Applied Nonparametric Hierarchical Bayes

Michael I. Jordan

Computer Science Division and Department of Statistics
University of California, Berkeley

http://www.cs.berkeley.edu/∼jordan

Joint work with: Yee Whye Teh and David Blei

Statistical machine learning

- Kernel methods
  - generally frequentist
  - generally nonparametric
  - kernel matrices as a lingua franca
- Probabilistic graphical models
  - generally Bayesian
  - generally parametric
  - modular combination of heterogeneous parts
- An increasingly flexible toolbox with increasingly rigorous theoretical foundations
  - but still far from "flexible data structures"
  - and still a patchy theory

Clustering

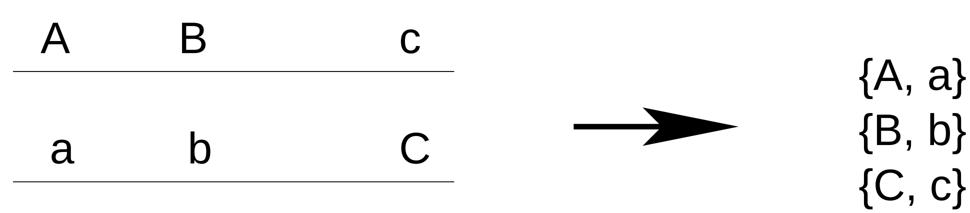
- Many adhoc approaches (e.g., hierarchical clustering)
  - little theoretical understanding
  - hard to use as a building block

- Kernel methods
  - spectral clustering
  - how to choose K?

- Probabilistic graphical models
  - mixture models
  - how to choose K?

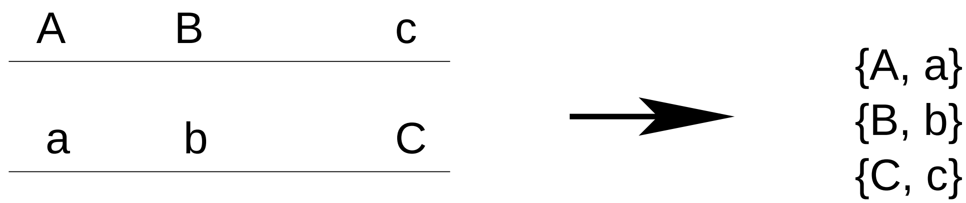
Haplotype modeling

- 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
- Given a sample of *genotypes* (unordered sets of pairs of markers)



  - estimate the underlying haplotypes
- This is a clustering problem

Haplotype modeling (cont.)



- The genotype is a mixture over the population haplotypes:

$$p(g) = \sum_{h_1, h_2 \in \mathcal{H}} p(h_1)p(h_2)p(g \,|\, h_1, h_2),$$

(assuming Hardy-Weinberg equilibrium)

- So this is naturally treated as a mixture modeling problem

- But what is the cardinality of $\mathcal{H}$?

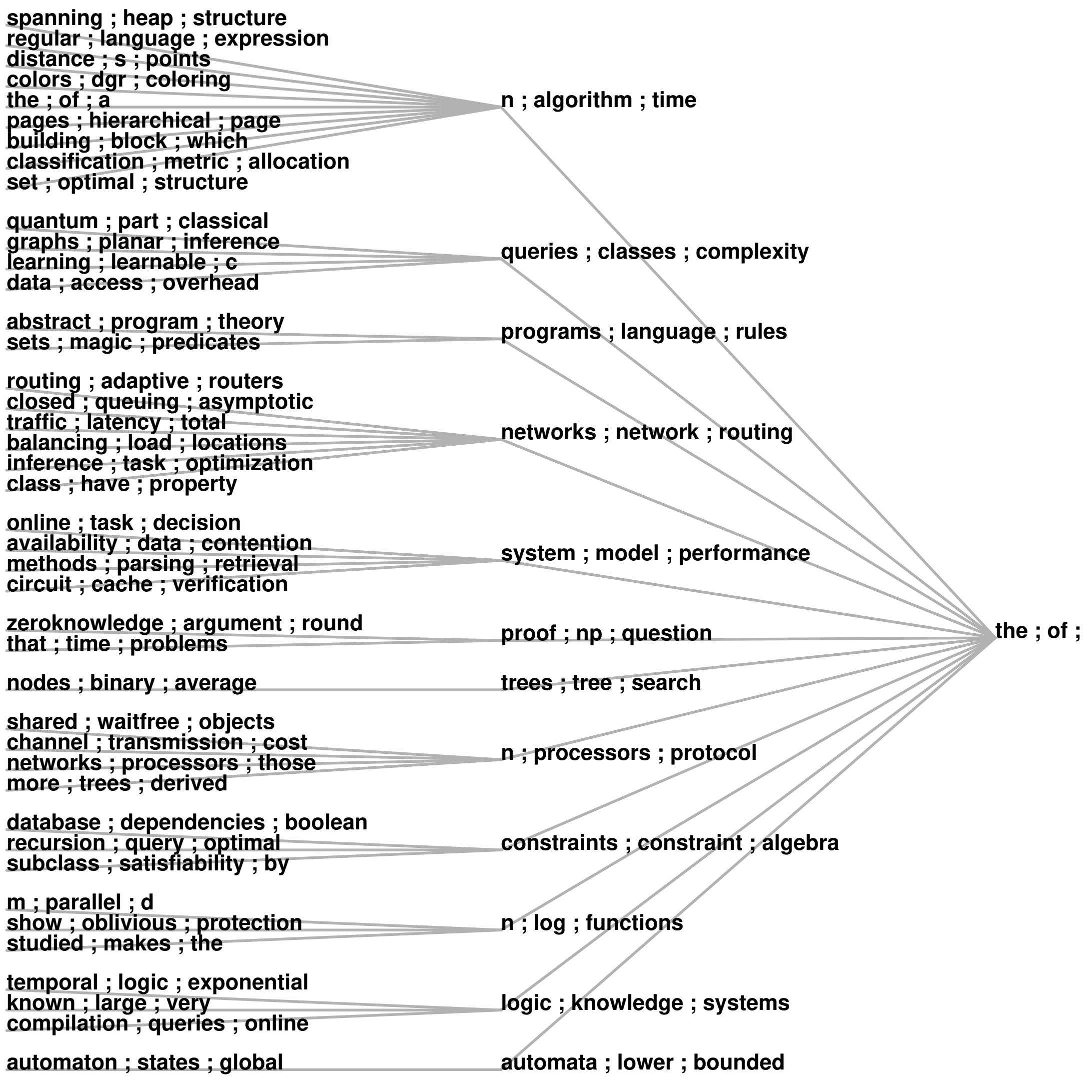
Subpopulations

- Suppose that we stratify the populations by ethnic group (e.g., African, Asian, European)
- Interesting to try to discover what haplotypes they have in common
  - thus we have multiple, linked clustering problems
  - need to choose the number of clusters in each group, and want to share clusters among groups?
- How to do this?

Document modeling

- Ingredients of a statistical document model
  - "bag of words assumption"—exchangeability for the words within a document
  - "topic"—probability distribution across words
  - "corpus"—collection of documents
- Desiderata
  - want to capture relationships among documents within a corpus
  - want to capture relationships between corpora
  - want to do this with hierarchical corpora
- Again, we have multiple, linked clustering problems—how to do this?

Document modeling (cont.)

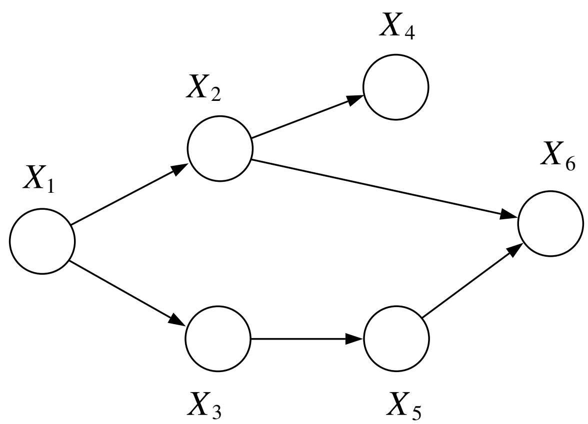


This talk

- What's new
  - hierarchical Dirichlet processes
  - nested Dirichlet processes
  - variational inference for Dirichlet processes
- What's not
  - hierarchical Bayesian modeling
  - Dirichlet processes
  - inference for Dirichlet processes

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:

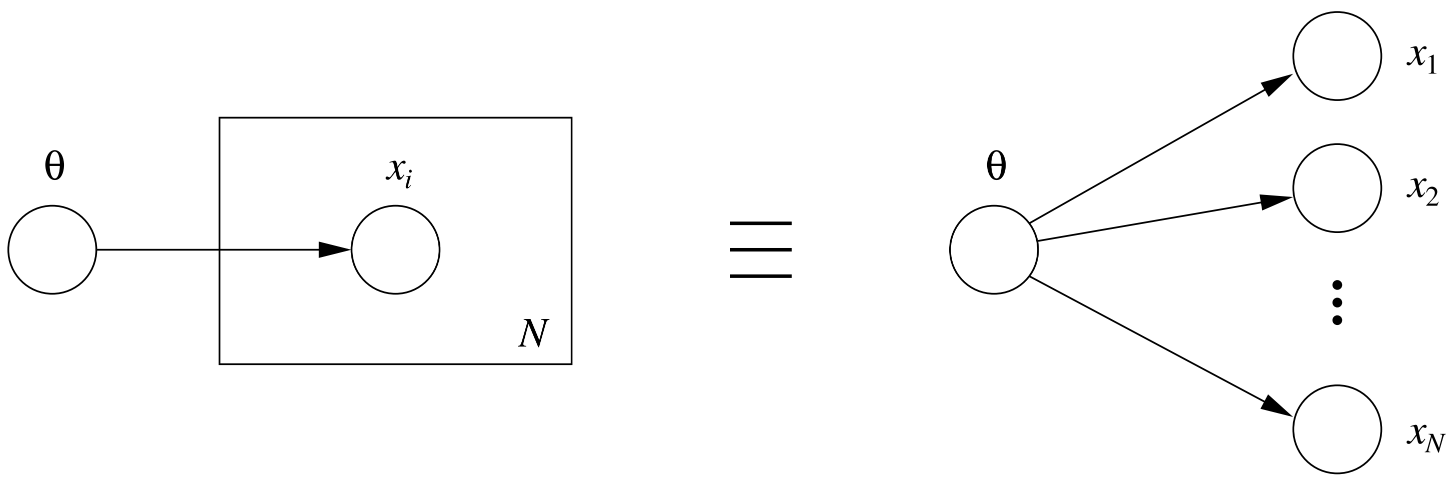


- The joint distribution on $(X_1, X_2, \ldots, X_N)$ factorizes according to the "parent-of" relation defined by the edges $\mathcal{E}$:

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

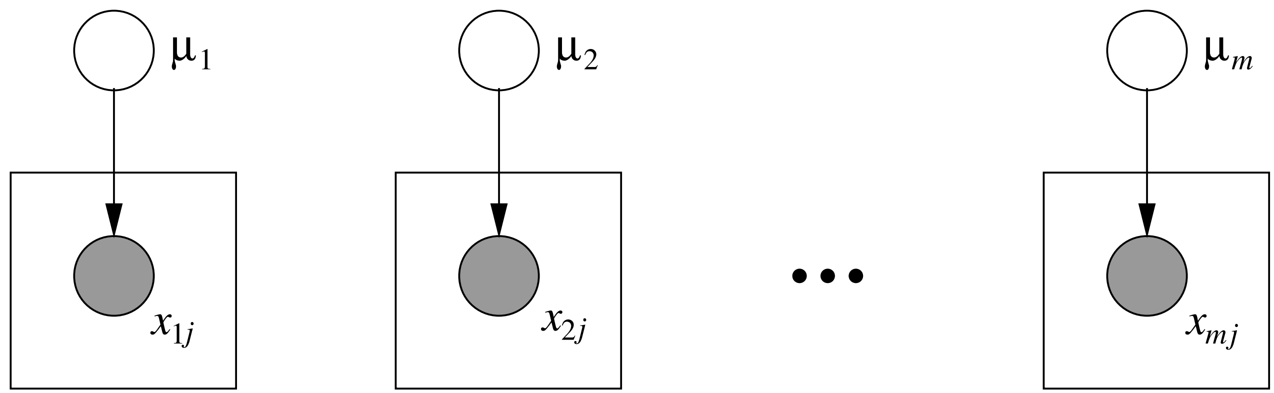
Plates

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



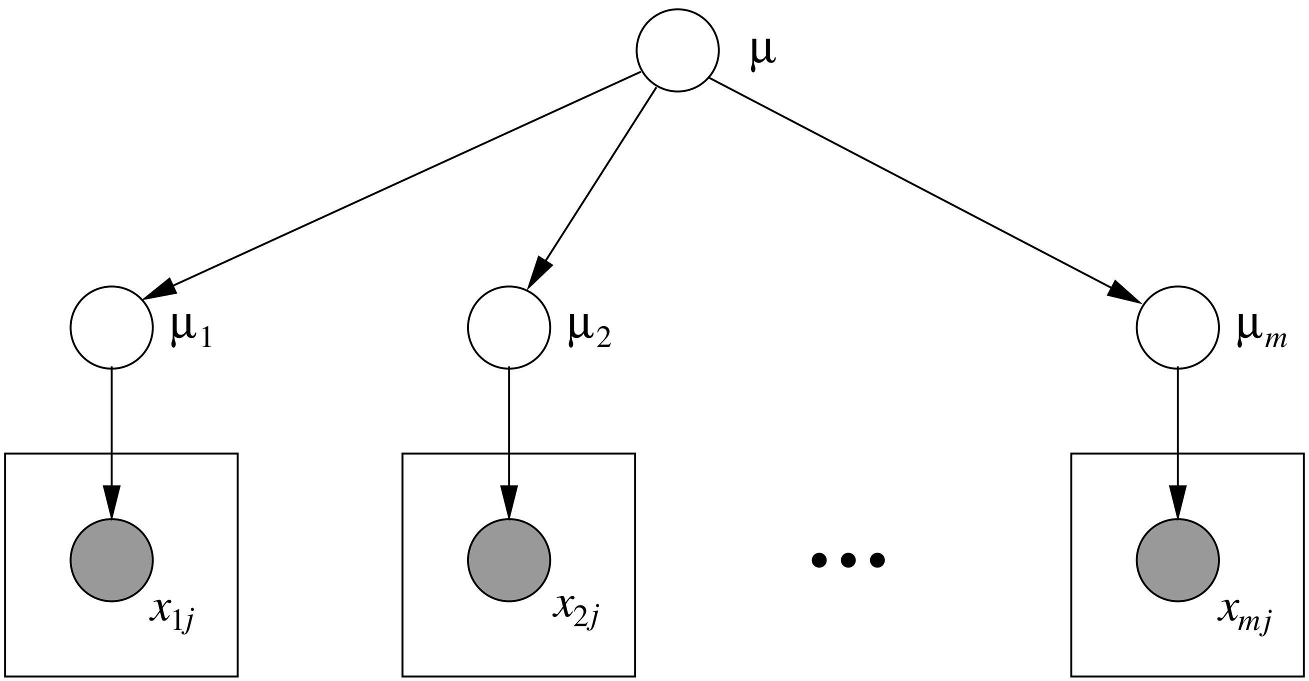
- Graphical representation of an exchangeability assumption on $(X_1, X_2, \ldots, X_N)$

Multiple inference problems



- Multiple Gaussian means
- $x_{ij} \sim N(\mu_i, \sigma_i^2)$
- Want to “share statistical strength”

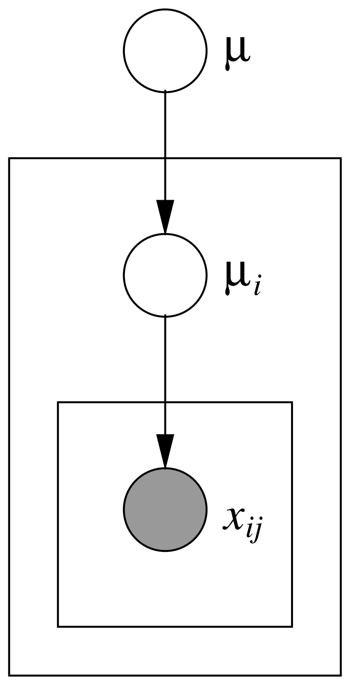
Hierarchical Bayesian modeling



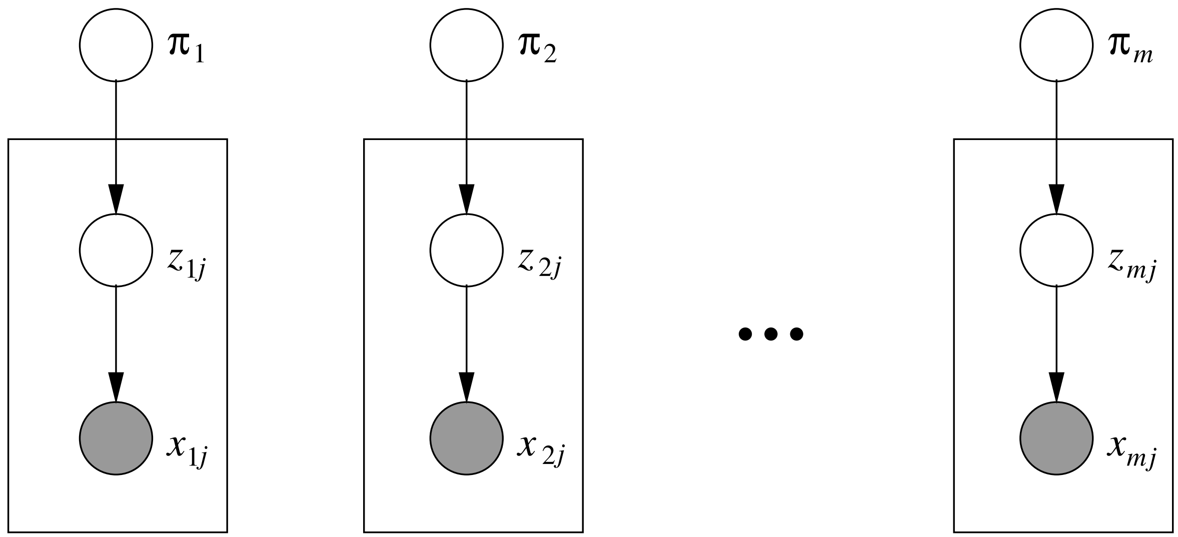
- Posterior mean is a shrinkage estimator

Hierarchical modeling

- Using the plate notation:



Multiple clustering problems

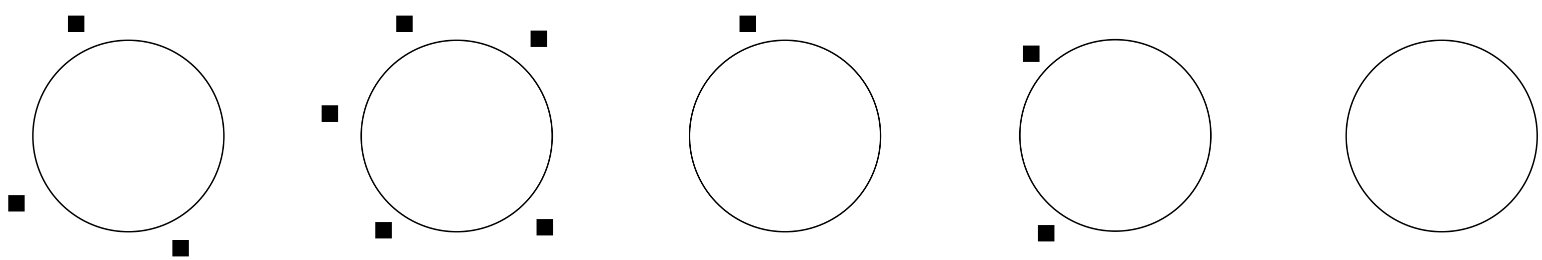


- mixture models with explicit allocations z_{ij}
- for each i: $p(x_{ij}|\pi_i, \theta_i, k_i) = \sum_{l=1}^{k_i} p(z_{ij} = l|\pi_i)p(x_{ij}|z_{ij} = l, \theta_i)$
- what to share: π_i?, θ_i?, k_i?
- how to share?

Chinese Restaurant Process (CRP)

- A 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:

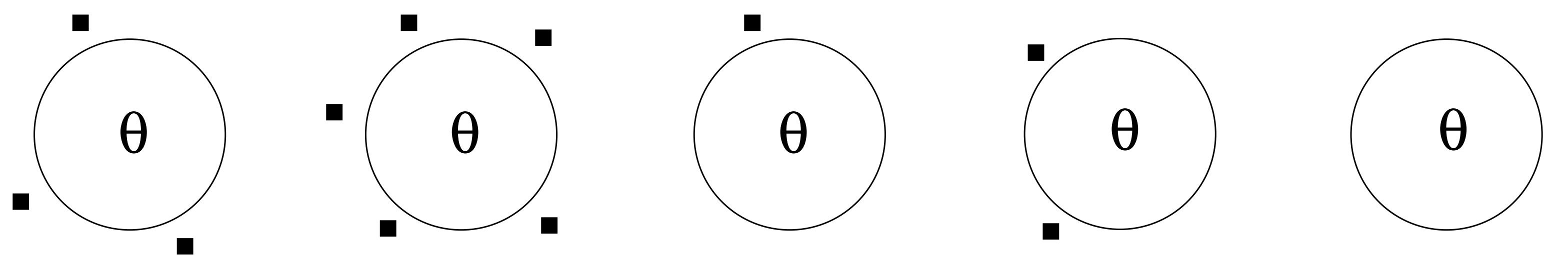
$$\begin{aligned} p(\text{previously occupied table } i \mid \mathcal{F}_{m-1}) &= \frac{m_i}{\alpha_0+m-1} \\ p(\text{the next unoccupied table} \mid \mathcal{F}_{m-1}) &= \frac{\alpha_0}{\alpha_0+m-1}, \end{aligned} \tag{1}$$



- Defines an exchangeable distribution on partitions of customers

The CRP and mixture models

- Associate a mixture component with each table
  - the first customer to sit at a table draws a parameter from the prior

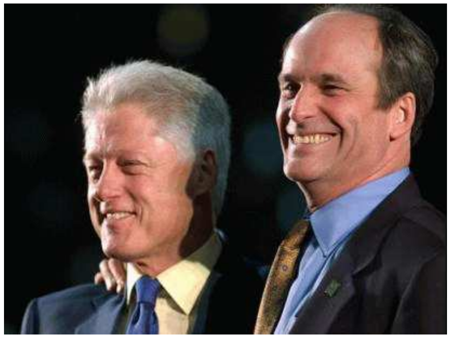


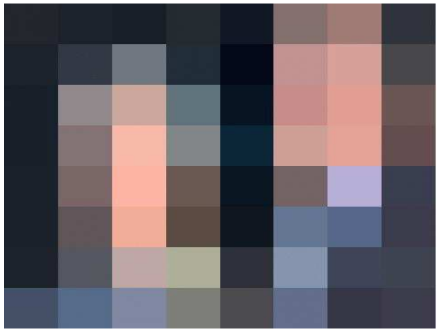
- This defines a prior on number of clusters and on the parameters associated with each cluster

- The likelihood is the usual mixture likelihood, but with an infinite number of mixture components

- Posterior inference can be performed via (e.g.) a Gibbs sampler

Example





- 5000 images from the Associated Press
- Each image represented as an 8×8 matrix of RGB values
- Modeled as a 192 dimensional Gaussian
- Clustered using a nonparametric Bayes mixture model
  - 70 clusters in the 5000 images
  - captures something about the layout of news photos

Sample clusters

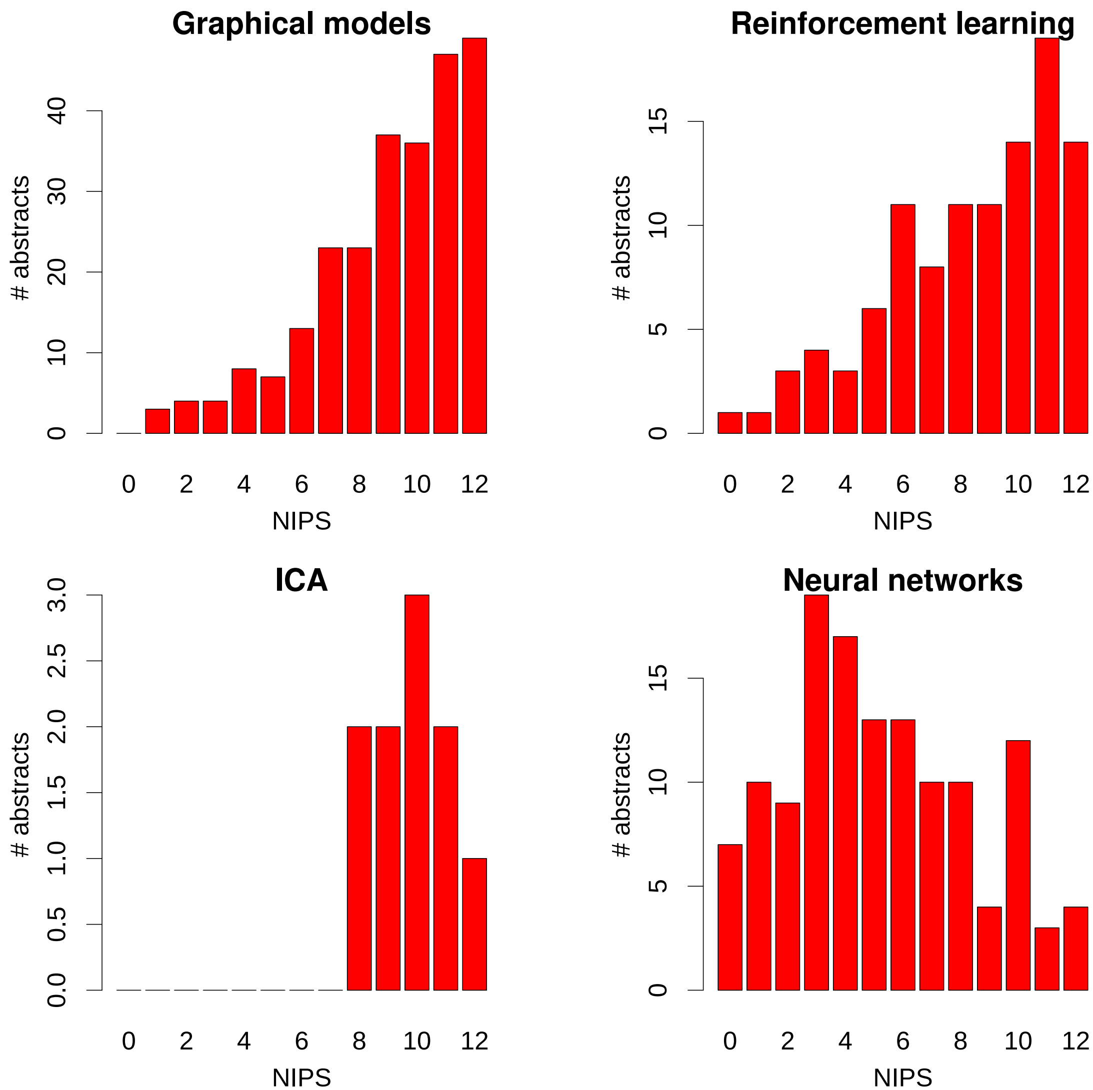
NIPS data

Increasing attention has recently been paid to algorithms based on dynamic programming (DP) due to the suitability of DP for learning problems involving control. In stochastic environments where the system being controlled is only incompletely known, however, a unifying theoretical account of these methods has been missing. In this paper we relate DP-based learning algorithms to the powerful techniques of stochastic approximation via a new convergence theorem, enabling us to establish a class of convergent algorithms to which both TD(λ) and Q-learning belong.

- 1718 abstracts from NIPS 0 – 12
- Modeled as a 4100-dimensional multinomial distribution
  - 26 clusters
  - capture different "topics" prevalent in the conference

"GM"	"RL"	"NN"	"ICA"
data	learning	network	eeg
model	reinforcement	networks	ica
algorithm	control	learning	channel
learning	algorithm	neural	data
models	function	paper	signals
problem	policy	time	source
networks	problem	training	artifacts
show	optimal	recurrent	independent
method	paper	input	changes
approach	state	method	results
based	problems	architecture	problem
paper	value	structure	components
new	algorithms	rules	time
results	methods	units	analysis
bayesian	model	problem	cell

Visualizing trends



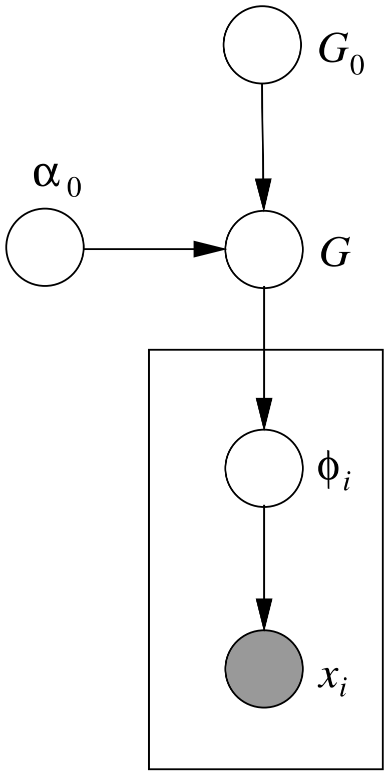
Dirichlet process

Definition 1. *Let* $(\Theta, \mathcal{B})$ *by a measurable space, with* G_0 *a probability measure on the space, and let* α_0 *be a positive real number. A* Dirichlet process *is the distribution of a random probability measure* G *over* $(\Theta, \mathcal{B})$ *such that, for any finite partition* $(A_1, \ldots, A_r)$ *of* Θ*, the random vector* $(G(A_1), \ldots, G(A_r))$ *is distributed as a finite-dimensional Dirichlet distribution:*

$$(G(A_1), \ldots, G(A_r)) \sim \mathrm{Dir}\big(\alpha_0 G_0(A_1), \ldots, \alpha_0 G_0(A_r)\big) \tag{2}$$

We write $G \sim \mathrm{DP}(\alpha_0, G_0)$ *if* G *is a random probability measure distributed according to the Dirichlet process. Call* G_0 *the base measure of* G *and call* α_0 *the concentration parameter.*

Dirichlet process mixture models



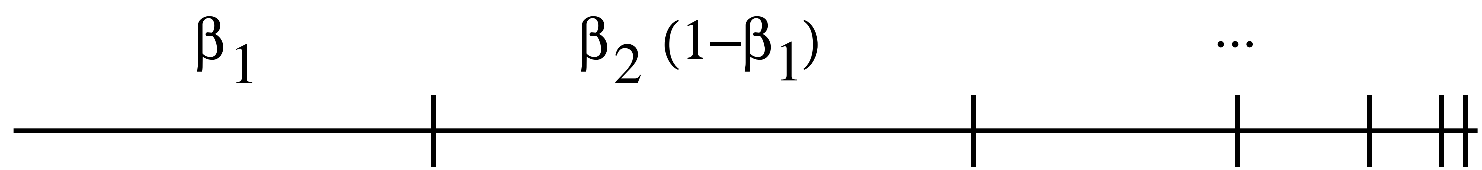
$$
\begin{array}{rcll}
G & \sim & \mathrm{DP}(\alpha_0, G_0) & \\
\phi_i \,|\, G & \sim & G & \quad i \in 1, \ldots, n \\
x_i \,|\, \phi_i & \sim & F(x_i \,|\, \phi_i) & \quad i \in 1, \ldots, n
\end{array}
$$

Stick-breaking representation

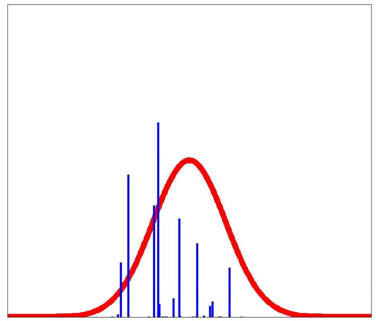
$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$$

where

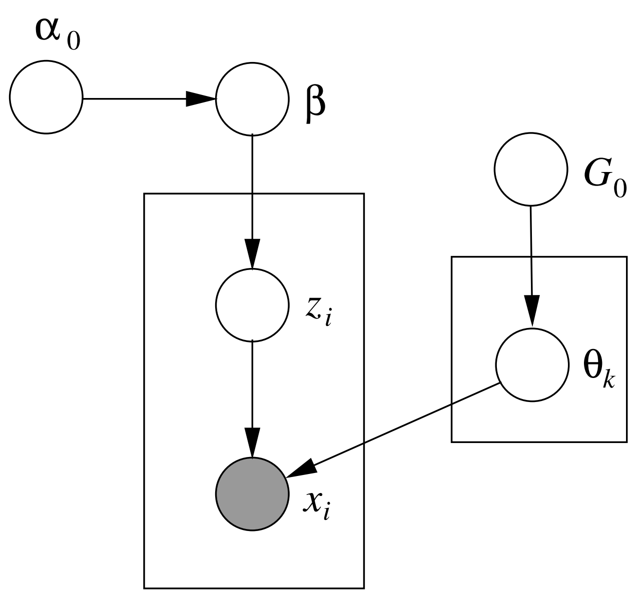
$$\theta_k \sim G_0 \qquad \beta_k \sim \mathrm{Beta}(1, \alpha_0) \qquad \pi_k = \beta_k \textstyle\prod_{l=1}^{k-1}(1-\beta_k)$$



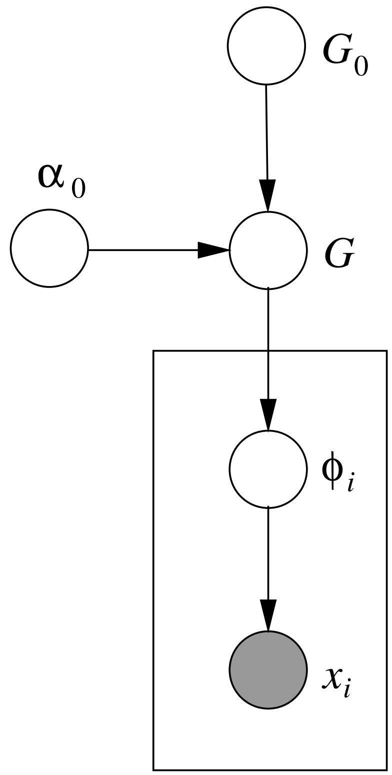
Stick-breaking representation

- A representation of G: 

- A graphical model representation:



Integrating out G



- Integrating out G yields a joint distribution on $\phi_1, \phi_2, \ldots, \phi_N$

- This distribution is precisely that given by the Chinese restaurant process

Inference for Dirichlet process mixtures

- Gibbs sampling
  - based on the Chinese restaurant process
  - based on the stick-breaking representation
- Variational inference
  - based on the stick-breaking representation

Sharing clusters among multiple groups

- Idea: associate a Dirichlet process with each group
- How to link them?
  - let them share the same underlying G_0?
- Problem: the atoms generated by these processes will be distinct with probability one
  - i.e., there will be sharing of statistical strength, but no sharing of clusters
- Need to have the base measure G_0 be discrete
  - but also need it to be flexible and random

Hierarchical Dirichlet process

- Let G_0 itself be distributed according to a DP:

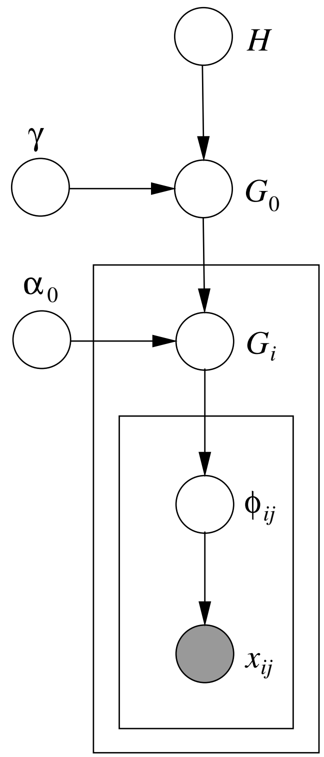
$$G_0 \,|\, \gamma, H \;\sim\; \mathrm{DP}(\gamma, H)$$

- Then

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

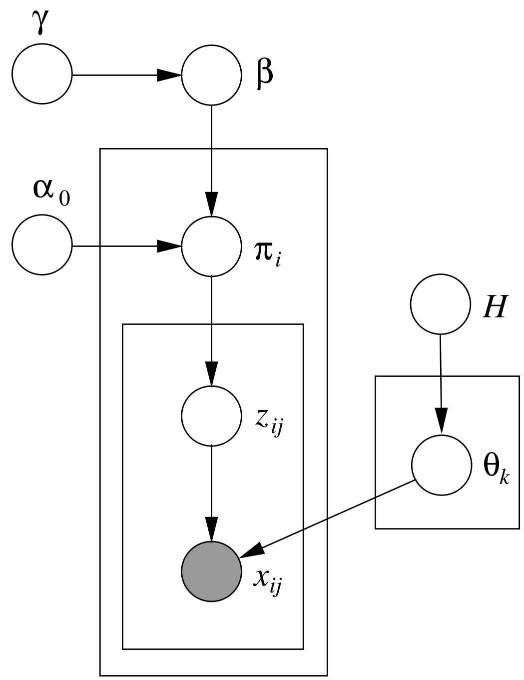
  has as its base measure an atomic distribution—samples of G will resample from these atoms

Hierarchical Dirichlet process mixture



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

Stick-breaking representation



$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k} \qquad\qquad G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\theta_k}$$

$$\pi_{jk} = \pi'_{jk} \prod_{l=1}^{k-1} (1 - \pi'_{jl}) \qquad\qquad \pi'_{jk} \sim \mathrm{Beta}\left(\alpha_0 \beta_k, \alpha_0 \left(1 - \sum_{l=1}^{k} \beta_l\right)\right)$$

Results: Nematode database

- data set of 5838 abstracts on nematode biology
- removed standard stop words and words appearing less than 10 times: yields a vocabulary size of 5699, with 476441 words in total
- the basic model is a multinomial mixture, one per abstract

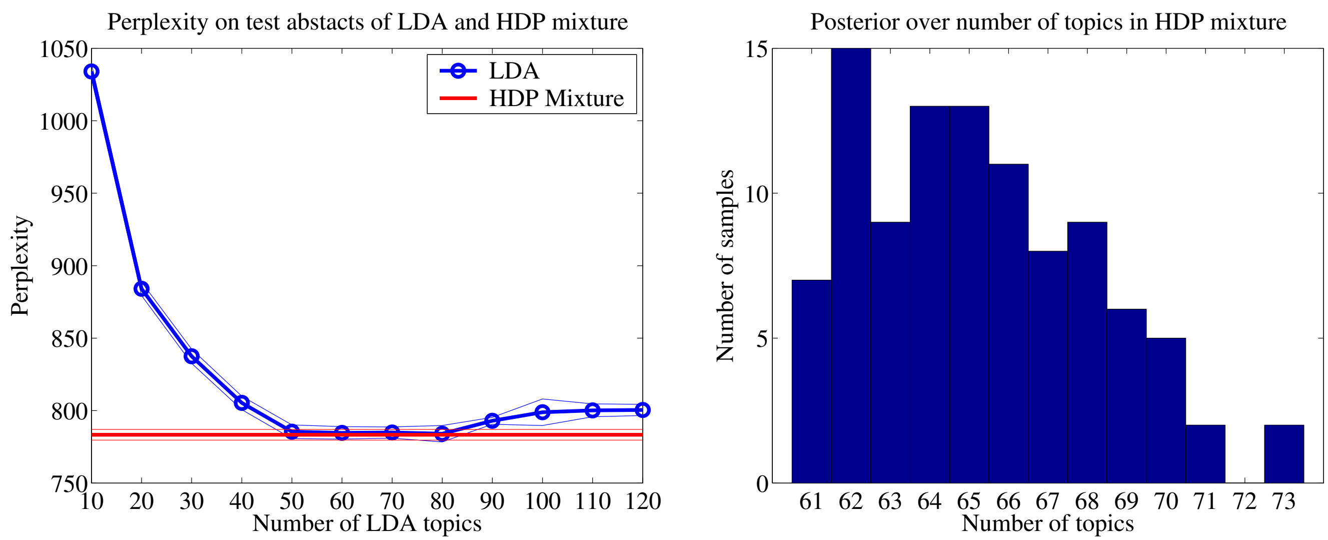


Figure 1: Left: perplexity on test abstracts of LDA and HDP mixture. Right: histogram of the number of topics over 100 posterior samples.

NIPS Conference articles (1988-2001)

- articles from the conference are divided into sections:

AA	algorithms and architectures
AP	applications
CS	cognitive science
CN	control and navigation
IM	implementations
NS	neuroscience
SP	signal processing
LT	learning theory
VS	vision

- each article is represented as a mixture model (over words in the vocabulary)
- an HDP is used to discover and share clusters ("topics") among articles within each section
- want to examine relationships among the sections

Models

- Each article is a DP mixture model
- Each section is a collection of mixture models—thus a section is modeled via an HDP mixture
- We have multiple sections
  - thus we require another level of the hierarchy to link the section HDPs—easily done
- Models:

"*none*"	a separate HDP for each section
"*flat*"	a single HDP for all sections
"*hier*"	a linked set of HDPs

- In presenting the results, we focus on one section (VS) and consider one other section at a time

Results

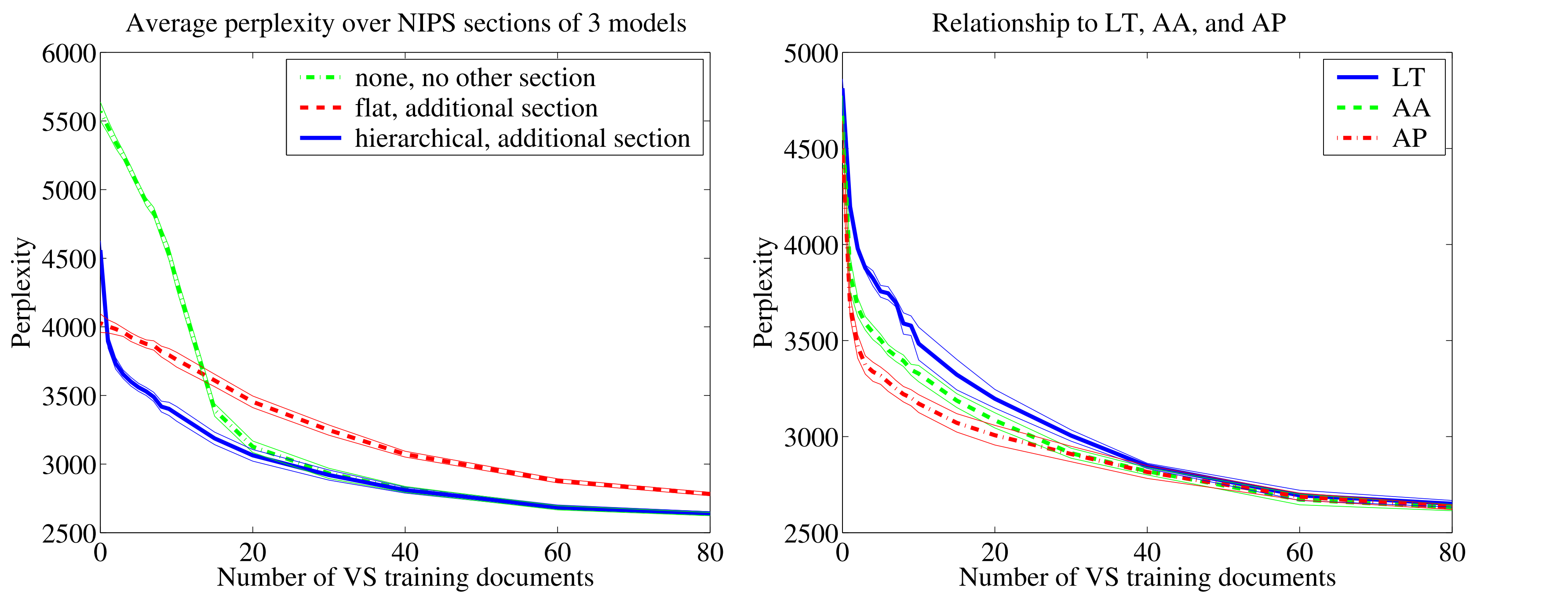


Figure 2: Left: Average perplexity of test VS documents given training documents from VS and another section for 3 different models. curves shown are averaged over the other sections and other 5 runs. Right: Average perplexity of test VS documents given LT, AA and AP documents respectively using M3, averaged over 5 runs.

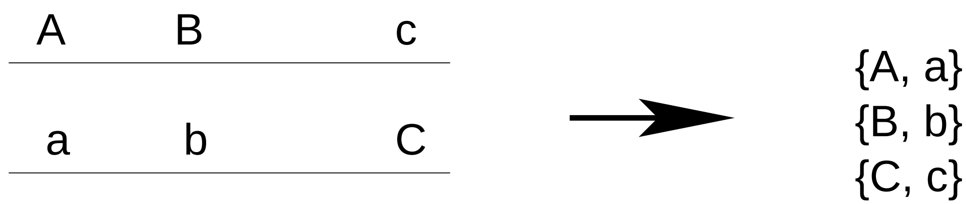
Shared topics

- Topics shared between VS and the other sections
  - the two highest probability topics are displayed

CS	NS	LT	AA	IM	SP	AP	CN
task representation pattern processing trained representations three process unit patterns	cells cell activity response neuron visual patterns pattern single	signal layer gaussian cells figure nonlinear rate equation cell	algorithms test approach methods based point problems large paper	processing pattern approach architecture single shows simple based large	visual images video language image pixel acoustic delta lowpass	approach based trained test layer features table classification rate paper	tree pomdp observable strategy class stochastic history strategies density
examples concept similarity bayesian hypotheses generalization numbers positive classes hypothesis	visual cells cortical orientation receptive contrast spatial cortex stimulus tuning	large examples form point see parameter consider random small optimal	distance tangent image images transformation transformations pattern vectors convolution simard	motion visual velocity flow target chip eye smooth direction optical	signals separation signal sources source matrix blind mixing gradient eq	image images face similarity pixel visual database matching facial examples	policy optimal reinforcement control action states actions step problems goal

Haplotype modeling

(Xing, Sharan, & Jordan, 2004)



- 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
- Given a sample of *genotypes* of a sample from a population (pairs of alleles with the association of alleles to chromosomes unknown):
  - estimate the underlying haplotypes (the *haplotype phase*)
- This is a mixture modeling problem—use the DP prior

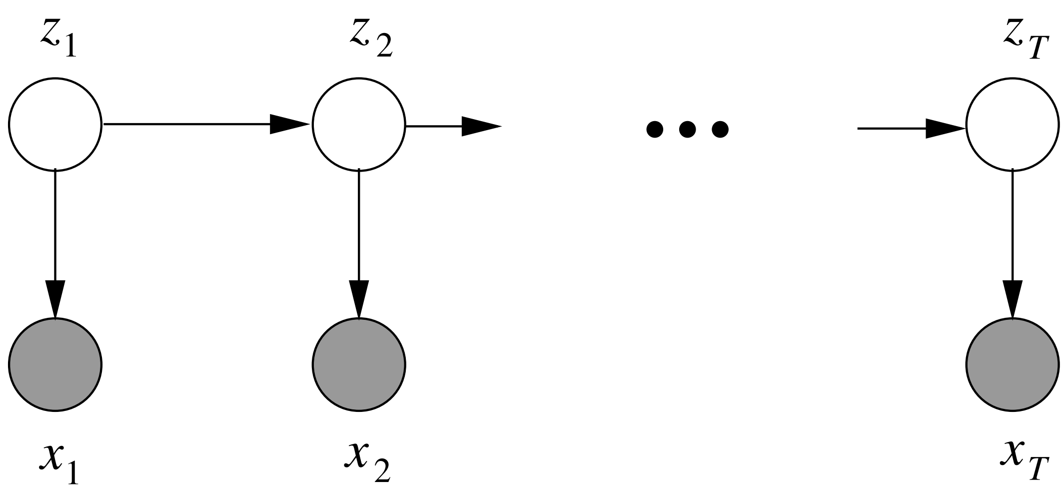
Haplotype modeling (cont.)

- Performance of DP prior on the data of Gabriel, et al (2002):

		DP			PHASE		
region	length	err_s	err_i	d_s	err_s	err_i	d_s
16a	13	0.185	0.480	0.141	0.174	0.440	0.130
1b	16	0.100	0.250	0.160	0.200	0.450	0.180
25a	14	0.135	0.353	0.115	0.212	0.588	0.212
7b	13	0.105	0.278	0.066	0.145	0.444	0.092

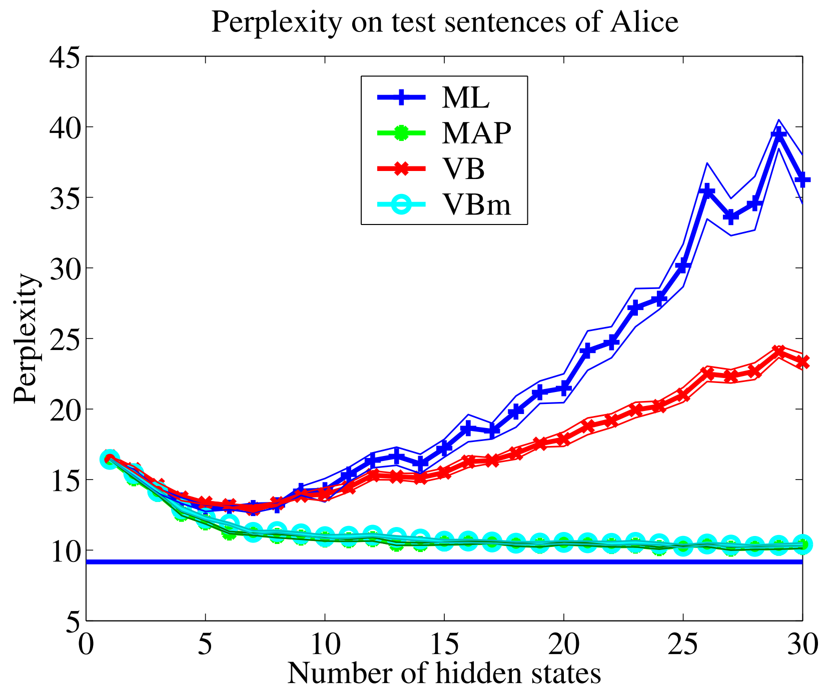
- Haplotype sharing among ethnic groups using the HDP prior:
  - order of magnitude more haplotypes among Africans than Europeans and Asians
  - transfer between Europeans and Asians via the Africans

Hidden Markov models



- The "infinite hidden Markov model"—an HMM with an unbounded number of states
- Straightforward to use 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

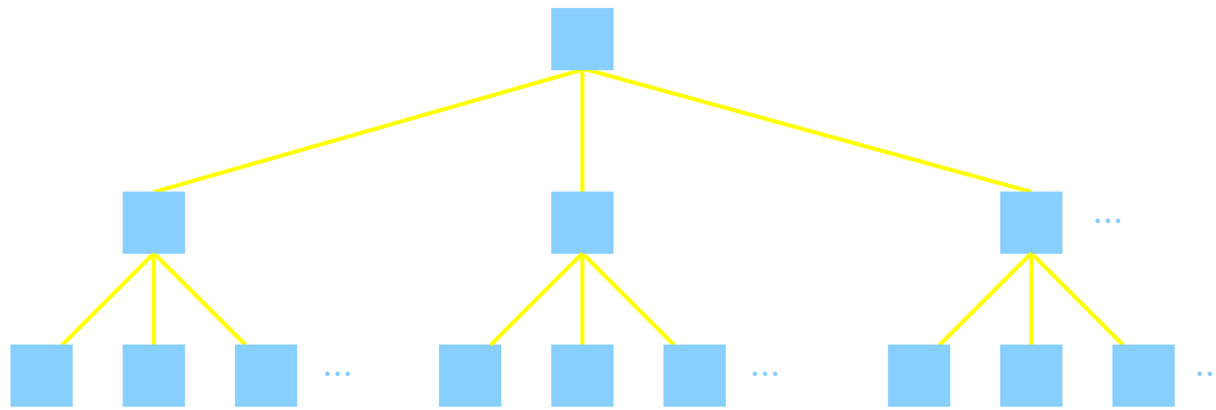
Alice in Wonderland



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

Hierarchical topic models

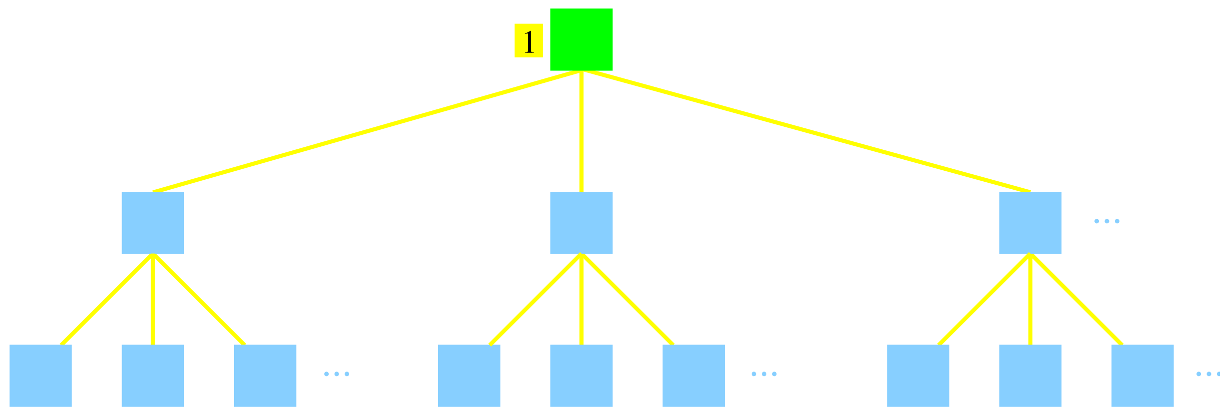
(Blei, Griffiths, Jordan, & Tenenbaum, 2004)



- Infinite number of restaurants in a city:
  - One restaurant is the root restaurant and 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

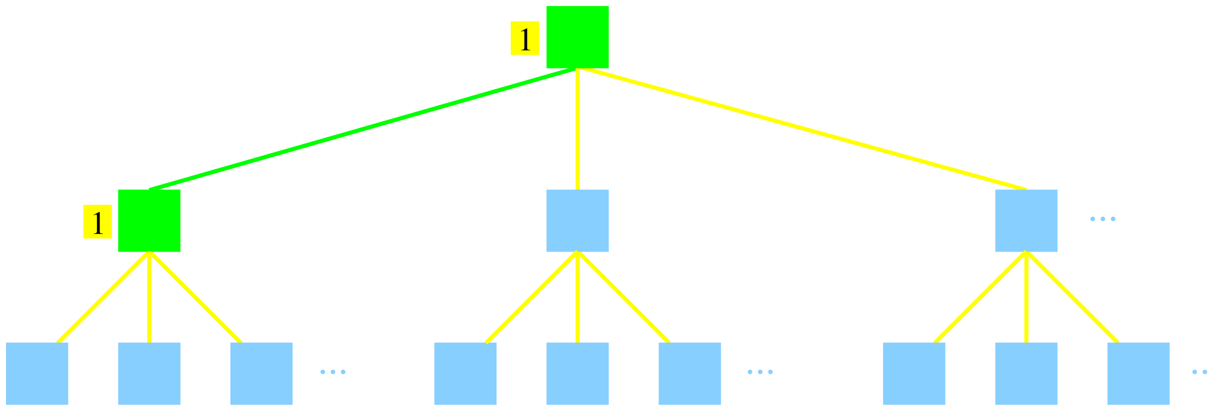
- Restaurants are organized into an infinitely branching tree

Nested Chinese restaurant process



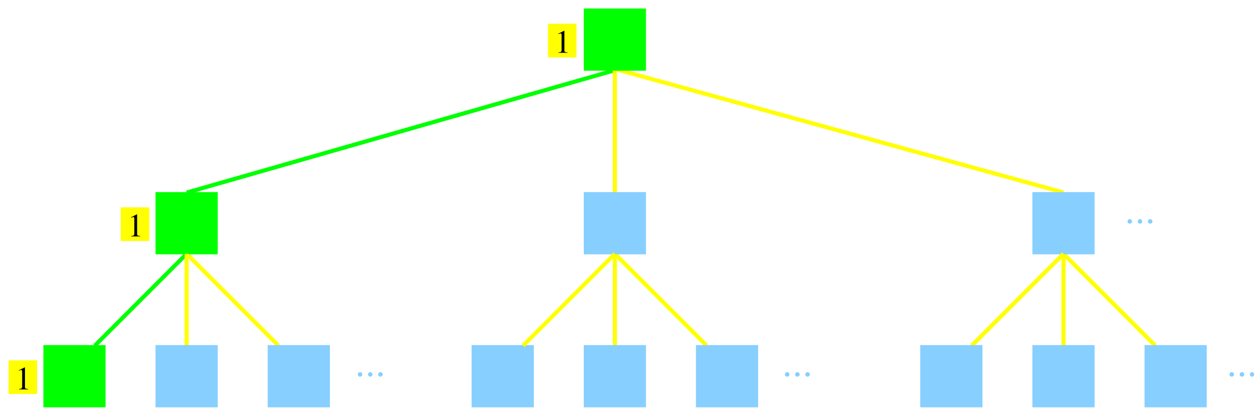
- A tourist arrives in the city for a culinary vacation
  - On the first evening, he enters the root restaurant and chooses a table
  - On the second evening, he goes to the restaurant identified by the first night's table and chooses another table
  - He repeats this process for L days

Nested Chinese restaurant process



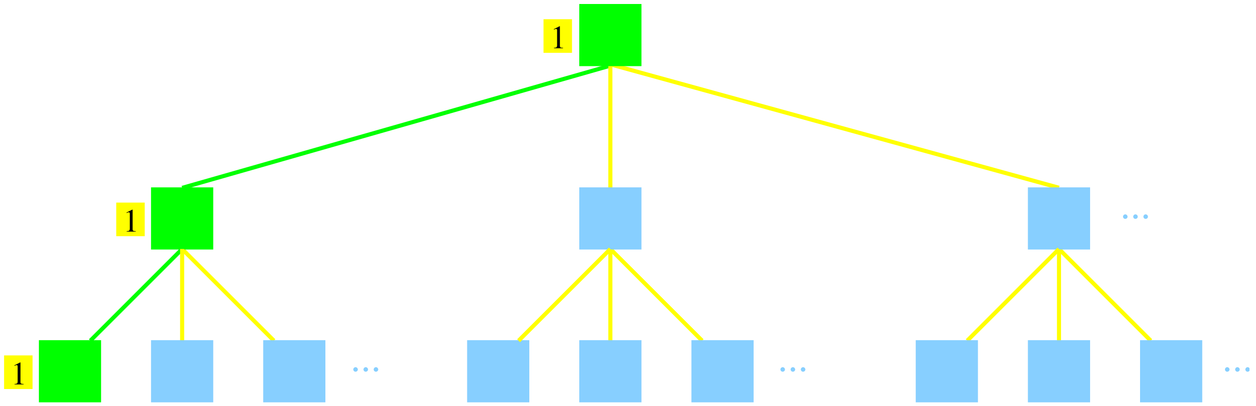
- A tourist arrives in the city for a culinary vacation
  - On the first evening, he enters the root restaurant and chooses a table
  - On the second evening, he goes to the restaurant identified by the first night's table and chooses another table
  - He repeats this process for L days

Nested Chinese restaurant process



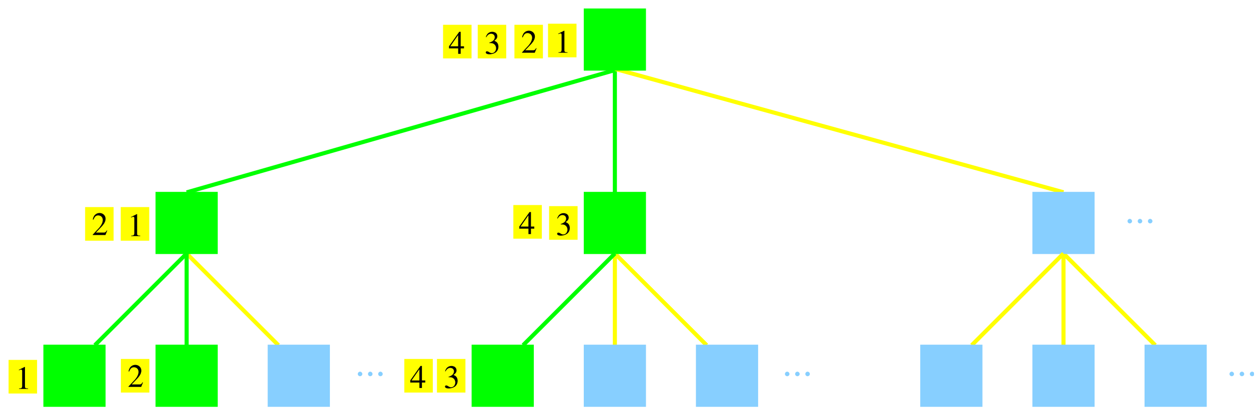
- A tourist arrives in the city for a culinary vacation
  - On the first evening, he enters the root restaurant and chooses a table
  - On the second evening, he goes to the restaurant identified by the first night's table and chooses another table
  - He repeats this process for L days

Nested Chinese restaurant process



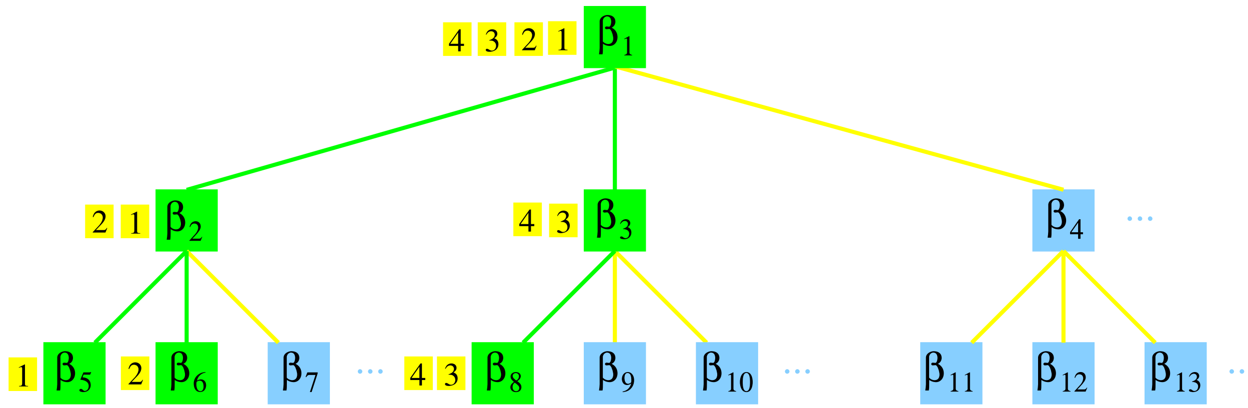
- The L chosen restaurants constitute a path from the root to a restaurant at the Lth level of the infinite tree

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 M tourists take L-day vacations, the collection of paths describe a particular L-level subtree of the infinite tree

Nested Chinese restaurant process

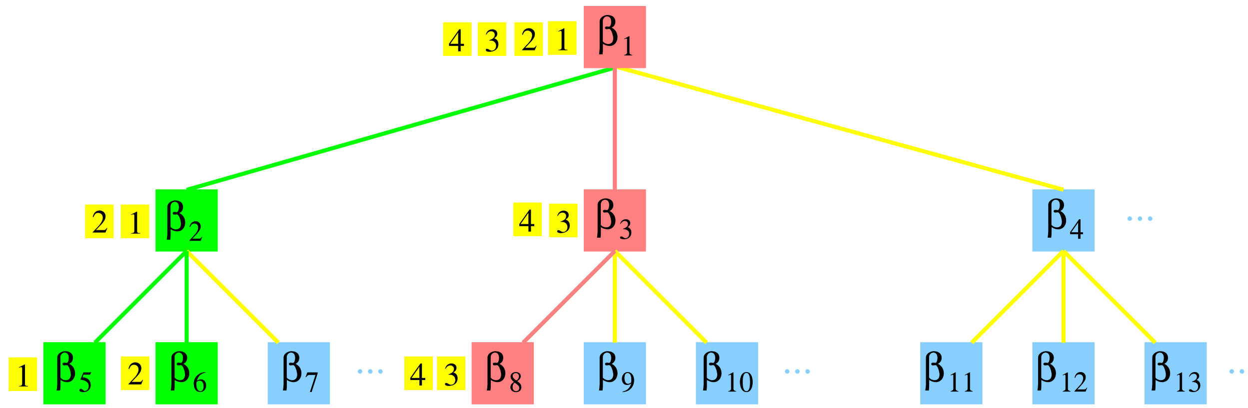


- The L chosen restaurants constitute a path from the root to a restaurant at the Lth level of the infinite tree

- After M tourists take L-day vacations, the collection of paths describe a particular L-level subtree of the infinite tree

- Assigning each restaurant to a parameter, we can use each tour as a topic path for a document

Nested Chinese restaurant process

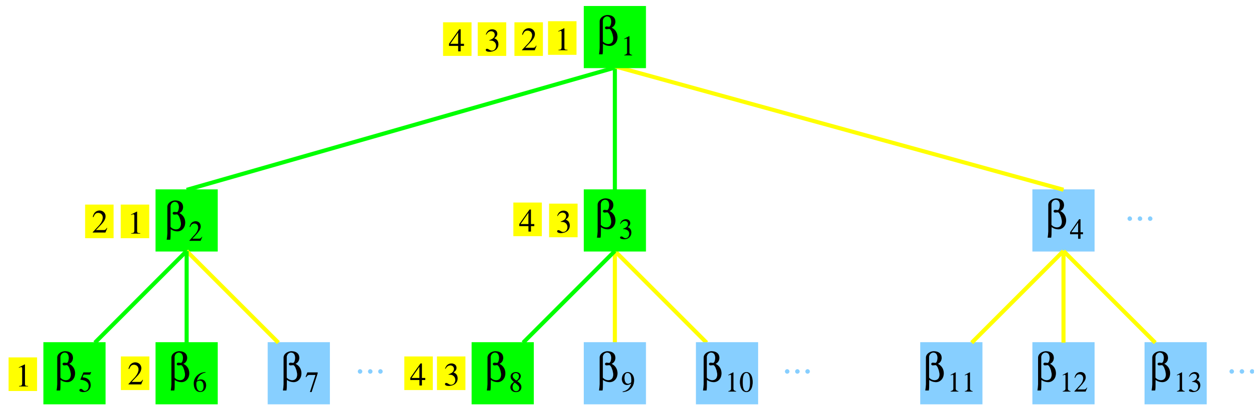


- The L chosen restaurants constitute a path from the root to a restaurant at the Lth level of the infinite tree

- After M tourists take L-day vacations, the collection of paths describe a particular L-level subtree of the infinite tree

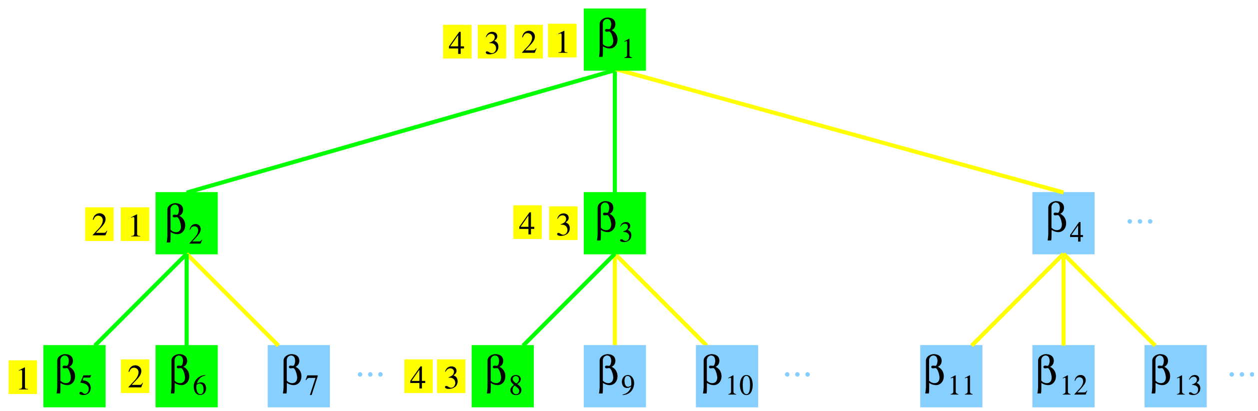
- Assigning each restaurant to a parameter, we can use each tour as a topic path for a document

Hierarchical LDA



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

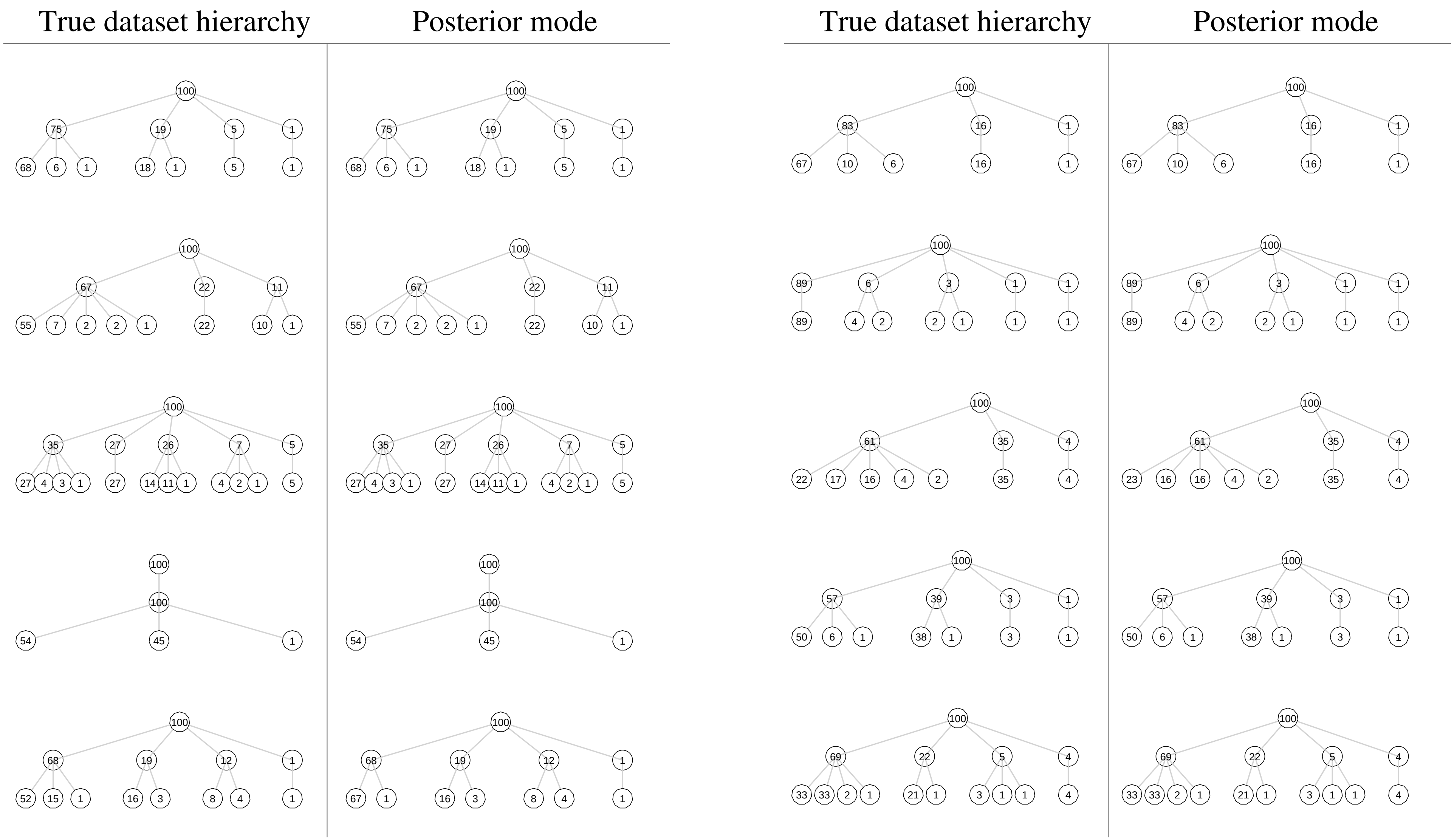
Hierarchical LDA



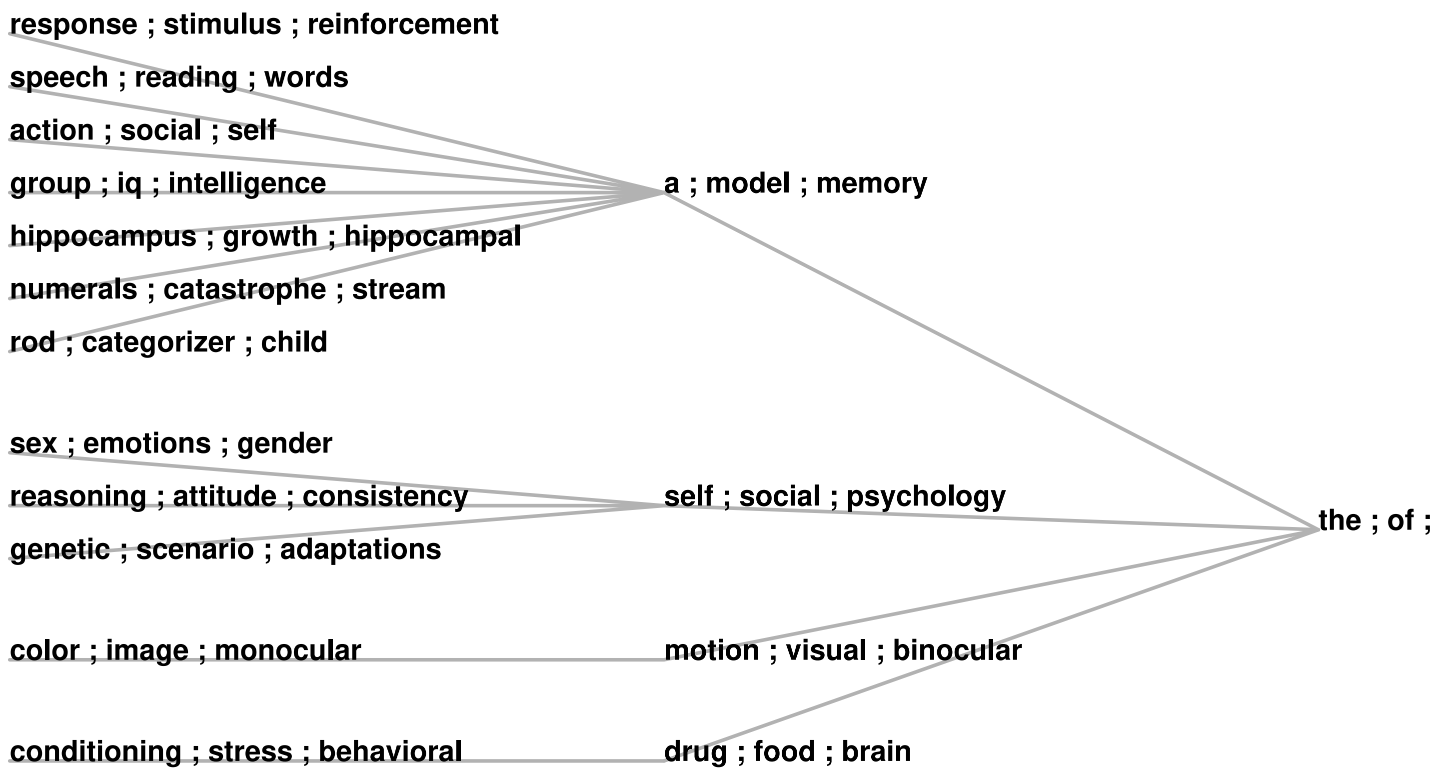
- Given a document collection, posterior is a distribution of
  - The structure of the hierarchy
  - Assignment of documents to paths, words to levels
  - Topics which populate the hierarchy

- Allows new documents to fill unoccupied parts of the tree

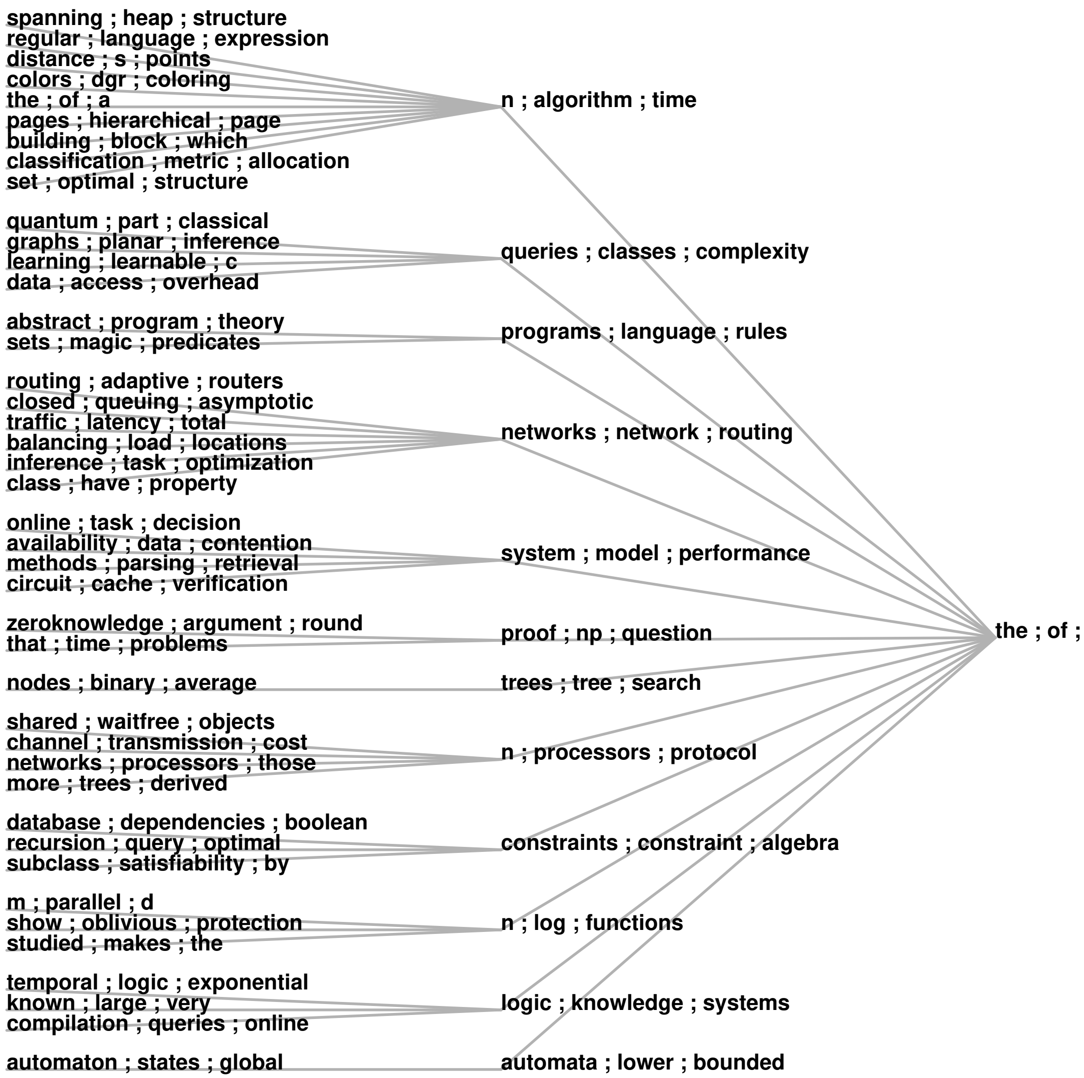
Estimating the correct hierarchy



Topic hierarchy from *Psychology*



Topic hierarchy from *JACM*

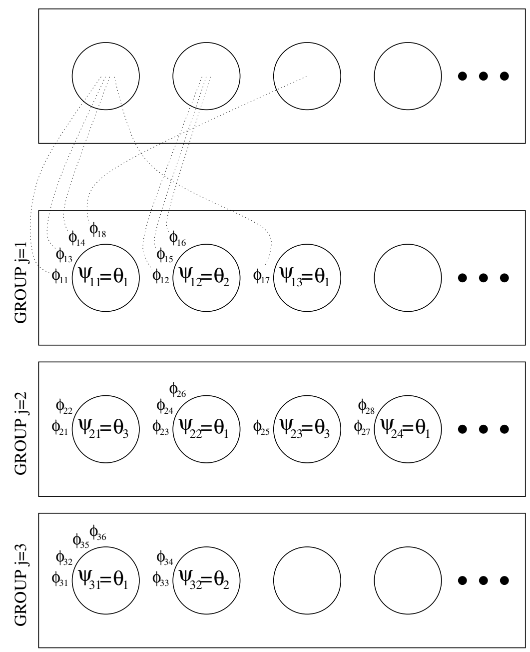


Inference for HDP mixtures

- Gibbs sampling
  - based on the Chinese restaurant process
  - based on the stick-breaking representation
- Variational inference
  - based on the stick-breaking representation

Chinese restaurant franchise (CRF)

- First integrate out the G_i, then integrate out G_0



- Set of *restaurants* with an unbounded number of *tables* in each restaurant
- One *menu* with an unbounded number of *dishes* on the 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

Gibbs sampler based on the CRF

$$p(t_{ji} = t \,|\, \boldsymbol{t} \backslash t_{ji}, \boldsymbol{k}, \boldsymbol{\theta}, \mathbf{x}) \quad \propto \quad \begin{cases} \alpha_0 f(x_{ji}|\theta_{k_{jt}}) & \text{if } t = t^{\text{new}} \\ n_{jt}^{-i} f(x_{ji}|\theta_{k_{jt}}) & \text{if } t \text{ currently used} \end{cases}$$

$$p(k_{jt} = k \,|\, \boldsymbol{t}, \boldsymbol{k} \backslash k_{jt}, \boldsymbol{\theta}, \mathbf{x}) \quad \propto \quad \begin{cases} \gamma \prod_{i:t_{ji}=t} f(x_{ji}|\theta_k) & \text{if } k = k^{\text{new}} \\ m_k^{-t} \prod_{i:t_{ji}=t} f(x_{ji}|\theta_k) & \text{if } k \text{ currently used} \end{cases}$$

$$p(\theta_k \,|\, \boldsymbol{t}, \boldsymbol{k}, \boldsymbol{\theta} \backslash \theta_k, \mathbf{x}) \quad \propto \quad h(\theta_k) \prod_{ji:k_{jt_{ji}}=k} f(x_{ji}|\theta_k)$$

Mean-field inference

- Roughly, an averaged version of a Gibbs sampler
  - deterministic algorithms (no coin-tossing)

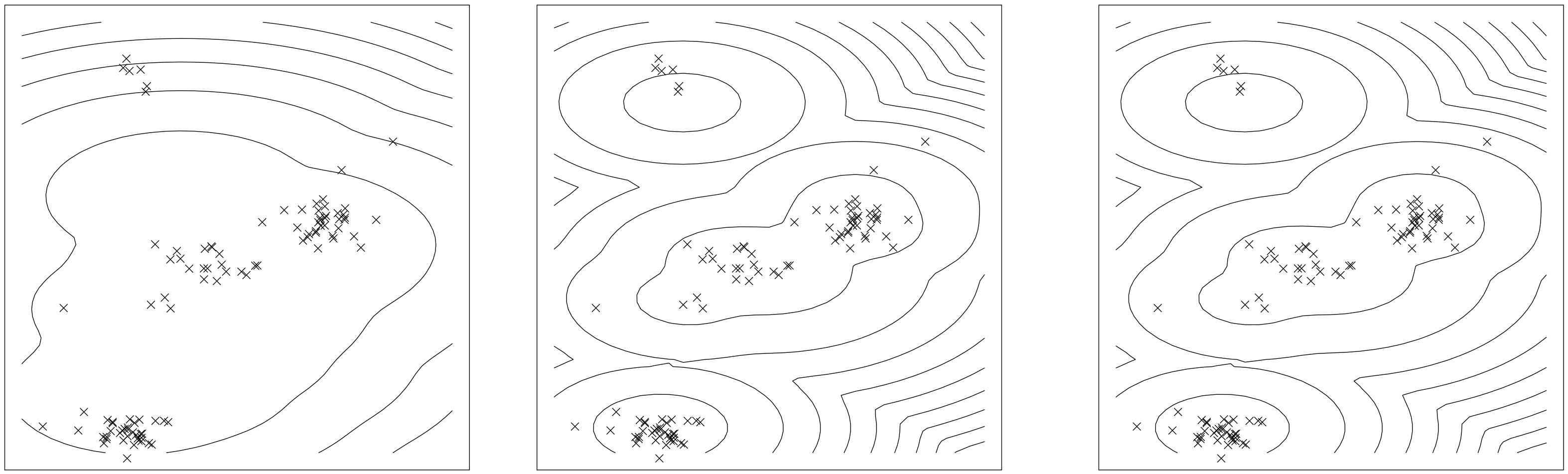
- Variational inference equations for a conjugate DP mixture in the exponential family

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

where

$$S = E[\log V_i \mid \gamma_i] + E[\eta_i \mid \tau_i]^T X_n - E[a(\eta_i) \mid \tau_i] - \textstyle\sum_{j=i+1}^{K} E[\log(1 - V_j) \mid \gamma_j].$$

Example: DP-Gaussian mixture



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

Figure 3: 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

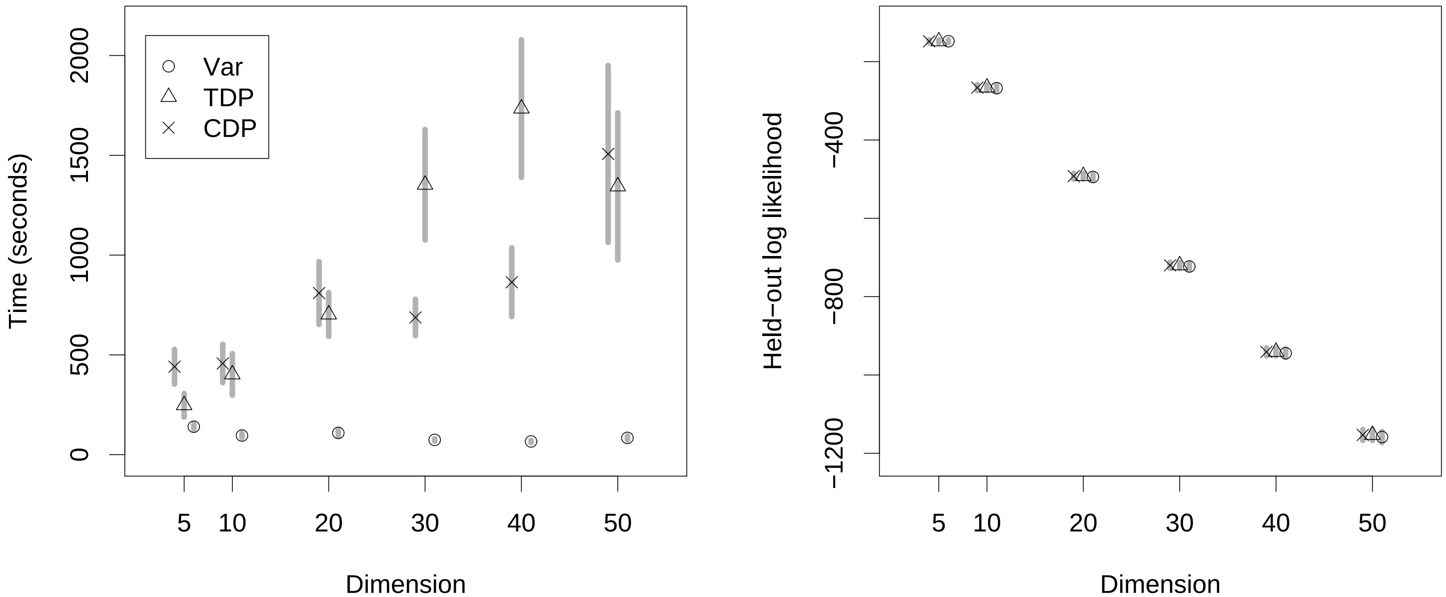


Figure 4: (Left) Convergence time per dimension across ten datasets for variational inference (Var), the TDP Gibbs sampler (TDP), and the collapsed Gibbs sampler (CDP). Grey bars are standard error. (Right) Average held-out log likelihood for the corresponding predictive distributions.

Conclusions

- Hierarchical Dirichlet processes provide an interesting, "combinatorial" notion of the sharing of statistical strength
- Mixture models are widely used in applications in speech, language, and bioinformatics
  - many opportunities for hierarchical Dirichlet processes
- Tip of the iceberg for applied nonparametric Bayesian modeling
- For details: **http://www.cs.berkeley.edu/~jordan**