Hierarchical Nonparametric Bayes

Michael I. Jordan

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

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

Acknowledgments: Yee Whye Teh, Romain Thibaux

Objective Bayes and Nonparametric Bayes

- The agenda for objective Bayes: *let the data speak*

Objective Bayes and Nonparametric Bayes

- The agenda for objective Bayes: *let the data speak*
- The agenda for nonparametric Bayes:

Objective Bayes and Nonparametric Bayes

- The agenda for objective Bayes: *let the data speak*
- The agenda for nonparametric Bayes: *let the data speak*

Objective Bayes and Nonparametric Bayes

- The agenda for objective Bayes: *let the data speak*
- The agenda for nonparametric Bayes: *let the data speak*
- Hmm, surely there must be relationships, but thus far the research efforts seem mainly detached
- I certainly *feel* that I'm being more "objective" when I work with a nonparametric prior than when I work with less flexible models
- In my view, a deeper understanding depends in part on understanding how these ideas interact with hierarchical modeling

Hierarchical Bayes

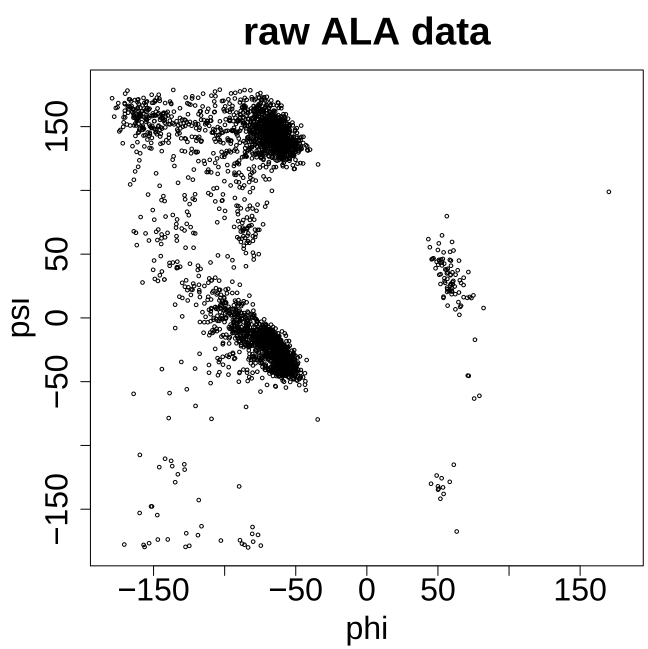
- The naturalness of hierarchies in the Bayesian formalism is the main reason I'm a Bayesian
  - provide both complexity and control
- Seemingly of particular relevance to nonparametric Bayesian work, where the emphasis is complexity and the need for control is great
- Of great help in the development of subjective priors; what can objective Bayes say about hierarchical priors?

Hierarchical Nonparametric Bayes

- Many nonparametric (or semiparametric) Bayesian models make use of classical parametric hierarchies
  - e.g., when using the Dirichlet process $\mathrm{DP}(\alpha_0, G_0)$, it's common to let G_0 lie in a parametric family, say $G_0 = N(\mu_0, \tau_0)$
- But in the spirit of nonparametric methods let's try to make fuller use of stochastic processes
  - e.g., in the Dirichlet process let G_0 be a random measure
- Why? *Because this construction allows us to solve a raft of practical problems that involve multiple, coupled clustering problems*

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 clustering 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

Document and Image Modeling

- Define a topic to be a probability distribution across *words* in some vocabulary

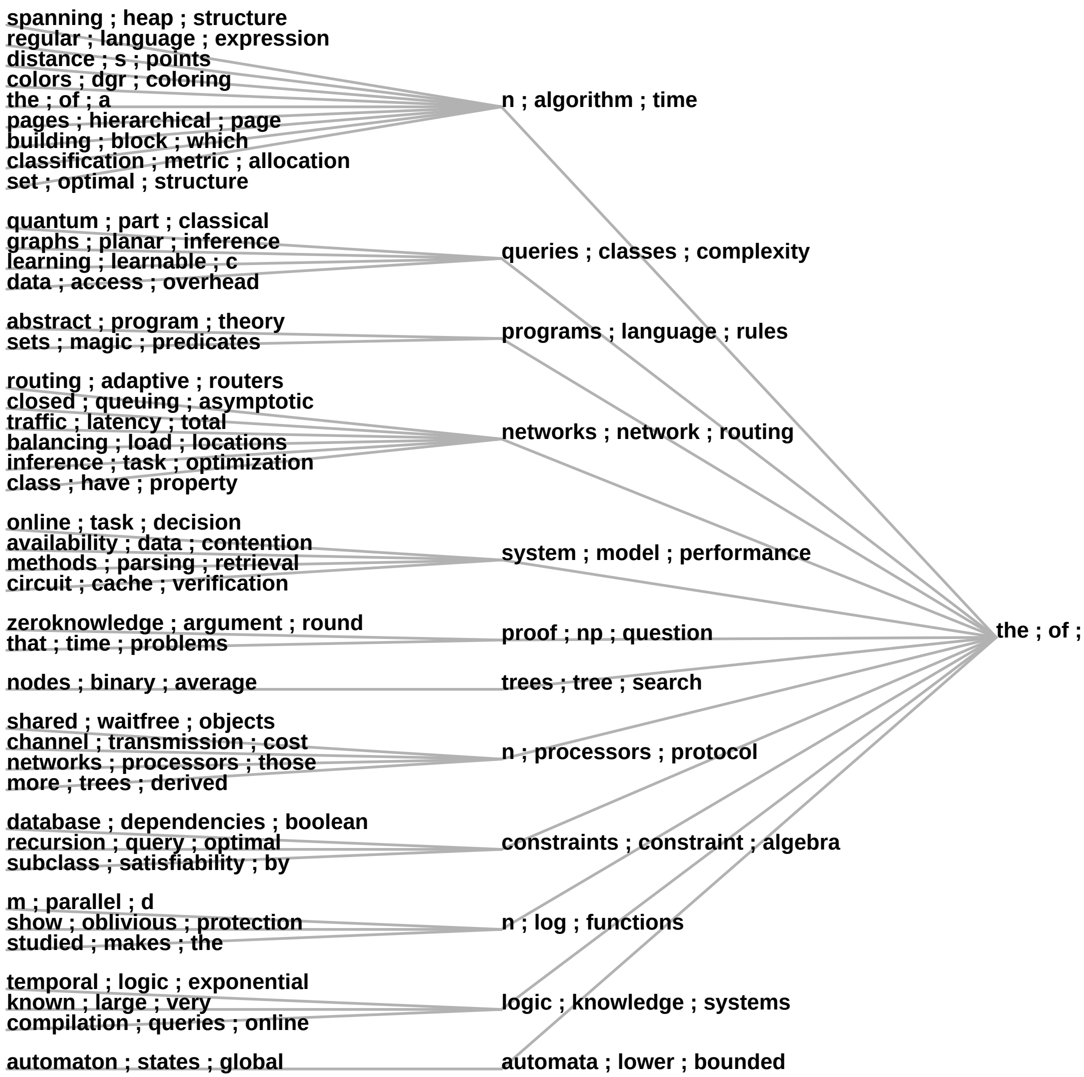
- Define a document to be a probability distribution across topics

- Given a corpus of documents, find the topics and find the patterns of usage of topics across documents

- Each document is a clustering problem; we must link multiple clusterings across a corpus

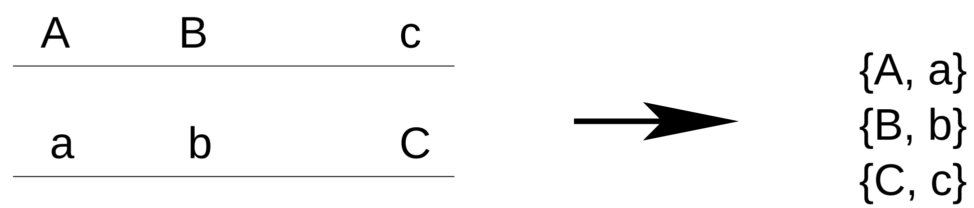
- Note that a "document" can be an image, where a "word" is a local image feature

Topic Hierarchies



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)



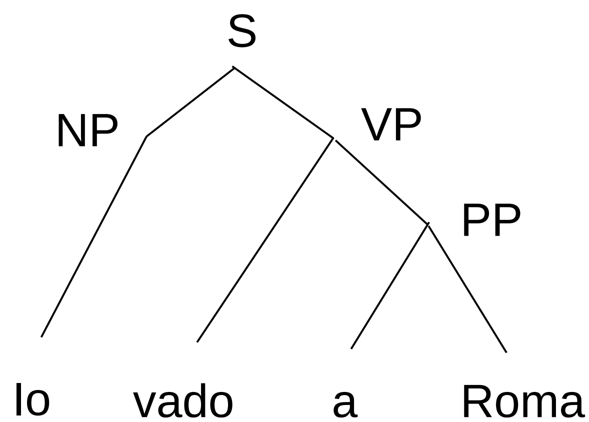
- Given a set of genotypes (multiple individuals), estimate the underlying haplotypes
- This is a clustering problem

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 all statistical

Natural Language Parsing (cont.)

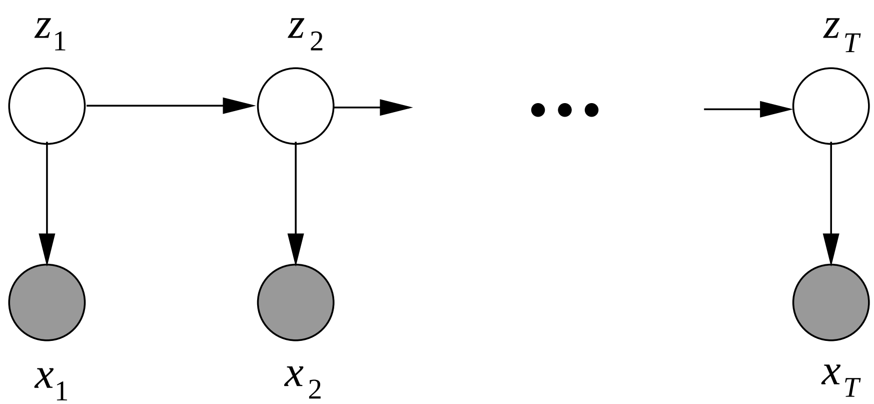
- Key idea: *lexicalization* of context-free grammars
  - the grammatical rules (S $\rightarrow$ NP VP) are conditioned on the specific lexical items (words) that they derive

- 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

Outline

- Dirichlet Processes (clusters)
- Hierarchical Dirichlet Processes (tied clusters)
- Beta Processes (features)
- Hierarchical Beta Processes (tied features)

Clustering—How to Choose K?

Clustering—How to Choose K?

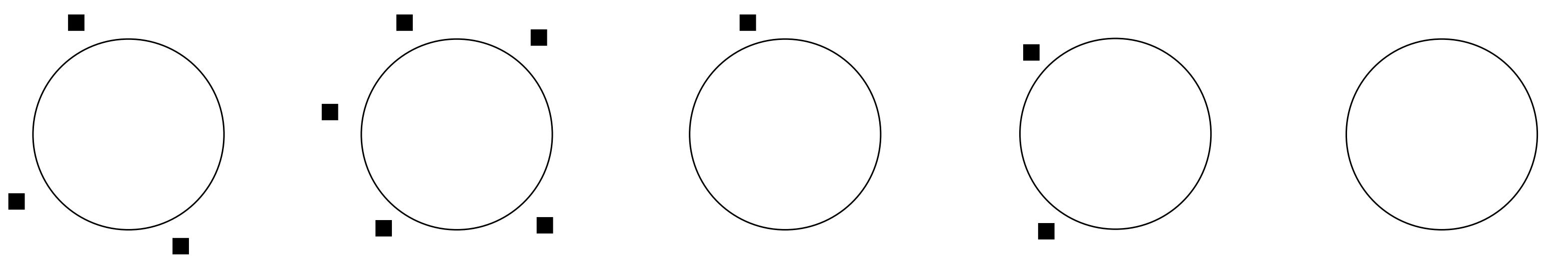
- Adhoc approaches (e.g., hierarchical clustering)
  - they do often yield a data-driven choice of K
  - but there is little understanding of how good these choices are
- Methods based on objective functions (M-estimators)
  - e.g., K-means, spectral clustering
  - do come with some frequentist guarantees
  - but it's hard to turn these into data-driven choices of K
- Parametric likelihood-based approaches
  - finite mixture models, Bayesian variants thereof
  - various model choice methods: hypothesis testing, cross-validation, bootstrap, AIC, BIC, DIC, Laplace, bridge sampling, reversible jump, etc
  - but do the assumptions underlying the method really apply to this setting? (not often)
- Let's try something different...

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:

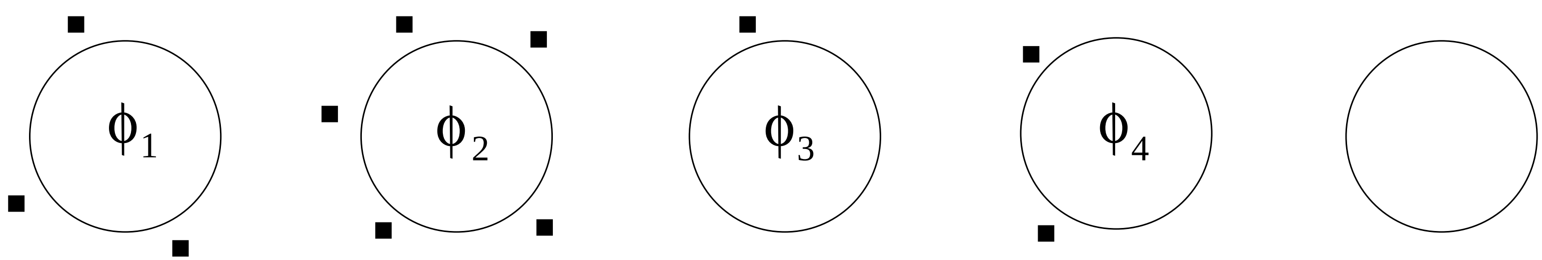
$$\begin{aligned} P(\text{previously occupied table } i \,|\, \mathcal{F}_{m-1}) &\propto n_i \\ P(\text{the next unoccupied table} \,|\, \mathcal{F}_{m-1}) &\propto \alpha_0 \end{aligned} \quad (1)$$

  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



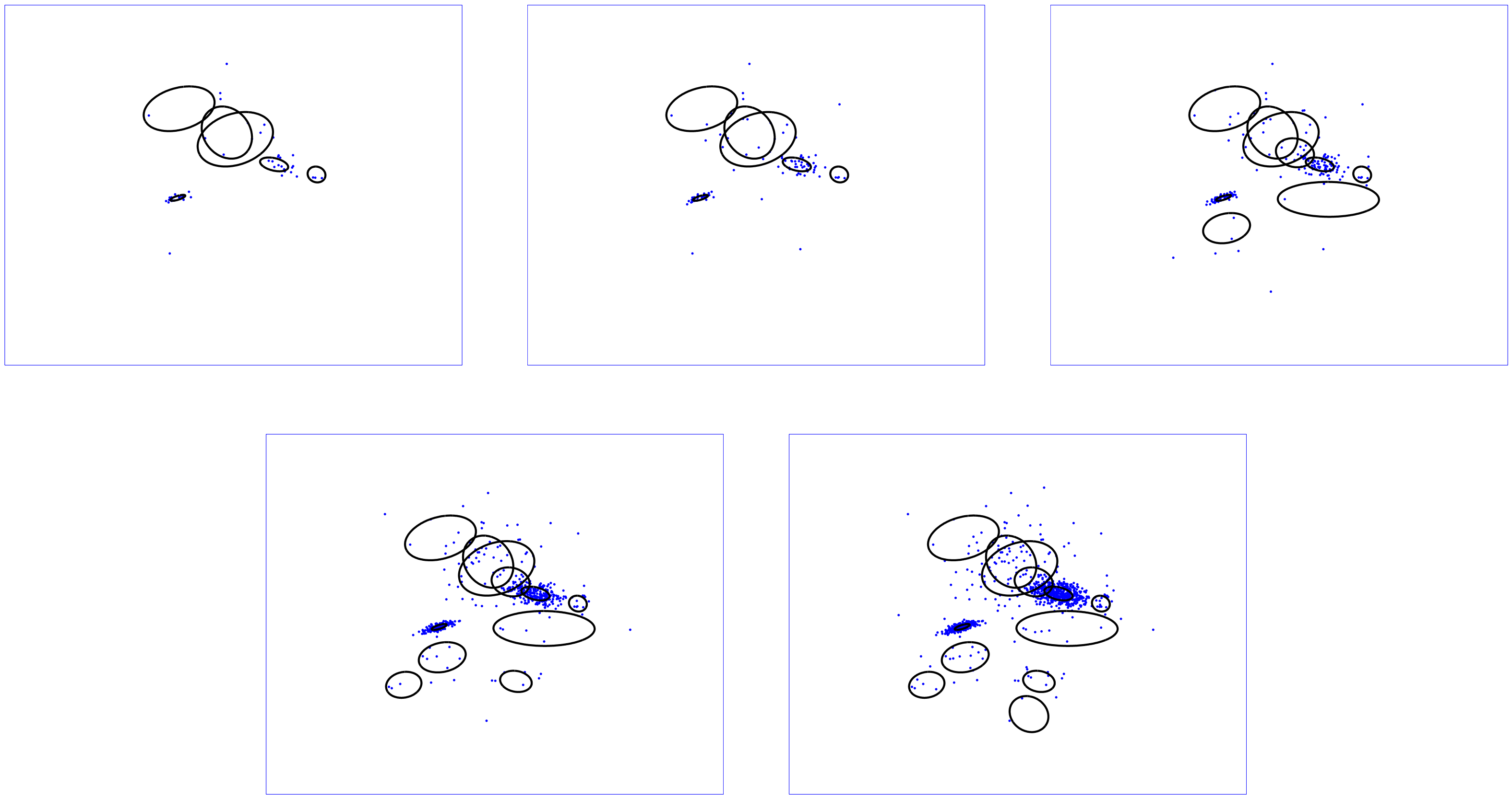
The CRP and Clustering

- 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



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

Inference for the CRP

- We've described how to generate data from the model; how do we go backwards and generate a model from data?

- A wide variety of variational, combinatorial and MCMC algorithms have been developed

- E.g., a Gibbs sampler is readily developed by using the (deep) fact that the Chinese restaurant process is exchangeable
  - to sample the table assignment for a given customer given the seating of all other customers, simply treat that customer as the last customer to arrive
  - in which case, the assignment is made proportional to the number of customers already at each table (cf. preferential attachment)
  - parameters are sampled at each table based on the customers at that table (cf. K means)

- (This isn't the state of the art, but it's easy to explain on one slide)

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

The De Finetti Theorem

- *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*:

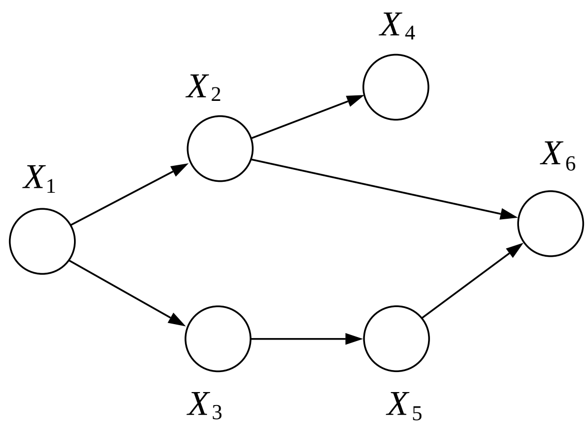
  $$p(x_1, x_2, \ldots, x_N) = \int \left(\prod_{i=1}^{N} p(x_i \,|\, G) \right) dP(G)$$

  for some random element G.

- The exchangeability of the CRP implies that there is an underlying "parameter" G and a distribution on that parameter. What are they?

Directed Graphical Models

- Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each node $v \in \mathcal{V}$ is associated with a random variable X_v:

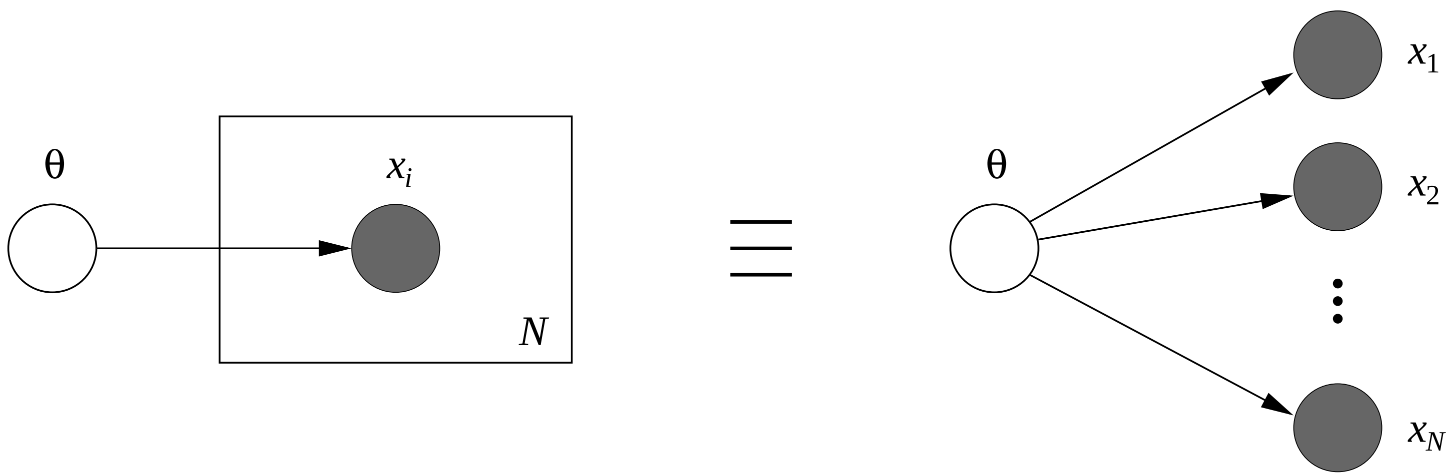


- The joint distribution on $(X_1, X_2, \ldots, X_N)$ factorizes according to the "parent-of" relation defined by the edges $\mathcal{E}$:

$$p(x_1, x_2, x_3, x_4, x_5, x_6; \theta) = p(x_1; \theta_1)\; p(x_2 \,|\, x_1; \theta_2)$$
$$p(x_3 \,|\, x_1; \theta_3)\; p(x_4 \,|\, x_2; \theta_4)\; p(x_5 \,|\, x_3; \theta_5)\; p(x_6 \,|\, x_2, x_5; \theta_6)$$

Plates

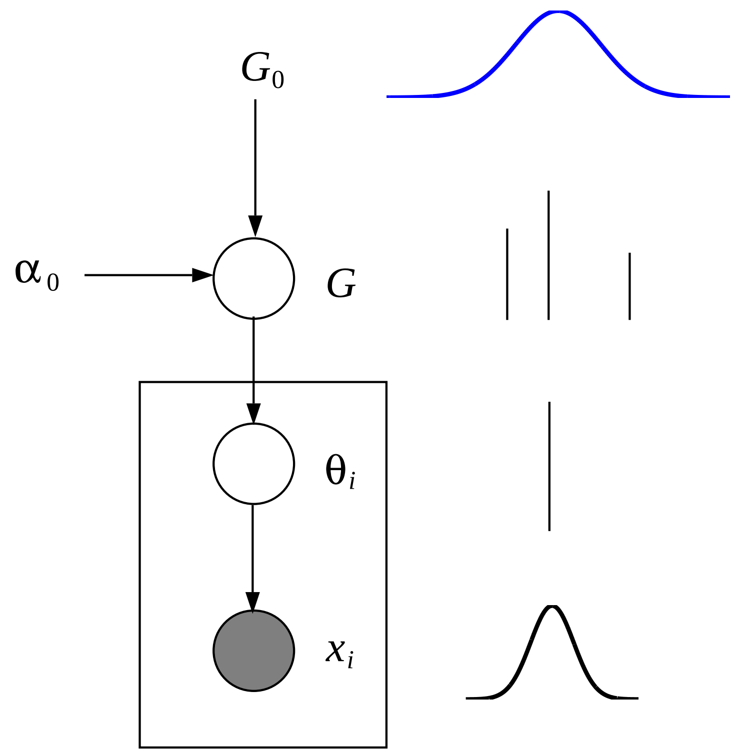
- A *plate* is a “macro” that allows subgraphs to be replicated:



- Shading denotes conditioning

Finite Mixture Models

$$
\begin{aligned}
\phi_k &\sim G_0 \\
\pi_k &\sim \mathrm{Dir}(\alpha_0/K, \ldots, \alpha_0/K) \\
G &= \sum_{k=1}^{K} \pi_k \, \delta_{\phi_k} \\
\theta_i &\sim G \\
x_i &\sim p(\cdot \,|\, \theta_i)
\end{aligned}
$$



- Note that G is a *random measure*

Going Nonparametric—A First Attempt

- Define a countably infinite mixture model by taking K to infinity and hoping that "$G = \sum_{k=1}^{\infty} \pi_k \, \delta_{\phi_k}$" means something, where

$$
\begin{aligned}
\phi_k &\sim G_0 \\
\pi_k &\sim \mathrm{Dir}(\alpha_0/K, \ldots, \alpha_0/K) \text{ as } K \to \infty
\end{aligned}
$$

- Several mathematical hurdles to overcome:
  - What is the distribution of any given π_k as $K \to \infty$? Does it stabilize at some fixed distribution?
  - Is $\sum_{k=1}^{\infty} \pi_k = 1$ under some suitable notion of convergence?
  - Do we get a few large mixing proportions, or are they all of similar "size"?
  - Do we get any "clustering" at all?

- This seems hard; let's approach the problem from a different point of view

Stick-Breaking

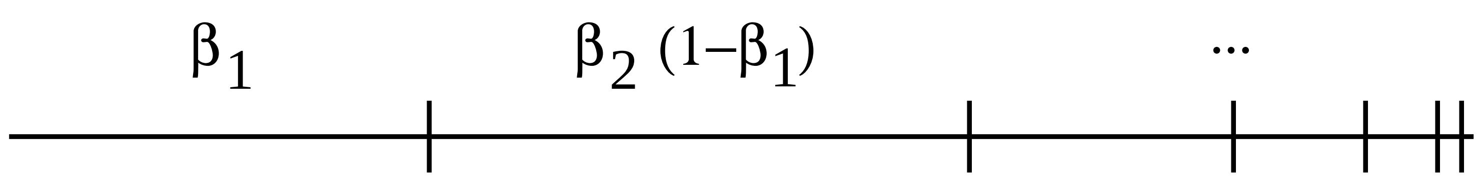
- Define an infinite sequence of Beta random variables:

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

- And then define an infinite sequence of mixing proportions as:

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

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



Stick-Breaking (cont)

- We now have an explicit formula for each π_k:

$$\pi_k = \beta_k \prod_{l=1}^{k-1} (1 - \beta_l)$$

- And now $G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$ has a clean definition as a random measure

- The distribution of G is known as a Dirichlet process

  – it can be shown that for any finite partition $(A_1, \ldots, A_r)$ of the sample space, the random vector $(G(A_1), \ldots, G(A_r))$ is distributed as a finite-dimensional Dirichlet distribution

- We write this as

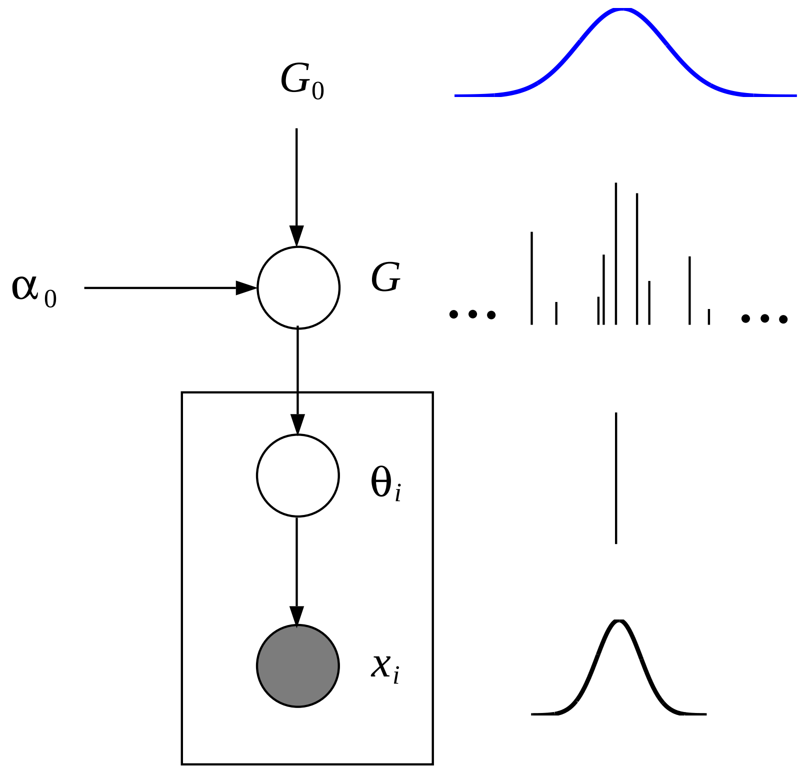
$$G \sim \mathrm{DP}(\alpha_0, G_0),$$

  where α_0 is known as the concentration parameter and G_0 is known as the base measure

Stick-Breaking (cont)

- An advantage of the stick-breaking perspective is that it permits numerous generalizations
  - e.g., using $\mathrm{Beta}(\alpha_1, \alpha_2)$ instead of $\mathrm{Beta}(1, \alpha_0)$ yields the heavier-tailed Pitman-Yor process
- Another advantage of the stick-breaking perspective is that it readily yields Bayesian hierarchies
  - as we'll see later

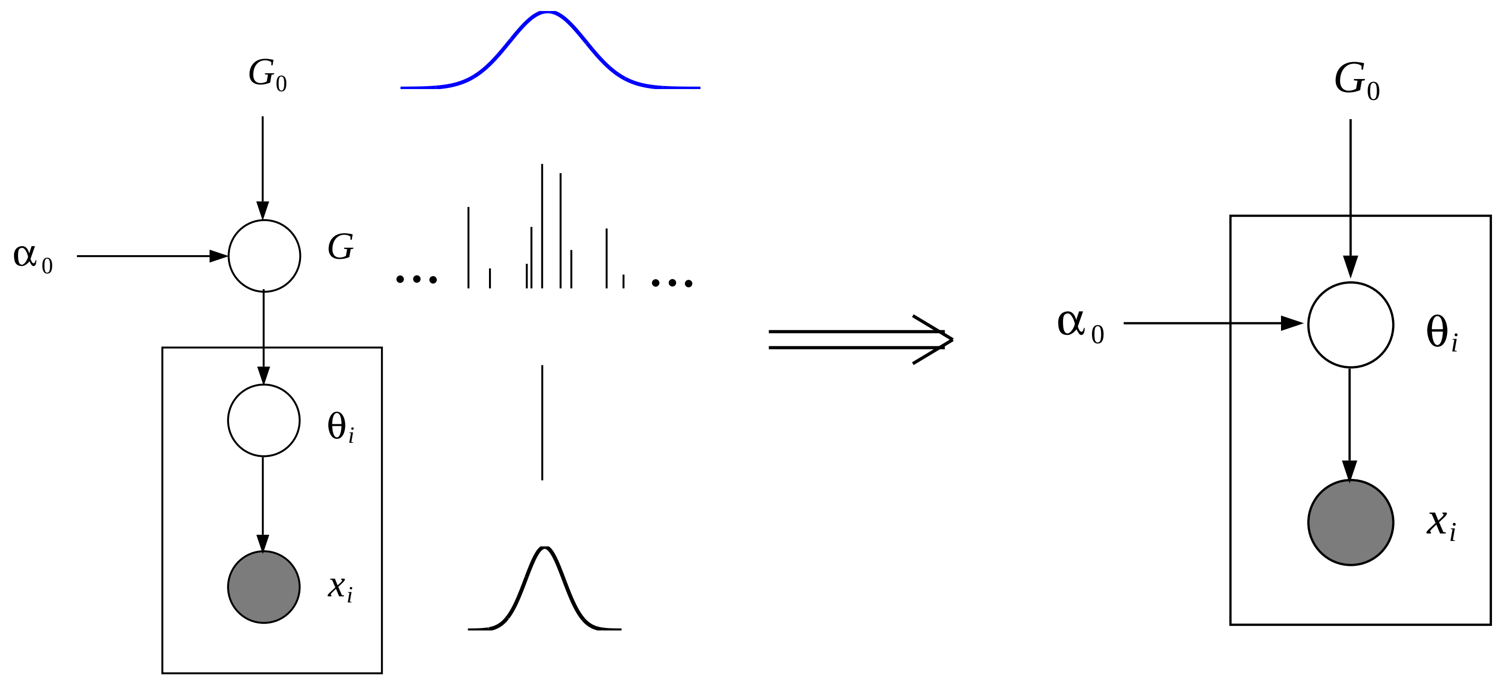
Dirichlet Process Mixture Models



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

Marginal Probabilities

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



Marginal Probabilities (cont)

- Dirichlet expectations:

$$\mathrm{E}[G(A) \,|\, \theta_1, \ldots, \theta_n] = \frac{\alpha_0 G_0(A) + \sum_{k=1}^{K} n_k \delta_{\phi_k}(A)}{\alpha_0 + n}$$

- This is just the Chinese restaurant process

- I.e., integrating over the random measure G, where $G \sim \mathrm{DP}(\alpha_0 G_0)$, yields the Chinese restaurant process

Summary Thus Far

- The Chinese restaurant process provides an elegant solution to the problem of "how many clusters?"
- The Chinese restaurant process yields an exchangeable distribution on data points
- De Finetti tells us that there must exist an underlying random measure
- That random measure is the Dirichlet process
- The Dirichlet process can be obtained explicitly via stick-breaking

Inference for Dirichlet Process Mixtures

- MCMC
  - based on the Chinese restaurant process or urn model
  - based on the stick-breaking representation
  - split-merge algorithms

- Variational inference
  - based on the stick-breaking representation

Truncated Dirichlet Processes

(e.g., Gelfand & Kottas; Ishawaran & James; Muliere & Tardella)

- Truncate the stick-breaking representation by fixing a value T and letting $\beta_T = 1$

- This implies $\pi_k = 0$ for $k > T$, and the distribution of

$$G_T = \sum_{k=1}^{T} \pi_k \delta_{\phi_k}$$

  is known as a *truncated Dirichlet process*

- Variational distance between distributions of marginals from a DP and from its truncation $\sim 4n \exp(-(T-1)/\alpha_0)$

  – T doesn't have to be very large to get a good approximation

Variational Inference

- The setup for (mean-field) variational inference:

- Given an intractable density P, consider a tractable family Q_μ, for *variational parameters* μ

- Define an optimization problem:

$$\mu^* = \arg\min D(Q_\mu \parallel P)$$

- Use Q_{μ^*} to approximate the desired marginals of P

- Almost all applications of this approach have been for parametric models (i.e., exponential family models)

Variational Inference for DP Mixtures

(Blei & Jordan, 2005)

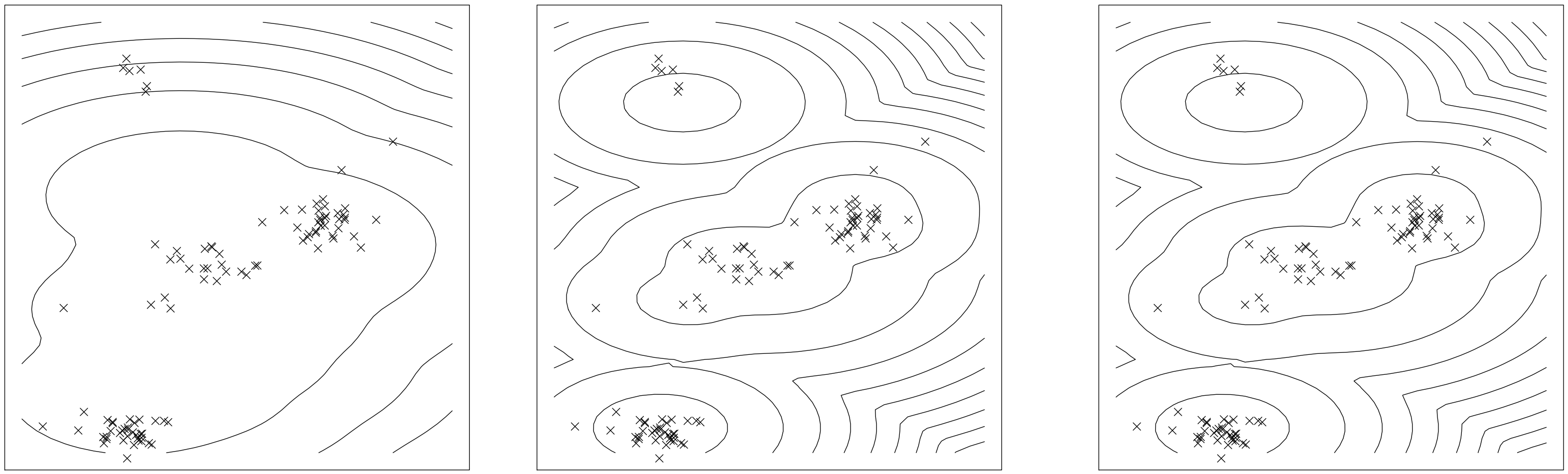
- The Q distribution is a truncated stick-breaking representation (note that P is *not* truncated)
- Variational inference equations for a conjugate DP mixture in the exponential family:

$$
\begin{aligned}
\gamma_{i,t} &= 1 + \textstyle\sum_n \phi_{n,t} \\
\gamma_{i,t} &= \alpha + \textstyle\sum_n \sum_{j=t+1}^{T} \phi_{n,j} \\
\tau_{t,1} &= \lambda_1 + \textstyle\sum_n \phi_{n,t} x_n \\
\tau_{t,2} &= \lambda_2 + \textstyle\sum_n \phi_{n,t} \\
\phi_{n,t} &\propto \exp(S),
\end{aligned}
$$

where (γ, τ, ϕ) are variational parameters and where:

$$S = E[\log V_t] + \textstyle\sum_{i=1}^{t-1} E[\log(1 - V_i)] + E[\eta_t^*]^T X_n - E[a(\eta_t^*)]$$

Example: DP-Gaussian Mixture



Initial state | 1st iteration | 5th (and last) iteration

Figure 1: The approximate predictive distribution given by variational inference at different stages of the algorithm. The data are 100 points generated by a Gaussian DP mixture model with fixed diagonal covariance.

Example: DP-Gaussian Mixture

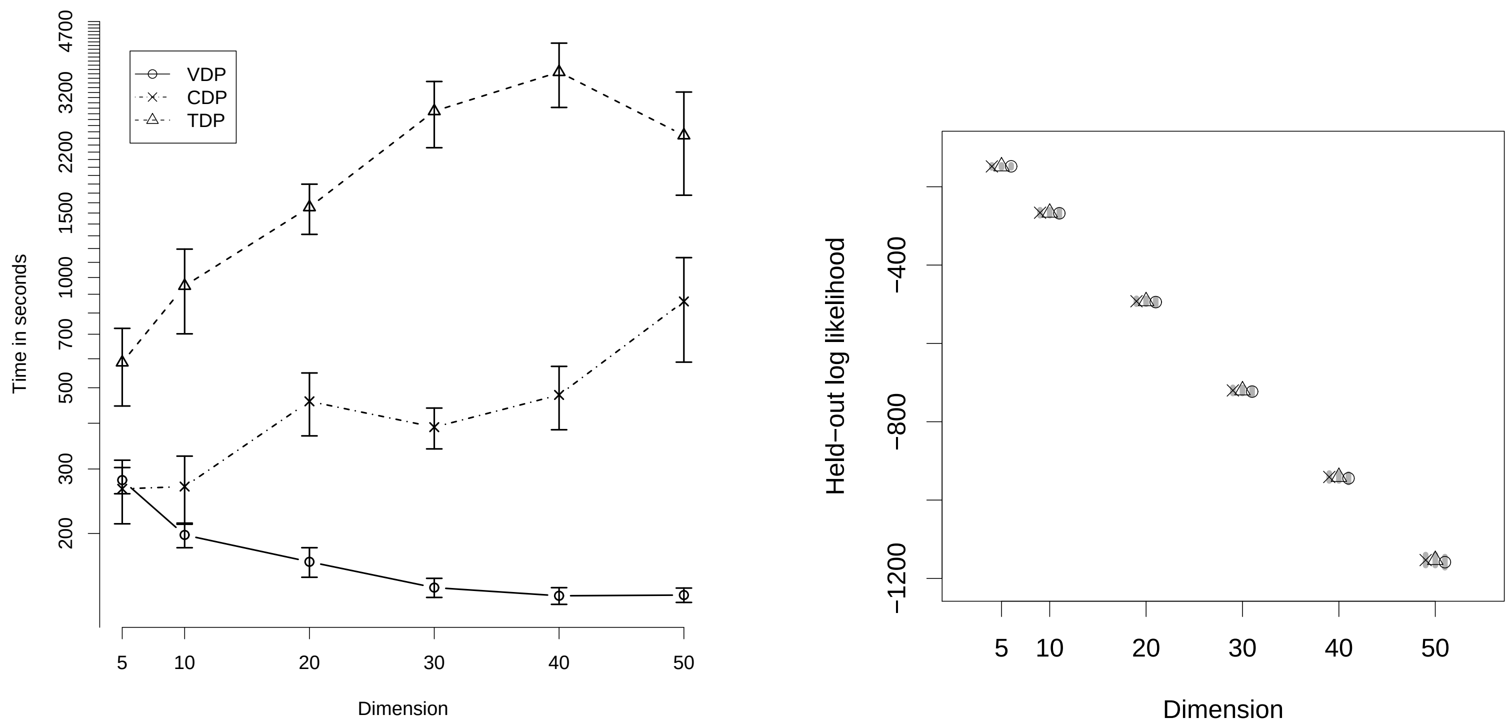


Figure 2: (Left) Convergence time per dimension across ten datasets for variational inference (VDP), the TDP Gibbs sampler (TDP), and the collapsed Gibbs sampler (CDP). Grey bars are standard error. (Right) Average held-out log likelihood for the corresponding predictive distributions.

DP-Based Haplotype Model

(Xing, Sharan, & Jordan, 2004)

- Recall the setup: for each individual we have a genotype (underordered set of genetic markers), and we want to recover the underlying chromosomes

- In the Chinese restaurant representation, each table is associated with the chromosome of a putative ancestral human

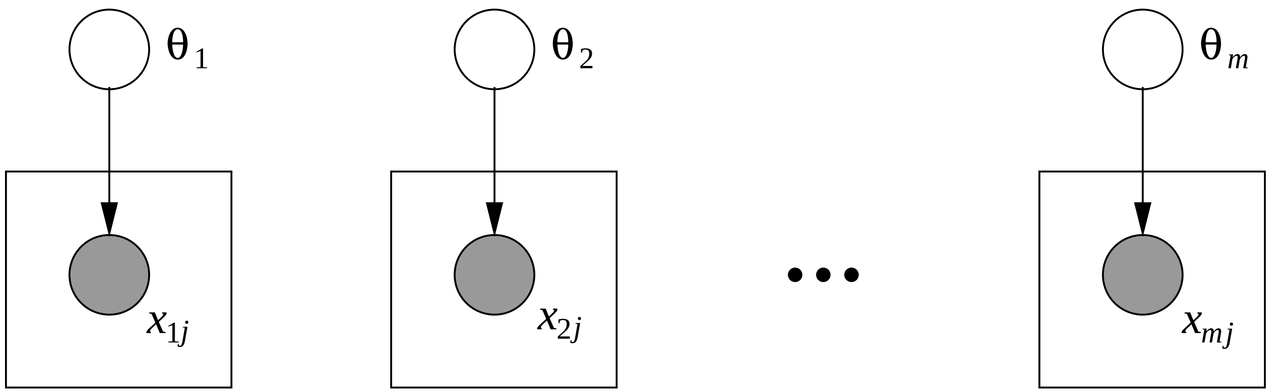
- Intuitively, we want individuals to sit at the table of their ancestor

- Comparative performance of model on the data of Gabriel, et al (2002):

region	length	DP	PHASE
16a	13	0.141	0.130
1b	16	0.160	0.180
25a	14	0.115	0.212
7b	13	0.066	0.092

Multiple Estimation Problems

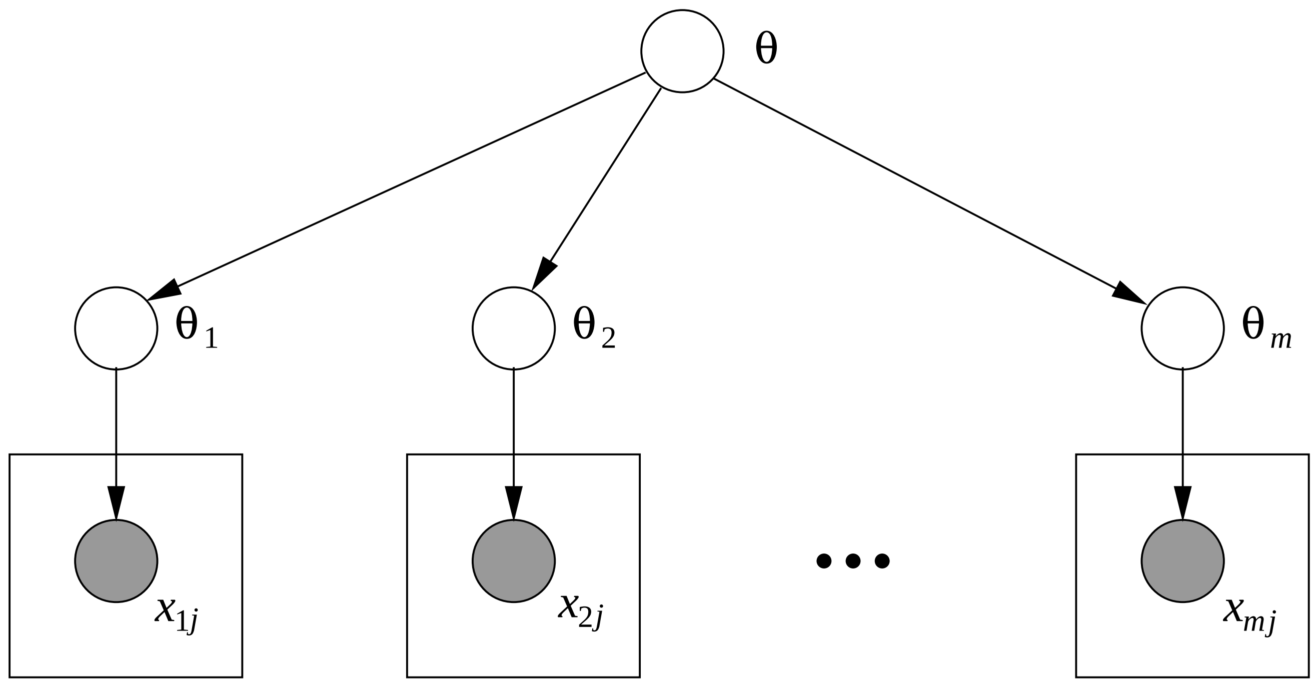
- We often face multiple, related estimation problems
- E.g., multiple Gaussian means: $x_{ij} \sim N(\theta_i, \sigma_i^2)$



- Maximum likelihood: $\hat{\theta}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$
- Maximum likelihood often doesn't work very well
  - want to "share statistical strength" (i.e., "smooth")

Hierarchical Bayesian Approach

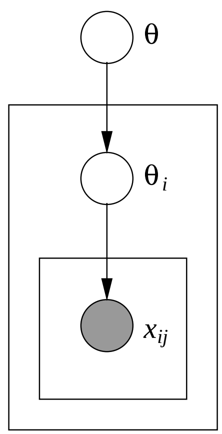
- The Bayesian or empirical Bayesian solution is to view the parameters θ_i as random variables, sampled from an underlying variable θ



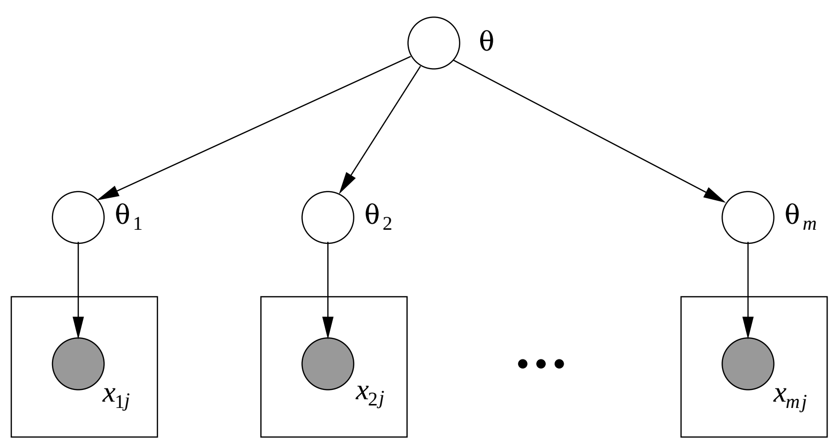
- Given this overall model, posterior inference yields *shrinkage*—the posterior mean for each θ_k combines data from all of the groups

Hierarchical Modeling

- Recall the plate notation:

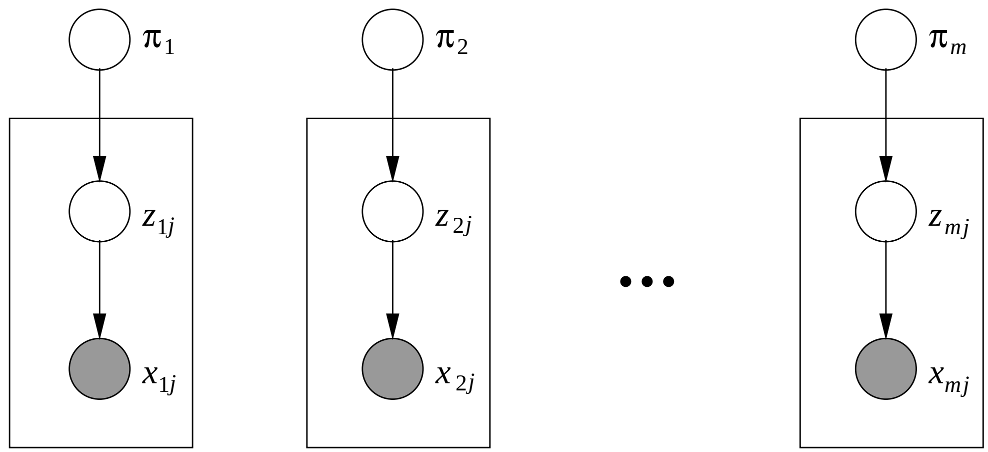


- Equivalent to:



Multiple Clustering Problems

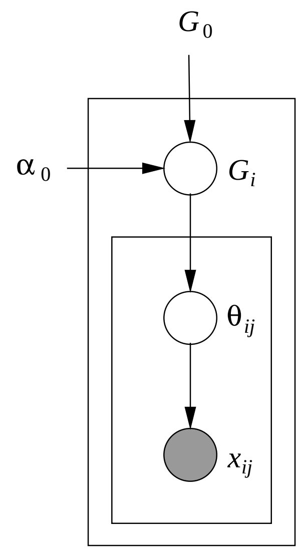
- What about the case in which we have multiple related clustering problems?
  - what to share? how to share?
- Mixture models: $p(x_{ij} \,|\, \pi_i, \theta_i) = \sum_{l=1}^{K_i} p(Z_{ij}^l = 1 \,|\, \pi_i)\; p(x_{ij} \,|\, Z_{ij}^l = 1, \theta_i)$



- What to share: π_i?, θ_i? What if we don't know the K_i?
- Model selection ideas seem unhelpful; let's consider a nonparametric Bayesian approach

A Nonparametric Approach—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:

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

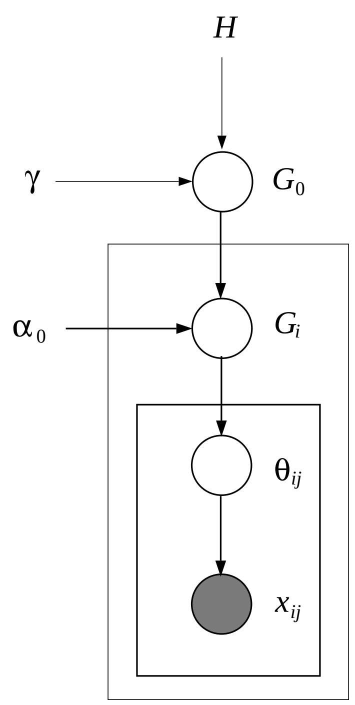
- Then

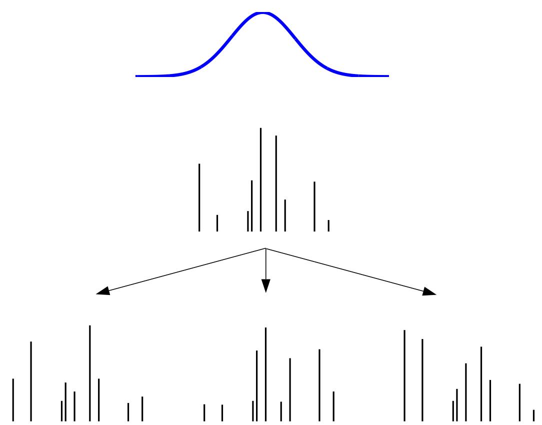
$$G_j \,|\, \alpha, G_0 \;\sim\; \mathrm{DP}(\alpha_0 G_0)$$

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

- I.e., just go to another level of the Bayesian hierarchy

Hierarchical Dirichlet Process Mixtures

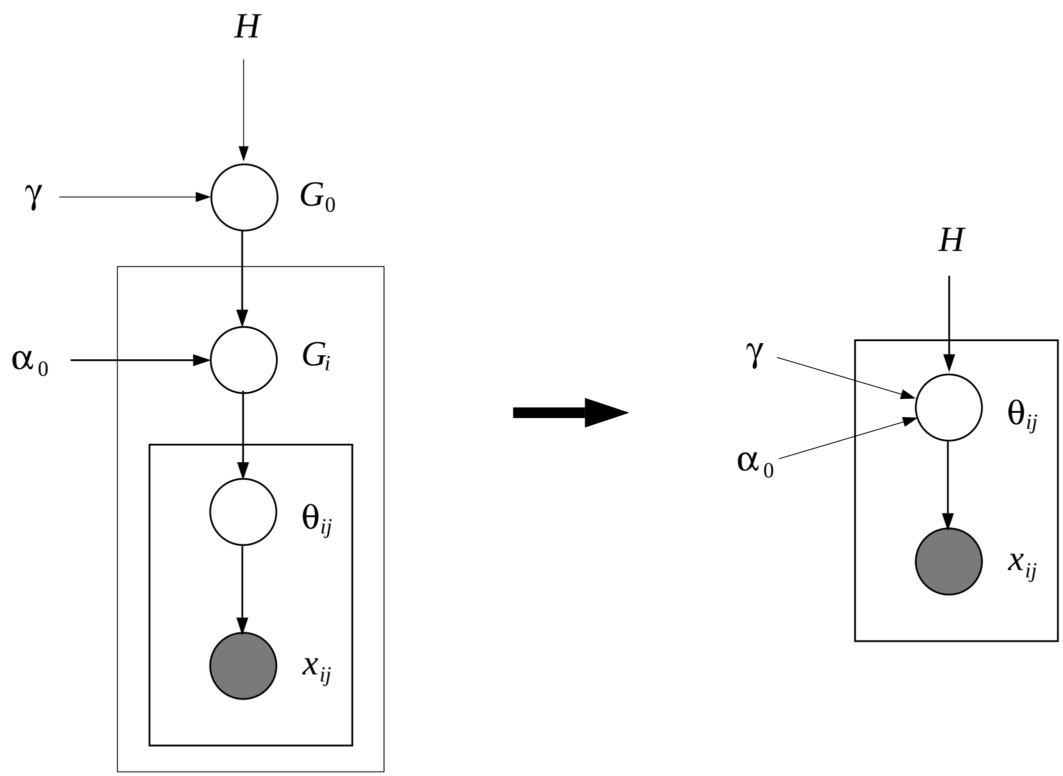




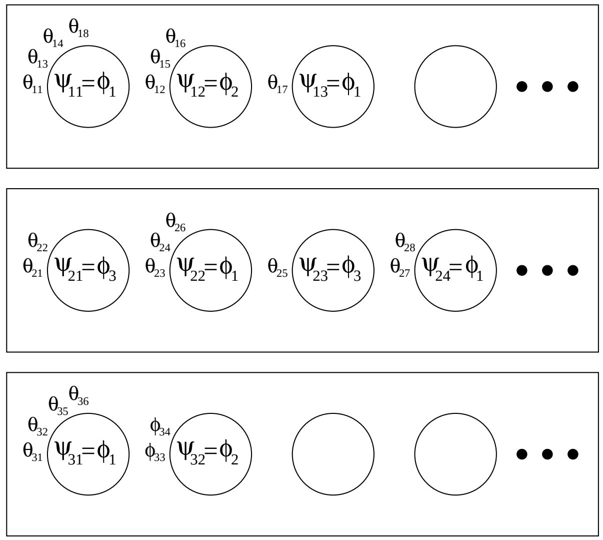
$$
\begin{aligned}
G_0 \mid \gamma, H &\sim \mathrm{DP}(\gamma H) \\
G_i \mid \alpha, G_0 &\sim \mathrm{DP}(\alpha_0 G_0) \\
\theta_{ij} \mid G_i &\sim G_i \\
x_{ij} \mid \theta_{ij} &\sim F(x_{ij}, \theta_{ij})
\end{aligned}
$$

Chinese Restaurant Franchise (CRF)

- First integrate out the G_i, then integrate out G_0



Chinese Restaurant Franchise (CRF)

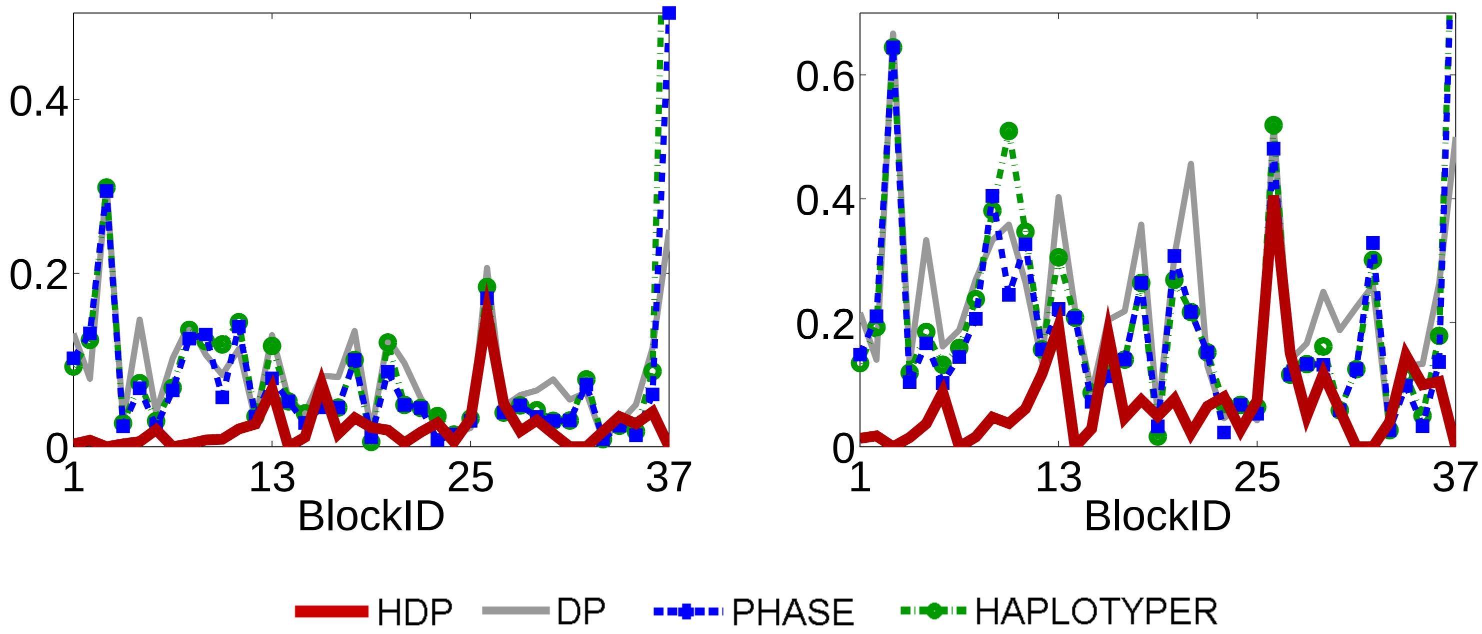


- To each group there corresponds a *restaurant*, with an unbounded number of *tables* in each restaurant
- There is a global *menu* with an unbounded number of *dishes* on the menu
- The first customer at a table selects a dish for that table from the global menu
- Reinforcement effects—customers prefer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

Haplotype Modeling (cont.)

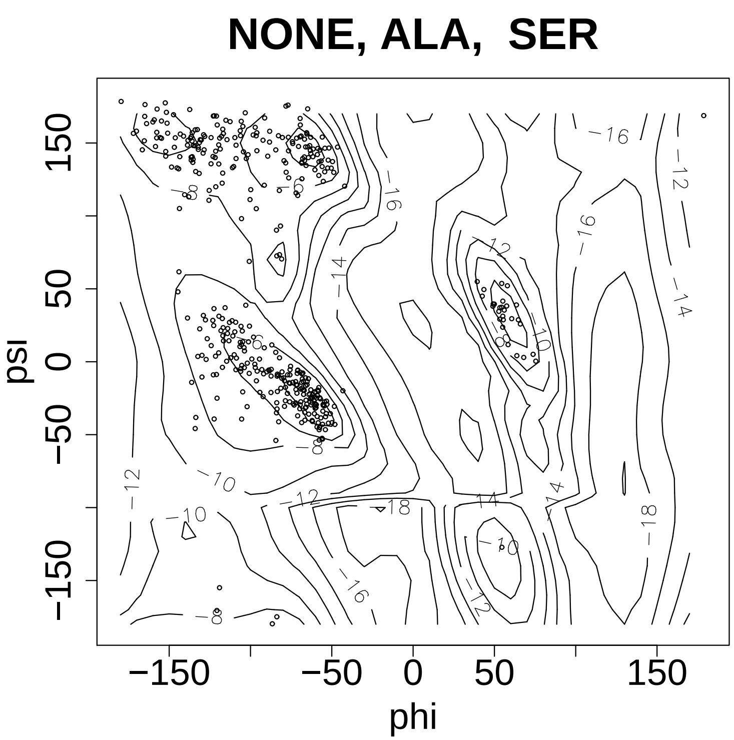
(Xing, Zhu, Jordan & Teh, 2006)

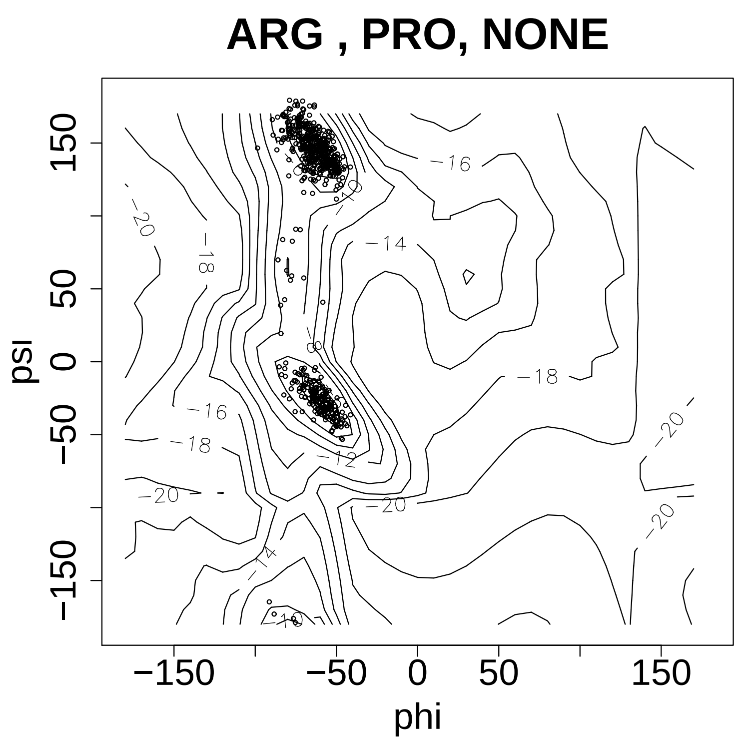
- HapMap data: two populations of CEPH (Utah residents with ancestry from northern and western Europe, CEU) and Yoruba in Ibadan, Nigeria (YRI)
  - these data contain 30 trios of genotypes and thus allow us to infer most of the true haplotypes



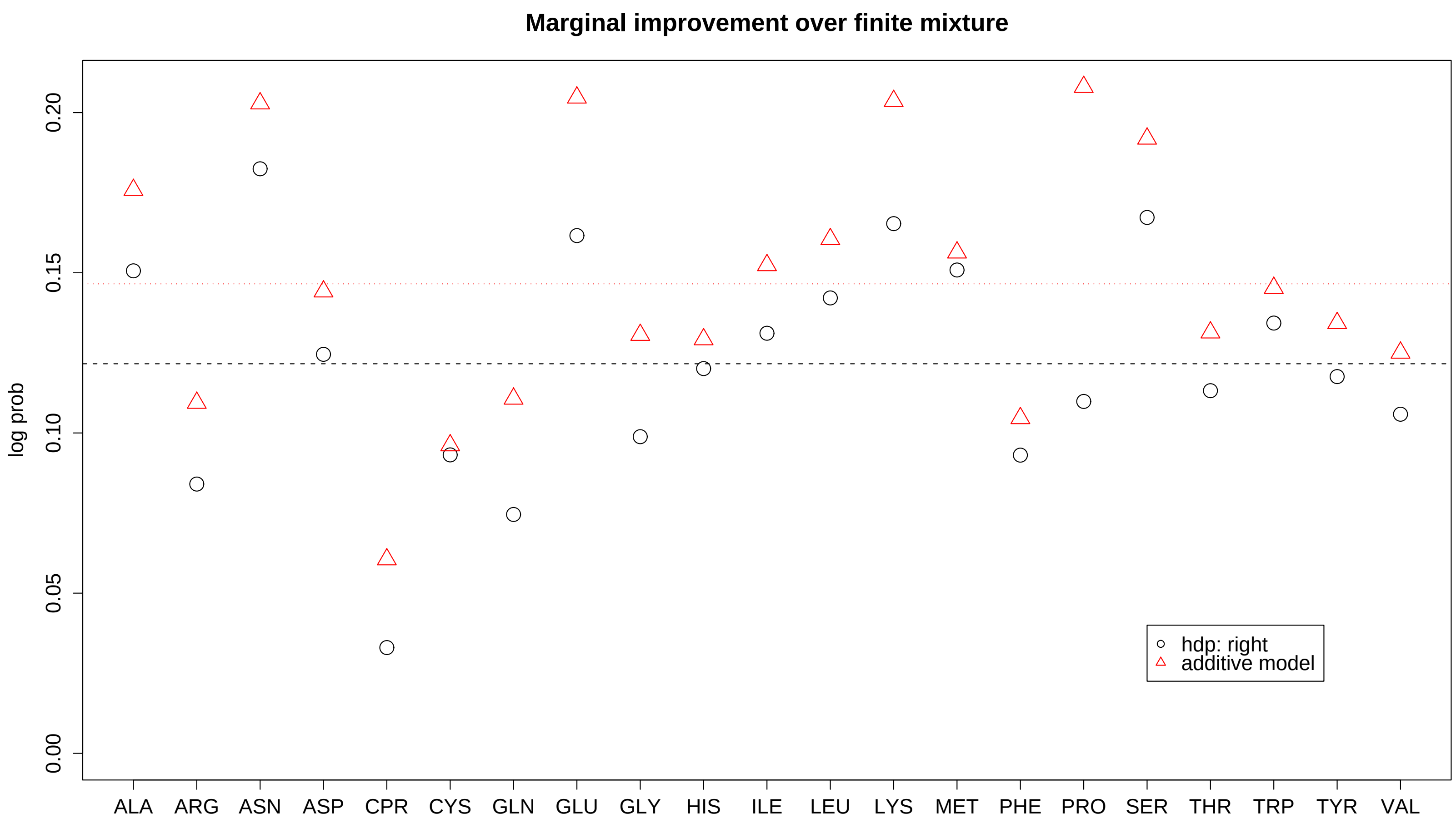
Protein Folding (cont.)

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

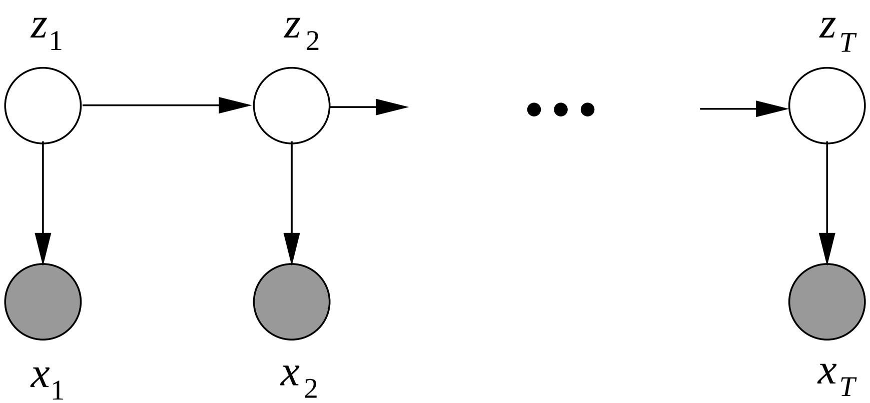




Protein Folding (cont.)



Nonparametric Hidden Markov models (cont.)

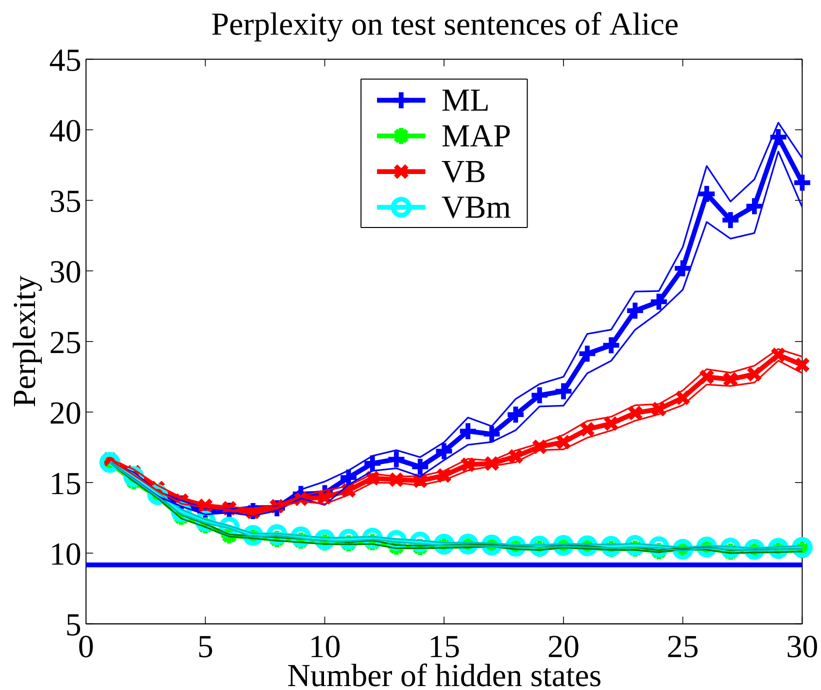


- An open problem—how to work with HMMs that have an unknown and unbounded number of states?

- 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

- Essentially the same idea can be used with hidden Markov trees

Alice in Wonderland



- Perplexity of test sentences taken from Lewis Carroll's *Alice in Wonderland*

Parsing (cont.)

(Liang, Petrov, Jordan & Klein, 2007)

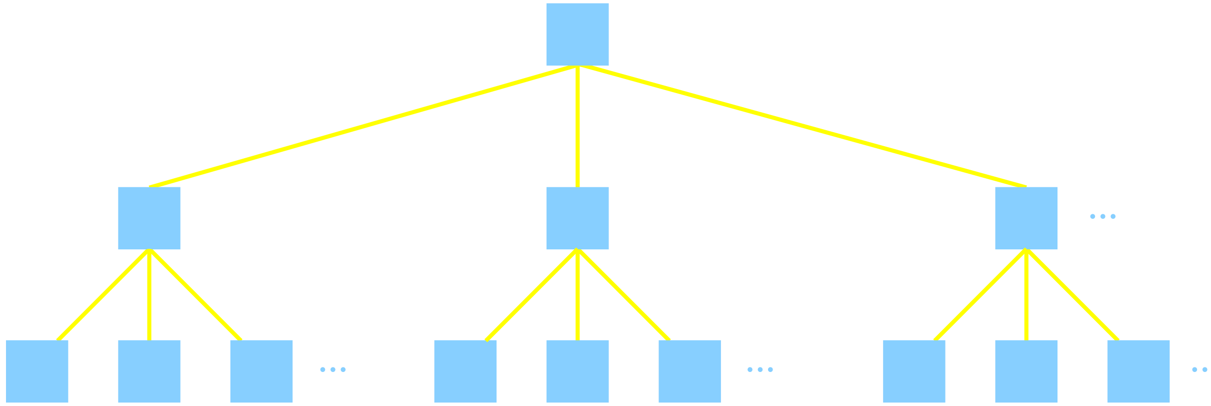
- Based on a training corpus, we build a lexicalized grammar in which the rules are based on word clusters

- 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

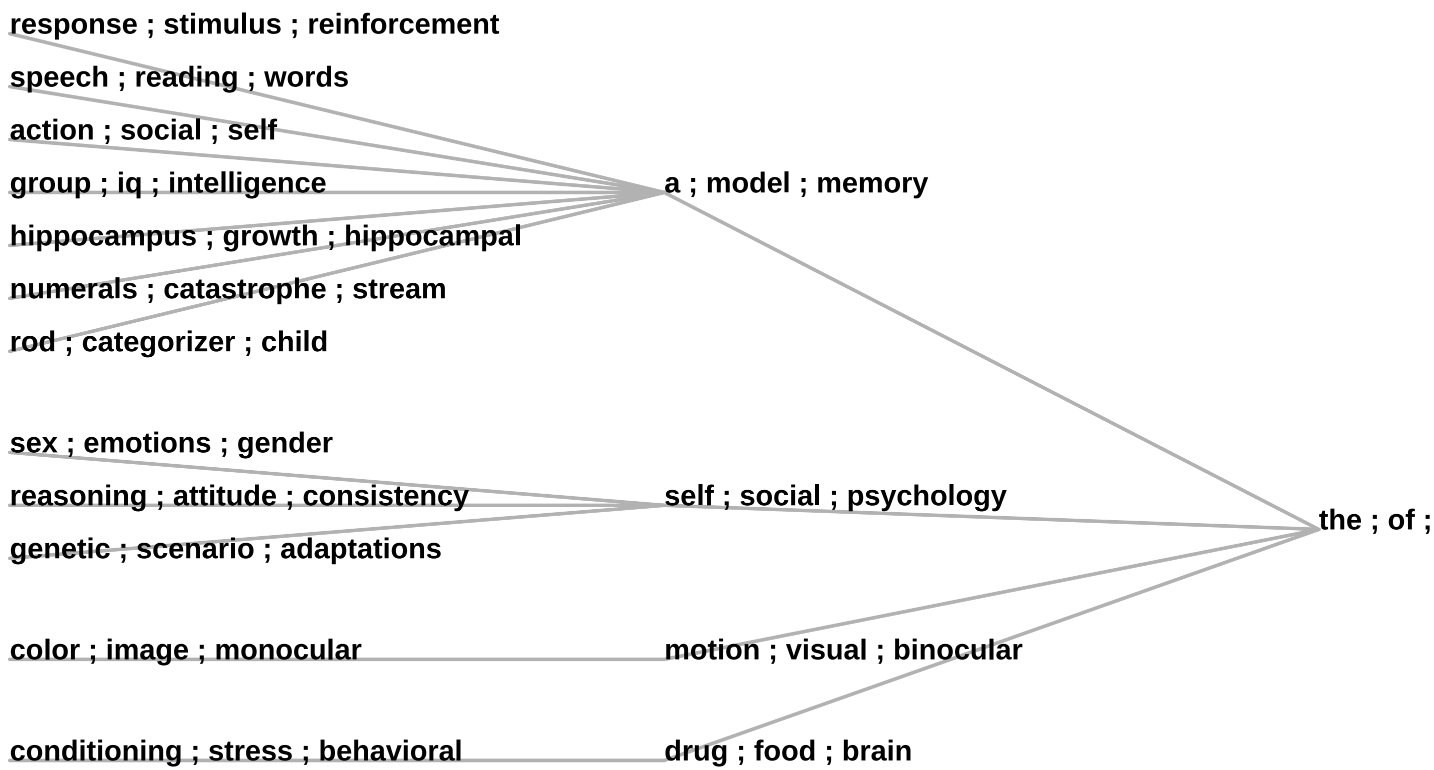
CRP-Based Hierarchical Topic Models

(Blei, et al., 2004)

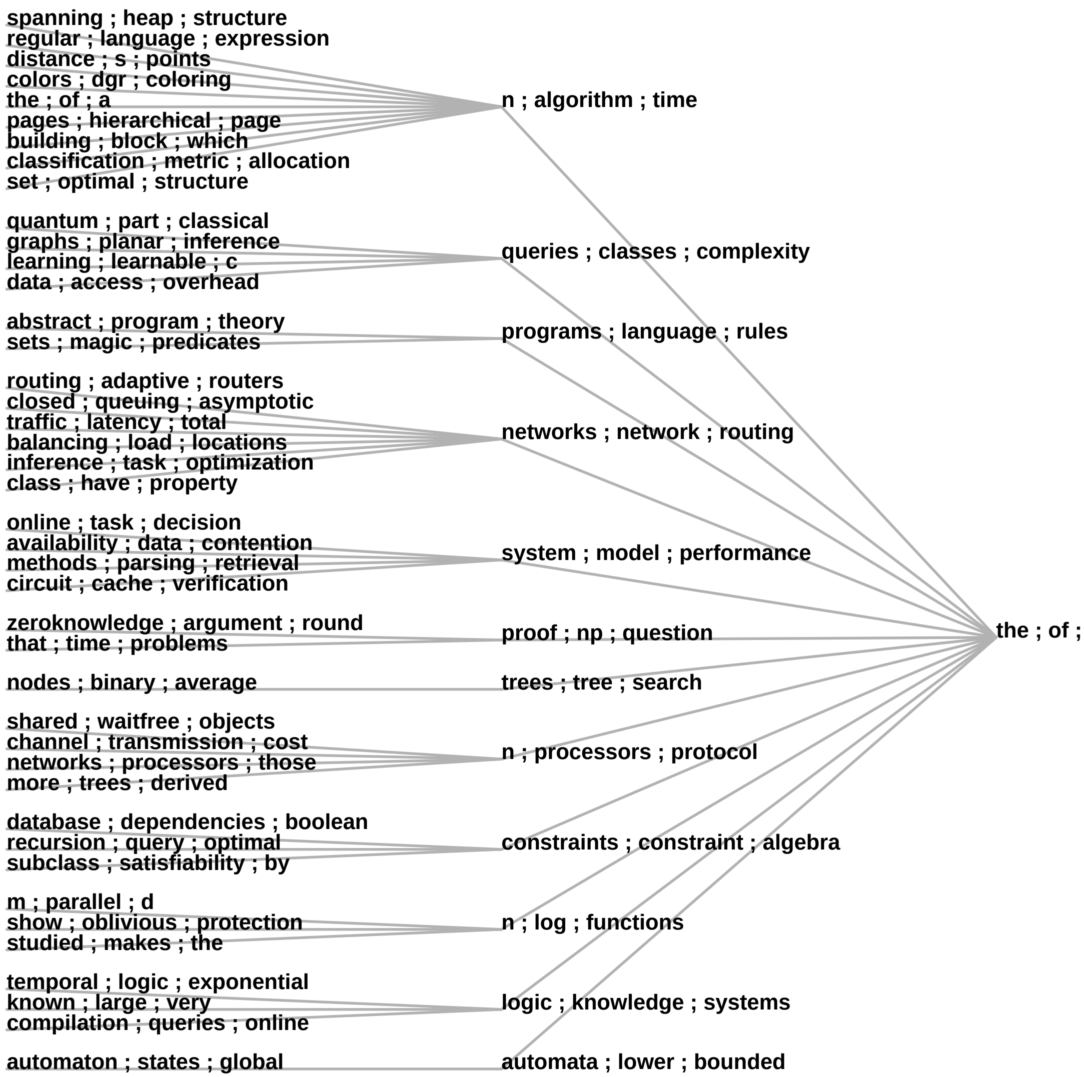


- Each node in the tree is a Chinese restaurant
- Each table in every restaurant has an associated distribution on words (a "topic") drawn from a prior
- Sitting at a table in a given restaurant also selects an outgoing branch, which provides access to further restaurants and further topics
  - we obtain a measure on trees of unbounded depth and unbounded branching factors

Topic Hierarchy from *Psychology Today*



Topic Hierarchy from *JACM*



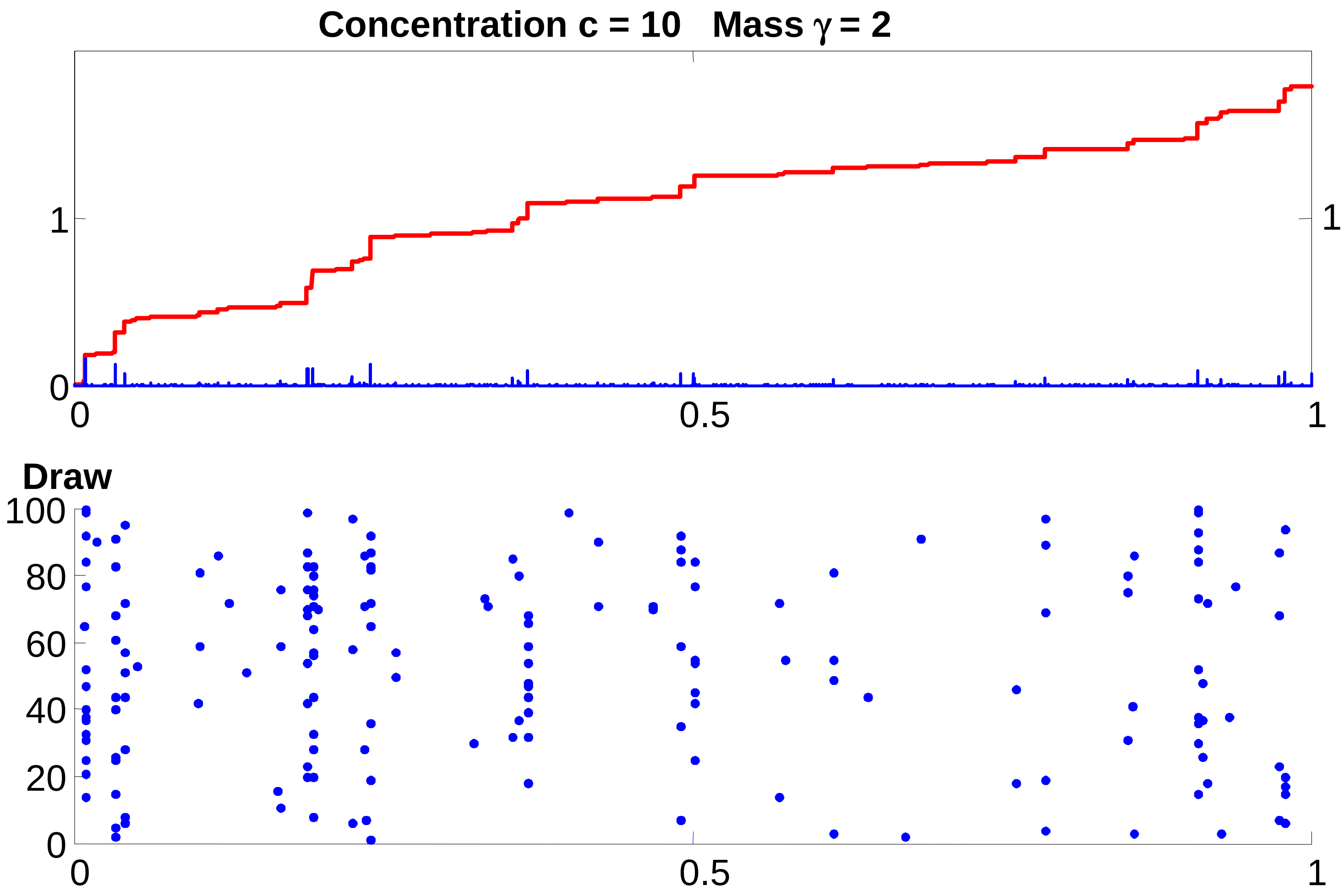
Beta Processes

- 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

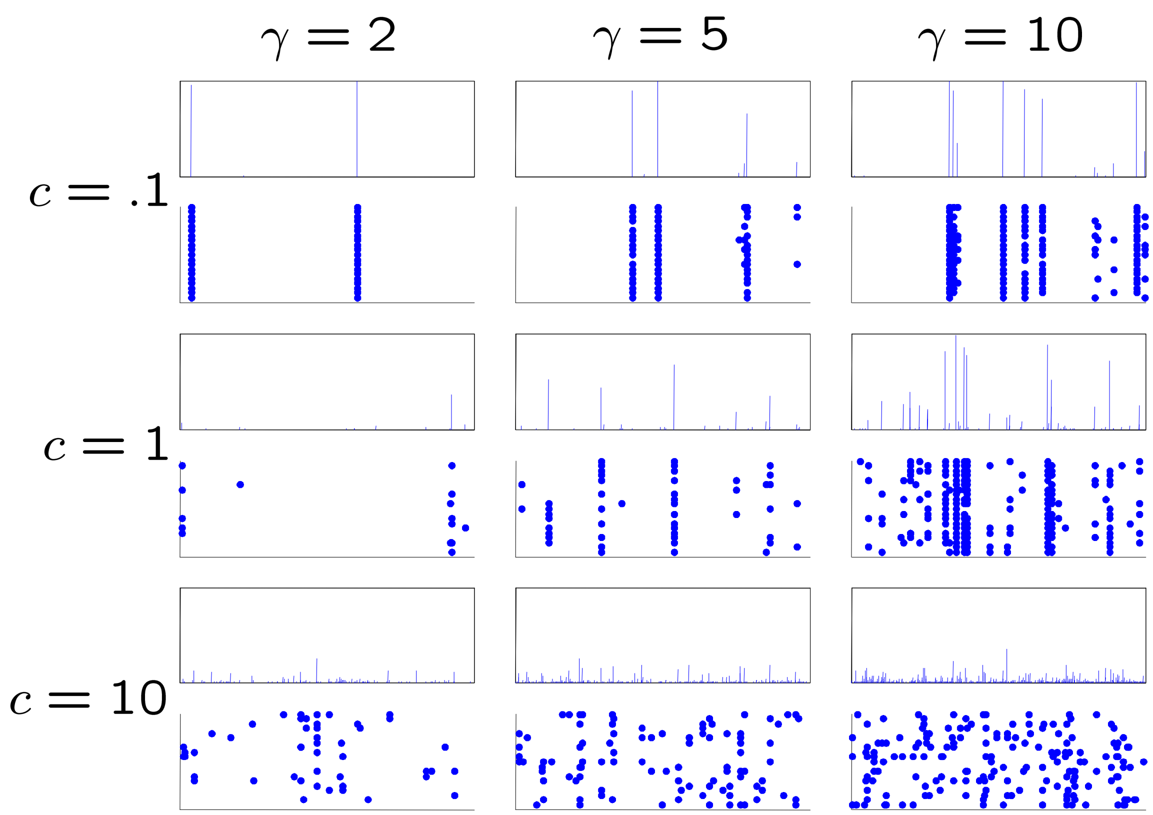
Lévy Processes

- 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

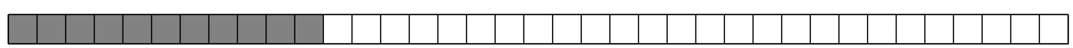
- The marginals of the Dirichlet process are characterized by the Chinese restaurant process

- What about the beta process?

Indian Buffet Process (IBP)

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

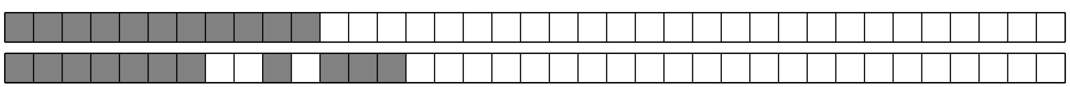
- Indian restaurant with infinitely many dishes in a buffet line
- N customers serve themselves
  - the first customer samples $\mathrm{Poisson}(\alpha)$ dishes
  - the ith customer samples a previously sampled dish with probability $\frac{m_k}{i+1}$ then samples $\mathrm{Poisson}(\frac{\alpha}{i})$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



Indian Buffet Process (IBP)

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

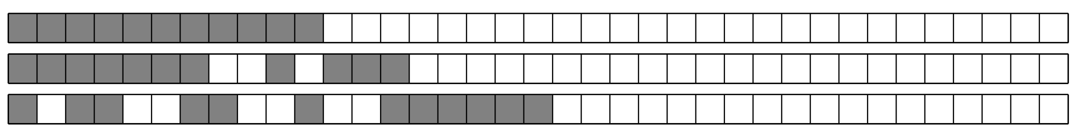
- Indian restaurant with infinitely many infinite dishes
- N customers serve themselves
  - the first customer samples $\mathrm{Poisson}(\alpha)$ dishes
  - the ith customer samples a previously sampled dish with probability $\frac{m_k}{i+1}$ then samples $\mathrm{Poisson}(\frac{\alpha}{i})$ new dishes



Indian Buffet Process (IBP)

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

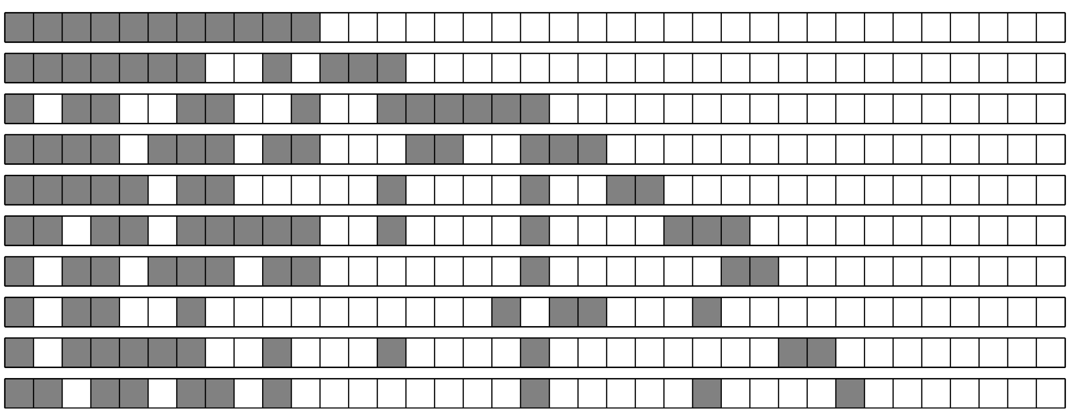
- Indian restaurant with infinitely many infinite dishes
- N customers serve themselves
  - the first customer samples $\mathrm{Poisson}(\alpha)$ dishes
  - the ith customer samples a previously sampled dish with probability $\frac{m_k}{i+1}$ then samples $\mathrm{Poisson}(\frac{\alpha}{i})$ new dishes



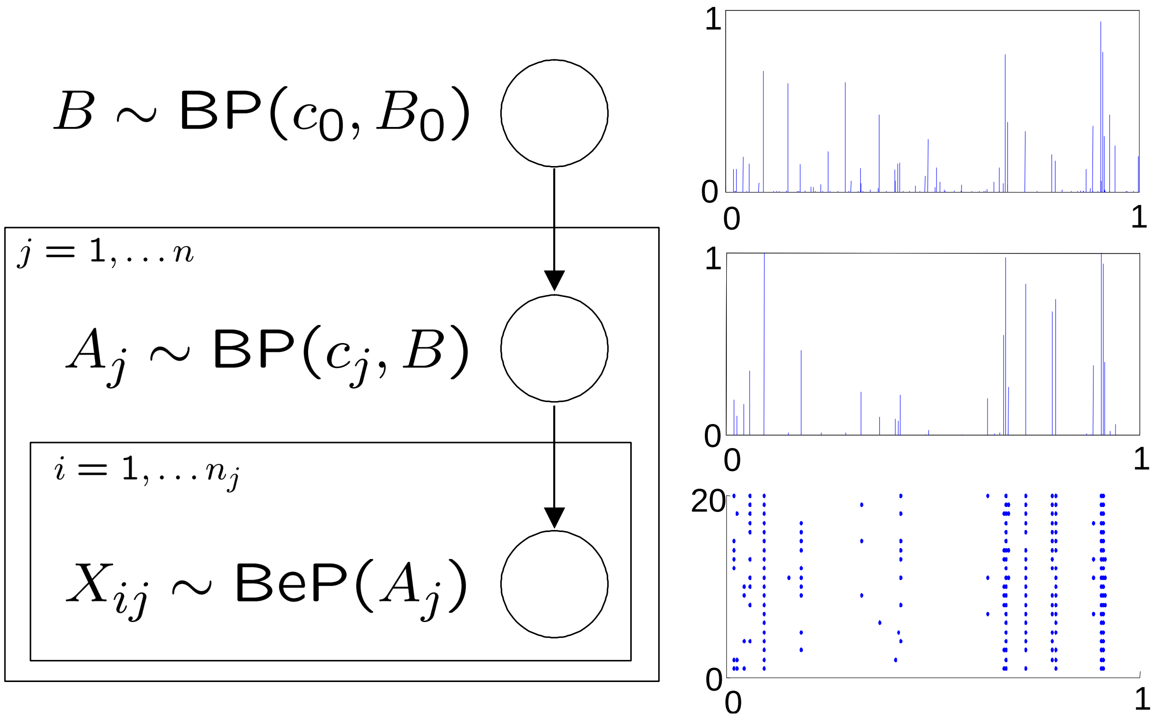
Indian Buffet Process (IBP)

(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.

Applications

- Parsing
  - describe nouns with features such as +`animate`, +`transitive`, +`plural`
- Text categorization
  - describe a document by the words appearing in the document
  - shrink between documents

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/$\sim$jordan