Hierarchical Diriclet Processes

Y. Teh, M. Jordan, M. Beal, D. Blei

Presented by Ben Swanson

Motivation

Problem

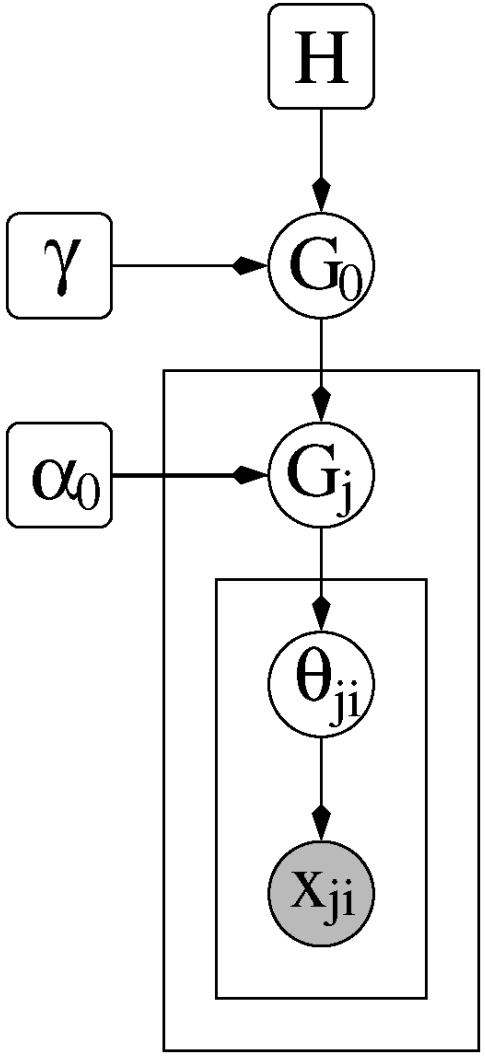
- Data is organized into groups or contexts
- Each group is generated with a mixture model
- Mixture components are from some parameterized family
- To compare groups, they must use the same mixture components

Hierarchical Dirichlet Process

Use a DP as the Base Distribution of a DP

- $G_0|\gamma, H \sim DP(\gamma, H)$
- $G_j|\alpha_0, G_0 \sim DP(\alpha_0, G_0)$

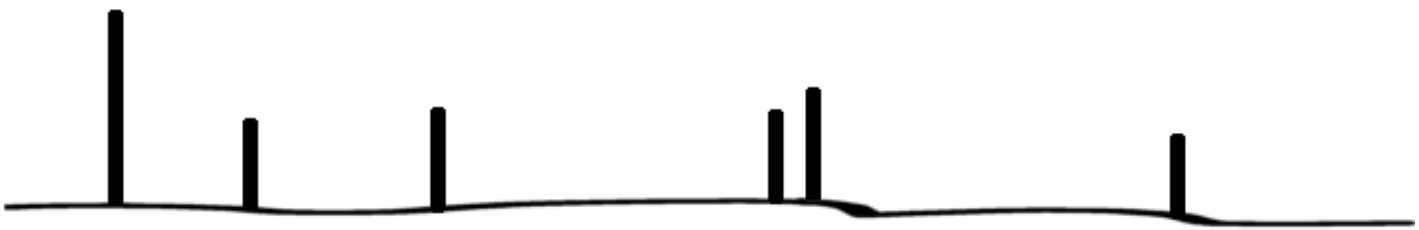
It's Double DP!



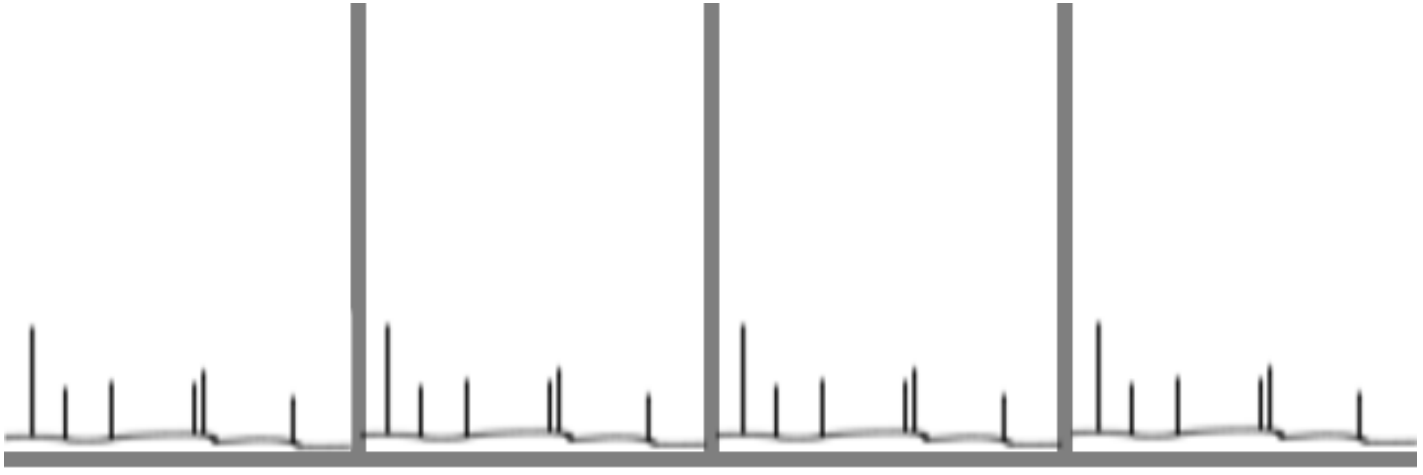
The Distribution

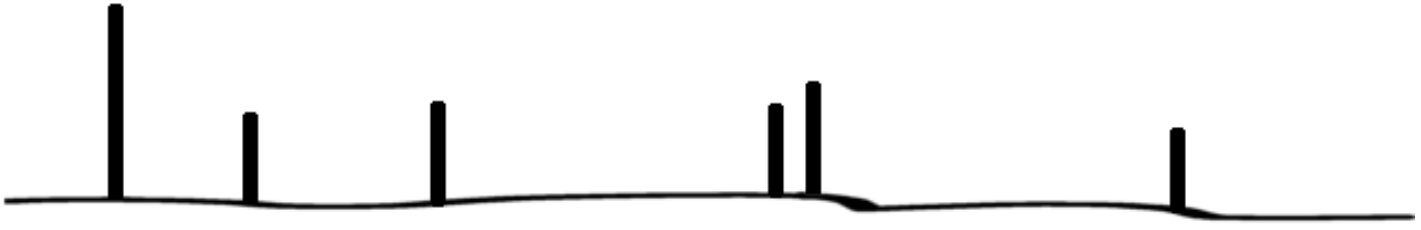
The Distribution

The Distribution

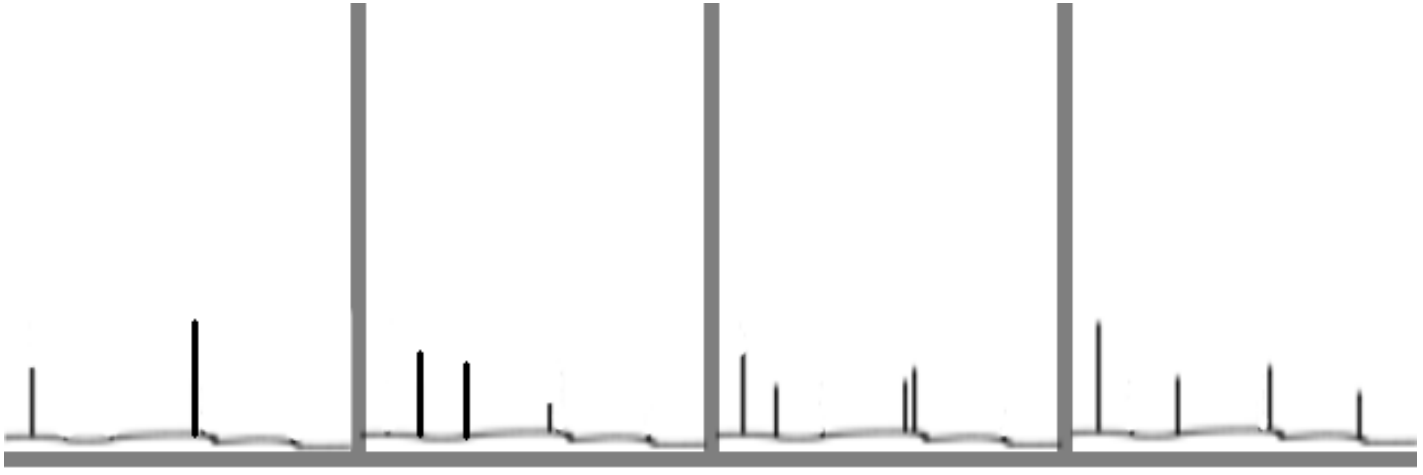


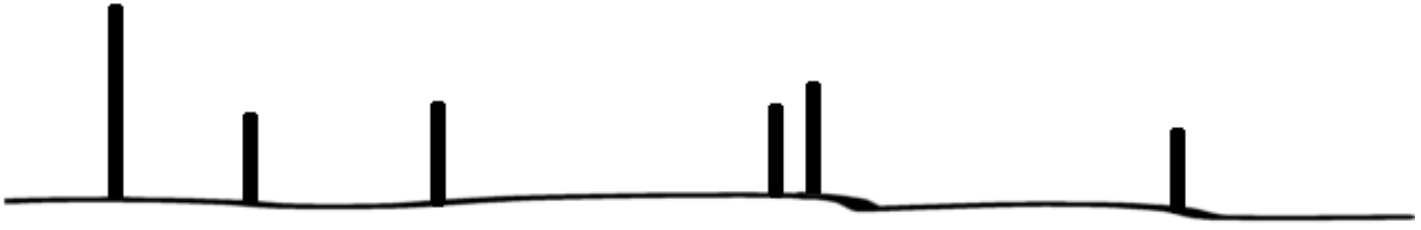
The Distribution



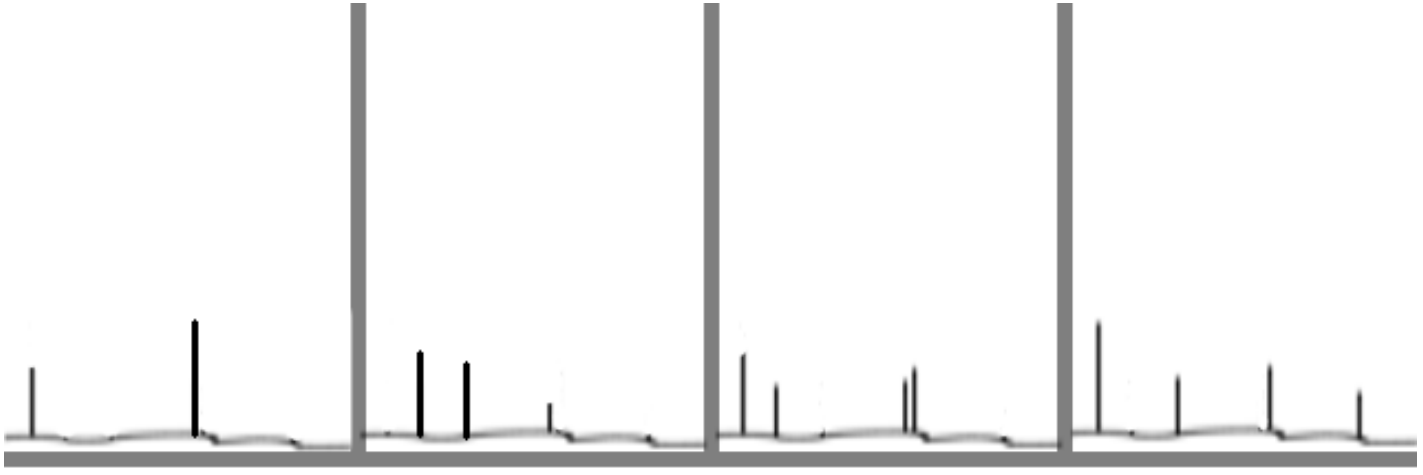


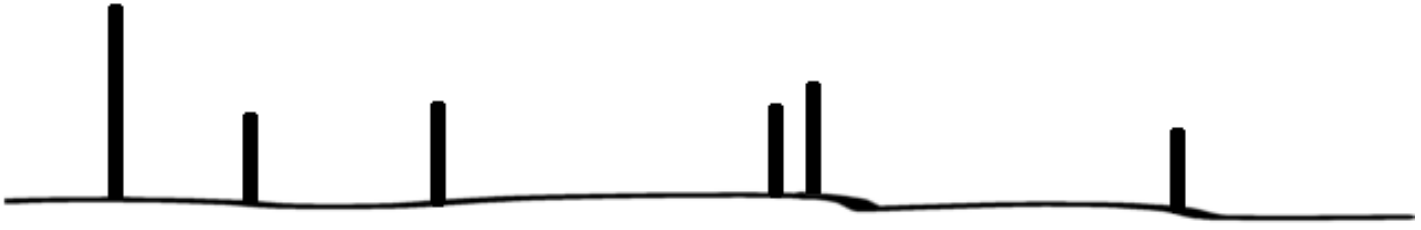
The Distribution





The Distribution

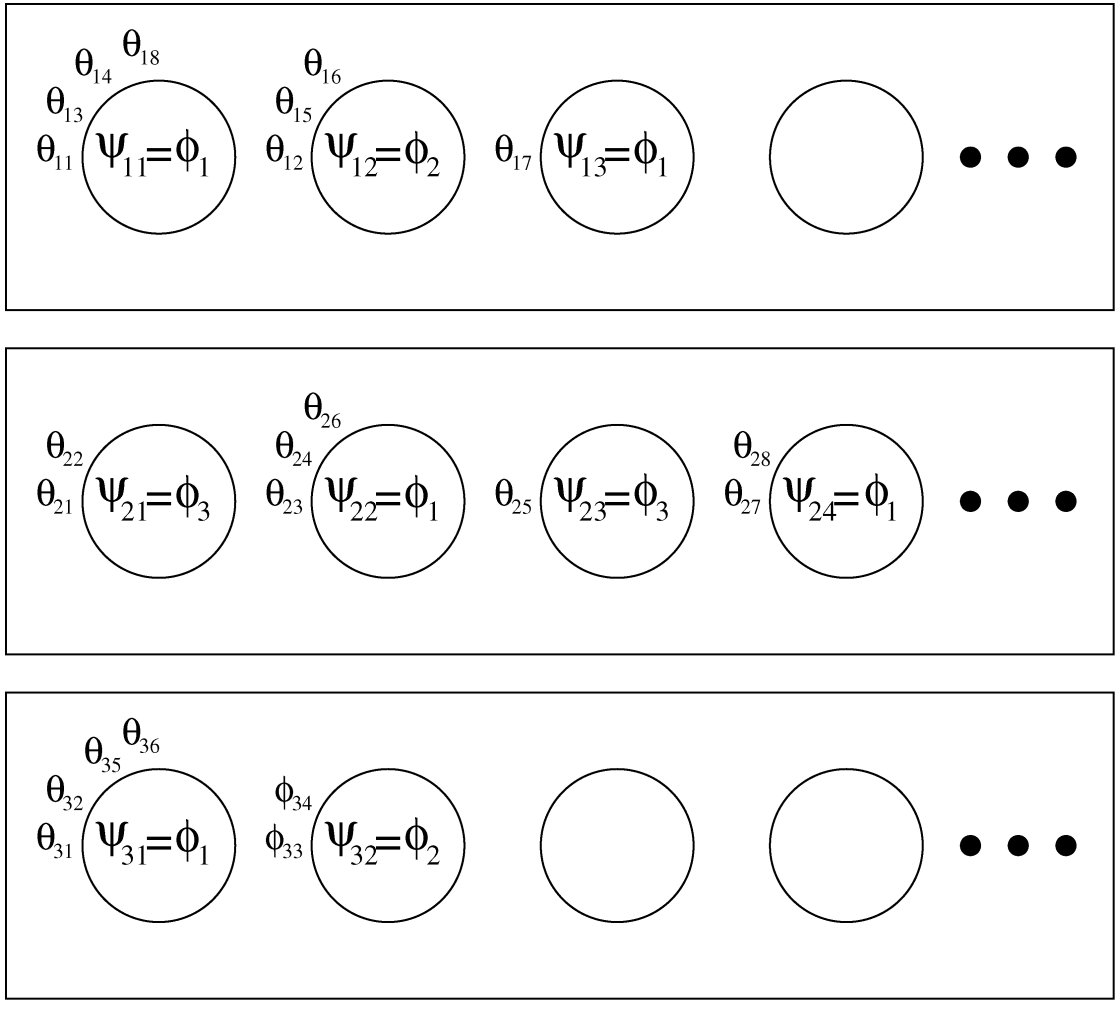




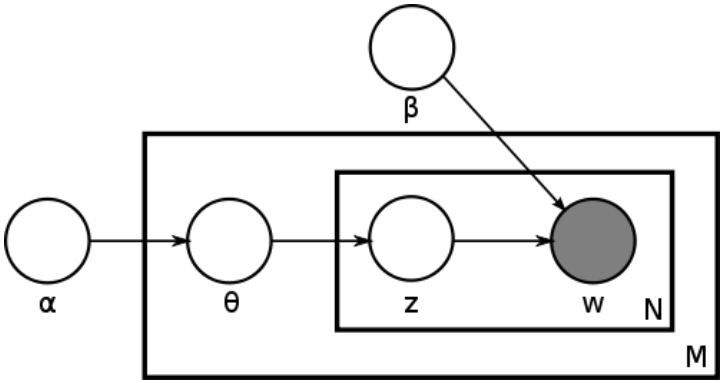
Stick Breaking Construction

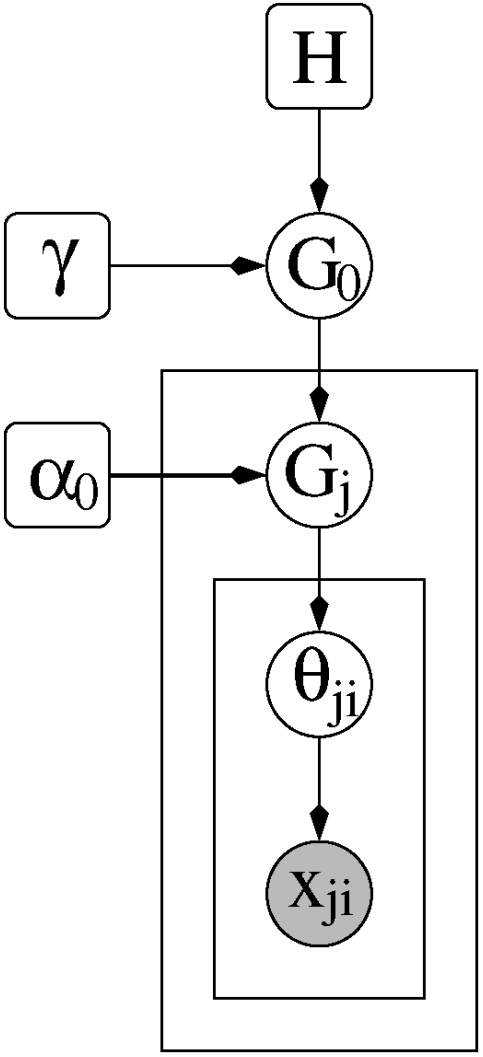
- $G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}$, $\phi_k \sim H$
- $G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_k}$, $\phi_k \sim G_0$
- $\pi'_{jk} \sim Beta(\alpha_0 \beta_k, \alpha_0 (1 - \sum_{i=1}^{k} \beta_i))$
- $\pi_{jk} = \pi'_{jk} \sum_{i=1}^{k-1} (1 - \pi'_{ji})$

Chinese Restaurant Franchise

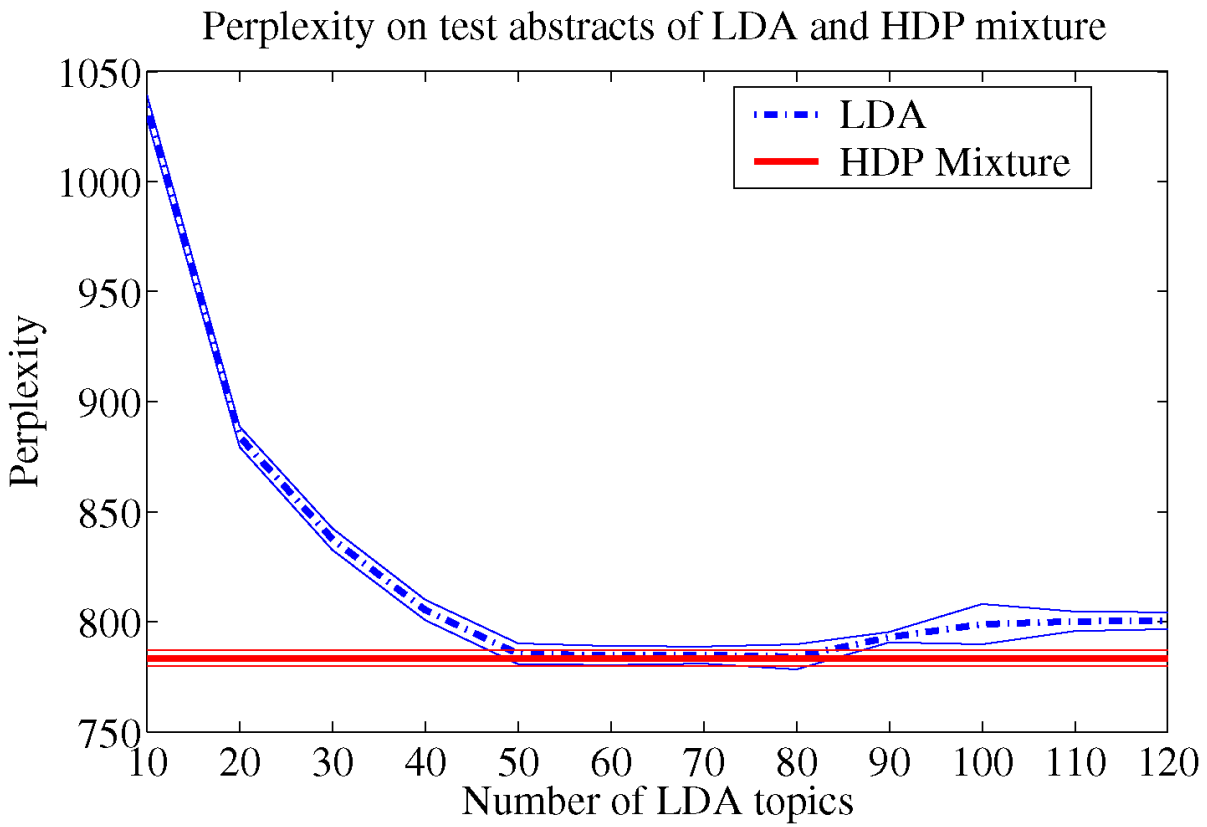


Document Modeling



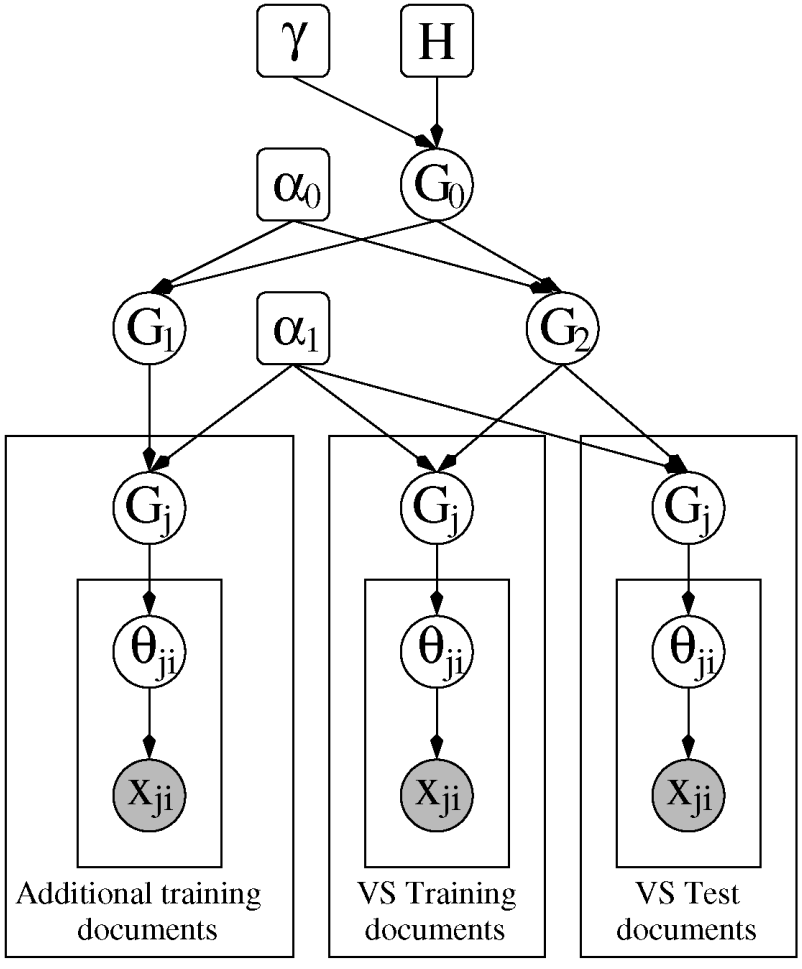


LDA vs HDP



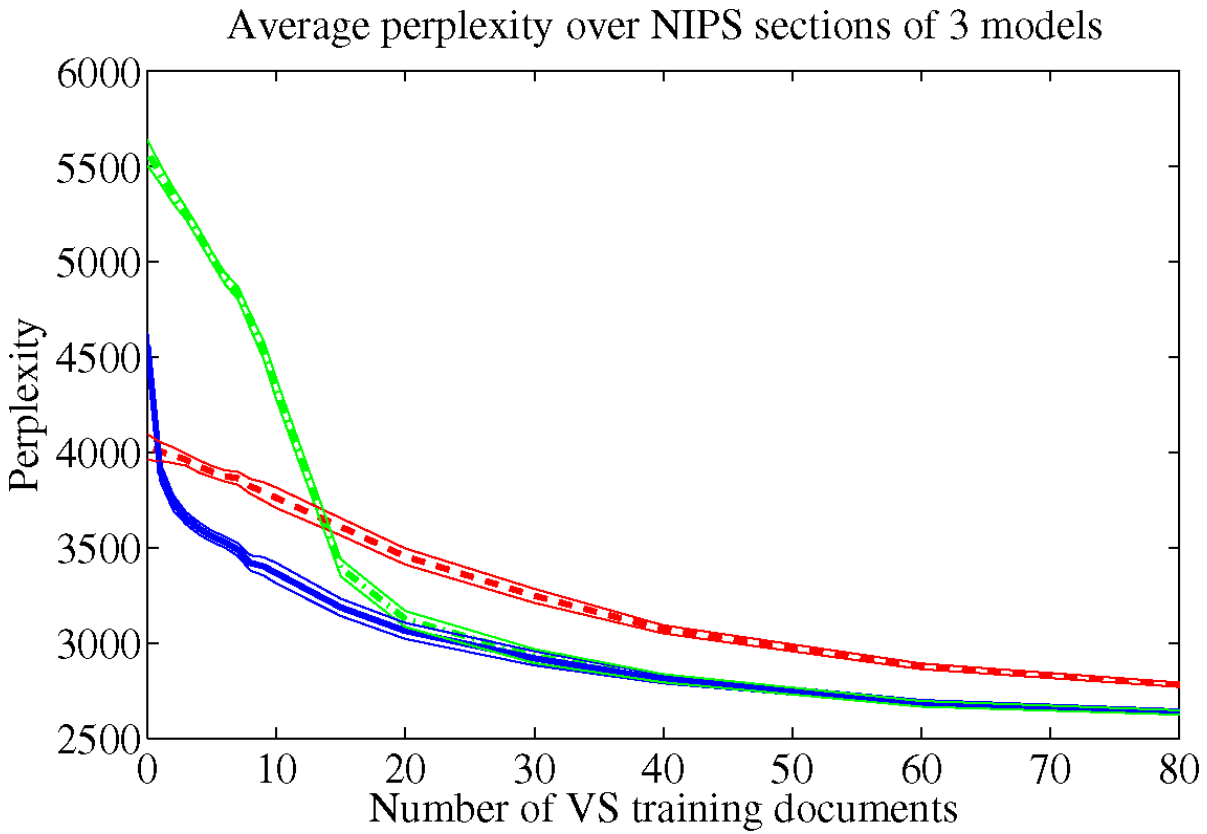
LDA used a symmetric prior

Speeding Sampling with Extra Data

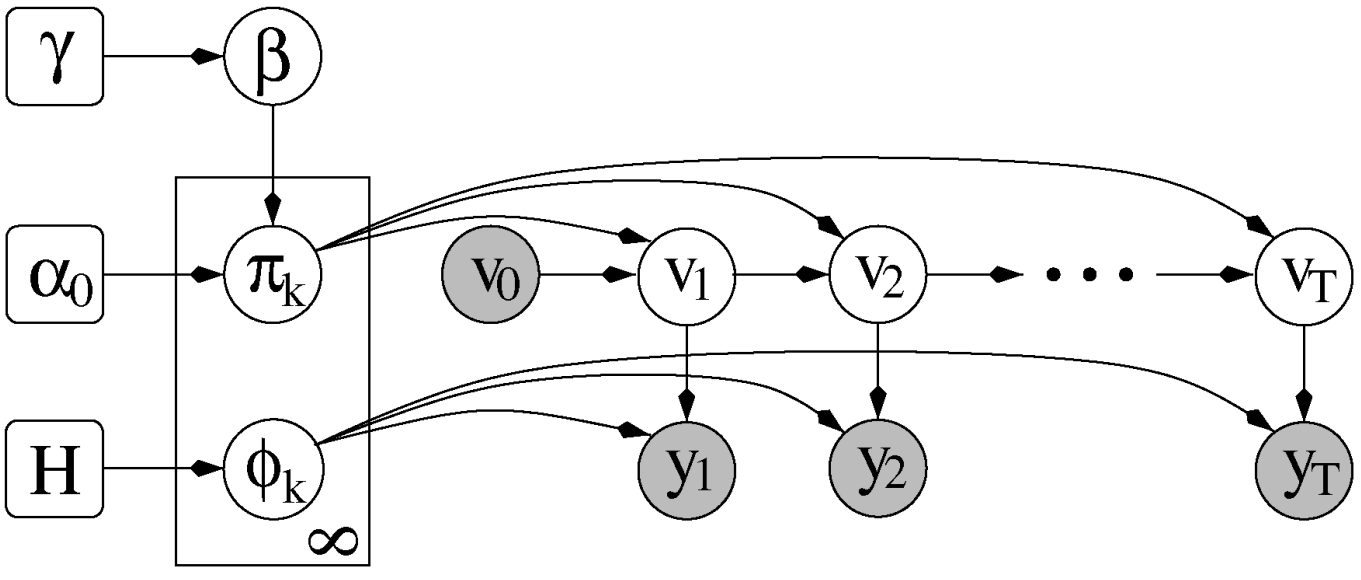


M3

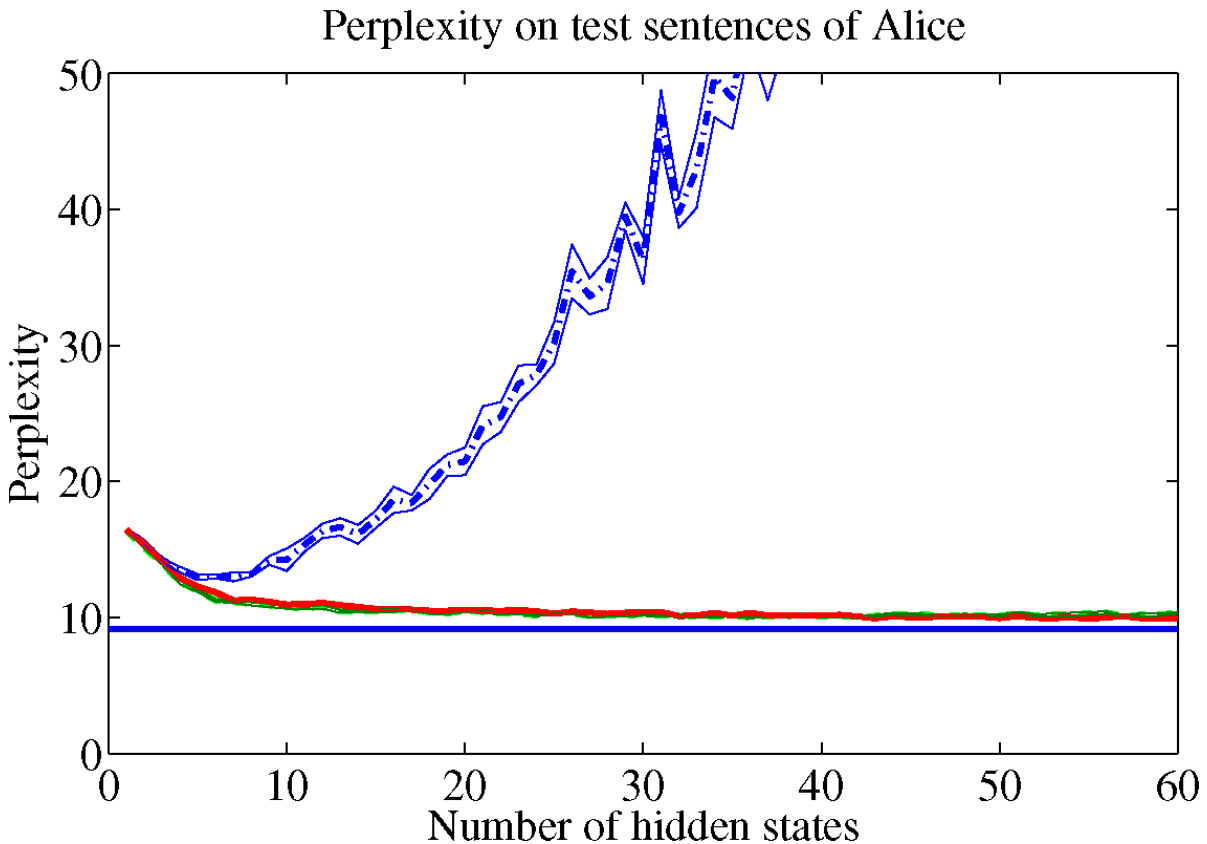
Results


Average perplexity over NIPS sections of 3 models
Perplexity
6000
5500
5000
4500
4000
3500
3000
2500
0
10
20
30
40
50
60
70
80
Number of VS training documents

HMM-HDP



Alice in Wonderland



Posterior Sampling

- Consider each table with probability $\propto n_{jt\bullet}^{-ji} f_{k_{jt}}^{-ji}(x_{ji})$
- Consider a new table $\propto \alpha_0 p(x_{ji}|\mathbf{t}^{-ji}, t = t^{new}, \mathbf{k})$
- If a new table is chosen, pick its ϕ_k using G_0
- If an existing table is chosen, resample its ϕ_k

HDP-HMM

- Keeping track of $m_{\bullet k}, t_{ji}, k_{jt}$

$$Data + Tables + Nodes \geq Size \leq Data + Tables + (Nodes * Nodes)$$

Augmented Posterior Sampling

- Instantiate G_0 using the Polya-Urn metaphor - $G_0 = \sum_{k=1}^{K} \beta_k \delta_{\phi_k} + \beta_u G_u$
- $\beta = (\beta_1, \beta_2, ..., \beta_K, \beta_u) \sim Dir(m_{\bullet 1}, m_{\bullet 2}, ..., \gamma)$
- Gets an estimate of G_0 in between samplings of $(\mathbf{t}, \mathbf{k})$

HDP-HMM

- Keep track of t_{ji}, k_{jt}
- Use β_k to represent all $m_{\bullet k}$'s

$Size = Data + Tables + Nodes$

Direct Assignment Posterior Sampling

- Same as above, but only keep track of how many times each ϕ_k is used
- Each datum x_{ji} gets an index z_{ji} into our current ϕ_k list
- Since k is not being resampled, every sample only changes one data item's mixture component
- The critical case is when a k is reestimated to a different previously chosen k

HDP-HMM

- Combine tables with the same k to get z_{ji}
- Using β_k to represent all $m_{\bullet k}$'s

$Size = Data + Nodes$

Parameter Resampling (Math)

Modification of Escobar and West 1995

- $q(\alpha_0|\mathbf{w}, \mathbf{s}) \propto Gamma(a - m_{\bullet\bullet} - \sum_{j=1}^{J} s_j, b - \sum_{j=1}^{J} \log w_j)$
- $q(w_j|\alpha_0) \propto Beta(\alpha_0 + 1, n_{j\bullet\bullet})$
- $q(s_j|\alpha_0) \propto Bin(\frac{n_{j\bullet\bullet}}{n_{j\bullet\bullet}+\alpha_0})$

Parameter Resampling (MATLAB)

From utilities/randconparam.m, in npbayes

```
function alpha = randconparam(alpha,numdata,numclass,aa,bb,numiter);

totalclass = sum(numclass);
num = length(numdata);

for ii = 1:numiter
  %make a num length vector of beta draws from a+1,n
  %this is getting w_j for each group
  xx = randbeta((alpha+1)*ones(1,num),numdata);

  %to sample s_j its a binary with prob (a / a+n) vs (n / a+n)
  %so this is a list of s_j's for each group
  zz = rand(1,num).*(alpha+numdata)<numdata;

  %totalclass is m.., the number of tables in all restaurants.
  gammaa = aa + totalclass - sum(zz);
  gammab = bb - sum(log(xx));

  %these should be the params of a gamma distribution we want to
  %sample from, but what is going on under the hood here?
  alpha = randgamma(gammaa)./gammab;

end
```

Unsupervised POS Tagging

Problem

- Given a corpus of sentences, assign each word a class (POS Tag)
- Assume Markov dependency and model with an HMM

Infinite Solution

- HDP Prior on state transitions
- x_{i-1} - Restaurant
- x_i - Table in x_{i-1}'s Restaurant
- $P(y_i|x_i)$ - ϕ_k

The infinite HMM for unsupervised POS Tagging, VanGael 2009

Infinte BiGram Model

Problem

- Recover segmentation given only unsegmented text (e.g. Chinese)
- Use a bigram language model of words w

Infinite Solution

- w_{i-1} - Restaurant
- w_i - Table
- δ_{w_i} - ϕ_k

Contextual dependencies in unsupervised word segmentation, Goldwater 2006 (actual model slightly more complex to encourage compact representation)

Infinte Tree

Problem

- Given a set of sentences, assign each word a class (POS Tag)
- Assume tree structure and model with a PCFG
- Every node is a preterminal, expanded with $A \longrightarrow Y\mathbf{B}$

The Infinite Tree, Finkel et al 2007

Other Infinite Trees

- Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models, Johnson et al 2007
- The infinite PCFG using hierarchical Dirichlet processes, Liang et al 2007

The End

- $G_0|\gamma, H \sim DP(\gamma, H)$
- $G_j|\alpha_0, G_0 \sim DP(\alpha_0, G_0)$