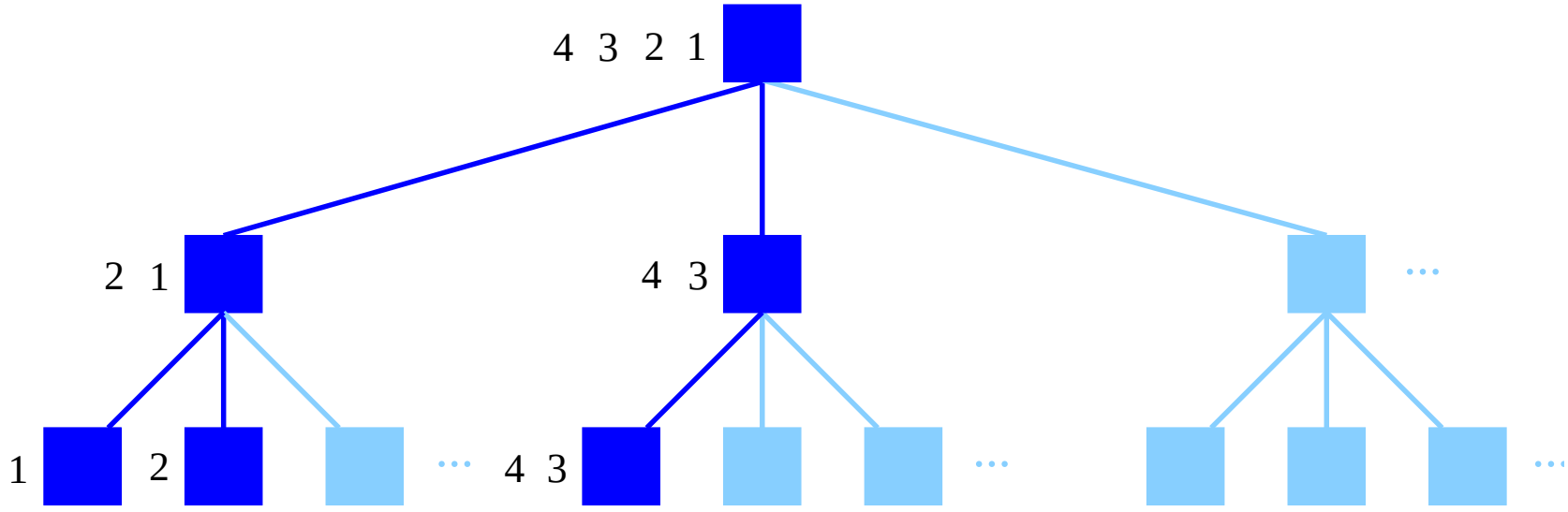
- A tourist arrives in the city for a culinary vacation
  - on the first evening, he enters the root restaurant and chooses a table, taking the card on that table
  - on the second evening, he goes to the restaurant identified on the card and chooses another table
  - he repeats this process for L days

Nested Chinese restaurant process



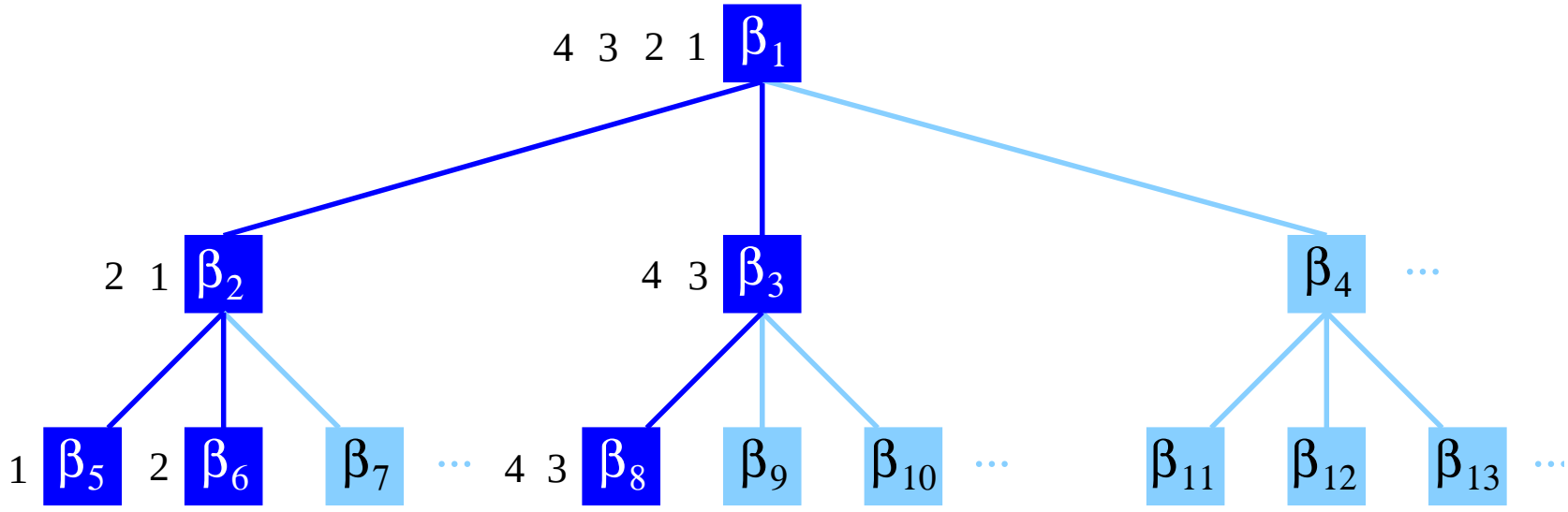
- The L chosen restaurants constitute a path from the root to a restaurant at the Lth level of the infinite tree
- After N tourists take L-day vacations, the collection of paths describe a particular L-level subtree of the infinite tree
- Can be used to define a statistical model in which each object is represented as a path through the tree

Nested Chinese restaurant process



- The L chosen restaurants constitute a path from the root to a restaurant at the Lth level of the infinite tree
- After N tourists take L-day vacations, the collection of paths describe a particular L-level subtree of the infinite tree
- Can be used to define a statistical model in which each object is represented as a path through the tree

Nested Chinese restaurant process



- The L chosen restaurants constitute a path from the root to a restaurant at the Lth level of the infinite tree
- After N tourists take L-day vacations, the collection of paths describe a particular L-level subtree of the infinite tree
- Can be used to define a statistical model in which each object is represented as a path through the tree

Latent Dirichlet Allocation (LDA)

- Model for text collections (Blei, Ng, & Jordan, 2003)
- Each document is a mixture of topics

$$P(w_i) = \sum_{k=1}^{K} P(w_i|z_i = k)P(z_i = k)$$

- Mixture weights (θ) vary across documents

$$\begin{aligned} \theta &\sim \text{Dirichlet}(\alpha) \\ z_i|\theta &\sim \text{Discrete}(\theta) \\ w_i|z_i &\sim \text{Discrete}(\beta_{z_i}) \\ \beta_k &\sim \text{Dirichlet}(\eta) \end{aligned}$$

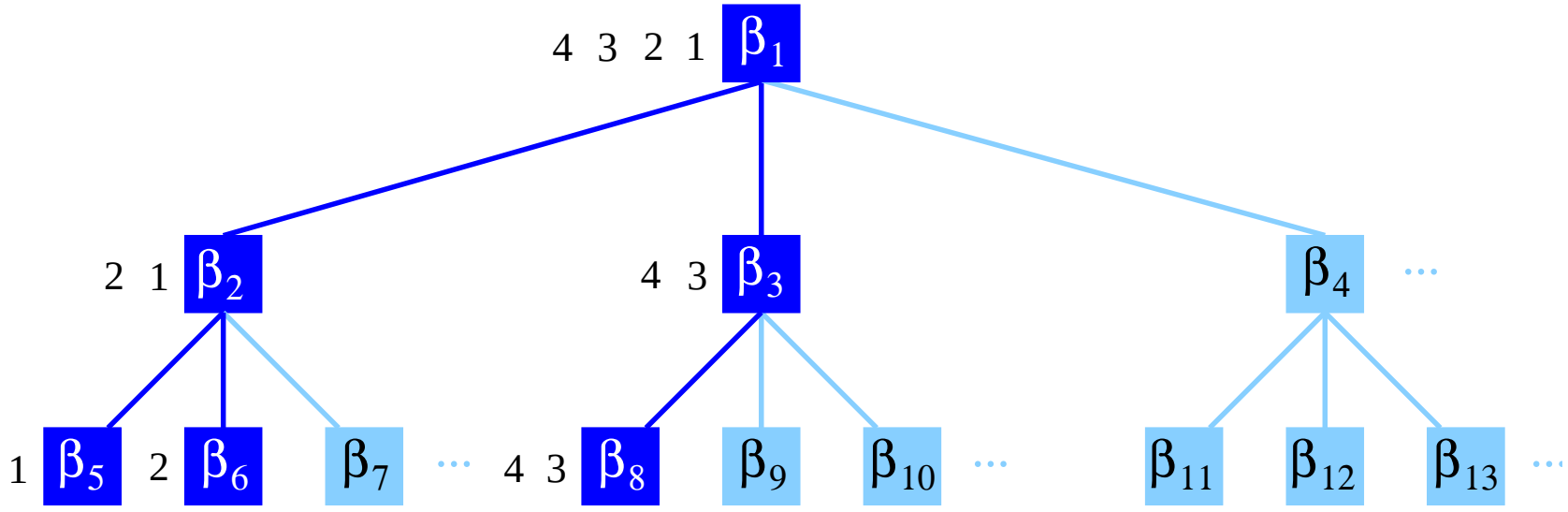
- c.f. grade-of-membership models (Erosheva, 2002)

Psychological Review topics

SIMILARITY	STIMULUS	SCALE	EMOTIONAL	MEMORY	PROCESSING	GROUP
CATEGORY	CONDITIONING	POWER	EMOTION	RETRIEVAL	MODEL	INDIVIDUAL
CATEGORIES	LEARNING	DISCRIMINATION	BASIC	RECALL	READING	GROUPS
RELATIONS	RESPONSE	LAW	EMOTIONS	ITEMS	WORD	OUTCOMES
DIMENSIONS	STIMULI	FUNCTION	AFFECT	INFORMATION	SEMANTIC	INDIVIDUALS
FEATURES	RESPONSES	PSYCHOPHYSICAL	STATES	TERM	LEXICAL	DIFFERENCES
STRUCTURE	AVOIDANCE	RATIO	EXPERIENCES	RECOGNITION	LANGUAGE	INTERACTION
SIMILAR	REINFORCEMENT	SENSORY	AFFECTIVE	ITEM	ACTIVATION	SOCIAL
REPRESENTATION	CLASSICAL	STIMULUS	AFFECTS	LIST	COMPREHENSION	PERSON
OBJECTS	DISCRIMINATION	FUNCTIONS	RESEARCH	ASSOCIATIVE	PHONOLOGICAL	LEVEL
PHENOMENA	GENERALIZATION	PHYSICAL	COGNITIVE	PROCESS	WORDS	MEMBERS
CONCEPTUAL	EFFECTS	SCALES	BIOLOGICAL	SERIAL	REPRESENTATION	CHANGE
CATEGORIZATION	EFFECT	MAGNITUDE	ESSENTIAL	STORAGE	NAMING	MATRIX
MATCHING	CS	MEASUREMENT	BASIS	SHORT	MODELS	LEVELS
REPRESENTED	CONTROL	RANGE	IDEA	EFFECTS	SENTENCE	CONTEXT
OBJECT	PHENOMENA	KNOWN	THEORY	TRACES	ACCOUNT	IMPRESSIONS
DIMENSIONAL	EXTINCTION	PSYCHOMETRIC	MOTIVATION	INTERFERENCE	PRODUCTION	FACTORS
MULTIDIMENSIONAL	ACQUISITION	VALUES	SYSTEMS	TRACE	SYNTACTIC	MAKING
ACCOUNT	CONDITIONED	PSYCHOLOGICAL	LOVE	POSITION	CONNECTIONIST	RISK
DIMENSION	TRAINING	OBSERVED	JAMES	STORED	NETWORK	CONSEQUENCES

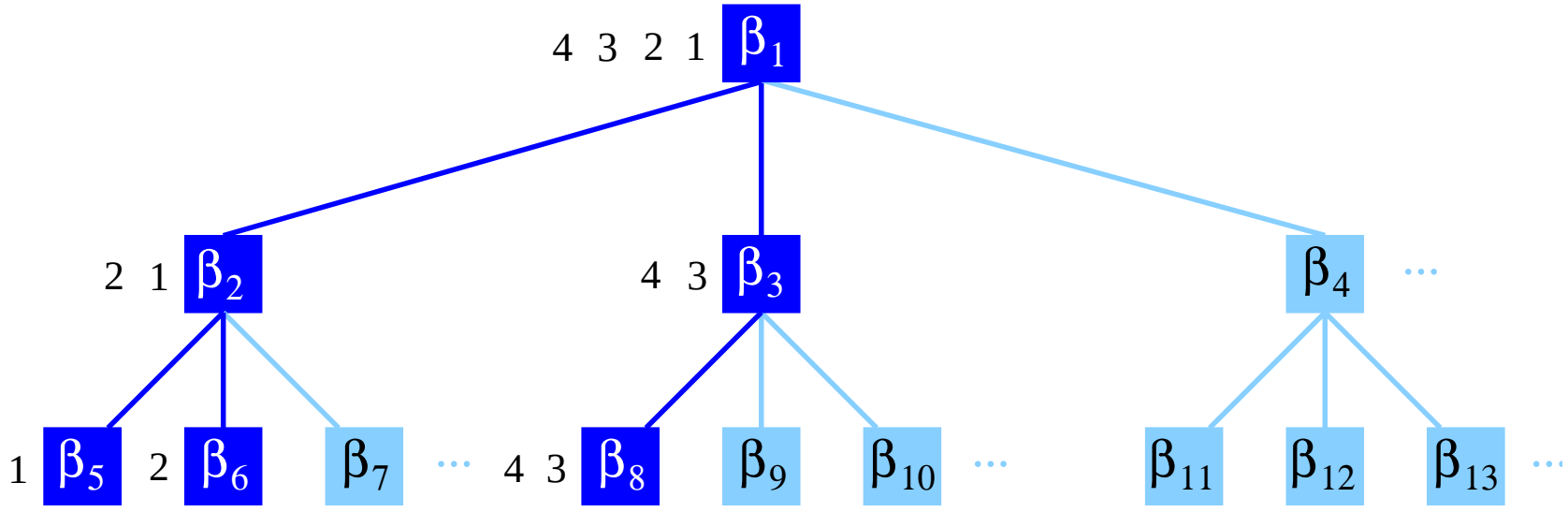
(words in each column are from one topic, sorted by β_k)

Hierarchical LDA



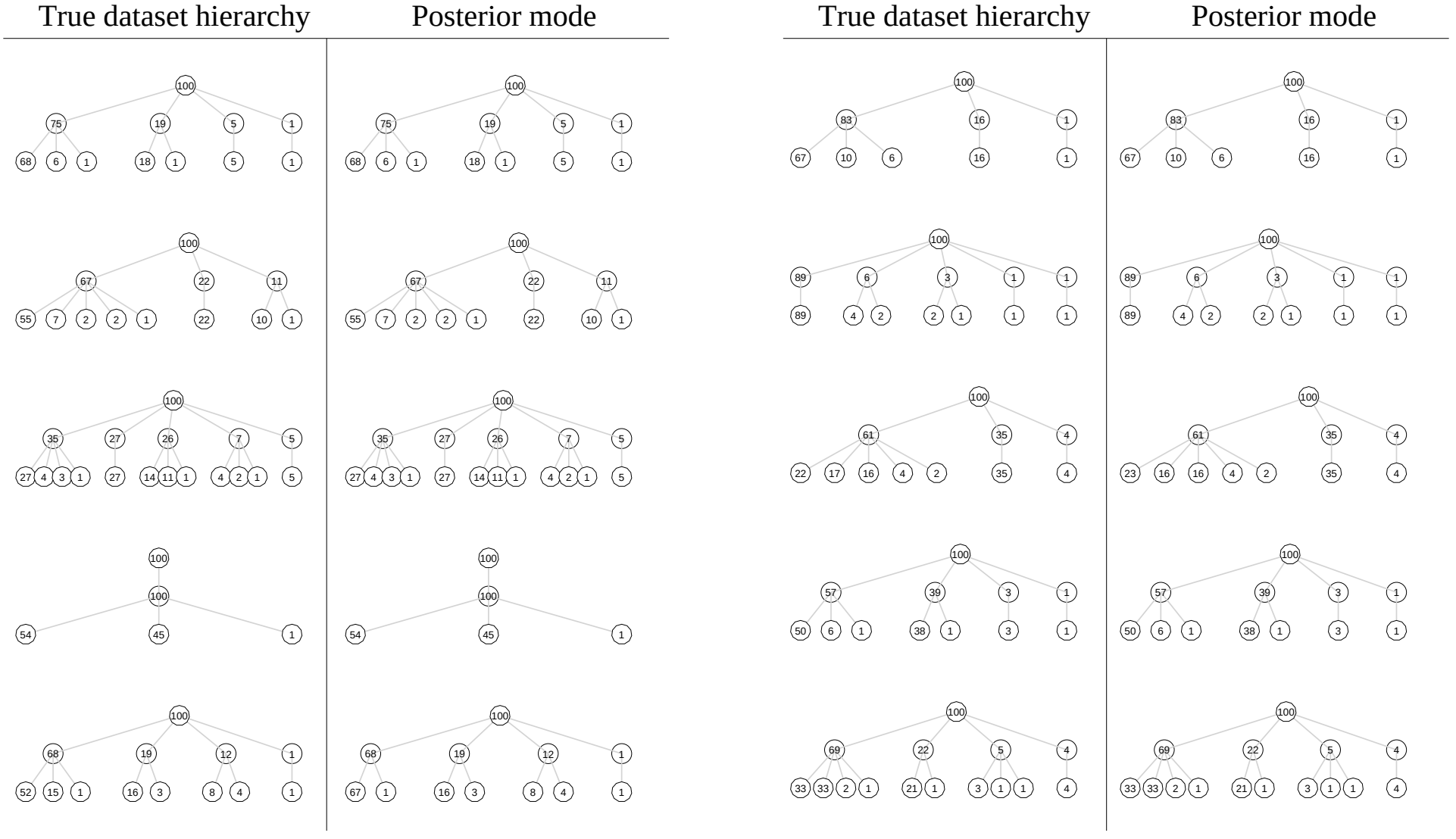
- Choose a path p through the infinite tree of restaurants
- Choose a distribution θ over levels
- For each word w_i
  - choose a level from $\mathrm{Discrete}(\theta)$
  - draw the word from the topic in the restaurant at that level

Hierarchical LDA

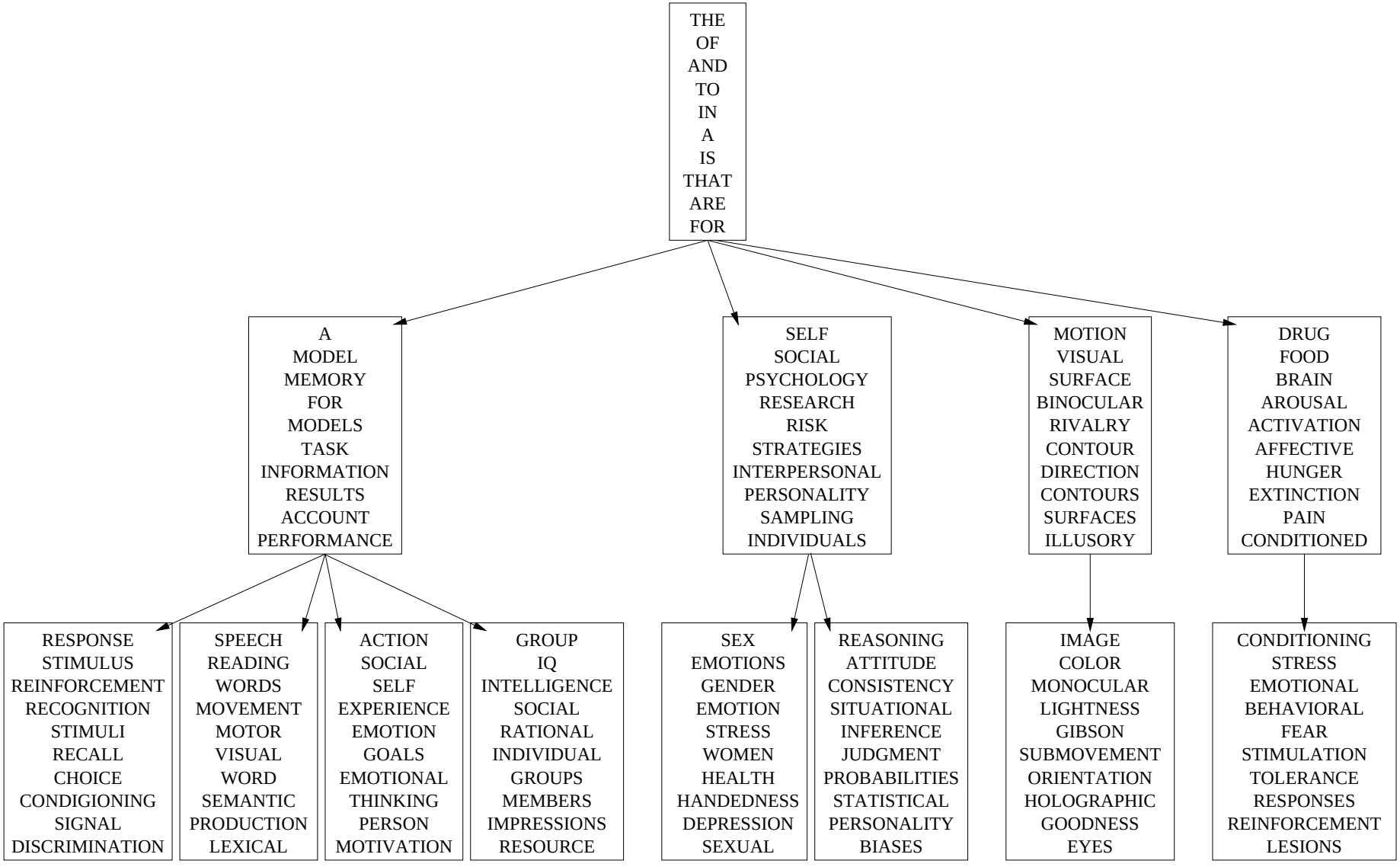


- Given a document collection, posterior is a distribution on
  - the structure of the hierarchy
  - assignment of documents to paths, words to levels
  - topics which populate the hierarchy
- Posterior inference via Gibbs sampling (on paths and $\mathbf{z}$)
- Allows new documents to fill unoccupied parts of the tree

Prior and posterior samples



Psychological Review hierarchy



Latent feature representations

- Many statistical models represent objects with latent features

Latent feature representations

- Many statistical models represent objects with latent features
  - binary features

Latent feature representations

- Many statistical models represent objects with latent features
  - binary features
  - factorial structures

Latent feature representations

- Many statistical models represent objects with latent features
  - binary features
  - factorial structures
  - continuous dimensions

Latent feature representations

- Many statistical models represent objects with latent features
  - binary features
  - factorial structures
  - continuous dimensions

- A common assumption: sparsity

Latent feature representations

- Many statistical models represent objects with latent features
  - binary features
  - factorial structures
  - continuous dimensions

- A common assumption: sparsity

- Define a prior for sparse latent feature representations by defining a prior on (infinite column) binary matrices

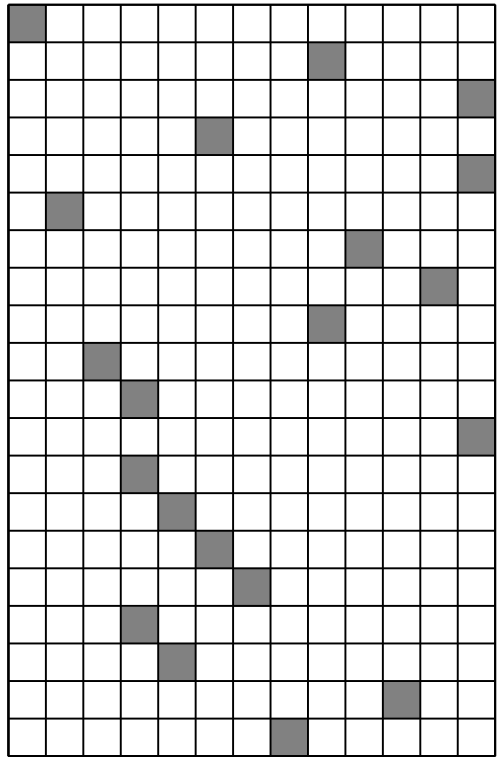
Priors on binary matrices

- Start with priors on $N \times K$ matrices, take $K \to \infty$

- Two cases:
  - "class matrices": one 1 per row
  - "feature matrices": general binary matrices

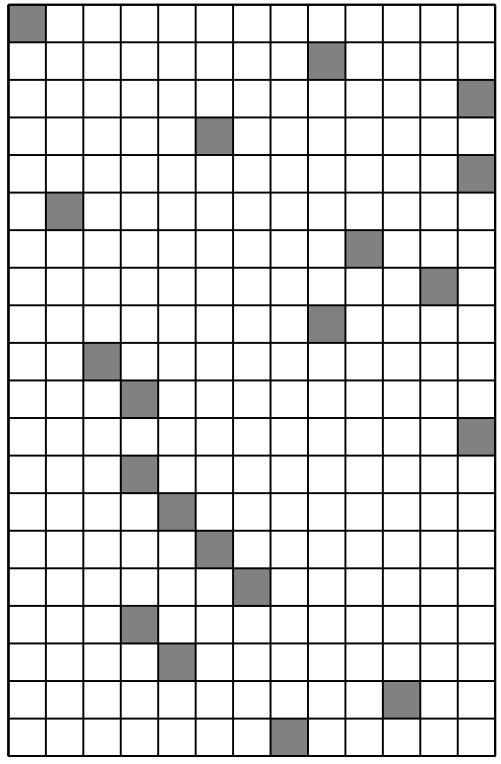
- Two priors:
  - the Chinese restaurant process
  - the Indian buffet process

Class matrices



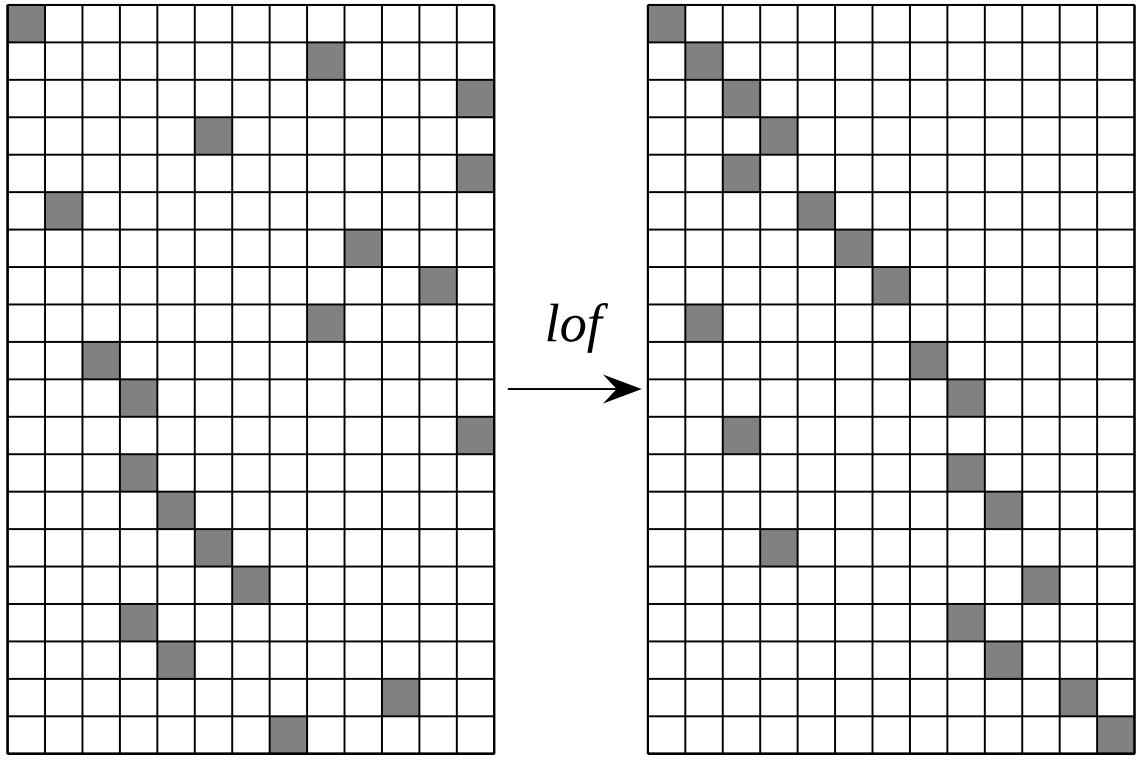
$$\begin{aligned} \mathbf{z}_i|\theta &\sim \text{Discrete}(\theta) \\ \theta &\sim \text{Dirichlet}(\alpha) \end{aligned}$$

Class matrices



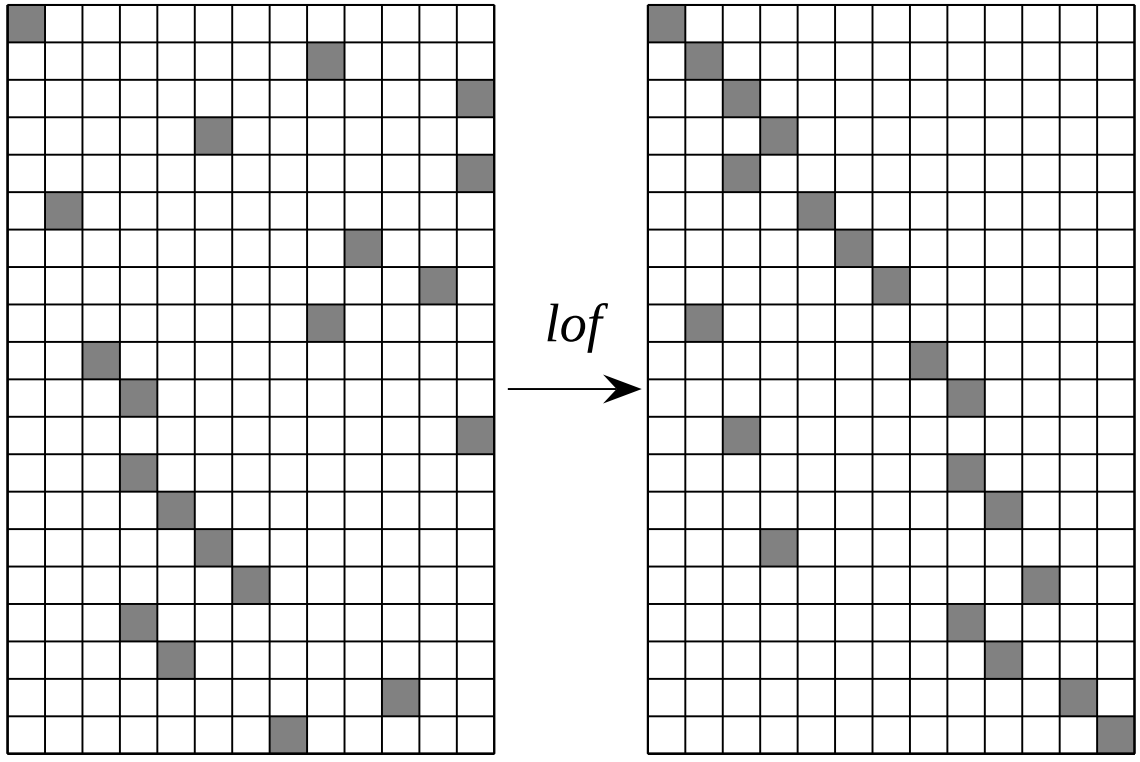
$$p(\mathbf{Z}) = \int_{\Delta} \prod_{i=1}^{N} p(\mathbf{z}_i|\theta)p(\theta)\ d\theta$$

Left-ordered form



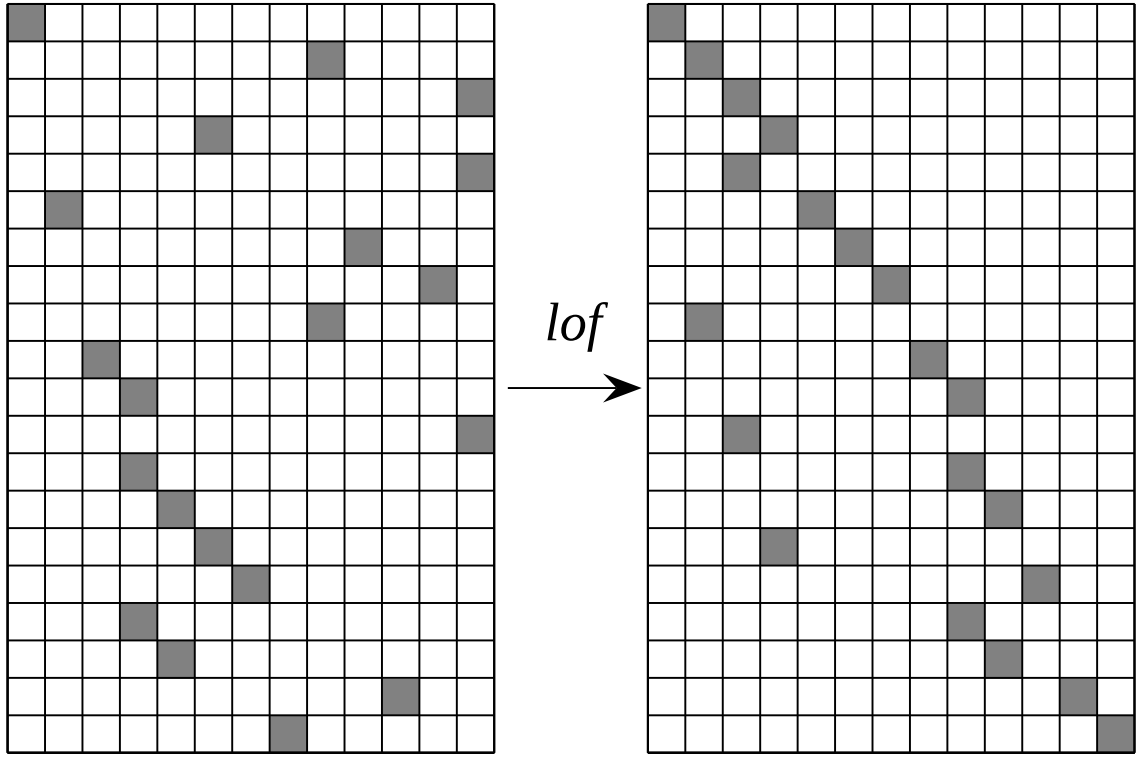
- History h of each class: binary column vector
- lof orders columns by values of binary histories

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



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

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

Feature matrices

- For general binary matrices

$$
\begin{aligned}
z_{ik} &\sim \text{Bernoulli}(\theta_k) \\
\theta_k &\sim \text{Beta}(\frac{\alpha}{K}, 1)
\end{aligned}
$$

Feature matrices

- For general binary matrices

$$\begin{aligned} z_{ik} &\sim \text{Bernoulli}(\theta_k) \\ \theta_k &\sim \text{Beta}(\frac{\alpha}{K}, 1) \end{aligned}$$

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

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

Feature matrices

- For general binary matrices

$$\begin{aligned} z_{ik} &\sim \text{Bernoulli}(\theta_k) \\ \theta_k &\sim \text{Beta}(\frac{\alpha}{K}, 1) \end{aligned}$$

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

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

- Taking the limit as $K \to \infty$ ...

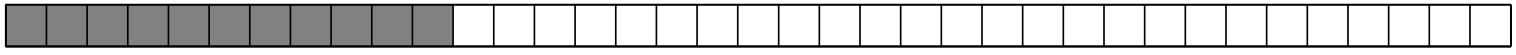
$$p([\mathbf{Z}]) = \exp\{-\alpha \sum_{i=1}^{N} \frac{1}{i}\} \frac{\alpha^{K_+}}{\prod_{h>0} K_h!} \prod_{k \leq K_+} \frac{(N - m_k)!(m_k - 1)!}{N!}$$

Indian buffet process (IBP)

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

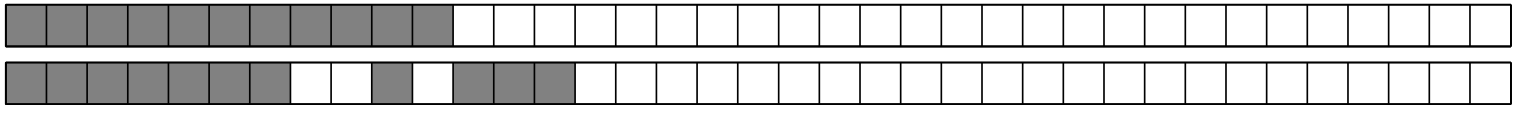
Indian buffet process (IBP)

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



Indian buffet process (IBP)

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

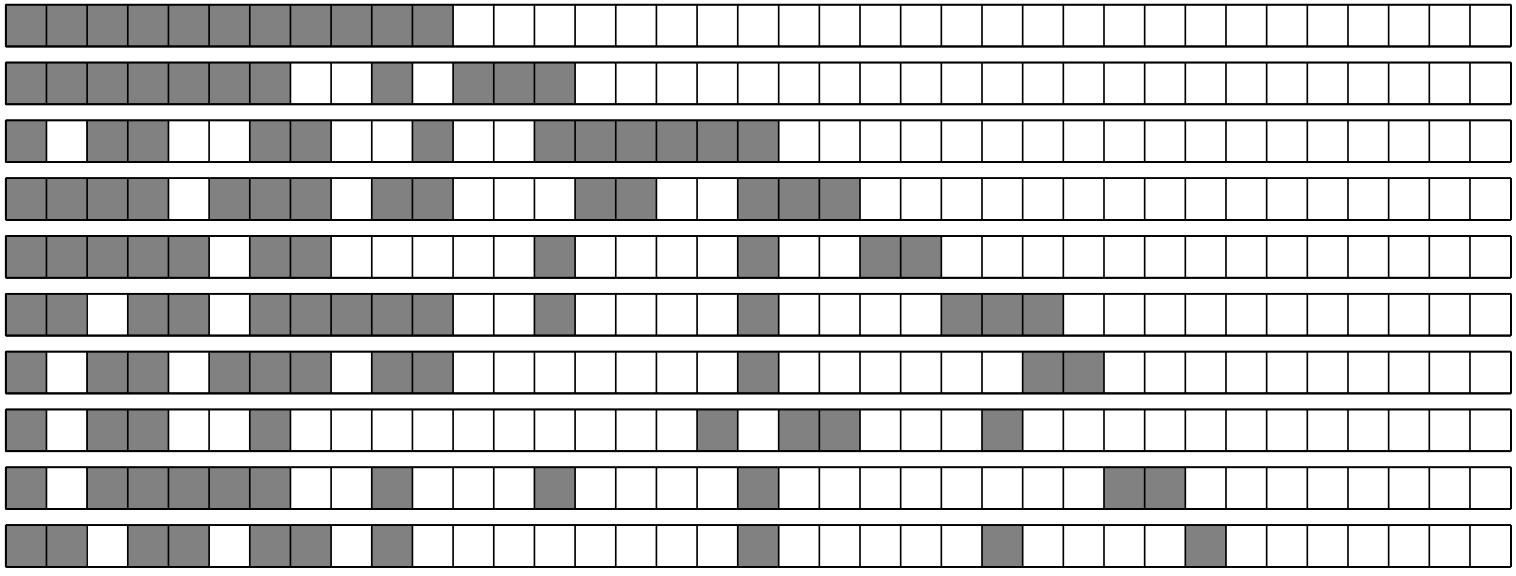


Indian buffet process (IBP)

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

Indian buffet process (IBP)

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



Another generating process

- lof-equivalence classes can be represented as vectors of history counts

$$\begin{array}{lcccccc} h & : & (& 1 & 2 & \cdots & 2^N - 1 &) \\ K_h & : & (& K_1 & K_2 & \cdots & K_{2^N-1} &) \end{array}$$

- Generate binary matrices by sampling K_h directly

$$K_h \sim \text{Poisson}(\alpha B(m_h, N - m_h + 1))$$

where $B(r, s)$ is the beta function

Properties of the IBP

- Exchangeable
- Total number of dishes $K^+ \sim \text{Poisson}(\alpha \sum_{i=1}^{N} \frac{1}{i})$
- Number of dishes sampled by each customer $\sim \text{Poisson}(\alpha)$
- Expected number of non-zero entries in $\mathbf{Z}$ is $N\alpha$

A linear-Gaussian model

- Data matrix $\mathbf{X} = [\mathbf{x}_1 \dots \mathbf{x}_N]^T$

$$\begin{aligned} \mathbf{x}_i &\sim \text{Gaussian}(\mathbf{z}_i\beta, \sigma_X\mathbf{I}) \\ \beta &\sim \text{Gaussian}(\sigma_\beta\mathbf{I}) \end{aligned}$$

- For $\mathbf{Z} \sim \text{CRP}(\alpha)$, spherical Gaussian mixture model
- For $\mathbf{Z} \sim \text{IBP}(\alpha)$, binary factor analysis
- Gibbs sampling

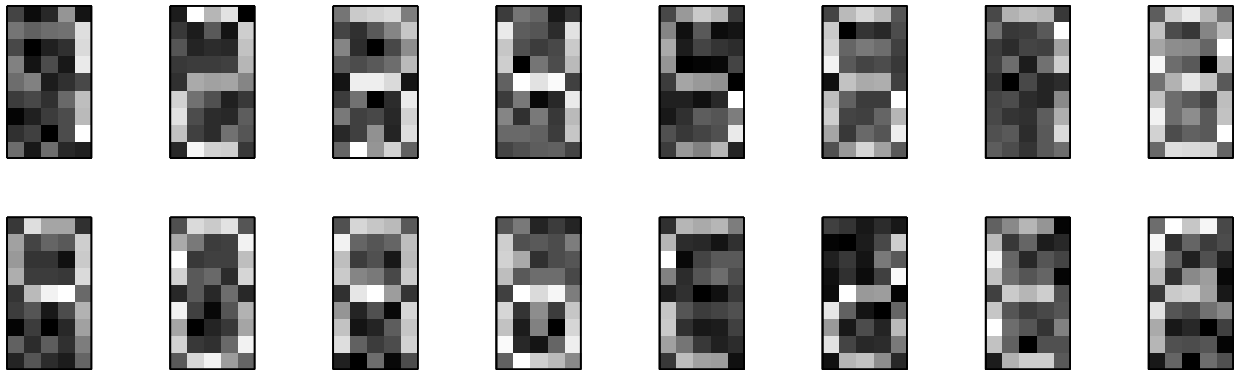
$$p(z_{ik}|\mathbf{X}, \mathbf{z}_{(-i)k}) \propto p(\mathbf{x}_i|\mathbf{X}_{-i}, \mathbf{Z})p(z_{ik}|\mathbf{z}_{(-i)k})$$

- Under the IBP

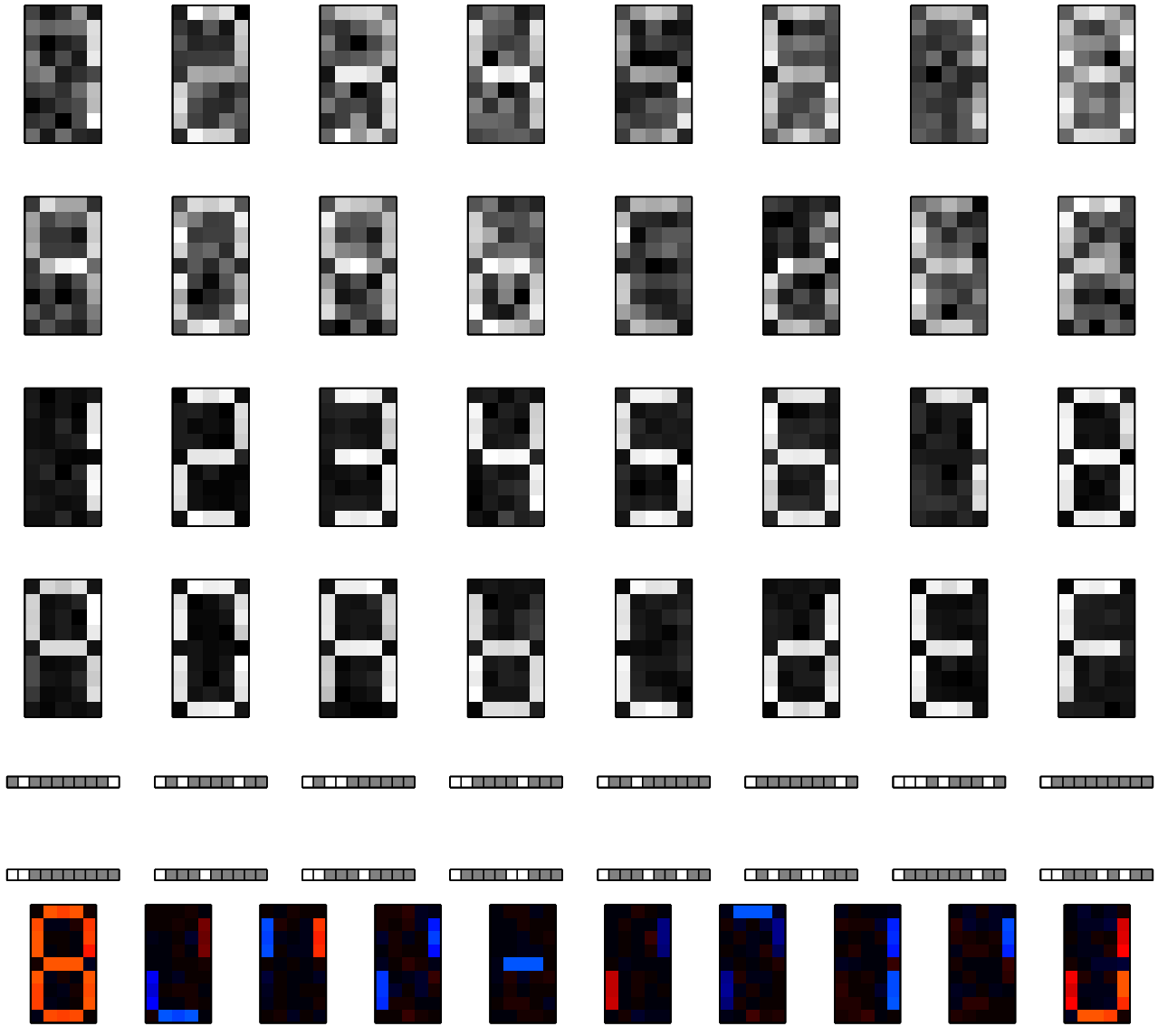
  $p(z_{ik}|\mathbf{z}_{(-i)k} = \frac{m_{k,-i}}{N}$

  $\text{Poisson}(\frac{\alpha}{N})$ prior on new features

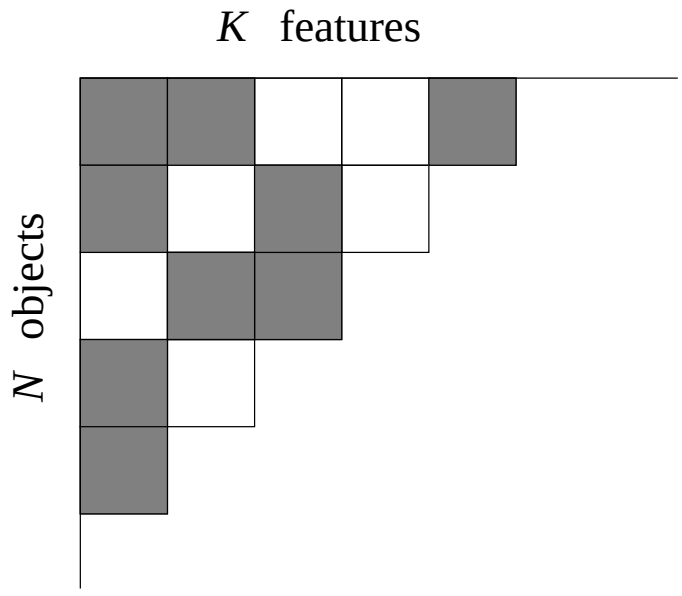
Seven segment LED displays



Seven segment LED displays

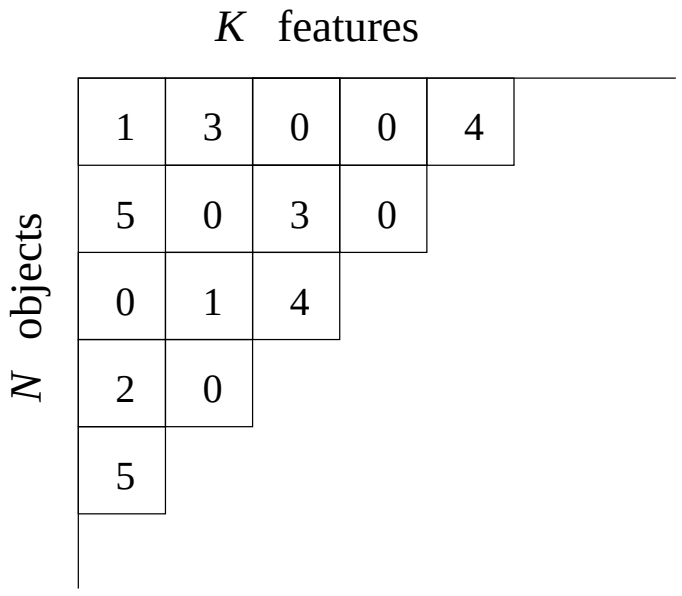


Different feature representations



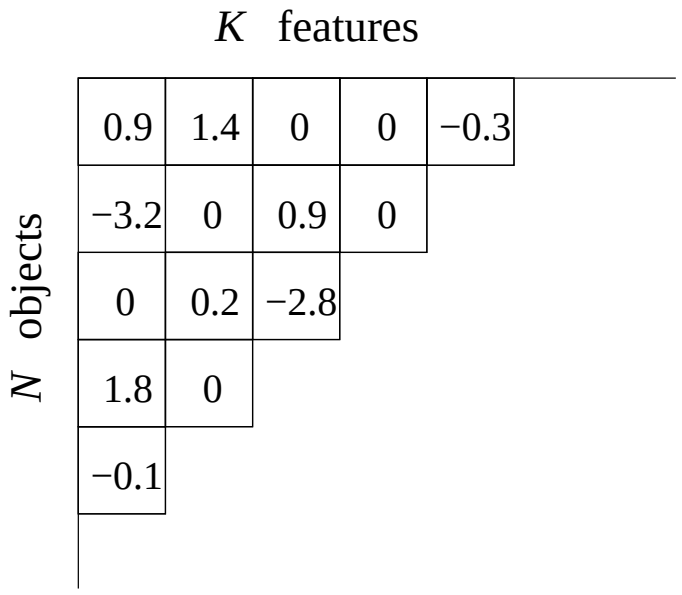
- Binary features

Different feature representations



- Binary features
- Factorial features

Different feature representations



- Binary features
- Factorial features
- Continuous features

Conclusion

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

- For mixture models, use the Chinese restaurant process
  - exchangeable distribution over partitions

- Same strategy can be extended to other representations
  - trees: nested Chinese restaurant process
  - binary matrices: Indian buffet process

- Still need good inference algorithms. . .