Bayesian Nonparametrics for Speech and Signal Processing

Michael I. Jordan
University of California, Berkeley

June 28, 2011

Acknowledgments: Emily Fox, Erik Sudderth, Yee Whye Teh, and Romain Thibaux

Computer Science and Statistics

- Separated in the 40's and 50's, but merging in the 90's and 00’s
- What computer science has done well: data structures and algorithms for manipulating data structures
- What statistics has done well: managing uncertainty and justification of algorithms for making decisions under uncertainty
- Bayesian nonparametrics brings the two threads together
  - stochastic processes for representing flexible data structures

Combinatorial Stochastic Processes

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

Bayesian Nonparametrics

- At the core of Bayesian inference is Bayes theorem:

$$posterior \propto likelihood \times prior$$

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

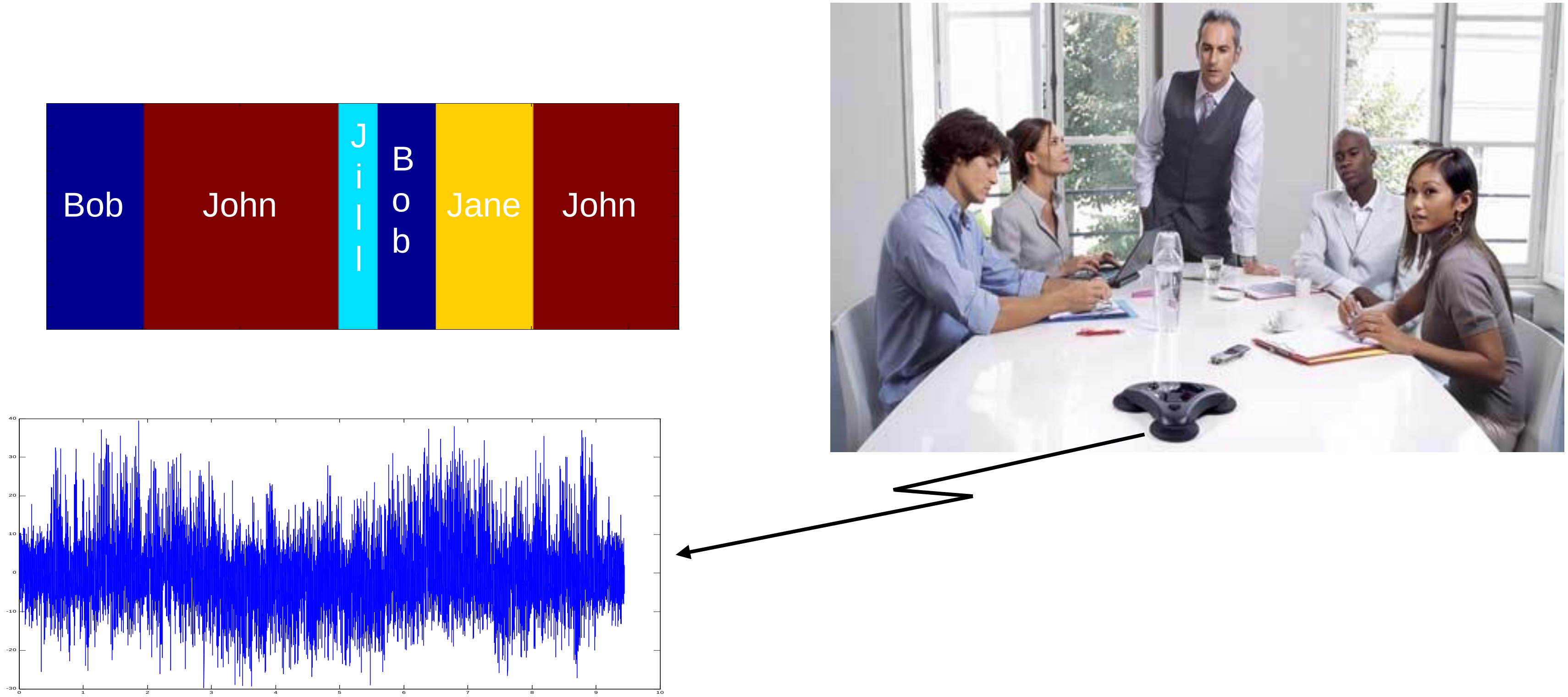
$$p(\theta \,|\, x) \propto p(x \,|\, \theta) p(\theta)$$

- For Bayesian nonparametric models, we let G be a general stochastic process (an “infinite-dimensional random variable”) and write (non-rigorously):

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

- This frees us to work with flexible data structures

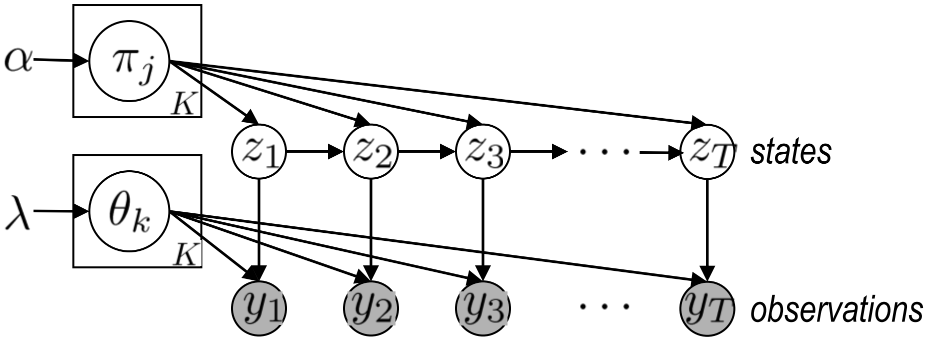
Speaker Diarization



Motion Capture Analysis

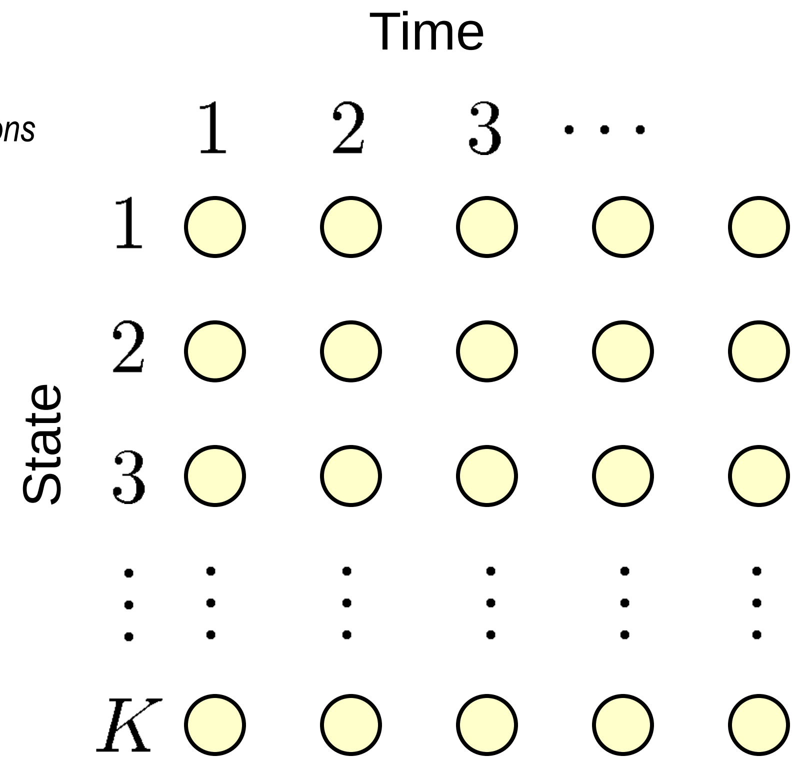
- Goal: Find coherent “behaviors” in the time series that transfer to other time series (e.g., jumping, reaching)

Hidden Markov Models

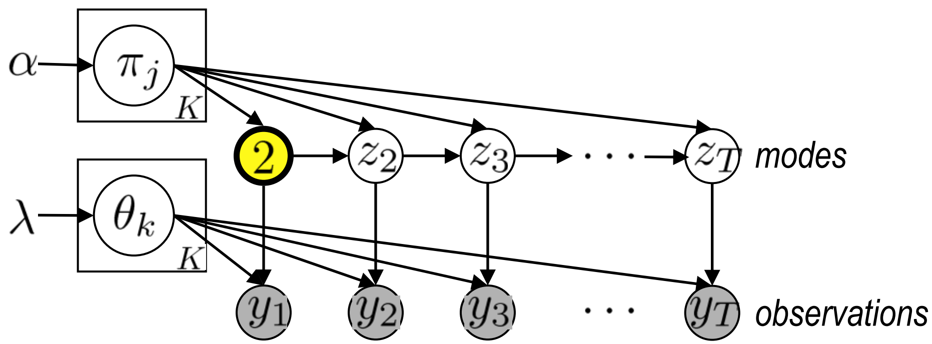


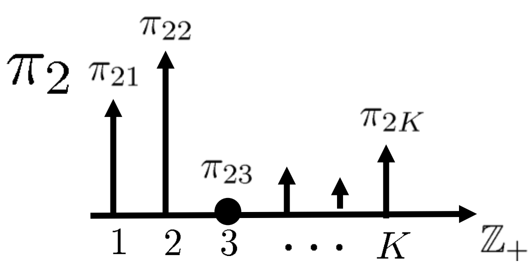
$$
\begin{aligned}
z_t &\sim \pi_{z_{t-1}} \\
y_t &\sim F(\theta_{z_t})
\end{aligned}
$$

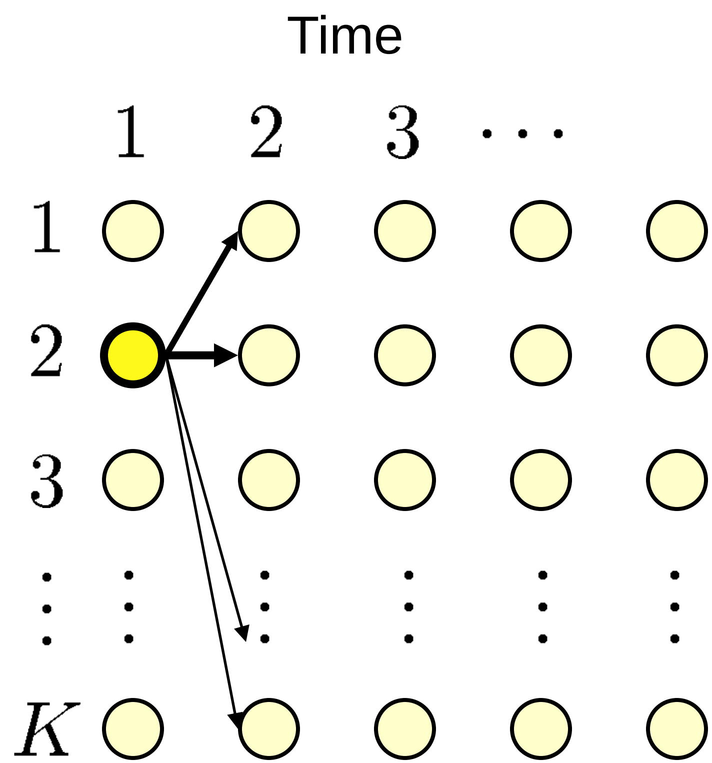
$$
P = \begin{bmatrix} \text{——} & \pi_1 & \text{——} \\ \text{——} & \pi_2 & \text{——} \\ & \vdots & \\ \text{——} & \pi_K & \text{——} \end{bmatrix}
$$



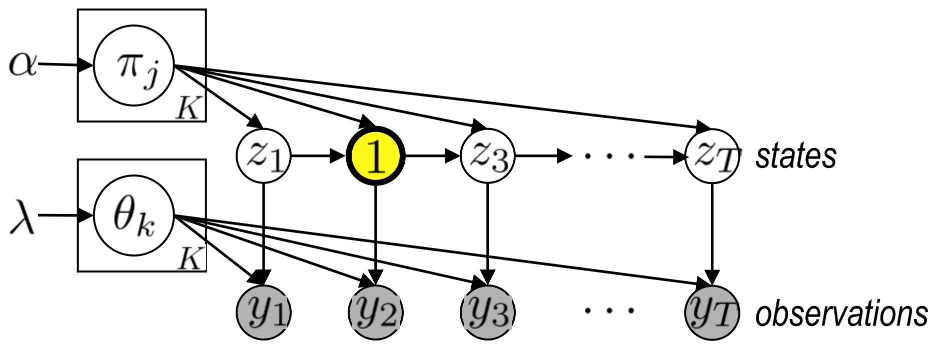
Hidden Markov Models

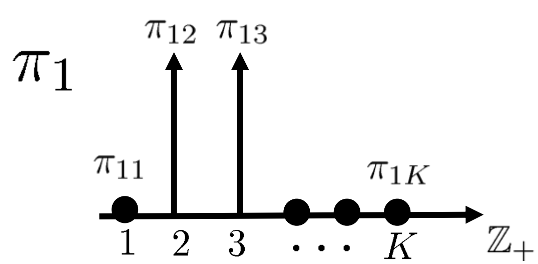


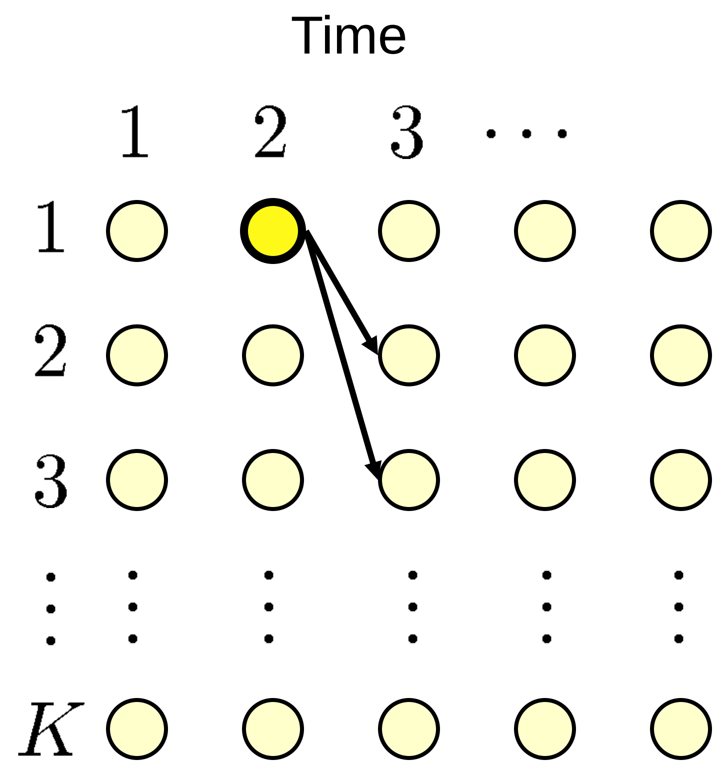




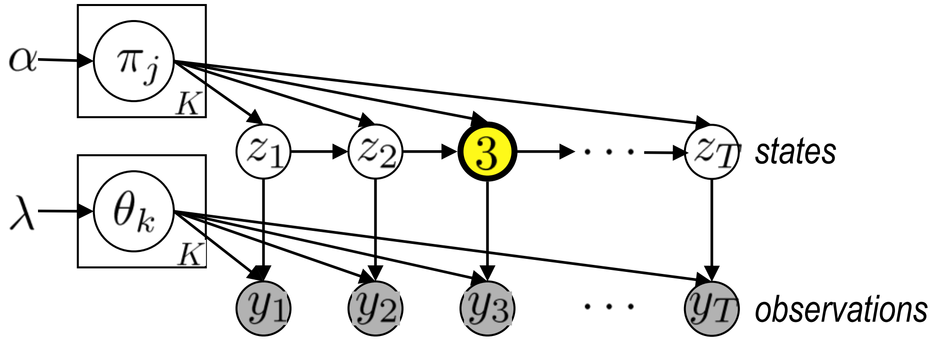
Hidden Markov Models

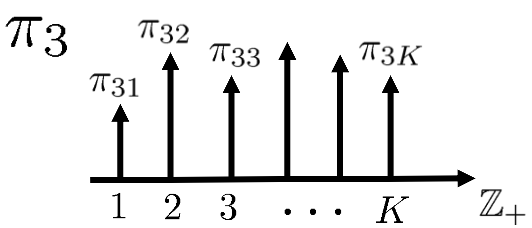


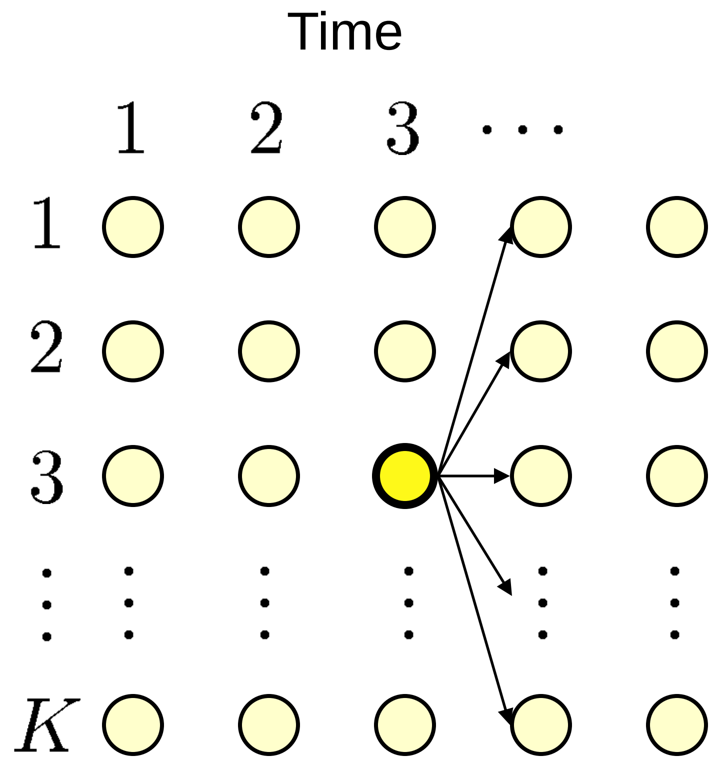




Hidden Markov Models







Issues with HMMs

- How many states should we use?
  - we don’t know the number of speakers a priori
  - we don’t know the number of behaviors a priori
- How can we structure the state space?
  - how to encode the notion that a particular time series makes use of a particular subset of the states?
  - how to share states among time series?

- We’ll develop a Bayesian nonparametric approach to HMMs that solves these problems in a simple and general way

Stick-Breaking

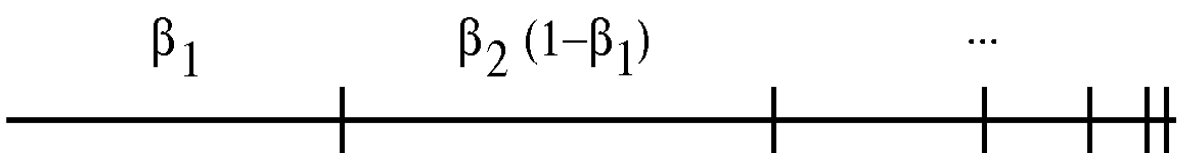
- A general way to obtain distributions on countably infinite spaces
- *The classical example*: Define an infinite sequence of beta random variables:

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

- And then define an infinite random sequence as follows:

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

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



Constructing Random Measures

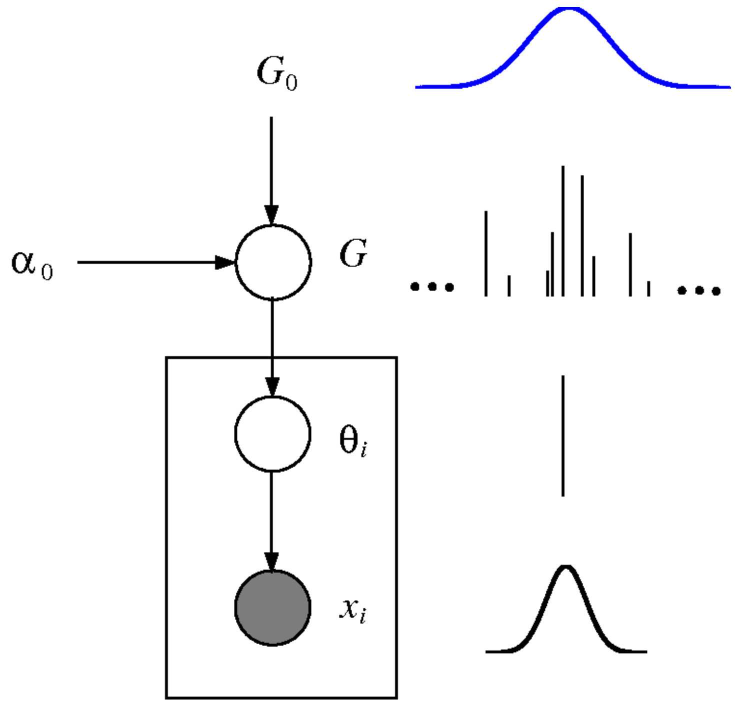
- It's not hard to see that $\sum_{k=1}^{\infty} \pi_k = 1$ (wp1)
- Now define the following object:

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

- where ϕ_k are independent draws from a distribution G_0 on some space
- Because $\sum_{k=1}^{\infty} \pi_k = 1$, G is a probability measure---it is a random measure
- The distribution of G is known as a Dirichlet process:

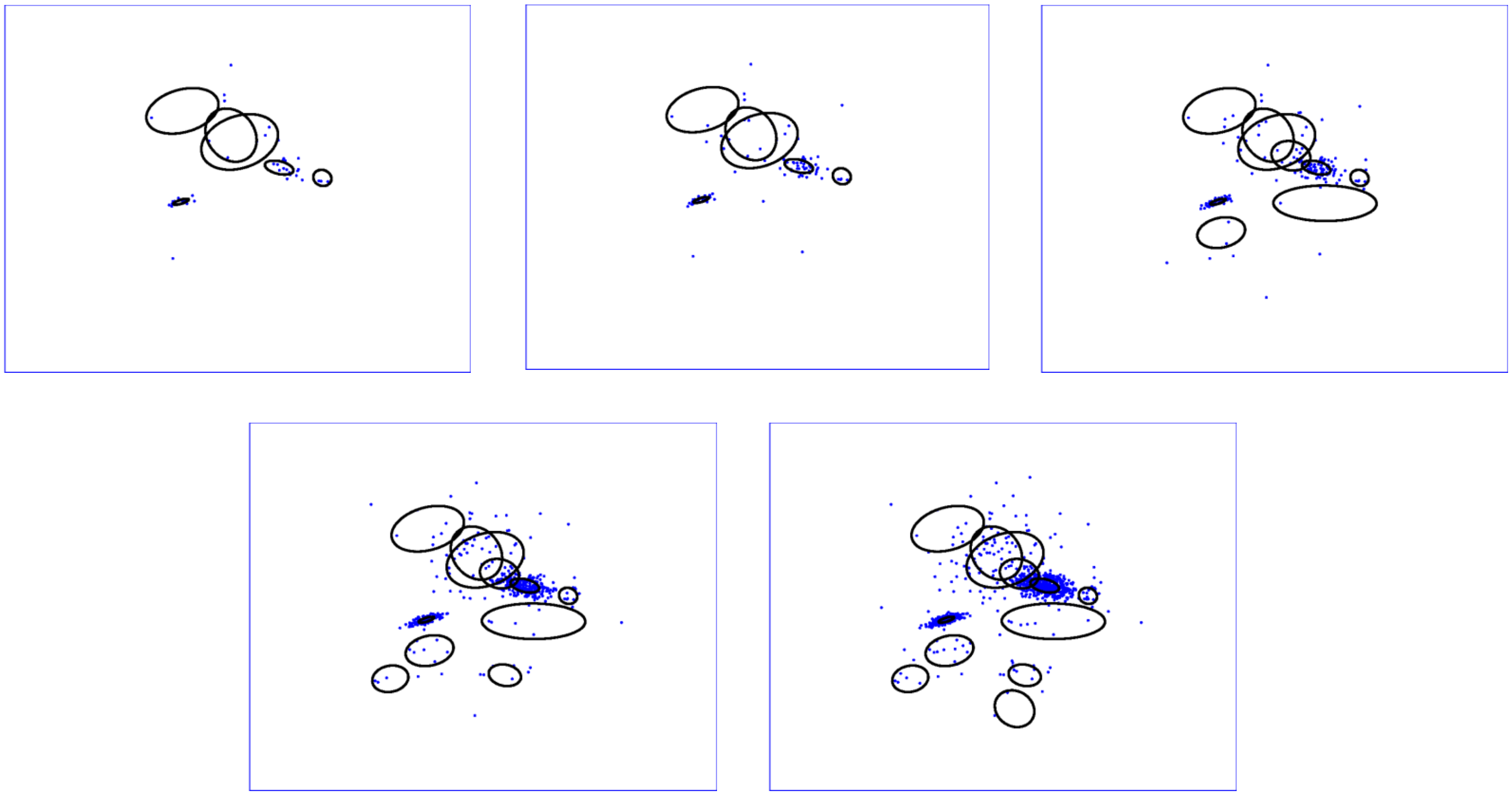
$$G \sim \mathrm{D}P(\alpha_0, G_0)$$

Dirichlet Process Mixture Models



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

CRP Prior, Gaussian Likelihood, Conjugate Prior



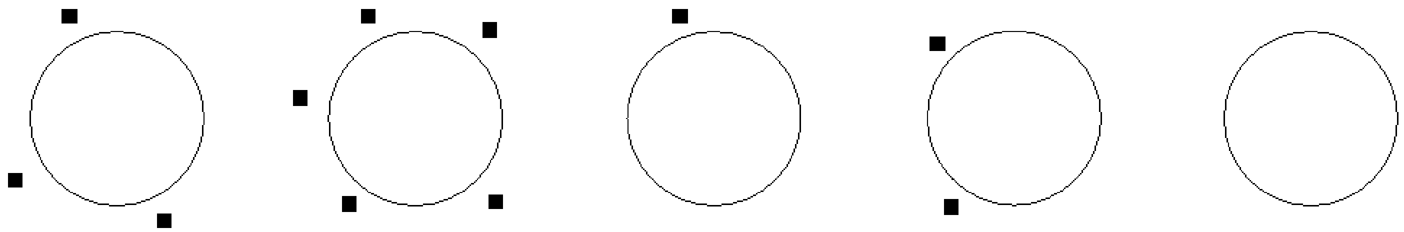
$$\begin{aligned} \phi_k &= (\mu_k, \Sigma_k) \sim N(a,b) \otimes IW(\alpha, \beta) \\ x_i &\sim N(\phi_k) \qquad \text{for a data point } i \text{ sitting at table } k \end{aligned}$$

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:

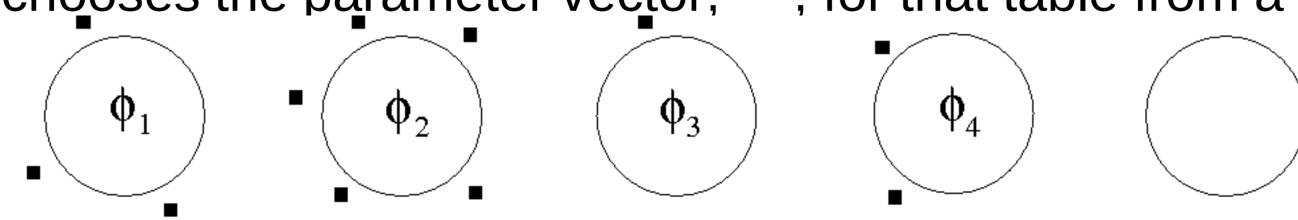
$$\begin{aligned} P(\text{previously occupied table } i \mid \mathcal{F}_{m-1}) &\propto n_i \\ P(\text{the next unoccupied table} \mid \mathcal{F}_{m-1}) &\propto \alpha_0 \end{aligned}$$

  - 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

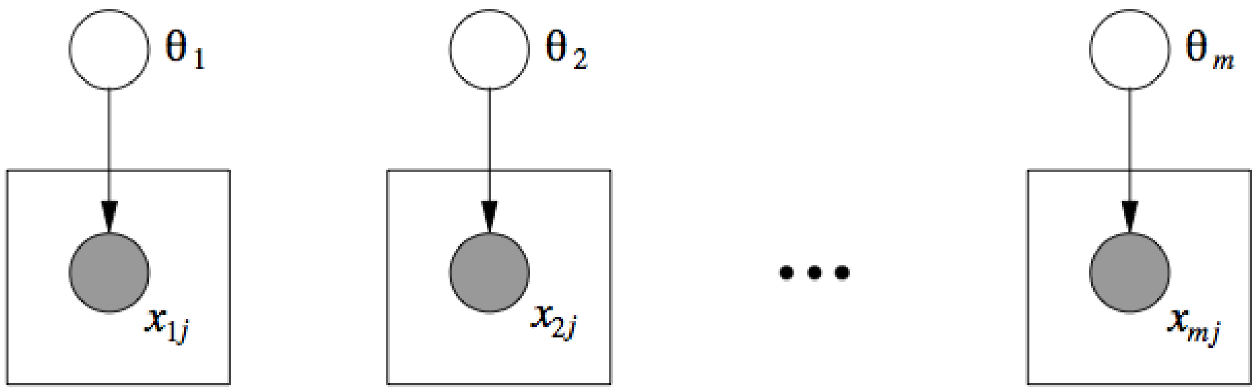
- Data points are customers; tables are mixture components
  - the CRP defines a prior distribution on the partitioning of the data and on the number of tables
- This prior can be completed with:
  - a likelihood---e.g., associate a parameterized probability distribution with each table
  - a prior for the parameters---the ϕ_kst customer to sit at table k G_0 chooses the parameter vector, , for that table from a prior



- So we now have defined a full Bayesian posterior for a mixture model of unbounded cardinality

Multiple Estimation Problems

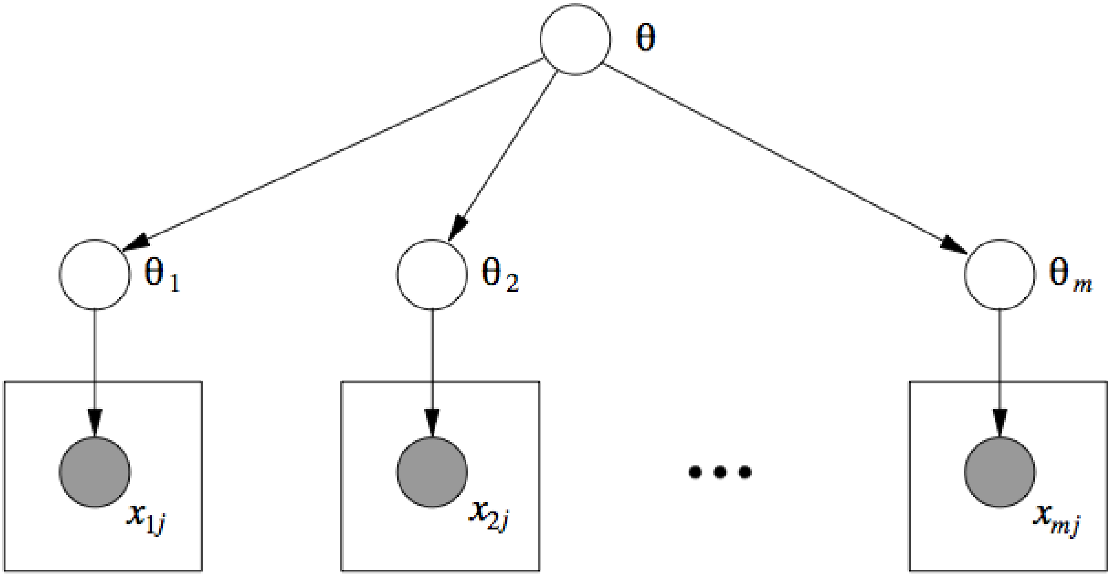
- We often face multiple, related estimation problems
- E.g., multiple Gaussian means: $x_{ij} \sim N(\theta_i, \sigma_i^2)$



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

Hierarchical Bayesian Approach

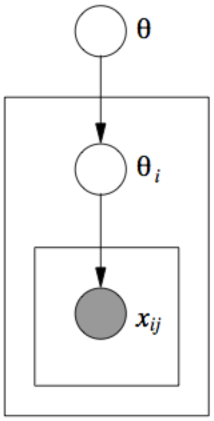
- The Bayesian or empirical Bayesian solution is to view the parameters θ_i as random variables, related via an underlying variable θ



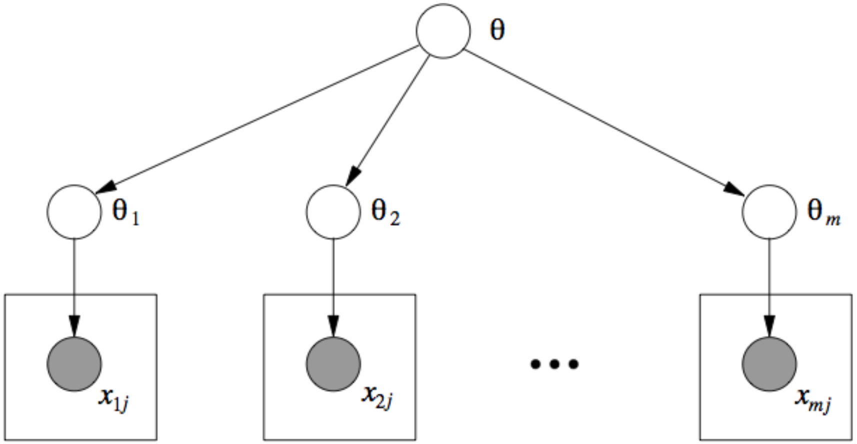
- Given this overall model, posterior inference yields shrinkage---the posterior mean for each θ_i combines data from all of the groups

Hierarchical Modeling

- The plate notation:

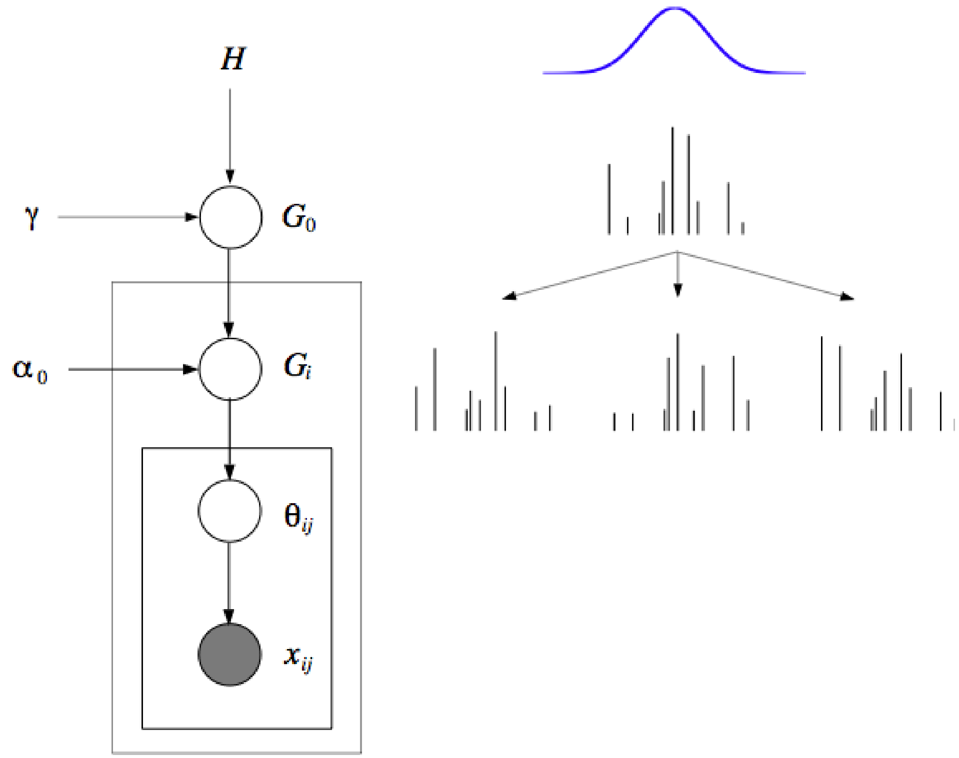


- Equivalent to:



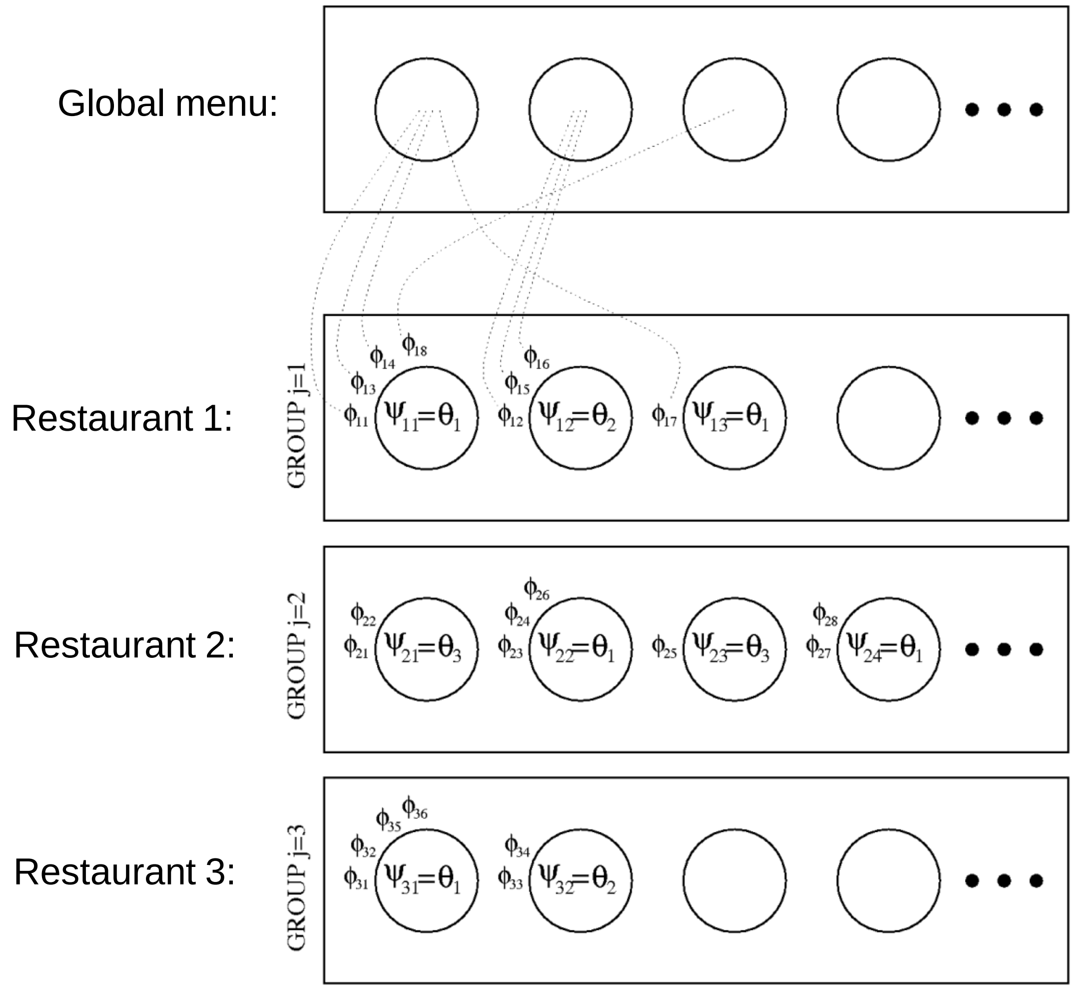
Hierarchical Dirichlet Process Mixtures

(Teh, Jordan, Beal, & Blei, JASA 2006)



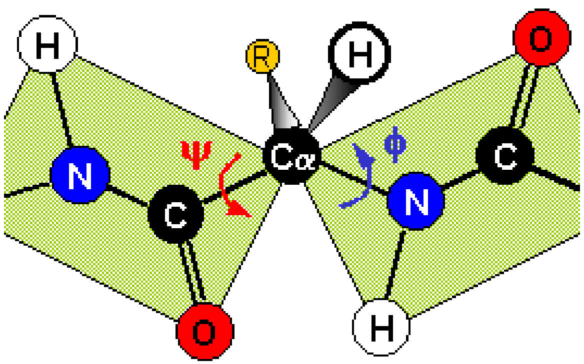
$$
\begin{aligned}
G_0|\gamma, H &\sim DP(\gamma H) \\
G_i|\alpha, G_0 &\sim DP(\alpha_0 G_0) \\
\theta_{ij}|G_i &\sim G_i \\
x_{ij}|\theta_{ij} &\sim F(x_{ij}|\theta_{ij})
\end{aligned}
$$

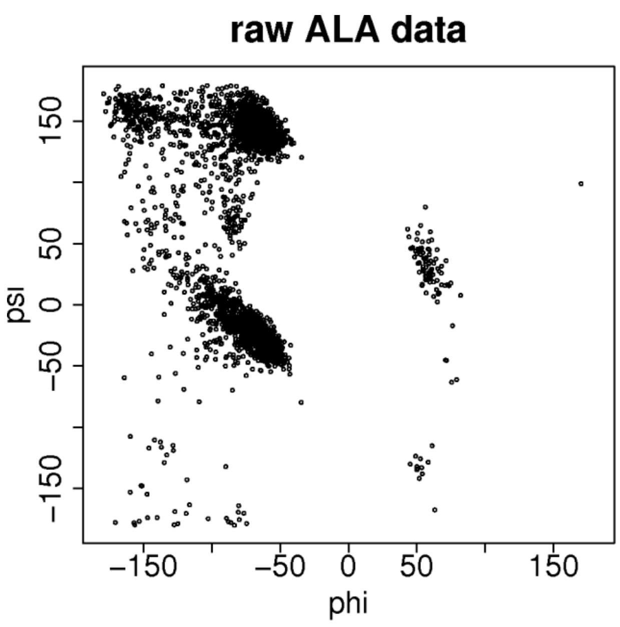
Chinese Restaurant Franchise (CRF)



Application: Protein Modeling

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



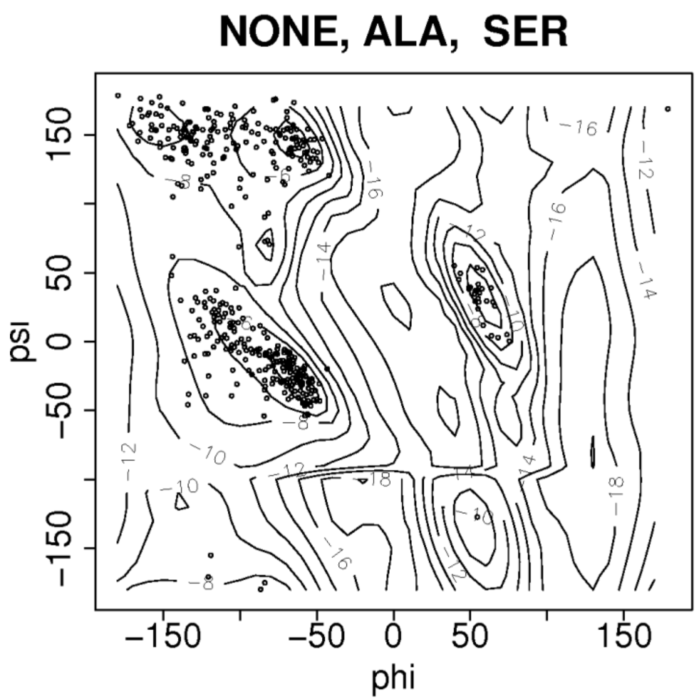


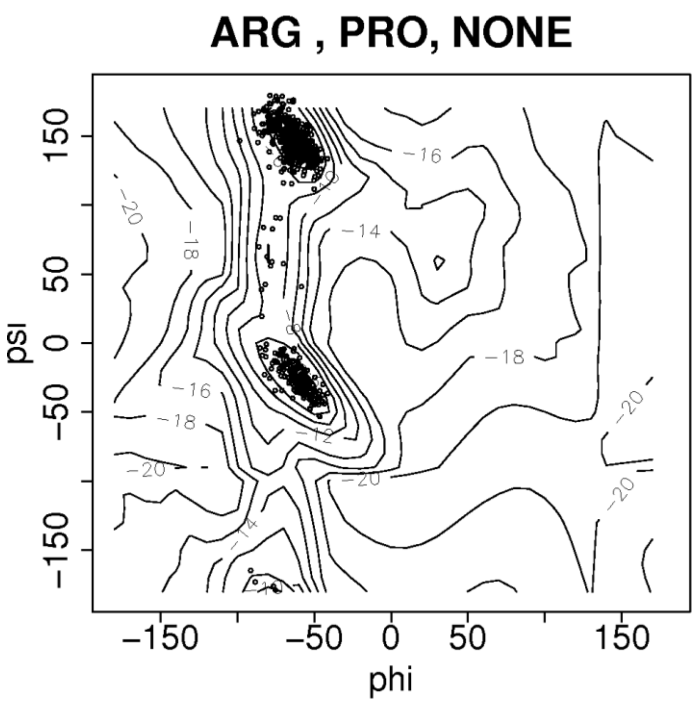
Application: Protein Modeling

- We want to model the density in the Ramachandran diagram to provide an energy term for protein folding algorithms
- We actually have a linked set of Ramachandran diagrams, one for each amino acid neighborhood
- We thus have a *linked set* of density estimation problems

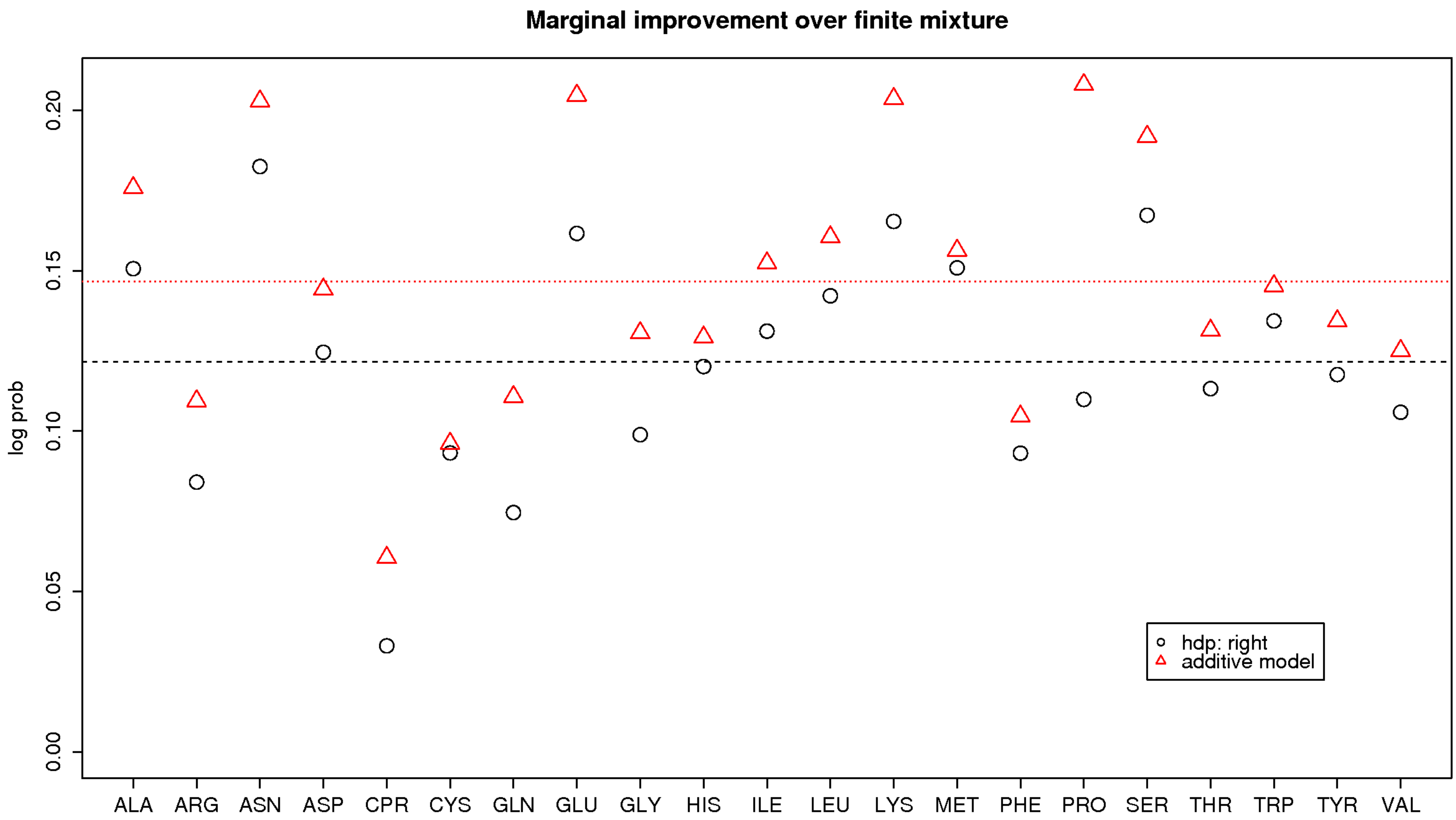
Protein Folding (cont.)

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

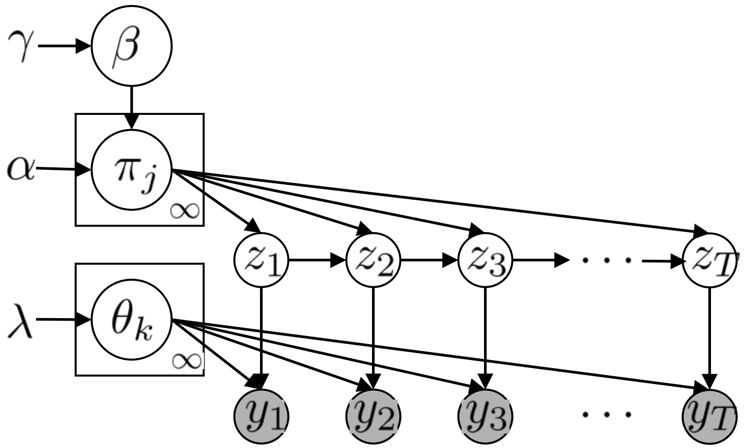


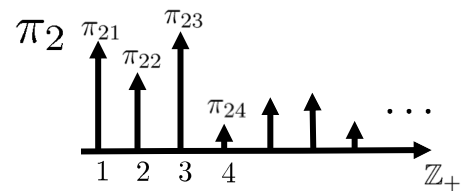


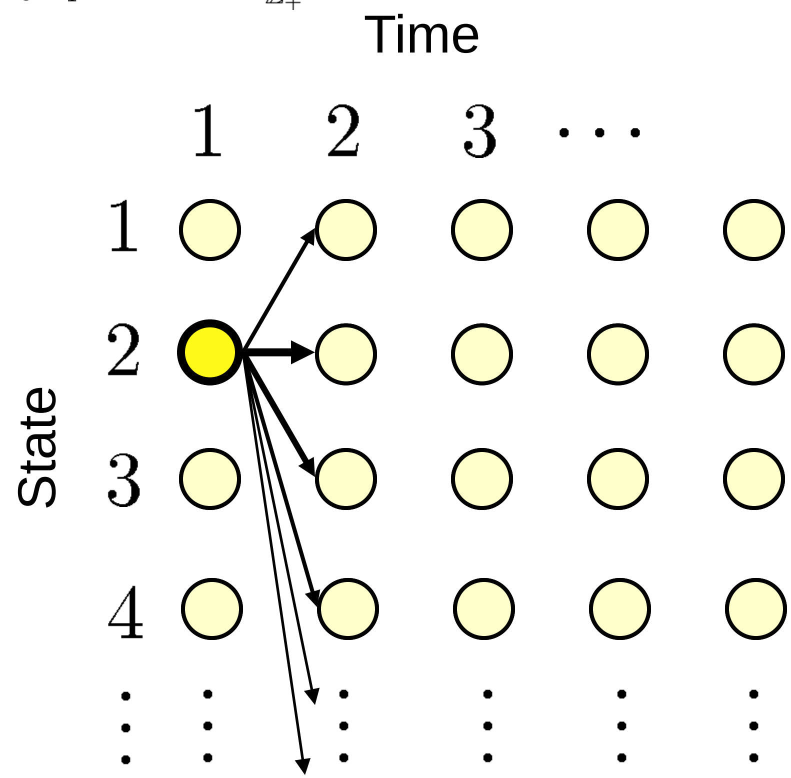
Protein Folding (cont.)



HDP-HMM

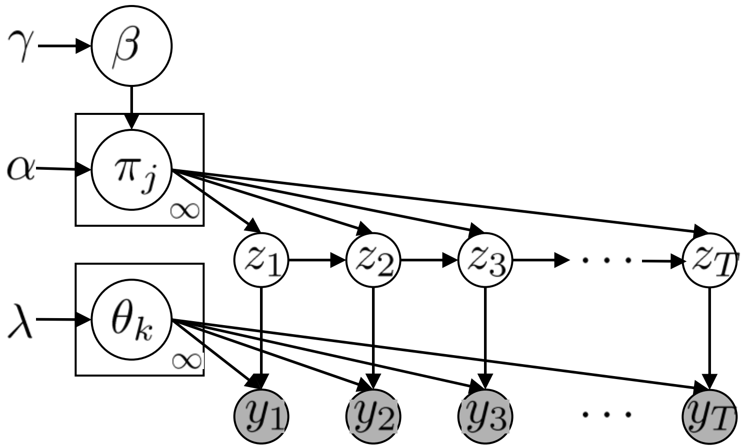


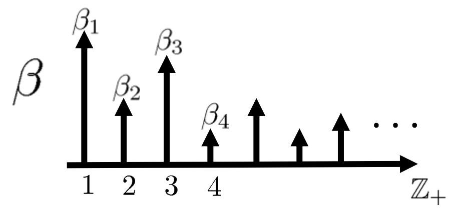


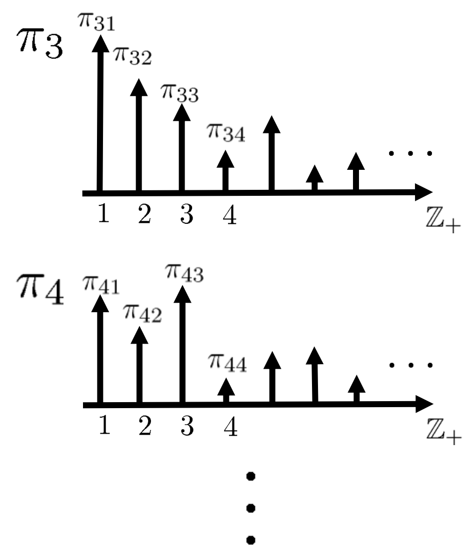


- Dirichlet process:
  - state space of unbounded cardinality
- Hierarchical DP:
  - ties state transition distributions

HDP-HMM







- Average transition distribution:

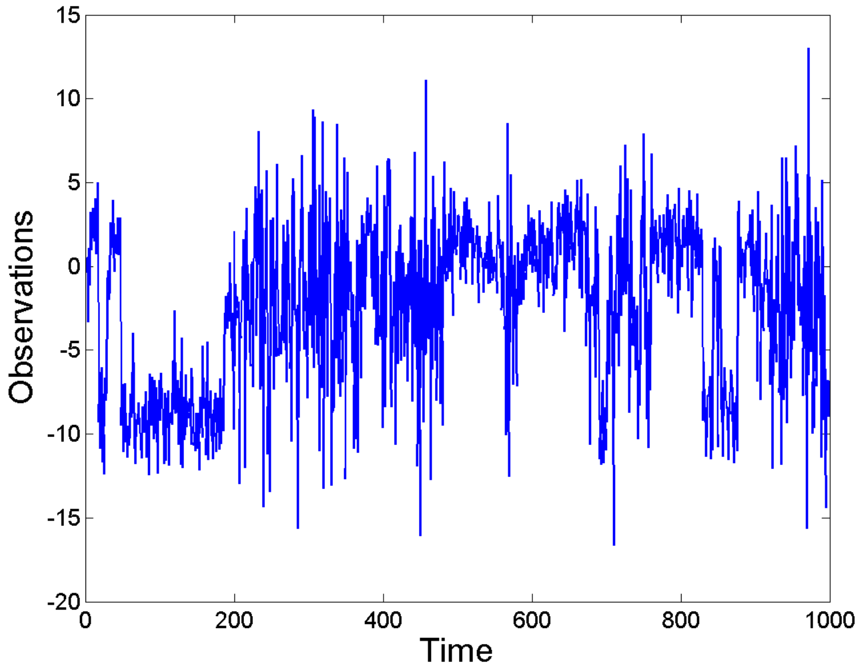
$$\beta \sim \mathrm{GEM}(\gamma)$$

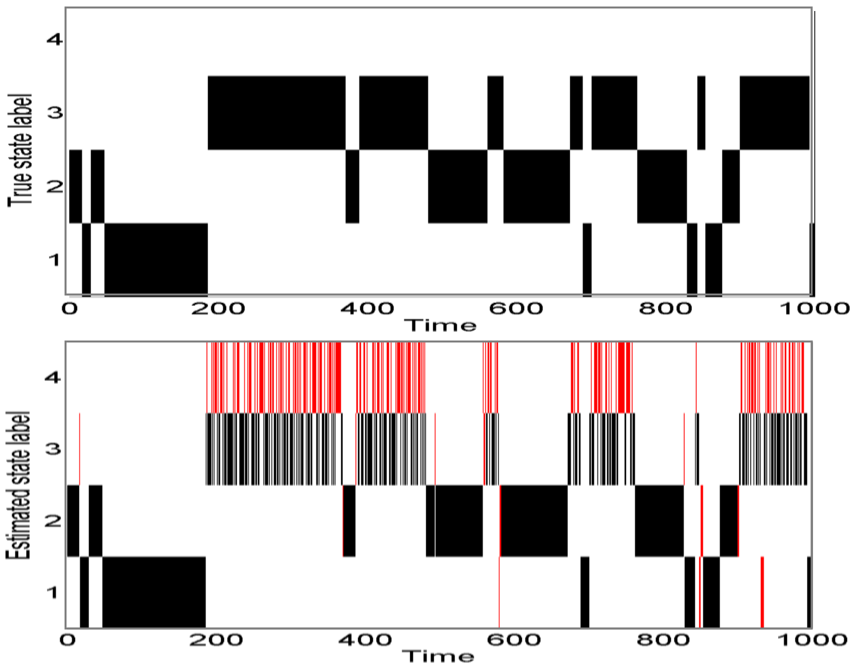
- State-specific transition distributions:

$$\pi_j \sim \mathrm{DP}(\alpha\beta) \quad j = 1, 2, 3, \ldots$$

sparsity of β is shared $\longrightarrow$ $\boxed{E[\pi_{jk}] = \beta_k}$

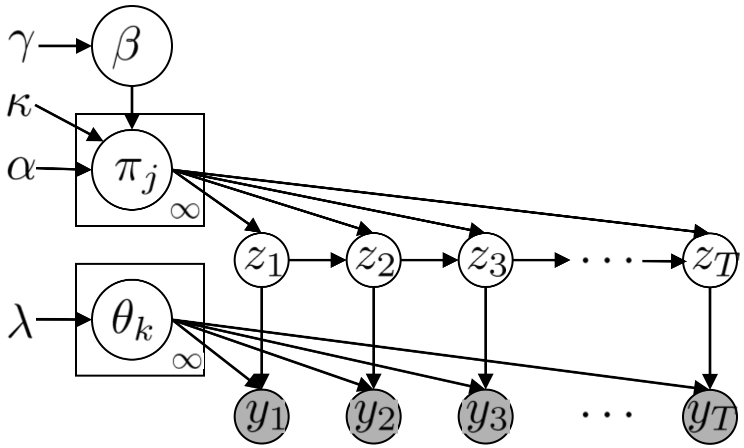
State Splitting





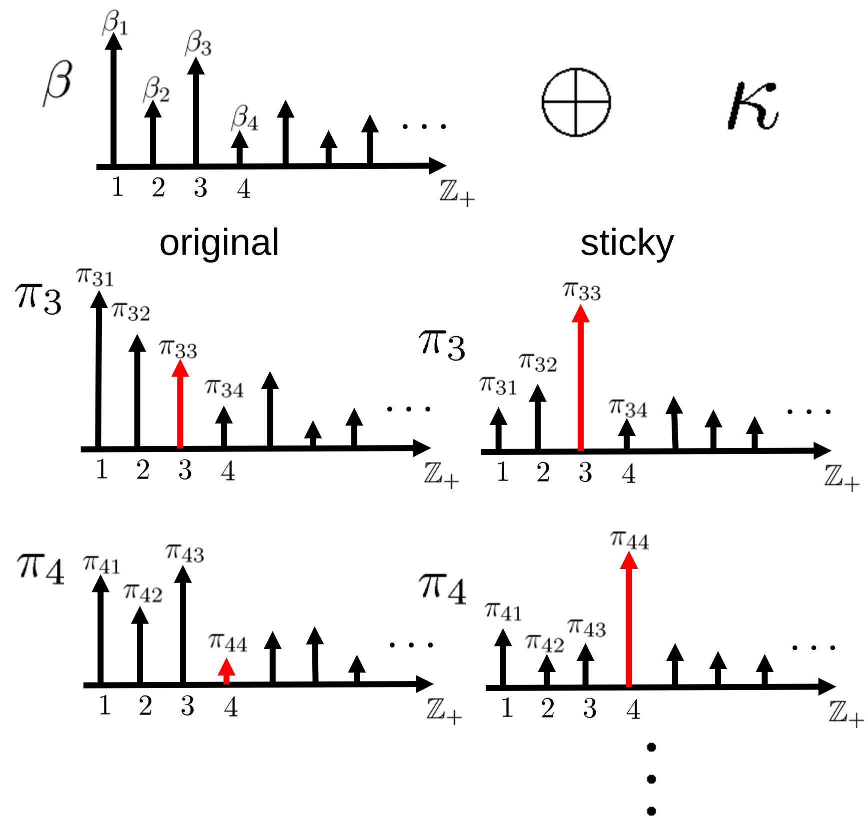
- HDP-HMM inadequately models temporal persistence of states
- DP bias insufficient to prevent unrealistically rapid dynamics
- Reduces predictive performance

"Sticky" HDP-HMM



$$
\begin{aligned}
\beta &\sim \mathrm{GEM}(\gamma) \\
\pi_j &\sim \mathrm{DP}\left(\alpha\beta + \kappa\delta_j\right)
\end{aligned}
$$

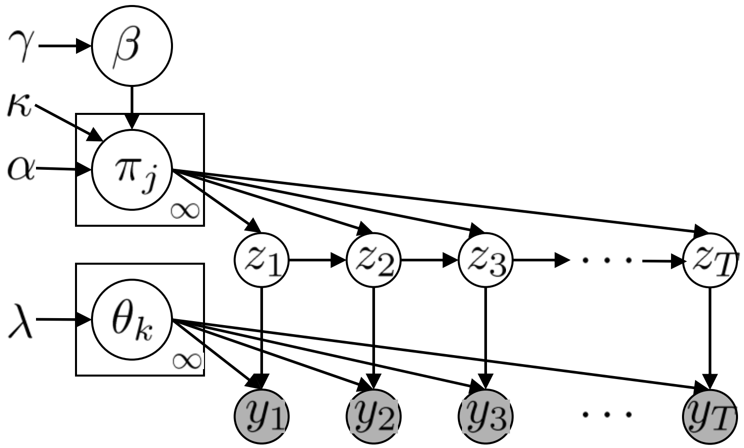
state-specific base measure

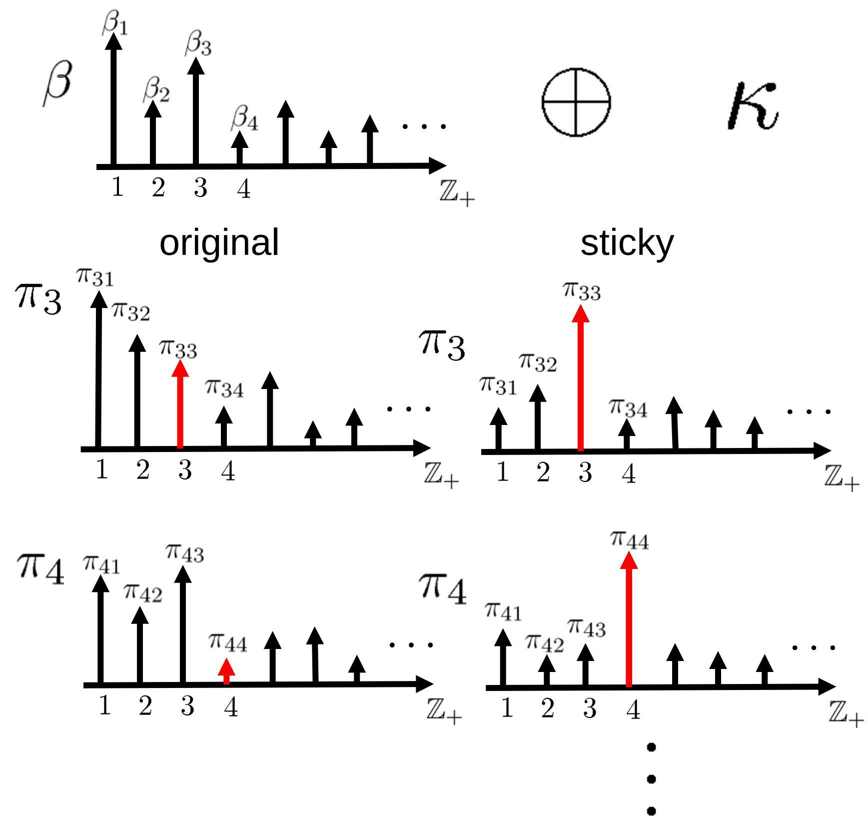


Increased probability of self-transition

$$E[\pi_{jk}] = \frac{\alpha\beta_k + \kappa\delta(j,k)}{\alpha + \kappa}$$

"Sticky" HDP-HMM





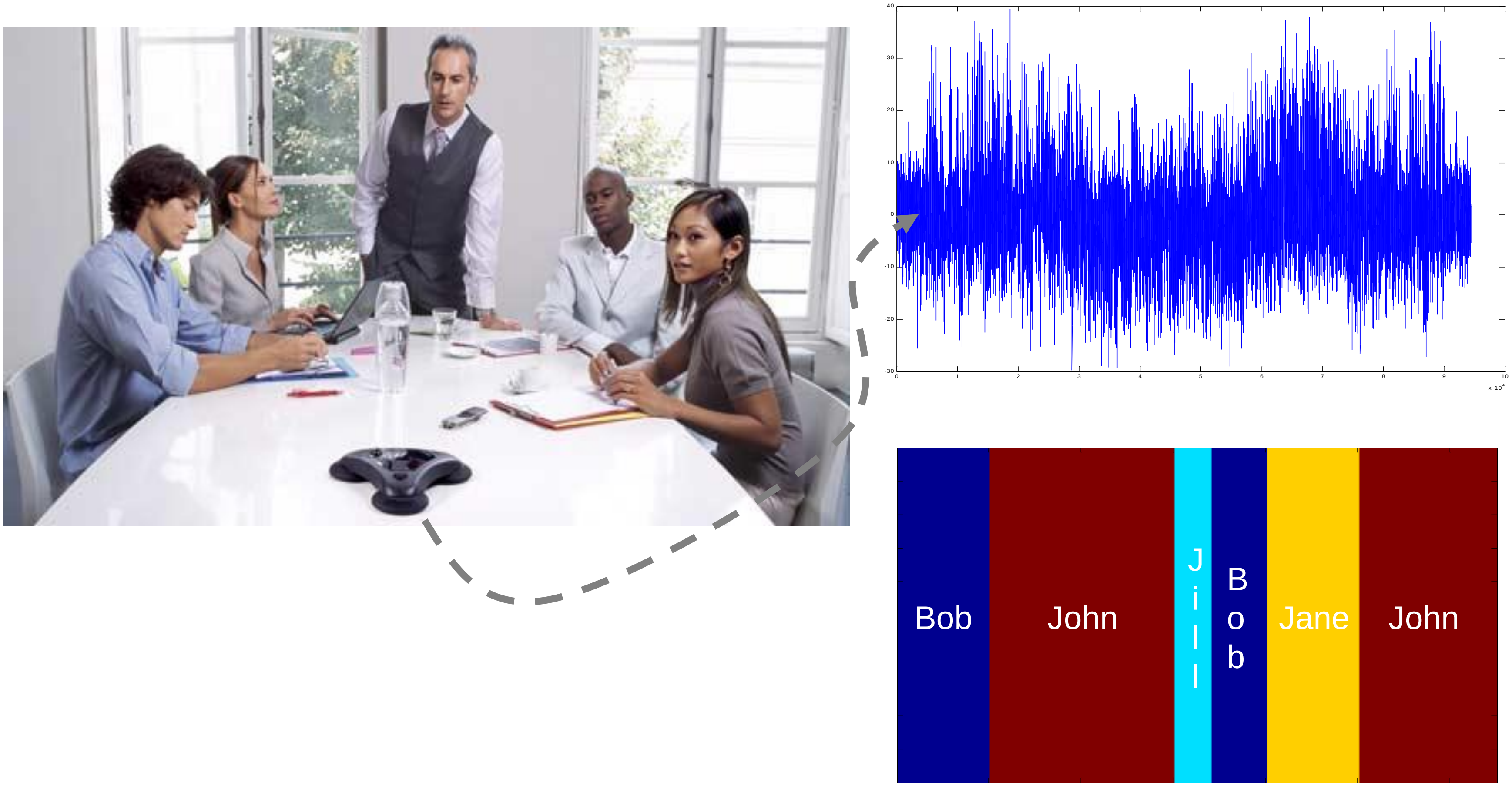
$$\beta \sim \mathrm{GEM}(\gamma)$$
$$\pi_j \sim \mathrm{DP}\left(\alpha\beta + \kappa\delta_j\right)$$

state-specific base measure

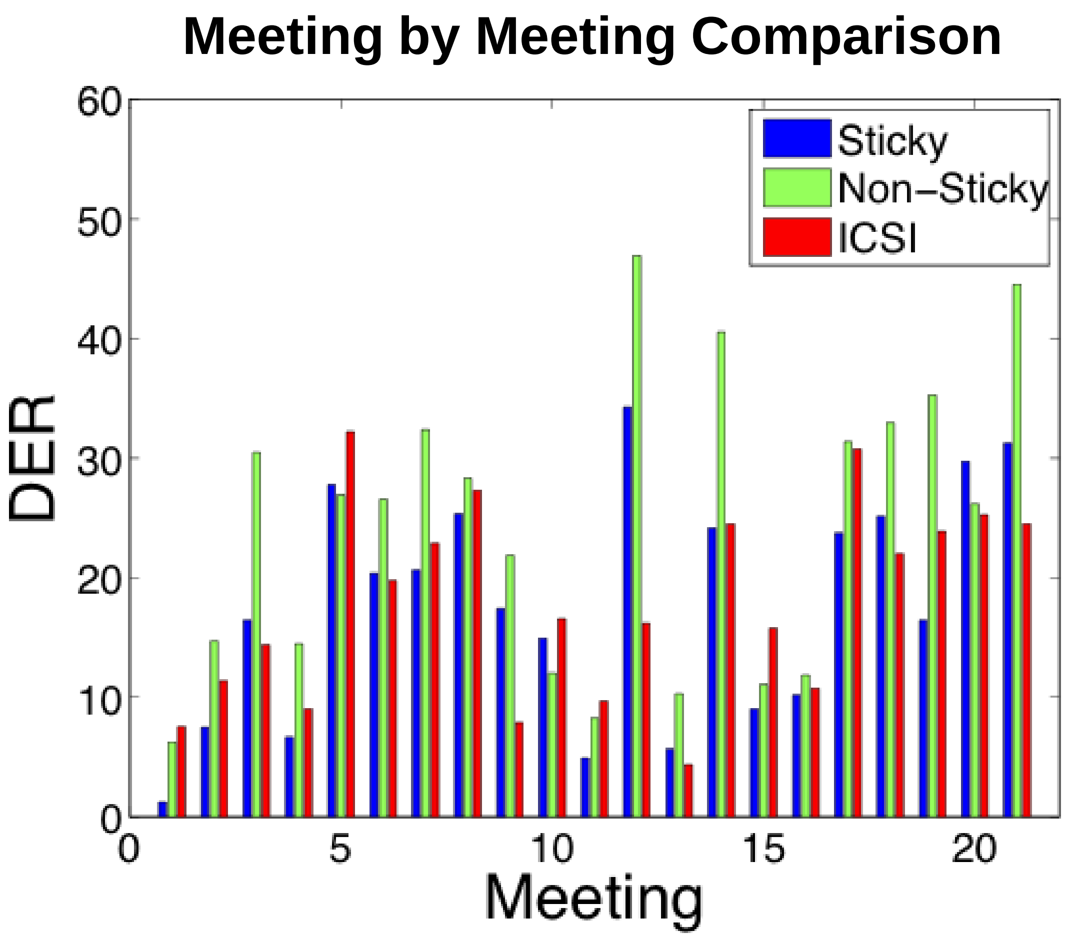
Increased probability of self-transition

$$E[\pi_{jk}] = \frac{\alpha\beta_k + \kappa\delta(j,k)}{\alpha + \kappa}$$

Speaker Diarization

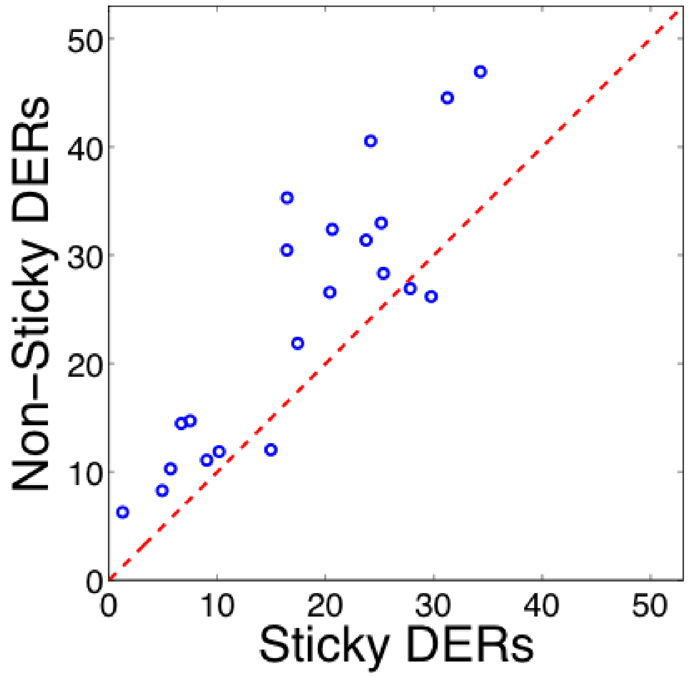


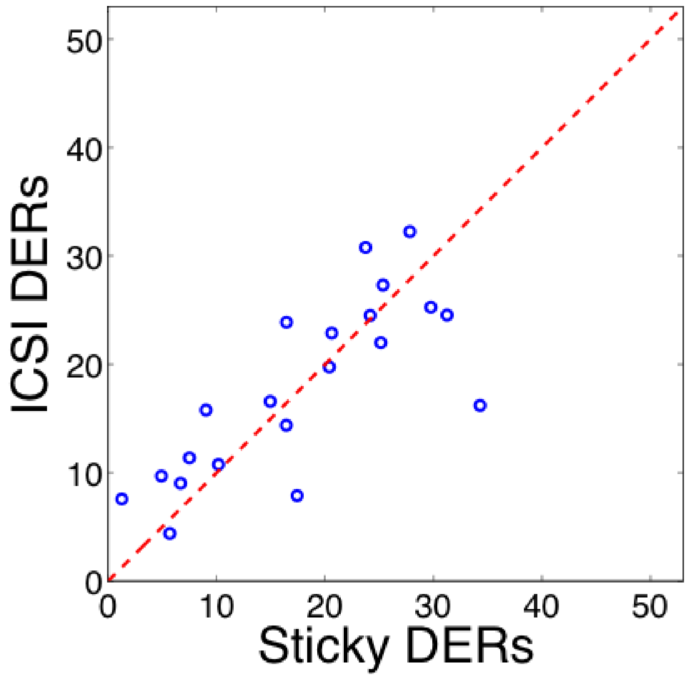
NIST Evaluations



- NIST Rich Transcription 2004-2007 meeting recognition evaluations
- 21 meetings
- ICSI results have been the current state-of-the-art

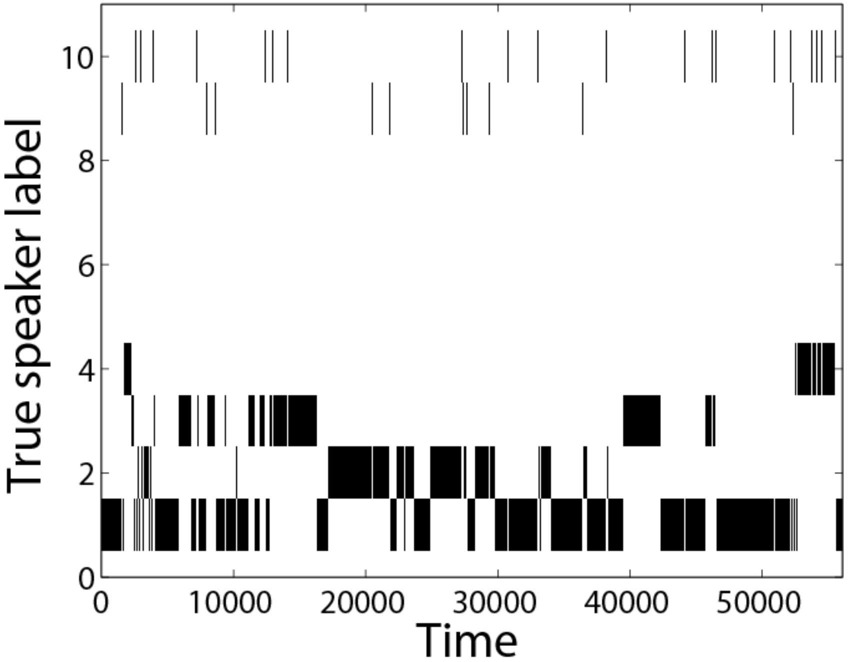
Results: 21 meetings

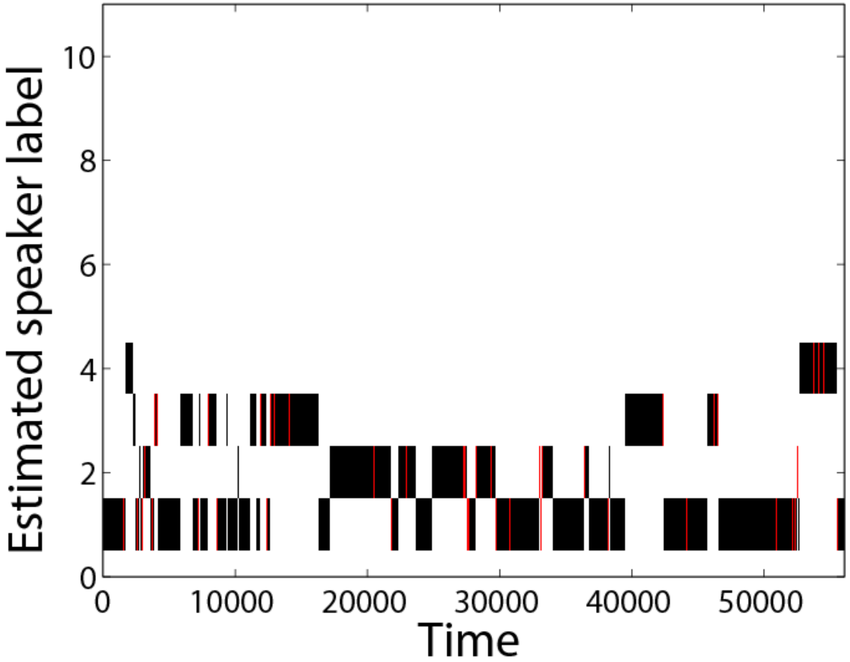




	Overall DER	Best DER	Worst DER
Sticky HDP-HMM	17.84%	1.26%	34.29%
Non-Sticky HDP-HMM	23.91%	6.26%	46.95%
ICSI	18.37%	4.39%	32.23%

Results: Meeting 1 (AMI_20041210-1052)

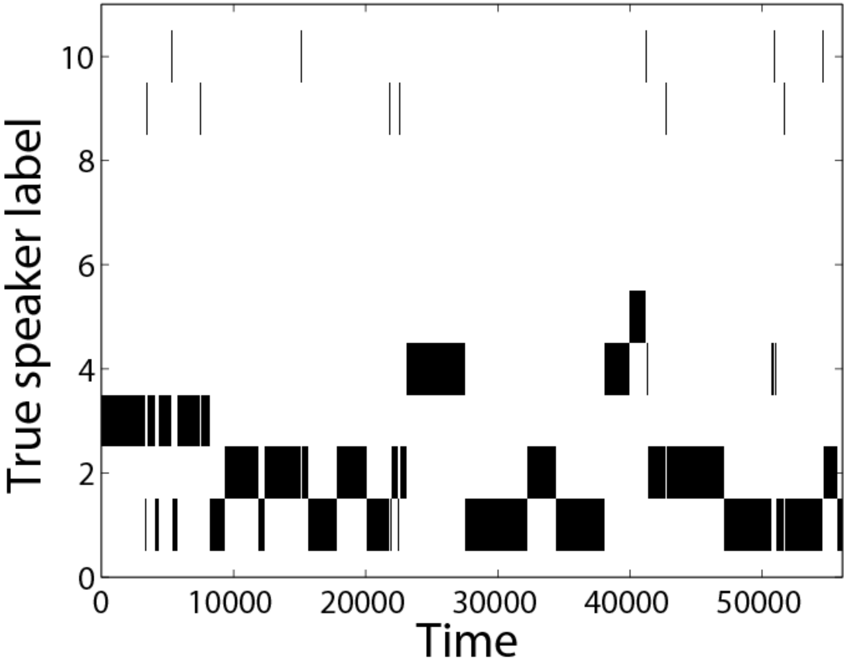


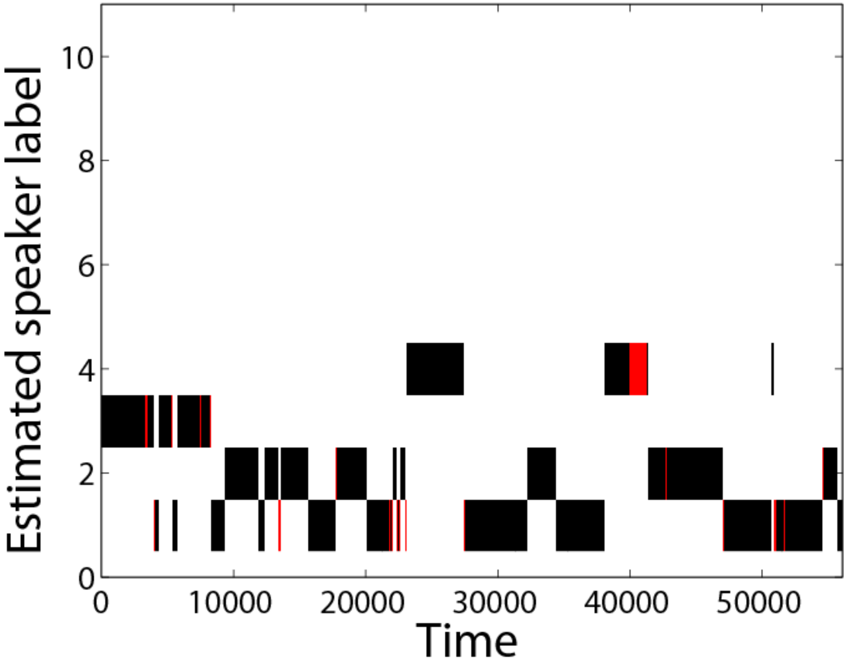


Sticky DER = 1.26%

ICSI DER = 7.56%

Results: Meeting 18 (VT_20050304-1300)



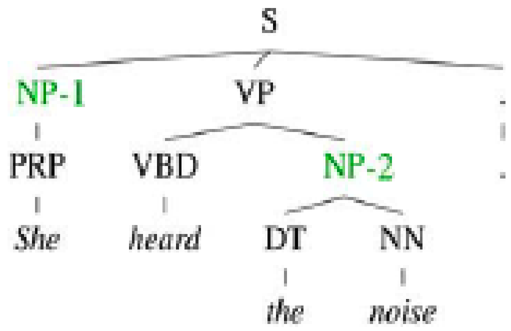


Sticky DER = 4.81%

ICSI DER = 22.00%

HDP-PCFG

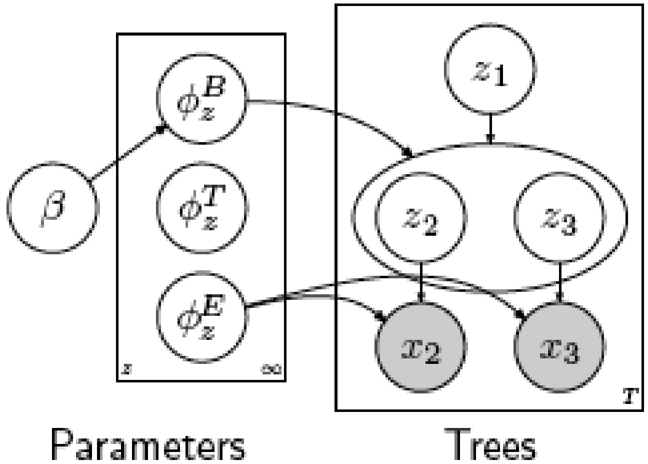
- The most successful parsers of natural language are based on *lexicalized probabilistic context-free grammars* (PCFGs)
- We want to learn PCFGs from data without hand-tuning of the number or kind of lexical categories

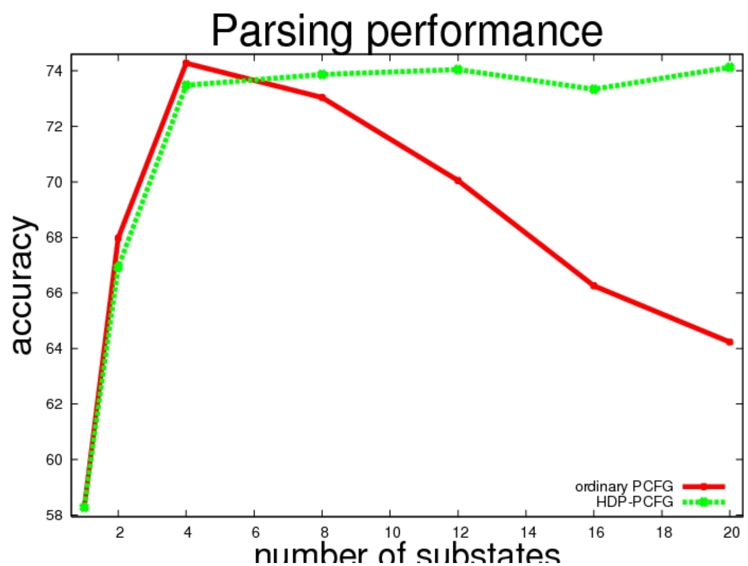


HDP-PCFG

HDP-PCFG

$\beta \sim \mathrm{GEM}(\alpha)$	[states]
$\phi^E_z \sim \mathrm{Dirichlet}(\alpha^E)$	[emissions]
$\phi^B_z \sim \mathrm{DP}(\alpha^B, \beta\beta^T)$	[productions]
$(z_{L(i)}, z_{R(i)}) \sim \mathrm{Multinomial}(\phi^B_{z_i})$	[nodes]





The Beta Process

- The Dirichlet process naturally yields a multinomial random variable (which table is the customer sitting at?)
- *Problem*: in many problem domains we have a very large (combinatorial) number of possible tables
  - using the Dirichlet process means having a large number of parameters, which may overfit
  - perhaps instead want to characterize objects as collections of attributes (“sparse features”)?
  - i.e., binary matrices with more than one 1 in each row

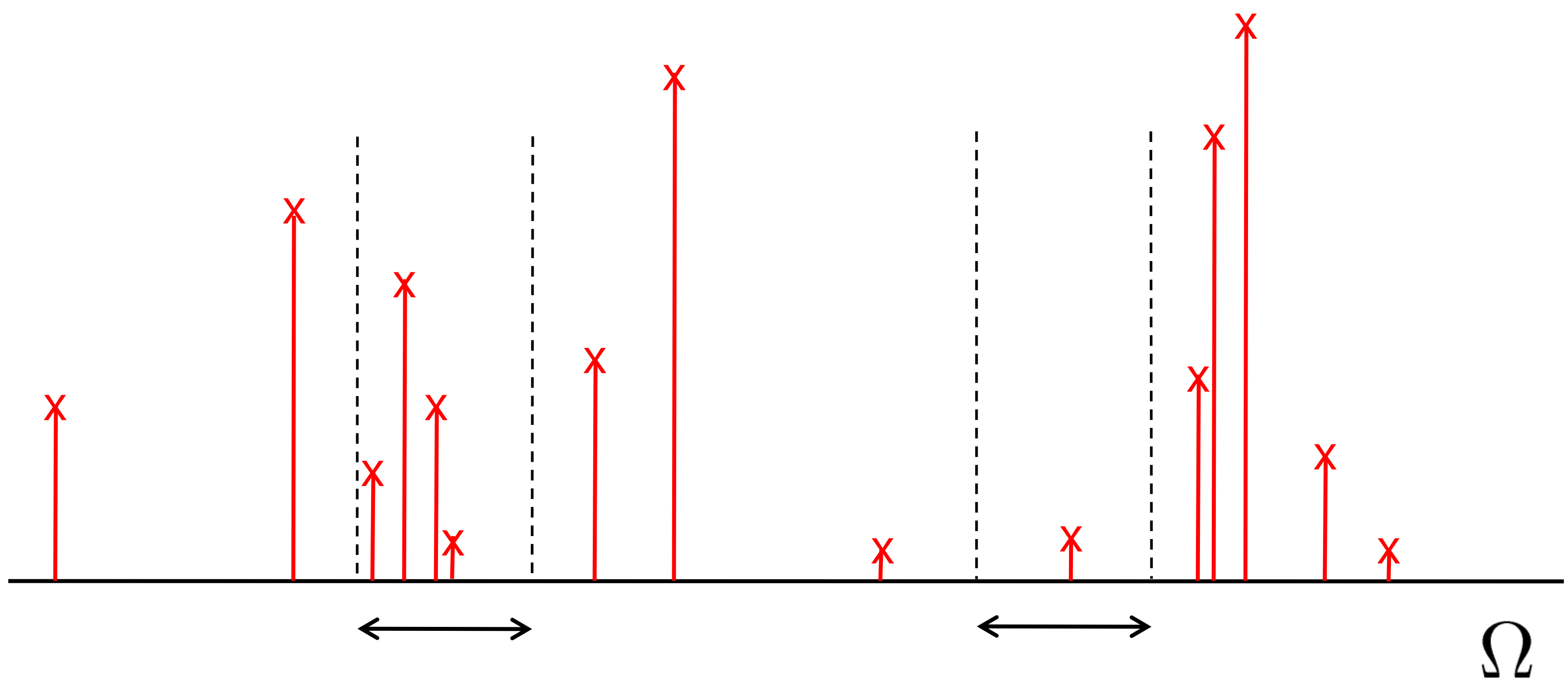
Completely Random Measures

(Kingman, 1968)

- Completely random measures are measures on a set Ω that assign independent mass to nonintersecting subsets of Ω
  - e.g., Brownian motion, gamma processes, beta processes, compound Poisson processes and limits thereof
- (The Dirichlet process is not a completely random process
  - but it's a normalized gamma process)
- Completely random processes are discrete wp1 (up to a possible deterministic continuous component)
- Completely random processes are random *measures*, not necessarily random *probability measures*

Completely Random Measures

(Kingman, 1968)

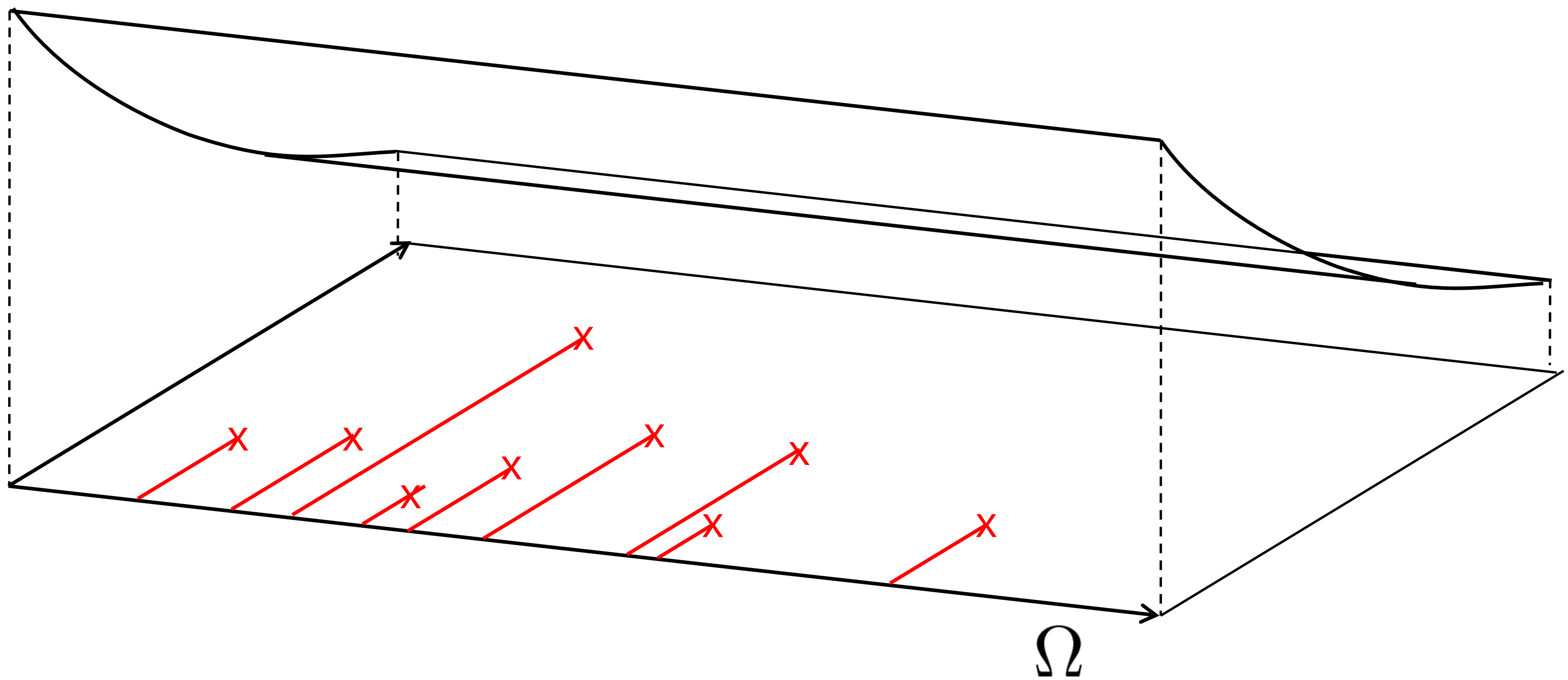


- Assigns independent mass to nonintersecting subsets of Ω

Completely Random Measures

(Kingman, 1968)

- Consider a non-homogeneous Poisson process on $\Omega \otimes R$ with rate function obtained from some product measure
- Sample from this Poisson process and connect the samples vertically to their coordinates in Ω



Beta Processes

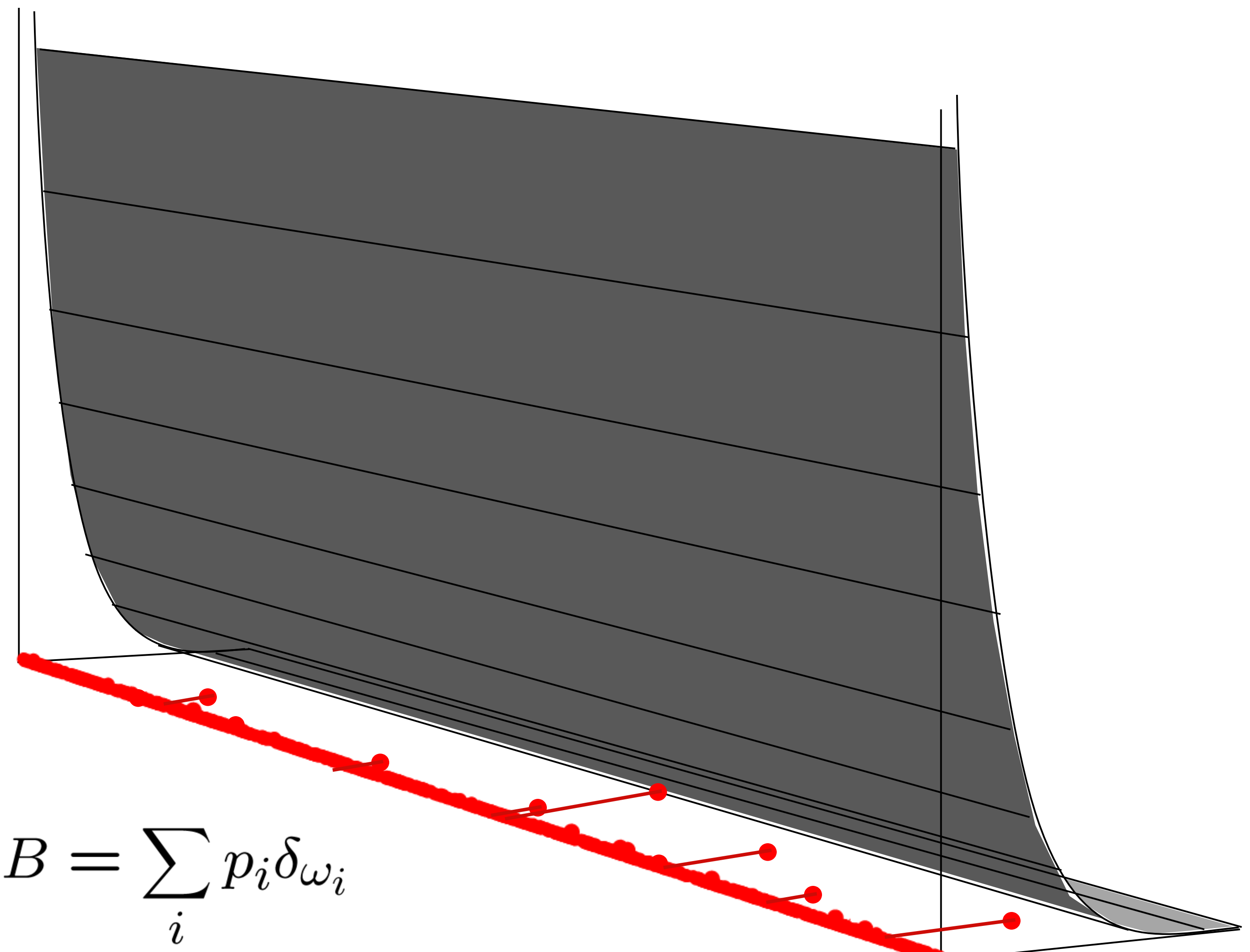
- The product measure is called a *Levy measure*
- For the beta process, this measure lives on $\Omega \otimes (0,1)$ and is given as follows:

$$\nu(d\omega, dp) = \underbrace{cp^{-1}(1-p)^{c-1}dp}_{\text{degenerate Beta(0,c) distribution}}\underbrace{B_0(d\omega)}_{\text{Base measure}}$$

- And the resulting random measure can be written simply as:

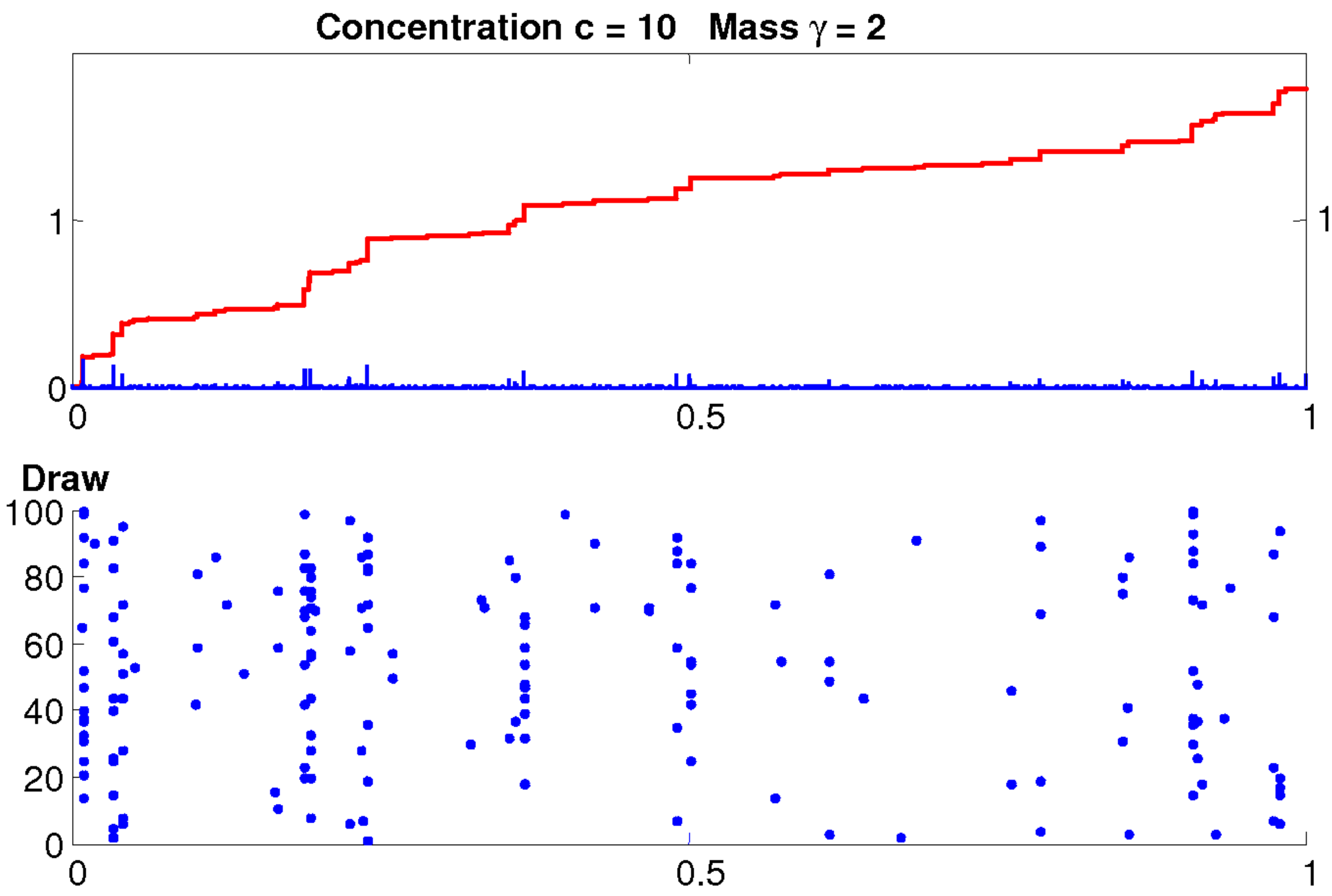
$$B = \sum_i p_i \delta_{\omega_i}$$

Beta Processes

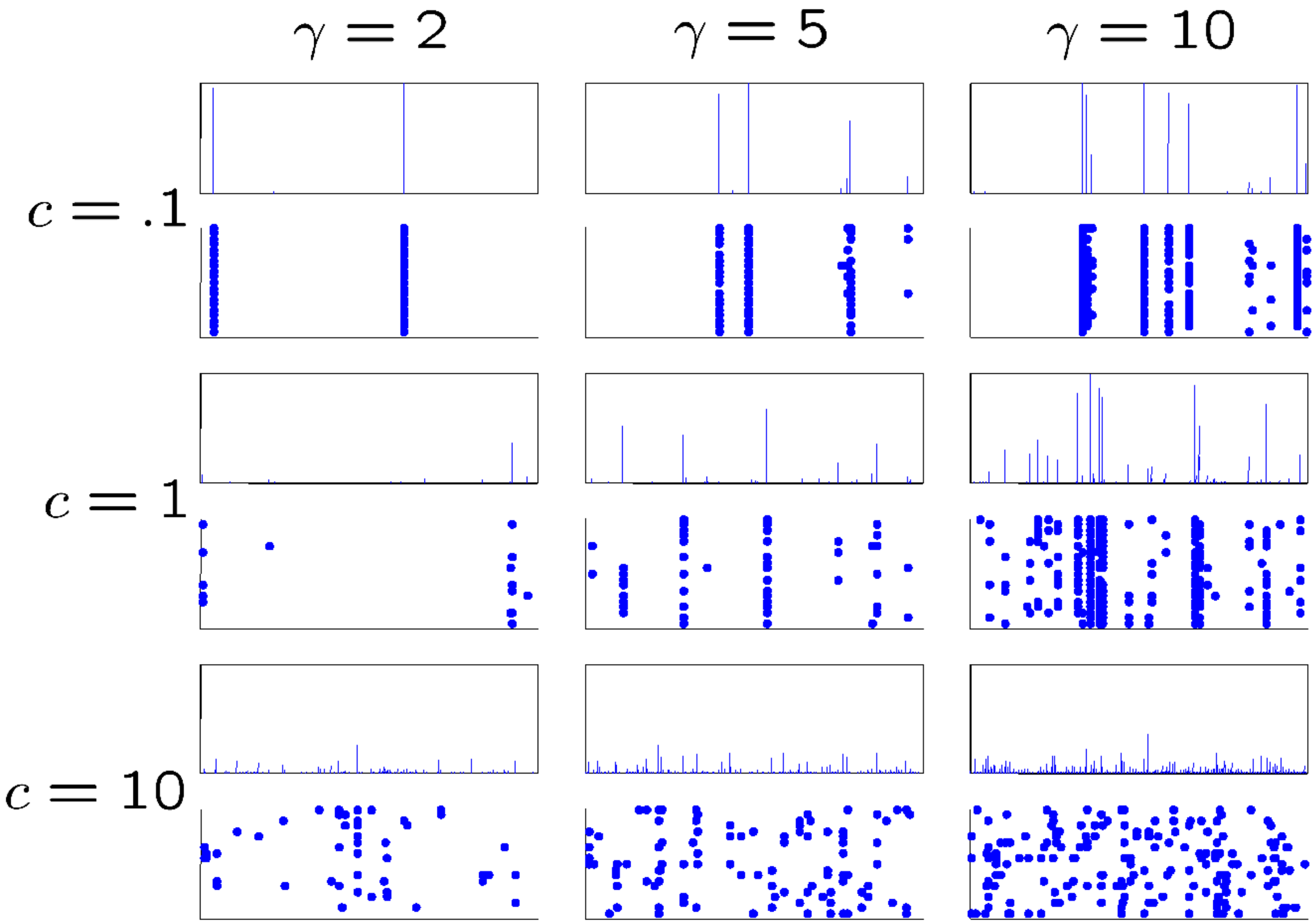


$$B = \sum_i p_i \delta_{\omega_i}$$

Beta Process and Bernoulli Process



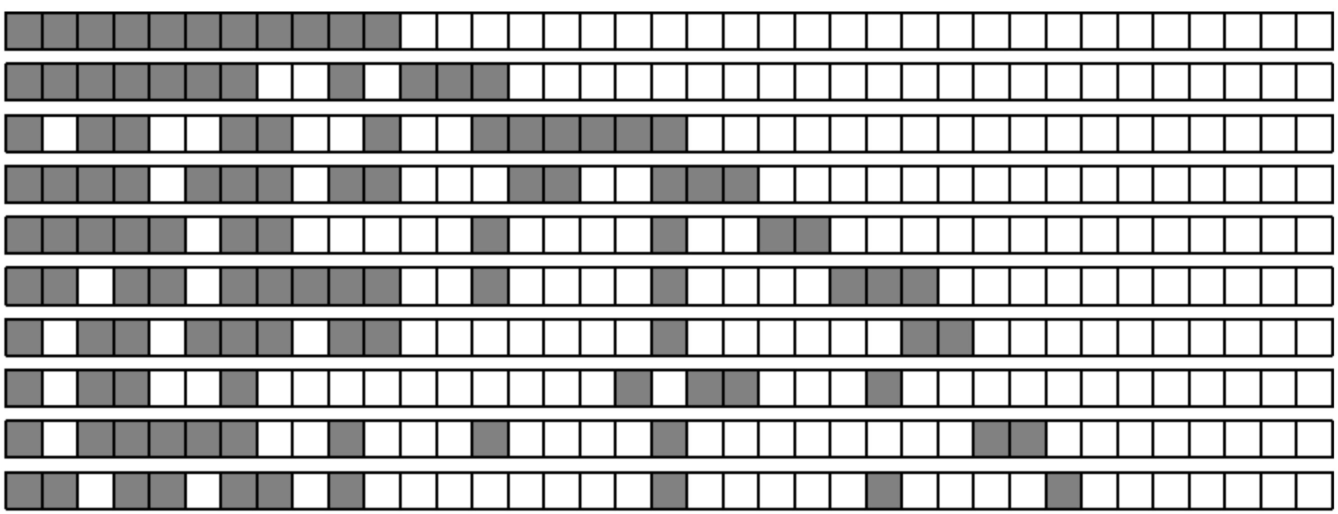
BP and BeP Sample Paths



Indian Buffet Process (IBP)

(Griffiths & Ghahramani, 2002)

- Indian restaurant with infinitely many dishes in a buffet line
- Customers 1 through n enter the restaurant
  - the first customer samples $\mathrm{Poiss}(\alpha)$ dishes
  - the ith customer samples a previously sampled dish with probability $m_k/(i+1)$ then samples $\mathrm{Poiss}(\alpha/i)$ new dishes



Beta Process Marginals

(Thibaux & Jordan, 2007)

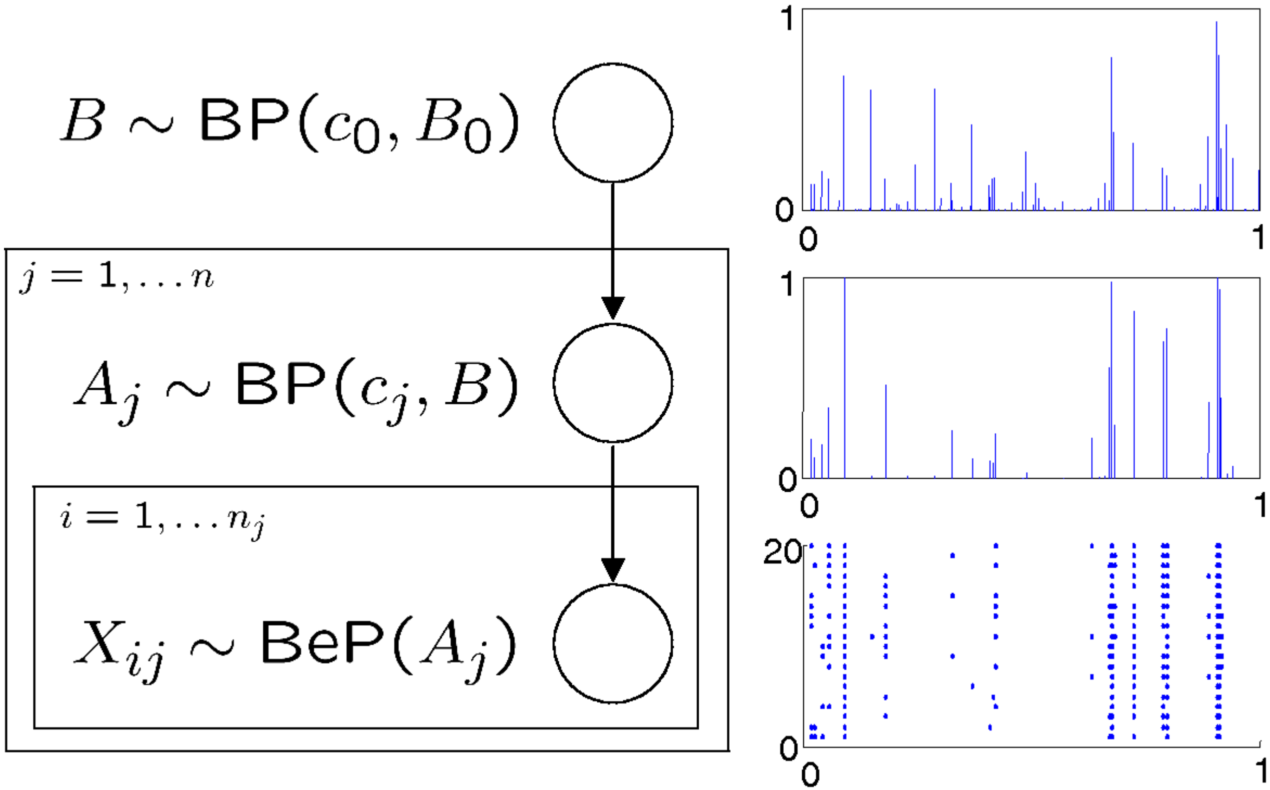
- *Theorem*: The beta process is the De Finetti mixing measure underlying the Indian buffet process (IBP)

Beta Process Point of View

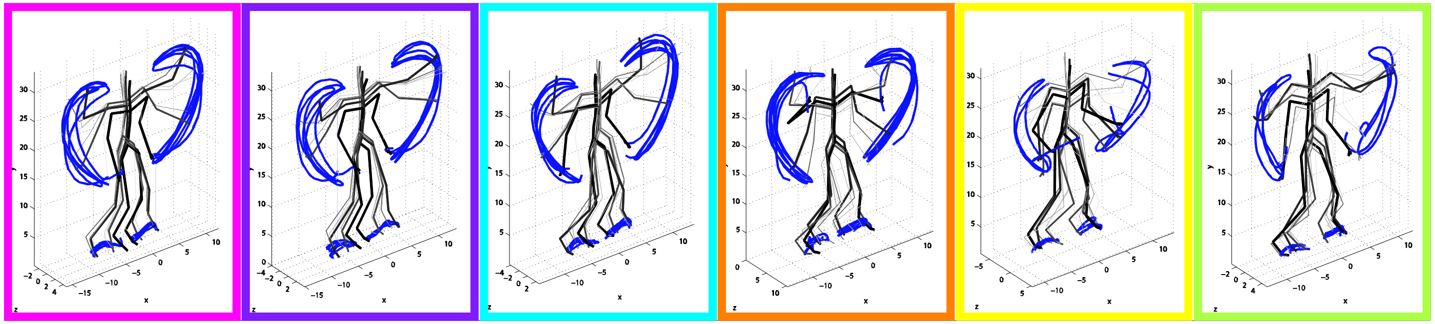
- The IBP is usually derived by taking a finite limit of a process on a finite matrix
- But this leaves some issues somewhat obscured:
  - is the IBP exchangeable?
  - why the Poisson number of dishes in each row?
  - is the IBP conjugate to some stochastic process?
- These issues are clarified from the beta process point of view
- A draw from a beta process yields a countably infinite set of coin-tossing probabilities, and each draw from the Bernoulli process tosses these coins independently

Hierarchical Beta Processes

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



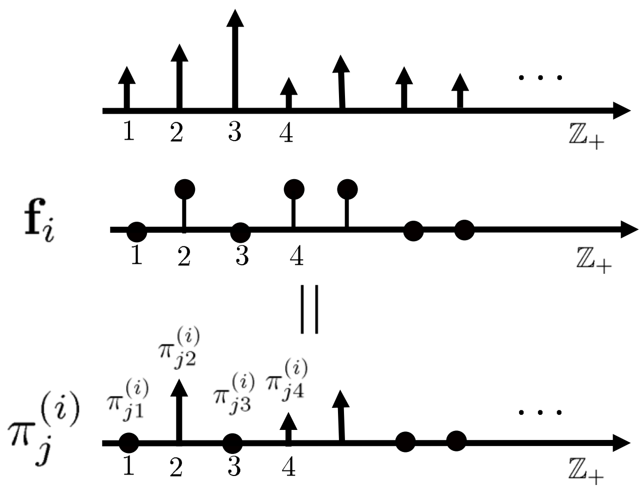
Multiple Time Series

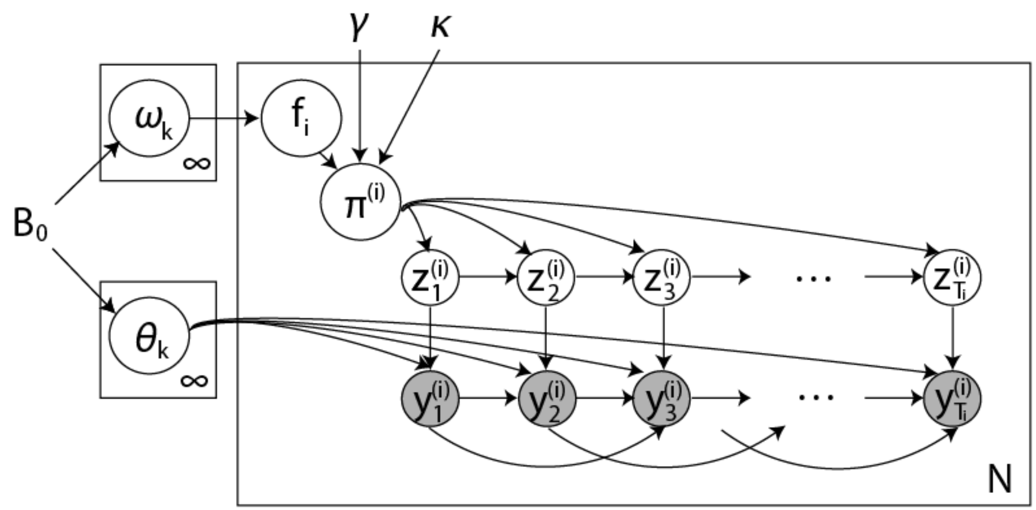


- Goals:
  - transfer knowledge among related time series in the form of a library of “behaviors”
  - allow each time series model to make use of an arbitrary subset of the behaviors
- Method:
  - represent behaviors as states in a nonparametric HMM
  - use the beta/Bernoulli process to pick out subsets of states

BP-AR-HMM

- Bernoulli process determines which states are used





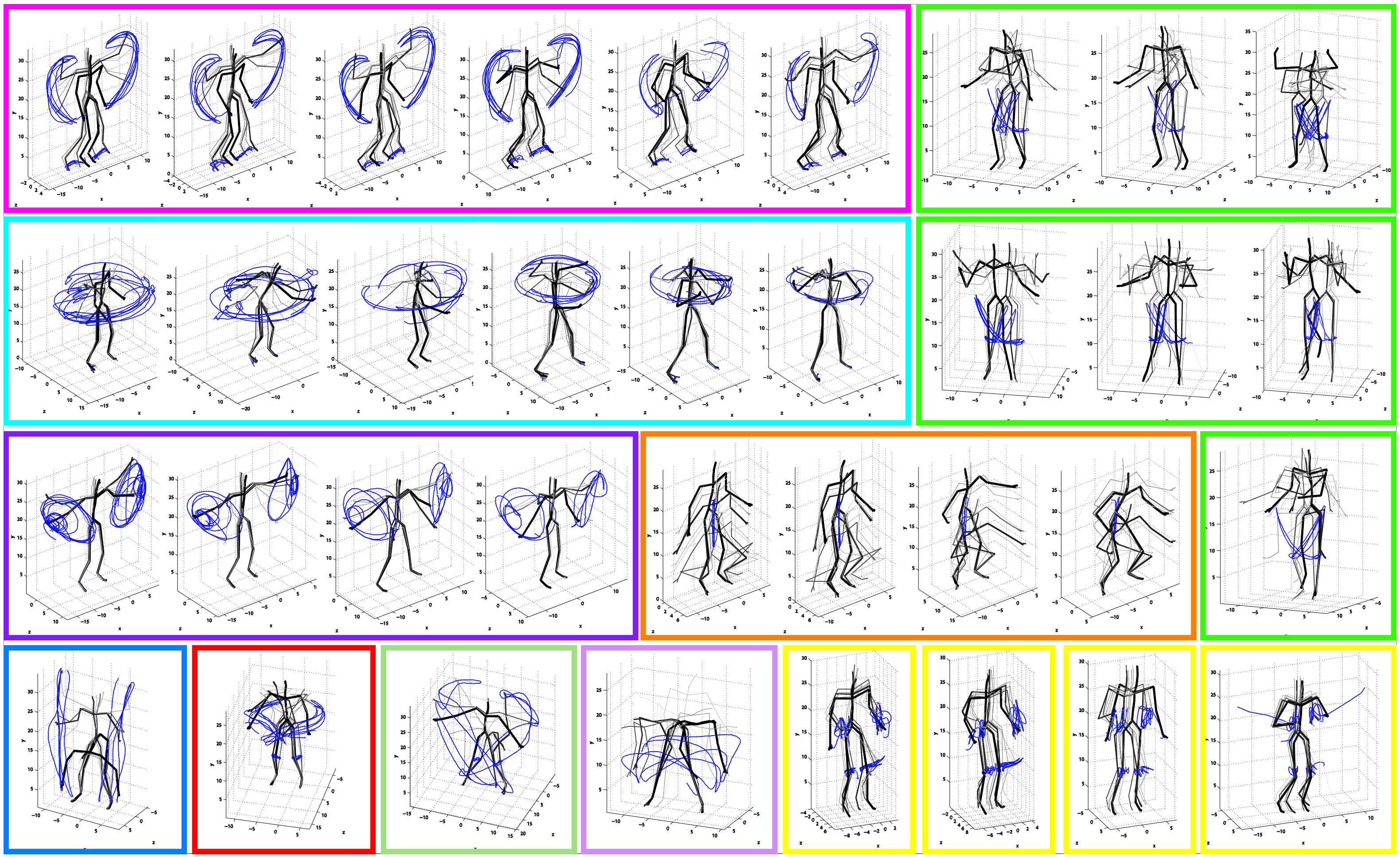
$$\pi_j^{(i)} \mid \mathbf{f}_i, \gamma, \kappa \sim \mathrm{Dir}([\gamma, \ldots, \gamma, \gamma + \kappa, \gamma, \ldots] \otimes \mathbf{f}_i)$$

$$z_t^{(i)} \sim \pi^{(i)}_{z_{t-1}^{(i)}}$$

$$\mathbf{y}_t^{(i)} = \sum_{j=1}^{r} A_{j, z_t^{(i)}} \mathbf{y}_{t-j}^{(i)} + \mathbf{e}_t^{(i)}(z_t^{(i)})$$

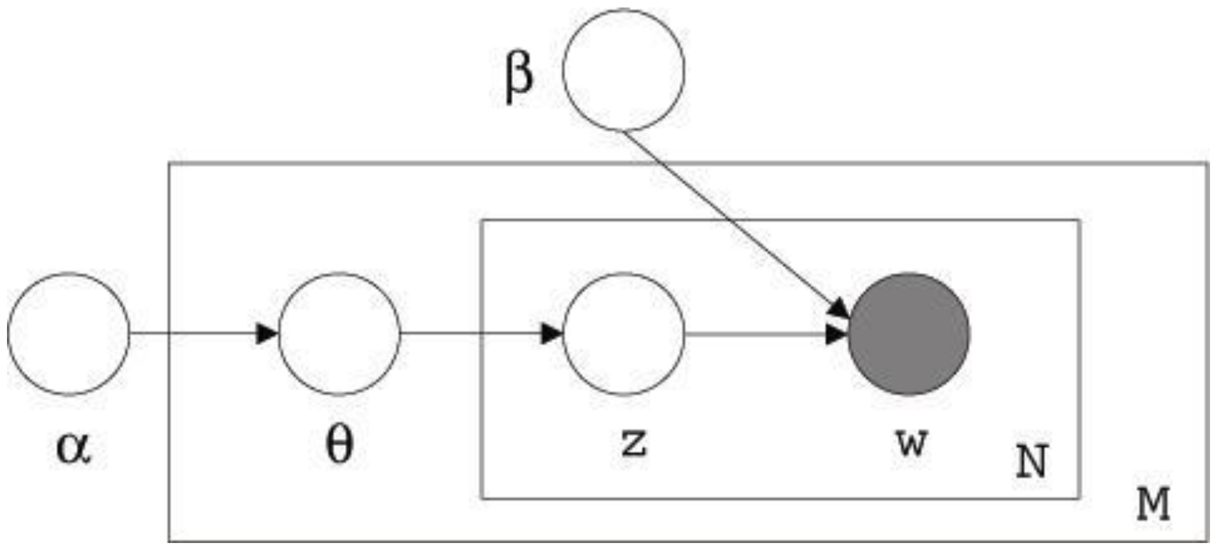
- Beta process prior:
  - encourages sharing
  - allows variability

Motion Capture Results



Latent Dirichlet Allocation

(Blei, Ng, and Jordan, 2003)



- A *word* is represented as a *multinomial* random variable *w*
- A *topic allocation* is represented as a *multinomial* random variable *z*
- A *document* is modeled as a *Dirichlet* random variable θ
- The variables α and β are *hyperparameters*

Finite Mixture Models

- The mixture components are distributions on individual words in some vocabulary (e.g., for text documents, a multinomial over lexical items)
  - often referred to as “topics”
- The generative model of a document:
  - select a mixture component
  - repeatedly draw words from this mixture component
- The mixing proportions are corpora-specific, not document-specific
- Major drawback: *each document can express only a single topic*

Finite Admixture Models

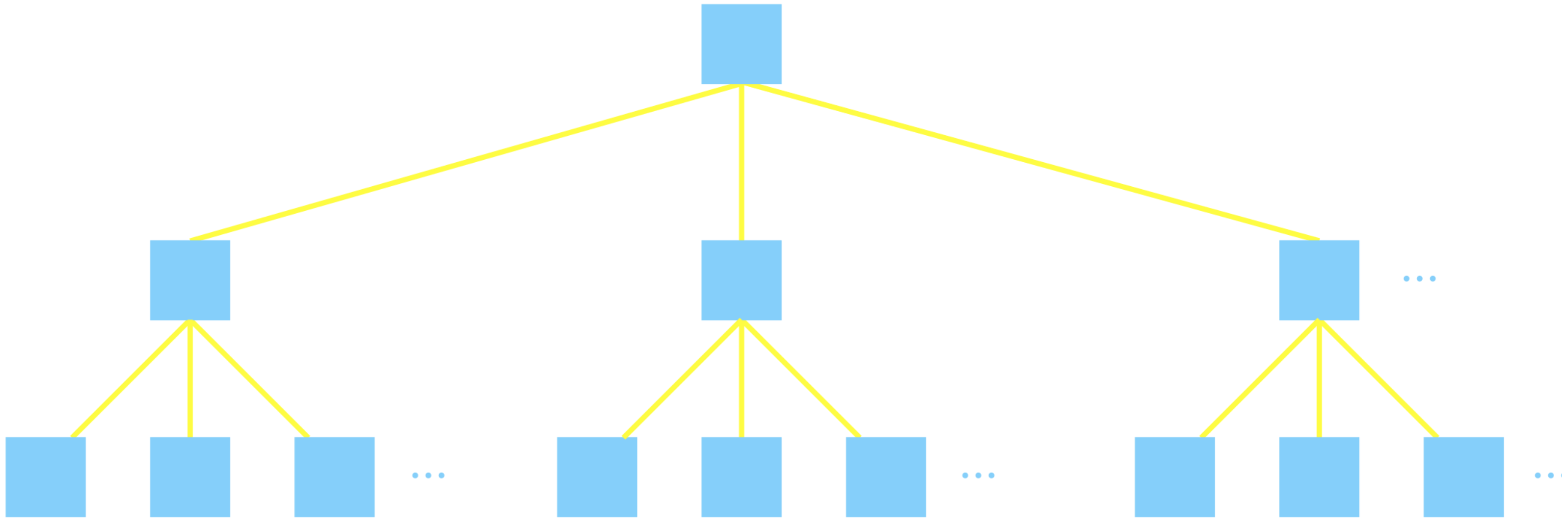
- The mixture components are distributions on individual words in some vocabulary (e.g., for text documents, a multinomial over lexical items)
  - often referred to as “topics”
- The generative model of a document:
  - repeatedly select a mixture component
  - draw a word from this mixture component
- The mixing proportions are document-specific
- Now each document can express multiple topics

Abstraction Hierarchies

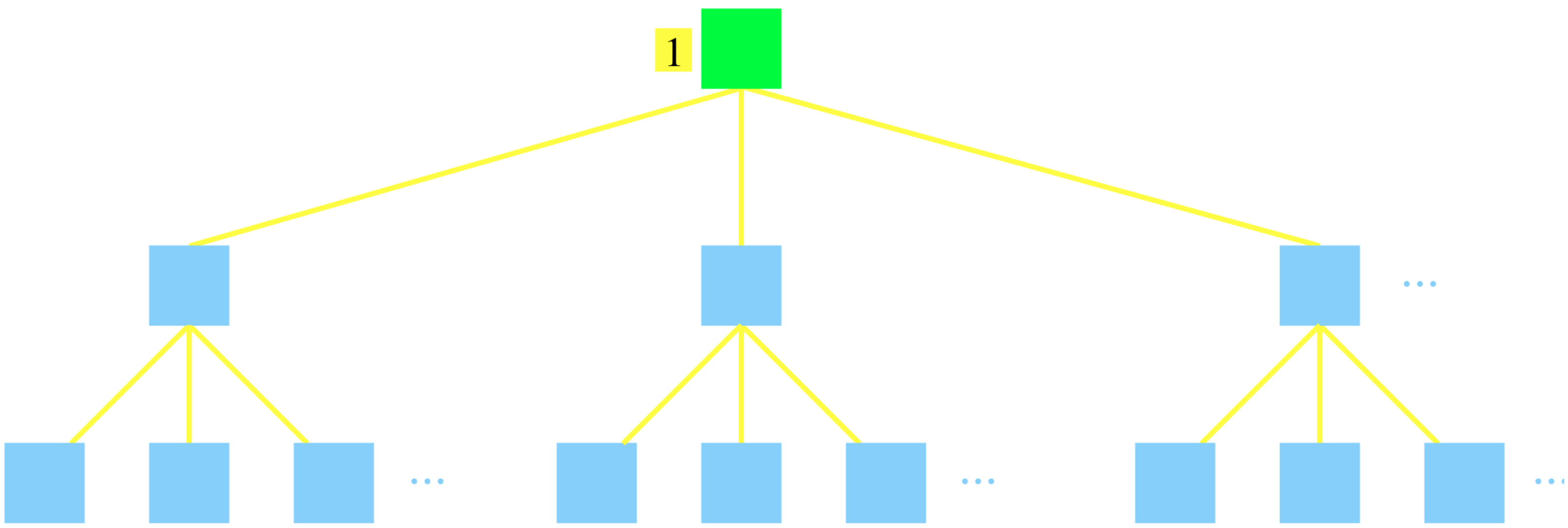
- Words in documents are organized not only by topics but also by level of abstraction
- Models such as LDA and the HDP don't capture this notion; common words often appear repeatedly across many topics
- Idea: *Let documents be represented as paths down a tree of topics, placing topics focused on common words near the top of the tree*

Nested Chinese Restaurant Process

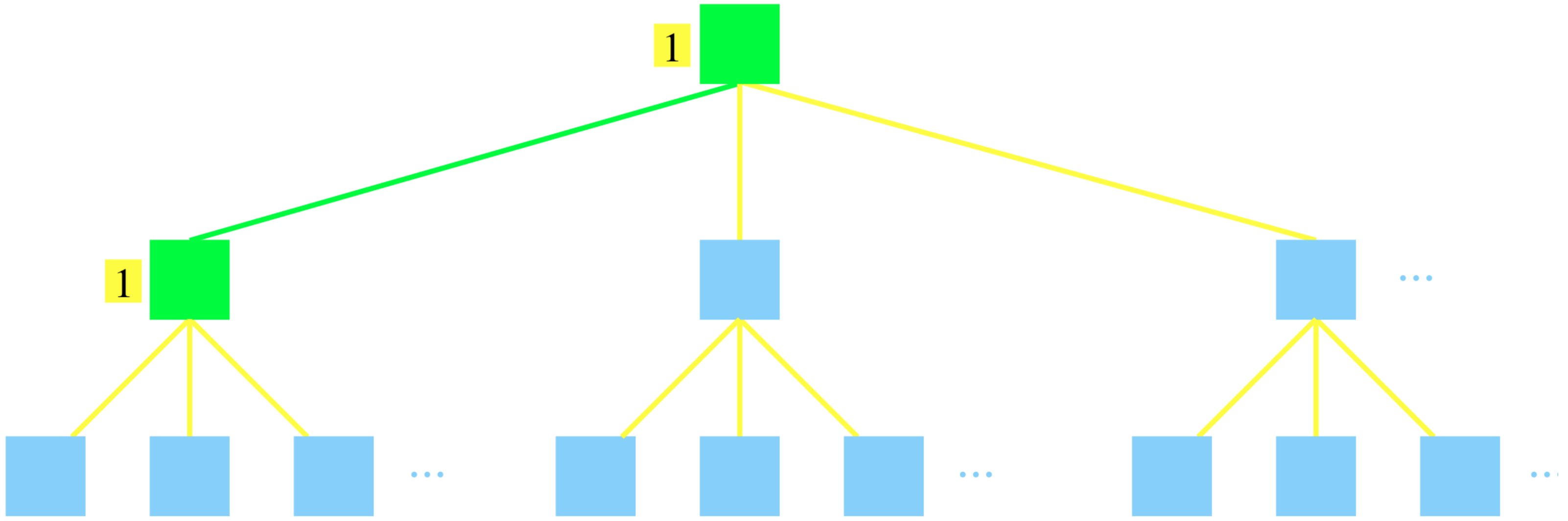
(Blei, Griffiths and Jordan, JACM, 2010)



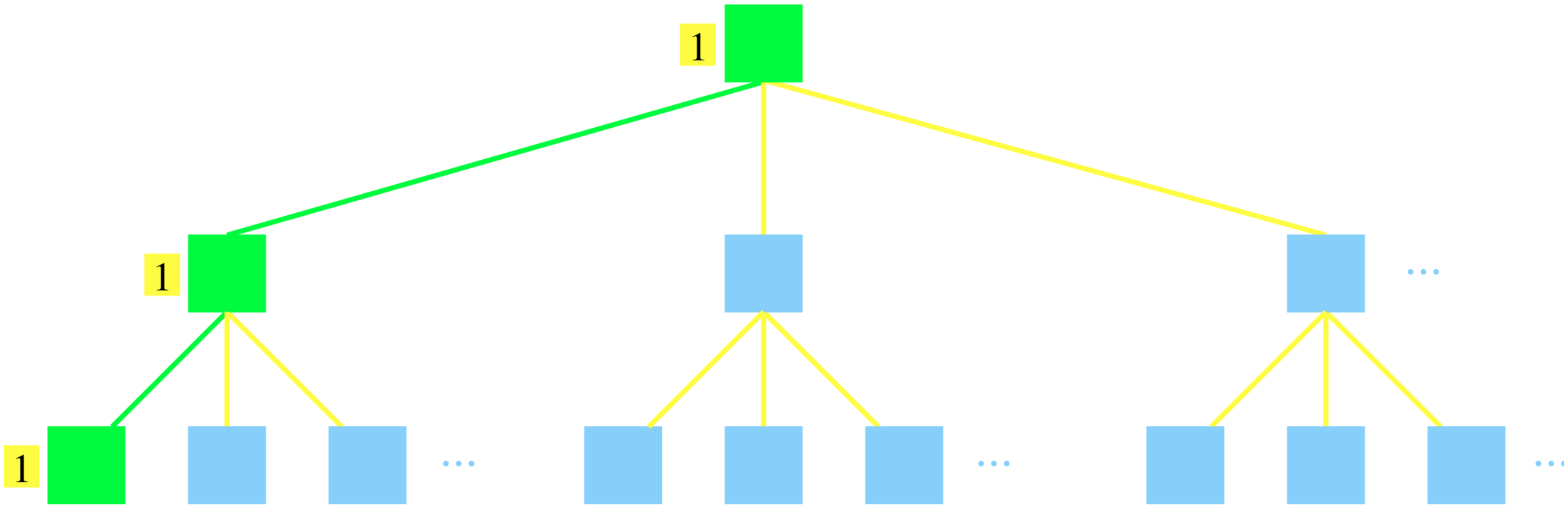
Nested Chinese Restaurant Process



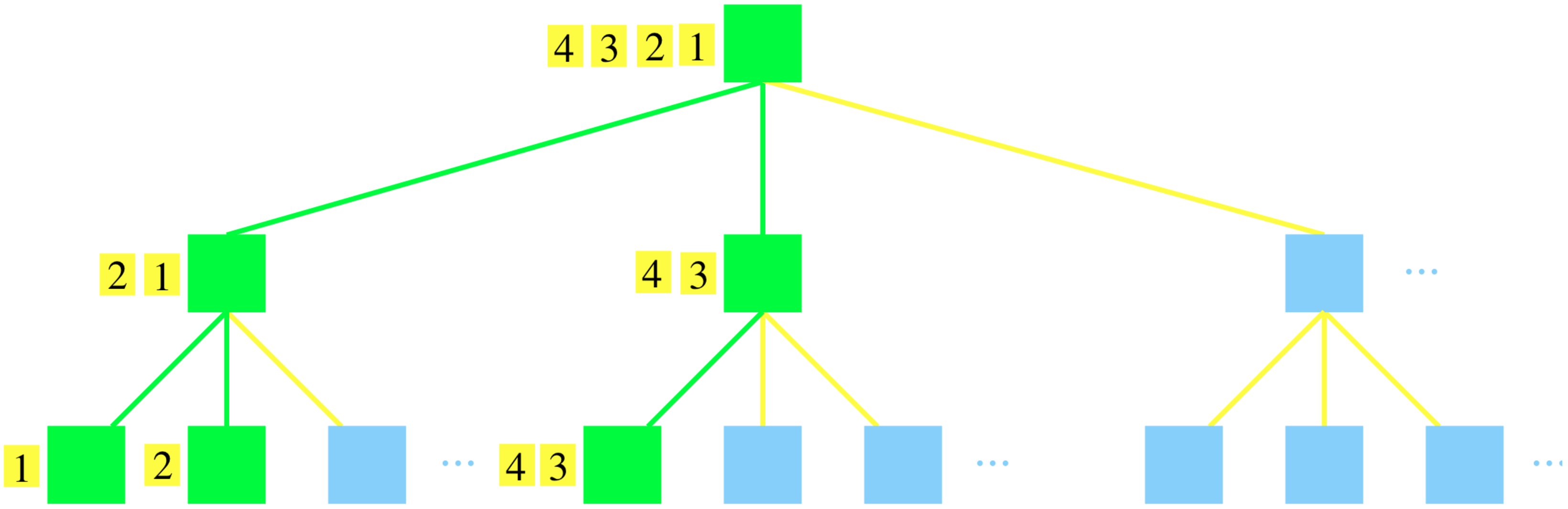
Nested Chinese Restaurant Process



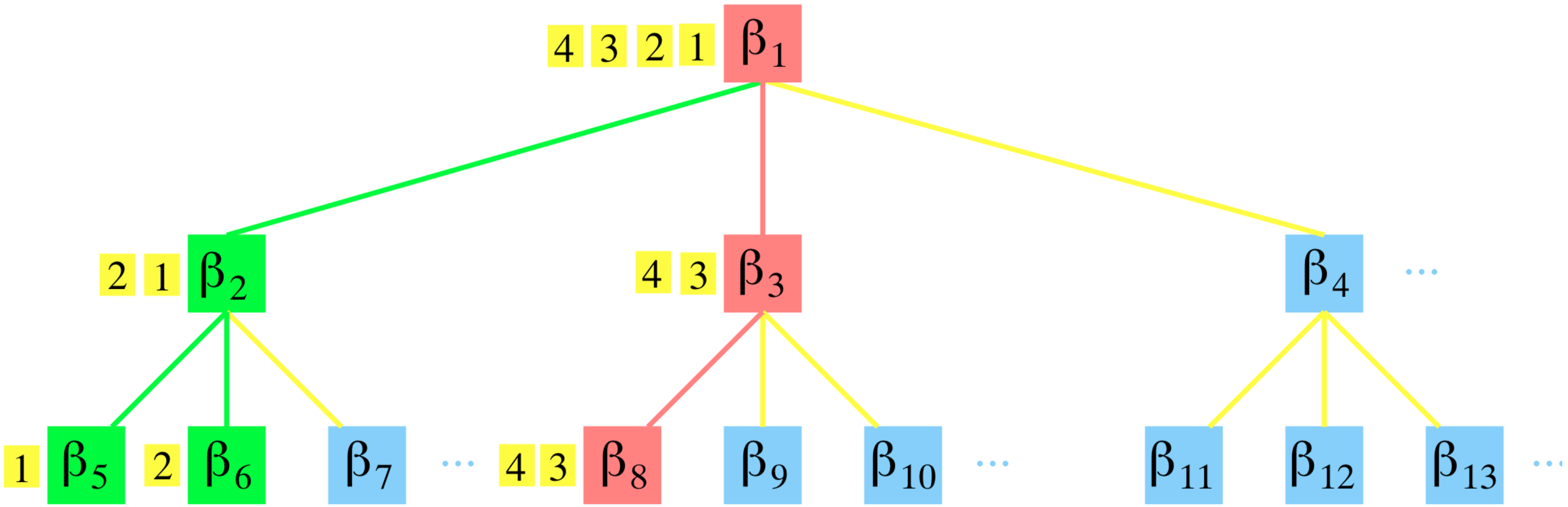
Nested Chinese Restaurant Process



Nested Chinese Restaurant Process



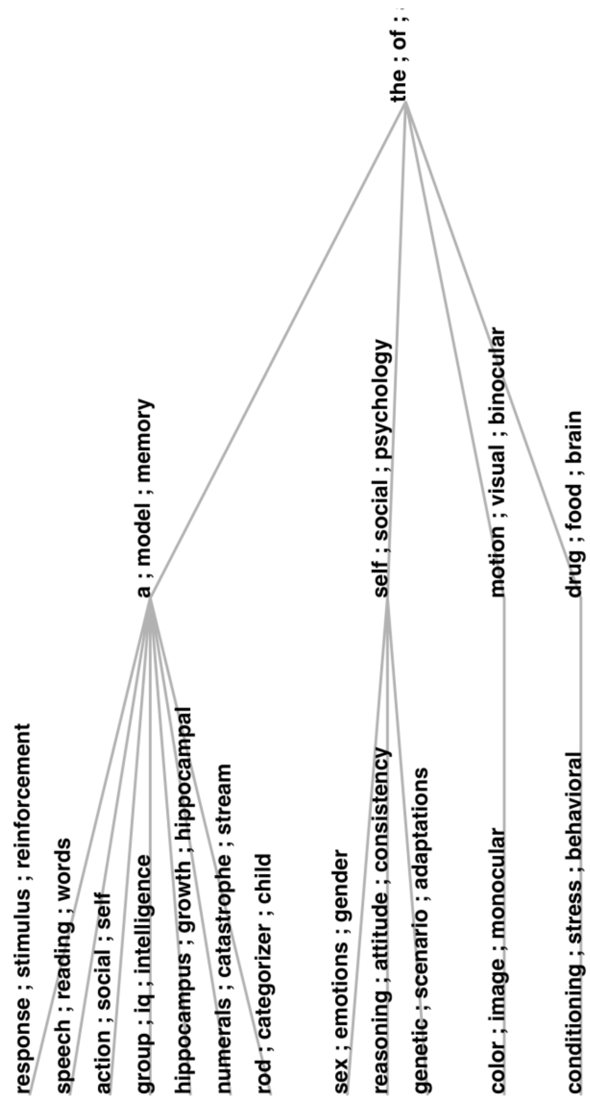
Nested Chinese Restaurant Process



Hierarchical Latent Dirichlet Allocation

- The generative model for a document is as follows:
  - use the nested CRP to select a path down the infinite tree
  - draw $\lambda \sim \mathrm{GEM}(\lambda_0)$ to obtain a distribution on levels along the path
  - repeatedly draw a level from λ and draw a word from the topic distribution at that level

Psychological Review Abstracts



Conclusions

- For papers, software, tutorials and more details:

  www.cs.berkeley.edu/~jordan/publications.html

- See, in particular, the papers:
  - "Hierarchical Models, Nested Models and Completely Random Measures"
  - "Bayesian Nonparametric Learning: Expressive Priors for Intelligent Systems"