Applied Nonparametric Bayes

Michael I. Jordan

Department of Electrical Engineering and Computer Science
Department of Statistics
University of California, Berkeley

http://www.cs.berkeley.edu/∼jordan

Acknowledgments: Yee Whye Teh, Romain Thibaux

Computer Science and Statistics

- Separated in the 40's and 50's, but merging in the 90's and 00's
- What computer science has done well: data structures and algorithms for manipulating data structures
- What statistics has done well: managing uncertainty and justification of algorithms for making decisions under uncertainty
- What machine learning attempts to do: hasten the merger along

Nonparametric Bayesian Inference (Theme I)

- At the core of Bayesian inference lies Bayes' theorem:

$$posterior \propto likelihood \times prior$$

- For parametric models, we let θ be a Euclidean parameter and write:

$$p(\theta|x) \propto p(x|\theta)p(\theta)$$

- For nonparametric models, we let G be a general stochastic process (an "infinite-dimensional random variable") and write:

$$p(G|x) \propto p(x|G)p(G)$$

which frees us to work with flexible data structures

Nonparametric Bayesian Inference (cont)

- Examples of stochastic processes we'll mention today include distributions on:
  - directed trees of unbounded depth and unbounded fan-out
  - partitions
  - sparse binary infinite-dimensional matrices
  - copulae
  - distributions

- A general mathematical tool: Lévy processes

Hierarchical Bayesian Modeling (Theme II)

- Hierarchical modeling is a key idea in Bayesian inference

- It's essentially a form of recursion
  - in the parametric setting, it just means that priors on parameters can themselves be parameterized
  - in our nonparametric setting, it means that a stochastic process can have as a parameter another stochastic process

- We'll use hierarchical modeling to build structured objects that are reminiscent of graphical models—but are nonparametric!
  - statistical justification—the freedom inherent in using nonparametrics needs the extra control of the hierarchy

What are "Parameters"?

- *Exchangeability*: invariance to permutation of the joint probability distribution of infinite sequences of random variables

  Theorem (De Finetti, 1935). *If* $(x_1, x_2, \ldots)$ *are infinitely exchangeable, then the joint probability* $p(x_1, x_2, \ldots, x_N)$ *has a representation as a mixture*:

$$p(x_1, x_2, \ldots, x_N) = \int \left(\prod_{i=1}^{N} p(x_i \,|\, G) \right) dP(G)$$

  for some random element G.

- The theorem would be false if we restricted ourselves to finite-dimensional G

Stick-Breaking

- A general way to obtain distributions on countably-infinite spaces

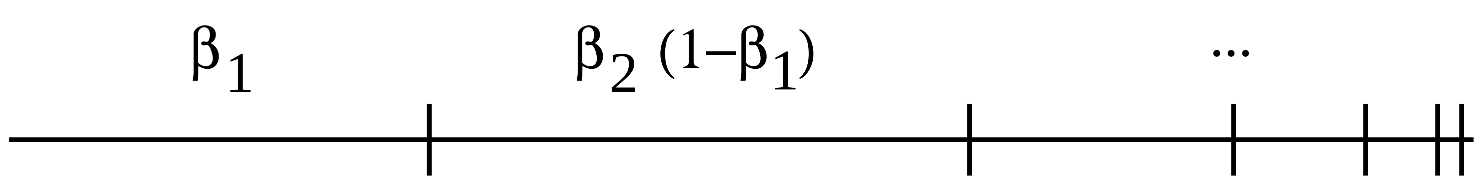
- *A canonical example*: Define an infinite sequence of beta random variables:

$$\beta_k \sim \mathrm{Beta}(1, \alpha_0) \qquad\qquad k = 1, 2, \ldots$$

- And then define an infinite random sequence as follows:

$$\pi_1 = \beta_1, \qquad \pi_k = \beta_k \prod_{l=1}^{k-1}(1 - \beta_l) \qquad\qquad k = 2, 3, \ldots$$

- This can be viewed as breaking off portions of a stick:



Constructing Random Measures

- It's not hard to see that $\sum_{k=1}^{\infty} \pi_k = 1$
- Now define the following object:

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k},$$

  where ϕ_k are independent draws from a distribution G_0 on some space

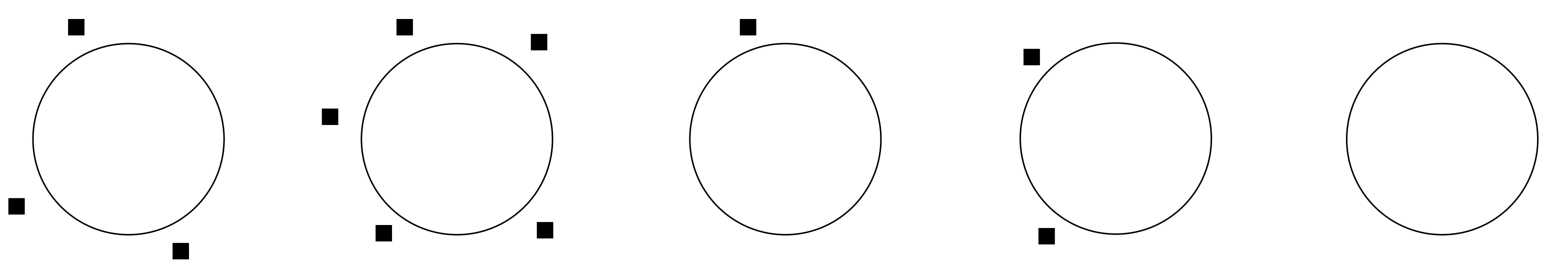
- Because $\sum_{k=1}^{\infty} \pi_k = 1$, G is a probability measure—it is a random measure
- The distribution of G is known as a Dirichlet process: $G \sim \mathrm{DP}(\alpha_0, G_0)$
- What exchangeable marginal distribution does this yield when integrated against in the De Finetti setup?

Chinese Restaurant Process (CRP)

- A random process in which n customers sit down in a Chinese restaurant with an infinite number of tables
  - first customer sits at the first table
  - mth subsequent customer sits at a table drawn from the following distribution:

$$\begin{aligned} P(\text{previously occupied table } i \,|\, \mathcal{F}_{m-1}) &\propto n_i \\ P(\text{the next unoccupied table} \,|\, \mathcal{F}_{m-1}) &\propto \alpha_0 \end{aligned} \tag{1}$$

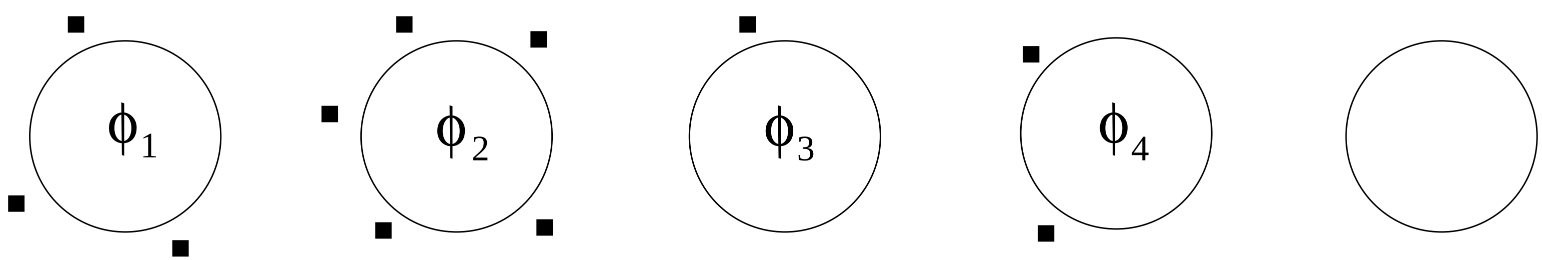
  where n_i is the number of customers currently at table i and where $\mathcal{F}_{m-1}$ denotes the state of the restaurant after $m-1$ customers have been seated



The CRP and Clustering

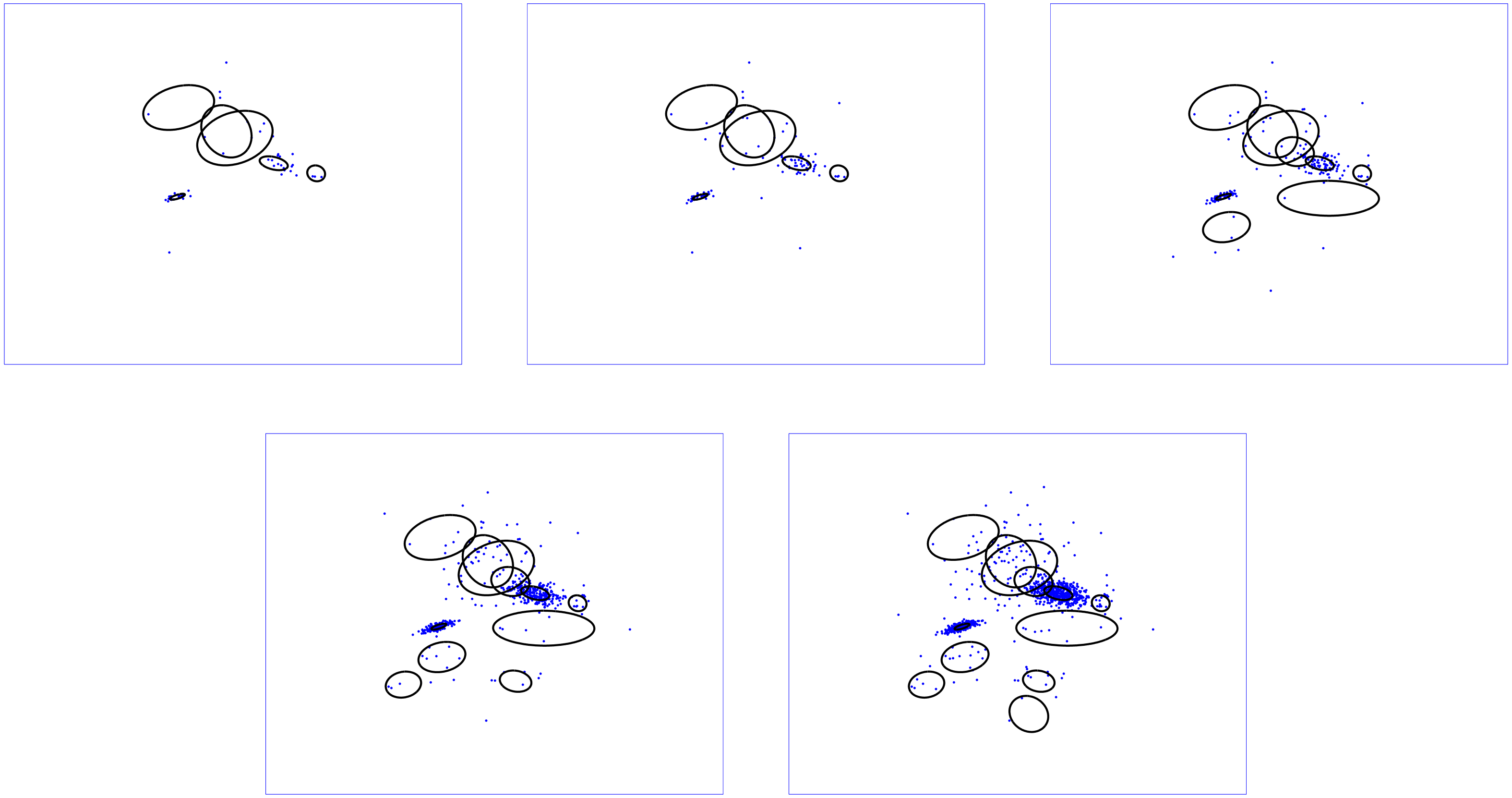
- Data points are customers; tables are clusters
  - the CRP defines a prior distribution on the partitioning of the data and on the number of tables

- This prior can be completed with:
  - a likelihood—e.g., associate a parameterized probability distribution with each table
  - a prior for the parameters—the first customer to sit at table k chooses the parameter vector for that table (ϕ_k) from a prior G_0



- So we now have a distribution—or can obtain one—for any quantity that we might care about in the clustering setting

CRP Prior, Gaussian Likelihood, Conjugate Prior



$$\begin{aligned} \phi_k &= (\mu_k, \Sigma_k) \sim N(a,b) \otimes IW(\alpha, \beta) \\ x_i &\sim N(\phi_k) \qquad \text{for a data point } i \text{ sitting at table } k \end{aligned}$$

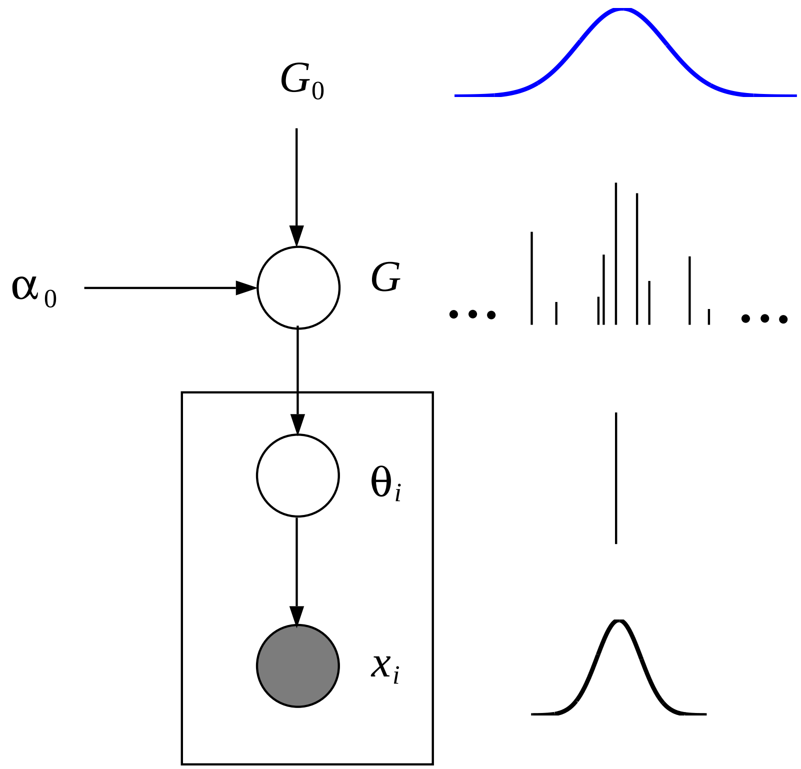
Exchangeability

- As a prior on the partition of the data, the CRP is exchangeable
- The prior on the parameter vectors associated with the tables is also exchangeable
- The latter probability model is generally called the Pólya urn model. Letting θ_i denote the parameter vector associated with the ith data point, we have:

$$\theta_i \,|\, \theta_1, \ldots, \theta_{i-1} \;\sim\; \alpha_0 G_0 + \sum_{j=1}^{i-1} \delta_{\theta_j}$$

- From these conditionals, a short calculation shows that the joint distribution for $(\theta_1, \ldots, \theta_n)$ is invariant to order (this is the exchangeability proof)
- As a prior on the number of tables, the CRP is nonparametric—the number of occupied tables grows (roughly) as $O(\log n)$—we're in the world of nonparametric Bayes

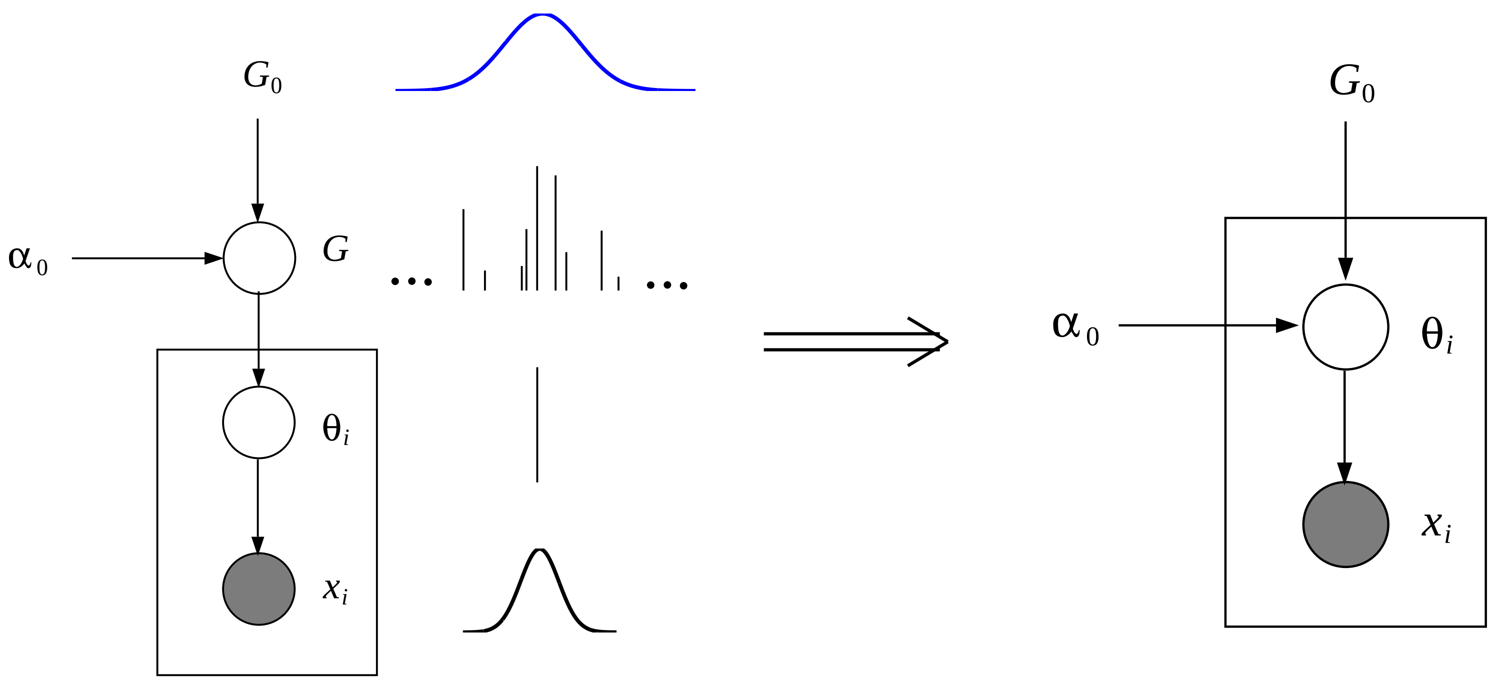
Dirichlet Process Mixture Models



$$
\begin{aligned}
G &\sim \mathrm{DP}(\alpha_0 G_0) & \\
\theta_i \,|\, G &\sim G & i \in 1, \ldots, n \\
x_i \,|\, \theta_i &\sim F(x_i \,|\, \theta_i) & i \in 1, \ldots, n
\end{aligned}
$$

Marginal Probabilities

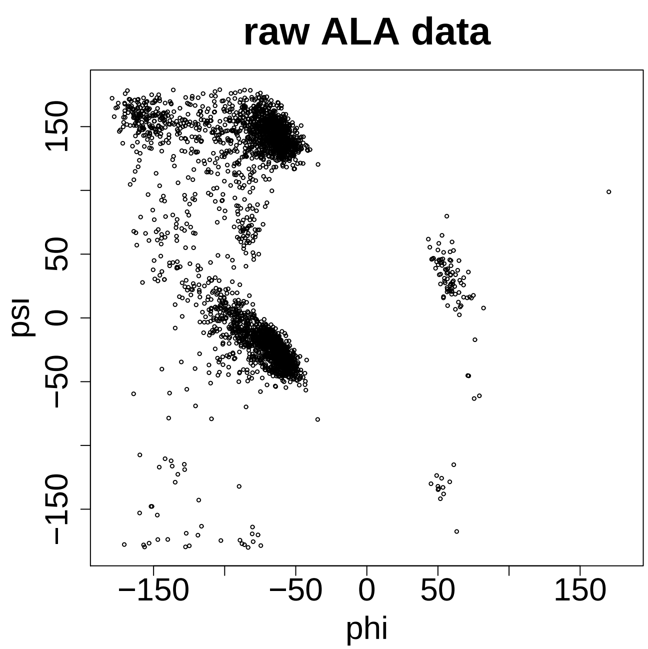
- To obtain the marginal probability of the parameters $\theta_1, \theta_2, \ldots$, we need to integrate out G



- This marginal distribution turns out to be the Chinese restaurant process (more precisely, it's the Pólya urn model)

Protein Folding

- A protein is a folded chain of amino acids
- The backbone of the chain has two degrees of freedom per amino acid (phi and psi angles)
- Empirical plots of phi and psi angles are called *Ramachandran diagrams*



Protein Folding (cont.)

- We want to model the density in the Ramachandran diagram to provide an energy term for protein folding algorithms

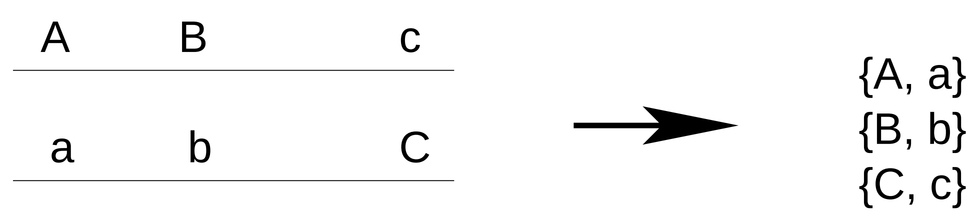
- We actually have a linked set of Ramachandran diagrams, one for each amino acid neighborhood

- We thus have a linked set of clustering problems

  – note that the data are *partially exchangeable*

Haplotype Modeling

- Consider M binary markers in a genomic region
- There are 2^M possible haplotypes—i.e., states of a single chromosome
  - but in fact, far fewer are seen in human populations
- A genotype is a set of unordered pairs of markers (from one individual)



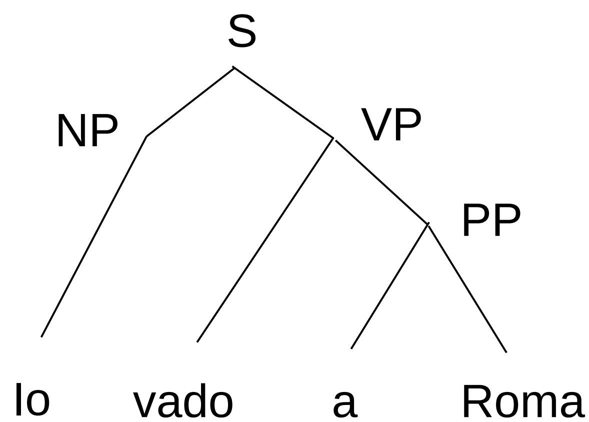
- Given a set of genotypes (multiple individuals), estimate the underlying haplotypes
- This is a clustering problem

Haplotype Modeling (cont.)

- A key problem is inference for the number of clusters
- Consider now the case of multiple groups of genotype data (e.g., ethnic groups)
- Geneticists would like to find clusters within each group but they would also like to share clusters between the groups

Natural Language Parsing

- Given a corpus of sentences, some of which have been parsed by humans, find a grammar that can be used to parse future sentences

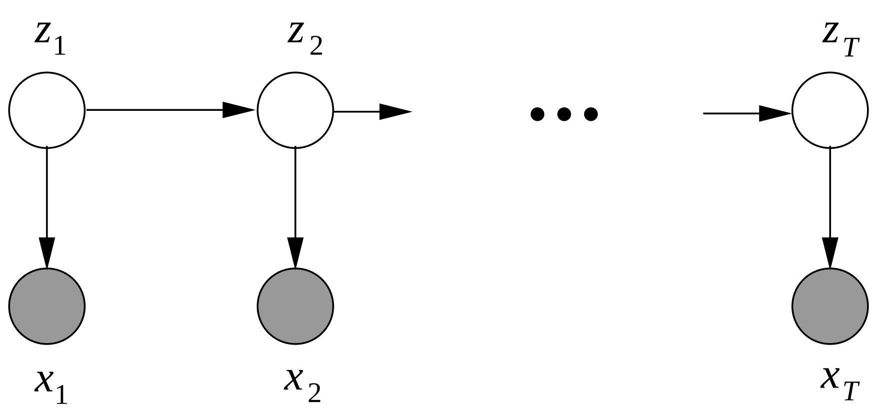


- Much progress over the past decade; state-of-the-art methods are statistical

Natural Language Parsing (cont.)

- Key idea: *lexicalization* of context-free grammars
  - the grammatical rules (S $\rightarrow$ NP VP) are conditioned on the specific lexical items (words) that they derive
- This leads to huge numbers of potential rules, and (adhoc) shrinkage methods are used to control the counts
- Need to control the numbers of clusters (model selection) in a setting in which many tens of thousands of clusters are needed
- Need to consider related groups of clustering problems (one group for each grammatical context)

Nonparametric Hidden Markov Models



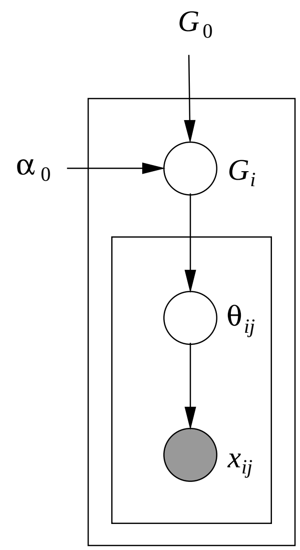
- An open problem—how to work with HMMs and state space models that have an unknown and unbounded number of states?
- Each row of a transition matrix is a probability distribution across "next states"
- We need to estimation these transitions in a way that links them across rows

Image Segmentation

- Image segmentation can be viewed as inference over partitions
  - clearly we want to be nonparametric in modeling such partitions
- Standard approach—use relatively simple (parametric) local models and relatively complex spatial coupling
- Our approach—use a relatively rich (nonparametric) local model and relatively simple spatial coupling
  - for this to work we need to combine information across images; this brings in the hierarchy

Hierarchical Nonparametrics—A First Try

- Idea: Dirichlet processes for each group, linked by an underlying G_0:



- Problem: the atoms generated by the random measures G_i will be distinct
  - i.e., the atoms in one group will be distinct from the atoms in the other groups—no sharing of clusters!
- Sometimes ideas that are fine in the parametric context fail (completely) in the nonparametric context... :-(

Hierarchical Dirichlet Processes

(Teh, Jordan, Beal & Blei, 2006)

- We need to have the base measure G_0 be discrete
  - but also need it to be flexible and random

Hierarchical Dirichlet Processes

(Teh, Jordan, Beal & Blei, 2006)

- We need to have the base measure G_0 be discrete
  - but also need it to be flexible and random
- The fix: Let G_0 itself be distributed according to a DP:

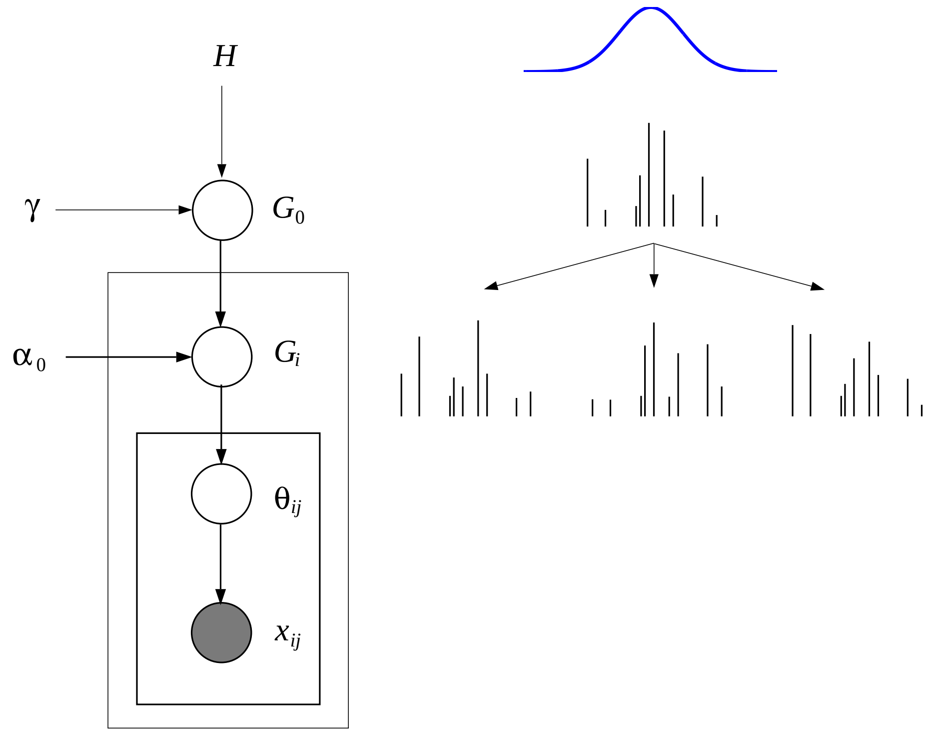
$$G_0 \,|\, \gamma, H \;\sim\; \mathrm{DP}(\gamma H)$$

- Then

$$G_j \,|\, \alpha, G_0 \;\sim\; \mathrm{DP}(\alpha_0 G_0)$$

has as its base measure a (random) atomic distribution—samples of G_j will resample from these atoms

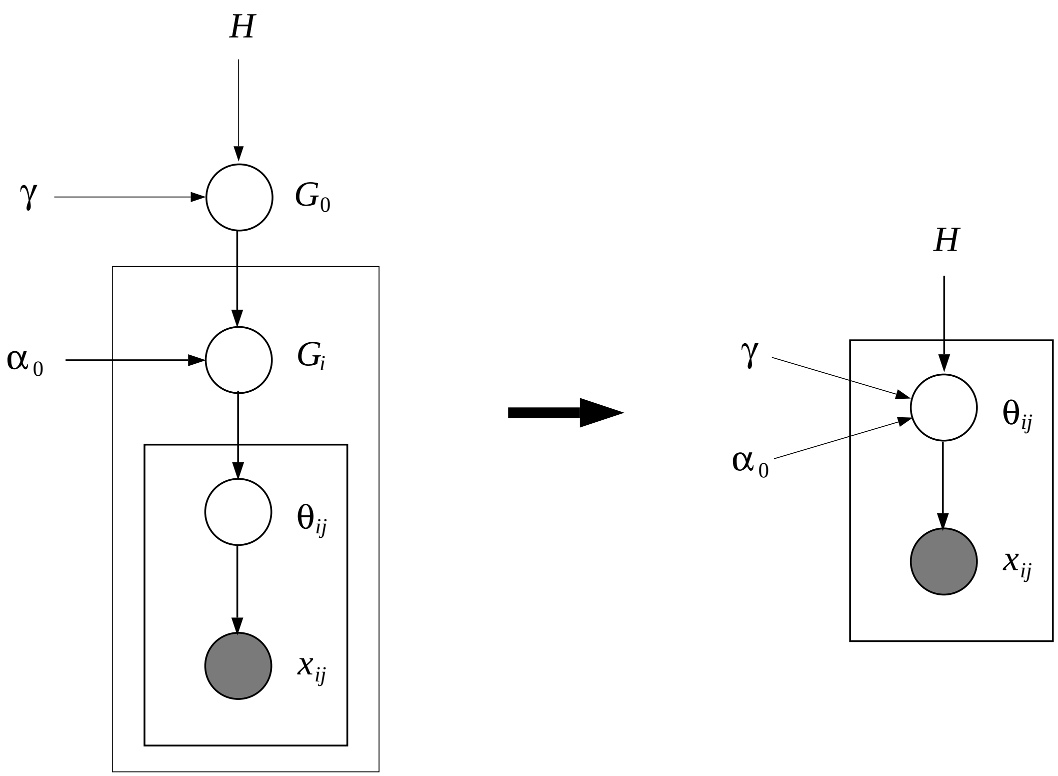
Hierarchical Dirichlet Process Mixtures



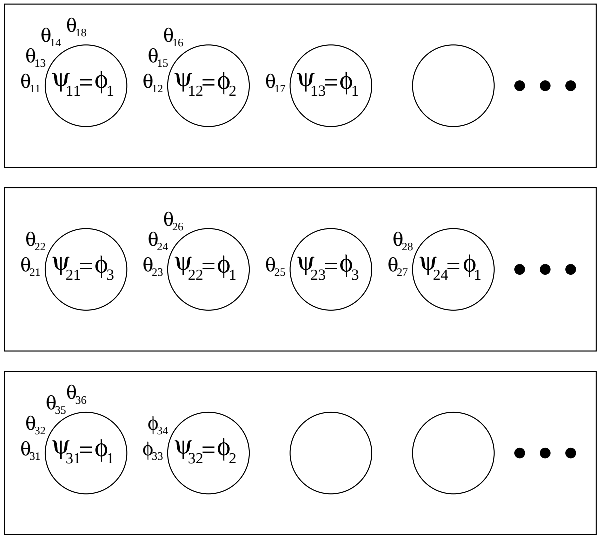
$$
\begin{aligned}
G_0 \,|\, \gamma, H &\sim \mathrm{DP}(\gamma H) \\
G_i \,|\, \alpha, G_0 &\sim \mathrm{DP}(\alpha_0 G_0) \\
\theta_{ij} \,|\, G_i &\sim G_i \\
x_{ij} \,|\, \theta_{ij} &\sim F(x_{ij}, \theta_{ij})
\end{aligned}
$$

Chinese Restaurant Franchise (CRF)

- First integrate out the G_i, then integrate out G_0



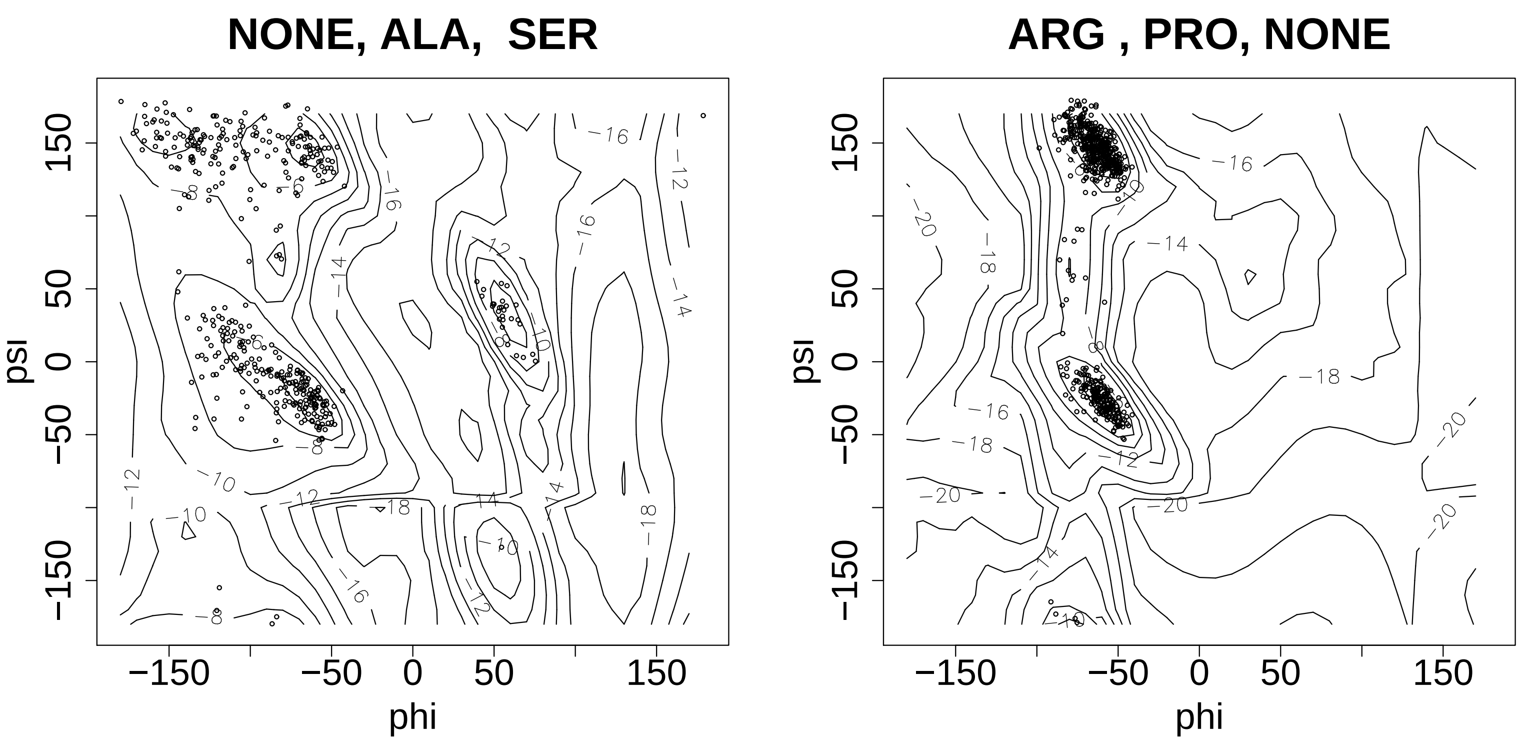
Chinese Restaurant Franchise (CRF)



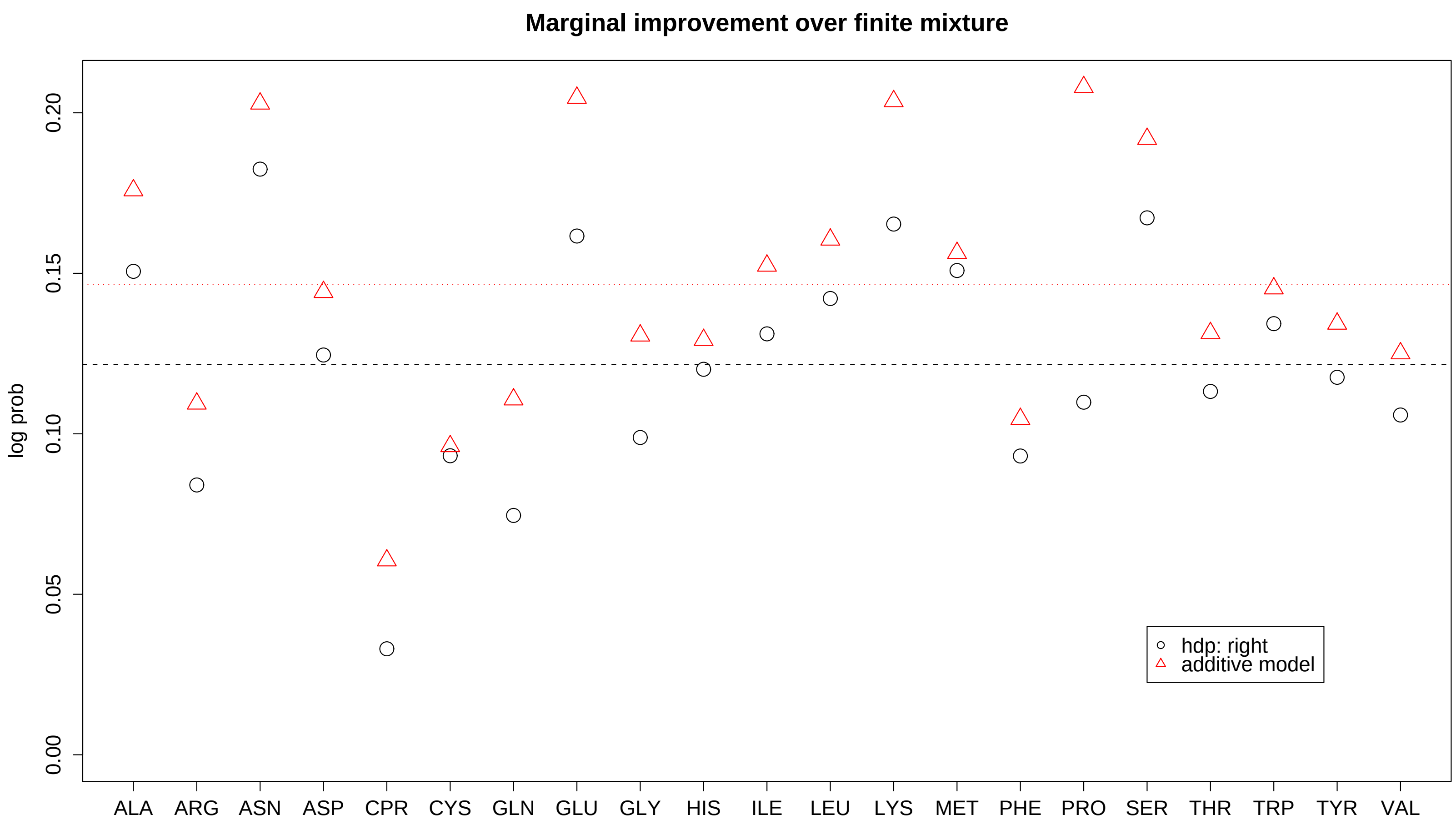
- To each group there corresponds a *restaurant*, with an unbounded number of *tables* in each restaurant
- There is a global *menu* with an unbounded number of *dishes* on the menu
- The first customer at a table selects a dish for that table from the global menu
- Reinforcement effects—customers prefer to sit at tables with many other customers, and prefer to choose dishes that are chosen by many other customers

Protein Folding (cont.)

- We have a linked set of Ramachandran diagrams, one for each amino acid neighborhood



Protein Folding (cont.)



Natural Language Parsing

- Key idea: *lexicalization* of context-free grammars
  - the grammatical rules (S $\rightarrow$ NP VP) are conditioned on the specific lexical items (words) that they derive

- This leads to huge numbers of potential rules, and (adhoc) shrinkage methods are used to control the choice of rules

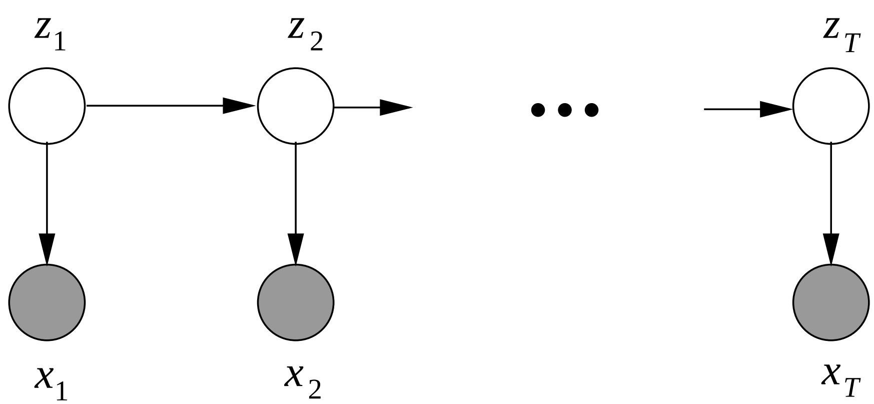
HDP-PCFG

(Liang, Petrov, Jordan & Klein, 2007)

- Based on a training corpus, we build a lexicalized grammar in which the rules are based on word clusters
- Each grammatical context defines a clustering problem, and we link the clustering problems via the HDP

T	PCFG		HDP-PCFG	
	F_1	Size	F_1	Size
1	60.4	2558	60.5	2557
4	76.0	3141	77.2	9710
8	74.3	4262	79.1	50629
16	66.9	19616	78.2	151377
20	64.4	27593	77.8	202767

Nonparametric Hidden Markov models



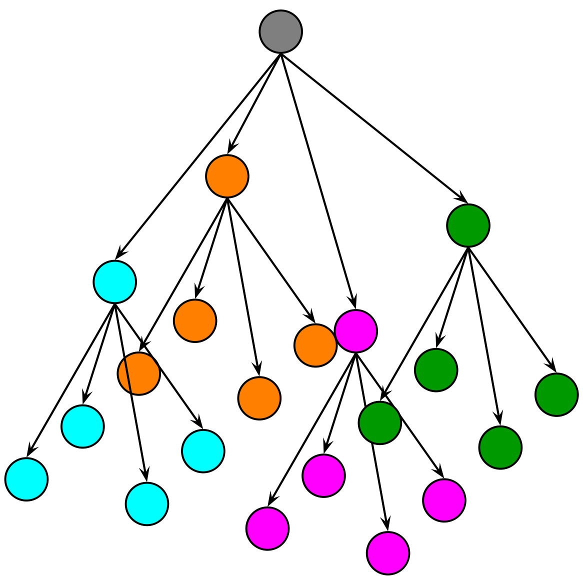
- A perennial problem—how to work with HMMs that have an unknown and unbounded number of states?

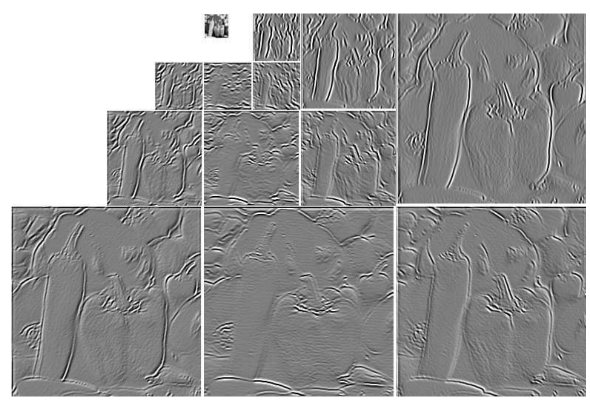
- A straightforward application of the HDP framework
  - multiple mixture models—one for each value of the "current state"
  - the DP creates new states, and the HDP approach links the transition distributions

Nonparametric Hidden Markov Trees

(Kivinen, Sudderth & Jordan, 2007)

- Hidden Markov trees in which the cardinality of the states is unknown a priori
- We need to tie the parent-child transitions across the parent states; this is done with the HDP





Nonparametric Hidden Markov Trees (cont.)

- Local Gaussian Scale Mixture (31.84 dB)

Nonparametric Hidden Markov Trees (cont.)

- Hierarchical Dirichlet Process Hidden Markov Tree (32.10 dB)

Image Segmentation

(Sudderth & Jordan, 2008)

- Image segmentation can be viewed as inference over partitions
  - clearly we want to be nonparametric in modeling such partitions

- Image statistics are better captured by the Pitman-Yor stick-breaking processes than by the Dirichlet process

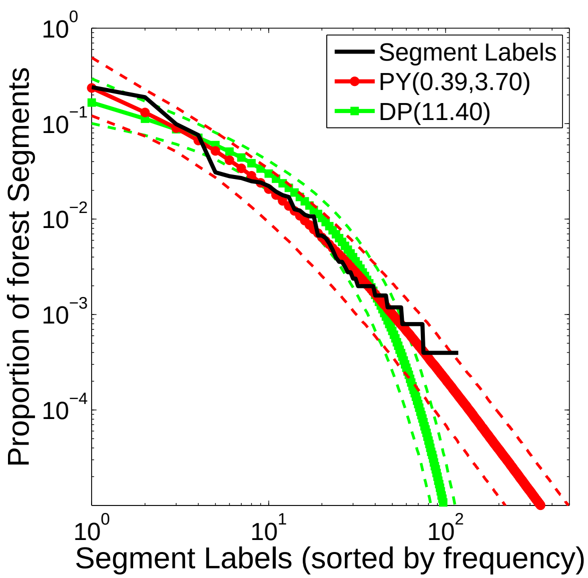


Image Segmentation (cont)

(Sudderth & Jordan, 2008)

- So we want Pitman-Yor marginals at each site in an image
- The (perennial) problem is how to couple these marginals spatially
  - to solve this problem, we again go nonparametric—we couple the PY marginals using Gaussian process copulae

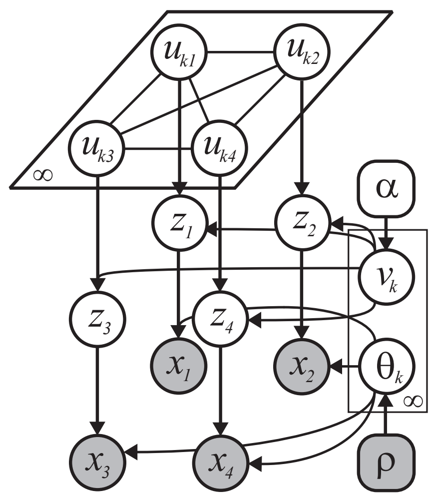


Image Segmentation (cont)

(Sudderth & Jordan, 2008)

- A sample from the coupled HPY prior:

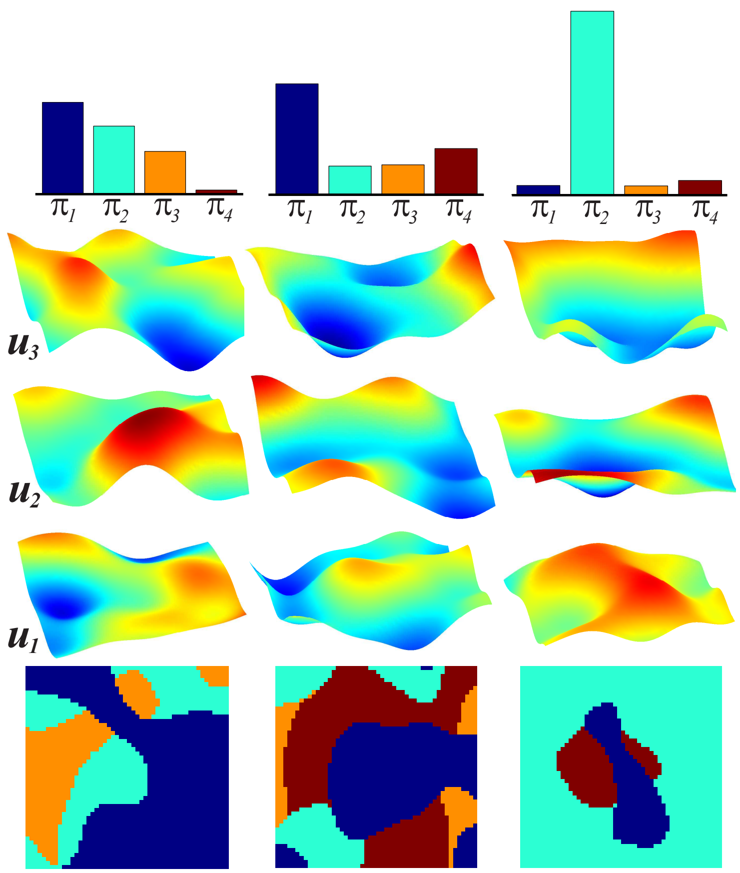


Image Segmentation (cont)

(Sudderth & Jordan, 2008)

- Comparing the HPY prior to a Markov random field prior

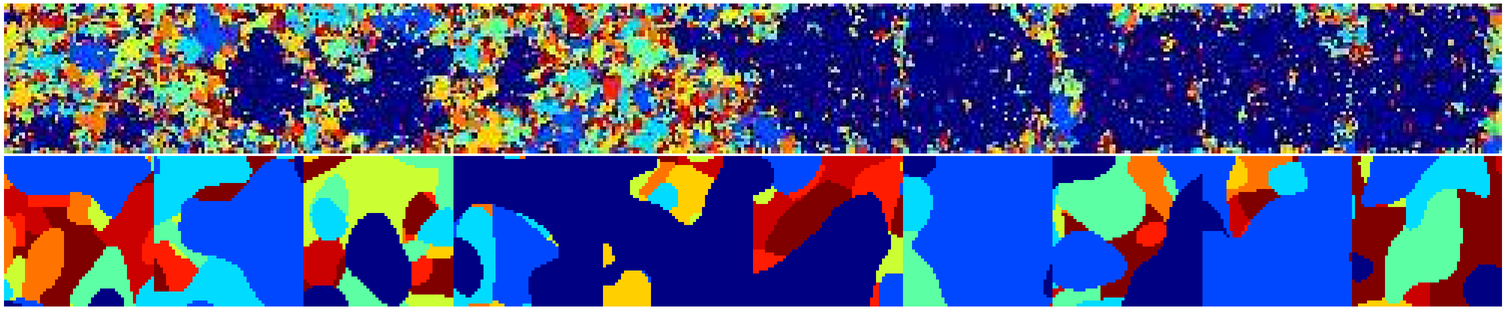
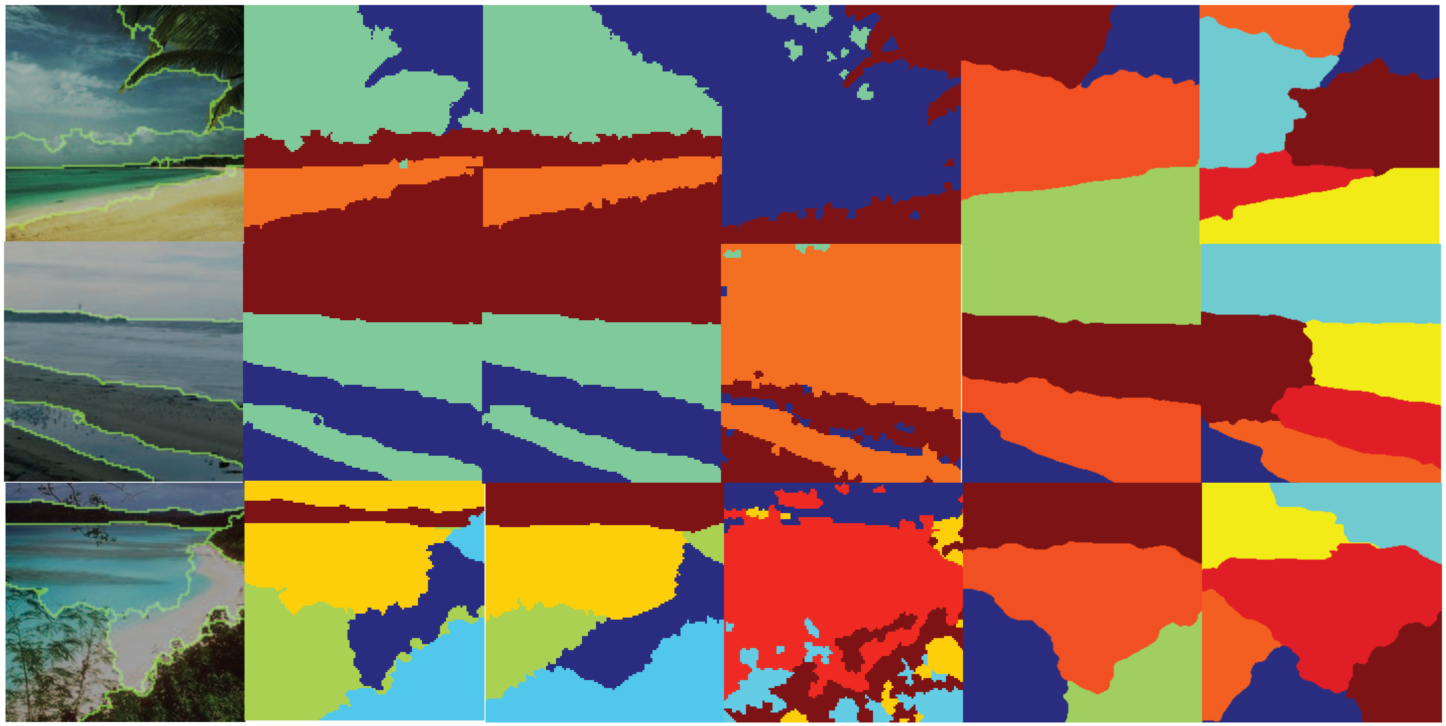


Image Segmentation (cont)

(Sudderth & Jordan, 2008)



Beta Processes

- The Dirichlet process yields a multinomial random variable (which table is the customer sitting at?)
- *Problem*: in many problem domains we have a very large (combinatorial) number of possible tables
  - it becomes difficult to control this with the Dirichlet process
- What if instead we want to characterize objects as collections of attributes ("sparse features")?
- Indeed, instead of working with the sample paths of the Dirichlet process, which sum to one, let's instead consider a stochastic process—the beta process—which removes this constraint
- And then we will go on to consider hierarchical beta processes, which will allow features to be shared among multiple related objects

Lévy Processes

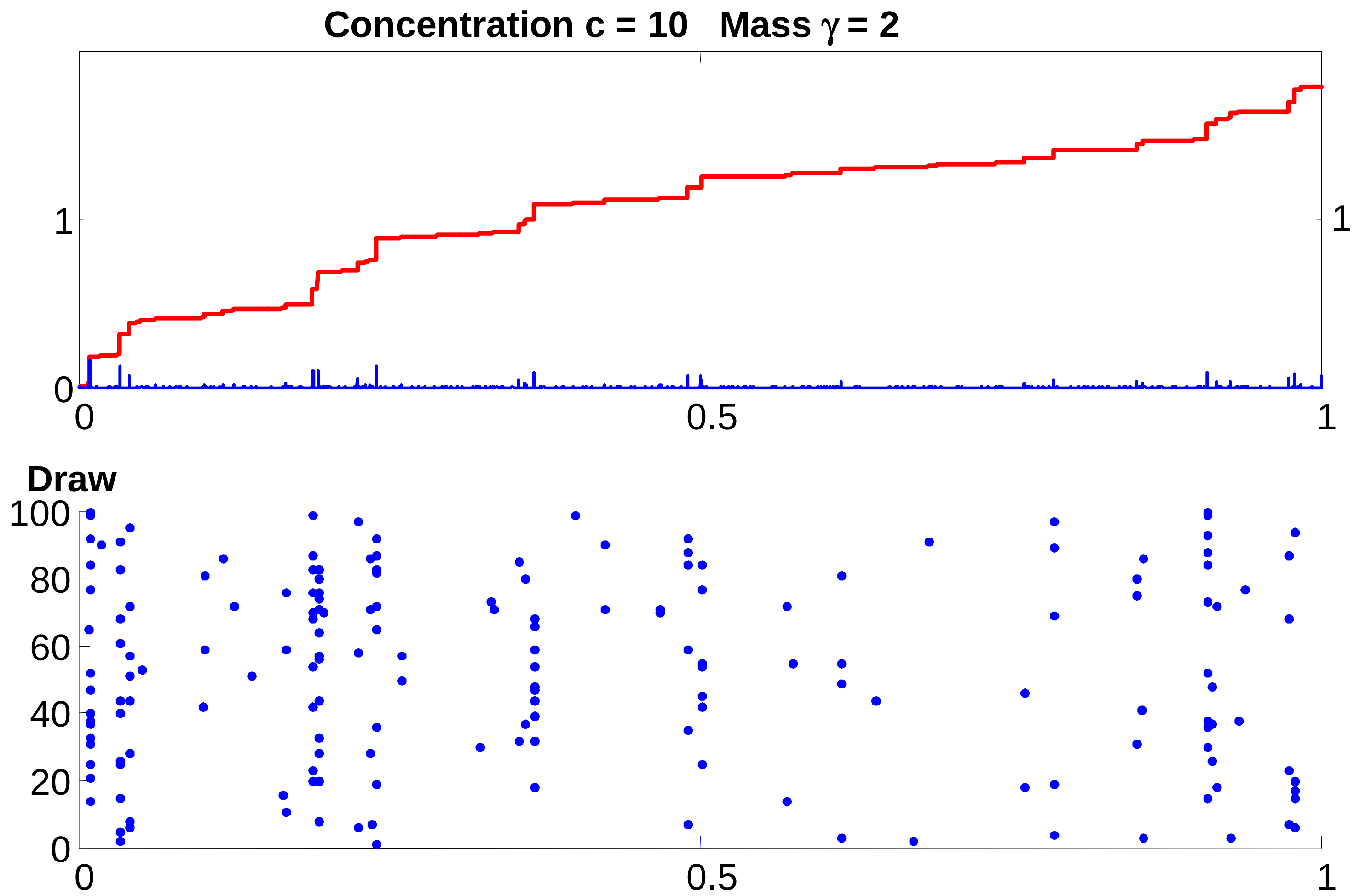
- Stochastic processes with independent increments
  - e.g., Gaussian increments (Brownian motion)
  - e.g., gamma increments (gamma processes)
  - in general, (limits of) compound Poisson processes

- The Dirichlet process is not a Lévy process
  - but it's a normalized gamma process

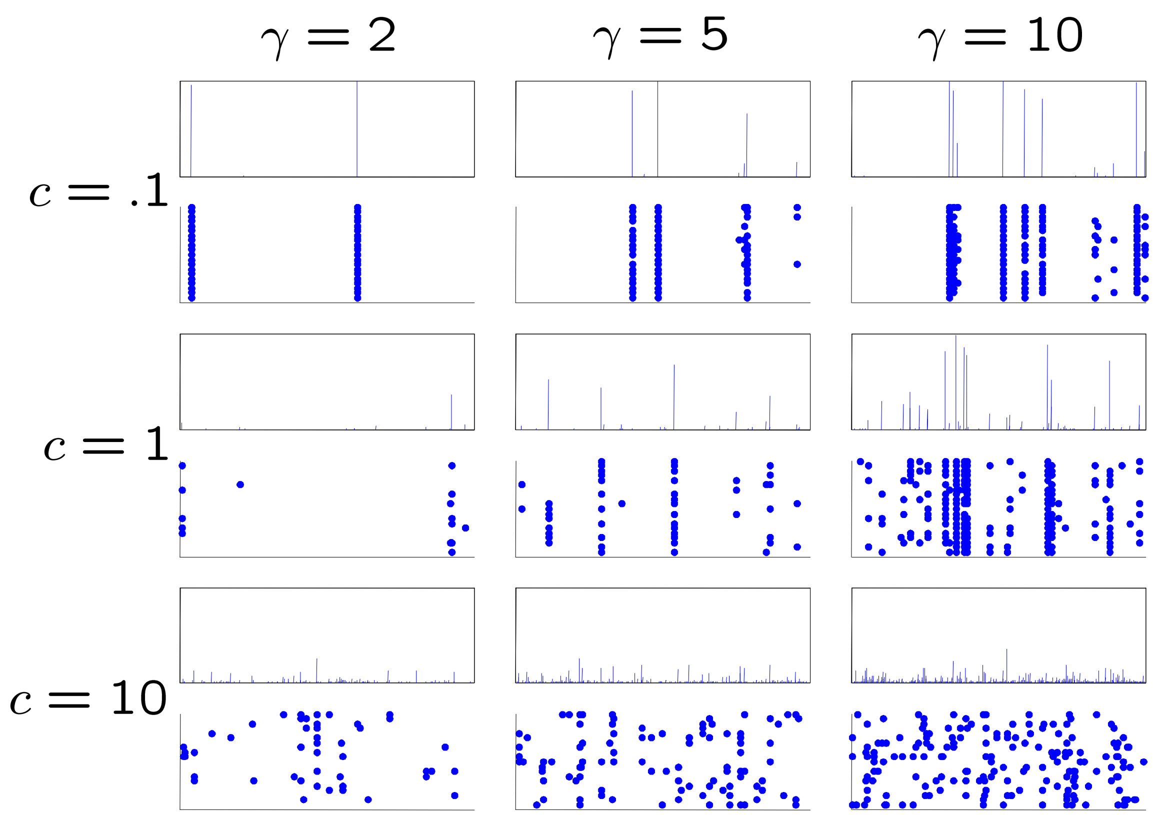
- The beta process assigns beta measure to small regions

- Can then sample to yield (sparse) collections of Bernoulli variables

Beta Processes



Examples of Beta Process Sample Paths



- Effect of the two parameters c and γ on samples from a beta process.

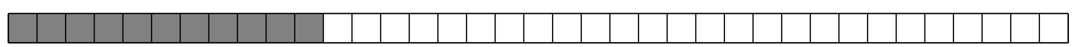
Beta Processes

- The marginals of the Dirichlet process are characterized by the Chinese restaurant process
- What about the beta process?

Indian Buffet Process (IBP)

(Griffiths & Ghahramani, 2005; Thibaux & Jordan, 2007)

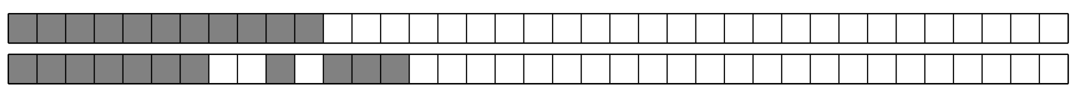
- Indian restaurant with infinitely many dishes in a buffet line
- N customers serve themselves
  - the first customer samples $\mathrm{Poisson}(\alpha)$ dishes
  - the ith customer samples a previously sampled dish with probability $\frac{m_k}{i+1}$ then samples $\mathrm{Poisson}(\frac{\alpha}{i})$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



Indian Buffet Process (IBP)

(Griffiths & Ghahramani, 2005; Thibaux & Jordan, 2007)

- 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

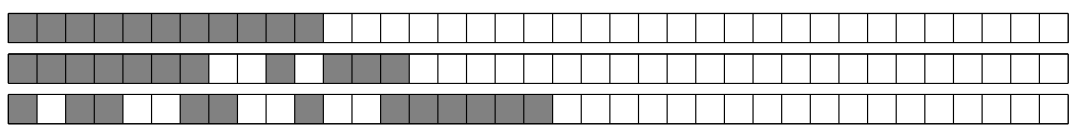


Indian Buffet Process (IBP)

(Griffiths & Ghahramani, 2005; Thibaux & Jordan, 2007)

- Indian restaurant with infinitely many infinite dishes

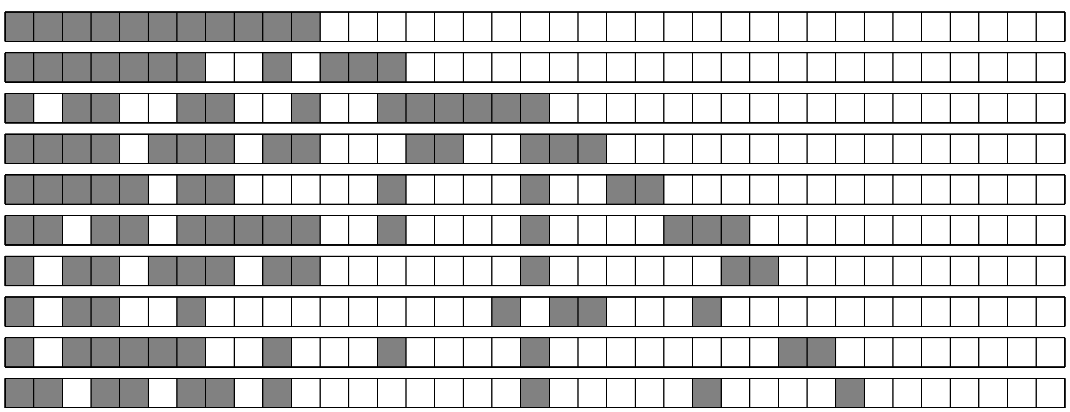
- N customers serve themselves
  - the first customer samples $\mathrm{Poisson}(\alpha)$ dishes
  - the ith customer samples a previously sampled dish with probability $\frac{m_k}{i+1}$ then samples $\mathrm{Poisson}(\frac{\alpha}{i})$ new dishes



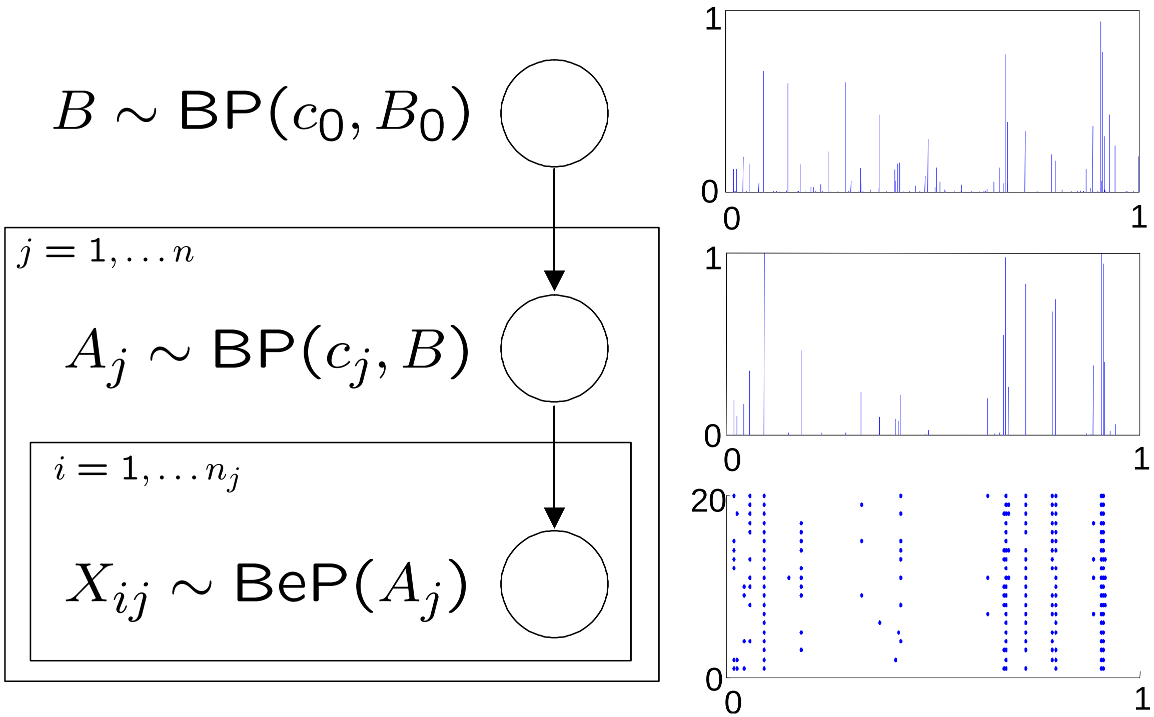
Indian Buffet Process (IBP)

(Griffiths & Ghahramani, 2005; Thibaux & Jordan, 2007)

- 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



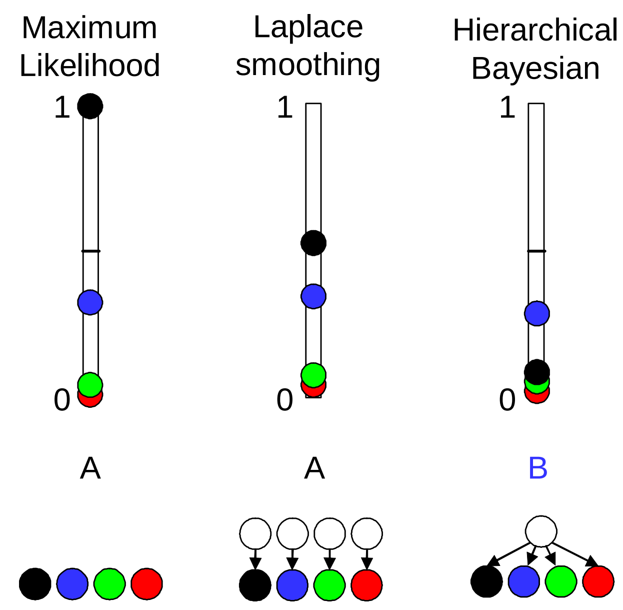
Hierarchical Beta Process



- A hierarchical beta process is a beta process whose base measure is itself random and drawn from a beta process.

Fixing Naive Bayes

	#Documents	#Documents with the word “epilepsy”
Topic A	1	1
Topic B	100	30
Topic C	1000	1
Topic D	1000	0



Topic of which “epilepsy” is most indicative: A A B

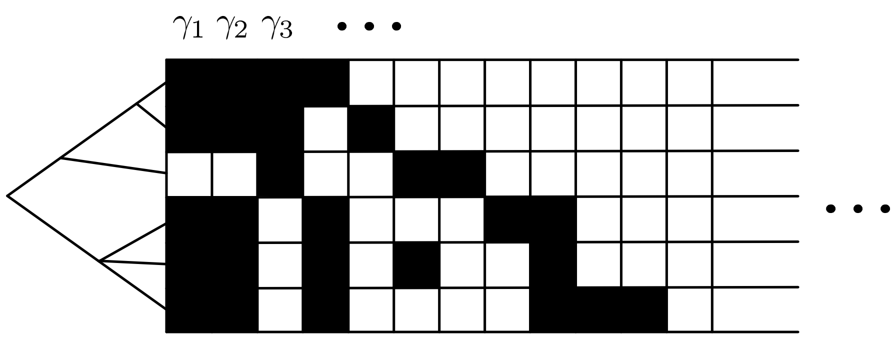
Graphical model:

- A hierarchical Bayesian model correctly takes the weight of the evidence into account and matches our intuition regarding which topic should be favored when observing this word.
- This can be done nonparametrically with the hierarchical beta process.

The Phylogenetic IBP

(Miller, Griffiths & Jordan, 2008)

- We don't always want objects to be exchangeable; sometimes we have side information to distinguish objects
  - but if we lose exchangeability, we risk losing computational tractability
- In the phylo-IBP we use a tree to represent various forms of partial exchangeability



- The process stays tractable (belief propagation to the rescue!)

Conclusions

- The underlying principle in this talk: exchangeability
- Leads to nonparametric Bayesian models that can be fit with computationally efficient algorithms
- Leads to architectural and algorithmic building blocks that can be adapted to many problems
- For more details (including tutorial slides):

http://www.cs.berkeley.edu/~jordan