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

Introduction

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 (a focus point for machine learning research)
IV: Graphical models, hierarchical 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

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.

- I.e., if you assert exchangeability, it is reasonable to act as if there is an underlying parameter, there is a prior on that parameter, and the data are conditionally IID given that parameter

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 random variables θ ranging over Euclidean vector spaces

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

Bayesian Nonparametrics

- There are Bayesian nonparametric approaches to many of the main problems in statistics:
  - regression
  - classification
  - clustering
  - survival analysis
  - time series analysis
  - spatial data analysis
  - etc
- These generally involve assumptions of exchangeability or partial exchangeability
  - and corresponding distributions on random objects of various kinds (functions, monotone functions, partitions, measures, etc)
- We'll focus on one problem for concreteness—clustering

Clustering—How to Choose K?

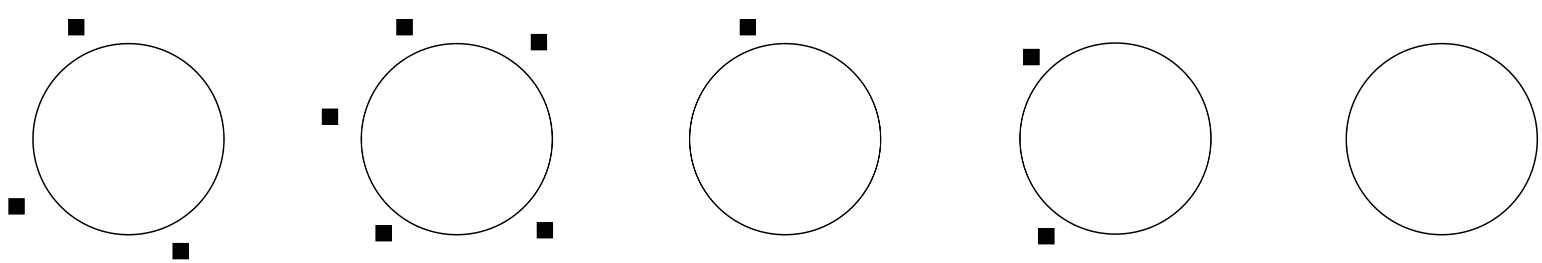
- Adhoc approaches (e.g., hierarchical clustering)
  - they do often yield a data-driven choice of K
  - but there is little understanding of how good these choices are
- Methods based on objective functions (M-estimators)
  - e.g., K-means, spectral clustering
  - do come with some frequentist guarantees
  - 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:

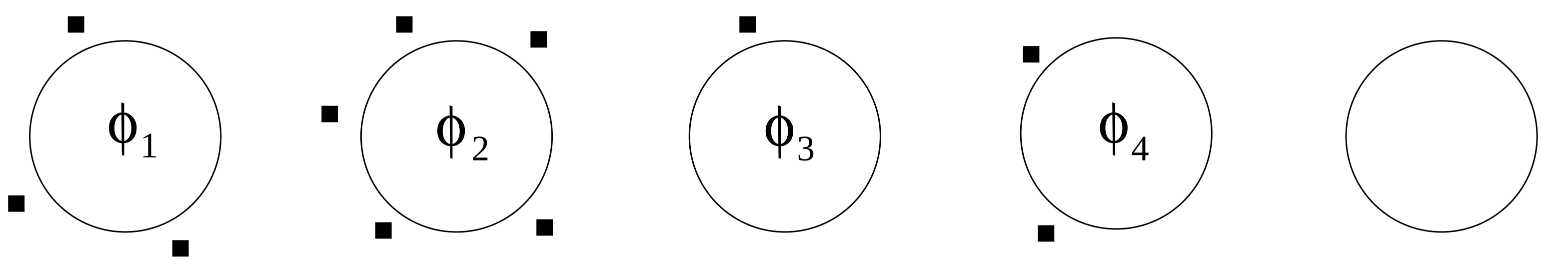
$$\begin{aligned} P(\text{previously occupied table } i \,|\, \mathcal{F}_{m-1}) &\propto n_i \\ P(\text{the next unoccupied table} \,|\, \mathcal{F}_{m-1}) &\propto \alpha_0 \end{aligned} \tag{1}$$

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



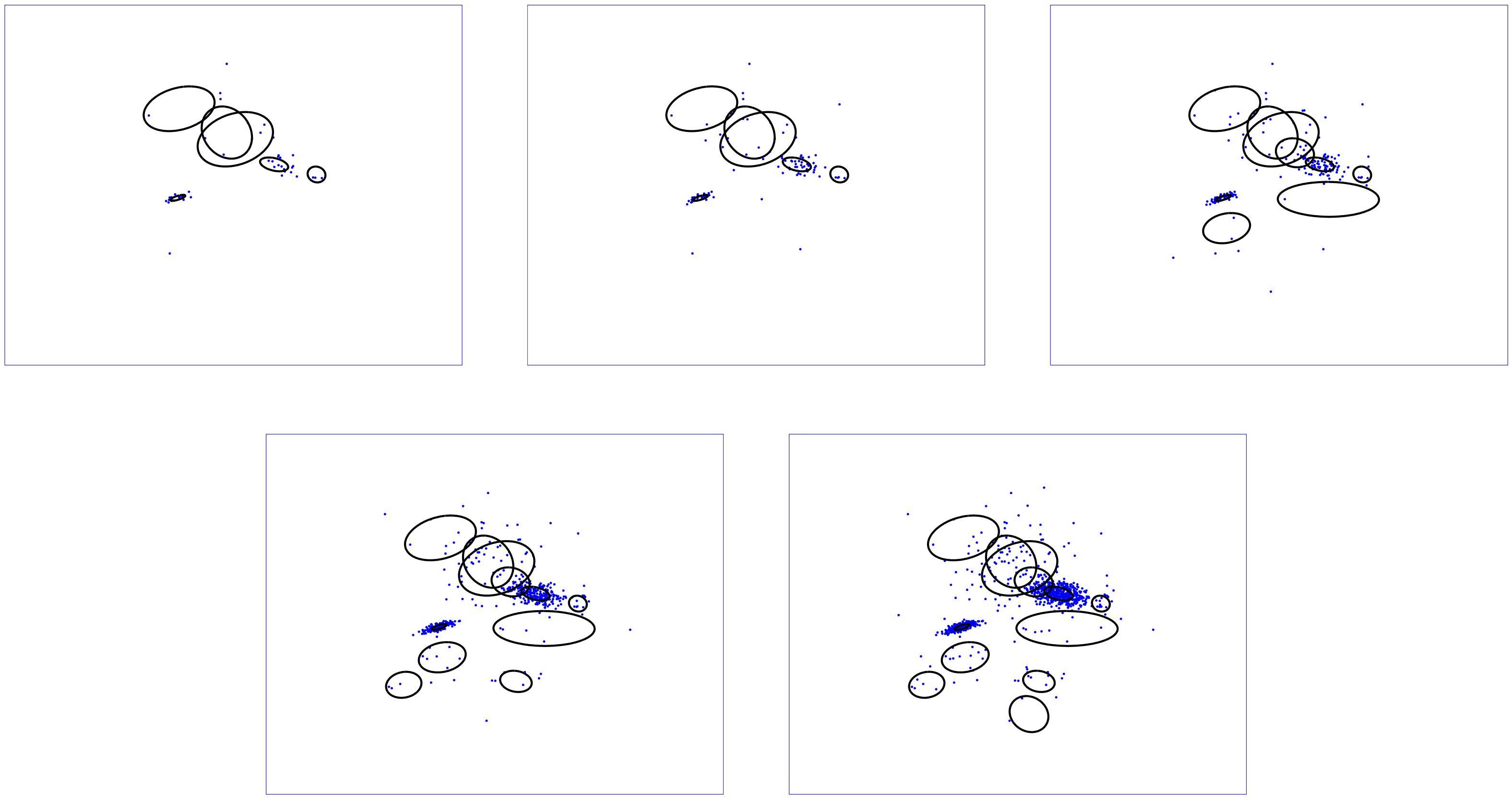
The CRP and Clustering

- Data points are customers; tables are clusters
  - the CRP defines a prior distribution on the partitioning of the data and on the number of tables
- This prior can be completed with:
  - a likelihood—e.g., associate a parameterized probability distribution with each table
  - a prior for the parameters—the first customer to sit at table k chooses the parameter vector for that table (ϕ_k) from the prior



- 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



$$\phi_k = (\mu_k, \Sigma_k) \sim N(a,b) \otimes IW(\alpha,\beta)$$

$$x_i \sim N(\phi_k)$$ for a data point i sitting at table k

Exchangeability

(Blackwell & MacQueen; Kingman; Aldous; Pitman)

- 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)
  - De Finetti implies that there is an underlying random "parameter" and a distribution on that parameter. What are they?

The CRP (cont)

- An additional fact about the CRP:
  - 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

- How do we do inference with a CRP?

- How does this relate to more standard model-based clustering?

- Any theory behind this?

- What can we do that's new with this setup?

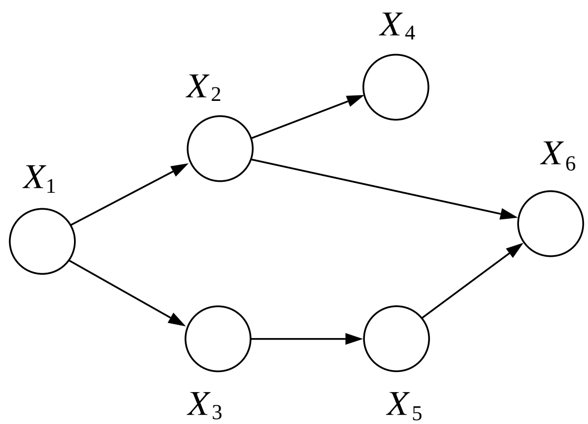
Outline

- Background
- Bayesian mixture models, stick-breaking, Dirichlet processes
- Inference
- Hierarchical and dependent Dirichlet processes
- Semiparametric models
- Applications
- Further directions in nonparametric Bayes: tail-free processes, neutral-to-the-right processes, Polya trees, diffusion trees, Pitman-Yor processes

Background

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:

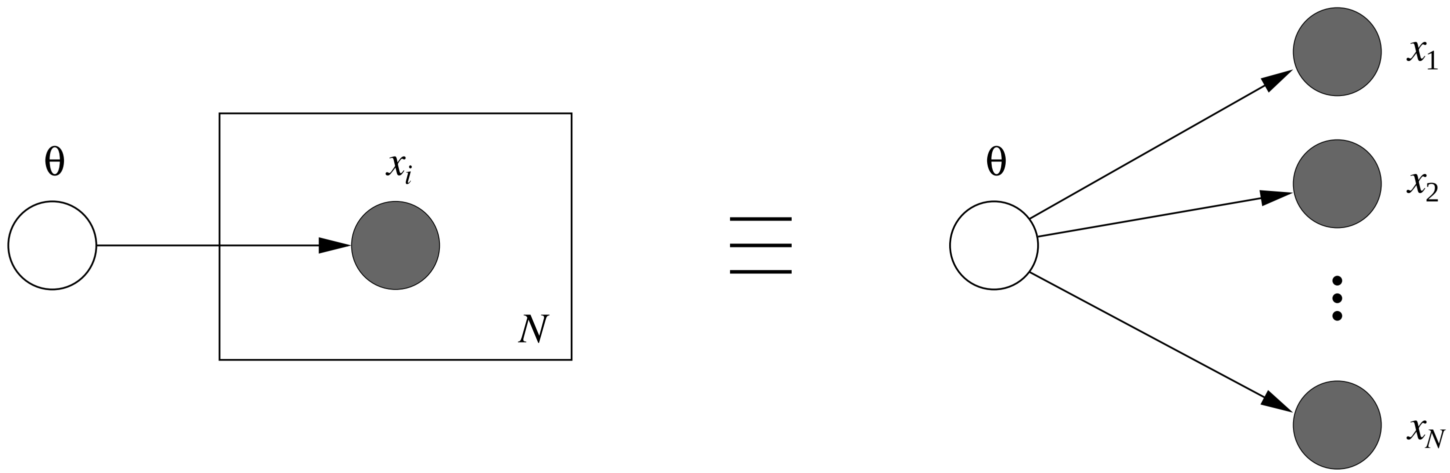


- The joint distribution on $(X_1, X_2, \ldots, X_N)$ factorizes according to the "parent-of" relation defined by the edges $\mathcal{E}$:

$$
\begin{aligned}
p(x_1, x_2, x_3, x_4, x_5, x_6; \theta) = {} & p(x_1; \theta_1)\; p(x_2 \,|\, x_1; \theta_2) \\
& p(x_3 \,|\, x_1; \theta_3)\; p(x_4 \,|\, x_2; \theta_4)\; p(x_5 \,|\, x_3; \theta_5)\; p(x_6 \,|\, x_2, x_5; \theta_6)
\end{aligned}
$$

Plates

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



- Shading denotes observed variables; i.e., conditioning

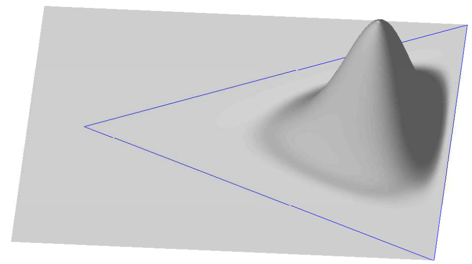
- Note that this graph represents the following marginal probability for the observations $(x_1, x_2, \ldots, x_N)$:

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

Gibbs Sampling

- A Markov chain Monte Carlo (MCMC) method
- Consider a set of variables X_V, with distribution $p(x_V)$ (which may be a conditional distribution)
- 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
- Under (usually) easily-checkable conditions, this scheme has $p(x_V)$ as its equilibrium distribution

Dirichlet Distribution



- Let $\pi = (\pi_1, \pi_2, \ldots, \pi_m)$ be a point in the $(m-1)$-simplex
  - i.e., $0 < \pi_i < 1$ and $\sum_{i=1}^{m} \pi_i = 1$

- Let $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m)$ be a set of parameters, where $\alpha_i > 0$

- The Dirichlet density is defined as:

$$p(\pi \mid \alpha) = \frac{\Gamma(\sum_{i=1}^{m} \alpha_i)}{\prod_{i=1}^{m} \Gamma(\alpha_i)} \pi_1^{\alpha_1 - 1} \pi_2^{\alpha_2 - 1} \cdots \pi_m^{\alpha_m - 1}$$

where $\pi_m = 1 - \sum_{k=1}^{m-1} \pi_k$. This defines an exponential family distribution on the simplex, with:

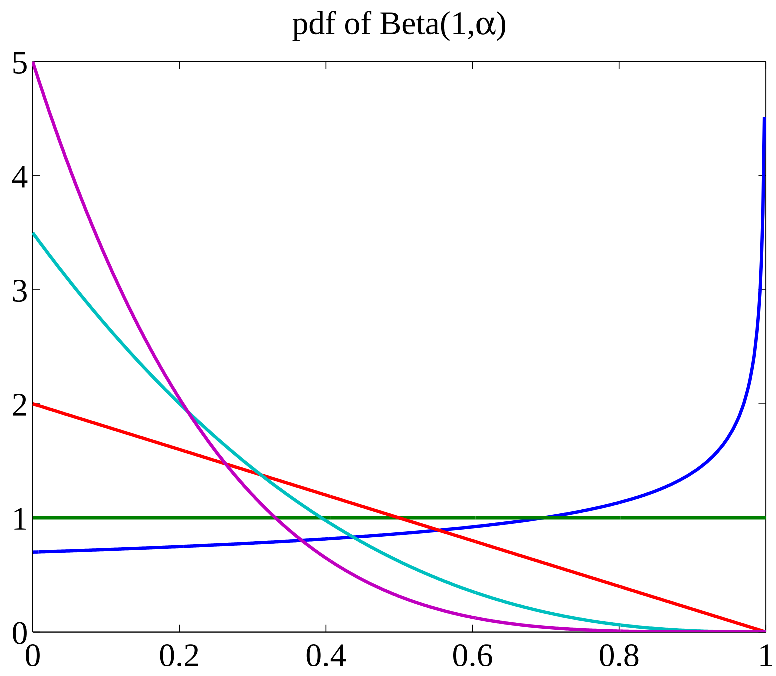
$$\mathbb{E}(\pi_i) = \frac{\alpha_i}{\sum_{i=1}^{m} \alpha_i}$$

Beta Distribution

- A special case of the Dirichlet distribution, where $m = 2$:

$$p(\pi \mid \alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \, \pi^{\alpha_1 - 1}(1 - \pi)^{\alpha_2 - 1}$$

- E.g., $\mathrm{Beta}(1, \alpha)$:



Aggregation Property of Dirichlet Distribution

- What is the distribution of $\pi = (\pi_1, \ldots, \pi_i + \pi_{i+1}, \ldots, \pi_m)$?
  - it is $\mathrm{Dir}(\alpha_1, \ldots, \alpha_i + \alpha_{i+1}, \ldots, \alpha_m)$
  - prove this by using a representation of the Dirichlet as a normalized set of independent gamma random variables

- In general, aggregation of any subset of Dirichlet variables yields a Dirichlet, with corresponding aggregation of the parameters

- E.g.,

$$\sum_k \pi_{i_k} \sim \mathrm{Beta}(\sum_k \alpha_{i_k}, \bar{\alpha} - \sum_k \alpha_{i_k})$$

for a subsequence $\{i_k\}$, where $\bar{\alpha} = \sum_i \alpha_i$

Multinomial-Dirichlet Conjugacy

- Let Z be a multinomial random variable with $p(Z^k = 1) = \pi_k$
- *Posterior*:

$$
\begin{aligned}
p(\pi \,|\, z) &\propto p(z \,|\, \pi)\; p(\pi) \\
&\propto \left(\pi_1^{z^1 - 1} \cdots \pi_m^{z^m - 1}\right) \left(\pi_1^{\alpha_1 - 1} \cdots \pi_m^{\alpha_m - 1}\right) \\
&= \left(\pi_1^{z^1 + \alpha_1 - 1} \cdots \pi_m^{z^m + \alpha_m - 1}\right)
\end{aligned}
$$

  which is $\mathrm{Dir}(z + \alpha)$.

- I.e., the distribution of $\pi \,|\, z$ is $\mathrm{Dir}(\alpha^{post})$, where

$$
\alpha_k^{post} = \begin{cases} \alpha_k + 1 & \text{for } z^k = 1 \\ \alpha_k & \text{otherwise} \end{cases}
$$

Bayesian mixture models, stick-breaking, and Dirichlet processes

Model-Based Clustering

- A generative approach to clustering:
  - pick one of K clusters from a distribution $\pi = (\pi_1, \pi_2, \ldots \pi_K)$
  - generate a data point from a cluster-specific probability distribution

- This yields a finite mixture model:

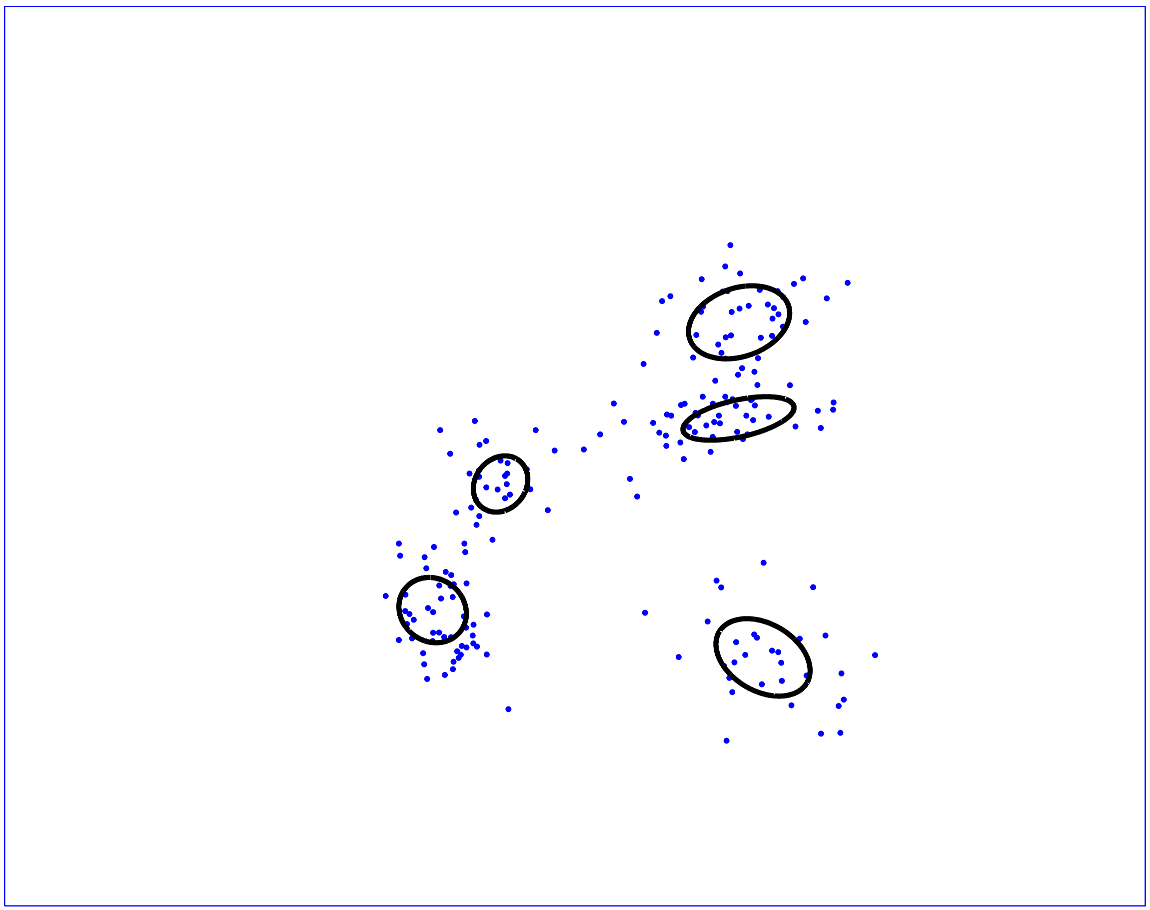
$$p(x \mid \phi, \pi) = \sum_{k=1}^{K} \pi_k \, p(x \mid \phi_k),$$

  where π and $\phi = (\phi_1, \phi_2, \ldots \phi_K)$ are the parameters, and where we've assumed the same parameterized family for each cluster (for simplicity)

- Data $\{x_i\}_{i=1}^{n}$ are assumed to be generated conditionally IID from this mixture

Finite Mixture Models

- E.g., for Gaussian mixtures, $\phi_k = (\mu_k, \Sigma_k)$ and $p(x \mid \phi_k)$ is a Gaussian density with mean μ_k and covariance matrix Σ_k



Finite Mixture Models (cont)

- Mixture models make the assumption that each data point arises from a single mixture component
  - the kth cluster is by definition the set of data points arising from the kth mixture component

- Can capture this explicitly via a latent multinomial variable Z:

$$
\begin{aligned}
p(x \mid \phi, \pi) &= \sum_{k=1}^{K} p(Z^k = 1 \mid \pi)\, p(x \mid Z^k = 1, \phi) \\
&= \sum_{k=1}^{K} \pi_k \, p(x \mid \phi_k)
\end{aligned}
$$

Finite Mixture Models (cont)

- Another way to express this: define an underlying measure

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

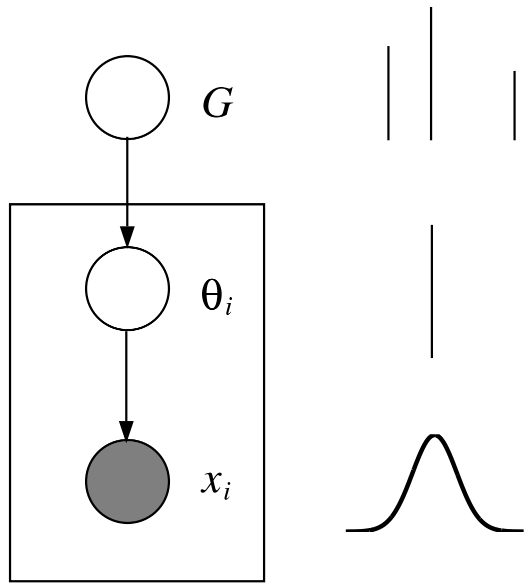
  where δ_{ϕ_k} is an *atom* at ϕ_k

- And define the process of obtaining a sample from a finite mixture model as follows. For $i = 1, \ldots, n$:

$$\begin{aligned} \theta_i &\sim G \\ x_i &\sim p(\cdot \,|\, \theta_i) \end{aligned}$$

- Note that each θ_i is equal to one of the underlying ϕ_k
  - indeed, the subset of $\{\theta_i\}$ that maps to ϕ_k is exactly the kth cluster

Finite Mixture Models (cont)



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

Bayesian Finite Mixture Models

(e.g., Lo; Ferguson; Escobar & West; Robert; Green & Richardson; Neal; Ishawaran & Zarepour)

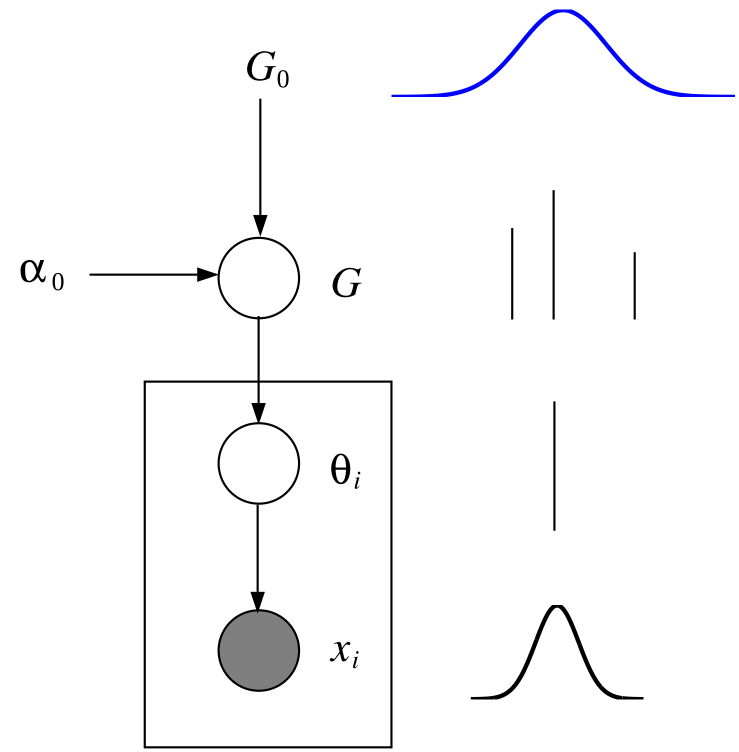
- Need to place priors on the parameters ϕ and π

- The choice of prior for ϕ is model-specific; e.g., we might use conjugate normal/inverse-gamma priors for a Gaussian mixture model
  - let's denote this prior as G_0

- Place a symmetric Dirichlet prior, $\mathrm{Dir}(\alpha_0/K, \ldots, \alpha_0/K)$, on the mixing proportions π
  - the symmetry accords with the (usual) assumption that we could scramble the labels of the mixture components and not change the model
  - the scaling (α_0/K) gives α_0 the semantics of a concentration parameter; the prior mean of ϕ_k is equal to $1/K$

Bayesian Finite Mixture Models (cont)

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



- Note that G is now a *random measure*

Inference

- Posterior distributions (for ϕ, Z and/or π) can't be found analytically; nor can predictive distributions (for future x)

- However, a variety of MCMC sampling algorithms are available

- E.g., use the indicators Z within a Gibbs sampler. Given Z, we know which data points belong to which cluster, so:
  - $p(\pi \,|\, Z, \phi)$: standard multinomial-Dirichlet conjugacy
  - $p(\phi \,|\, Z, \pi)$: separate updates for each cluster; i.e., for each ϕ_k (and conjugacy of G_0 and $p(\cdot \,|\, \phi)$ can make this easy)
  - $p(Z \,|\, \pi, \phi)$: multinomial classification

Model Choice for Finite Mixture Models

- How to choose K, the number of mixture components?

- Various generic model selection methods can be considered: e.g., cross-validation, bootstrap, AIC, BIC, DIC, MDL, covariance penalties, Laplace, bridge sampling, etc

- Or can place a parametric prior on K (e.g., Poisson) and can use Bayesian (model selection or model averaging) methods

- The Dirichlet process and Chinese restaurant process provide a nonparametric Bayesian alternative

Going Nonparametric—A First Perspective

(e.g., Kingman; Waterson; Patil & Taillie; Liu; Ishwaran & Zarepour)

- 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

A Second Perspective—Stick-Breaking

(e.g., Connor & Mosimann; Doksum; Freedman; Kingman; Pitman; Sethuraman)

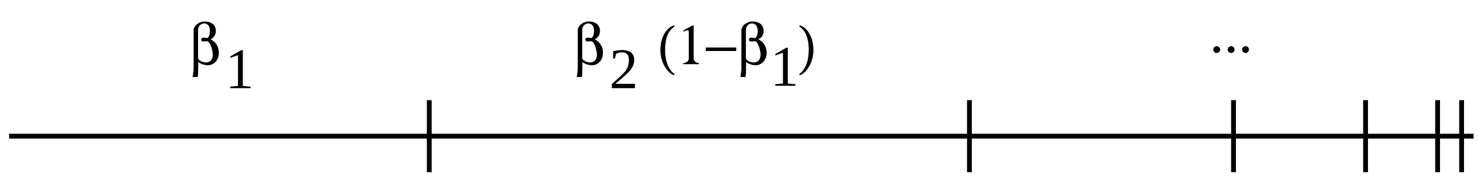
- Define an infinite sequence of Beta random variables:

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

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

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

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



Stick-Breaking (cont)

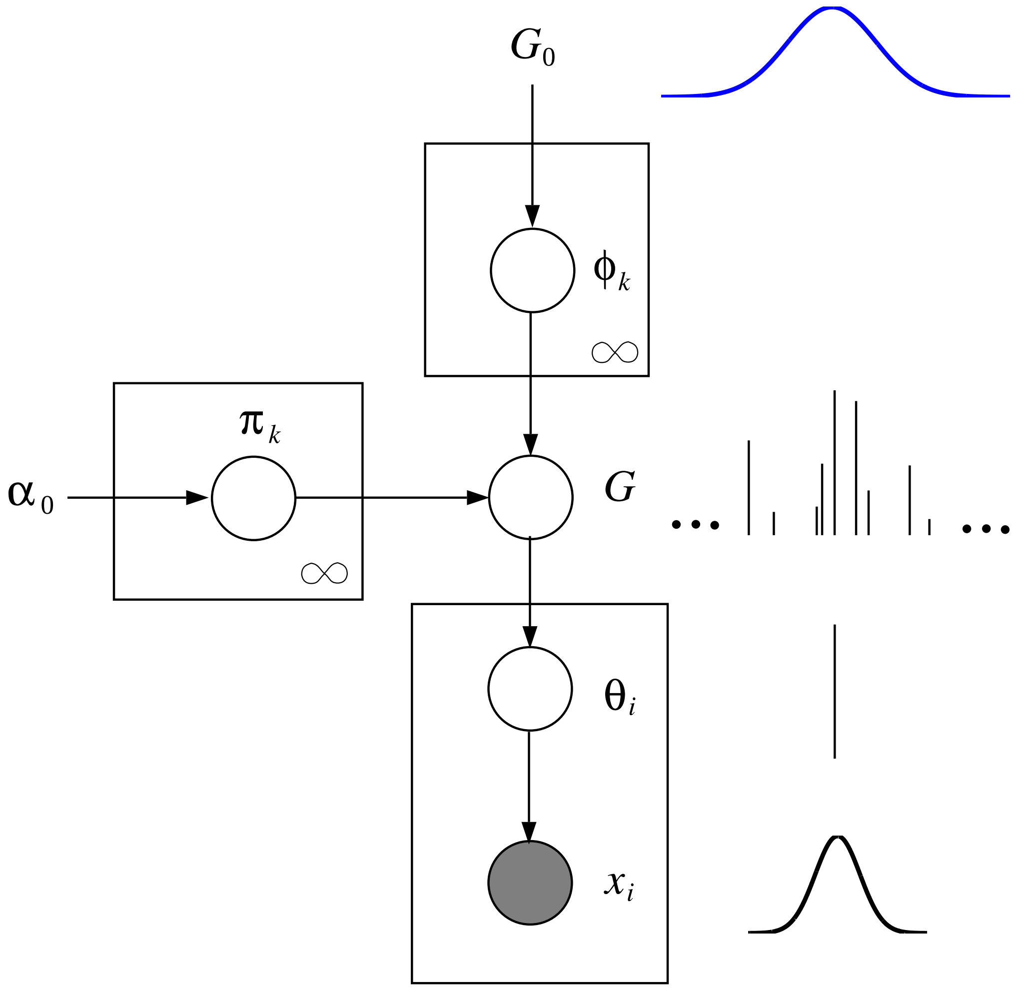
- We now have an explicit formula for each π_k: $\beta_k \prod_{l=1}^{k-1}(1-\beta_l)$

- We can also easily see that $\sum_{k=1}^{\infty}\pi_k = 1$ (wp1):

$$
\begin{aligned}
1 - \sum_{k=1}^{K}\pi_k &= 1 - \beta_1 - \beta_2(1-\beta_1) - \beta_3(1-\beta_1)(1-\beta_2) - \cdots \\
&= (1-\beta_1)(1-\beta_2 - \beta_3(1-\beta_2) - \cdots) \\
&= \prod_{k=1}^{K}(1-\beta_k) \\
&\to 0 \qquad (\text{wp1 as } K \to \infty)
\end{aligned}
$$

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

Graphical Model Representation



Size-Biased Sampling

(Patil & Taillie; Pitman)

- Under an appropriate definition of convergence, the random Dirichlet weights $\mathrm{Dir}(\alpha_0/K, \ldots, \alpha_0/K)$ converge as $K \to \infty$

- View these weights as adjacent subintervals of $(0, 1)$

- Generate a new random sequence π as follows:
  - throw darts uniformly at random at $(0, 1)$ (sequentially)
  - whenever a new subinterval is hit, set π_i to the corresponding weight and increment i

- This *size-biased sampling* yields the stick-breaking distribution

- Rank-ordering a draw from the stick-breaking distribution yields the *Poisson-Dirichlet distribution* (which provides the "appropriate definition of convergence" referred to above)

Towards Inference

- We now have a well-defined notion of a prior random measure; how about a posterior random measure?

- I.e., suppose that I sample a point $\theta \sim G$; how do I update $P(G)$ into $P(G \mid \theta)$?

- Or suppose that I sample a large number of vectors $\theta_i \sim G$; is there a law of large numbers for the empirical measure?

- What about the marginal probabilities? Can I say something about $p(\theta_i \mid \theta_1, \ldots, \theta_{i-1})$ when G has been integrated out (in some sense)?

- Onward and upward into some more serious nonparametrics...

Stochastic Processes

- In elementary probability theory, random variables are defined as functions whose ranges are the reals

- In more advanced probability theory, one lets random variables range over more general spaces, including function spaces and spaces of measures

- Stochastic process theory provides a unifying perspective on these more general random objects

Gaussian Processes

- Suppose that we wish to place a distribution on a function space
- Gaussian processes provide one way to do this
- Recall the recipe (for functions $x(t)$ on $\mathbb{R}$):
  - specify a Gaussian distribution for the finite-dimensional distributions; i.e., for each finite collection of values $\{x(t_i)\}$
  - do so in a way that makes these specifications consistent with each other (use a covariance function)
- Kolmogorov's theorem says that consistency suffices to yield a meaningful limit
  - we obtain a distribution on functions—the Gaussian process
- Similarly, we can obtain distributions on measures...

Dirichlet Process

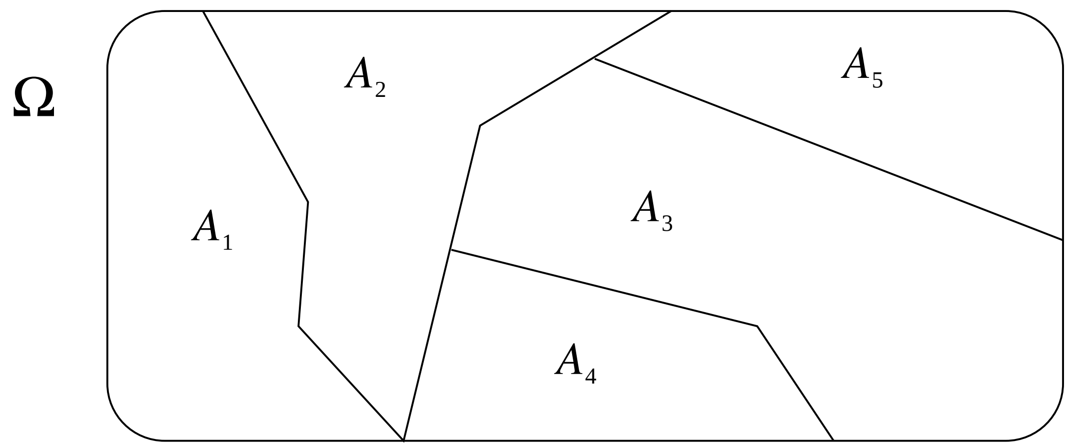
- A measure is a function from subsets to the nonnegative reals
- The finite-dimensional distributions will be defined on finite collections of subsets
  - it suffices to consider partitions
- We need to ensure consistency—probabilities need to add when we collapse cells in a partition
  - the Dirichlet distribution is one way to achieve this

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_1 from G. What is the posterior process?

- For a fixed partition, we get a standard Dirichlet update (for the cell that contains θ_1 the exponent increases by one; stays the same for all other cells)
  - this is true for even the tiniest cell
  - suggests that the posterior is a Dirichlet process in which the base measure has an atom at θ_1

- Indeed, we have (for a proof, see, e.g., Schervish, 1995):

$$G \,|\, \theta_1 \sim \mathrm{DP}(\alpha_0 G_0 + \delta_{\theta_1})$$

- Iterating the posterior update yields:

$$G \,|\, \theta_1, \ldots, \theta_n \sim \mathrm{DP}(\alpha_0 G_0 + \sum_{i=1}^{n} \delta_{\theta_i})$$

Relationship to Stick-Breaking

- Recalling the formula for the expectation of a Dirichlet random variable, for any set $A \subseteq \Omega$, we have:

$$\mathbb{E}[G(A) \mid \theta_1, \ldots, \theta_n] = \frac{\alpha_0 G_0(A) + \sum_{i=1}^n \delta_{\theta_i}(A)}{\alpha_0 + n} \to \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}(A)$$

where ϕ_k are the unique values of the θ_i, where $\pi_k = \lim_{n\to\infty} n_k/n$, and where n_k is the number of repeats of ϕ_k in the sequence $(\theta_1, \ldots, \theta_n)$

  – assuming that the posterior concentrates, this suggests that the random measures $G \sim \mathrm{DP}(\alpha_0 G_0)$ are discrete (wp1)

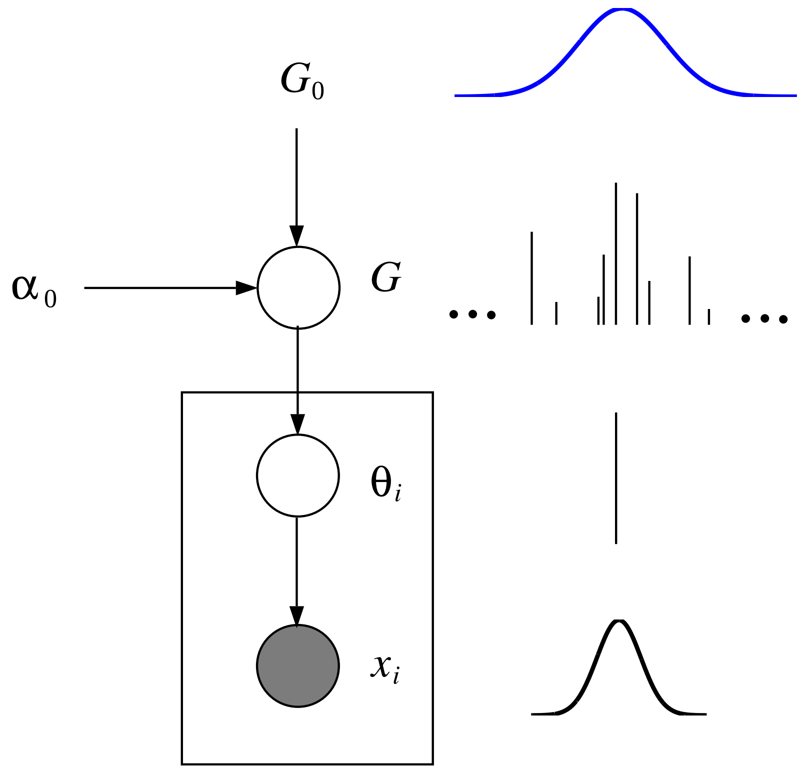
- Is there an infinite sum of the form $G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$ that obeys the definition of the Dirichlet process?

  – yes, the stick-breaking random measure!
  – this important result is not hard to prove; it follows from elementary facts about the Dirichlet distribution (Sethuraman, 1994)

Back to Mixture Models

- In the mixture model setting, θ_i is the parameter associated with the ith data point x_i
  - i.e., it's not observed
- We use the Dirichlet process to induce a prior on the θ_i, and then complete the model by introducing a likelihood
  - just as in finite mixture models
- This yields a model known as a Dirichlet process mixture model

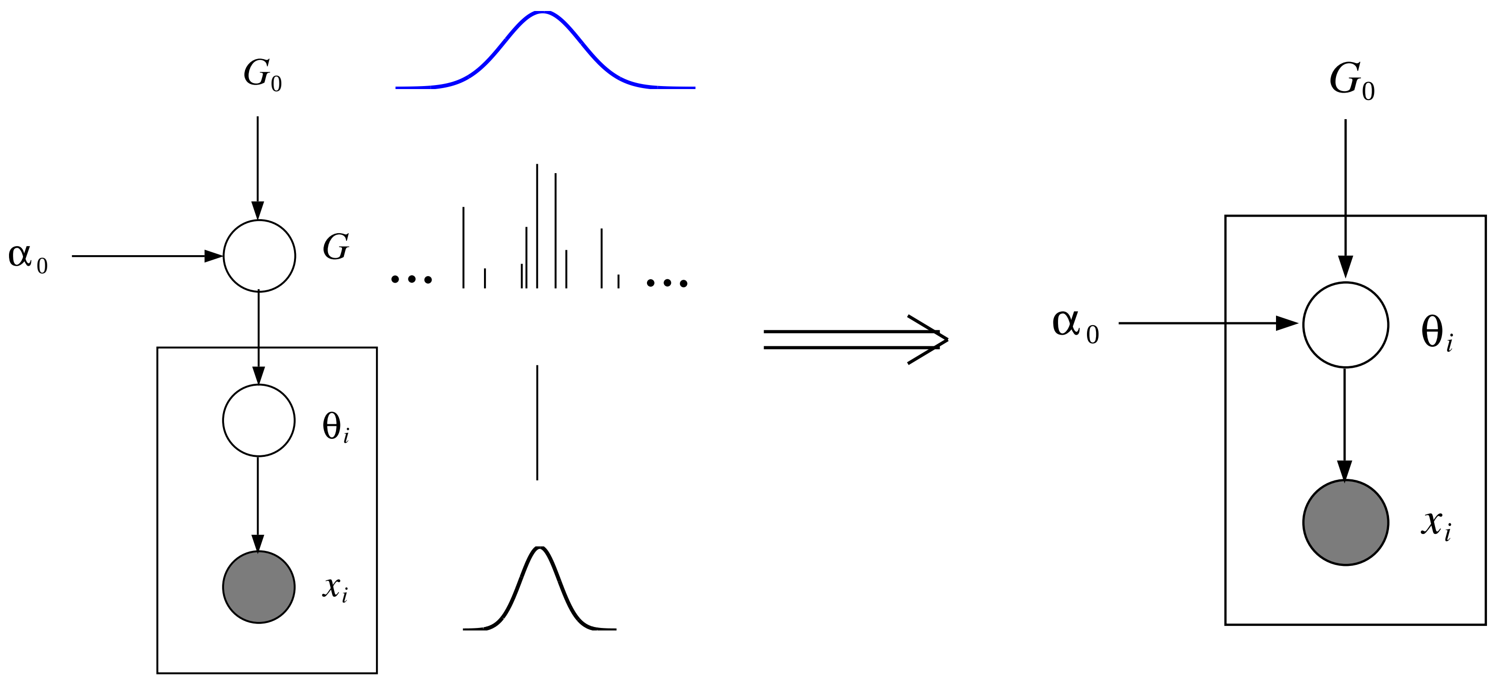
Dirichlet Process Mixture Models



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

Marginal Probabilities

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



Marginal Probabilities (cont)

- Recall the formula

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

- Let A be a singleton set equal to one of the ϕ_k. The formula says that the marginal probability of observing ϕ_k again is proportional to n_k.

- And the marginal probability of observing a new ϕ vector is proportional to α_0.

- This is just the Pólya urn scheme!

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

Inference

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

MCMC for Dirichlet Process Mixtures

(e.g., Escobar; MacEachern; Doss; West; Müller; Neal; Ishwaran, James & Zarepour; Jain; Gelfand & Kottas)

- Key insight: *take advantage of exchangeability*

- E.g., in the Chinese restaurant prior, where should customer i sit, conditional on the seating choices of all other customers?
  - easy when customer i is the last customer to arrive
  - (seemingly) hard otherwise

- But by exchangeability, can always swap customer i with the final customer; the joint probability is invariant to such swaps

Collapsed Gibbs Sampling

- For each data point:
  - pretend that it is the last point (by exchangeability)
  - the prior is just the Chinese restaurant dynamics
  - the likelihood is just the usual mixture likelihood
  - this yields a straightforward formula for the conditional for Gibbs sampling (in the conjugate case)

- If parameters are also desired, 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 the posterior distribution on partitions and parameters

Alternative MCMC Algorithms

- Collapsed Gibbs isn't always feasible
  - need conjugacy for collapsed Gibbs
  - slow to mix, due to moving a single data point at a time

- Metropolis-Hastings (Neal)

- Gibbs sampling with auxiliary parameters (MacEachern & Müller; Neal)

- Split-merge algorithms (Green & Richardson; Jain & Neal)
  - allow groups of points to move
  - non-conjugacy

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

Gibbs Sampling based on the TDP (Blocked Gibbs)

- State of the Markov chain: (β, θ, Z)
- For $i \in \{1, 2, \dots, n\}$, sample Z_i from:

$$p(Z_i^k = 1 \,|\, \beta, \theta, x) \propto \pi_k \, p(x_i \,|\, \theta_k)$$

- For $k \in \{1, 2, \dots, T\}$, sample β_k from $\mathrm{Beta}(\gamma_{k1}, \gamma_{k2})$, where:

$$\begin{aligned} \gamma_{k1} &= 1 + \textstyle\sum_{i=1}^{n} z_i^k \\ \gamma_{k2} &= \alpha_0 + \textstyle\sum_{j=k+1}^{T} \sum_{i=1}^{n} z_i^j \end{aligned}$$

- For $k \in \{1, 2, \dots, T\}$, sample θ_k from its posterior

Variational Inference

- Recall 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)

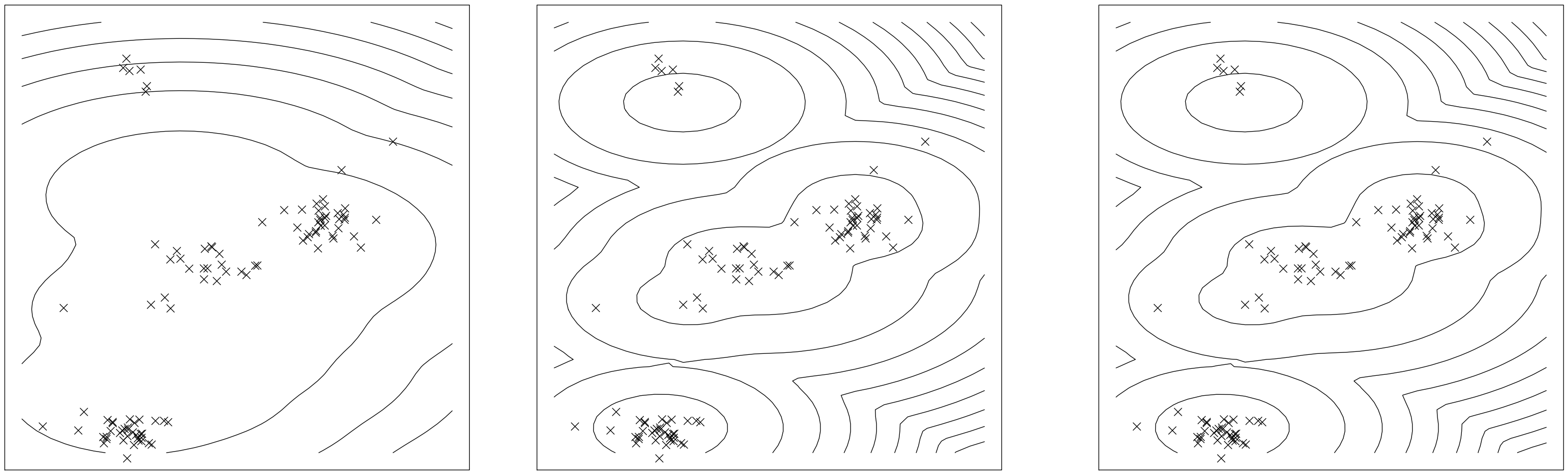
- The Q distribution is a truncated stick-breaking representation (note that P is *not* truncated)
- Variational inference equations for a conjugate DP mixture in the exponential family:

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

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

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

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

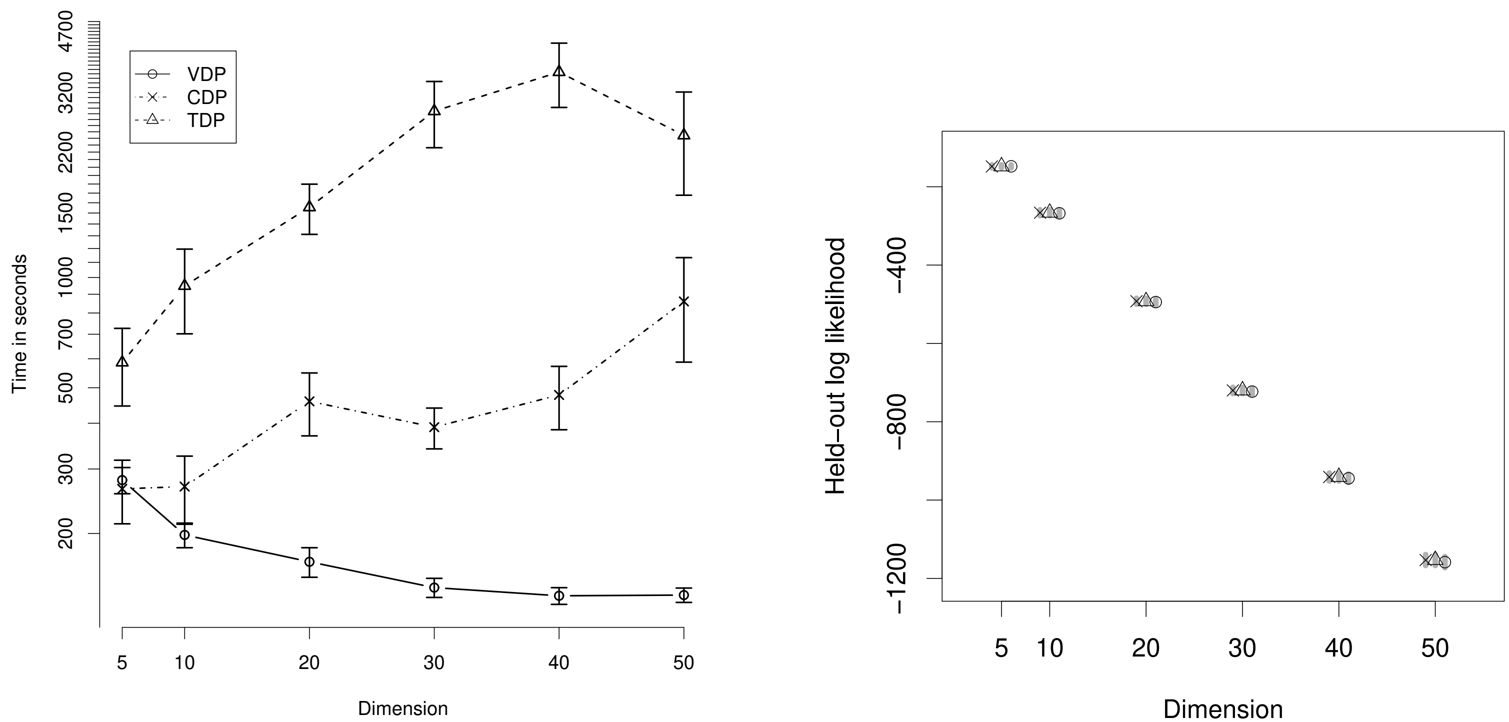


Figure 2. (Left) Convergence time per dimension across ten datasets for variational inference (VDP), the TDP Gibbs sampler (TDP), and the collapsed Gibbs sampler (CDP). Grey bars are standard error. (Right) Average held-out log likelihood for the corresponding predictive distributions.

Bayesian Inference for α_0

(West, 1992)

- The prior for the number of tables K depends (strongly) on α_0; must take care in the choice of this parameter

- Can integrate over α_0 via a nice trick

- Antoniak (1974) showed that

$$p(k \mid \alpha_0) \propto \alpha_0^k \, \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + n)}$$

- Rewrite the ratio of gammas:

$$\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + n)} = \frac{(\alpha_0 + n) B(\alpha_0 + 1, n)}{\alpha_0 \, \Gamma(n)}$$

where $B(\alpha_1, \alpha_2) = \int x^{\alpha_1 - 1}(1 - x)^{\alpha_2 - 1} dx$ is the beta function.

Bayesian Inference for α_0 (cont)

(West, 1992)

- This yields

$$p(\alpha_0 \,|\, k) \propto p(\alpha_0)\alpha_0^{k-1}(\alpha_0 + n) \int x^{\alpha_0}(1-x)^{n-1}dx,$$

which is a marginal probability under joint that involves a beta distribution for x and a distribution that's conjugate to a gamma for α_0

- So introduce x as an explicit variable (this is called *augmentation*) and let the prior $p(\alpha_0)$ be a gamma distribution

- This yields simple conditionals for Gibbs sampling

Empirical Bayes for Dirichlet Process Mixtures

(McAuliffe, Blei, & Jordan, 2005)

- Full Bayesians always attempt to integrate over hyperparameters, often by introducing distributions with hyper-hyperparameters
  - as we've seen, this isn't hard to do for α_0, but what about G_0?
  - it's common to adopt a parametric model for G_0, with a free hyperparameter
  - this isn't very appealing—it puts mass on a small parametrized subset of distributions at the core of a nonparametric model

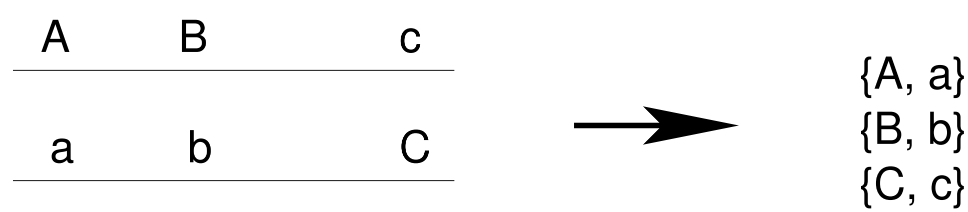
- The alternative is *empirical Bayes*, which (usually) means fixing the hyperparameters to values that maximize the (marginal) likelihood
  - can use kernel density estimation to estimate G_0
  - although we don't observe data from G_0, we impute samples from G_0; use these in kernel density estimation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

Applications

- Haplotype modeling
- Latent Dirichlet allocation with unknown numbers of topics
- Infinite hidden Markov model

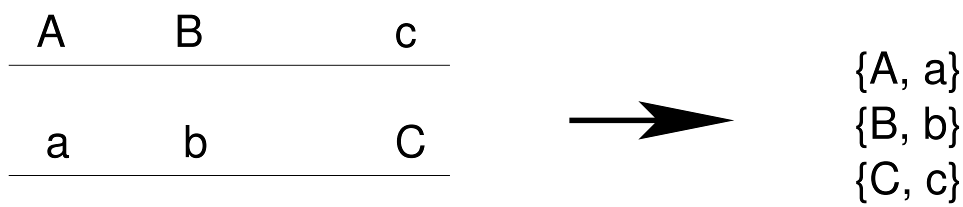
Haplotype Modeling

- 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.)



- 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 the genetics leads to a mixture modeling problem

- But what is the cardinality of $\mathcal{H}$?
  - Dirichlet process mixtures to the rescue

DP-Based Haplotype Model

(Xing, Sharan, & Jordan, 2004)

- In the Chinese restaurant representation, each table is associated with an underlying haplotype (the chromosome of a putative ancestral human)
- 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

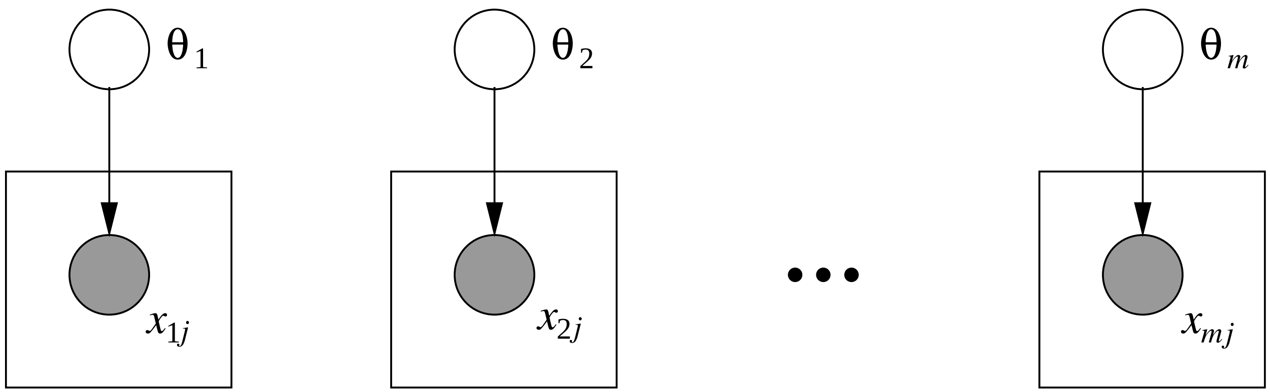
Hierarchical and dependent Dirichlet processes

Hierarchical Bayesian Modeling

- A "hierarchy" (in the Bayesian sense of the term) means a model in which parameters are treated as random variables
  - this provides a way to express prior and posterior uncertainty
  - this provides a way to link related quantities
- Personal opinion—the inferential consequences of hierarchical modeling (see below) are often the main practical argument for taking a Bayesian approach
- The other rationale are often less meaningful in practice
  - as often as not, priors are a nuisance
  - "optimality" is of questionable relevance when models are rough approximations (as is often the case)
  - going Bayesian doesn't free you from the responsibility of post-hoc evaluation (particularly if you're foregoing frequentist-style evaluation)

Multiple Learning Problems

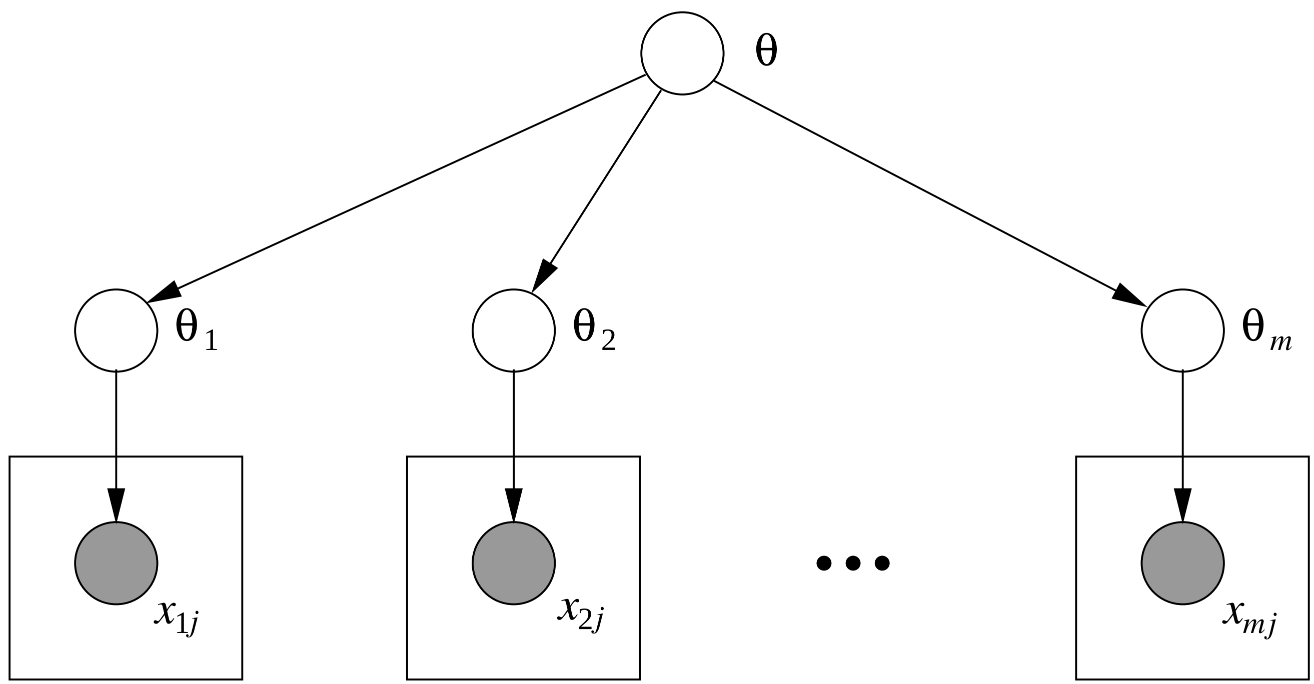
- We often face multiple, related learning problems ("transfer learning")
- E.g., multiple Gaussian means: $x_{ij} \sim N(\theta_i, \sigma_i^2)$



- 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

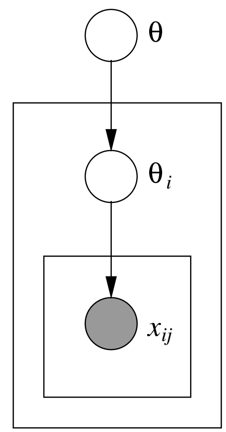
- The Bayesian solution is to view the parameters θ_i as random variables, sampled from an underlying variable θ



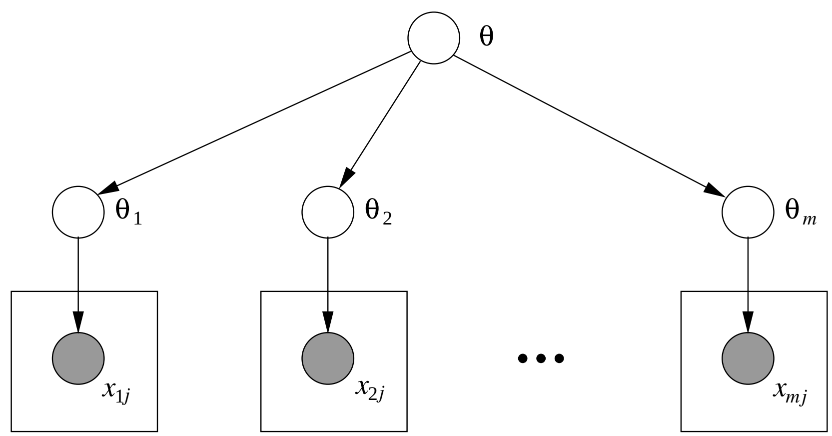
- Given this overall model, Bayesian inference yields *shrinkage*—the posterior mean for each θ_k combines data from all of the groups, without simply lumping the data into one group

Hierarchical Modeling

- Recall the plate notation:



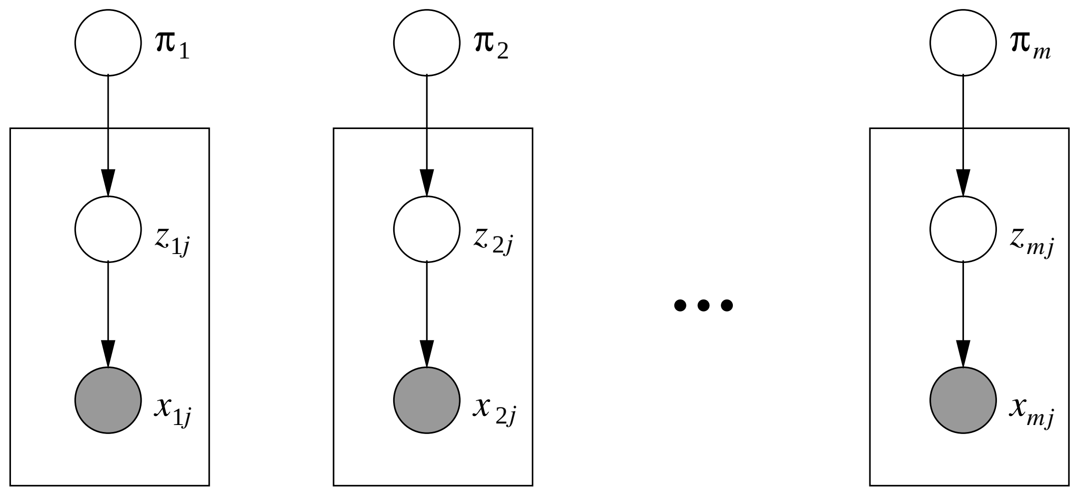
- Equivalent to:



Multiple Clustering Problems

- What about the case in which we have multiple related clustering problems?
  - what to share? how to share?

- Mixture models: $p(x_{ij}\,|\,\pi_i, \theta_i) = \sum_{l=1}^{K_i} p(Z_{ij}^l = 1\,|\,\pi_i)\; p(x_{ij}\,|\,Z_{ij}^l = 1, \theta_i)$

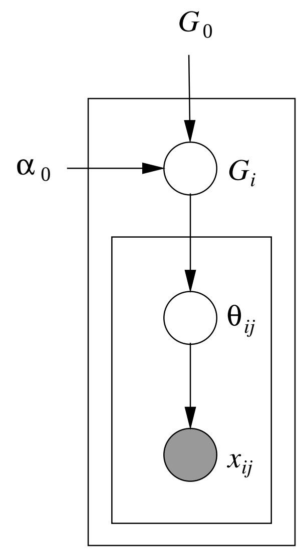


- What to share: π_i?, θ_i? What if we don't know the K_i?

- A nonparametric Bayesian approach can clean this up quite nicely...

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 work well in the parametric context fail (completely) in the nonparametric context... :-(

Hierarchical Dirichlet Processes

(Teh, Jordan, Beal & Blei, 2004)

- 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, 2004)

- 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:

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

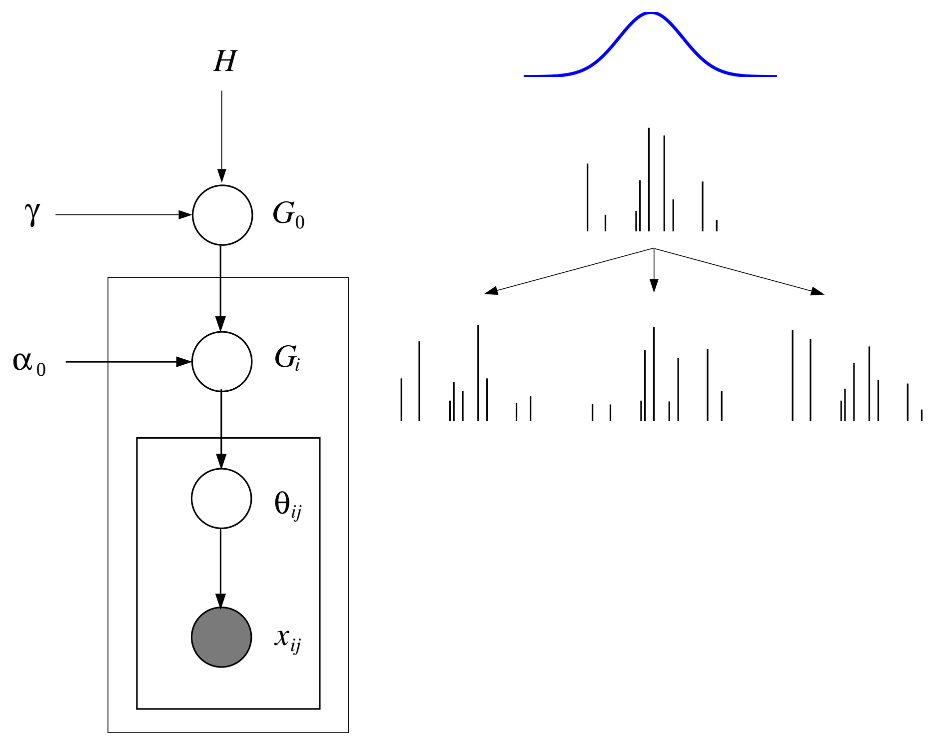
- Then

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

  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



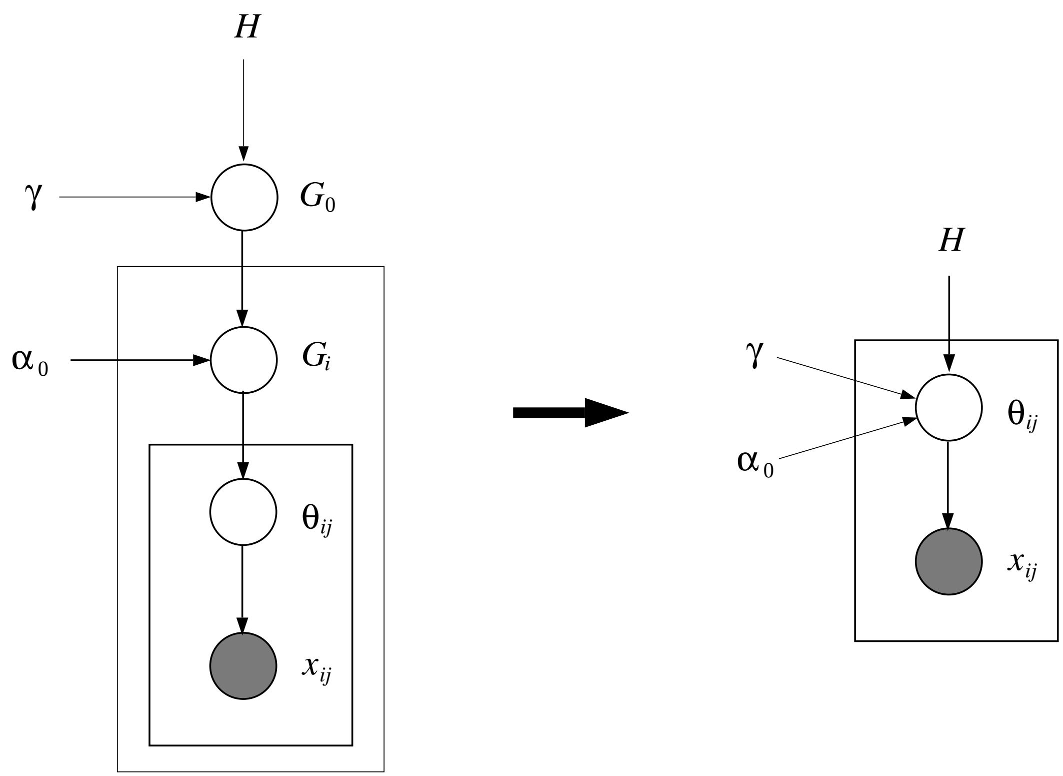
$$
\begin{aligned}
G_0 \mid \gamma, H &\sim \mathrm{DP}(\gamma H) \\
G_i \mid \alpha, G_0 &\sim \mathrm{DP}(\alpha_0 G_0) \\
\theta_{ij} \mid G_i &\sim G_i \\
x_{ij} \mid \theta_{ij} &\sim F(x_{ij}, \theta_{ij})
\end{aligned}
$$

Stick-Breaking Representation

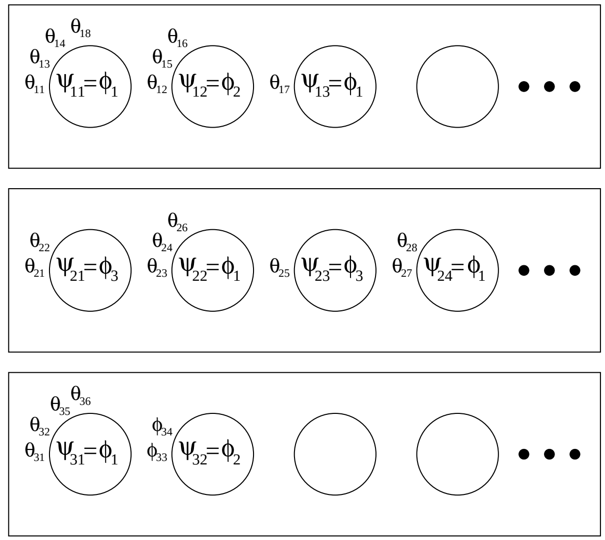
$$
\begin{aligned}
\beta'_k &\sim \mathrm{Beta}(1, \gamma) \\
\beta_k &= \beta'_k \prod_{l=1}^{k-1} (1 - \beta'_l) \\
\pi'_{jk} &\sim \mathrm{Beta}\left(\alpha_0 \beta_k, \alpha_0 \left(1 - \sum_{l=1}^{k} \beta_l\right)\right) \\
\pi_{jk} &= \pi'_{jk} \prod_{l=1}^{k-1} (1 - \pi'_{jl}) \\
G_0 &= \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k} \\
G_j &= \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_k}
\end{aligned}
$$

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

From HDPs to DDPs

- HDPs are a special case of a more general framework
- Recall the stick-breaking representation of the Dirichlet process:

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

- The HDP defines a set of groups indexed by j and allows the stick lengths $\{\pi_k\}$ to vary across groups:

$$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_k}$$

- The atom locations are constant over groups so as to allow sharing of clusters (that's the whole idea of the HDP)

Dependent Dirichlet Processes

(MacEachern, 2000)

- More generally (and with goals other than sharing clusters in mind), we could allow both the stick lengths and the atom locations to vary across groups

- Even more generally, there's no reason to restrict ourselves to a finite index set—we can index the stick lengths and the atom locations by a continuous variable:
$$G_x = \sum_{k=1}^{\infty} \pi_{xk} \delta_{\phi_{xk}}$$
for $x \in \mathcal{X}$, where $\mathcal{X}$ might be the real line or some more general space
  - i.e., holding k fixed, the stick lengths π_{xk} and the atom locations ϕ_{xk} become general stochastic processes (i.e., indexed collections of random variables)

- We get collections of DPs that vary smoothly as a function of some covariate

Examples

- Spatial DDPs (MacEachern, Kottas and Gelfand, 2001)
  - constant stick lengths π_k
  - each atom ϕ_{xk} is a Gaussian process indexed by $x \in \mathbb{R}^2$
- ANOVA DDPs (De Iorio, Müller and Rosner, 2004)
  - consider factors A and B and define random effects ρ_{ak} and τ_{bk}
  - for $x = (a, b)$, let $\phi_{xk} = \mu_k + \rho_{ak} + \tau_{bk}$
  - constant stick lengths π_k
  - this yields row-wise and column-wise linkage of nonparametric distributions in the cells of an ANOVA table

Semiparametric Models

(e.g., Bush and MacEachern, 1996; Mukhopadhyay and Gelfand, 1997)

- "Semiparametric" generally refers to a class of models that include both a parametric component and a nonparametric component

- Often the interest is in inference about the parametric component—the nonparametric component is a *nuisance parameter*

  – to be estimated and "projected out" in a frequentist setting
  – to be integrated out in a Bayesian setting

- Independent component analysis is an example (we're interested in the demixing matrix—a parameter—and want to find it whatever the unknown source distributions, about which we make few assumptions)

- Dirichlet processes are useful as nonparametric components in semiparametric models

Example: Poisson Regression

(Carota and Parmigiani, 2002)

- Suppose that we don't really trust the Poisson model (e.g., over-dispersion)
- Replace the Poisson with a DP with a Poisson base measure:

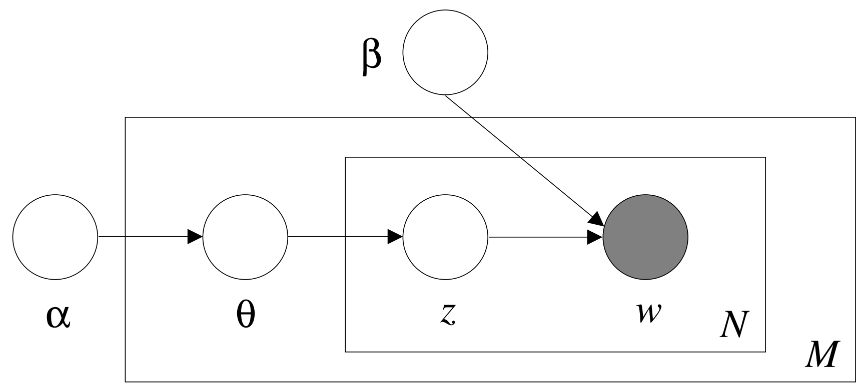
$$
\begin{aligned}
h_0(\alpha_{0k}) &= \gamma^T x_k \\
h(\theta_k) &= \theta^T x_k \\
G_k(\alpha_{0k}, \theta_k) &= \mathrm{DP}(\alpha_{0k} G_0(\theta_k)) \\
y \,|\, G &\sim \prod_{i=1}^{n} G_{k(i)}(y_i),
\end{aligned}
$$

  where x is the covariate (assumed discrete), k indexes the levels of the covariate, $k(i)$ is the level corresponding to the ith data point, y_i is the ith response (an integer), and h_0 and h are link functions
- The goal of inference is (often) a posterior for the parameters θ_k (the effect of some treatment); the distributions G_k are nuisance parameters

Applications

Latent Dirichlet Allocation (LDA) Model

(Blei, Ng, & Jordan, 2003)

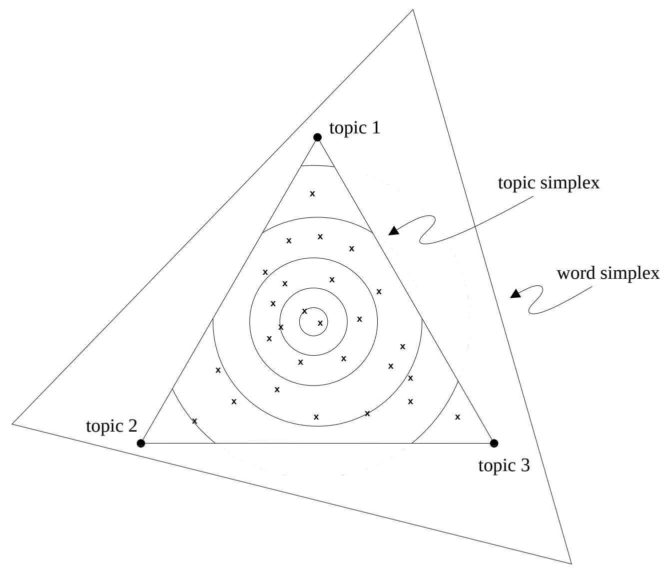


- *Random variables*:
  - A word is represented as a *multinomial* random variable w
  - A topic is represented as a *multinomial* random variable z
  - A document is represented as a *Dirichlet* random variable θ

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

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

The Number of Topics in LDA

- Most applications of LDA have picked the number of topics via some form of model selection method; can we use DPs instead?

- *Problem*: LDA is an "admixture model"—the mixture indicator isn't selected once per document, but rather once per word within a document
  - i.e., each document is its own clustering problem
  - we have *multiple* clustering problems—one per document
  - we want topics from one document to transfer to other documents

- This is exactly the setup for the hierarchical Dirichlet process

- I.e., the HDP solves the problem of choosing the number of topics for LDA (by integration)

Taking the Hierarchy to Another Level

- The HDP formalism makes sense for an arbitrary number of levels of the hierarchy

- At each level, resample from the base measure and pass the resulting atoms as a base measure to the children

- This is practically quite useful
  - e.g., LDA when there are multiple corpora (want to transfer topics among documents and among corpora)
  - e.g., hidden Markov models with multiple speakers

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

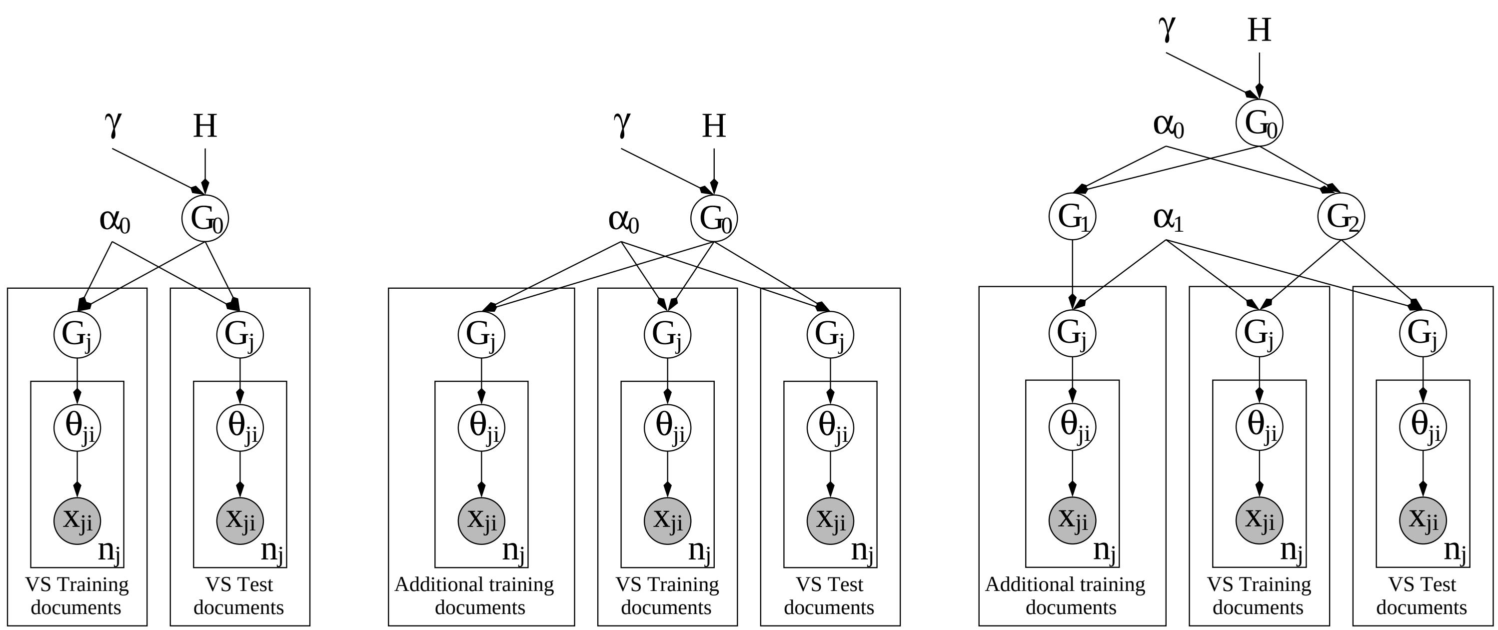
- each article is represented as a mixture model (over words in the vocabulary)
- an HDP is used to discover and share clusters ("topics") among articles within each section
- want to examine relationships among the sections

Models

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

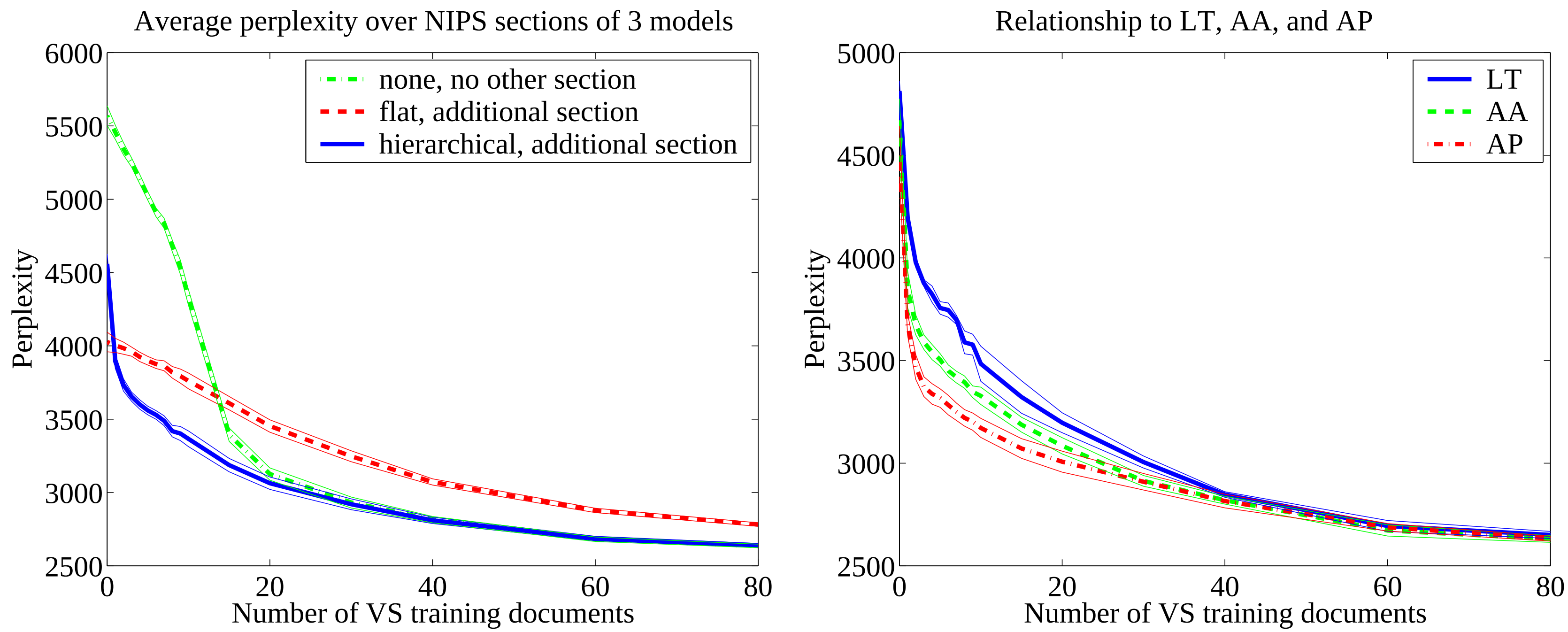
- Models:

“none”	a separate HDP for each section
“flat”	a single HDP for all sections
“hierarchical”	a linked set of HDPs



“none” *“flat”* *“hierarchical”*

Results



- 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.

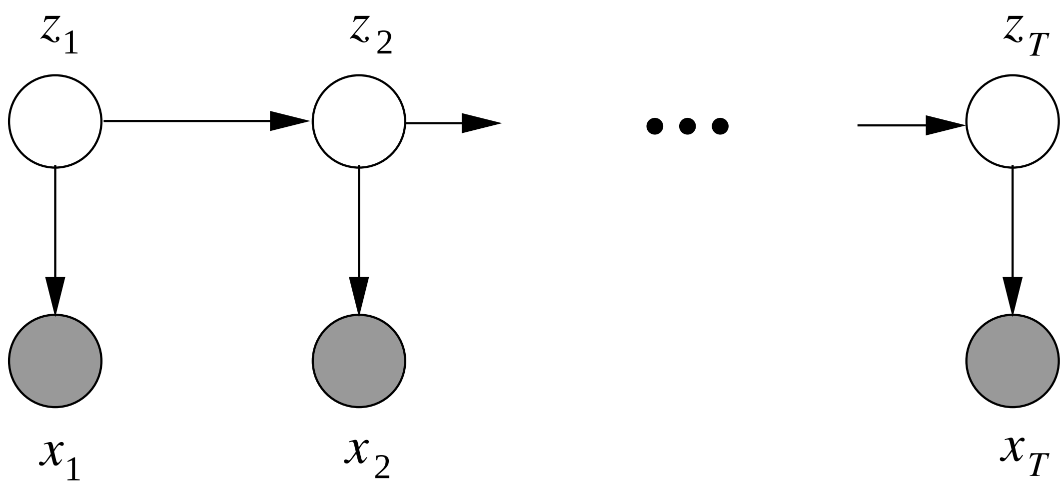
- Right: Average perplexity of test VS documents given LT, AA and AP documents respectively using HDP model, averaged over 5 runs.

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

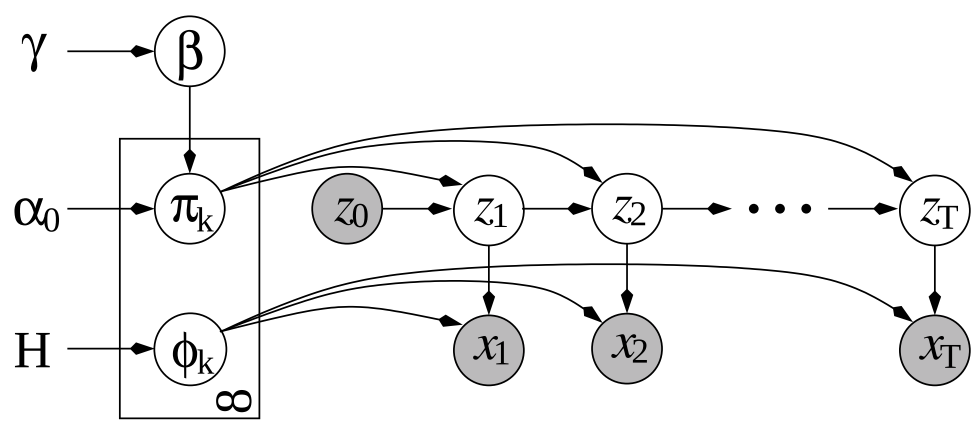


- The hidden Markov model—a dynamic variant of a mixture model
  - each row of the transition matrix is a set of current-state-dependent mixing proportions for the choice of the next state

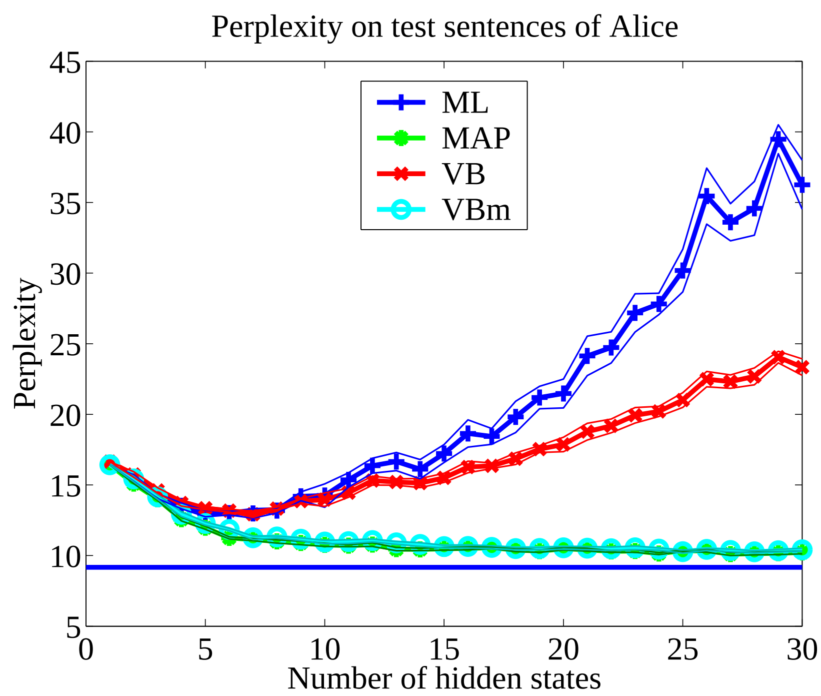
- The "infinite hidden Markov model" (iHMM)—an HMM with an unbounded number of states (Beal, Ghahramani, Rasmussen, 2002)

Hierarchical Dirichlet Process HMMs

- It is straightforward to treat countable state spaces—and in particular the iHMM—within the HDP framework
- An HMM can be viewed as a set of mixture models—one mixture model for each value of the "current state"
- When a new state arises, the HDP machinery "shares" this new state among all of the current states
  - states are dishes in the CRF; they are shared among the restaurants



Alice in Wonderland

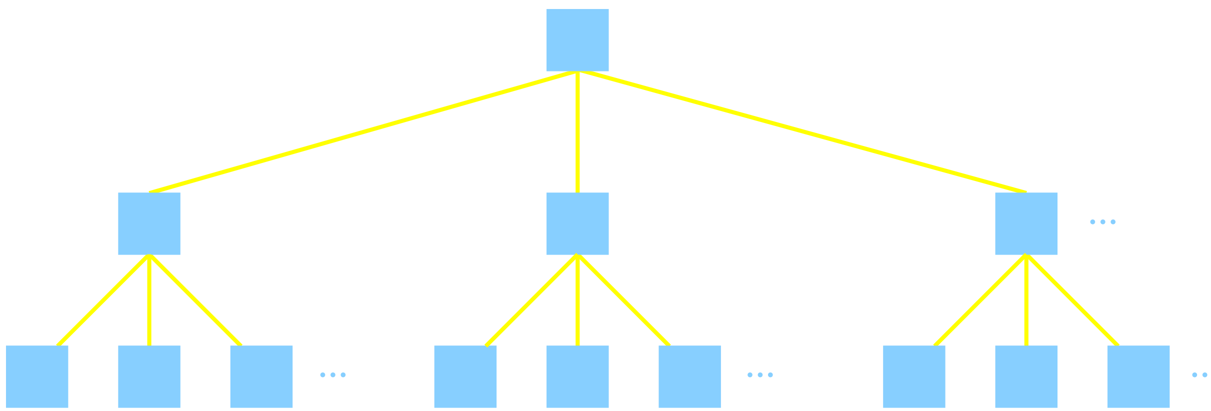


- Perplexity of test sentences taken from Lewis Carroll's *Alice in Wonderland*

Hierarchical Topic Models

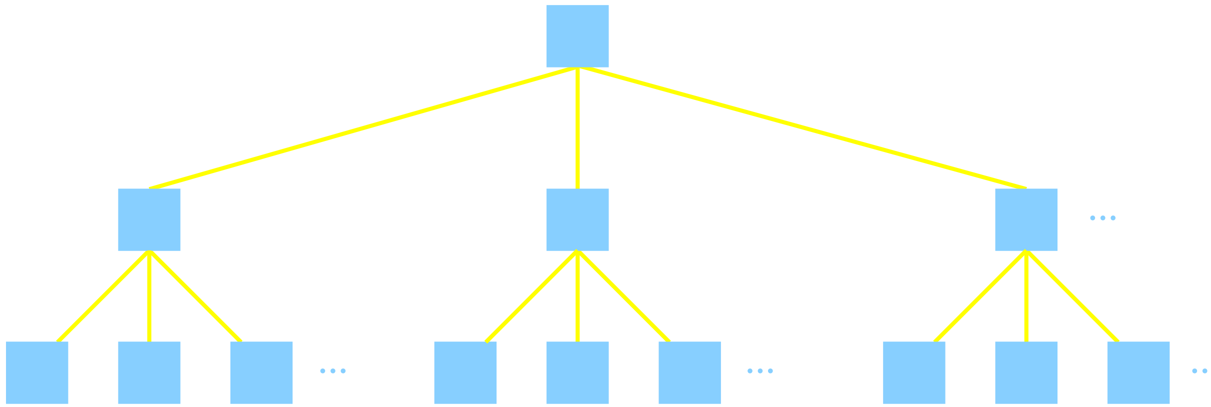
- Topic models such as LDA are "flat"—the same set of topics are available to all documents

- A hierarchical topic model organizes topics in a tree (Hofmann, 2000)
  - note that the word "hierarchical" is not being used in its Bayesian sense here



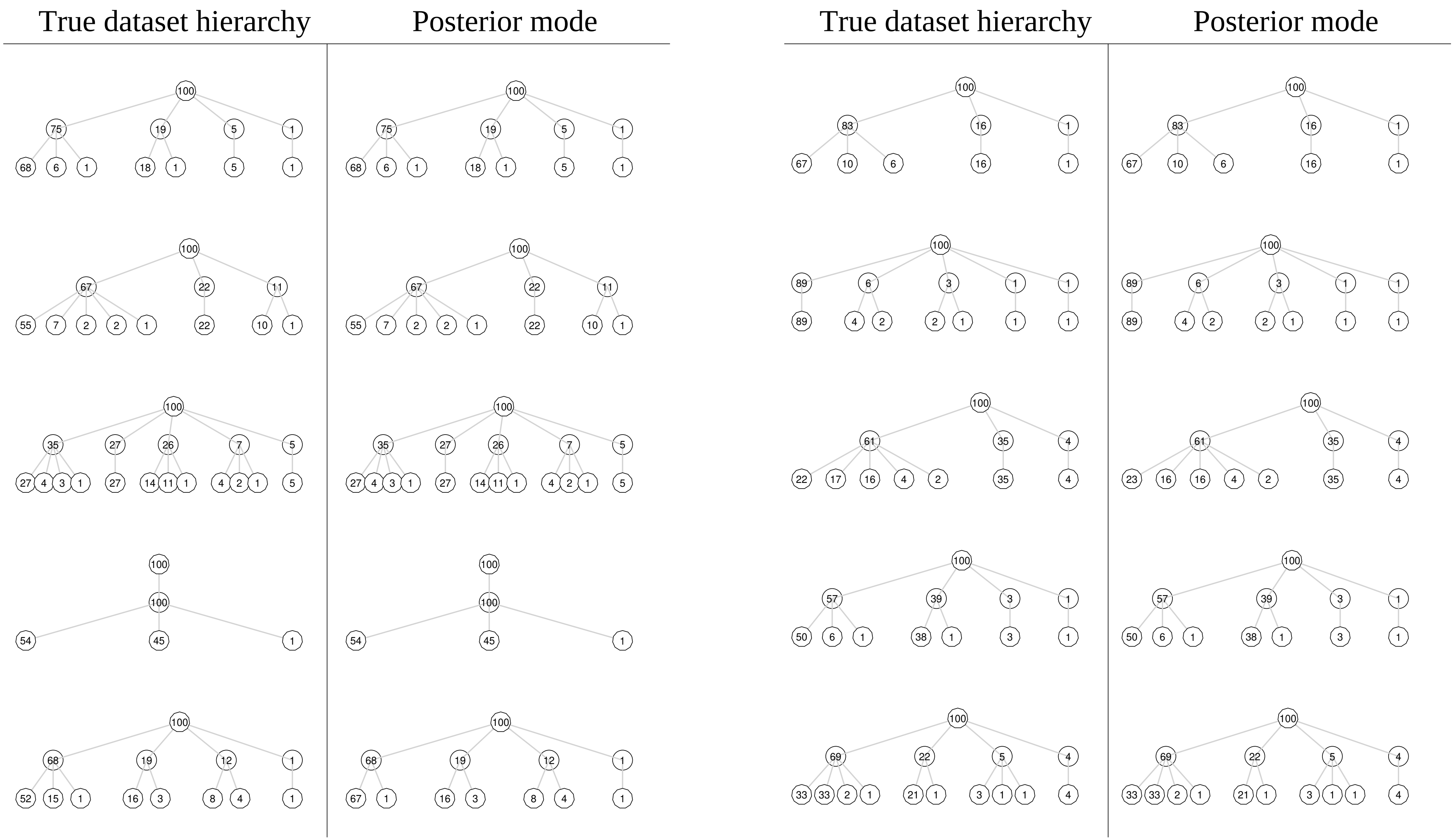
CRP-Based Hierarchical Topic Models

(Blei, et al., 2004)

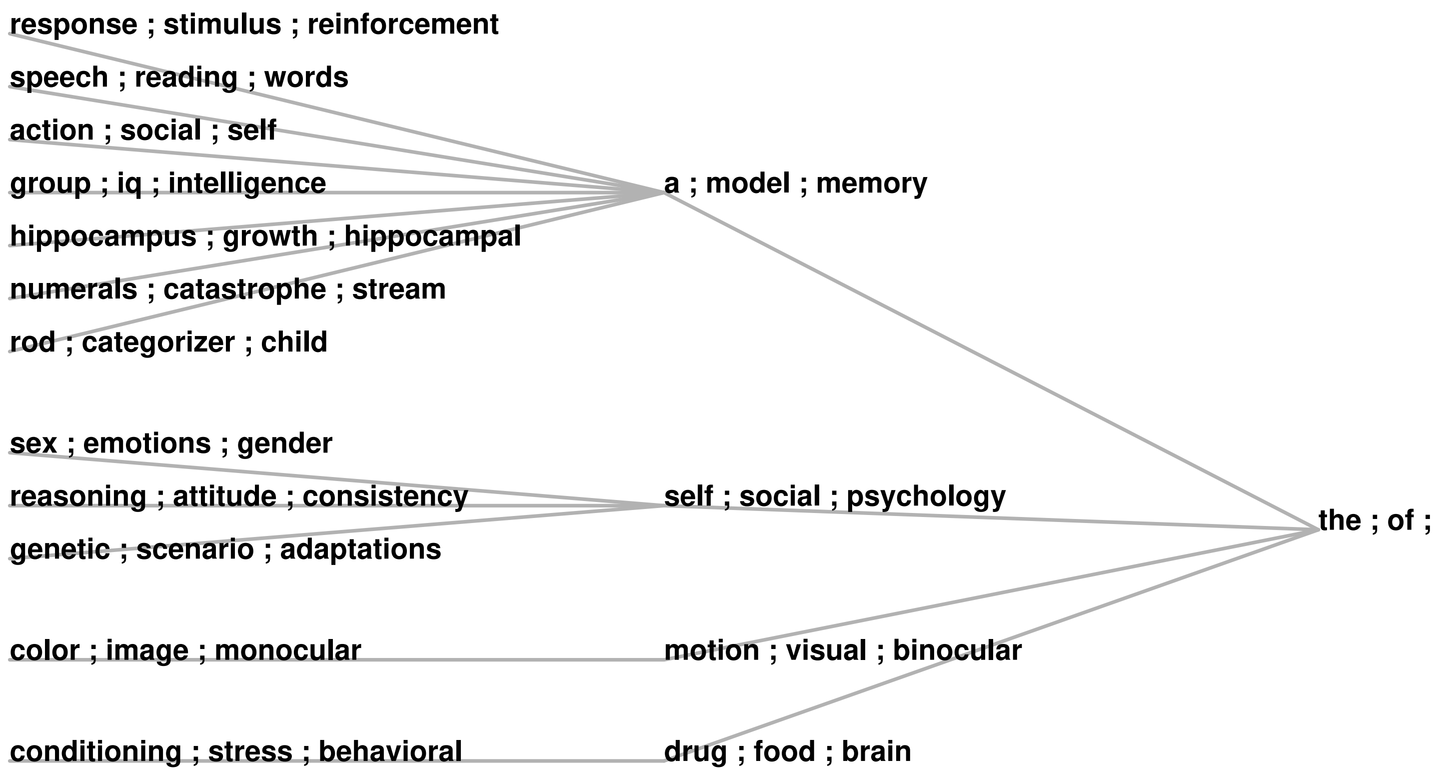


- Each node in the tree is a Chinese restaurant
- Each table in every restaurant has an associated distribution on words (a "topic") drawn from a prior
- Sitting at a table in a given restaurant also selects an outgoing branch, which provides access to further restaurants and further topics
  - we hope for more specialized topics to emerge as we descend the tree

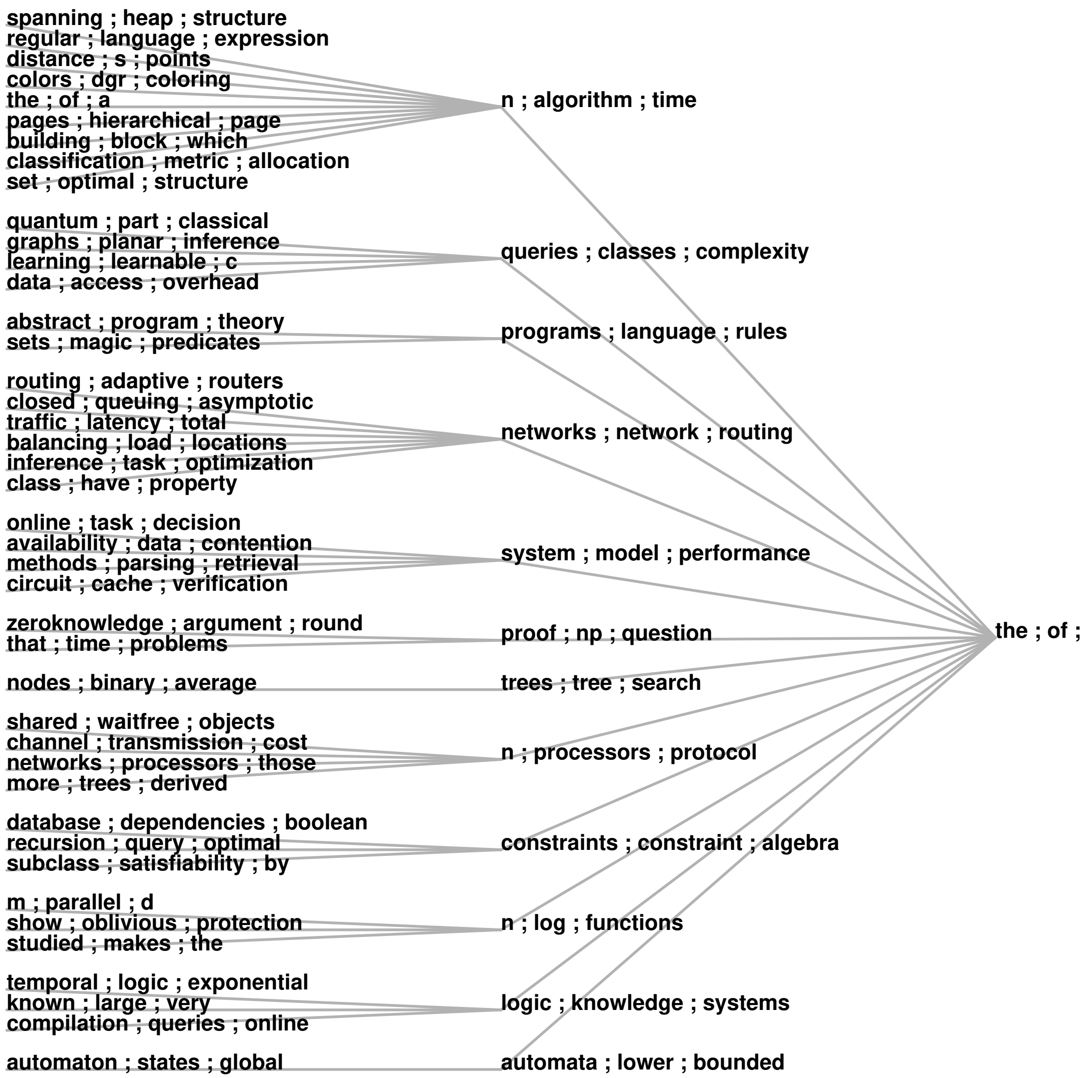
Maximum A Posteriori Hierarchy



Topic Hierarchy from *Psychology Today*



Topic Hierarchy from *JACM*



Further directions in nonparametric Bayes

Tailfree Processes

(Freedman; Fabius)

- Given a sample space Ω, let $\{\rho_i\}$ be a nested sequence of countable partitions of Ω
  - this defines an infinite tree in which the nodes are subsets of Ω
- At each level of the tree, define variables $\{W_{i,A} : A \in \rho_i\}$—independent across levels.
- For any subset $B_n \in \rho_n$, define:

$$P(B_n) = \prod_{i=1}^{n} W_{i,B_i}$$

where $B_n \subseteq B_{n-1} \subseteq \cdots \subseteq B_1$ denotes the path in the tree

Tailfree Processes (cont)

- Under some regularity conditions on the $\{W_{i,A}\}$, this defines a random measure known as a *tailfree process*

- *Example*: Can place a distribution on the class of distributions symmetric around a point by choosing partitions that are symmetric around that point and enforcing symmetry of the $\{W_{i,A}\}$

- *Theorem*: A tailfree prior induces a tailfree posterior

- Dirichlet processes are a special case of tailfree processes; indeed:

- *Theorem*: A Dirichlet process is tailfree with respect to every sequence of partitions

- In general, tailfree processes are not discrete; indeed, conditions can be given that make them absolutely continuous wrt Lebesgue measure

Neutral-to-the-Right Processes

(Doksum; Ferguson; Hjort; Walker & Muliere)

- Many applications (e.g., survival analysis) require priors on distribution functions on $(0, \infty)$

- A random distribution function $F(t)$ on $(0, \infty)$ is *neutral to the right* (NTR) if for every m and $0 < t_1 < \cdots < t_m$ there exist independent random variables $\{V_i\}$ such that

$$\mathbb{P}(1 - F(t_1), 1 - F(t_2), \ldots, 1 - F(t_n)) = \mathbb{P}(V_1, V_1V_2, \ldots, \prod_{j=1}^{m} V_j)$$

- *Theorem*: F is NTR iff $Z(t) = -\log[1 - F(t)]$ is a Lévy process

- *Theorem*: A NTR prior induces an NTR posterior

- Example: Beta-Stacy process, for which a generalized urn model is known

Polya Trees

(Ferguson; Lavine; Mauldin; Sudderth & Williams)

- A special case of tailfree processes in which:
  - the tree has k branches at every non-terminal
  - given a node $A \in \rho_n$, and letting $\mathrm{ch}(A)$ denote the children of A, the distribution of the variables $W_{i,\mathrm{ch}(A)}$ is Dirichlet (independently for different $A \in \rho_n$),

- Again, Dirichlet processes are a special case, but Polya tree distributions are more general—they can be absolutely continuous wrt Lebesgue measure

- Posterior updating: standard multinomial-Dirichlet updating up the tree (the posterior is a Polya tree distribution)

- Quantile pyramids: an alternative in which the tree lives on $(0, 1)$, which induces random quantiles

Dirichlet Diffusion Trees

(Neal, 2001)

- A random measure obtained from a random branching process
- Given a sample space $\mathcal{X}$, define a Brownian path $Z(t) : t \in (0,1)$ on $\mathcal{X}$
  - an observable data point is located at $Z(1)$
- Successive Brownian paths begin by following this same path, but branch off according to a hazard function
  - when arriving at a previous branch point, select among the branches with probability proportional to the number of data that have previously selected that branch (urn-like behavior)
  - the hazard function is inversely proportional to the number of data that have followed that path (yields clumpiness)
- Particular choices of hazard functions can yield non-discrete measures

General Stick-Breaking Processes

(e.g., Ishwaran & James, 2001)

- Recall that the stick-breaking construction is based on $\beta_k \sim \mathrm{Beta}(1, \alpha_0)$
- We can retain exchangeability while using more general definitions of the β_k variables
- E.g., the *Pitman-Yor process* $\mathrm{PY}(a, b)$ defines:

$$\beta_k \sim \mathrm{Beta}(1 - a, b + ka) \qquad\qquad k = 1, 2, \ldots$$

- Mass is taken away from large values of the stick lengths and placed on smaller values; this yields a longer tail
- A Chinese restaurant characterization of marginals is still available
- Teh (2005) notes the close relationship of this process to smoothing rules in language modeling

Some theoretical and philosophical considerations

Statistical Inference

- Just writing down probability distributions and turning a Bayesian crank isn't all there is to statistical inference
  - don't be misled by optimality claims; in real-life problems, models are usually cartoons

- One is somewhat safe in the parametric world, in that a poor prior can always be overcome by data (e.g., Doob)
  - but still, one should always do a posteriori checking

- One isn't safe in the nonparametric world, where Bayesian inferences can be quite wrong

- It's hard to do a posteriori checking when one is "quite wrong," so one needs some help from a priori analysis, aka, frequentist evaluation

Consistency

(e.g.; Diaconis & Freedman; Doss; Barron; Ghosal, Ghosh & Ramamoorthi; Ghosal & van der Vaart; Wasserman; Zhao)

- Does the posterior concentrate in neighborhoods of the true underlying nonparametric distribution as the number of data points goes to infinity?
  - not in general, and not even for the Dirichlet process
  - e.g., a Dirichlet process with a Student t base measure and a Gaussian location model (a setting in which consistent frequentist estimates exist)

- Negative results for the weak neighborhoods; positive results for Hellinger and Kullback-Leibler

- Robins-Ritov paradox (Bayes vs. Horvitz-Thompson estimator)

Consistency

- "Unfortunately, in high-dimensional problems, arbitrary deatils of the prior can really matter; indeed, the prior can swamp the data, no matter how much data you have. That is what our examples suggest, and that is why we advise against the mechanical use of Bayesian nonparametric techniques" (Diaconis & Freedman, 1986).

Conclusions

- Bayesians can do nonparametrics, by putting measures on measures
- Infinite-dimensional problems are tricky to work with, particularly the (unavoidable) uncountably infinite-dimensional case
  - it's not enough to just write out probabilities and compute
- There is a healthy and large literature on these problems; read the literature!
- Consider attending the biannual Bayesian nonparametrics workshop: type "Bayesian nonparametrics workshop" into Google
  - not to be confused with the NIPS Bayesian nonparametrics workshop, which you might also consider attending
- For more details on this talk: http://www.cs.berkeley.edu/~jordan

References

Aldous, D. (1985). Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII–1983*, pages 1–198. Springer, Berlin.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics*, 2(6):1152–1174.

Beal, M. J., Ghahramani, Z., and Rasmussen, C. E. (2002). The infinite hidden Markov model. In Dietterich, T. G., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems*, volume 14, pages 577–584, Cambridge, MA. MIT Press.

Berry, D. A. and Christensen, R. (1979). Empirical Bayes estimation of a binomial parameter via mixtures of Dirichlet processes. *Annals of Statistics*, 7:558–568.

Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *Annals of Statistics*, 1:353–355.

Blei, D. M. and Jordan, M. I. (2005). Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–144.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Blei, D. M., Griffiths, T., Jordan, M. I., and Tenenbaum, J. (2004). Hierarchical topic models and the nested Chinese restaurant process. In Thrun, S., Saul, L., and Schölkopf, B., editors, *Advances in Neural Information Processing*, volume 16, Cambridge, MA: MIT Press.

Brown, E. R. and Ibrahim, J. G. (2003). A Bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics*, 59:221–228.

Brunner, L. J. (1995). Bayesian linear regression with error terms that have symmetric unimodal densities. *Journal of Nonparametric Statistics*, 4:335–348.

Bush, C. and MacEachern, S. N. (1996). A semi-parametric Bayesian model for randomised blocked designs. *Biometrika*, 83:175–185.

Carota, C. and Parmigiani, G. (2002). Semiparametric regression for count data. *Biometrika*, 89(2):265–281.

Cifarelli, D. and Regazzini, E. (1978). Problemi statistici non parametrici in condizioni di scambiabilità parziale e impiego di medie associative. Technical report, Quaderni Istituto Matematica Finanziaria dell'Università di Torino.

De Iorio, M., Müller, P., and Rosner, G. L. (2004). An ANOVA model for dependent random measures. *Journal of the American Statistical Association*, 99(465):205–215.

Diaconis, P. and Freedman, D. (1986a). On inconsistent Bayes estimates of location. *Annals of Statistics*, 14:68–87.

Diaconis, P. and Freedman, D. (1986b). On the consistency of bayes estimates. *Annals of Statistics*, 14:1–67.

Diaconis, P. and Freedman, D. (1993). Nonparametric binary Bayesian regression. *Annals of Statistics*, 21:2108–2137.

Doksum, K. (1974). Tailfree and neutral random probabilities and their posterior distributions. *Annals of Probability*, 2:183–201.

Doss, H. (1985). Bayesian nonparametric estimation of the median; Part i: Computation of the estimates. *Annals of Statistics*, 13:1432–1444.

Doss, H. (1994). Bayesian nonparametric estimation for incomplete data via successive substitution sampling. *Annals of Statistics*, 22(4):1763–1786.

Escobar, M. D. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association*, 89:268–277.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:577–588.

Ewens, W. (1972). The sampling theory of selective neutral alleles. *Theoretical Population Biology*, 3:87–112.

Fabius, J. (1964). Asymptotic behavior of Bayes estimates. *Annals of Mathematical Statistics*, 35:846–856.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1:209–230.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *Annals of Statistics*, 2:615–629.

Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. In Rizvi, H., and Rustagi, J., editors, *Recent Advances in Statistics*. New York: Academic Press.

Fong, D. K. H., Pammer, S. E., Arnold, S. F., and Bolton, G. E. (2002). Reanalyzing ultimatum bargaining—comparing nondecreasing curves without shape constraints. *Journal of Business and Economic Statistics*, 20:423–440.

Freedman, D. (1963). On the asymptotic behavior of Bayes estimates in the discrete case. *Annals of Mathematical Statistics*, 34:1386–1403.

Gelfand, A. E. and Kottas, A. (2002). A computational approach for full nonparametric Bayesian inference under Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 11:289–305.

Gelfand, A. E., Kottas, A., and MacEachern, S. N. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association*, 100:1021–1035.

Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *Annals of Statistics*, 27:143–158.

Ghosal, S. and van der Vaart, A. (2001). Entropies and rates of convergence for Bayes and maximum likelihood estimation for mixture of normal densities. *Annals of Statistics*, 29:1233–1263.

Green, P. and Richardson, S. (2001). Modelling heterogeneity with and without the Dirichlet process. *Scandinavian Journal of Statistics*, 28:355–377.

Guglielmi, A., Holmes, C. C., and Walker, S. G. (2003). Perfect simulation involving functionals of a Dirichlet process. *Journal of Computational and Graphical Statistics*.

Hjort, N. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *Annals of Statistics*, 18:1259–1294.

Ishwaran, H. and James, L. (2003). Generalized weighted Chinese restaurant processes for species sampling mixture models. *Statistica Sinica*, 13:1211–1235.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173.

Ishwaran, H. and James, L. F. (2002). Approximate Dirichlet process computing in finite normal mixtures: Smoothing and prior information. *Journal of Computational and Graphical Statistics*, 11:508–532.

Ishwaran, H. and James, L. F. (2004). Computational methods for multiplicative intensity models using weighted gamma processes: Proportional hazards, marked point processes and panel count data. *Journal of the American Statistical Association*, 99:175–190.

Ishwaran, H. and Zarepour, M. (2000). Markov chaine Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika*, 87:371–390.

Ishwaran, H. and Zarepour, M. (2002a). Dirichlet prior sieves in finite normal mixtures. *Statistica Sinica*, 12:941–963.

Ishwaran, H. and Zarepour, M. (2002b). Exact and approximate sum-representations for the Dirichlet process. *Canadian Journal of Statistics*, 30:269–283.

Jain, S. and Neal, R. (2004). A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, 13(1):158–182.

Kingman, J. F. C. (1974). Random discrete distributions. *Journal of the Royal Statistical Society*, 37:1–22.

Kleinman, K. P. and Ibrahim, J. G. (1998). A semi-parametric Bayesian approach to generalized linear mixed models. *Statistics in Medicine*, 17:2579–2596.

Kottas, A. (2006). Nonparametric Bayesian survival analysis using mixtures of Weibull distributions. *Journal of Statistical Planning and Inference*, 136:578–596.

Lavine, M. (1992). Some aspects of Polya tree distributions for statistical modeling. *Annals of Statistics*, 20:1222–1235.

Liu, J. S. (1996). Nonparametric hierarchical Bayes via sequential imputations. *Annals of Statistics*, 24:911–930.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *Annals of Statistics*, 12(1):351–357.

MacEachern, S. N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics B*, 23:727–741.

MacEachern, S. N. (1998). Computational methods for mixture of Dirichlet process models. In Dey, D., Muller, P., and Sinha, D., editors, *Practical Nonparametric and Semiparametric Bayesian Statistics*, pages 23–44. Springer.

MacEachern, S. N. (1999). Dependent nonparametric processes. In *Proceedings of the Section on Bayesian Statistical Science*. American Statistical Association.

MacEachern, S. N. and Müller, P. (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, 7:223–238.

Mallick, B. K. and Walker, S. G. (1997). Combining information from several experiments with nonparametric priors. *Biometrika*, 84:697–706.

Mauldin, R. D., Sudderth, W. D., and Williams, S. C. (1992). Polya trees and random distributions. *Annals of Statistics*, 20:1203–1221.

McAuliffe, J., Blei, D. M., and Jordan, M. I. (to appear). Nonparametric empirical Bayes for the Dirichlet process mixture model. *Statistics and Computing*.

Mukhopadhyay, S. and Gelfand, A. E. (1997). Dirichlet process mixed generalized linear models. *Journal of the American Statistical Association*, 92:633–639.

Muliere, P. and Petrone, S. (1993). A Bayesian predictive approach to sequential search for an optimal dose: Parametric and nonparametric models. *Journal of the Italian Statistical Society*, 2:349–364.

Muliere, P., Secchi, P., and Walker, S. G. (2000). Urn schemes and reinforced random walks. *Stochastic Processes and their Applications*, 88:59–78.

Muliere, P. and Tardella, S. (1998). Approximating distributions of random functionals of Ferguson-Dirichlet priors. *Canadian Journal of Statistics*, 26:283–297.

Müller, P., Quintana, F., and Rosner, G. (2004). A method for combining inference across related nonparametric Bayesian models. *Journal of the Royal Statistical Society*, 66:735–749.

Neal, R. M. (1992). Bayesian mixture modeling. In *Proceedings of the Workshop on Maximum Entropy and Bayesian Methods of Statistical Analysis*, volume 11, pages 197–211.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9:249–265.

Neal, R. M. (2001). Defining priors for distributions using Dirichlet diffusion trees. Technical Report 0104, Department of Statistics, University of Toronto.

Patil, G. P. and Taillie, C. (1977). Diversity as a concept and its implications for random communities. *Bulletin of the International Statistical Institute*, 47:497–515.

Pitman, J. (1996). Some developments of the Blackwell-MacQueen urn scheme. In Ferguson, T. S., Shapley, L. S., and MacQueen, J. B., editors, *Statistics, Probability and Game Theory*, pages 245–267. Institute of Mathematical Statistics, Hayward, CA.

Pitman, J. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, 25:855–900.

Pitman, J. (2002a). Combinatorial stochastic processes. Technical Report 621, Department of Statistics, University of California at Berkeley. Lecture notes for St. Flour Summer School.

Pitman, J. (2002b). Poisson-dirichlet and GEM invariant distributions for split-and-merge transformations of an interval partition. *Combinatorics, Probability and Computing*, 11:501–514.

Robert, C. (1996). Mixtures of distributions: Inference and estimation. In Gilks, W. R., Richardson, S., and Spiegelhalter, D., editors, *Practical Markov Chain Monte Carlo*. Chapman and Hall.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2004). Hierarchical Dirichlet processes. Technical Report 653, Department of Statistics, University of California at Berkeley.

Tomlinson, G. and Escobar, M. (1999). Analysis of densities. Technical report, University of Toronto.

Walker, S. G. and Muliere, P. (1995). Beta-Stacy processes and a generalisation of the Polya urn scheme. *Annals of Statistics*, 25:1762–1780.

Wasserman, L. (1998). Asymptotic properties of nonparametric Bayesian procedures. In Dey, D., Muller, P., and Sinha, D., editors, *Practical Nonparametric and Semiparametric Bayesian Statistics*, pages 293–304. Springer.

West, M. (1992). Hyperparameter estimation in Dirichlet process mixture models. Technical Report 92-A03, Institute of Statistics and Decision Sciences, Duke University.

West, M., Müller, P., and Escobar, M. (1994). Hierarchical priors and mixture models, with applications in regression and density estimation. In Freeman, P. R. and Smith, A. F. M., editors, *Aspects of Uncertainty*, pages 363–386. John Wiley.

Xing, E. P., Sharan, R., and Jordan, M. I. (2004). Haplotype modeling via the Dirichlet process. *International Conference on Machine Learning*, New York: ACM Press.

Zhao, L. (2000). Bayesian aspects of some nonparametric problems. *Annals of Statistics*, 28:532–552.