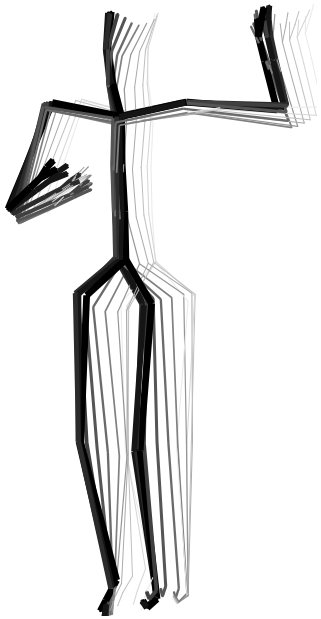
Toward Reliable Bayesian Nonparametric Learning



Erik Sudderth

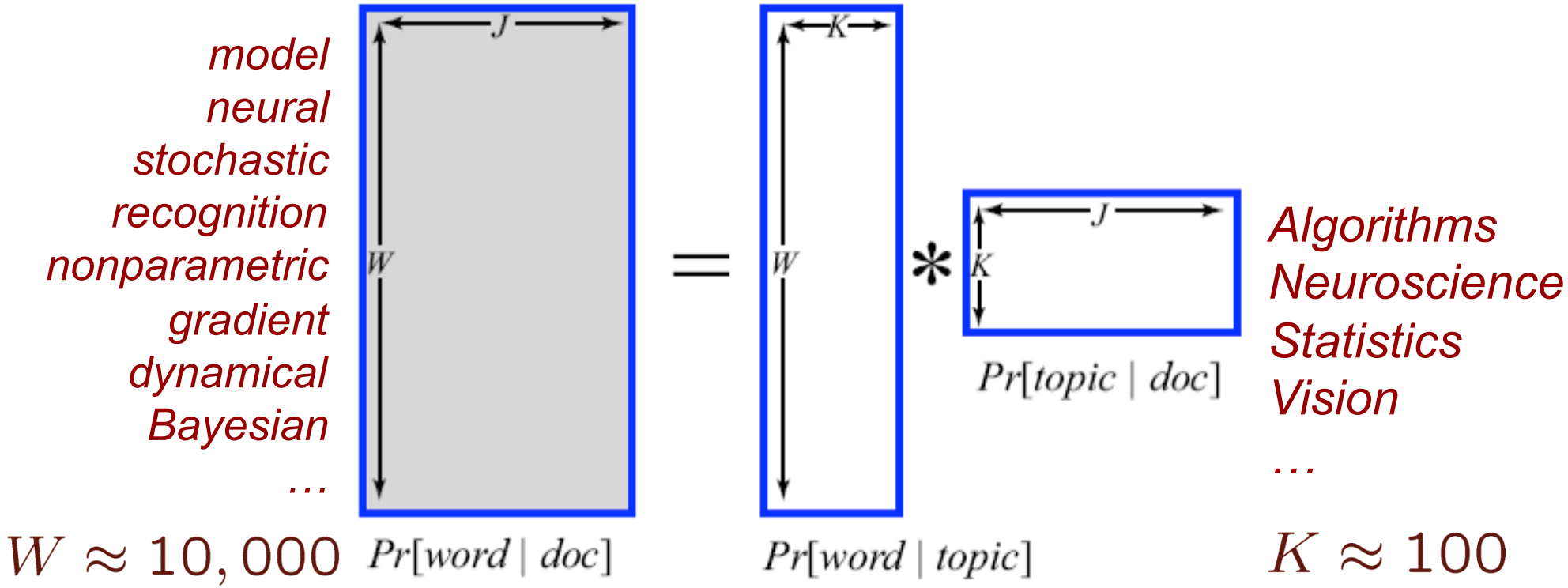
Brown University
Department of Computer Science

Joint work with

Donglai Wei & Michael Bryant (HDP topics)
Michael Hughes & Emily Fox (BP-HMM)

Documents & Topic Models

Framework for unsupervised discovery of *low-dimensional* latent structure from *bag of word* representations



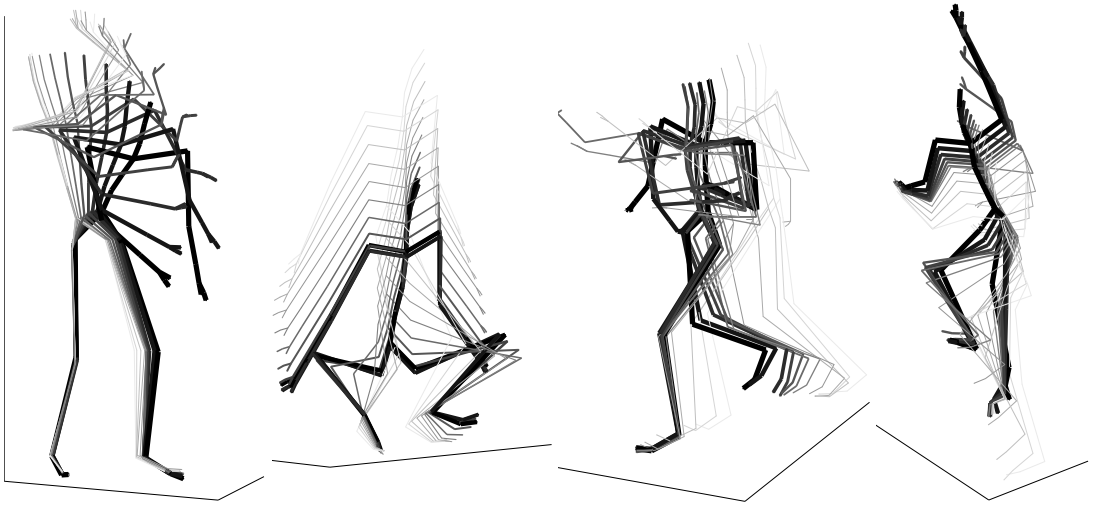
- **pLSA**: Probabilistic Latent Semantic Analysis *(Hofmann 2001)*
- **LDA**: Latent Dirichlet Allocation *(Blei, Ng, & Jordan 2003)*
- **HDP**: Hierarchical Dirichlet Processes *(Teh, Jordan, Beal, & Blei 2006)*

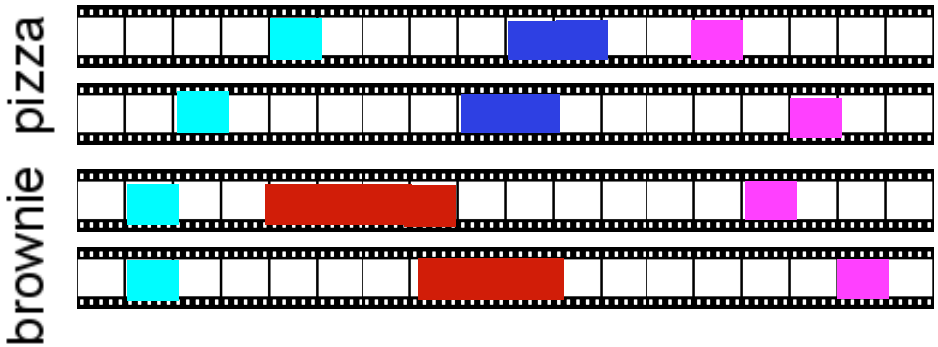
Temporal Activity Understanding

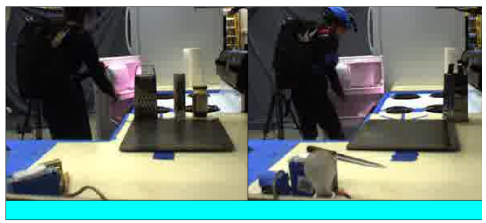
To organize large time series collections, an essential task is to
Identify segments whose visual content arises from same physical cause

GOAL: Set of temporal behaviors

- Detailed segmentations
- *Sparse* behavior sharing
- Nonparametric recovery & growth of model complexity
- Reliable general-purpose tool across domains






Open Fridge


Grate Cheese


Set Oven Temp.


Stir Brownie Mix

Learning Challenges

Can local updates uncover global structure?

- ***MCMC:*** Local Gibbs and Metropolis-Hastings proposals
- ***Variational:*** Local coordinate ascent optimization
- Do these algorithms live up to our complex models???

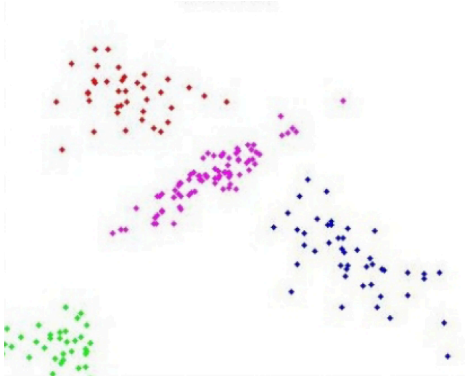
Non-traditional modeling and inferential goals

- ***Nonparametric:*** Model structure grows and adapts to new data, no need to specify number of topics/objects/etc.
- ***Reliable:*** Our primary goal is often not prediction, but correct recovery of latent cluster/feature structure
- ***Simple:*** Often want just a single “good” model, not samples or a full representation of posterior uncertainty

Outline

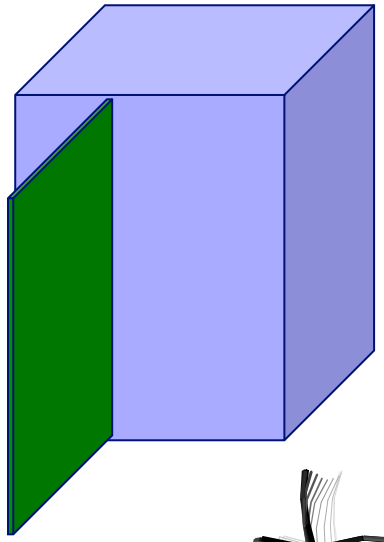
Bayesian Nonparametrics

- Dirichlet process (DP) mixture models
- Variational methods and the ME algorithm



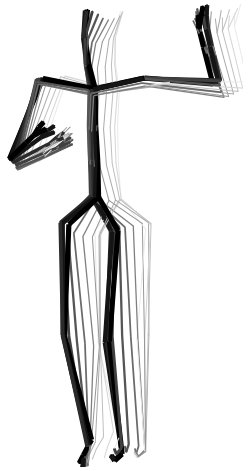
Reliable Nonparametric Learning

- Hierarchical DP topic models
- ME search in a collapsed representation
- Non-local online variational inference



Nonparametric Temporal Models

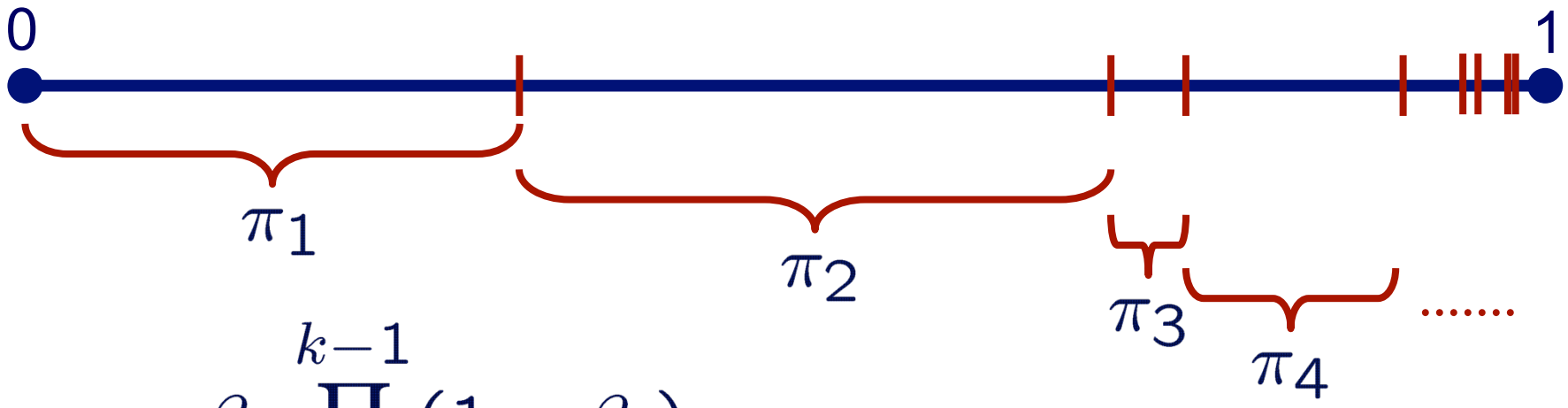
- Beta Process Hidden Markov Models (BP-HMM)
- Effective split-merge MCMC methods



Stick-Breaking and DP Mixtures

$$p(x) = \sum_{k=1}^{\infty} \pi_k f\left(x \mid \theta_k\right)$$

Dirichlet process implies a prior distribution on the weights of a countably infinite mixture:



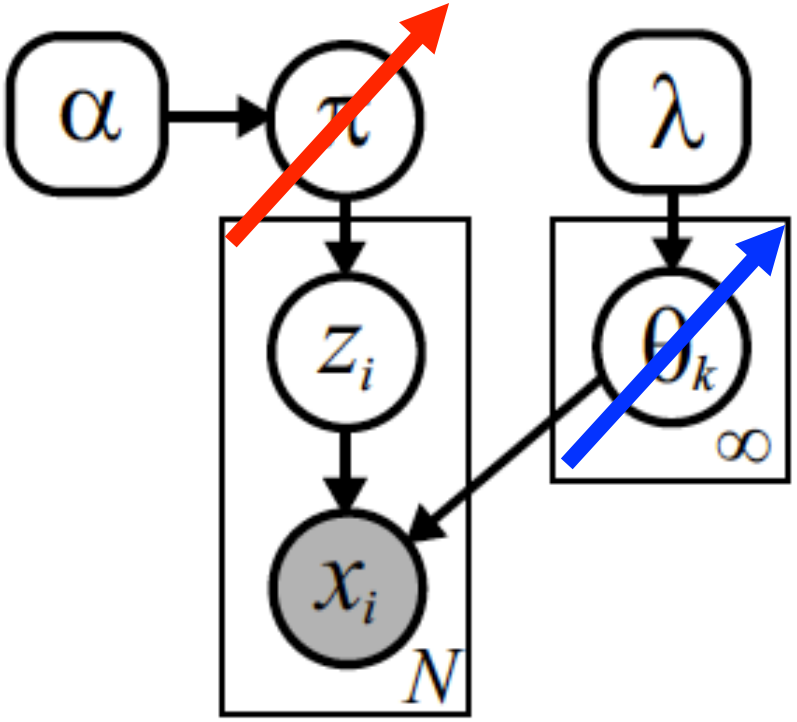
$$\pi_k = \beta_k \prod_{\ell=1}^{k-1} (1 - \beta_\ell)$$

$$\beta_k \sim \text{Beta}(1, \alpha)$$

α → concentration parameter

Sethuraman, 1994

Clustering and DP Mixtures



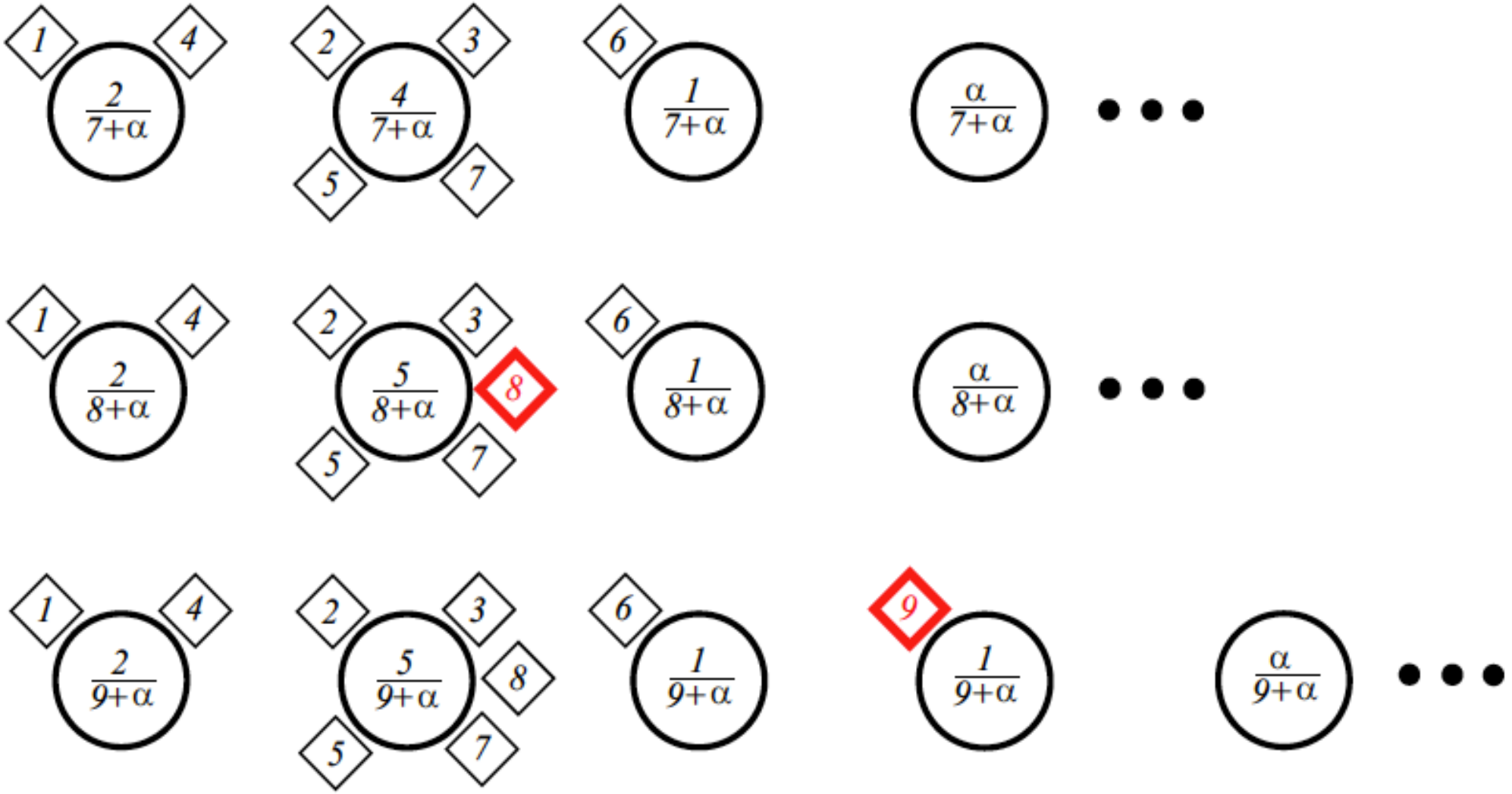
$$\pi \sim \mathrm{GEM}(\alpha)$$
$$\theta_k \sim H(\lambda) \qquad k = 1, 2, \ldots$$

$z_i \sim \pi$ *Indicates which cluster generated each observation*

$x_i \sim F(\theta_{z_i})$ *N data points observed*

- Conjugate priors allow marginalization of cluster parameters
- Marginalized cluster sizes induce Chinese restaurant process

Chinese Restaurant Process



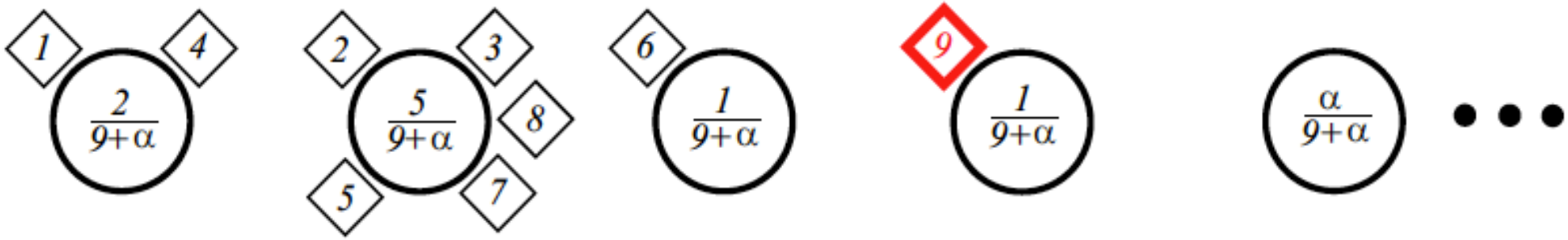
$$p(z_{N+1} = z \mid z_1, \ldots, z_N, \alpha) = \frac{1}{\alpha + N} \left(\sum_{k=1}^{K} N_k \delta(z, k) + \alpha \delta(z, \bar{k}) \right)$$

DP Mixture Marginal Likelihood

Closed form probability for any hypothesized partition of N observations into K clusters:

$$\log p(x, z) = \log \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)} + \sum_{k=1}^{K} \left\{ \log \alpha + \log \Gamma(N_k) + \log \int_{\Theta} \prod_{i|z_i=k} f(x_i \mid \theta_k)\, dH(\theta_k) \right\}$$

$$\Gamma(N_k) = (N_k - 1)!$$



$$p(z_{N+1} = z \mid z_1, \ldots, z_N, \alpha) = \frac{1}{\alpha + N} \left(\sum_{k=1}^{K} N_k \delta(z, k) + \alpha \delta(z, \bar{k}) \right)$$

DP Mixture Inference

Monte Carlo Methods

- Stick-breaking representation: Truncated or slice sampler
- CRP representation: Collapsed Gibbs sampler
- Split-merge samplers, retrospective samplers, …

Variational Methods

$$\log p(x \mid \alpha, \lambda) \geq H(q) + \mathbb{E}_q[\log p(x, z, \theta \mid \alpha, \lambda)]$$

- Valid for any hypothesized distribution $q(z, \theta)$
- Mean field variational methods optimize in tractable family
- Truncated stick-breaking representation: *Blei & Jordan, 2006*
- Collapsed CRP representation: *Kurihara, Teh, & Welling 2007*

Maximization Expectation

EM Algorithm

- E-step: Marginalize latent variables (approximate)
- M-step: Maximize likelihood bound given model parameters

ME Algorithm

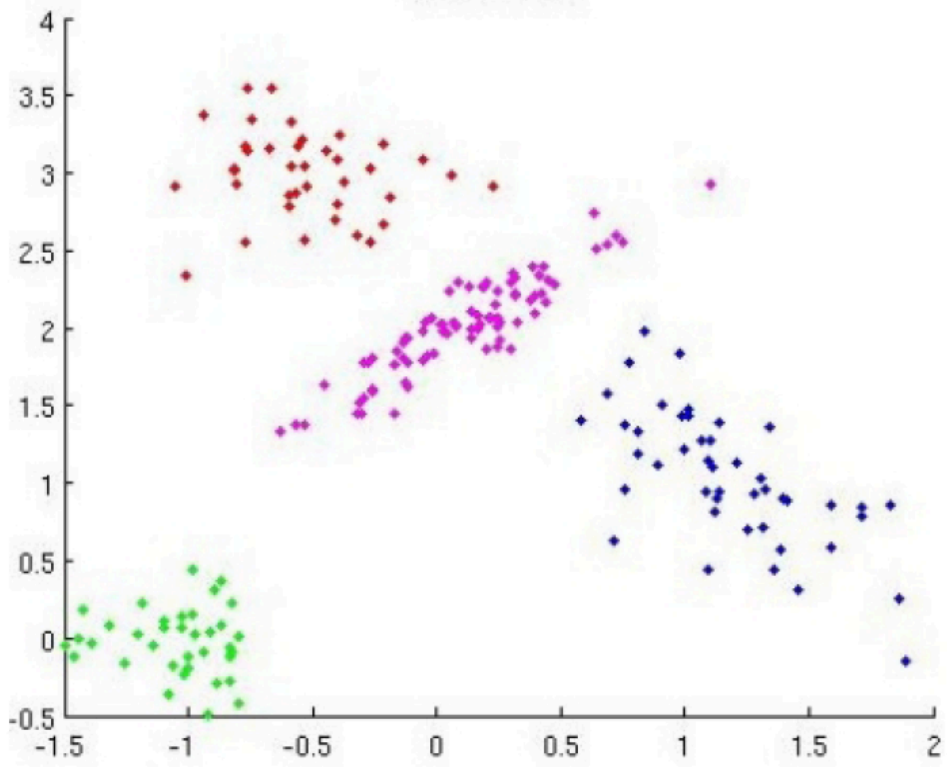
Kurihara & Welling, 2009

- M-step: Maximize likelihood given latent assignments
- E-step: Marginalize random parameters (exact)

Why Maximization-Expectation?

- Parameter marginalization allows Bayesian “model selection”
- Hard assignments allow efficient algorithms, data structures
- Hard assignments consistent with clustering objectives
- *No need for finite truncation of nonparametric models*

A Motivating Example

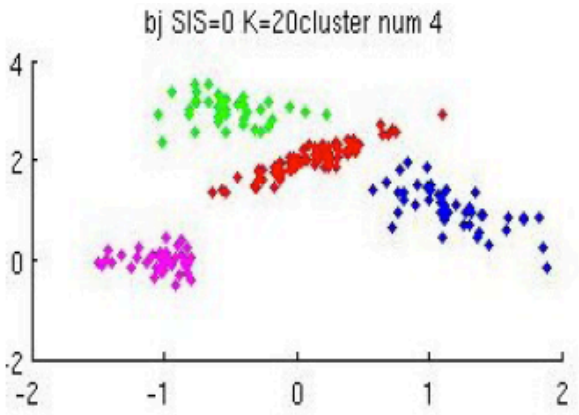


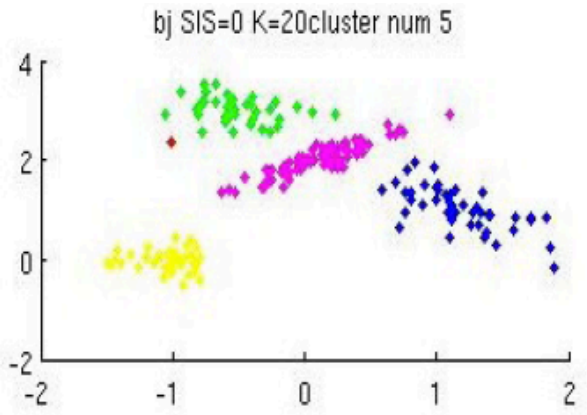
200 samples from a mixture of 4 two-dimensional Gaussians

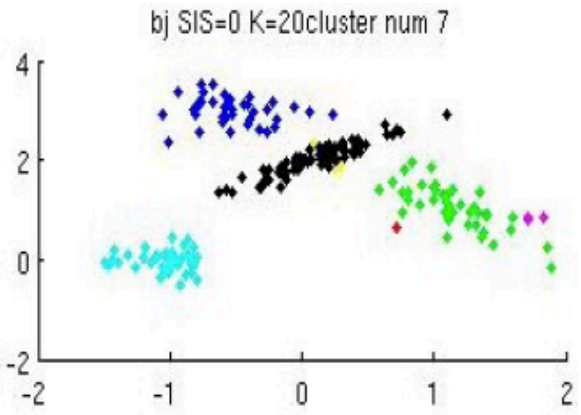
- Stick-breaking variational: Truncate to K=20 components
- CRP collapsed variational: Truncate to K=20 components
- ME local search: No finite truncation required

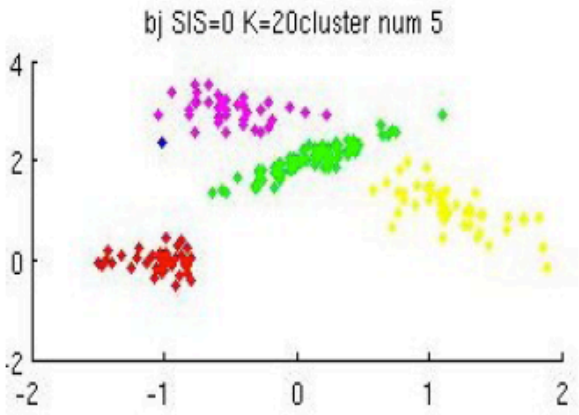
$$\log p(x,z) = \log \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)} + \sum_{k=1}^{K} \left\{ \log \alpha + \log \Gamma(N_k) + \log \int_{\Theta} \prod_{i|z_i=k} f(x_i \mid \theta_k)\, dH(\theta_k) \right\}$$

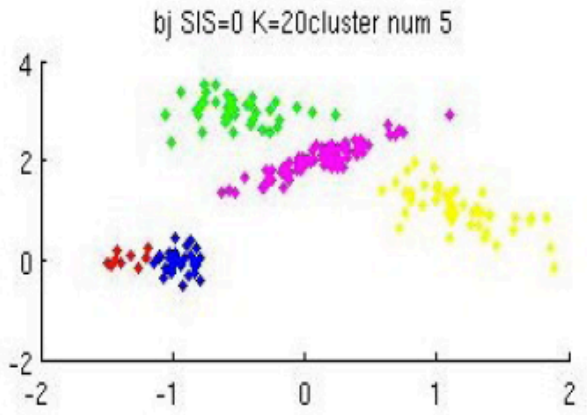
Stick-Breaking Variational

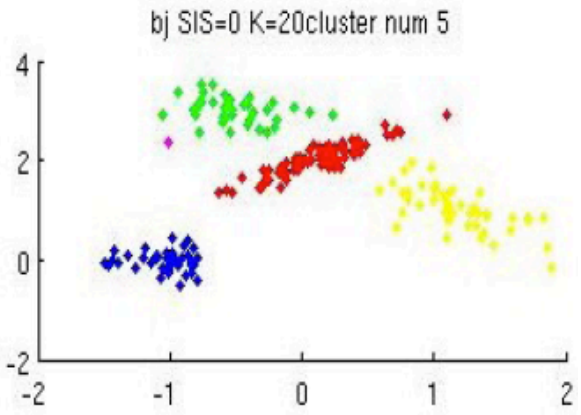


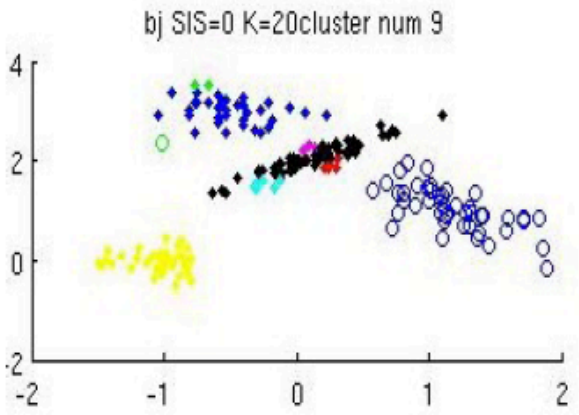


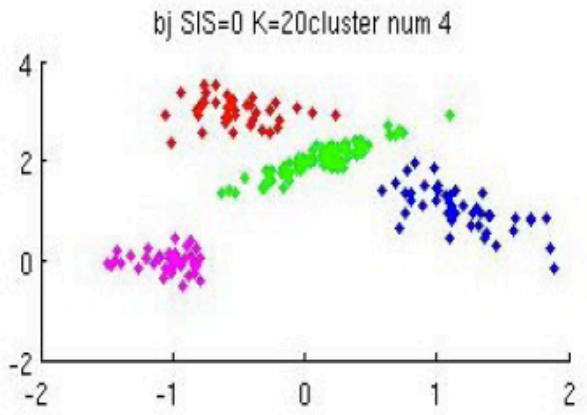


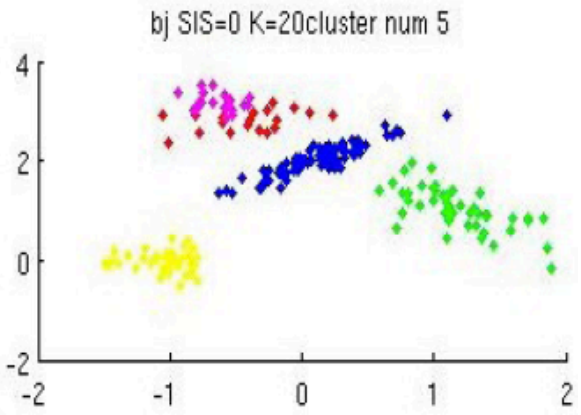




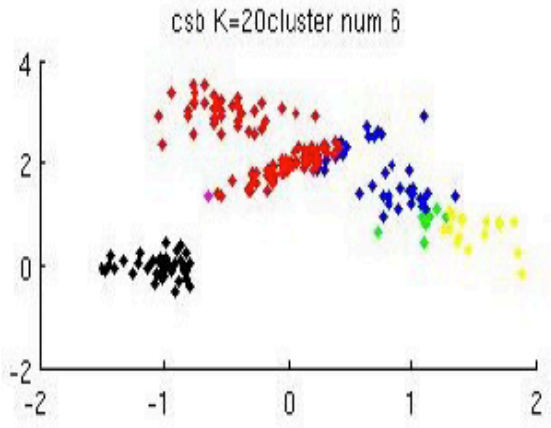


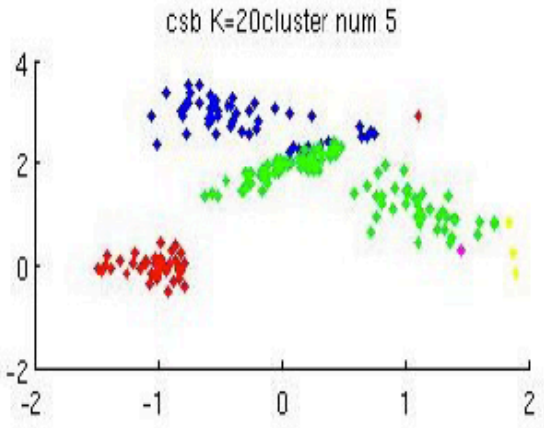


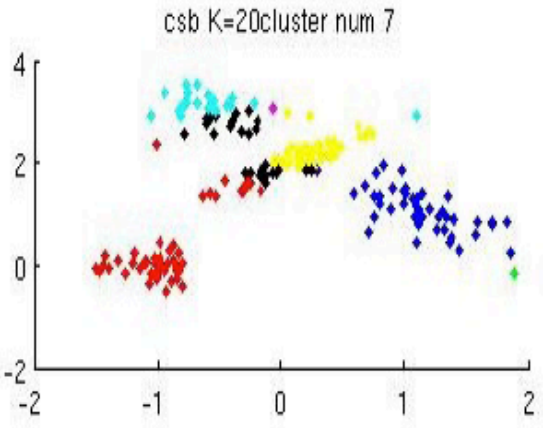


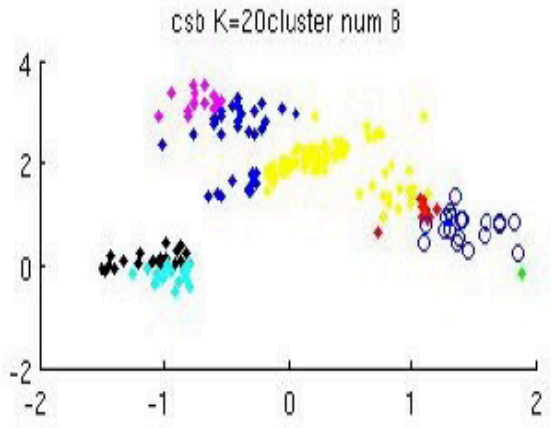


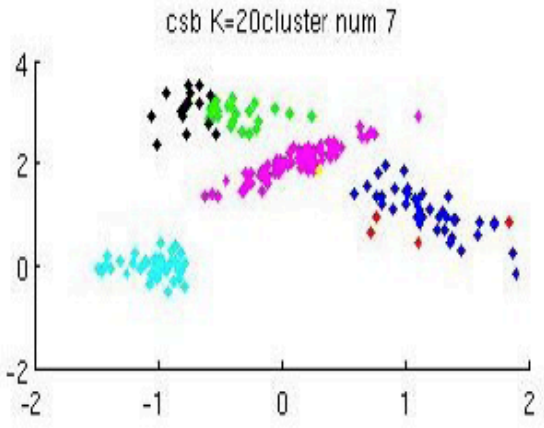
Collapsed Variational

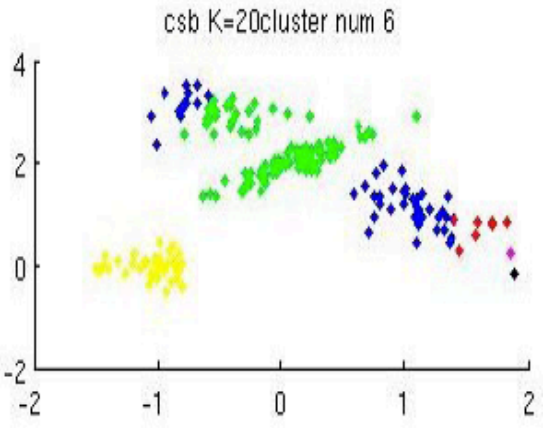


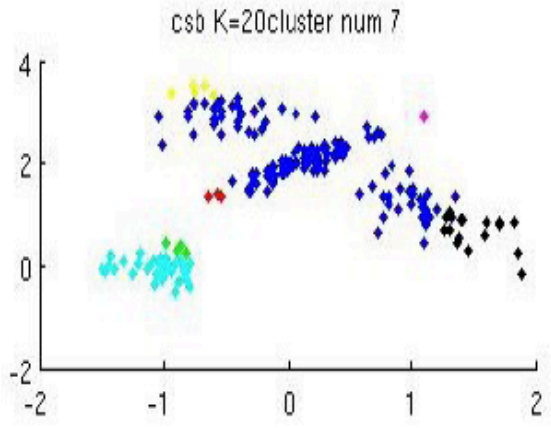


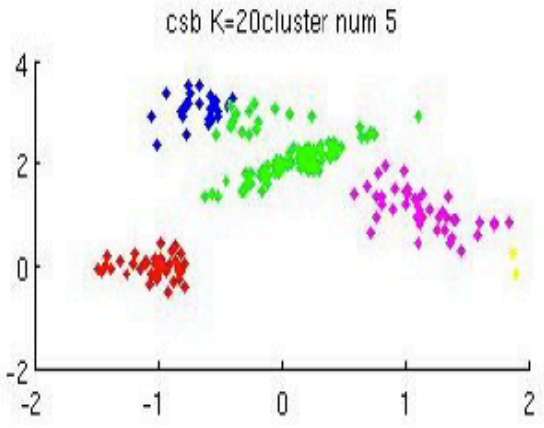


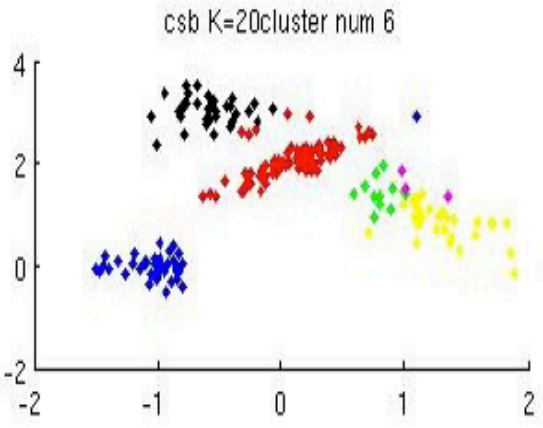




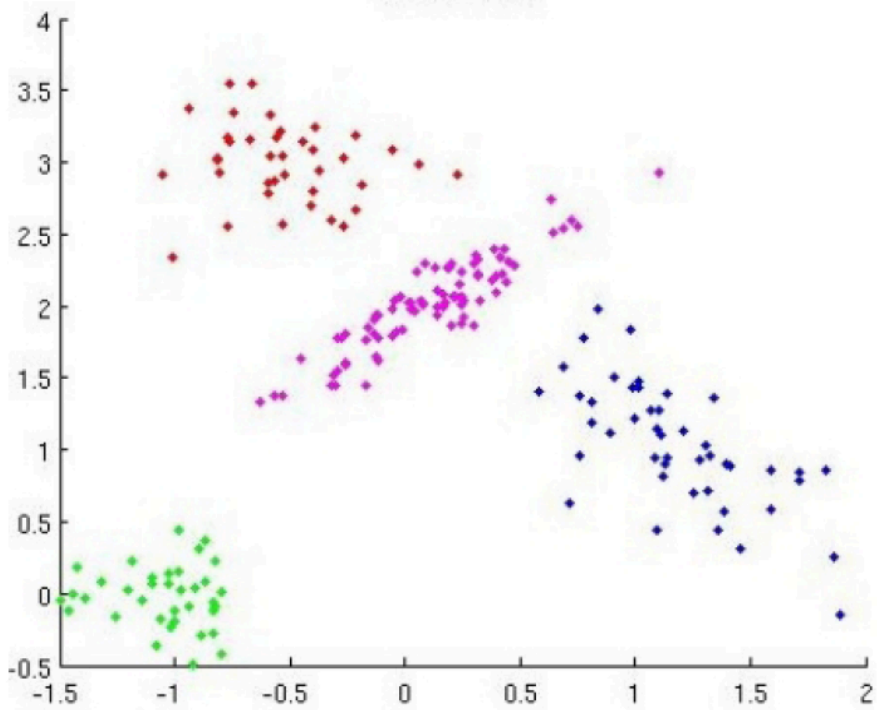








ME Local Search with Merge



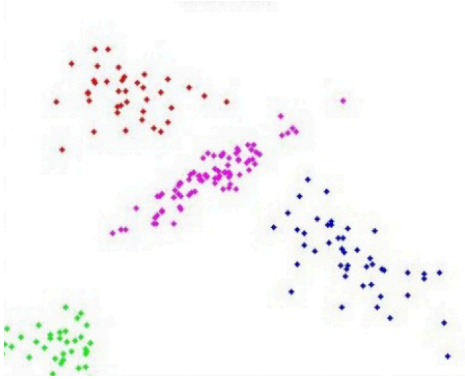
Every run, from hundreds of initializations, produces the same (optimal) partition

- Dynamics of inference algorithm often matter more in practice than choice of model representation/approximation
  - True for MCMC as well as variational methods
- Easier to design complex algorithms for simple objectives

Outline

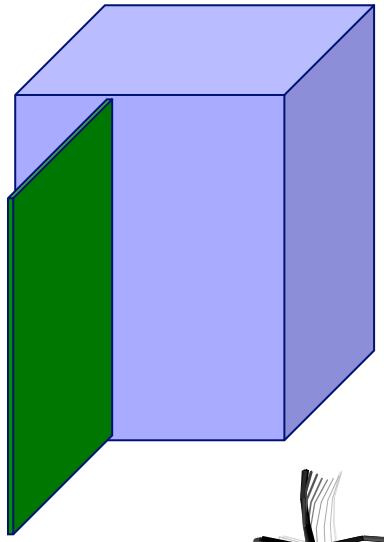
Bayesian Nonparametrics

- Dirichlet process (DP) mixture models
- Variational methods and the ME algorithm



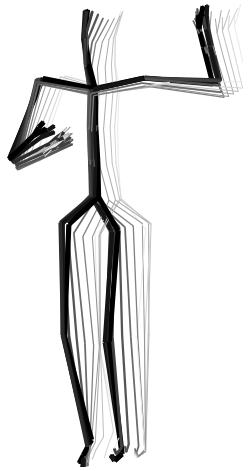
Reliable Nonparametric Learning

- Hierarchical DP topic models
- ME search in a collapsed representation
- Non-local online variational inference

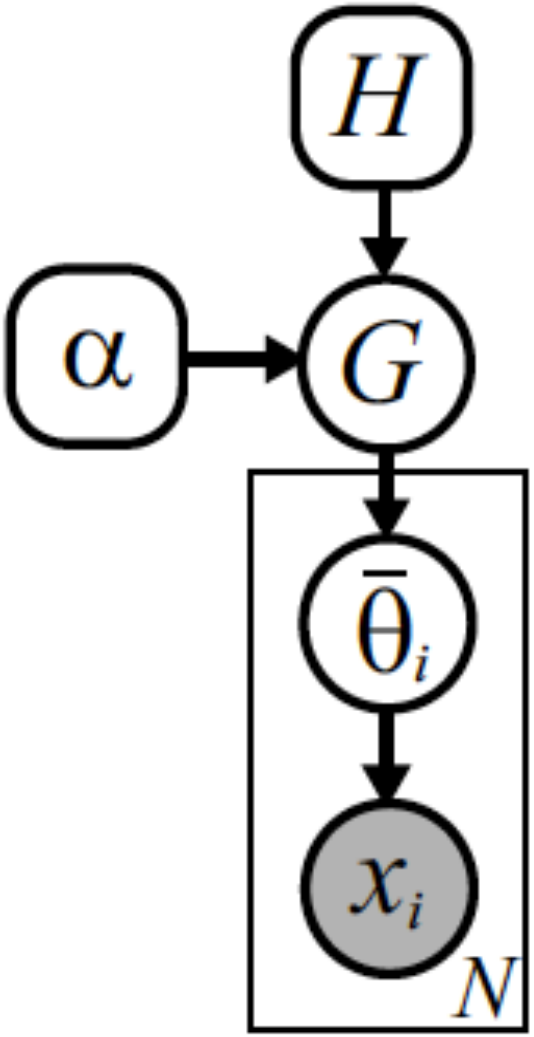


Nonparametric Temporal Models

- Beta Process Hidden Markov Models (BP-HMM)
- Effective split-merge MCMC methods



Distributions and DP Mixtures



$$G \sim \mathrm{DP}(\alpha, H)$$

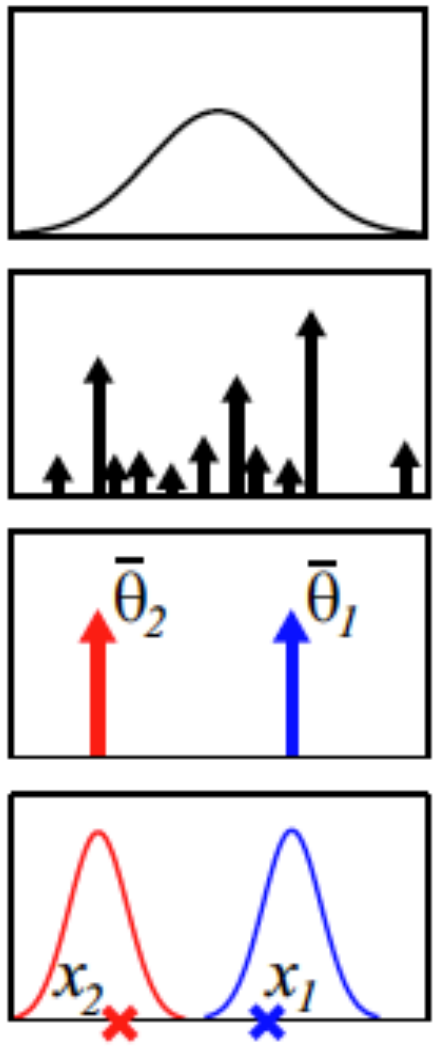
$$G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta(\theta, \theta_k)$$

$$\pi \sim \mathrm{GEM}(\alpha)$$

$$\theta_k \sim H(\lambda) \qquad k = 1, 2, \ldots$$

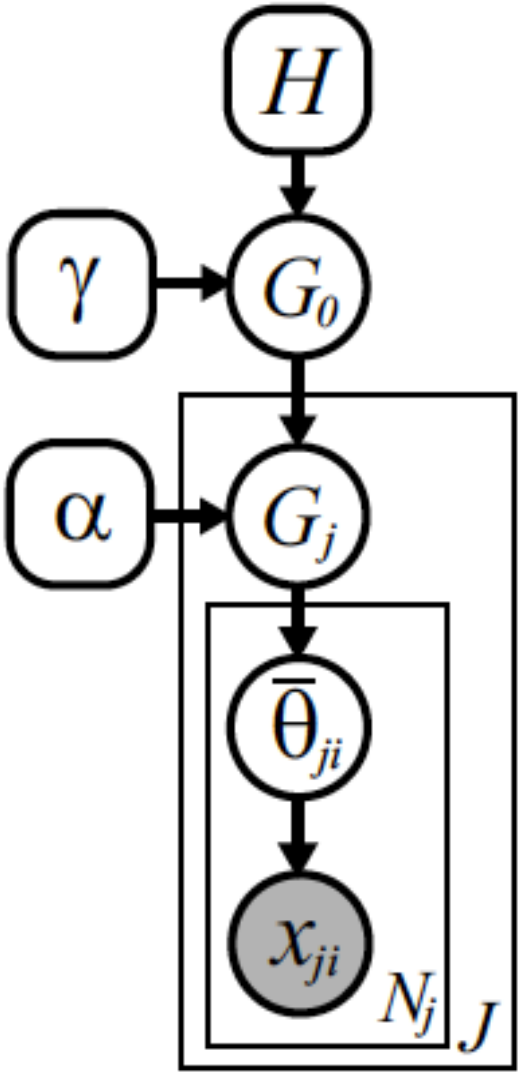
$$\bar{\theta}_i \sim G$$

$$x_i \sim F(\bar{\theta}_i)$$



Ferguson, 1973
Antoniak, 1974

Distributions and HDP Mixtures



Global discrete measure:

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

Atom locations define topics, atom masses their frequencies.

For each of J groups:

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

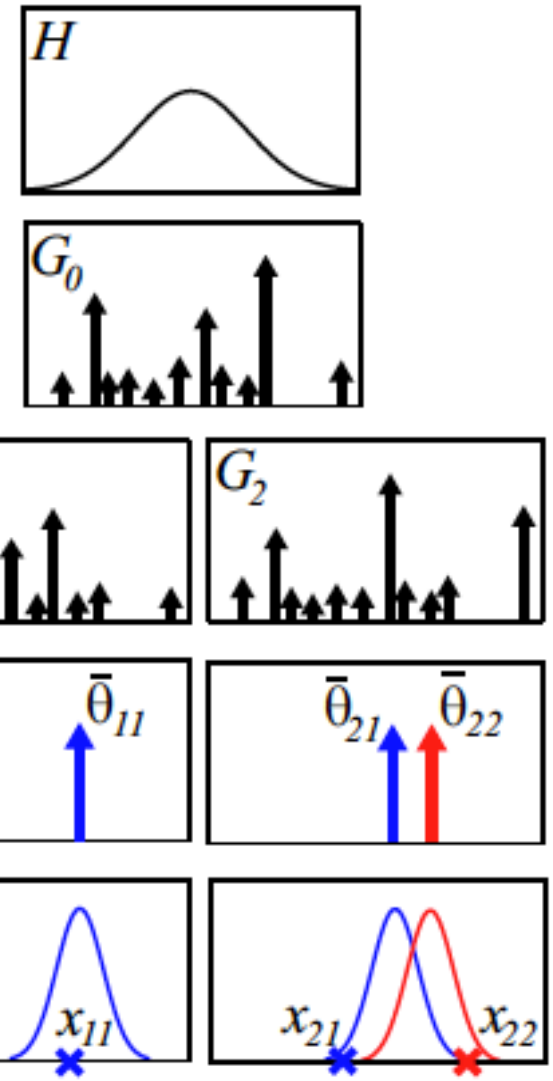
Each document has its own topic frequencies.

For each of N_j data:

$$\bar{\theta}_{ji} \sim G_j$$

$$x_{ji} \sim F(\bar{\theta}_{ji})$$

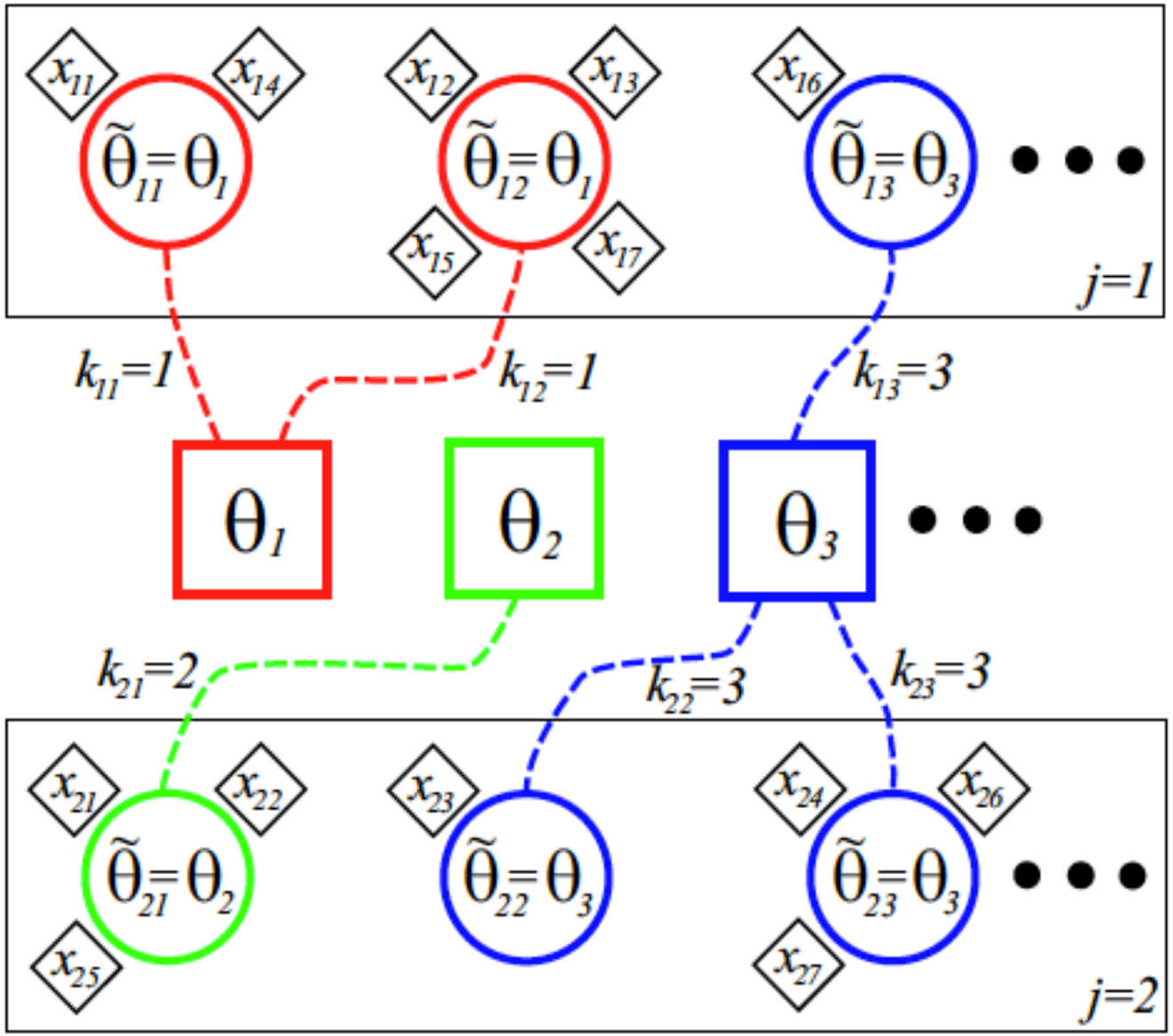
Bag of word tokens.



Hierarchical Dirichlet Process *(Teh, Jordan, Beal, & Blei 2004)*

- Instance of a dependent Dirichlet process *(MacEachern 1999)*
- Closely related to Analysis of Densities *(Tomlinson & Escobar 1999)*

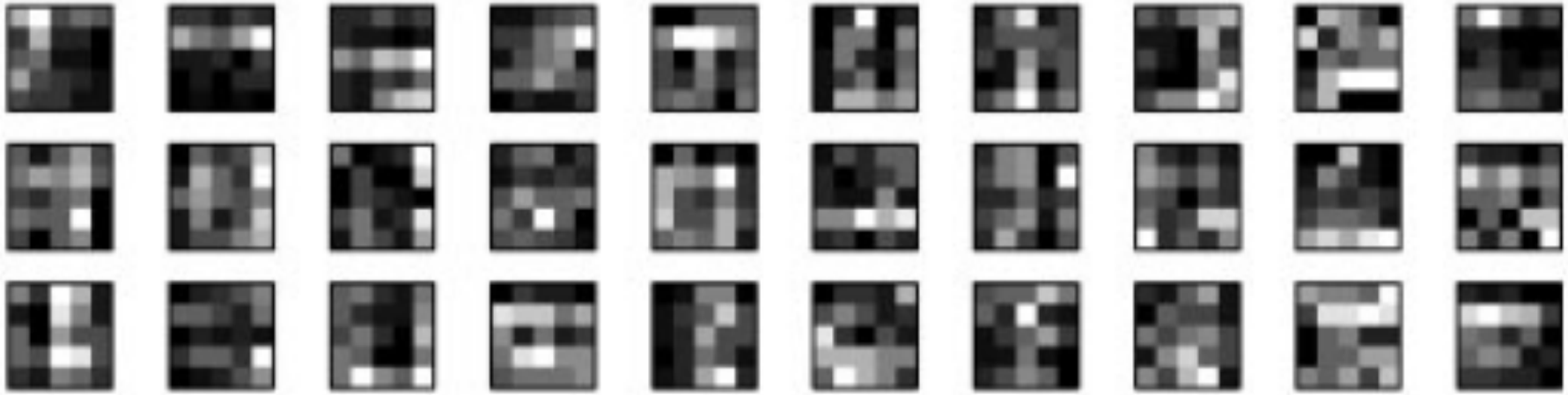
Chinese Restaurant Franchise



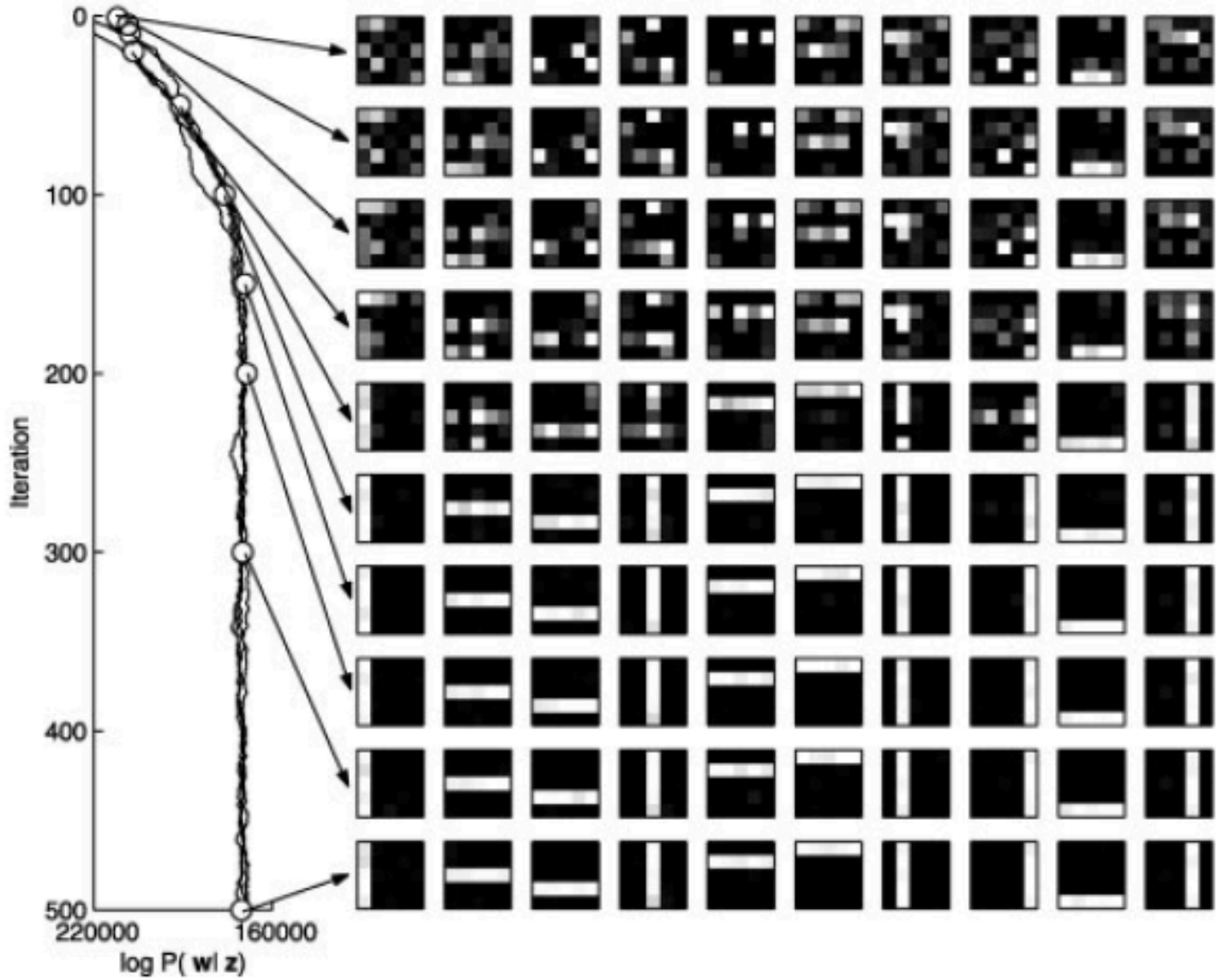
The Toy Bars Dataset

- Latent Dirichlet Allocation *(LDA, Blei et al. 2003)* is a parametric topic model (a finite Dirichlet approximation to the HDP)
- Griffiths & Steyvers *(2004)* introduced a collapsed Gibbs sampler, and demonstrated it on a toy "bars" dataset:

10 topic distributions on 25 vocabulary words, and example documents

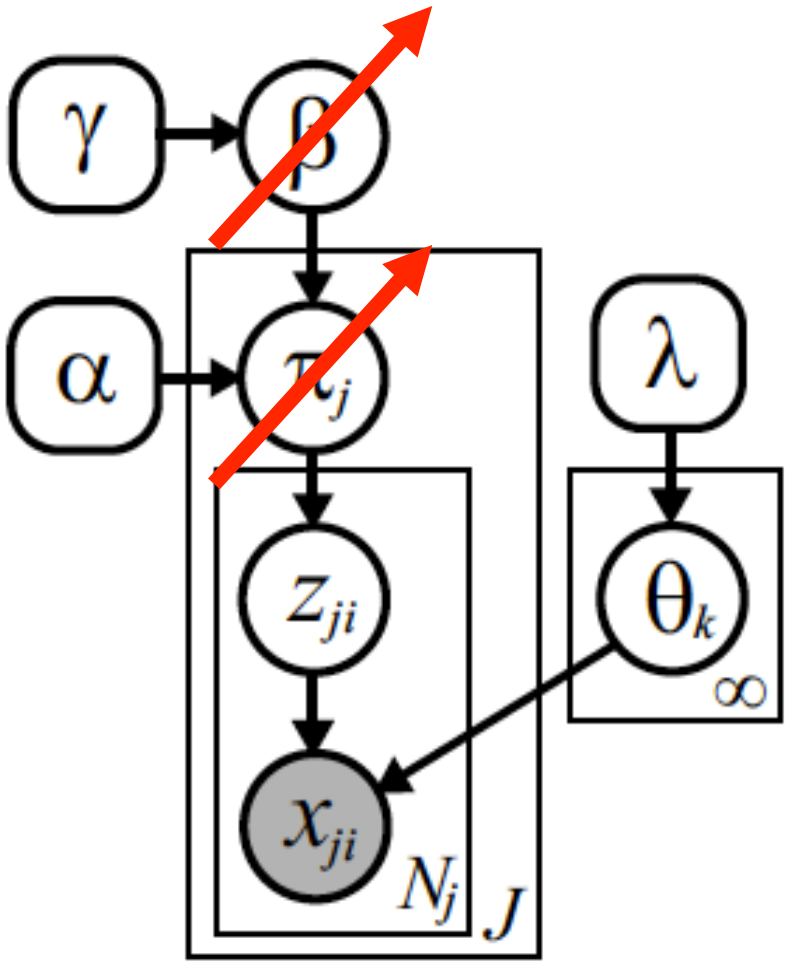


The Perfect Sampler?



$$P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

Direct Cluster Assignments



Can we marginalize both global and document-specific topic frequencies?

Global discrete measure:

$$G_0(\theta) = \sum_{k=1}^{\infty} \beta_k \delta(\theta, \theta_k)$$
$$\boldsymbol{\beta} \sim \mathrm{GEM}(\gamma)$$
$$\theta_k \sim H(\lambda) \qquad k = 1, 2, \ldots$$

For each of J groups:

$$G_j(\theta) = \sum_{k=1}^{\infty} \pi_{jk} \delta(\theta, \theta_k) \qquad \boldsymbol{\pi}_j \sim \mathrm{DP}(\alpha, \boldsymbol{\beta})$$

For each of N_j data:

$$z_{ji} \sim \pi_j$$
$$x_{ji} \sim F(\theta_{z_{ji}})$$

Direct Assignment Likelihood

n_{jtk} → Number of tokens in document *j*, assigned to table *t* and topic *k*

n^{w}_{jtk} → Number of tokens of type (word) *w* in document *j*, assigned to table *t* and topic *k*

m_{jk} → Number of tables in document *j* assigned to topic *k*

$$\log p(x, z, m \mid \alpha, \gamma, \lambda) =$$

$$\log \frac{\Gamma(\gamma)}{\Gamma(m_{..} + \gamma)} + \sum_{k=1}^{K} \left\{ \log \gamma + \log \Gamma(m_{.k}) + \log \frac{\Gamma(W\lambda)}{\Gamma(n_{..k} + W\lambda)} + \sum_{w=1}^{W} \log \frac{\Gamma(\lambda + n^{w}_{..k})}{\Gamma(\lambda)} \right\}$$

$$+ \sum_{j=1}^{J} \left\{ \log \frac{\Gamma(\alpha)}{\Gamma(n_{j..} + \alpha)} + m_{j.} \log \alpha + \sum_{k=1}^{K} \log \left[\begin{array}{c} n_{j.k} \\ m_{jk} \end{array} \right] \right\}$$

$\left[\begin{array}{c} n_{j.k} \\ m_{jk} \end{array} \right]$ = Number of permutations of $n_{j.k}$ items with m_{jk} disjoint cycles *(unsigned Stirling numbers of the first kind, Antoniak 1974)*

Sufficient statistics: Global topic assignments and counts of tables assigned to each topic

Permuting Identical Observations

n_{jtk} → Number of tokens in document *j*, assigned to table *t* and topic *k*

n^{w}_{jtk} → Number of tokens of type (word) *w* in document *j*, assigned to table *t* and topic *k*

m_{jk} → Number of tables in document *j* assigned to topic *k*

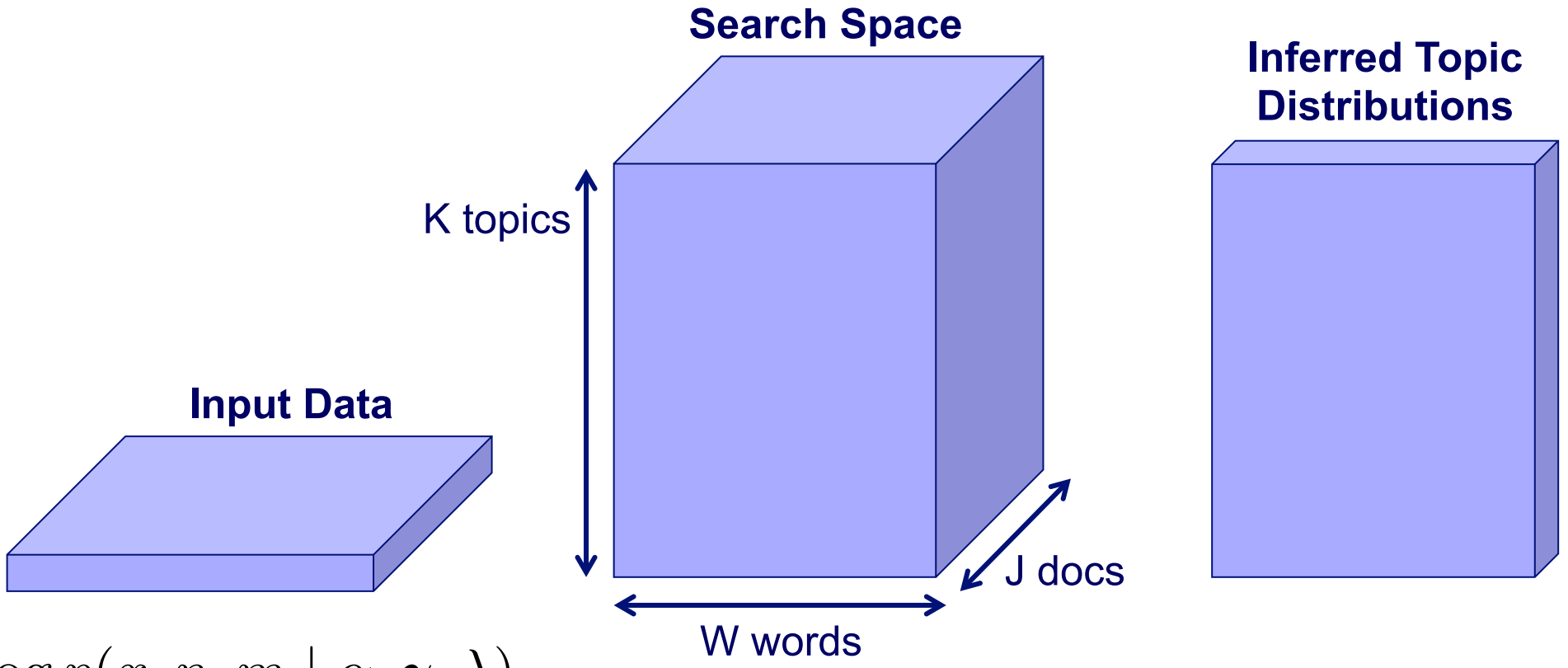
$$\log p(x, n, m \mid \alpha, \gamma, \lambda) =$$

$$\log \frac{\Gamma(\gamma)}{\Gamma(m_{..} + \gamma)} + \sum_{k=1}^{K} \left\{ \log \gamma + \log \Gamma(m_{.k}) + \log \frac{\Gamma(W\lambda)}{\Gamma(n_{..k} + W\lambda)} + \sum_{w=1}^{W} \log \frac{\Gamma(\lambda + n^{w}_{..k})}{\Gamma(\lambda)} \right\}$$

$$+ \sum_{j=1}^{J} \left\{ \log \frac{\Gamma(\alpha)}{\Gamma(n_{j..} + \alpha)} + m_{j.} \log \alpha + \sum_{k=1}^{K} \log \begin{bmatrix} n_{j.k} \\ m_{jk} \end{bmatrix} + \sum_{w=1}^{W} \log \frac{\Gamma(n^{w}_{j..} + 1)}{\prod_{k=1}^{K} \Gamma(n^{w}_{j.k} + 1)} \right\}$$

- When a word is repeated multiple times within a document, those instances (tokens) have identical likelihood statistics
- We sum all possible ways of allocating repeating tokens to produce a given set of counts $n^{w}_{j.k}$

HDP Optimization

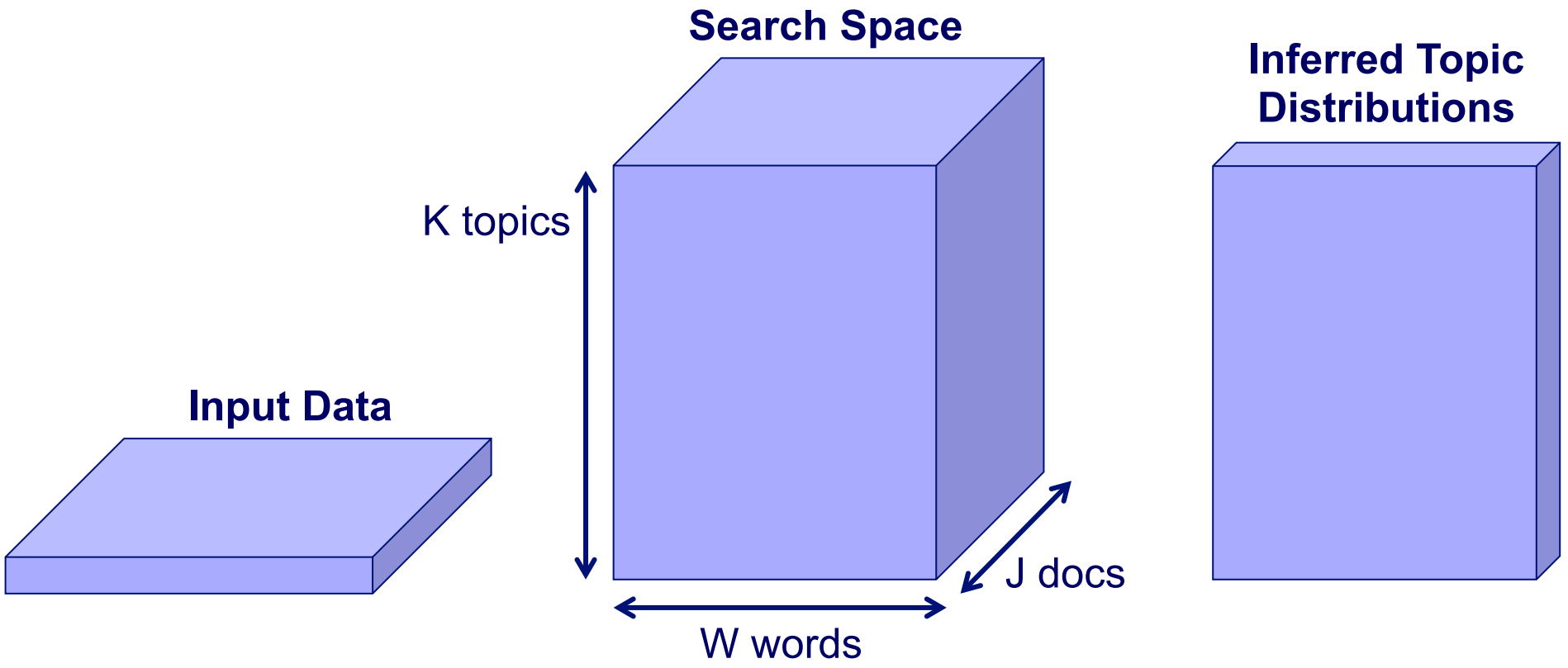


$$\log p(x, n, m \mid \alpha, \gamma, \lambda) =$$

$$\log \frac{\Gamma(\gamma)}{\Gamma(m_{..} + \gamma)} + \sum_{k=1}^{K} \left\{ \log \gamma + \log \Gamma(m_{.k}) + \log \frac{\Gamma(W\lambda)}{\Gamma(n_{..k} + W\lambda)} + \sum_{w=1}^{W} \log \frac{\Gamma(\lambda + n^{w}_{..k})}{\Gamma(\lambda)} \right\}$$

$$+ \sum_{j=1}^{J} \left\{ \log \frac{\Gamma(\alpha)}{\Gamma(n_{j..} + \alpha)} + m_{j.} \log \alpha + \sum_{k=1}^{K} \log \begin{bmatrix} n_{j.k} \\ m_{jk} \end{bmatrix} + \sum_{w=1}^{W} \log \frac{\Gamma(n^{w}_{j..} + 1)}{\prod_{k=1}^{K} \Gamma(n^{w}_{j.k} + 1)} \right\}$$

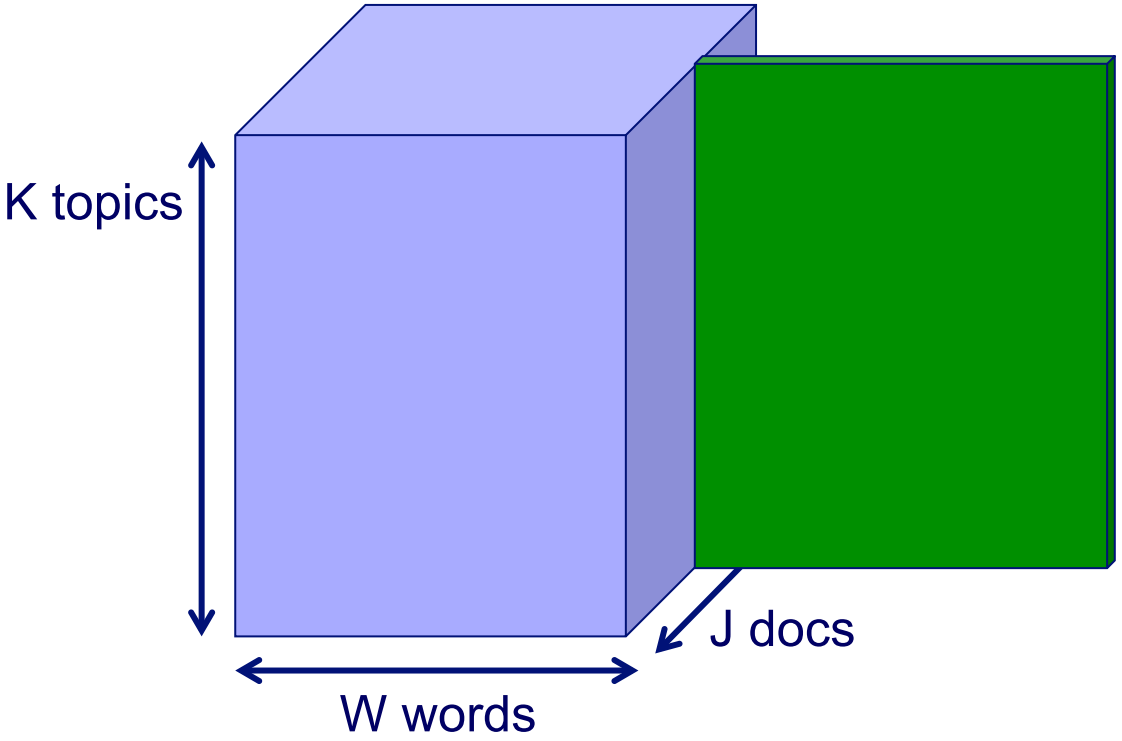
ME Search: Local Moves

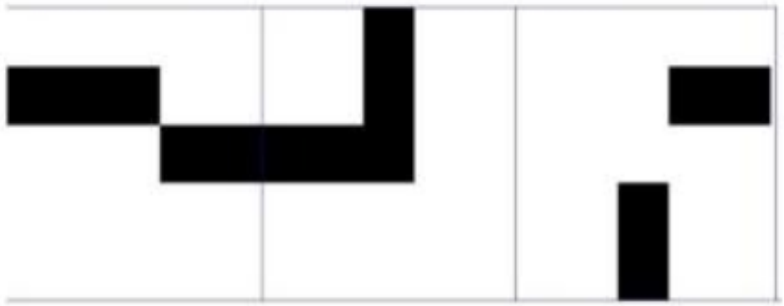


In some random order:

- Assign one word token to the optimal (possibly new) table
- Assign one table to the optimal (possibly new) topic
- Merge two tables, assign to the optimal (possibly new) topic

ME Search: Reconfigure Document

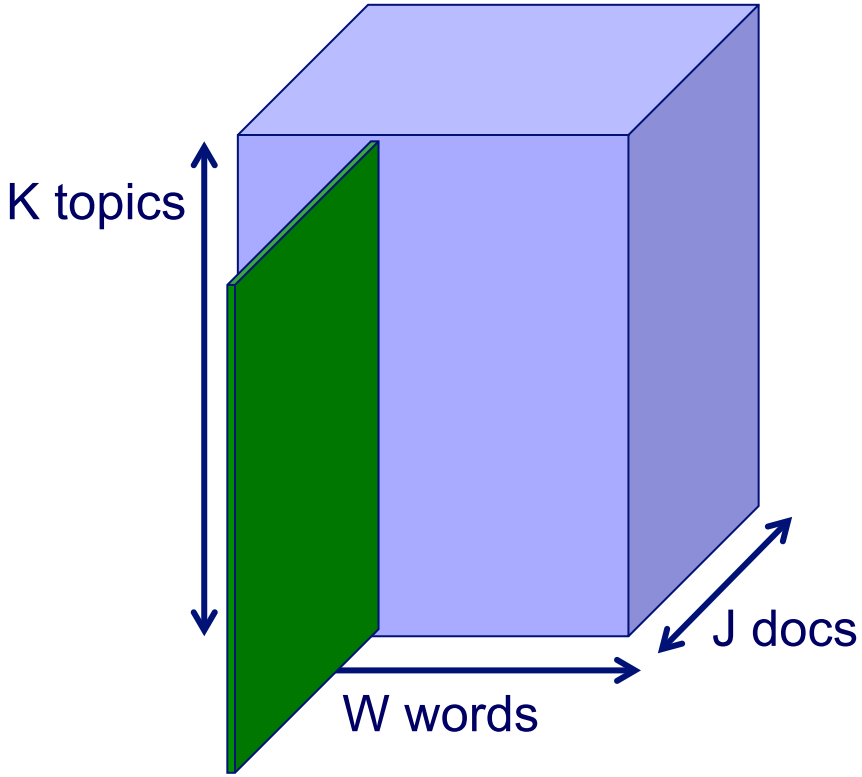


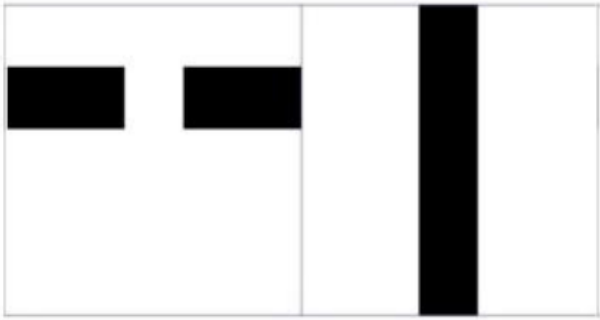


For some document, fixing configurations of all others:

- Remove all existing assignments, and sequentially assign tokens to topics via a conditional CRP sampler
- Refine configuration with local search (only this document)
- Reject if new configuration has lower likelihood

ME Search: Reconfigure Word

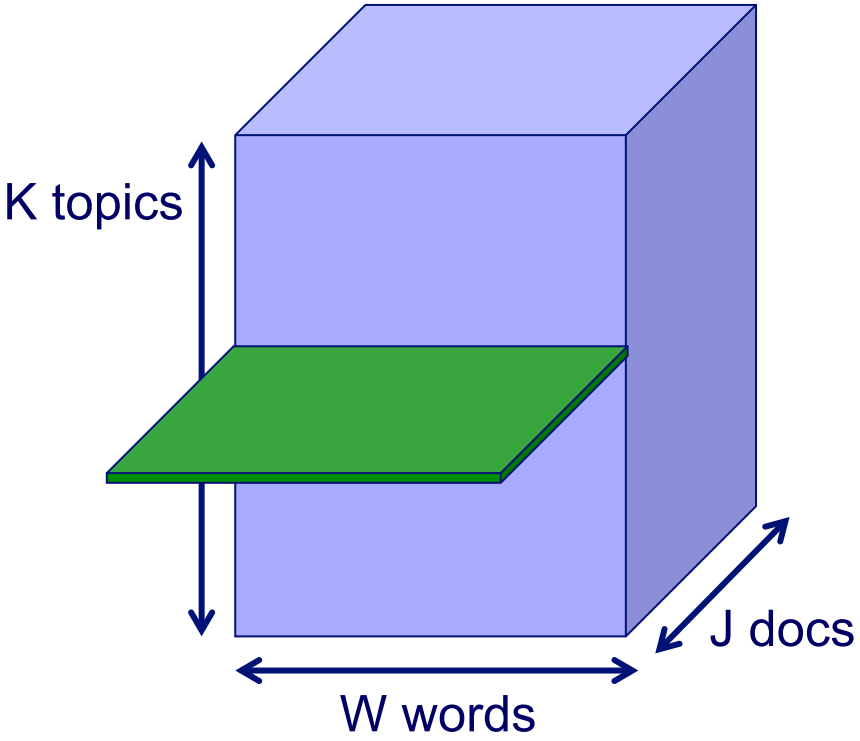




For some vocabulary word, fixing configurations of all others:

- Remove all existing assignments topic by topic, sequentially assign tokens to topics via a conditional CRP sampler
- Refine configuration with local search (only this word type)
- Reject if new configuration has lower likelihood

ME Search: Reconfigure Topic



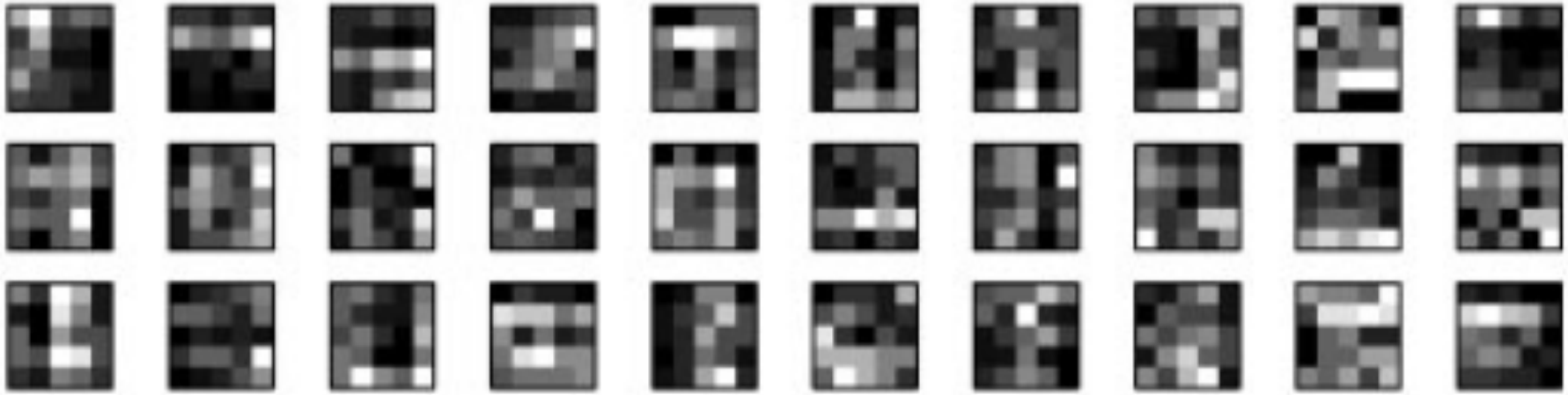
For some topic, fixing configurations of all others:

- Merge with another topic
- Refine or reject topic: Apply reconfigure-document and reconfigure-word moves to this topic's documents/words
- Reject if any new configuration has lower likelihood

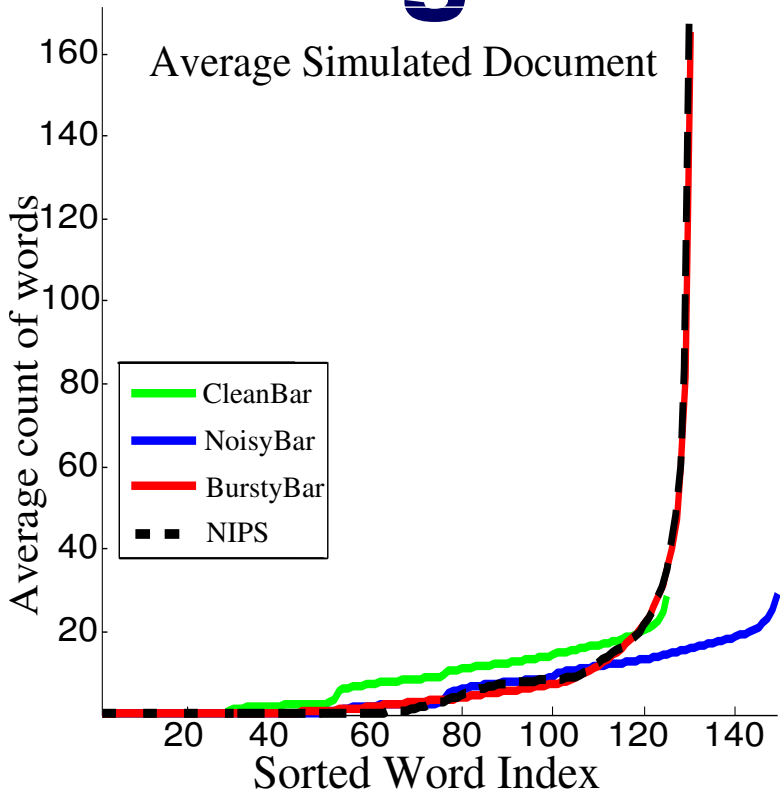
The Toy Bars Dataset

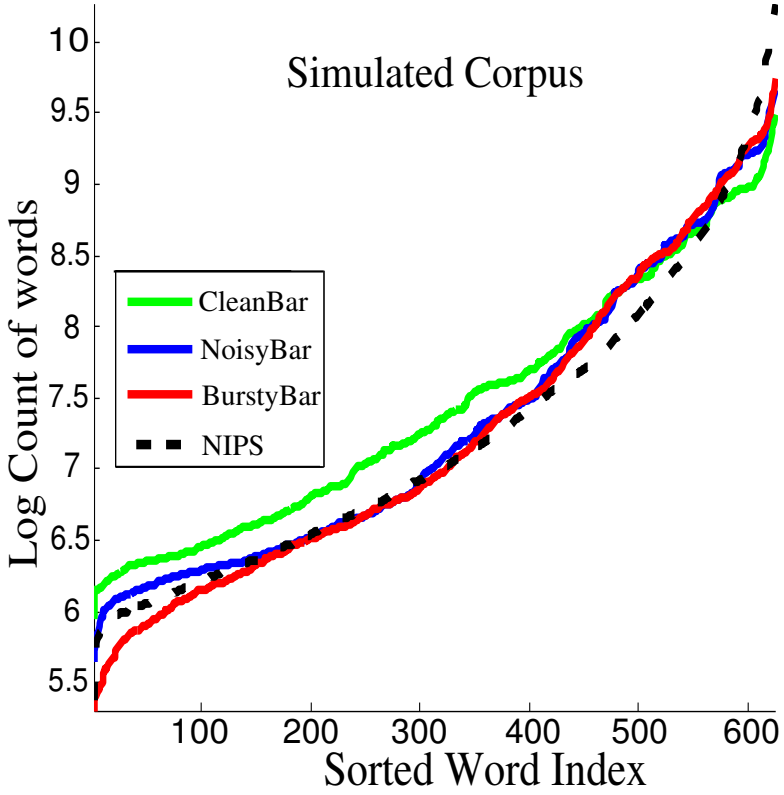
- Latent Dirichlet Allocation *(LDA, Blei et al. 2003)* is a parametric topic model (a finite Dirichlet approximation to the HDP)
- Griffiths & Steyvers *(2004)* introduced a collapsed Gibbs sampler, and demonstrated it on a toy "bars" dataset:

10 topic distributions on 25 vocabulary words, and example documents



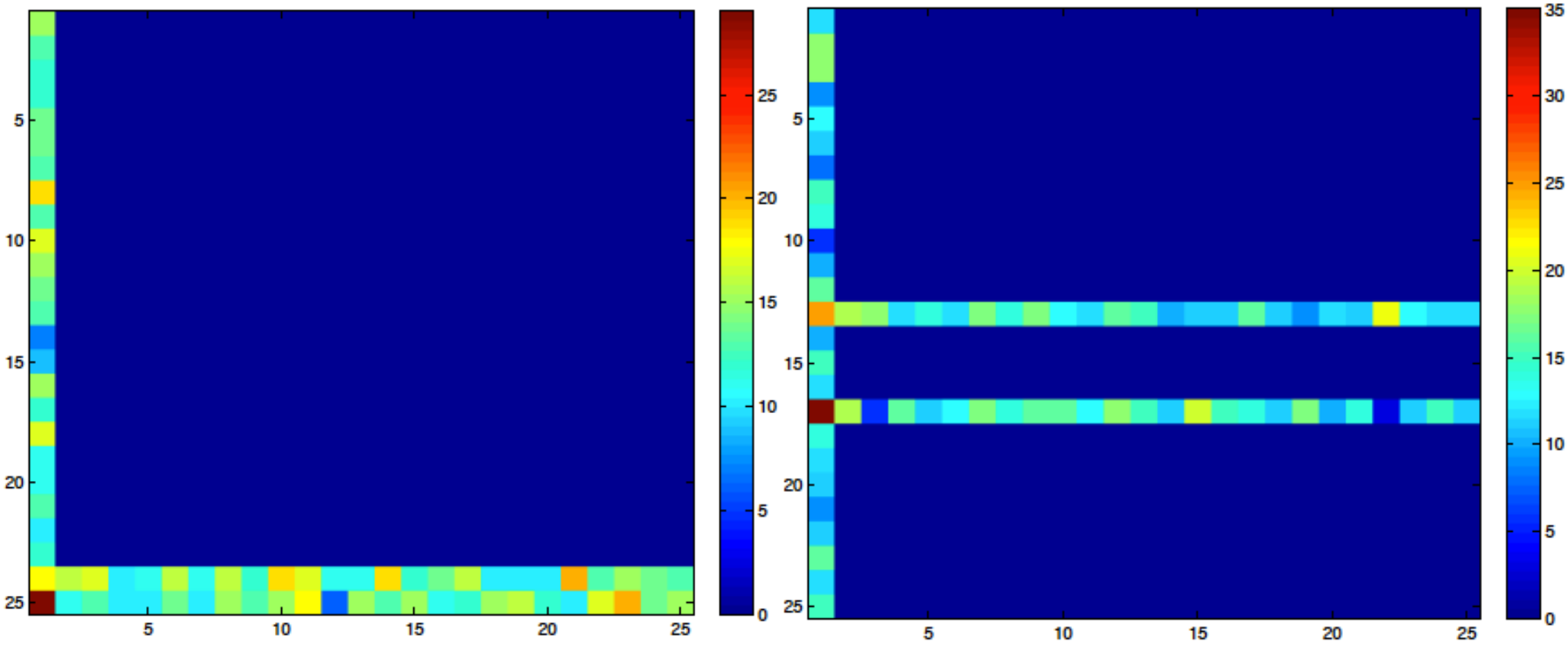
Making More Realistic Bars



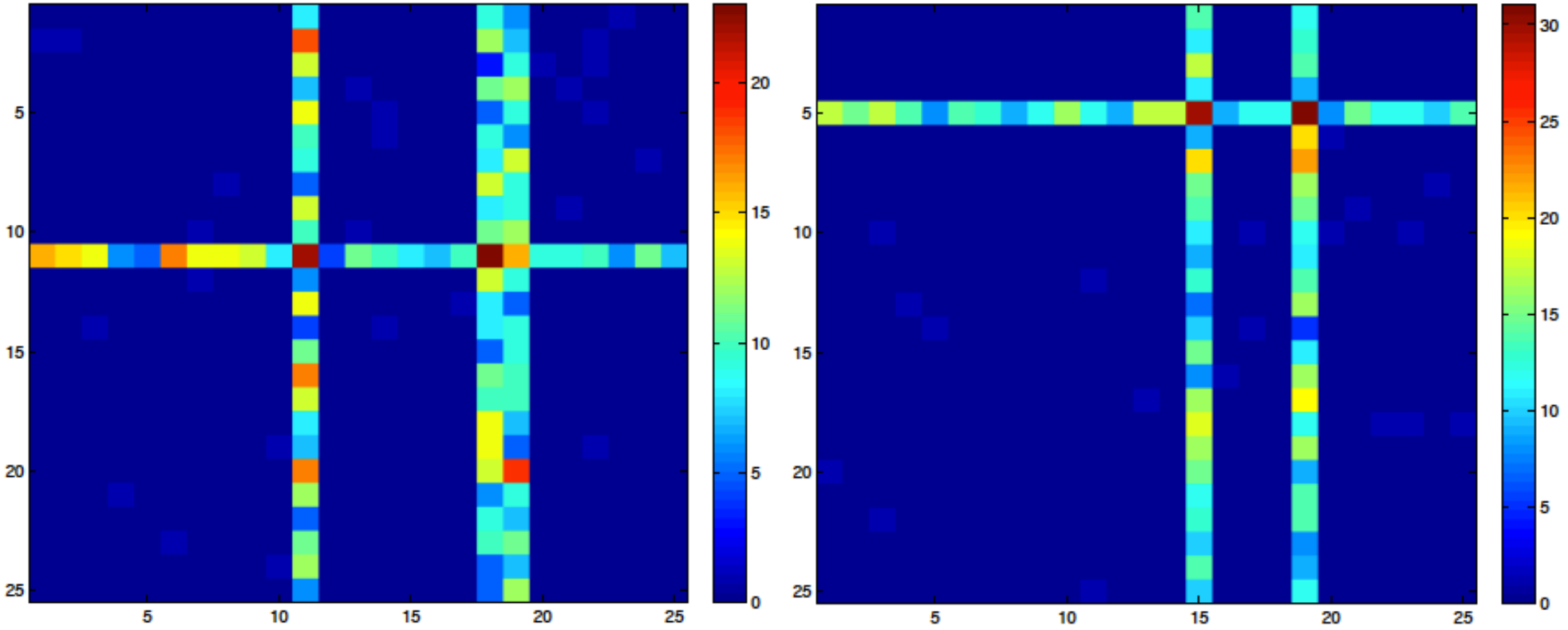


- Frequency: Assign geometrically decreasing rates of occurrence to topics, rather than making equally likely
- Noise: Generate portion of document words uniformly at random, rather than from the primary topics
- Burstiness: Increase the frequencies of a few randomly chosen words from the most likely topics

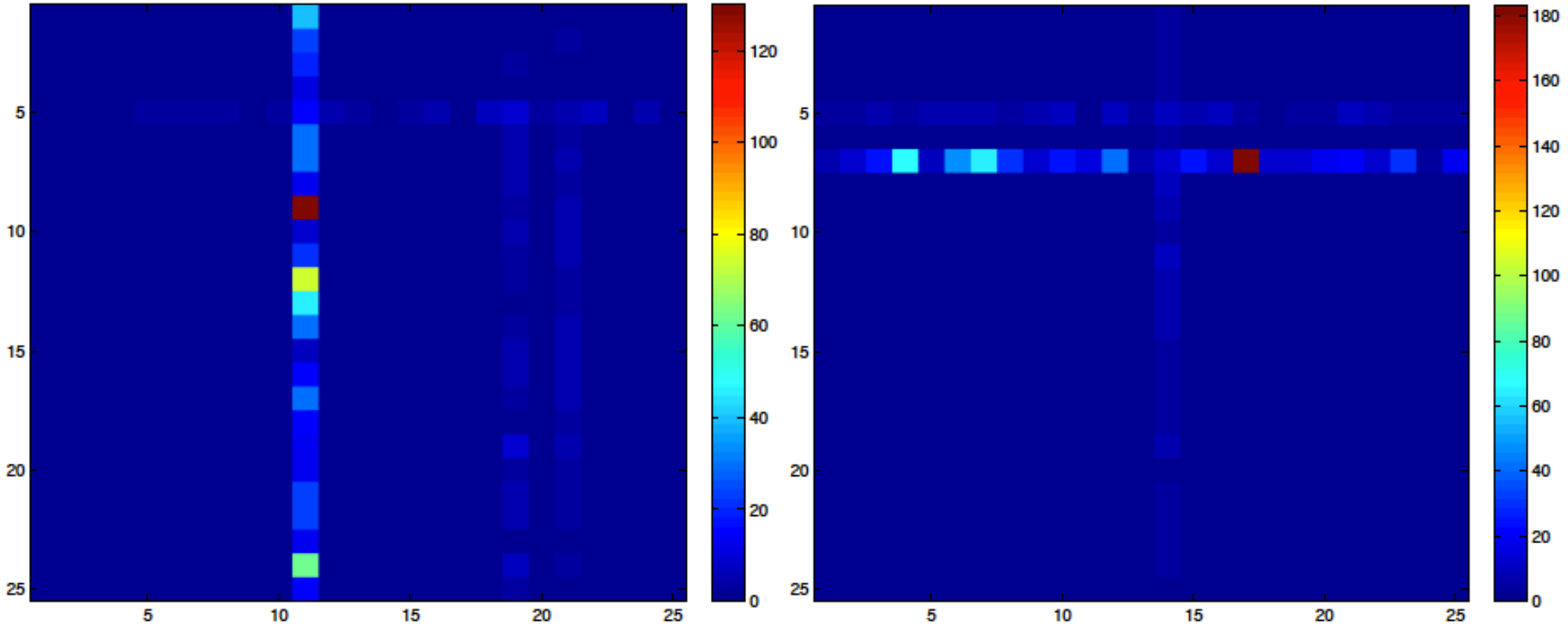
Bars with Unequal Frequencies



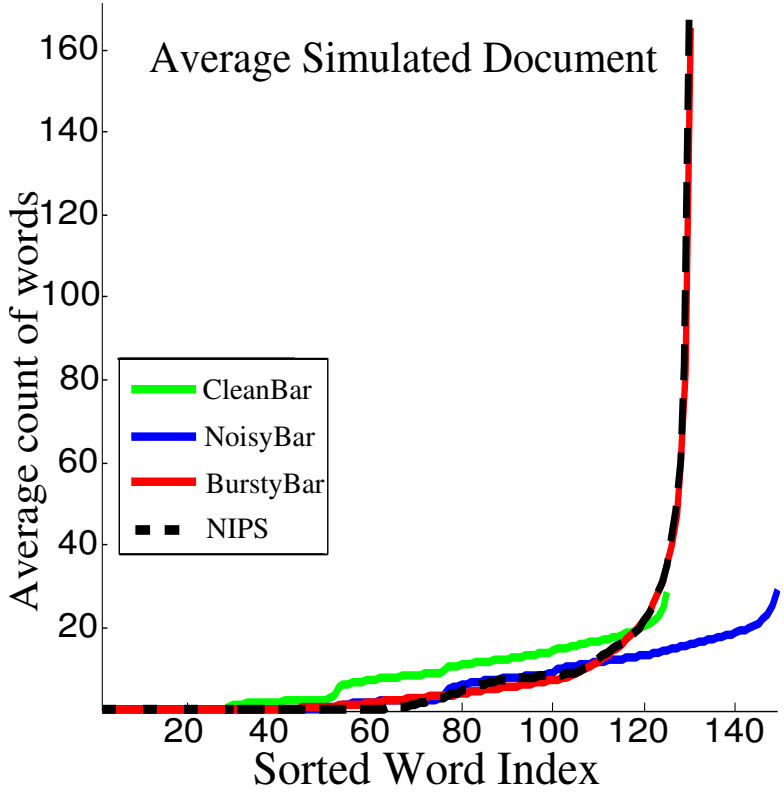
Unequal Bars with Noise

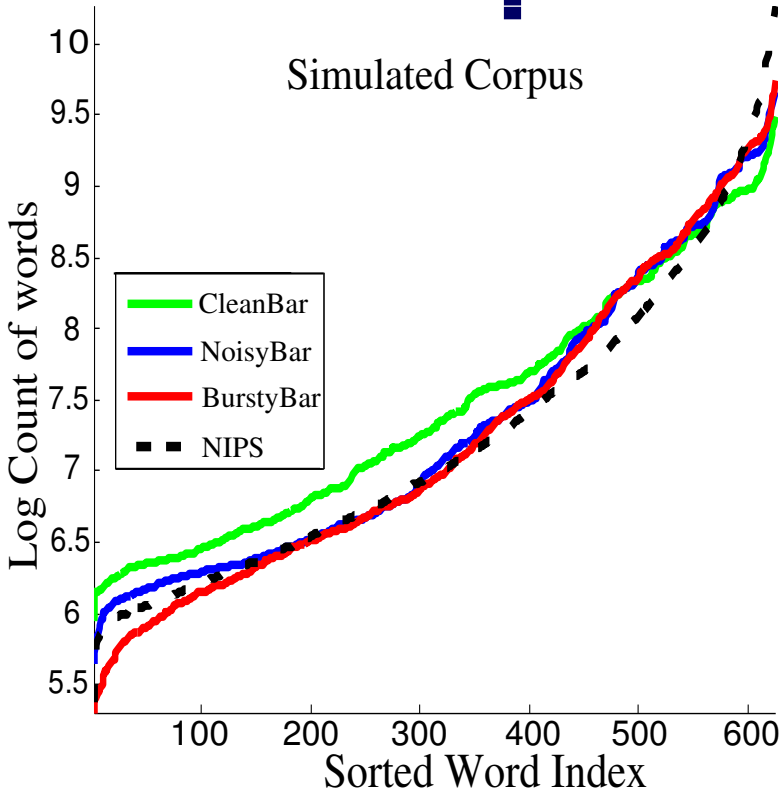


Bursty, Noisy, Unequal Bars



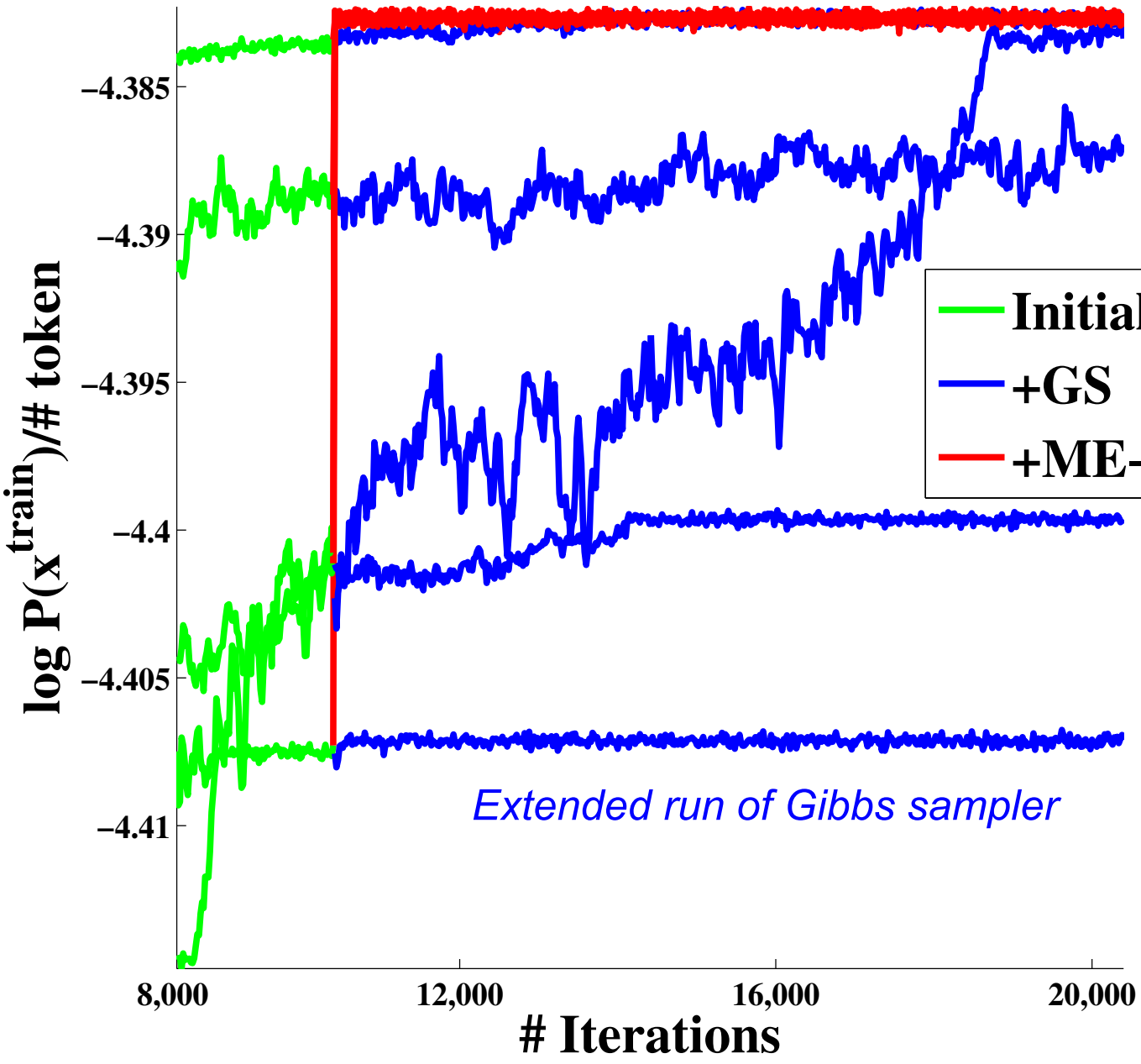
What Should the HDP Capture?





- Frequency: Well, via explicit parameters in base measure
- Noise: Weakly, by using many extraneous topics with small probability mass
- Burstiness: Completely unmodeled; topics are fixed multinomials with no document-specific variation

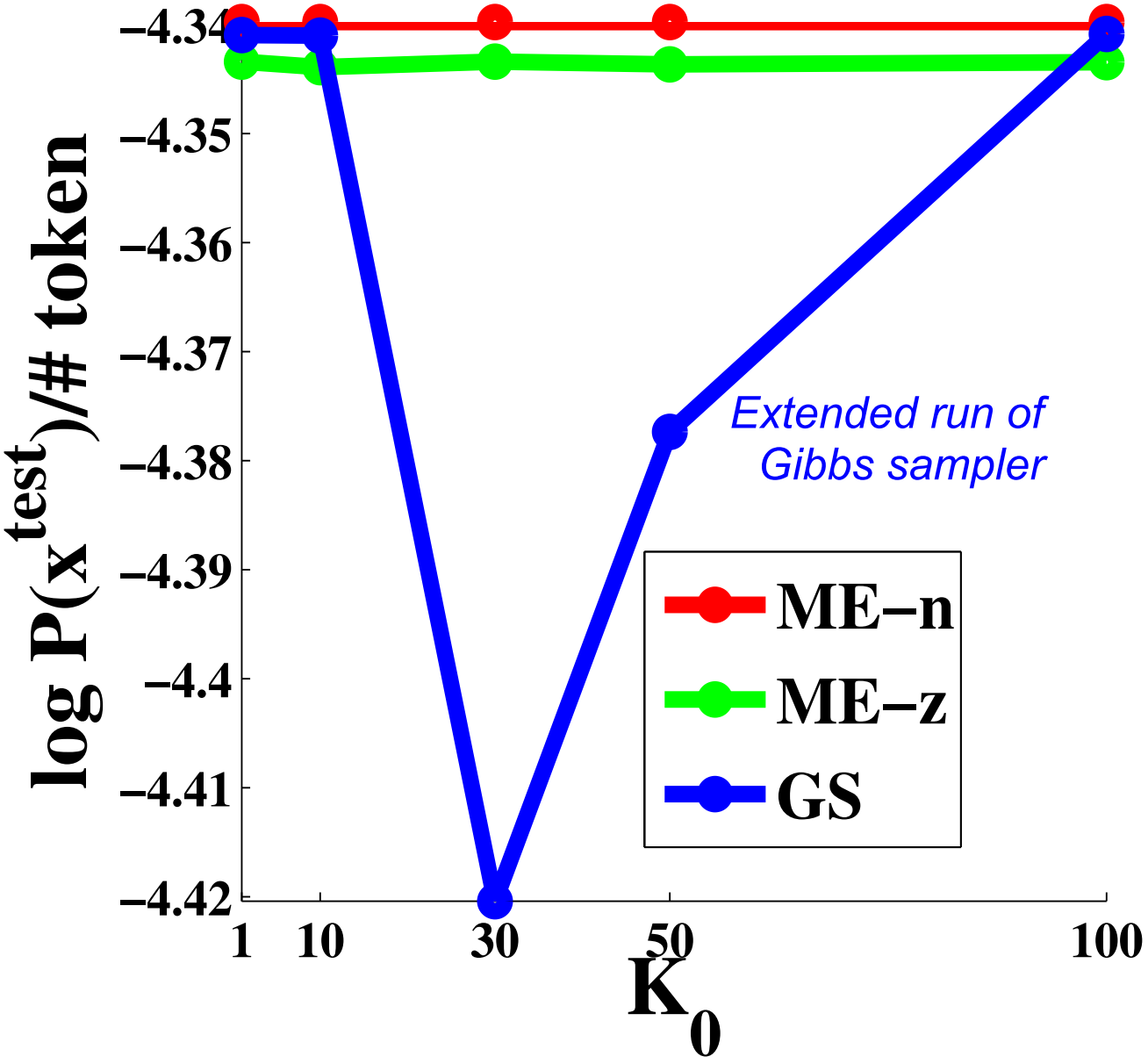
Unequal Bars: Mixing Rates



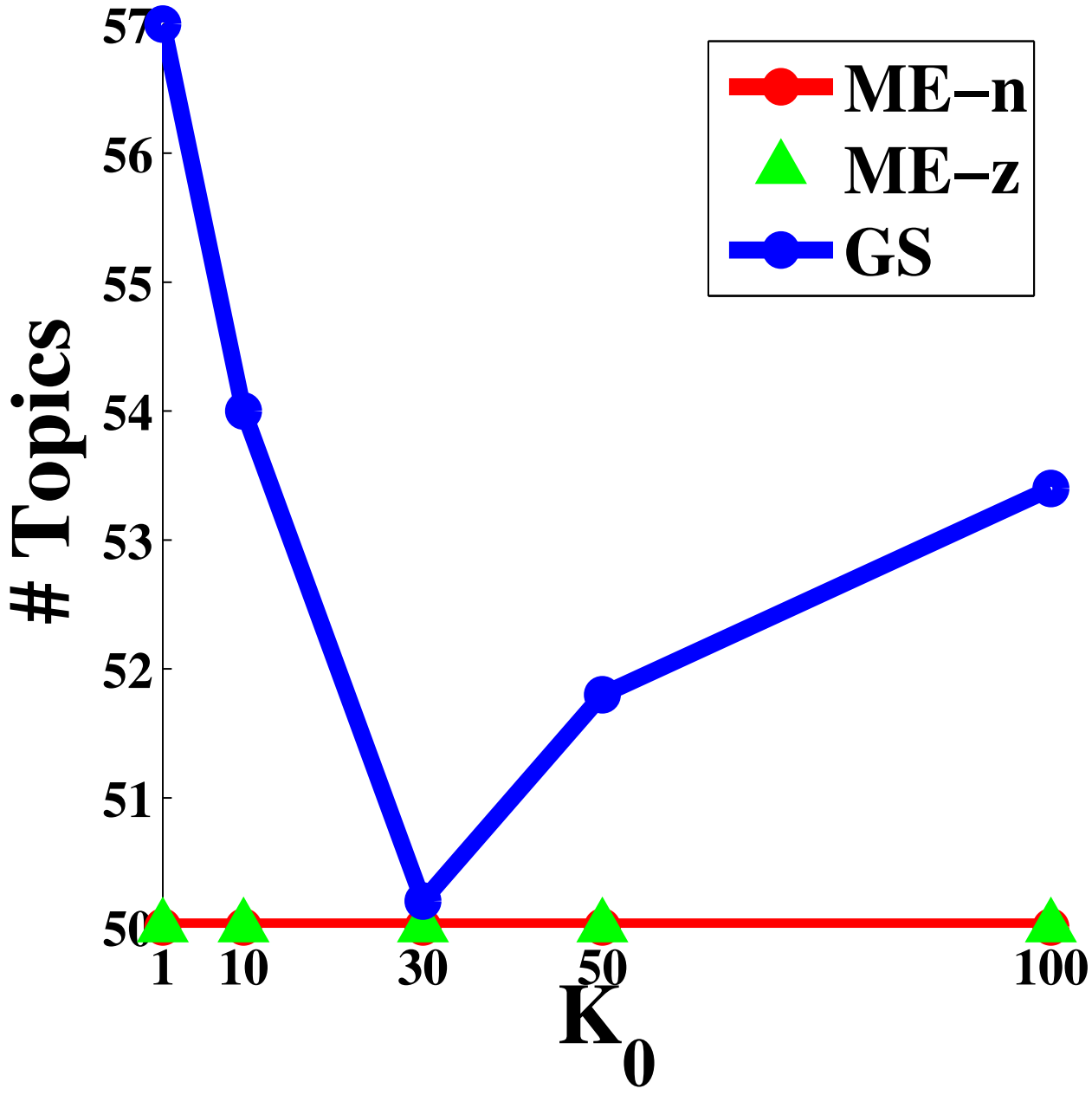
Dataset Scale:
1566 documents
625 words
50 true topics

Unequal Bars: Test Likelihoods

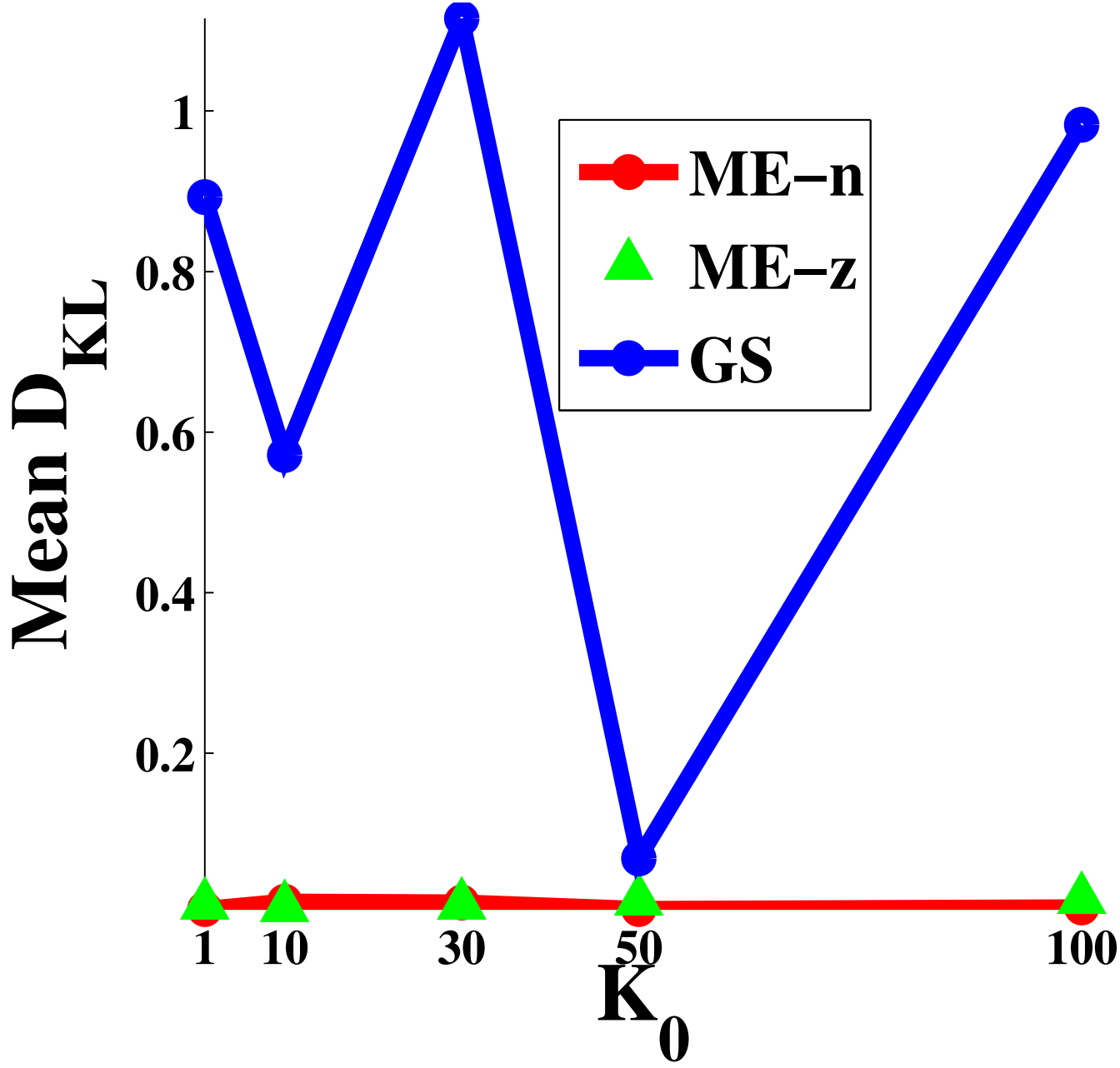
Initialize by short run of MCMC, run ME search, output solution



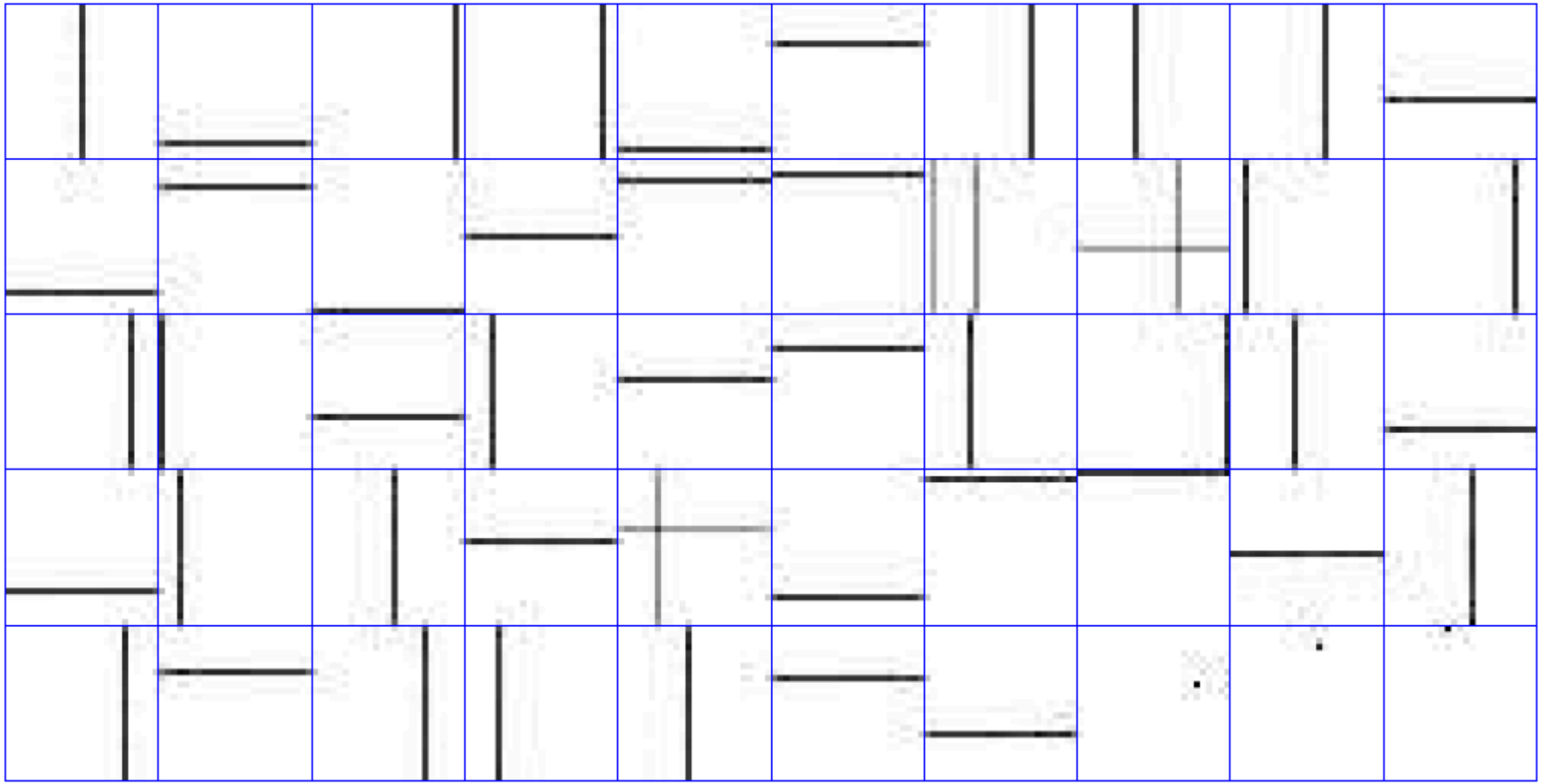
Unequal Bars: Number of Topics



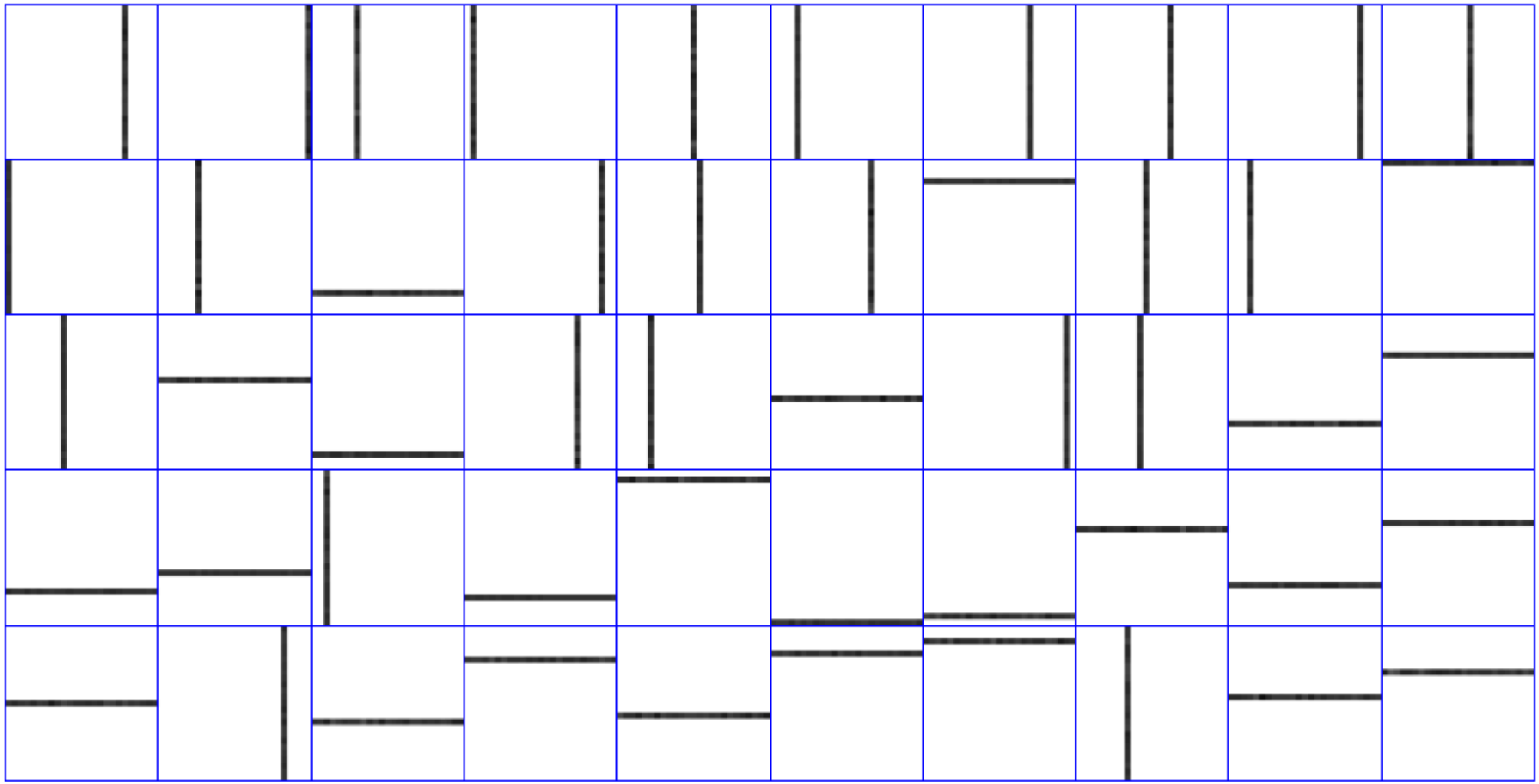
Unequal Bars: KL Divergence



Unequal Bars: Gibbs Topics

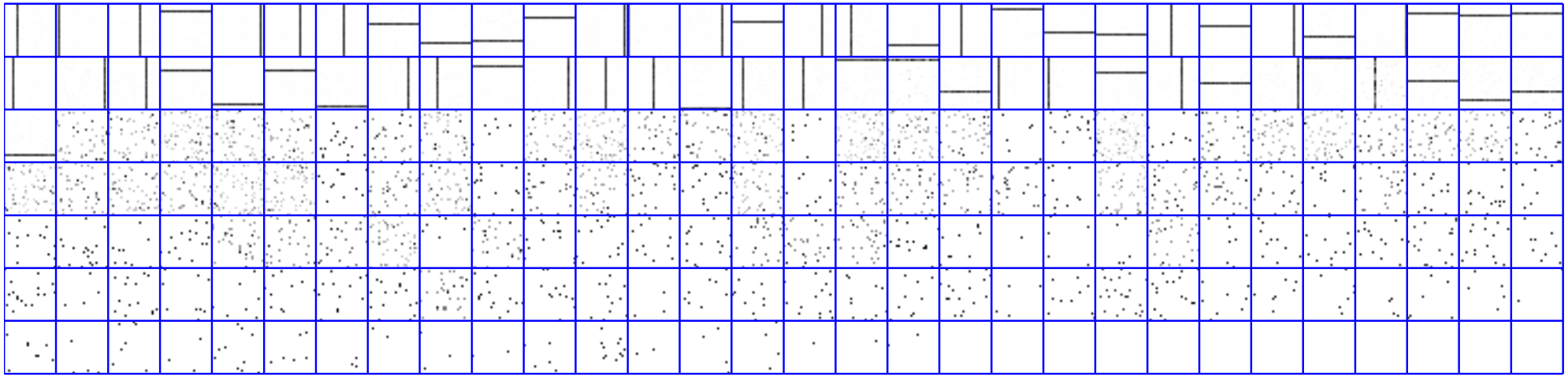


Unequal Bars: ME Topics

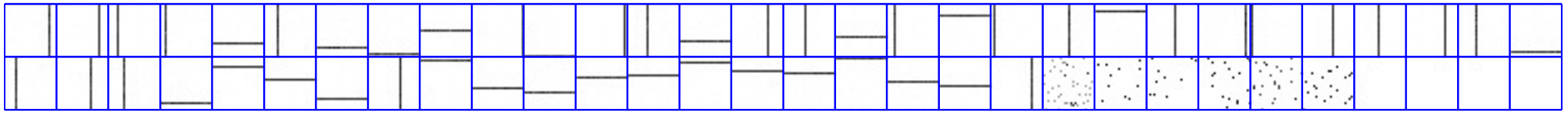


Noisy Bars: Topics

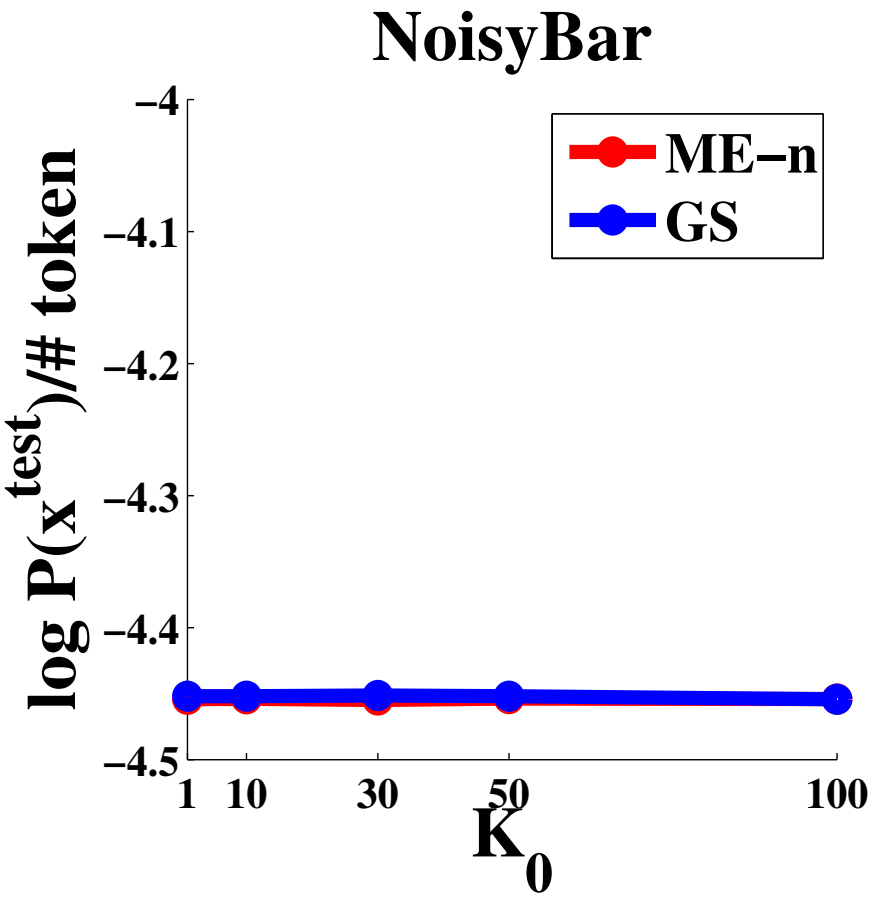
Gibbs Sampler

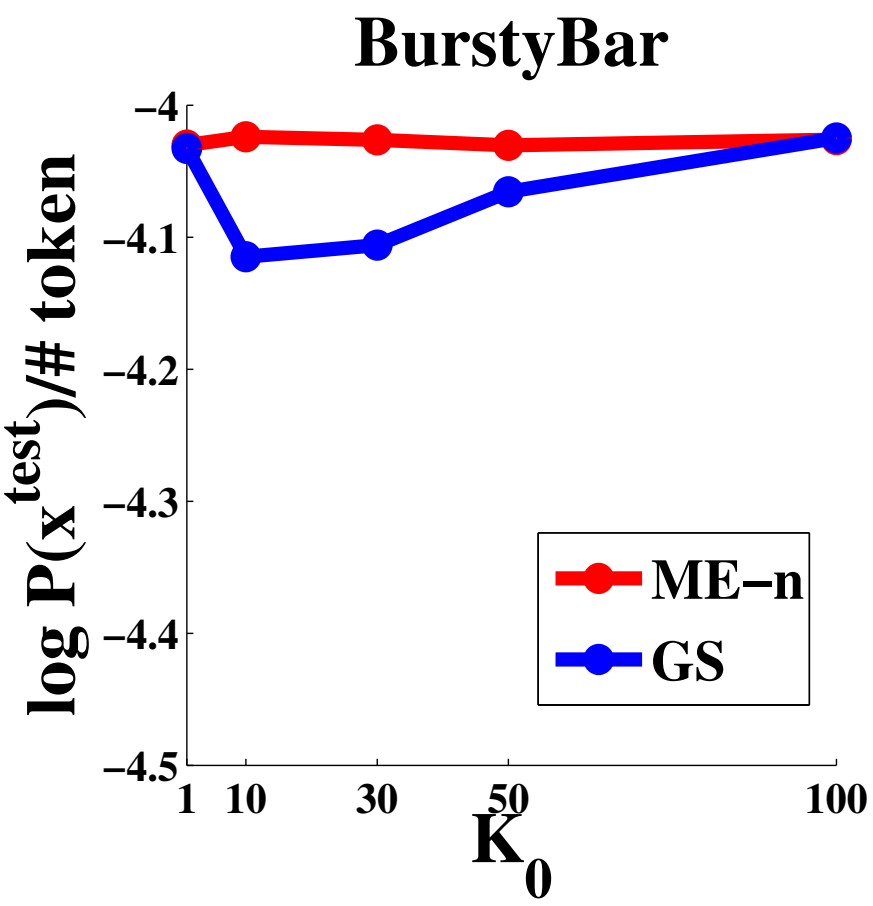


ME Search

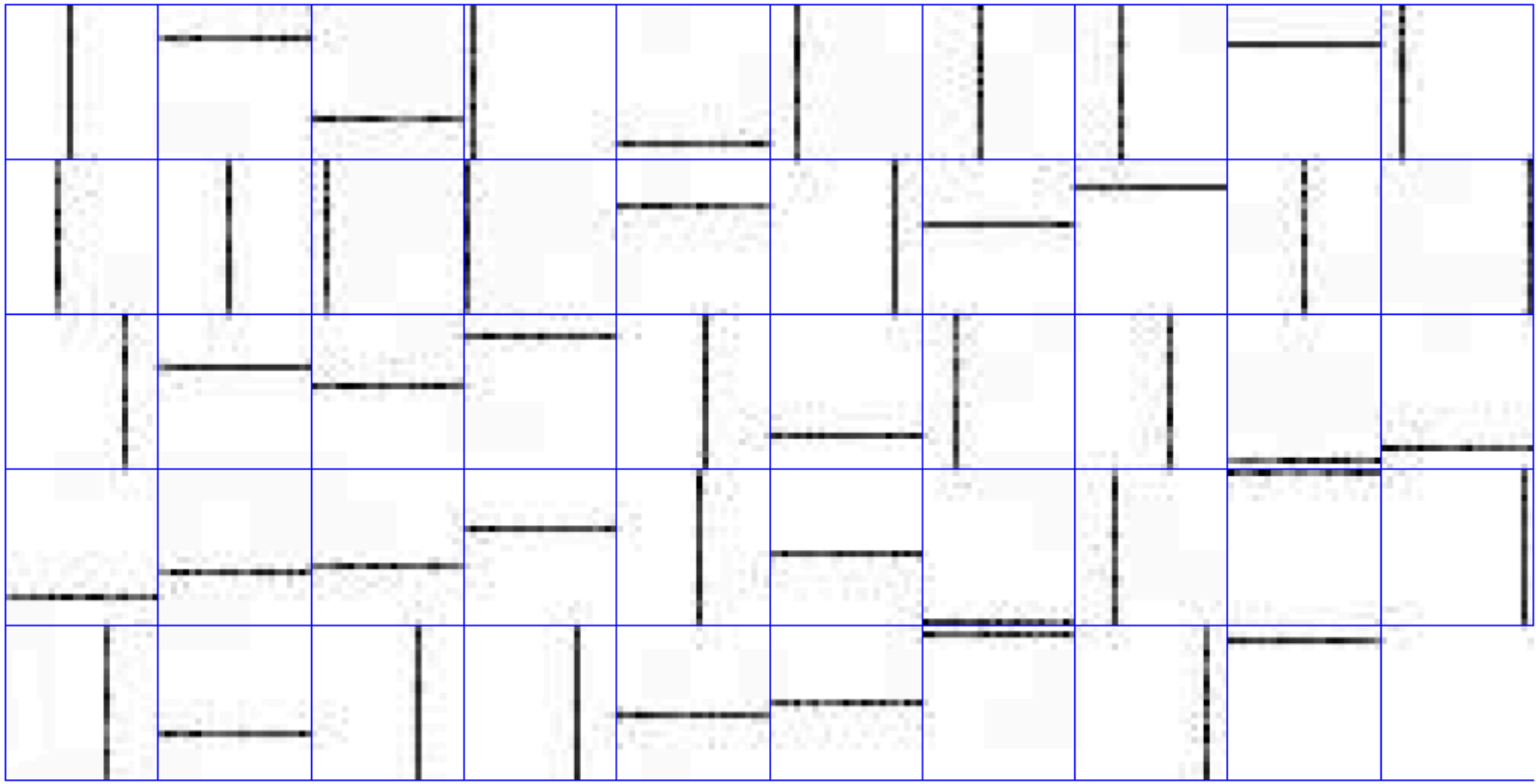


Noisy & Bursty Test Likelihoods



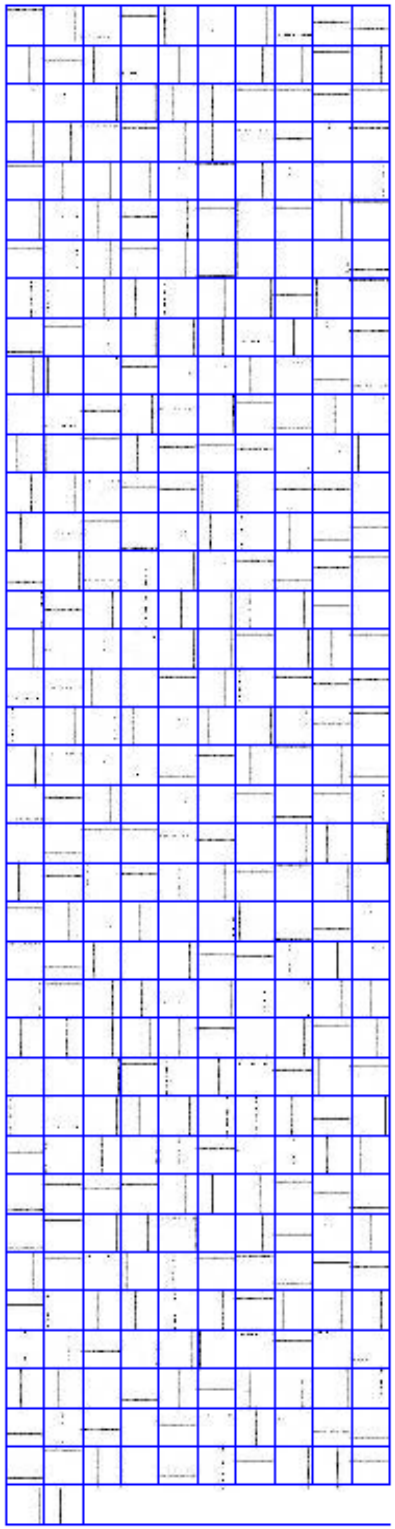


Bursty Bars: Topics



$\lambda = 4$

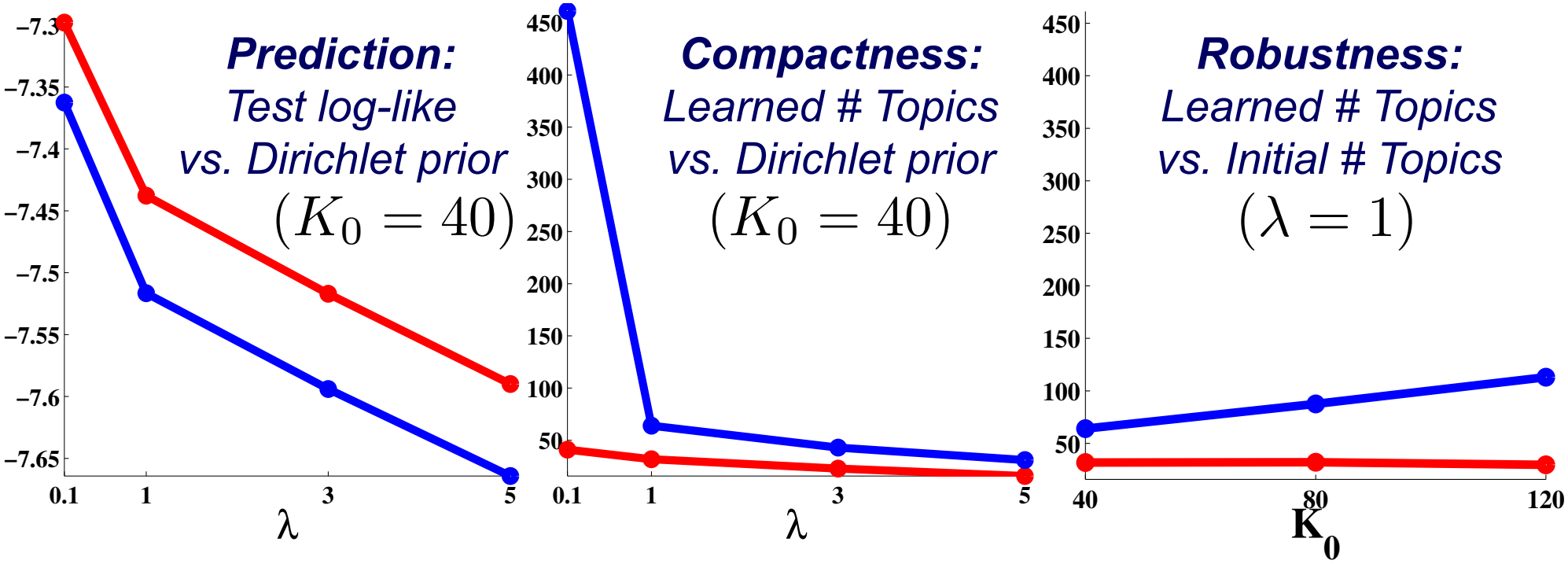
$\lambda = 0.1$



- Predictive likelihood and topic coherence are negatively correlated
- There is some work on modeling burstiness with parametric topic models *(Doyle & Elkan, 2009)*

NIPS Dataset

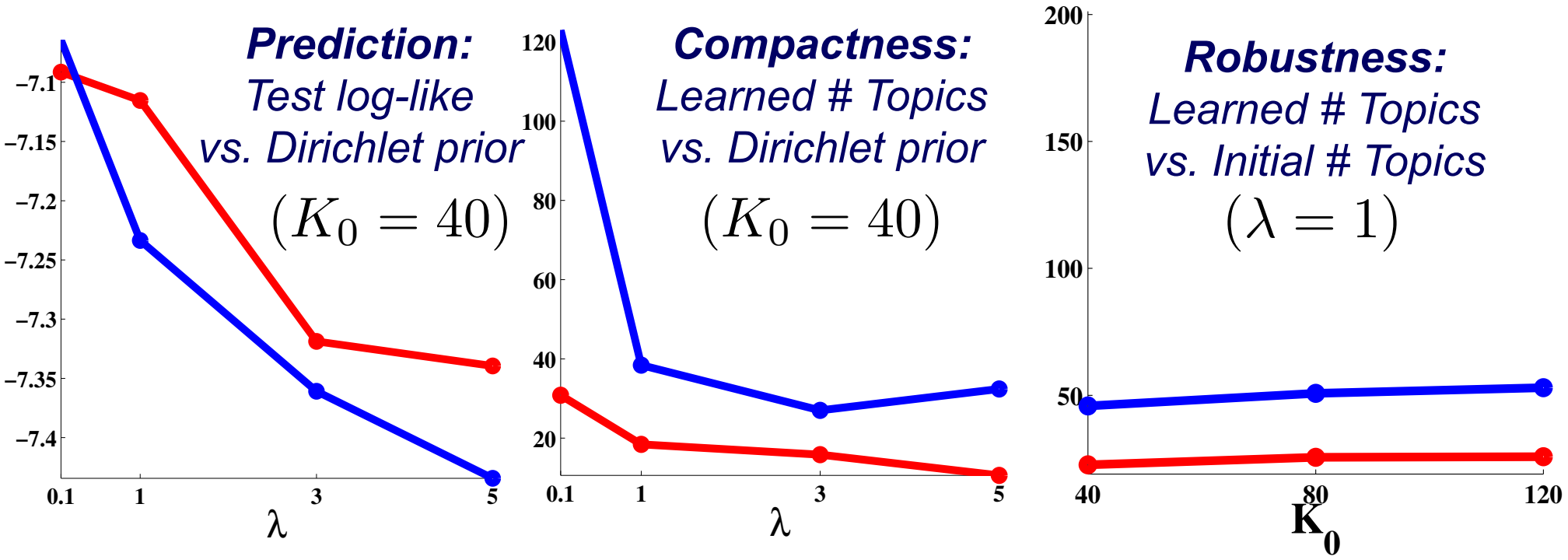
1,740 documents, ~1.7 million word tokens



- Red: ME-n Search (5 iterations initialized by brief Gibbs sampling)
- Blue: Gibbs Sampling (20,000 iterations)
- All results averaged over runs from 5 random initializations
- Predictive likelihoods computed via Chib-style MCMC estimator

20 Newsgroups Dataset

4,709 documents, ~435,000 word tokens

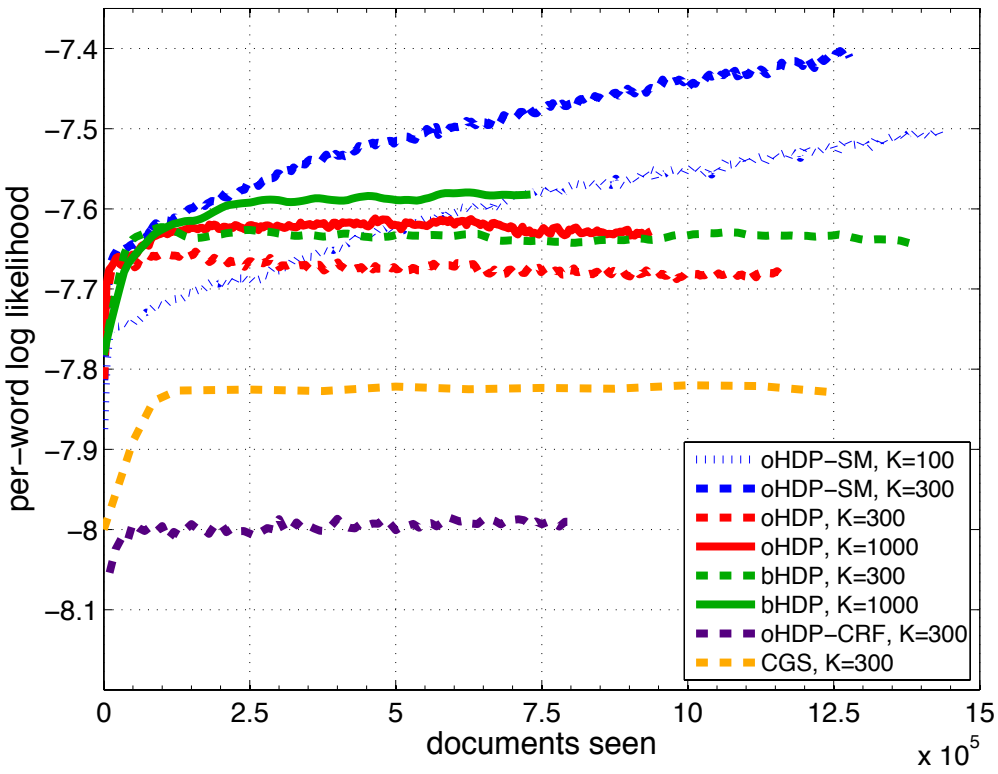


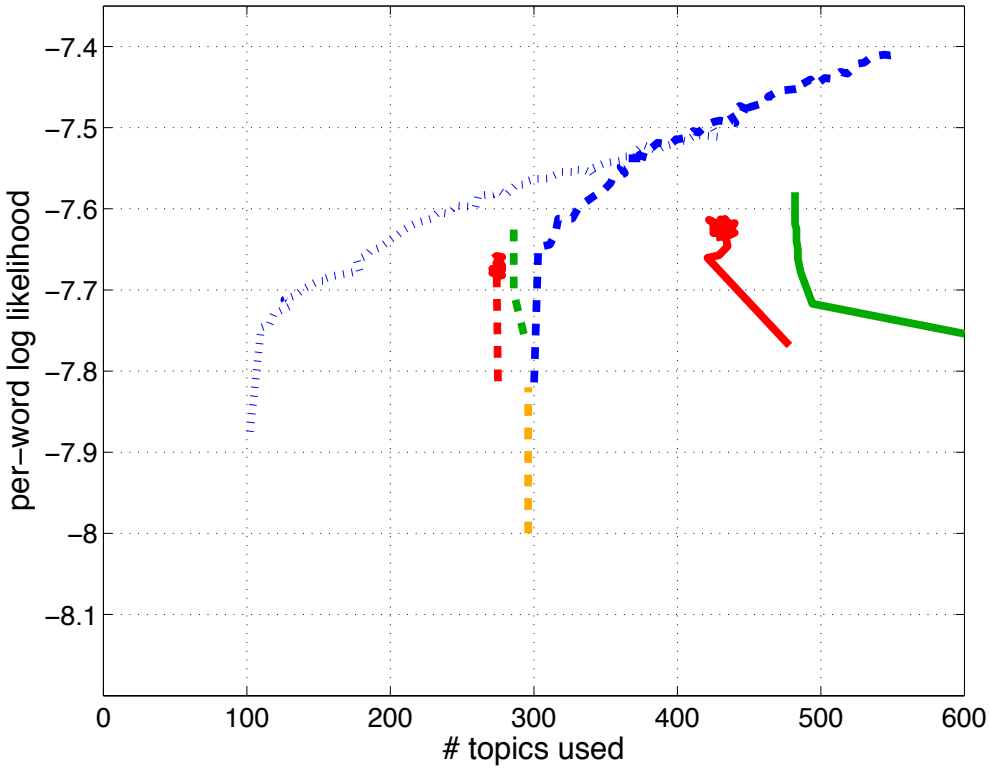
- Red: ME-n Search (5 iterations initialized by brief Gibbs sampling)
- Blue: Gibbs Sampling (20,000 iterations)
- All results averaged over runs from 5 random initializations
- Predictive likelihoods computed via Chib-style MCMC estimator

Online Variational Learning

[Bryant NIPS 2012]

NIPS (1740 documents)



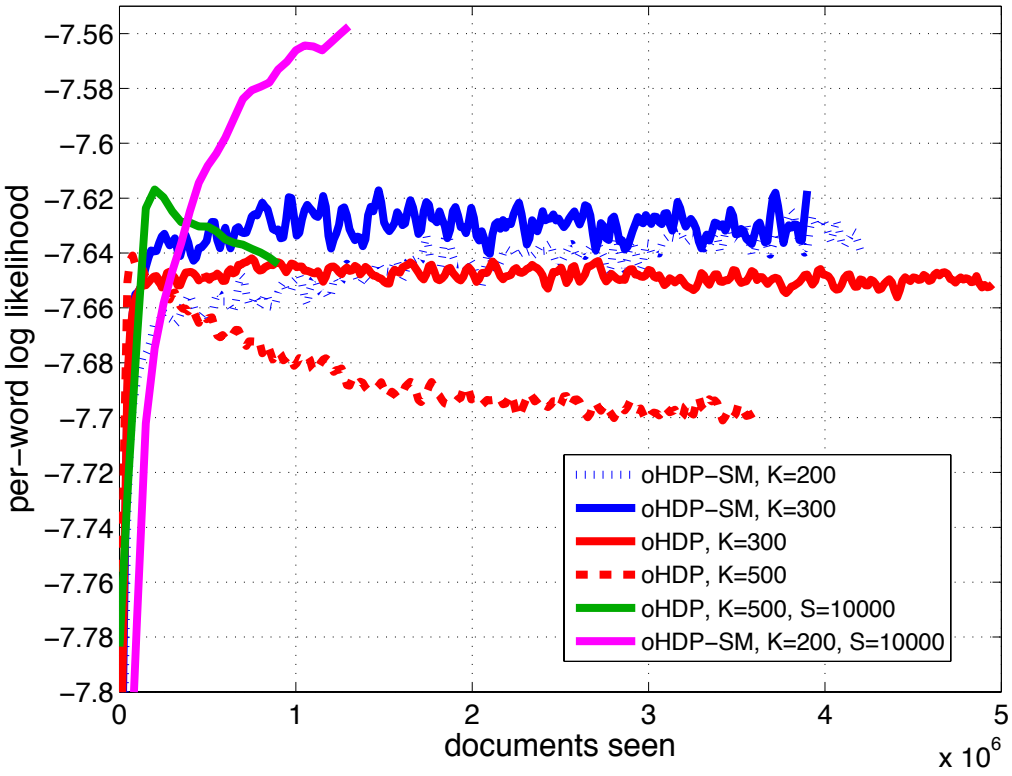


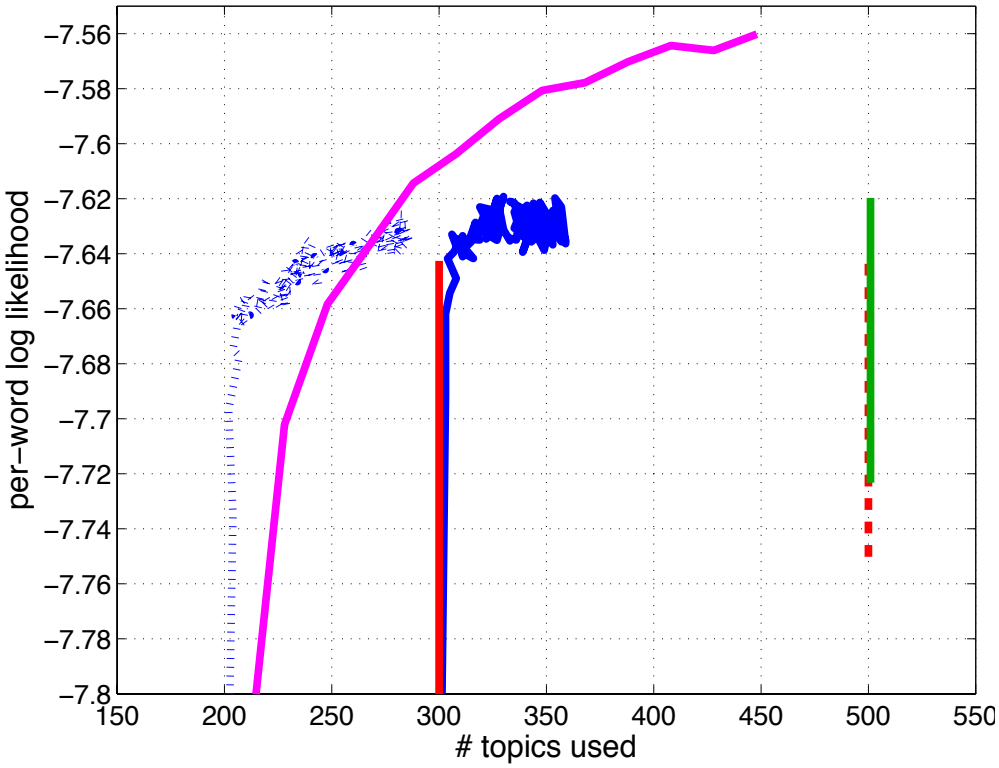
- Online variational inference with split-merge proposals
- Batch variational inference (direct assignment, local optimization)
- Online variational inference (direct assignment, local optimization)
- Collapsed Gibbs sampling (direct assignment)
- Online variational inference (Chinese restaurant franchise, Wang et al.)

Online Variational Learning

[Bryant NIPS 2012]

New York Times (1.8 million documents)





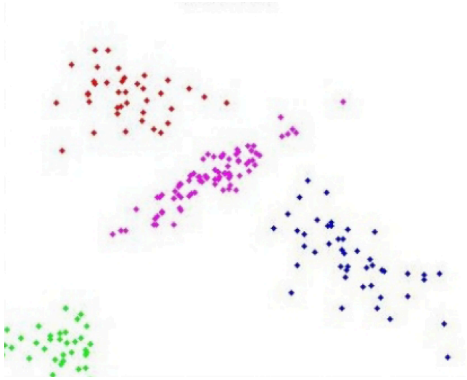
- Online variational inference with split-merge proposals (large mini-batches)
- Online variational inference with split-merge proposals
- Online variational inference (direct assignment, local optimization)

Split Topic Evolution

[Bryant NIPS 2012]

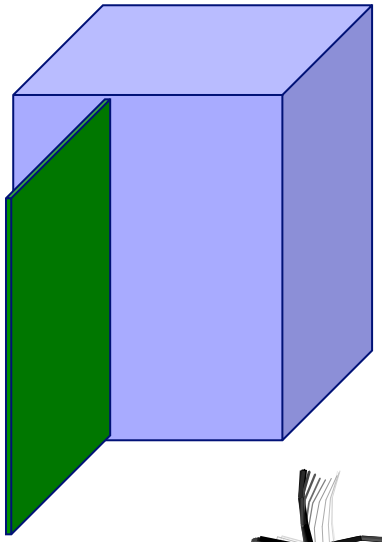
Original topic	40,000	80,000	120,000	160,000	200,000	240,000
patterns	patterns	patterns	patterns	patterns	patterns	patterns
pattern	pattern	pattern	pattern	pattern	pattern	pattern
cortex	cortex	cortex	cortex	cortex	cortex	cortex
neurons	neurons	neurons	neurons	neurons	neurons	responses
neuronal	neuronal	responses	responses	responses	types	
responses	responses	neuronal	type	type	type	
single	single	single	behavioral	behavioral	behavioral	
inputs	temporal	type	types	types	form	
temporal	inputs	number	neuronal	form	neurons	
activation	type	temporal	single	neuronal	areas	
patterns	neuronal	neuronal	neuronal	neuronal	neuronal	
neuronal	patterns	neurons	dendritic	dendritic	dendritic	
pattern	pattern	activation	peak	fire	postsynaptic	
neurons	neurons	cortex	activation	peak	fire	
cortex	cortex	dendrite	cortex	activation	cortex	
inputs	activation	preferred	pyramidal	msec	activation	
activation	dendrite	patterns	msec	pyramidal	peak	
type	inputs	peak	fire	cortex	msec	
preferred	peak	pyramidal	dendrites	postsynaptic	pyramidal	
peak	preferred	inputs	inputs	inputs	inputs	

Outline



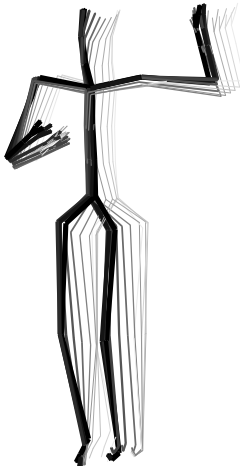
Bayesian Nonparametrics

- Dirichlet process (DP) mixture models
- Variational methods and the ME algorithm



Reliable Nonparametric Learning

- Hierarchical DP topic models
- ME search in a collapsed representation
- Non-local online variational inference



Nonparametric Temporal Models

- Beta Process Hidden Markov Models (BP-HMM)
- Effective split-merge MCMC methods

BP Hidden Markov Model

[Fox NIPS 2009]

z
Behavior Sequence

π
Transition Matrix

Video 1

Video 2

Video 3

Video 4

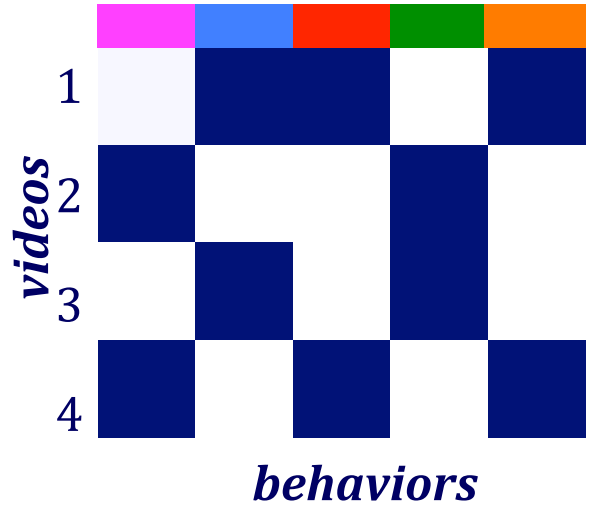
.8	.2	
.1	.9	

.8	.15	.05
.05	.9	.05
.1	.1	.8

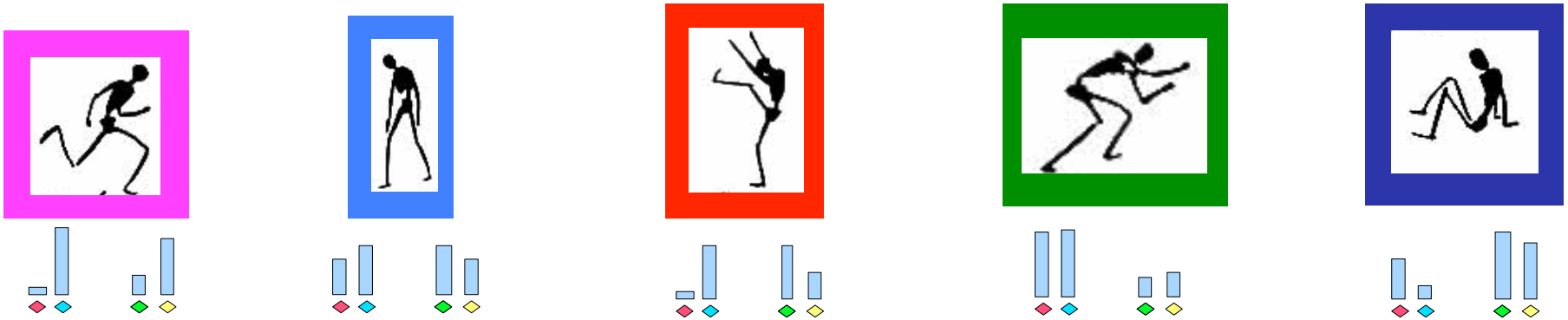
.8	.15	.05
.05	.8	.15
.05	.05	.9

.9	.1	
.1	.9	

F
Behavior Matrix



Behavior Library - *HMM emission parameters* θ



Beta Process HMM

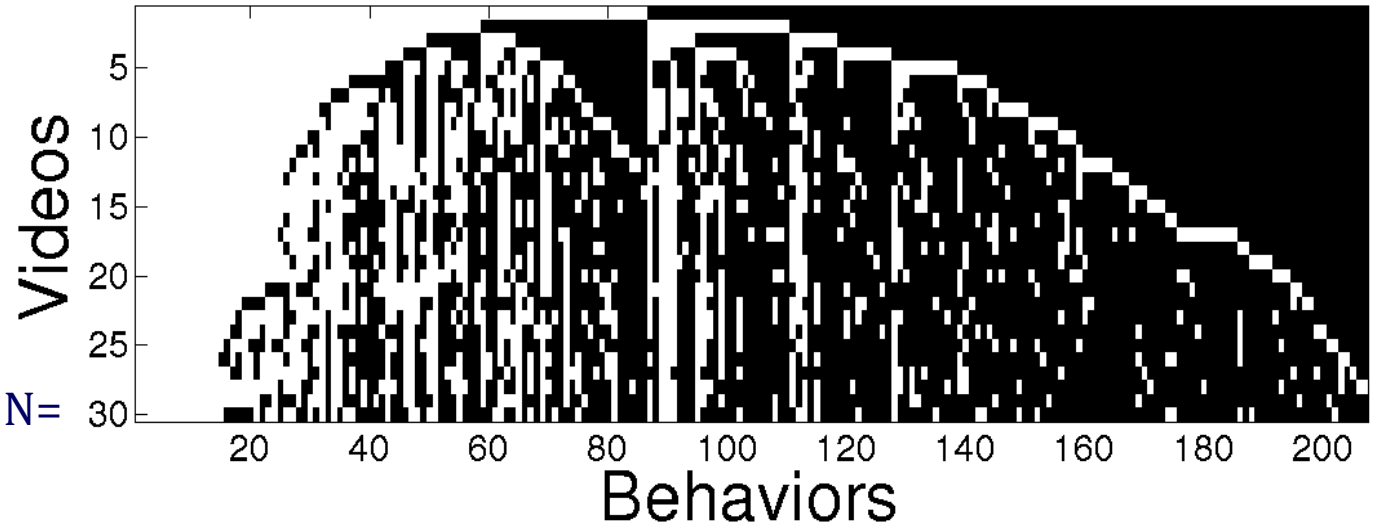
[Ghahramani 2006]
[Thibaux 2007]

Each video ***i*** has **sparse binary vector** indicating available behaviors

$$f_i = [\,0 \;\; 1 \;\; 0 \;\; 1 \;\; 0 \;\; 0 \;\; 0 \;\; \ldots\,]$$

Beta Process (BP): prior on sparse binary matrix

Alternative representation: Indian Buffet Process



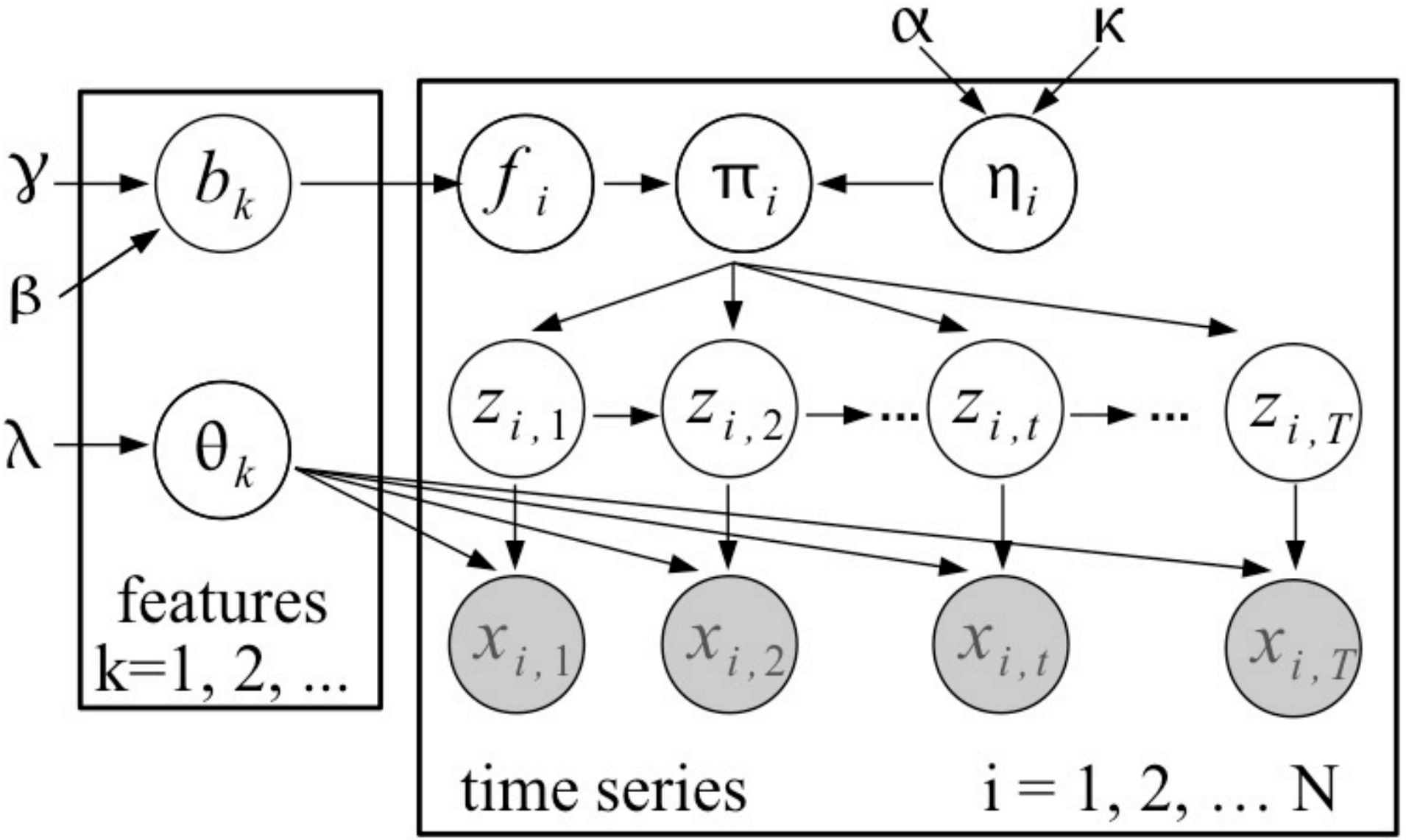
num. behaviors/video
Poisson(γ)

total num. behaviors
$O(\gamma \log N)$

often called "features" in machine learning literature

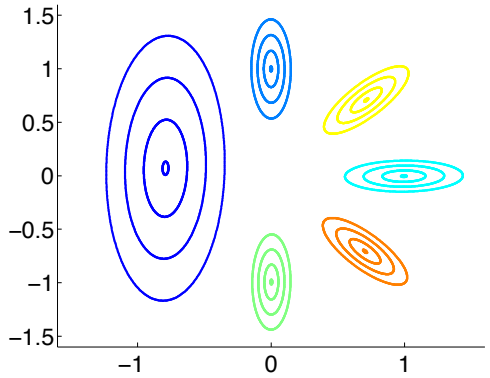
Encourage behavior sharing, but allow *sparsity* and rare behaviors

BP-HMM Graphical Model

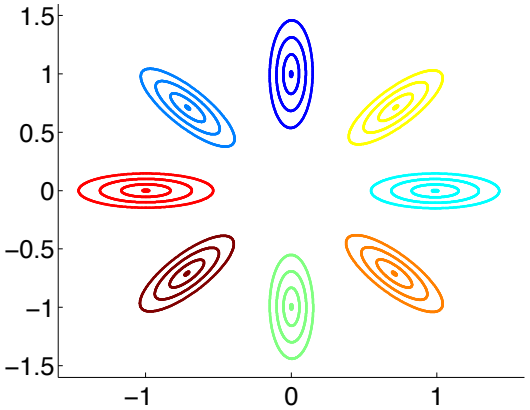


Baseline MCMC Learning

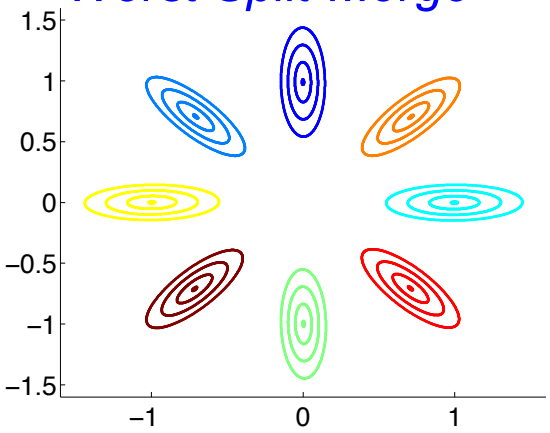
Best Prior

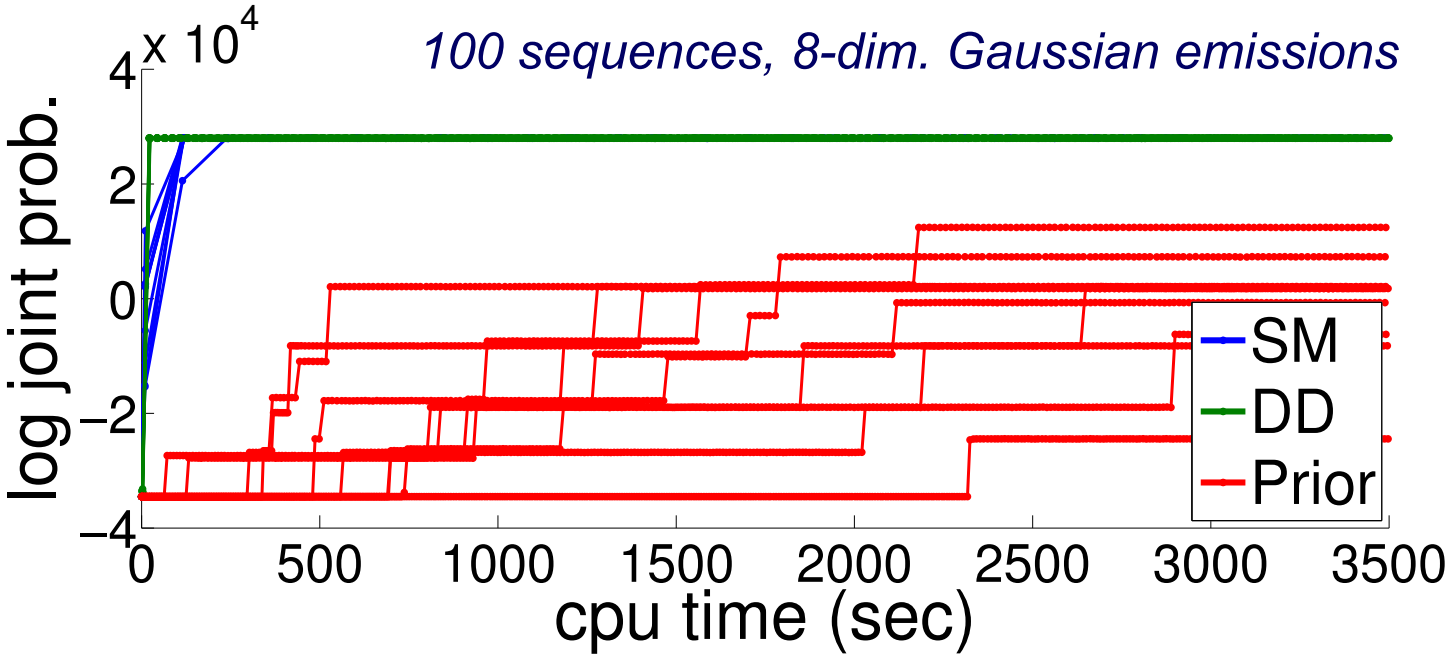


Worst Data-Driven



Worst Split-Merge



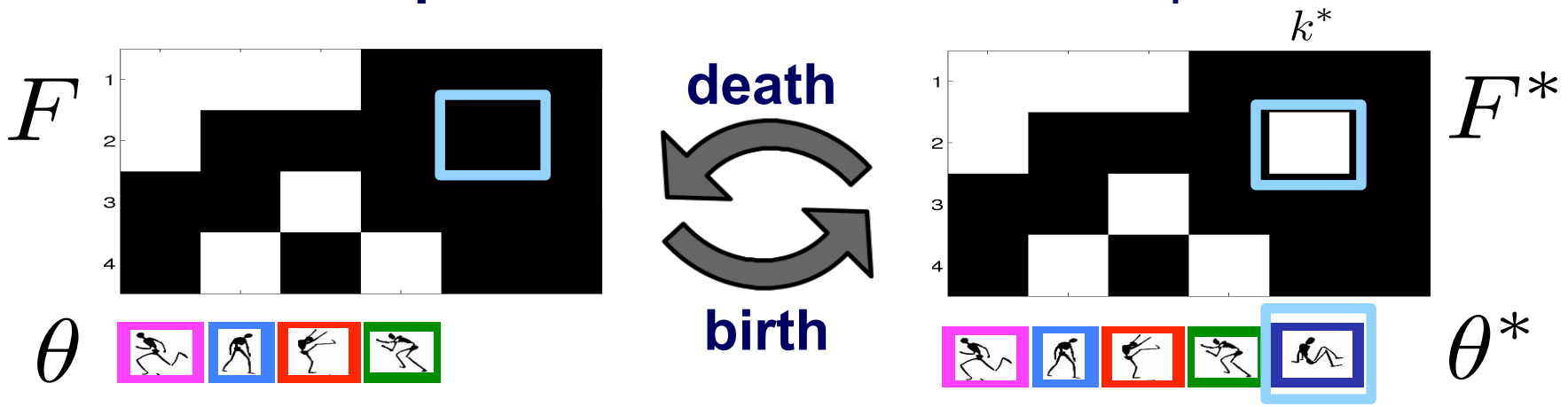


Prior proposals (Fox 2009) fairly sophisticated:

- Collapsed: Marginalize state sequences via dynamic programming in feature proposals
- Auxiliary variables: Blocked state sequence resampling to sample new transition and emission parameters
- But performance nevertheless very poor…

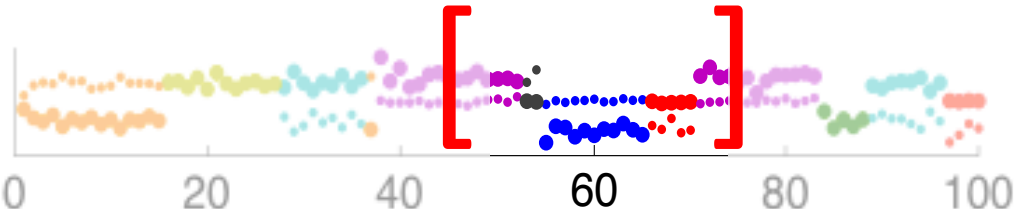
Data-Driven Birth/Death

Reversible Jump MCMC: Add or delete unique features



Propose from prior [Fox et al. NIPS 2009]

$$\theta^*_{k^*} \sim p(\theta)$$



Data-driven proposal [Hughes et al. NIPS 2012]

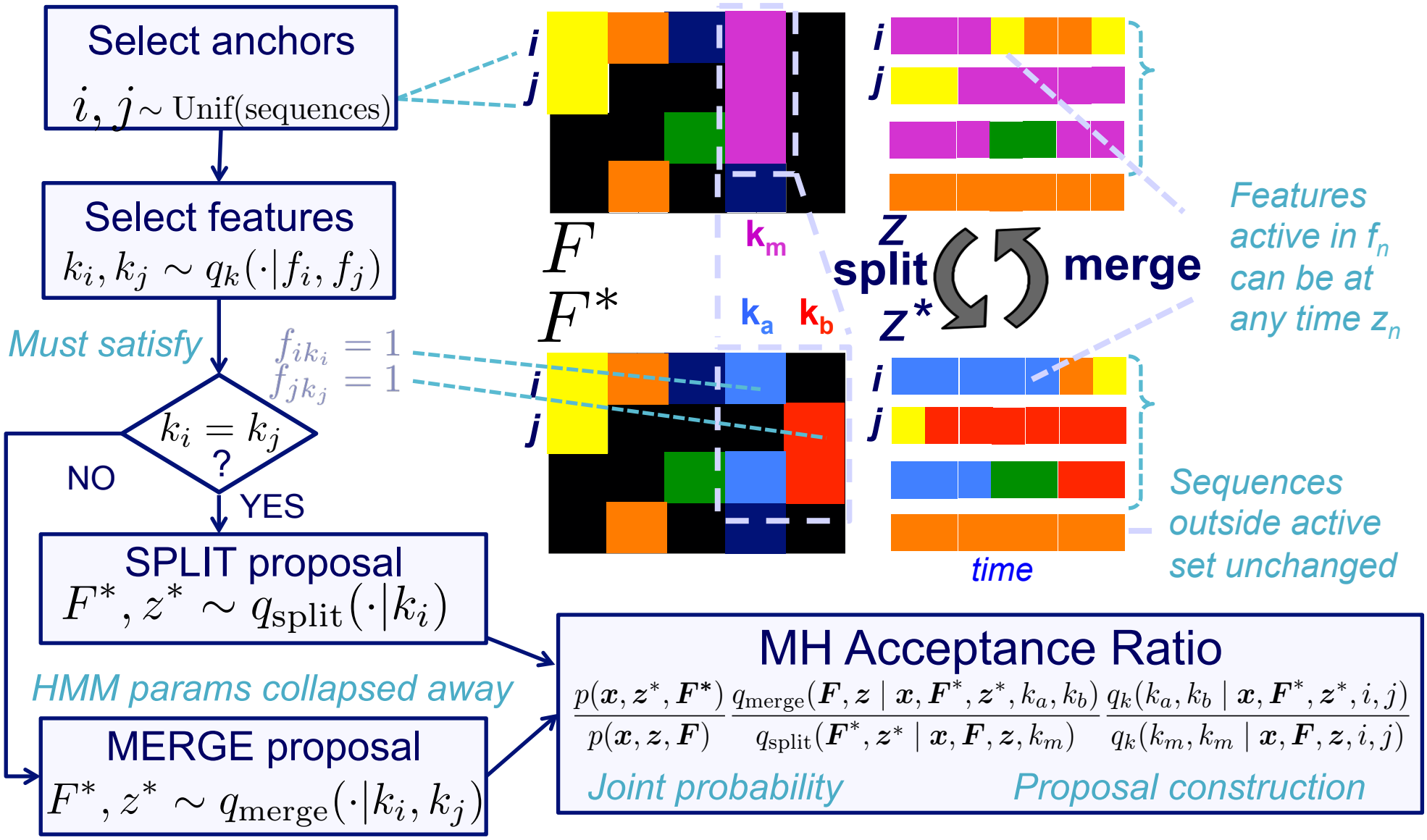
- select random window *W* of sequence
- proposal: mixture of prior and posterior over *W*

$$\theta^*_{k^*} \sim \frac{1}{2}p(\theta) + \frac{1}{2}p(\theta|x_{it} : t \in W)$$

Using mixture ensures good death move acceptance rate

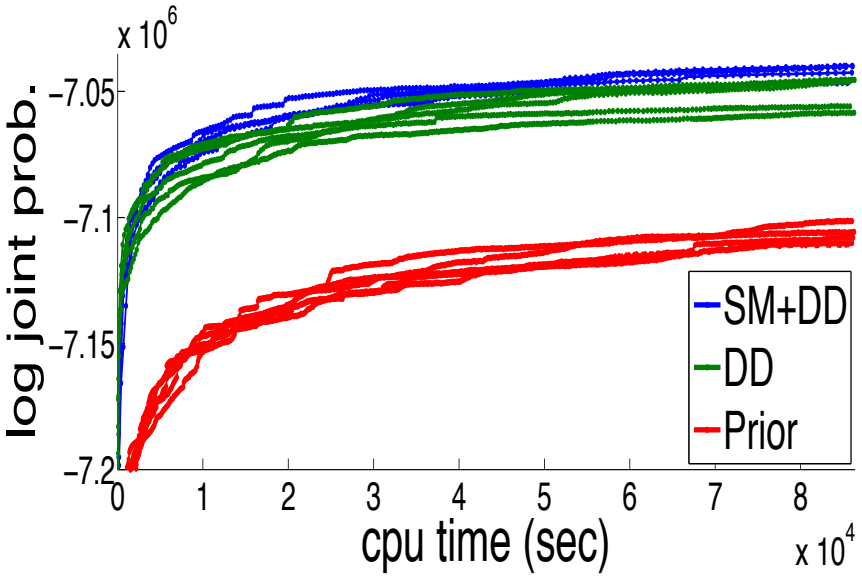
Sequential Split/Merge

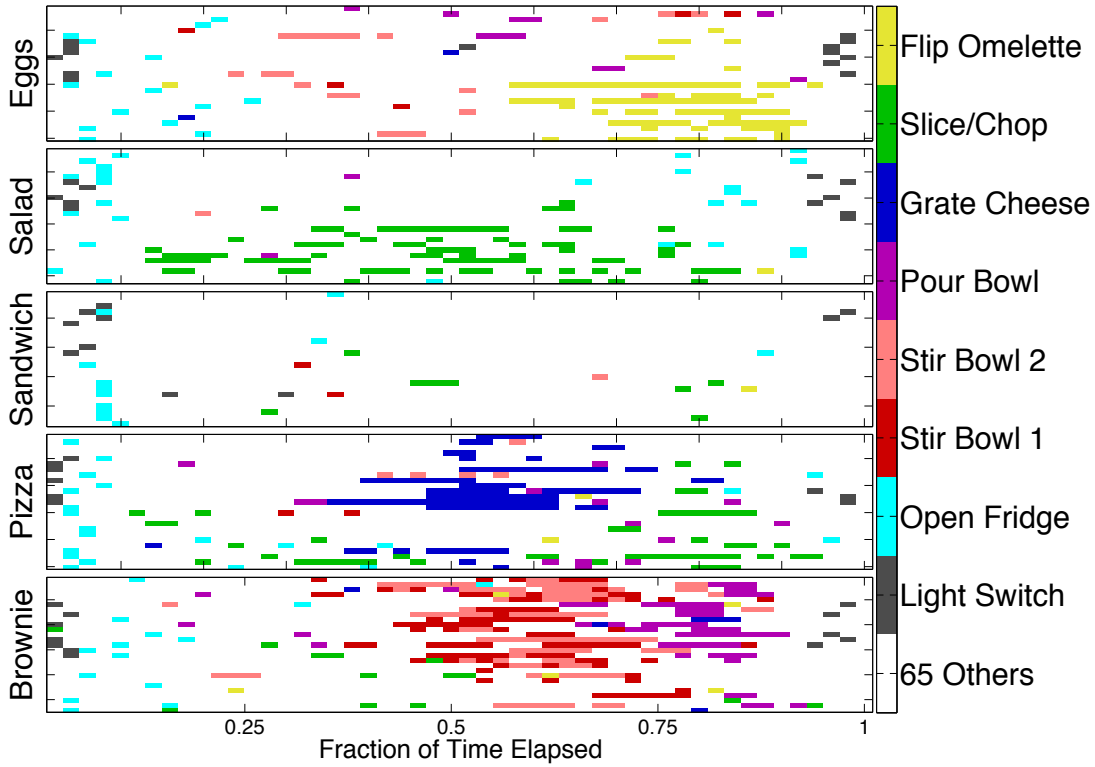
[Hughes NIPS 2012]



Sequential allocation [Dahl 2005] efficiently gives self-consistent split proposals

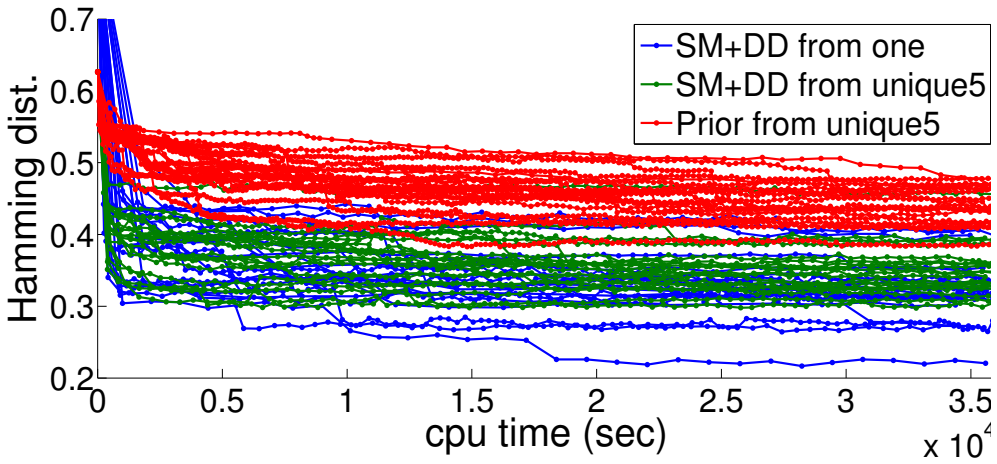
Video Activity Understanding

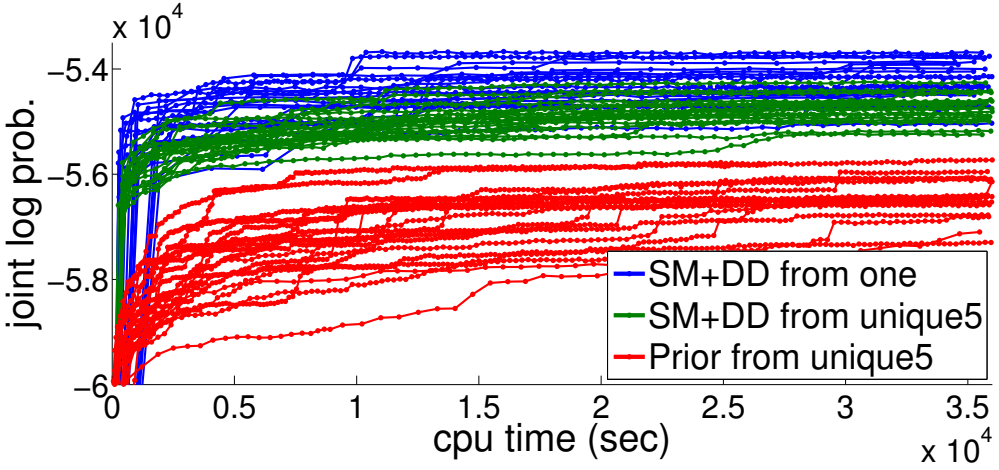


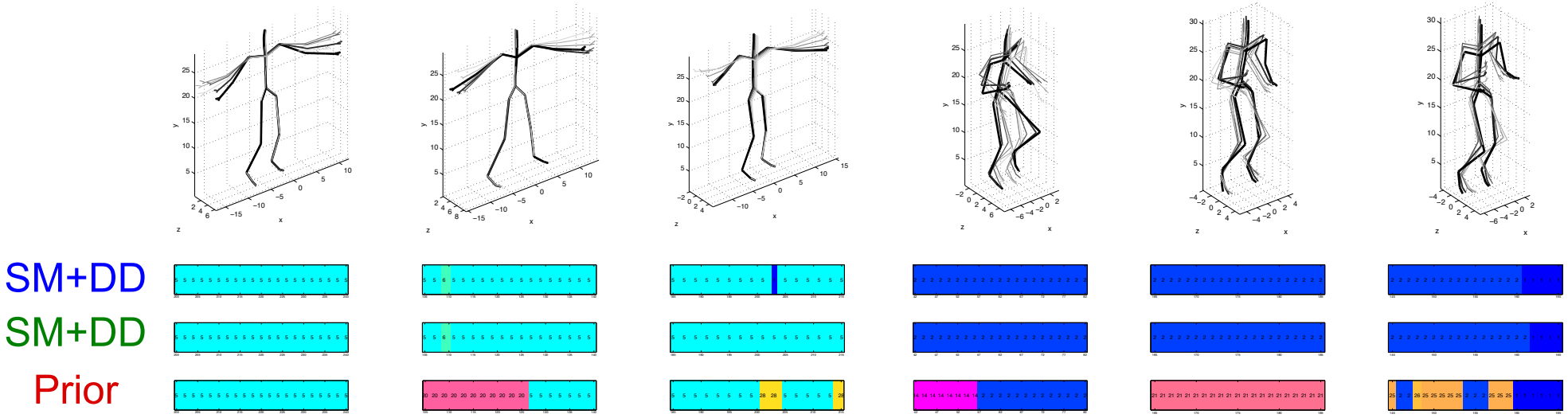


- 126 videos of recipe preparation (CMU Kitchen Database)
- Prior proposals are unusably poor
- Split-merge provides reasonable, but not complete, robustness to initialization

Mocap Analysis: 6 Sequences



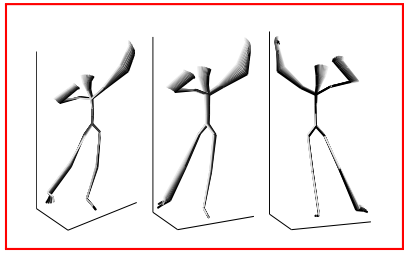


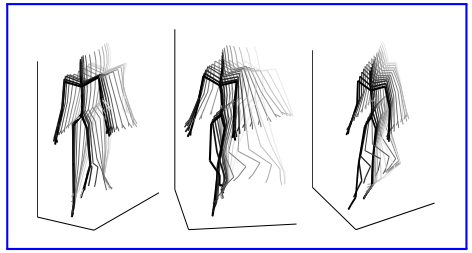


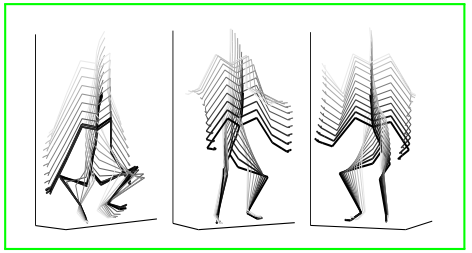
- 6 motion capture sequences (CMU Mocap Database)
- Human annotation of 12 partially shared exercises (ground truth validation)
- Huge difference in quality of “typical” chains for different algorithms

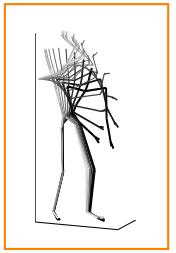
Mocap Analysis: 124 Sequences

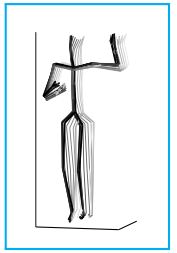
Analyzing all “Physical Activities & Sports” from CMU Mocap, here are 10 of 33 recovered behaviors:

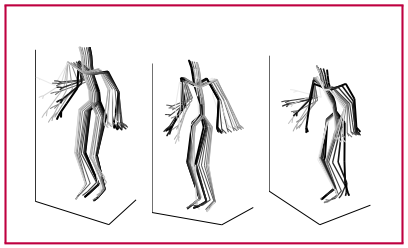

Ballet

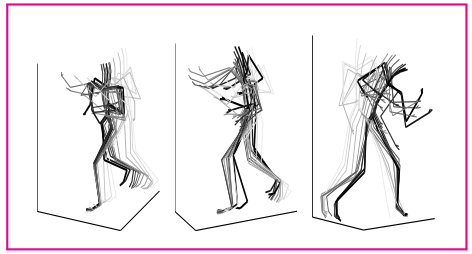

Walk

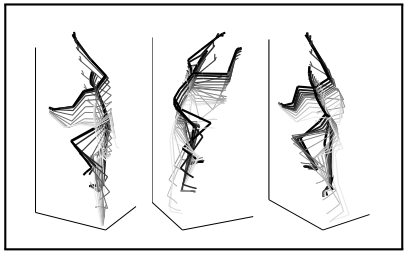

Squat

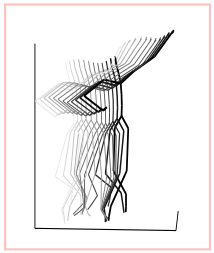

Sword

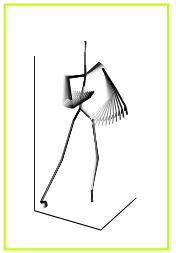

Lambada

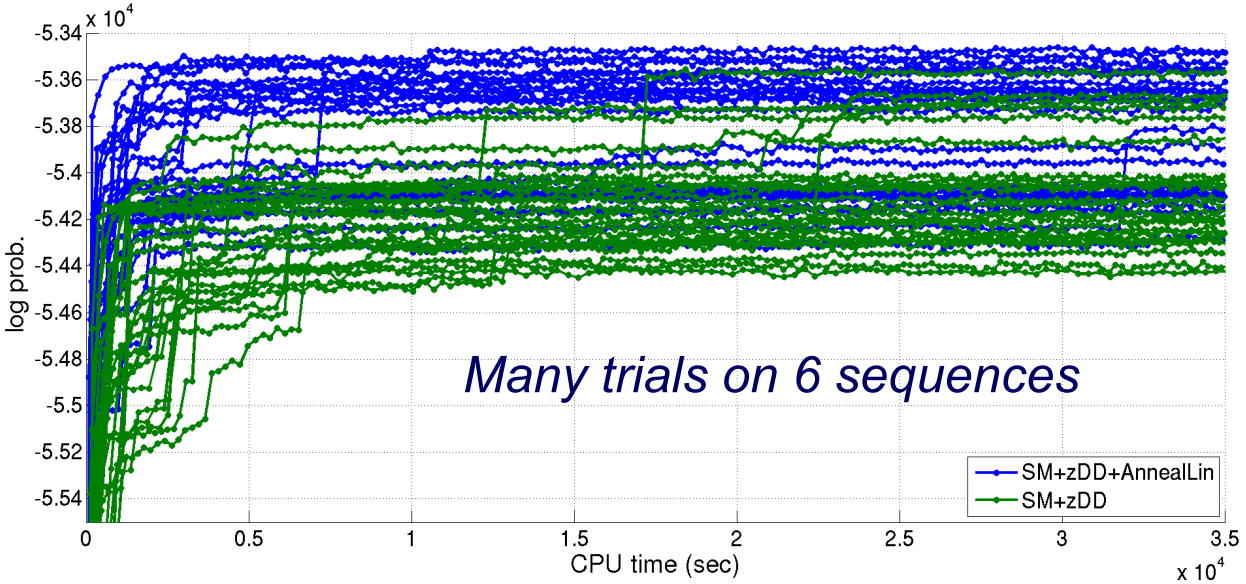

Dribble Basketball


Box


Climb


Indian Dance

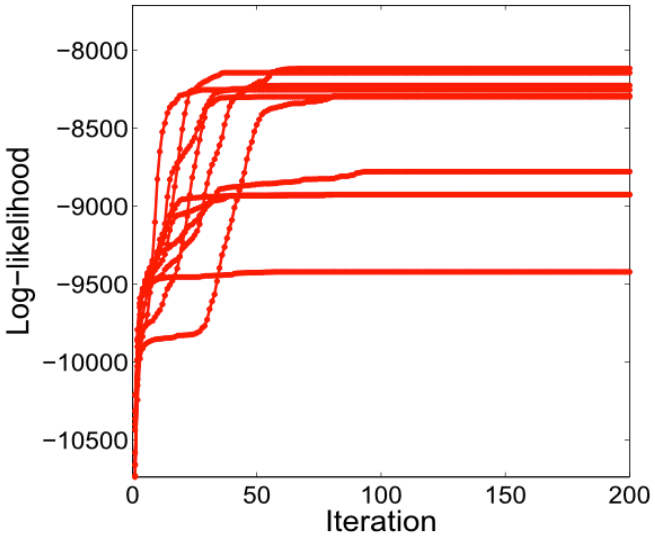

Tai Chi

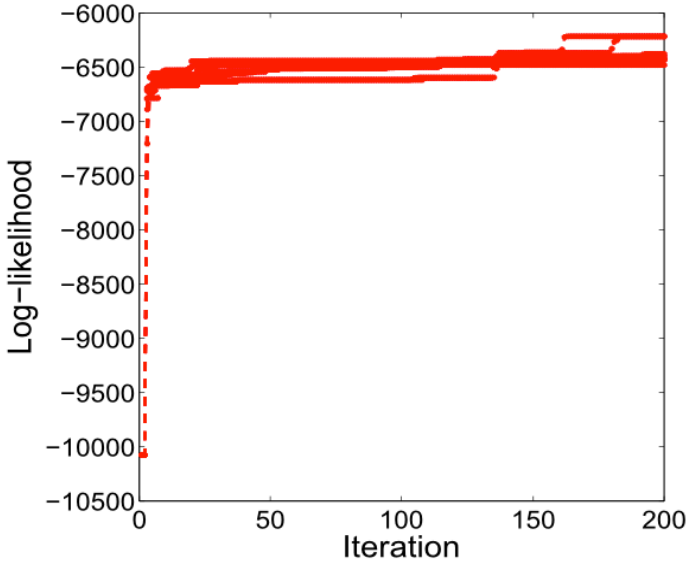


- Non-standard annealing: reduce proposal weight in MH acceptance ratio
- Hypothesis: Local reversibility is too strong for effective mixing of split-merge MCMC (widespread problem)

Spatial Image Segmentation

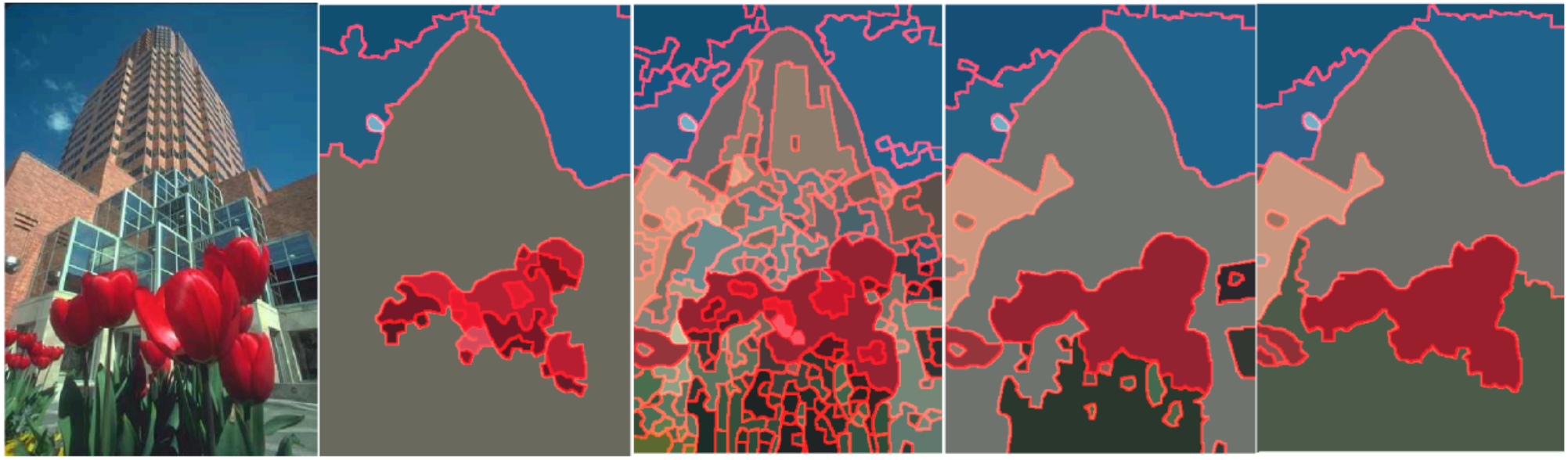
[Ghosh CVPR 2012]





Mean Field Variational

EP Stochastic Search



Best *Worst* *Best* *Worst*

Summary and Outlook

Toward Reliable Bayesian Nonparametric Learning

- Basic samplers, and conventional variational methods, are not as reliable as you've heard
- ***Maximization-Expectation search:*** Not the ultimate solution, but proof there's a problem
- ***Feasible:*** Split-Merge MCMC moves inspired by ME
  But local reversibility can still cause slow mixing…

Key Challenges

- New "default" learning algorithms, robust to initialization
- Automatic learning for more complex hierarchies, and rich temporal and spatial models of the world