Dirichlet Processes, Chinese Restaurant Processes and All That

Michael I. Jordan

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

http://www.cs.berkeley.edu/∼jordan

Acknowledgments: David Blei, Yee Whye Teh

Some Well-Worn but Still-Very-Useful Distinctions

	Frequentist	Bayesian
Parametric	I	IV
Semiparametric	II	V
Nonparametric	III	VI

I: Logistic regression, ANOVA, Fisher discriminant analysis, ARMA, etc
II: Independent component analysis, Cox model, nonmetric MDS, etc
III: Nearest neighbor, kernel methods, bootstrap, decision trees, neural nets, etc
IV: Conjugate analysis, hierarchical models, graphical models, etc
V: ?
VI: Gaussian processes, ?

Some Well-Worn but Still-Very-Useful Distinctions

	Frequentist	Bayesian
Parametric	I	IV
Semiparametric	II	V
Nonparametric	III	VI

- What do we mean by “parameters” anyway?
  - note in particular that “nonparametric” doesn’t mean “no parameters”
  - it means (very roughly) that the number of parameters grows with the number of data points

Decision-Theoretic Perspective on Parameters

- Define a loss function, with data X and “parameter” θ:

$$\begin{tikzcd}
 & l(X,\theta) \arrow[dl, "\text{frequentist expectation}"'] \arrow[dr, "\text{Bayesian expectation}"] & \\
R(\theta) = \mathbb{E}_\theta l(X,\theta) & & \rho(X) = \mathbb{E}[l(X,\theta) \mid X]
\end{tikzcd}$$

- For a full appreciation of the relationships between methods, it’s critical that we let θ range over something more than classical vector spaces
  - e.g., in this talk, θ will range over a space of measures
  - so we’d better be prepared to integrate over general spaces if we’re going to be nonparametric Bayesian

Decision-Theoretic Perspective on Parameters

- Define a loss function, with data X and "parameter" θ:

$$\begin{tikzcd}
 & l(X,\theta) \arrow[dl, "\text{frequentist expectation}"'] \arrow[dr, "\text{Bayesian expectation}"] & \\
R(\theta) = \mathbb{E}_\theta l(X,\theta) & & \rho(X) = \mathbb{E}[l(X,\theta) \,|\, X]
\end{tikzcd}$$

- For a full appreciation of the relationships between methods, it's critical that we let θ range over something more than classical vector spaces
  - e.g., in this talk, θ will range over a space of measures
  - so we'd better be prepared to integrate over general spaces if we're going to be nonparametric Bayesian

Decision-Theoretic Perspective on Parameters

- Define a loss function, with data X and "parameter" θ:

$$\begin{tikzcd}
& l(X,\theta) \arrow[dl, "\text{frequentist expectation}"'] \arrow[dr, "\text{Bayesian expectation}"] & \\
R(\theta) = \mathbb{E}_\theta l(X,\theta) & & \rho(X) = \mathbb{E}[l(X,\theta) \mid X]
\end{tikzcd}$$

- For a full appreciation of the relationships between methods, it's critical that we let θ range over something more than classical vector spaces
  - e.g., in this talk, θ will range over a space of measures
  - so we'd better be prepared to integrate over general spaces if we're going to be nonparametric Bayesian

The De Finetti Perspective on Parameters

- For Bayesians, the De Finetti theorem is a compelling motivation for both "parameters" and priors on parameters
  - but everyone should know what the De Finetti theorem is, not just Bayesians...

- What is the De Finetti theorem?

- It's the "bag-of-words theorem"

The De Finetti Perspective on Parameters (cont.)

- Suppose that we agree that if our data are reordered, it doesn't matter
- This is generally **not** an assertion of "independent and identically distributed"
- Rather, it is an assertion of "exchangeability"
  - *exchangeability*: the joint probability distribution underlying the data is invariant to permutation

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*:

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

for some random variable θ.

The De Finetti Perspective on Parameters (cont.)

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*:

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

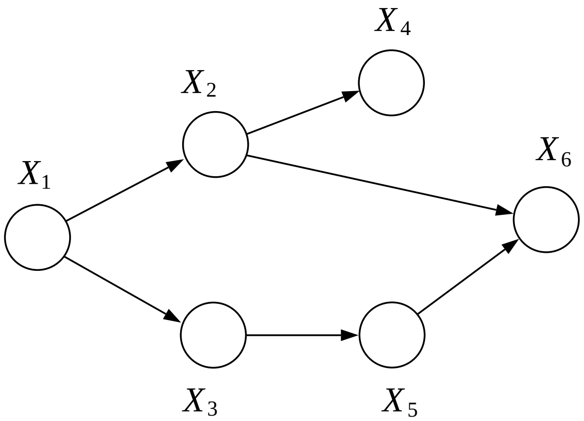
for some random variable θ.

- The theorem wouldn't be true if we limited ourselves to parameters θ ranging over Euclidean vector spaces

- In particular, we need to allow θ to range over measures, in which case $P(\theta)$ is a measure on measures
  - the Dirichlet process is an example of a measure on measures...

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:

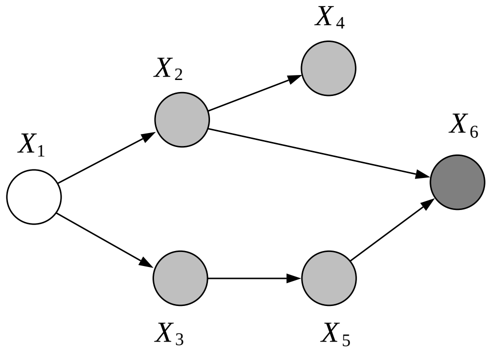


- The joint distribution on $(X_1, X_2, \ldots, X_N)$ factorizes according to the "parent-of" relation defined by the edges $\mathcal{E}$:

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

Inference—Computing Conditional Probabilities

- Conditioning



- Marginalization:

$$p(x_1, x_6) = \int_{x_2} \int_{x_3} \int_{x_4} \int_{x_5} p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_3)p(x_6|x_2, x_5)$$

- Conditional probabilities:

$$p(x_1 \,|\, x_6) = \frac{p(x_1, x_6)}{p(x_6)}$$

Inference Algorithms

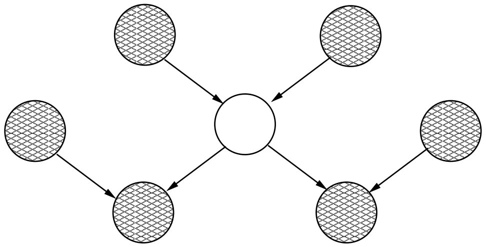
- *Exact algorithms*
  - elimination algorithm
  - sum-product algorithm
  - junction tree algorithm

- *Sampling algorithms*
  - importance sampling
  - Markov chain Monte Carlo (MCMC)

- *Variational algorithms*
  - mean field methods (e.g., Jordan et al., 1999; Opper & Saad, 2001)
  - sum-product algorithm and variations (e.g., Yedidia et al., 2001; Minka, 2001; McEliece & Yildirim, 2002)
  - semidefinite relaxations (Wainwright & Jordan, 2003)

Gibbs Sampling

- A widely-used Markov chain Monte Carlo (MCMC) method
- Given a set of variables X_V, we set up a Markov chain as follows:
  - initialize the X_i to arbitrary values
  - choose i randomly
  - sample from $p(x_i|x_{V\setminus i})$
  - iterate
- It is easy to prove that this scheme has $p(x_V)$ as its equilibrium distribution
- Gibbs sampling is often readily implemented in graphical models



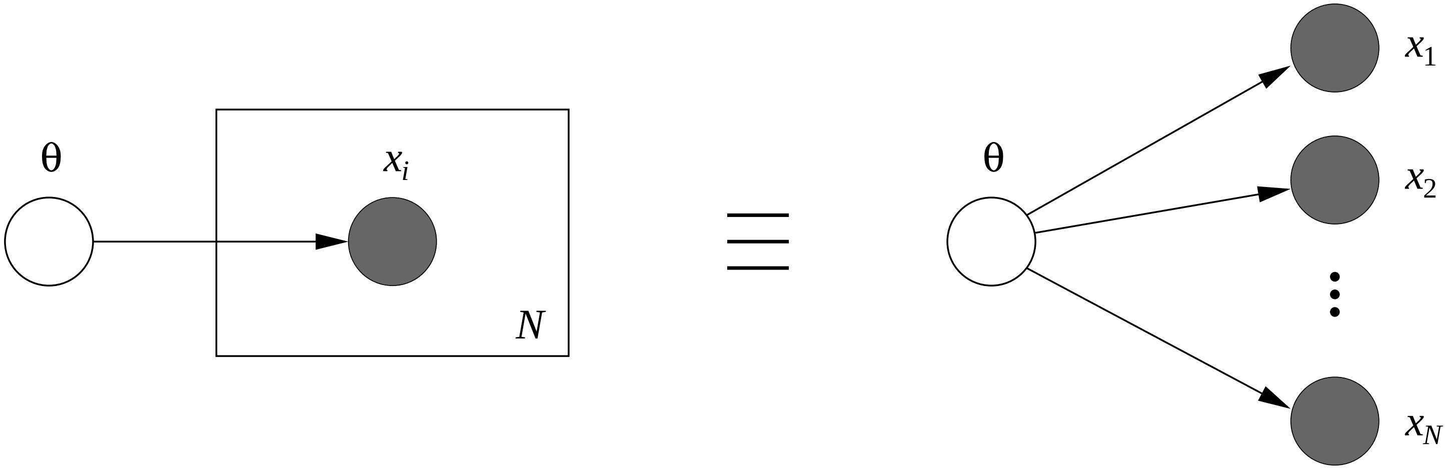
Variational Algorithms

- Three steps:
  - convert the inference problem into an optimization problem
  - relax the optimization problem into a simplified optimization problem
  - solve the relaxation

- Many variations
  - *mean field algorithms* (pretend the law of large numbers holds)
  - *sum-product algorithm* (pretend the graph is a tree)
  - *semidefinite relaxations* (pretend that second moments suffice)

Plates

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



- Note that this is a graphical representation of the De Finetti theorem

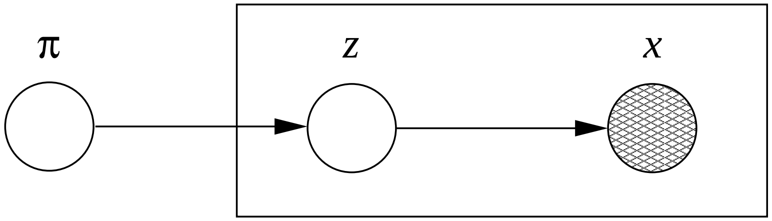
$$p(x_1, x_2, \ldots, x_N) = \int p(\theta) \left(\prod_{i=1}^{N} p(x_i \,|\, \theta) \right) d\theta$$

Outline

- Mixture models—how to choose K?
- Chinese restaurant process
- Latent Dirichlet allocation
- Hierarchical models
- Dirichlet processes
- Hierarchical Dirichlet processes
- Empirical Bayes for the Dirichlet process

Mixture Models

- The standard mixture modeling approach to clustering



- The latent multinomial variable Z represents the clusters:

$$\begin{aligned} p(x|\theta) &= \sum_k p(Z=k)p(x|Z=k,\theta) \\ &= \sum_k \pi_k f_k(x|\theta_k), \end{aligned}$$

where π_k are the *mixing proportions* and $f_k(x|\theta_k)$ are the *mixture components* (e.g., Gaussians, where $\theta_k = (\mu_k, \Sigma_k)$)

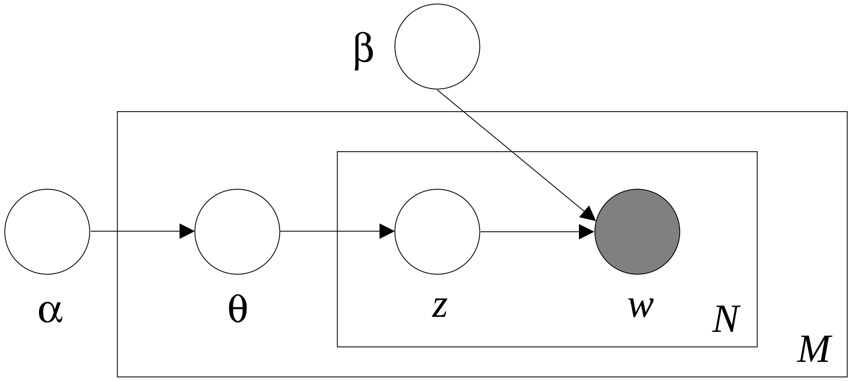
Model Selection for Mixture Models

$$p(x|\theta) = \sum_{k=1}^{K} \pi_k f_k(x|\theta_k),$$

- How to choose K, the number of mixture components?
  - similarly, for the HMM, how to choose the number of states?
- Various generic model selection methods can be used: e.g., cross-validation, bootstrap, AIC, BIC, DIC, MDL, Laplace, bridge sampling, reversible jump MCMC, etc
  - the Chinese restaurant process provides another alternative

Latent Dirichlet Allocation (LDA) Model

(Blei, Ng, & Jordan, 2003)



- *Random variables*:
  - A word is represented as a *multinomial* random variable w
  - A topic is represented as a *multinomial* random variable z
  - A document is represented as a *Dirichlet* random variable θ

- *Plates*:
  - *repeated* sampling of Dirichlet document variable within corpus
  - *repeated* sampling of multinomial topic variable within documents

Dirichlet Distribution

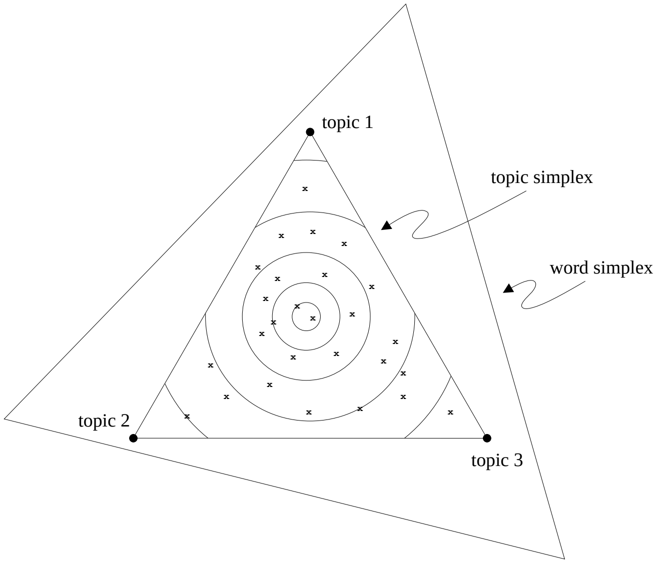
- Let $\theta = (\theta_1, \theta_2, \ldots, \theta_m)$ be a point in the $(m-1)$-simplex
  - i.e., $0 < \theta_i < 1$ and $\sum_{i=1}^m \theta_i = 1$
- Let $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m)$ be a set of parameters, where $\alpha_i > 0$
- The Dirichlet distribution:

$$p(\theta \,|\, \alpha) = \frac{\Gamma(\sum_{i=1}^m \alpha_i)}{\prod_{i=1}^m \Gamma(\alpha_i)} \, \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \cdots \theta_m^{\alpha_m - 1}$$

is an exponential family distribution on the simplex, with $\mathbb{E}(\theta_i) = \frac{\alpha_i}{\sum_{i=1}^m \alpha_i}$

The Topic Simplex

- Each corner of the simplex corresponds to a *topic*—a component of the vector z:

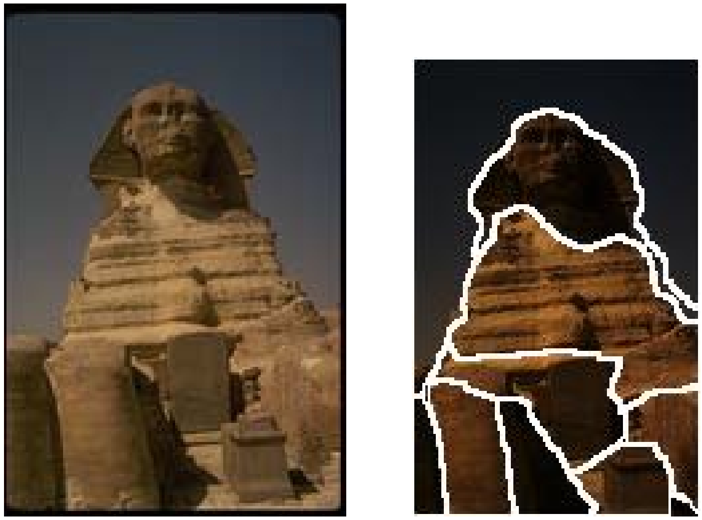


The topic simplex for $k = 3$.

- A document is modeled as a point in the simplex—a multinomial distribution over topics

- A corpus is modeled as a Dirichlet distribution on the simplex

Probabilistic Modeling of Documents/Images

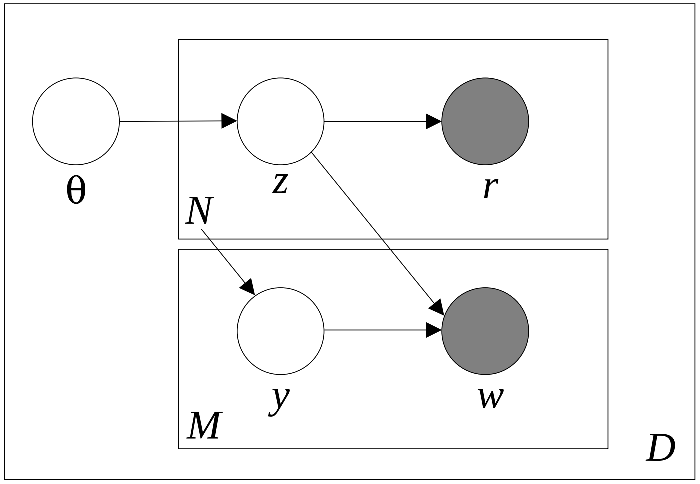


Sculpture, Statue, Stone

- Images are segmented into regions, and each region is represented as a 47-dimensional Gaussian vector
- Data are 11,000 images and their captions

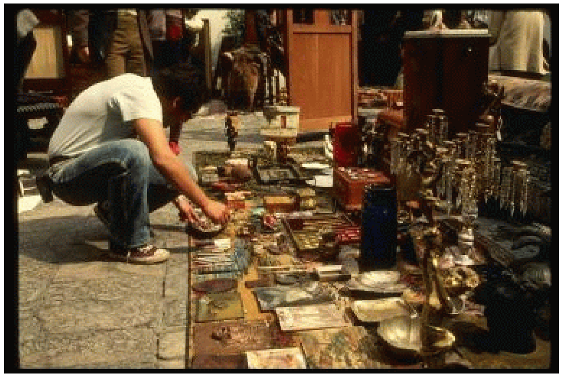
Correspondence LDA Model

(Blei & Jordan, 2003)



- Image-topics and word-topics
  - a word is represented as a *multinomial* random variable w
  - an image region is represented as a *Gaussian* random variable r
  - a word-topic is represented as a *multinomial* random variable z
  - an image-topic is represented as a *multinomial* random variable y

Automatic Annotation

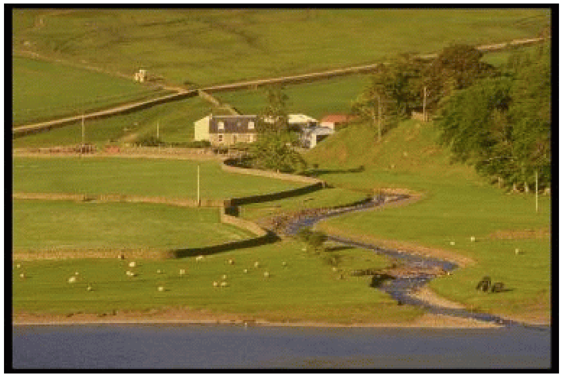


True caption
market people

Corr−LDA
people market pattern textile display

GM−LDA
people tree light sky water

GM−Mixture
people market street costume temple



True caption
scotland water

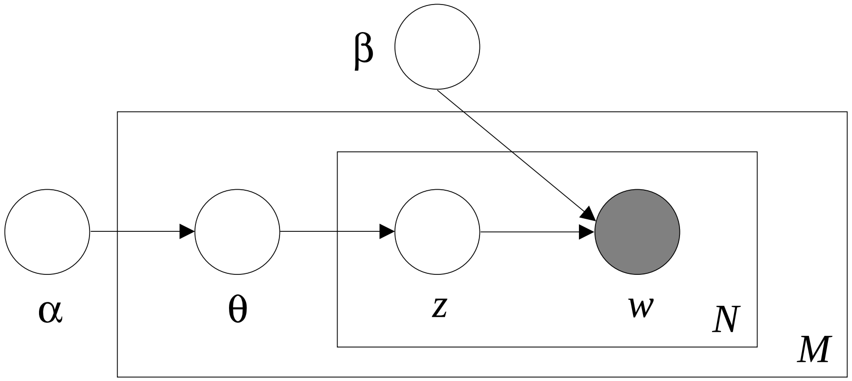
Corr−LDA
scotland water flowers hills tree

GM−LDA
tree water people mountain sky

GM−Mixture
water sky clouds sunset scotland

(Use the top five words from $p(w|\mathbf{r})$ to annotate an image.)

Model Selection Problem for LDA, CorrLDA, etc



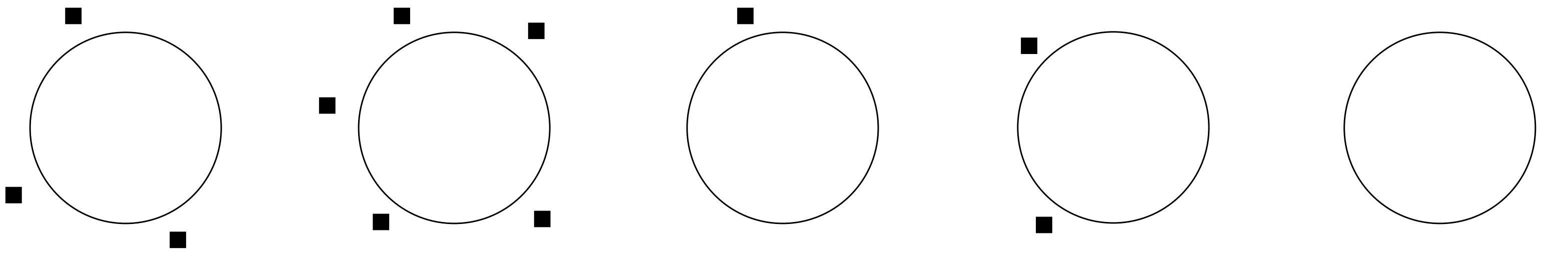
- How to choose the number of topics (the cardinality of z)?

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:

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

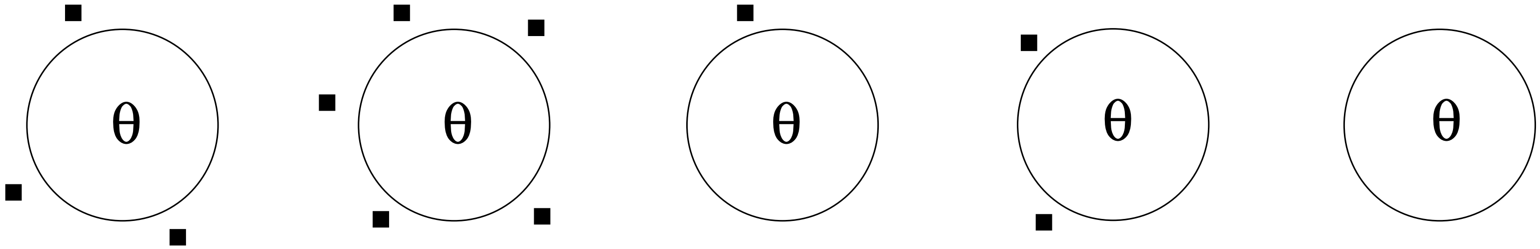
  where m_i is the number of customers currently at table i



- 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

Gibbs Sampling

- For each data point:
  - pretend that it is the last point (by exchangeability)
  - choose a table using the Chinese restaurant dynamics
- For each table:
  - resample the parameter vector at that table, conditioning on all of the data points sitting at the table
- This will converge to a posterior distribution on partitions and parameters

NIPS Data

- 1718 abstracts from NIPS 0 – 12

> 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.

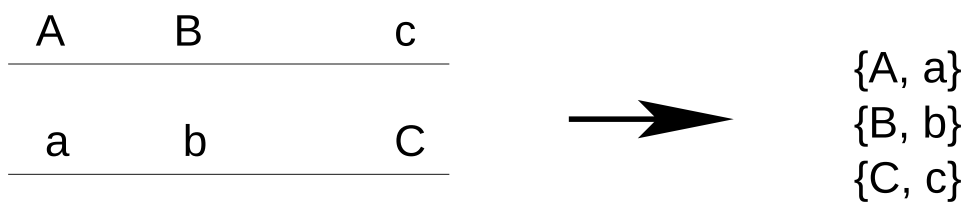
- Each cluster is a 4100-dimensional multinomial distribution

- The posterior mode was 26 clusters

“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

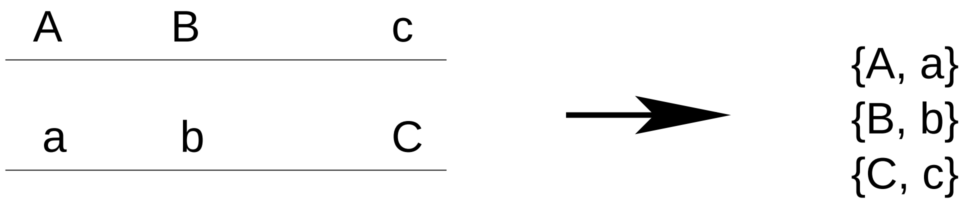
Haplotype Modeling

- 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.)



A B c

a b C

{A, a}
{B, b}
{C, c}

- 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

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

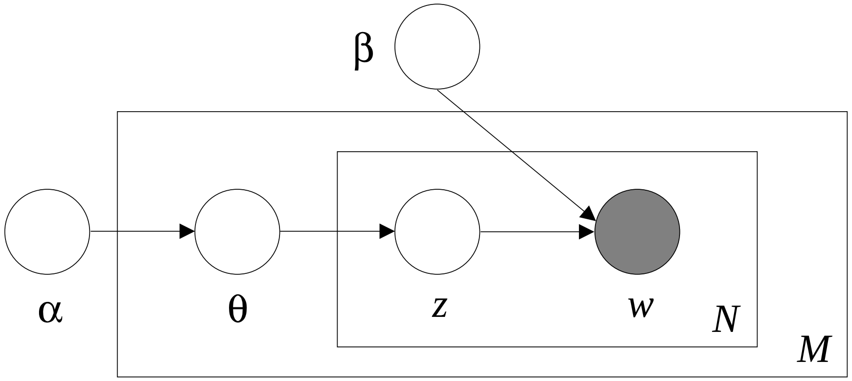
CRP-based Haplotype Model

(Xing, Sharan, & Jordan, 2004)

- Comparative performance of model 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

Model Selection Problem for LDA (cont.)



- Have we solved the problem of choosing the number of topics for LDA?
  - unfortunately, no
- We have *multiple* clustering problems—one per document
- We need to find a way to link multiple clustering problems

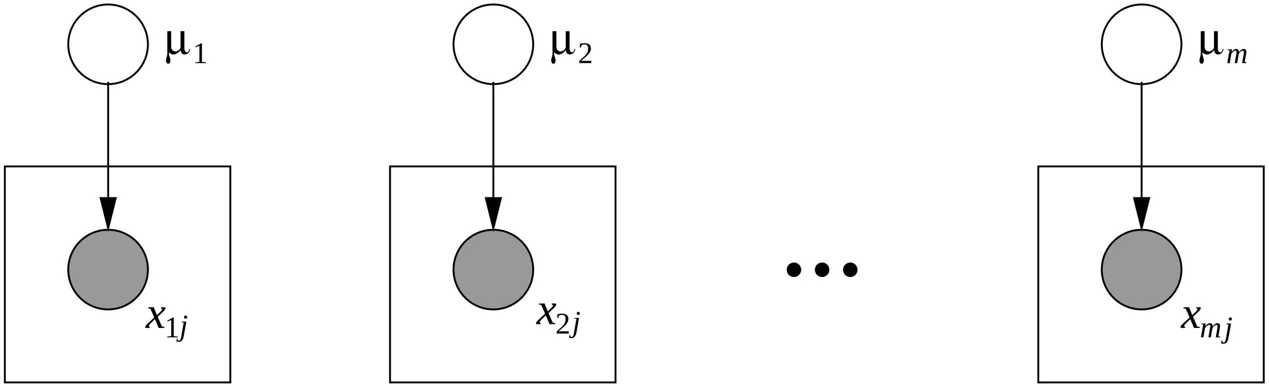
Haplotype Modeling (cont.)

- 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?

Multiple Inference Problems



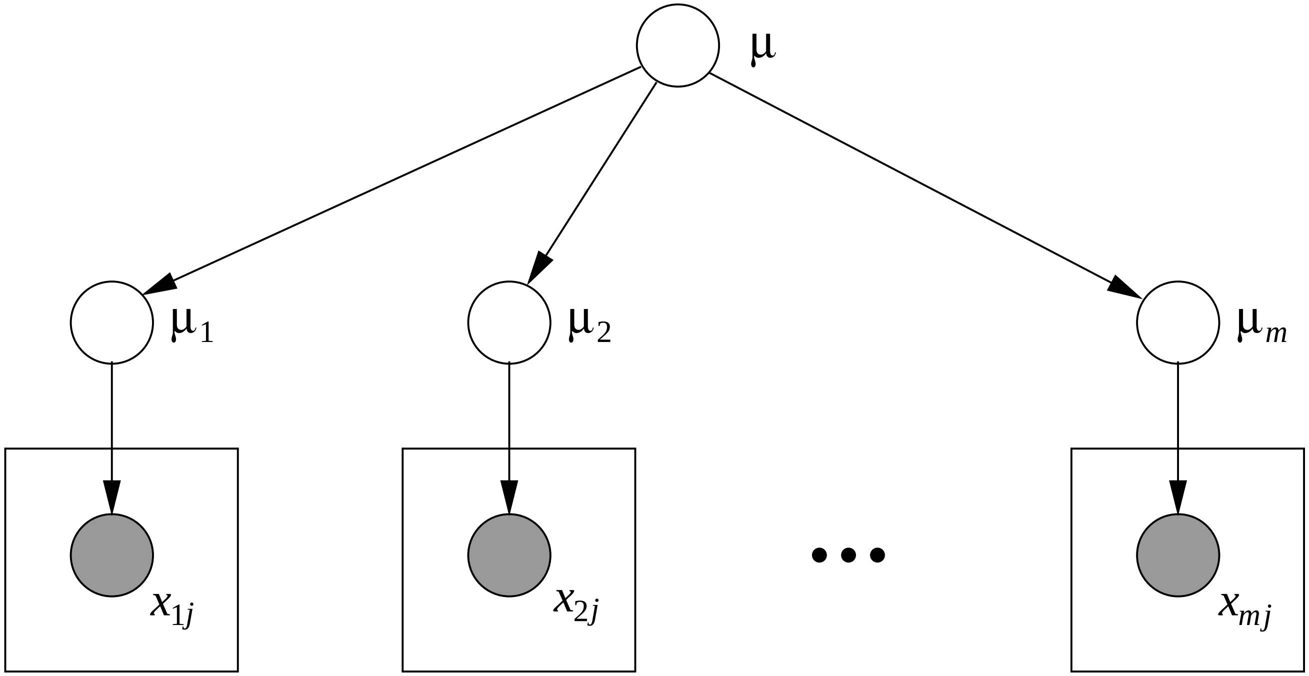
- Multiple Gaussian means (e.g., mean heights in various cities)

$$x_{ij} \sim N(\mu_i, \sigma_i^2)$$

- Maximum likelihood: $\hat{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$

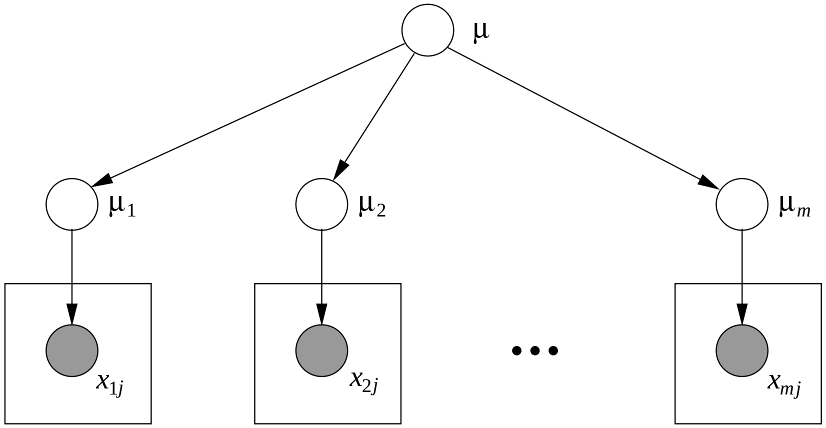
- Maximum likelihood often doesn't work very well
  - want to "share statistical strength" (i.e., "smooth")

Hierarchical Bayesian Modeling

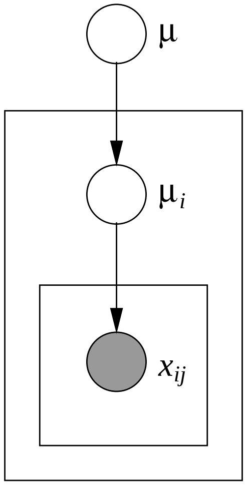


- Posterior mean is a shrinkage estimator—the posterior mean for each μ_k combines data from all of the cities, without simply lumping the data into one group

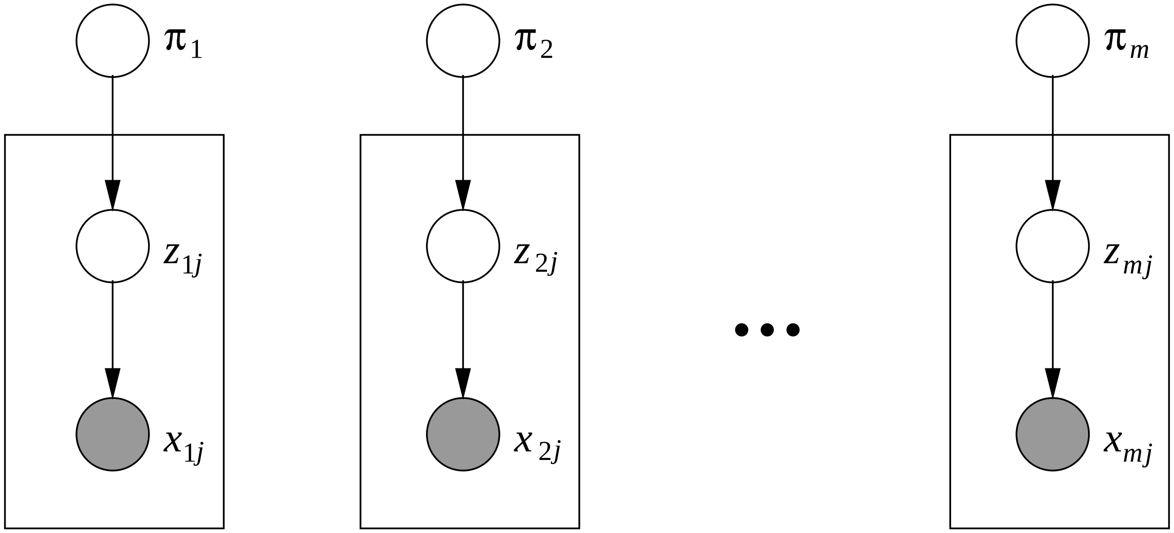
Hierarchical Modeling



- Recall the plate notation:



Multiple Clustering Problems



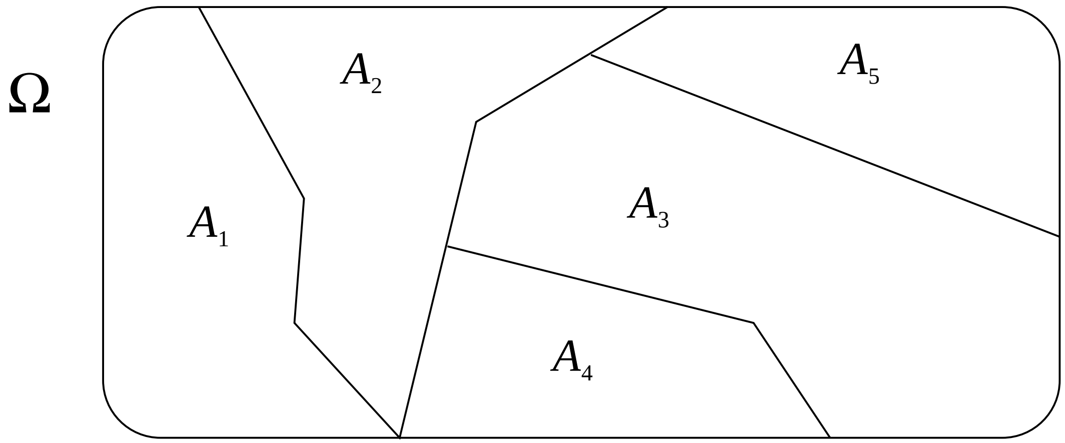
- E.g., LDA, the stratified haplotype problem
- For each i: $p(x_{ij}|\pi_i, \theta_i, k_i) = \sum_{l=1}^{k_i} p(z_{ij} = l|\pi_i)p(x_{ij}|z_{ij} = l, \theta_i)$
- What to share: π_i?, θ_i?, k_i?
- How to share?

Dirichlet Process

Definition 1. *Let $(\Omega, \mathcal{B})$ 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 $(\Omega, \mathcal{B})$ such that, for any finite partition $(A_1, \ldots, A_r)$ of Ω, 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.



The Posterior Dirichlet Process

- Suppose that we sample G from a Dirichlet process and then sample θ from G

- What is the posterior process?

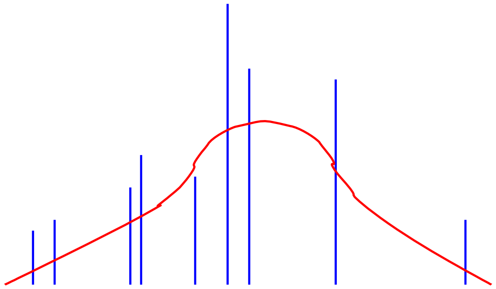
- For a fixed partition, we get a standard Dirichlet update (for the cell that contains θ the exponent increases by one; it stays the same for all other cells)
  - this is true for even the tiniest cell
  - suggests that the posterior process is Dirichlet with an atom at θ

- Continue doing this ad infinitum and we reveal G as a (random) infinite sum of atoms

Stick-Breaking Representation

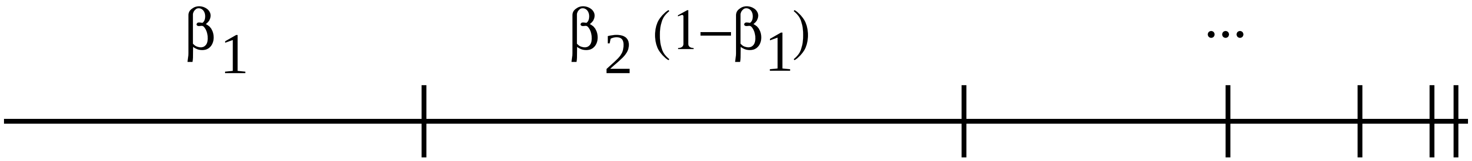
- Sethuraman (1994) gave an explicit representation for a draw from a Dirichlet process:

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

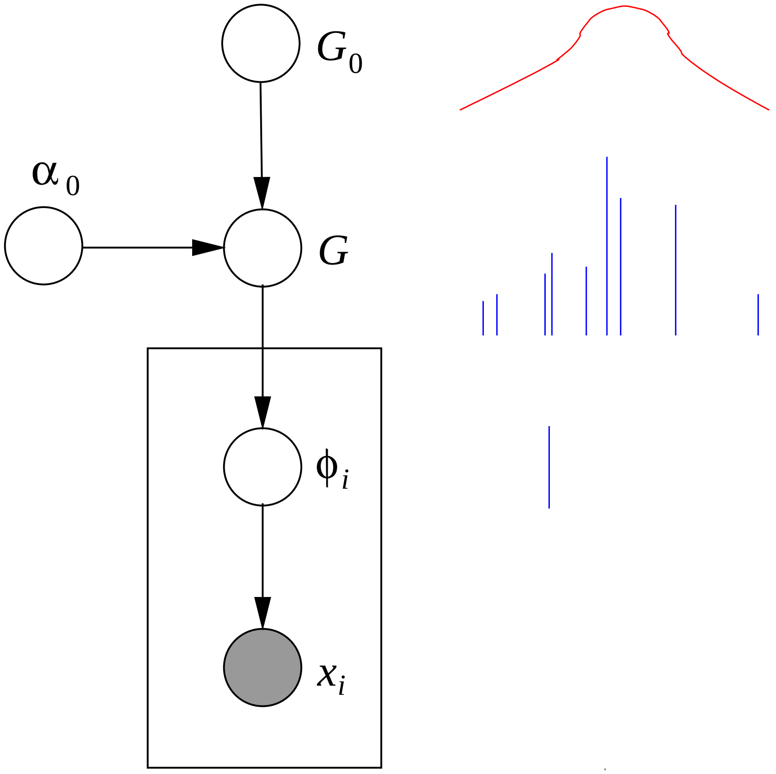


where

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

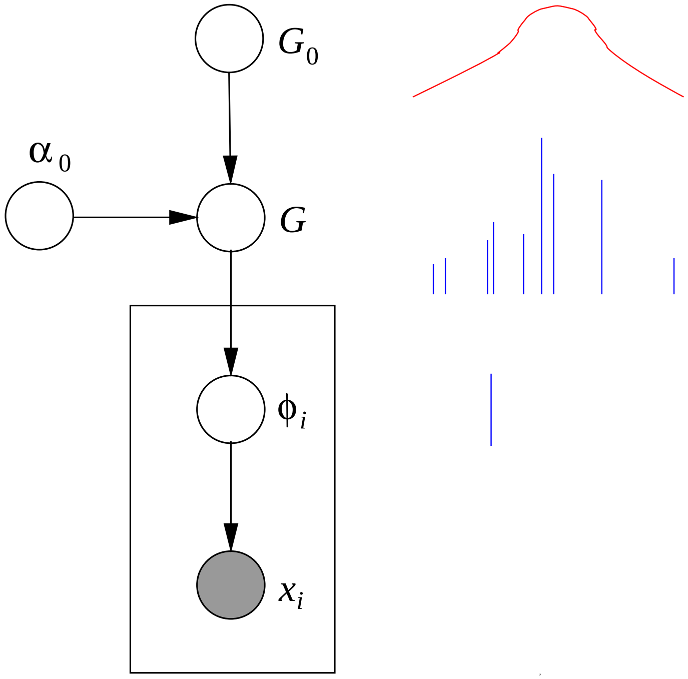


Dirichlet Process Mixture Models



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

Integrating Out G



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

$$\phi_i \,|\, \phi_1, \ldots, \phi_{i-1}, \alpha_0, G_0 \;\sim\; \sum_{l=1}^{i-1} \frac{1}{i-1+\alpha_0} \delta_{\phi_l} + \frac{\alpha_0}{i-1+\alpha_0} G_0$$

- 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

Variational Inference

(Blei & Jordan, 2005)

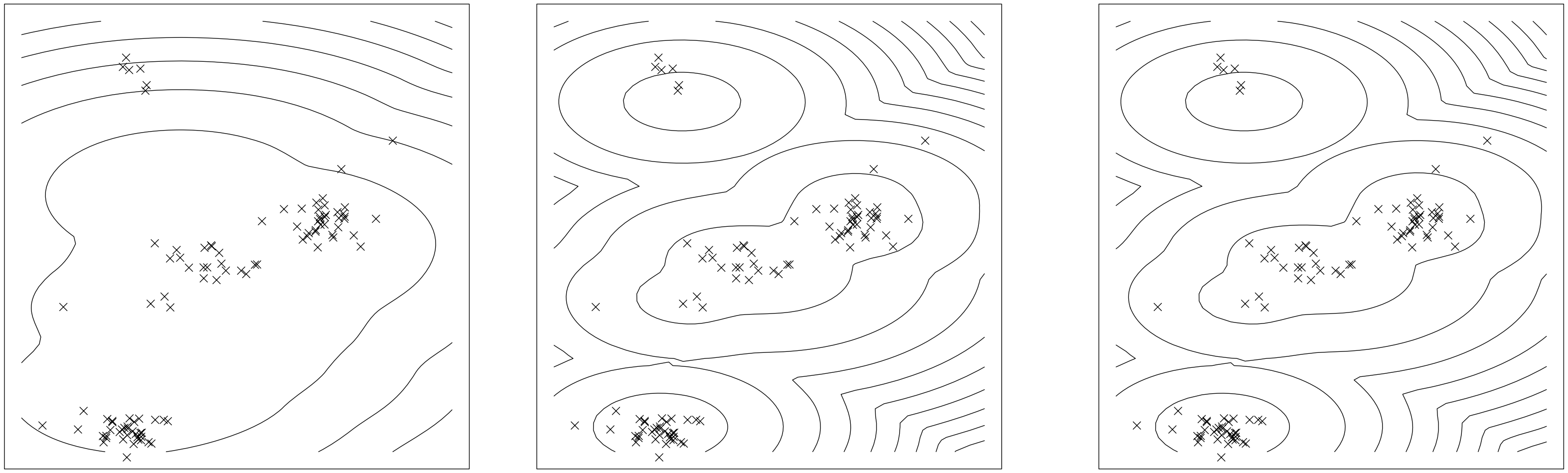
- The "Q distribution" is a truncated stick-breaking representation
- Variational inference equations for a conjugate DP mixture in the exponential family:

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

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 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

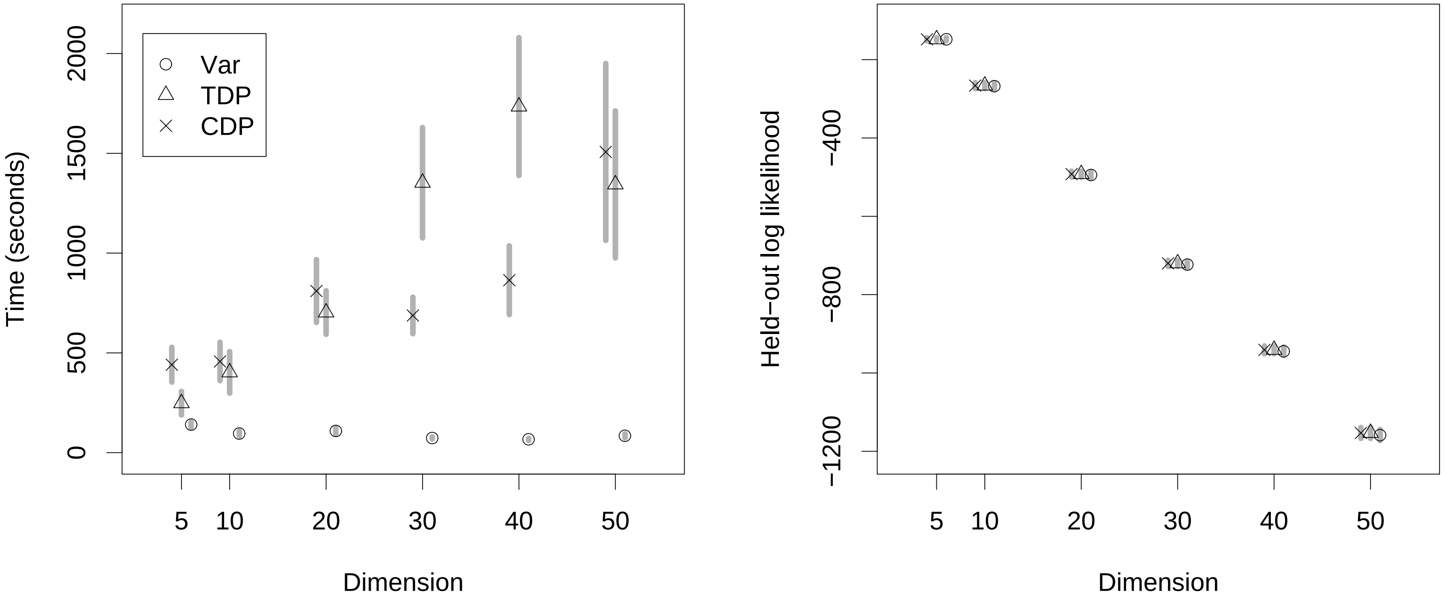
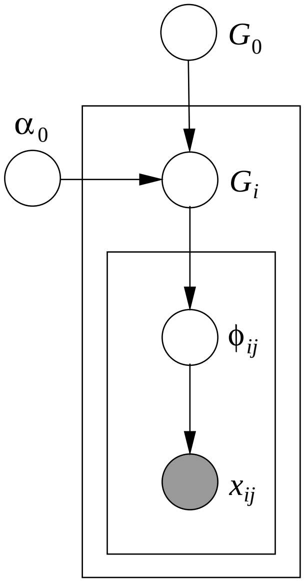


Figure 2. (Left) Convergence time per dimension across ten datasets for variational inference (Var), the TDP Gibbs sampler (TDP), and the collapsed Gibbs sampler (CDP). Grey bars are standard error. (Right) Average held-out log likelihood for the corresponding predictive distributions.

Multiple Clustering Problems (cont.)

- Idea: a Dirichlet process for each group, and share the underlying G_0:



- Problem: the atoms generated by these processes will be distinct w.p.1
  - 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

(Teh, Jordan, Beal & Blei, 2004)

- Let G_0 itself be distributed according to a DP:

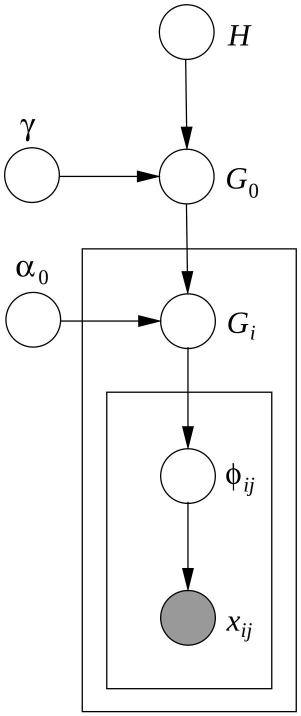
$$G_0 \,|\, \gamma, H \;\sim\; \mathrm{DP}(\gamma, H)$$

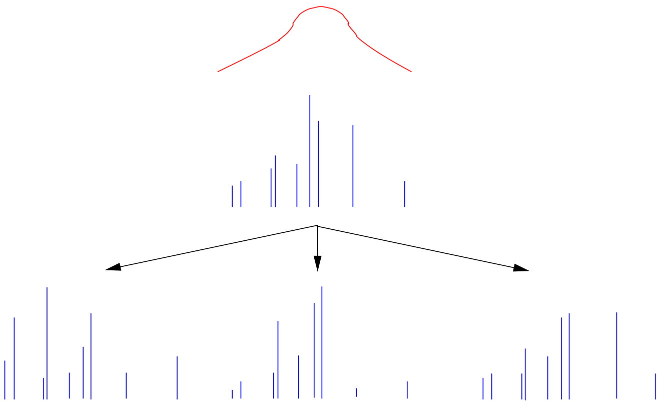
- Then

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

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

Hierarchical Dirichlet Process Mixture

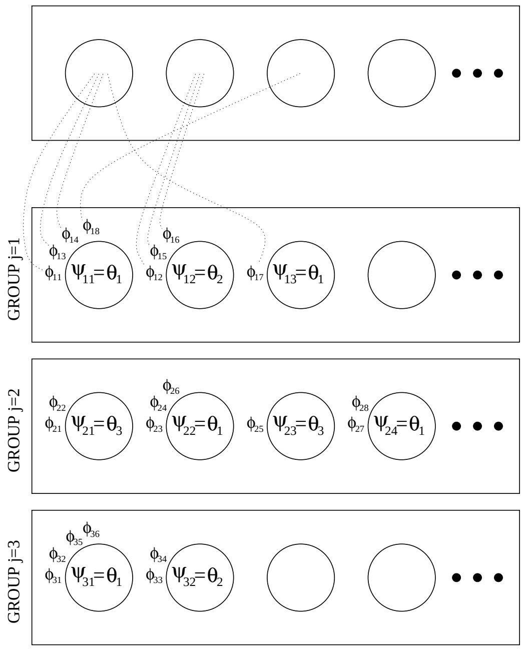




$$\begin{aligned} G_0 \,|\, \gamma, H &\sim \mathrm{DP}(\gamma, H) \\ G_i \,|\, \alpha, G_0 &\sim \mathrm{DP}(\alpha_0, G_0) \\ \phi_{ij} \,|\, G_i &\sim G_i \\ x_{ij} \,|\, \phi_{ij} &\sim F(x_{ij}, \phi_{ij}) \end{aligned}$$

Gibbs Sampling—the 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

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

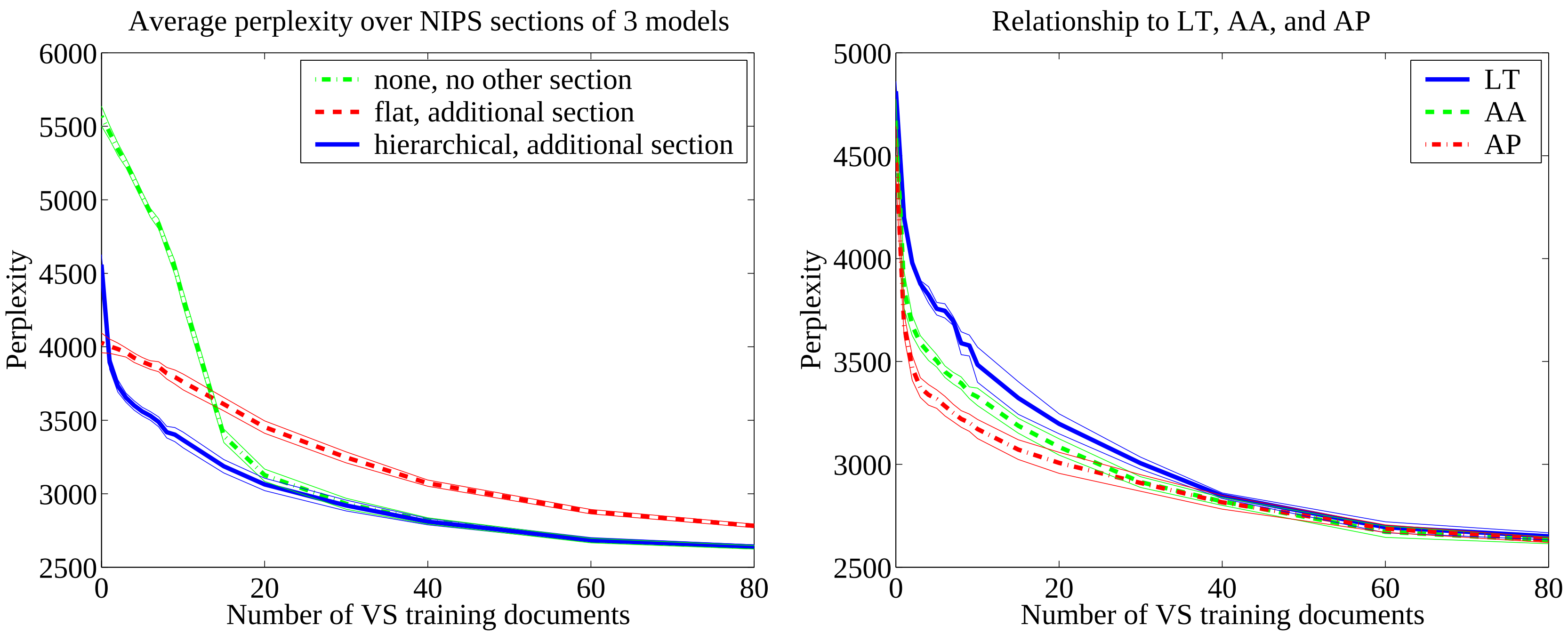


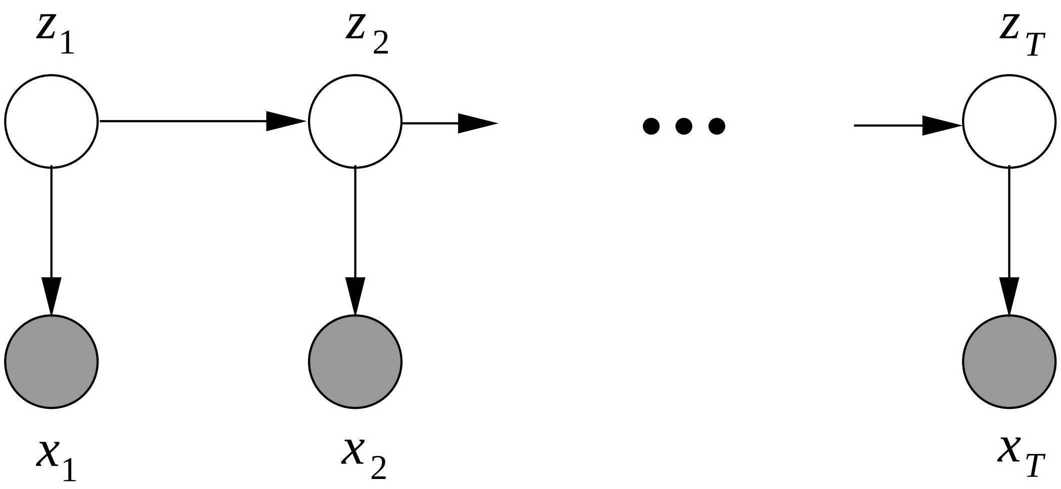
Figure 3. Left: Average perplexity of test VS documents given training documents from VS and another section for 3 different models. curves shown are averaged over the other sections and other 5 runs. Right: Average perplexity of test VS documents given LT, AA and AP documents respectively using M3,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

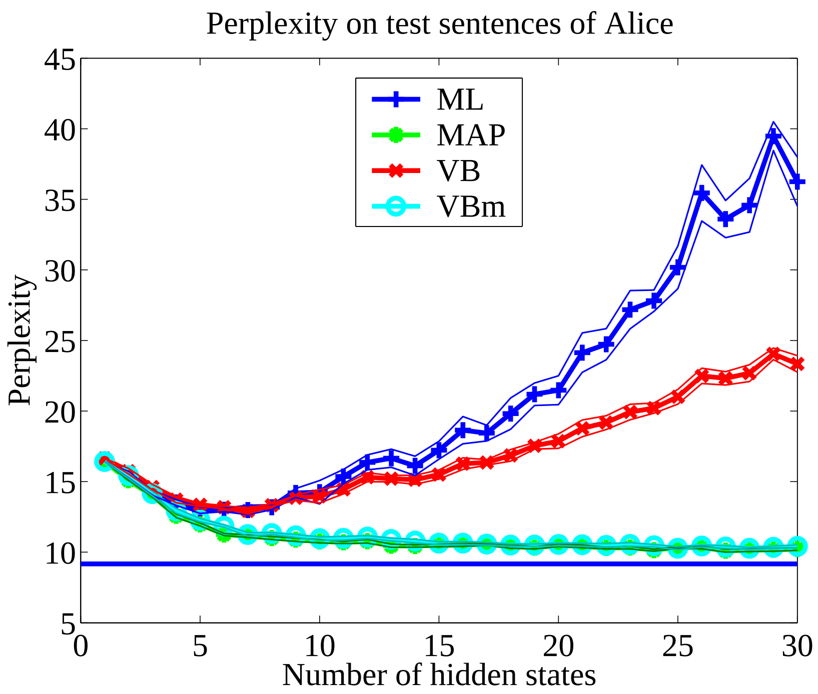
Hidden Markov Models



- The “infinite hidden Markov model”—an HMM with an unbounded number of states (Beal, Ghahramani, Rasmussen, 2002)

- 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

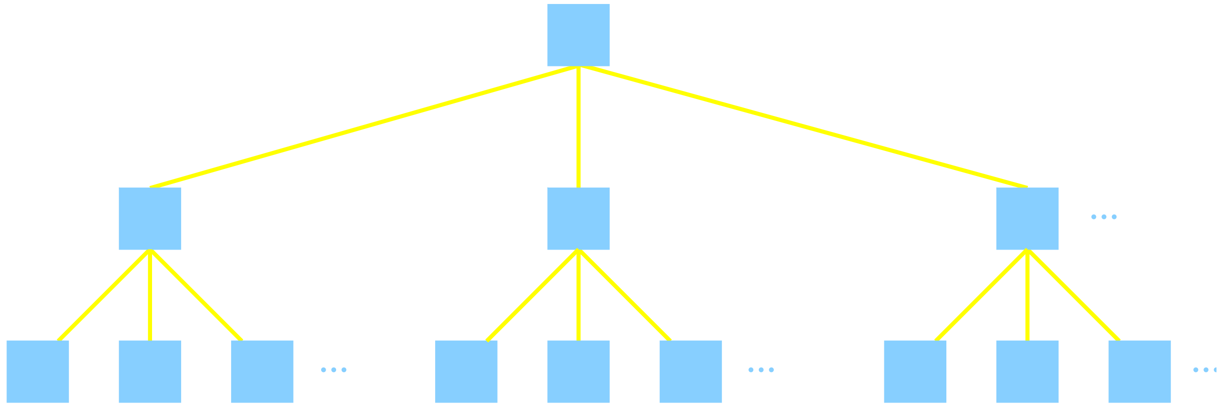
Alice in Wonderland



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

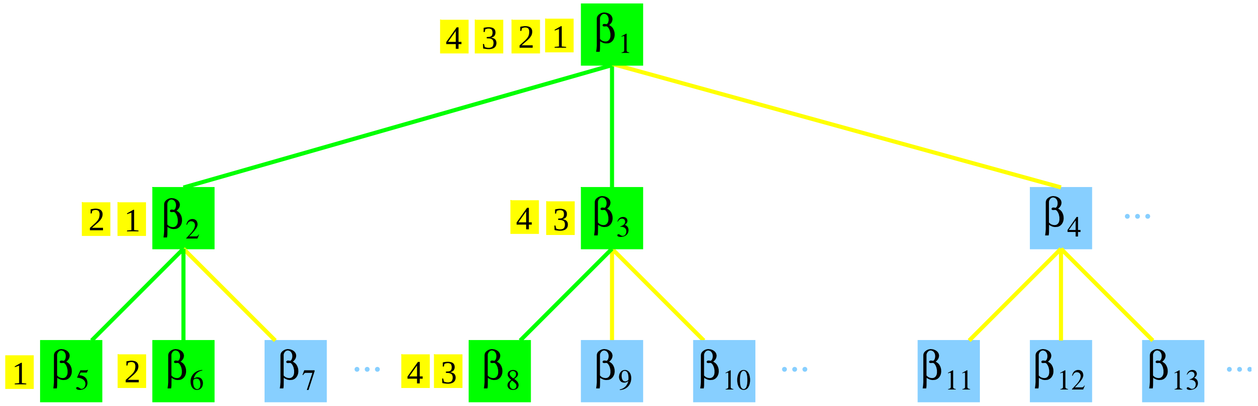
Hierarchical Topic Models

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



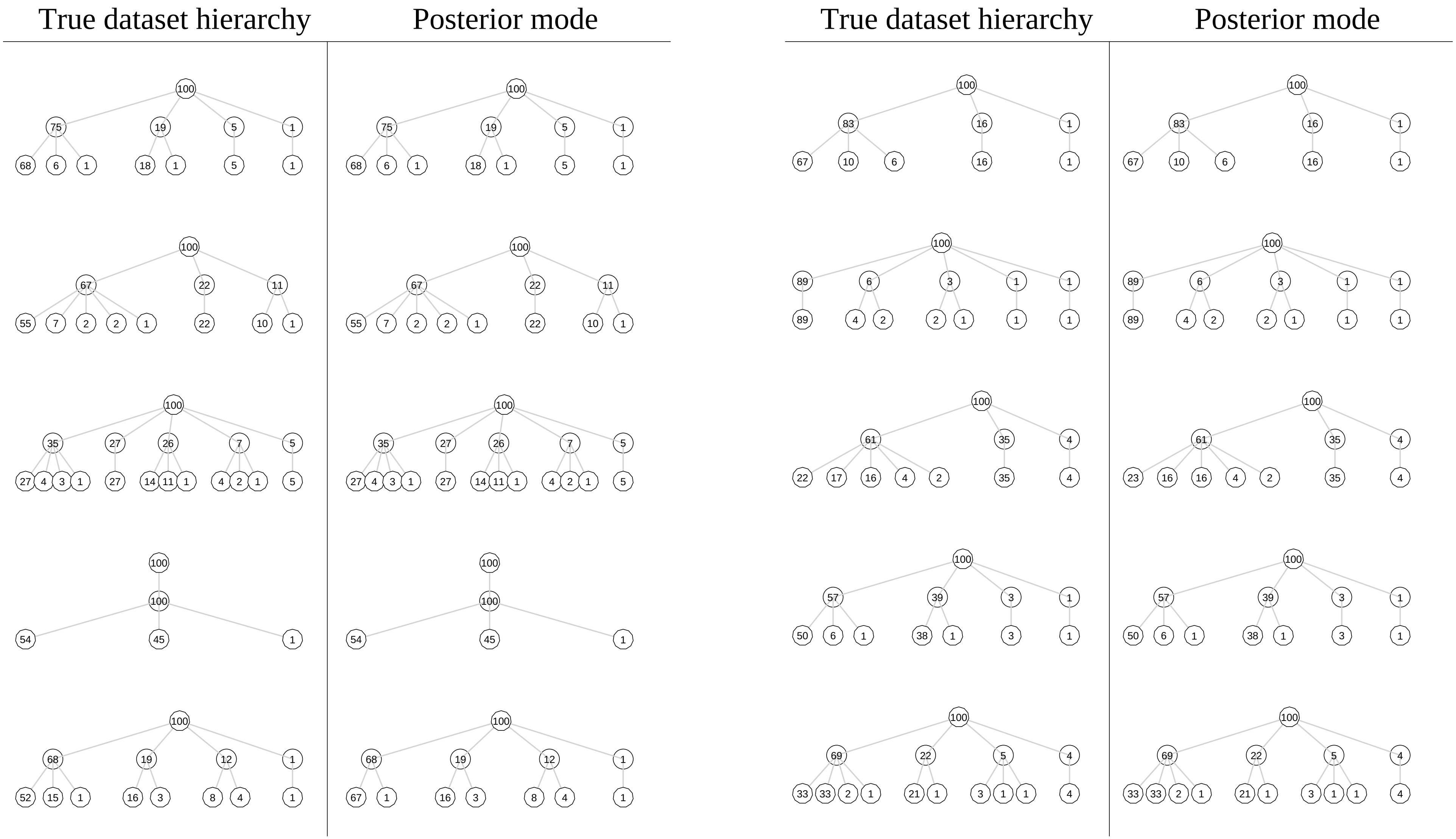
- Infinite number of restaurants in a city:
  - one restaurant is the root restaurant and on each of its infinite tables is a card with the name of another restaurant
  - on each of the tables in those restaurants are cards that refer to other restaurants, and this structure repeats
- Restaurants are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

Nested Chinese Restaurant Process

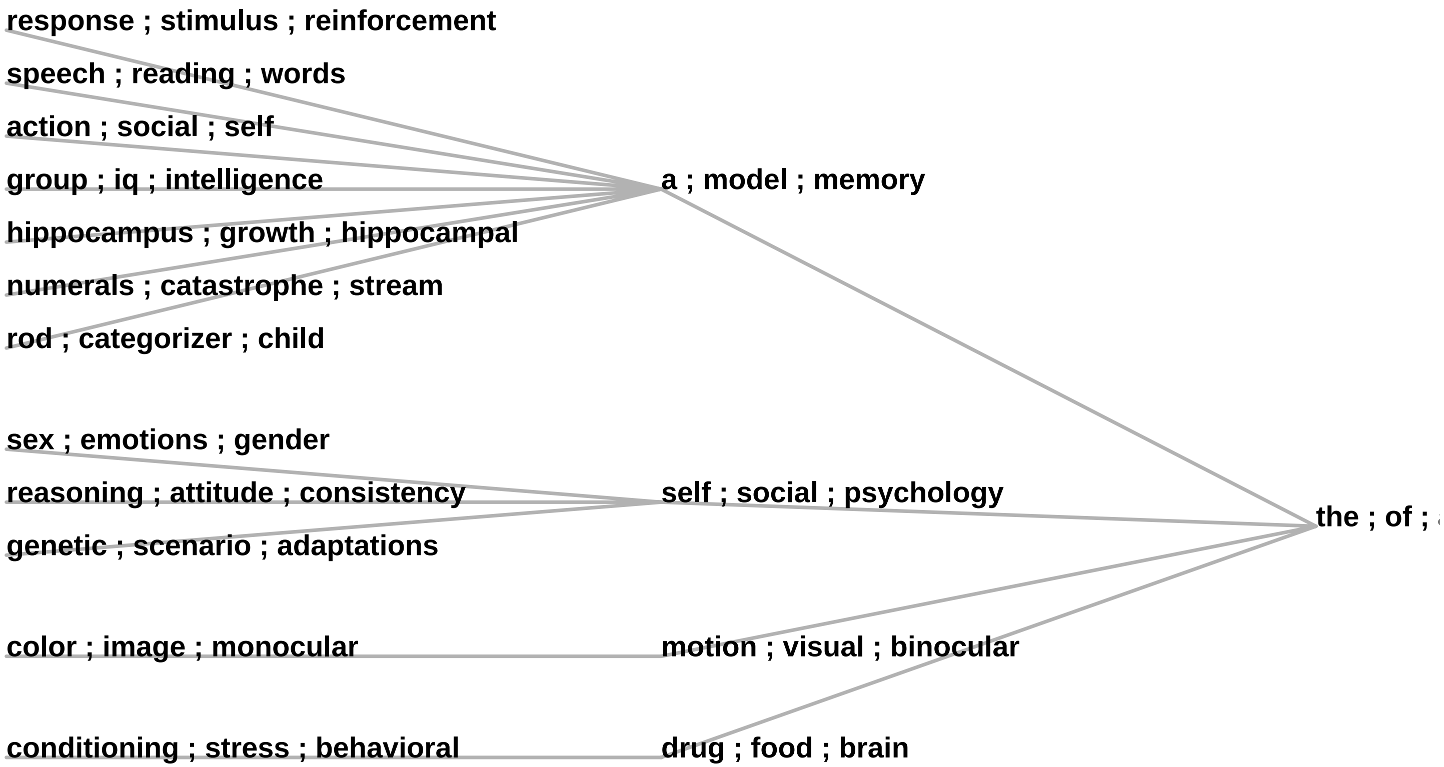


- A tourist arrives in the city for a culinary vacation
  - on each of L evenings, he enters a restaurant and chooses a table, which is labeled with the next evening's restaurant
- The L chosen restaurants constitute a path from the root to a restaurant at the Lth level of the infinite tree
- Assigning each restaurant to a parameter, we can use each tour as a topic path for a document

Estimating the Hierarchy


True dataset hierarchy | Posterior mode | True dataset hierarchy | Posterior mode

Topic Hierarchy from *Psychology Today*



response ; stimulus ; reinforcement

speech ; reading ; words

action ; social ; self

group ; iq ; intelligence

hippocampus ; growth ; hippocampal

numerals ; catastrophe ; stream

rod ; categorizer ; child

a ; model ; memory

sex ; emotions ; gender

reasoning ; attitude ; consistency

genetic ; scenario ; adaptations

self ; social ; psychology

the ; of ;

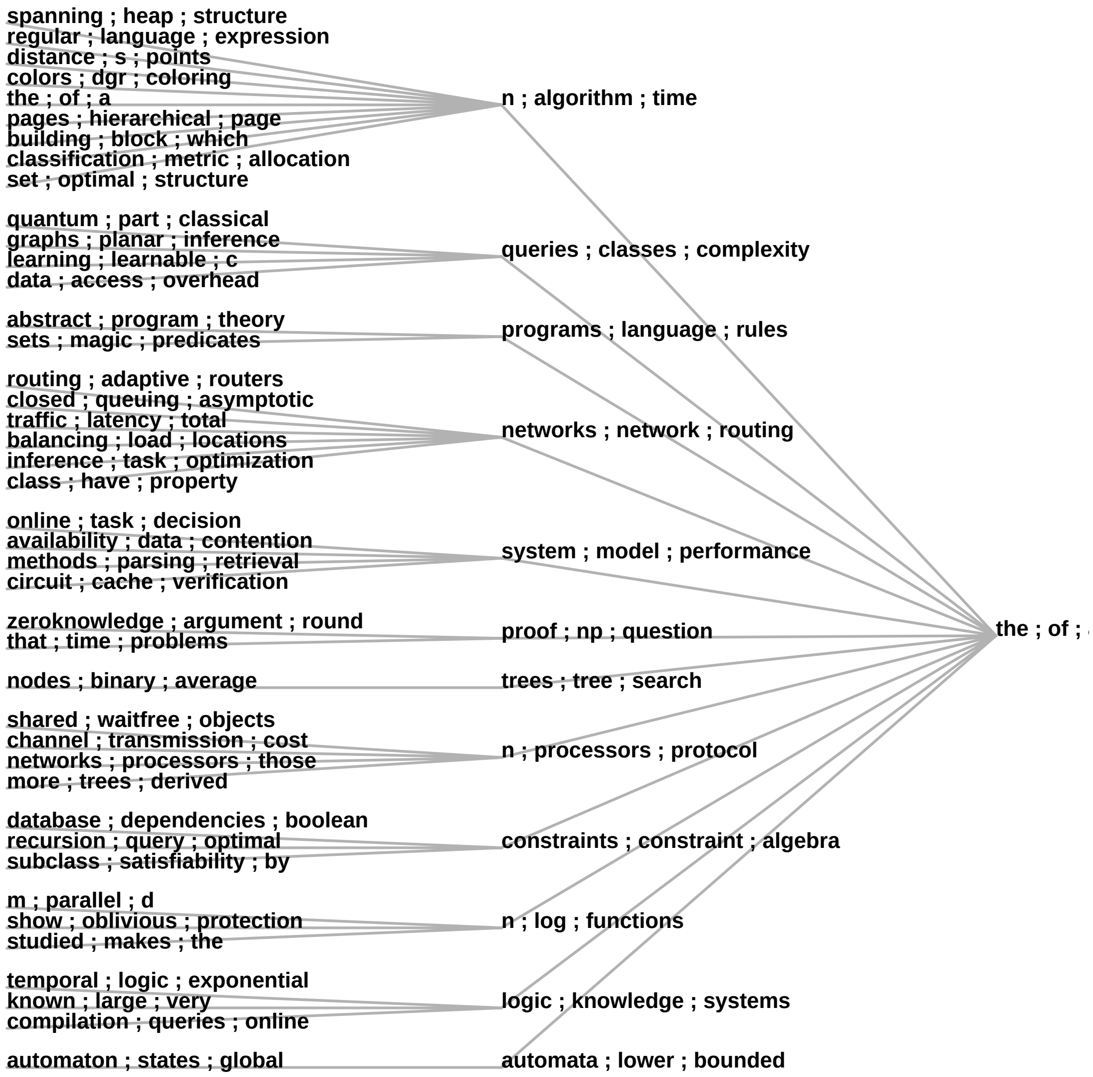
color ; image ; monocular

motion ; visual ; binocular

conditioning ; stress ; behavioral

drug ; food ; brain

Topic Hierarchy from *JACM*



spanning ; heap ; structure
regular ; language ; expression
distance ; s ; points
colors ; dgr ; coloring
the ; of ; a
pages ; hierarchical ; page
building ; block ; which
classification ; metric ; allocation
set ; optimal ; structure

quantum ; part ; classical
graphs ; planar ; inference
learning ; learnable ; c
data ; access ; overhead

abstract ; program ; theory
sets ; magic ; predicates

routing ; adaptive ; routers
closed ; queuing ; asymptotic
traffic ; latency ; total
balancing ; load ; locations
inference ; task ; optimization
class ; have ; property

online ; task ; decision
availability ; data ; contention
methods ; parsing ; retrieval
circuit ; cache ; verification

zeroknowledge ; argument ; round
that ; time ; problems

nodes ; binary ; average

shared ; waitfree ; objects
channel ; transmission ; cost
networks ; processors ; those
more ; trees ; derived

database ; dependencies ; boolean
recursion ; query ; optimal
subclass ; satisfiability ; by

m ; parallel ; d
show ; oblivious ; protection
studied ; makes ; the

temporal ; logic ; exponential
known ; large ; very
compilation ; queries ; online

automaton ; states ; global

n ; algorithm ; time
queries ; classes ; complexity
programs ; language ; rules
networks ; network ; routing
system ; model ; performance
proof ; np ; question
trees ; tree ; search
n ; processors ; protocol
constraints ; constraint ; algebra
n ; log ; functions
logic ; knowledge ; systems
automata ; lower ; bounded

the ; of ;

Empirical Bayes for DP Mixtures

(McAuliffe, Blei, & Jordan, 2005)

- How to set the hyperparameters?

- The religious-Bayesian answer—integrate over them, by introducing distributions with hyper-hyperparameters
  - this isn't hard to do for α_0, but what about G_0?
  - using a parametric model for G_0 with a free hyperparameter isn't very appealing

- The alternative—*empirical Bayes*, which (usually) means fixing the hyperparameters to values that maximize the (marginal) likelihood
  - can use kernel density estimation to estimate G_0
  - although we don't observe data from G_0, we impute samples from G_0; use these in kernel density estimation
  - can also estimate α_0 by maximum marginal likelihood

Conclusions

- Graphical models are said to be “expressive” compared to other machine learning techniques

- But essentially all deployed graphical models are parametric; this isn’t very “expressive”

- The Chinese restaurant and related nonparametric Bayesian methods open the door to much more flexible models

- For more details on this talk: http://www.cs.berkeley.edu/∼jordan

- For more information on the field: type “nonparametric Bayes workshop Rome” into Google