

Jurgen Van Gael, Yunus Saatci, Yee Whye Teh, Zoubin Ghahramani

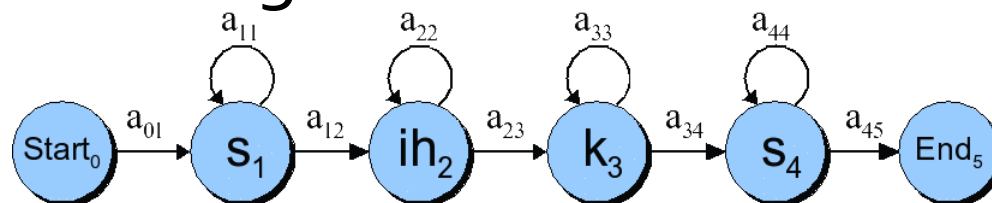
INFINITE HIDDEN MARKOV MODELS

Sequential Data (Time Series)

- Part-Of-Speech Tagging

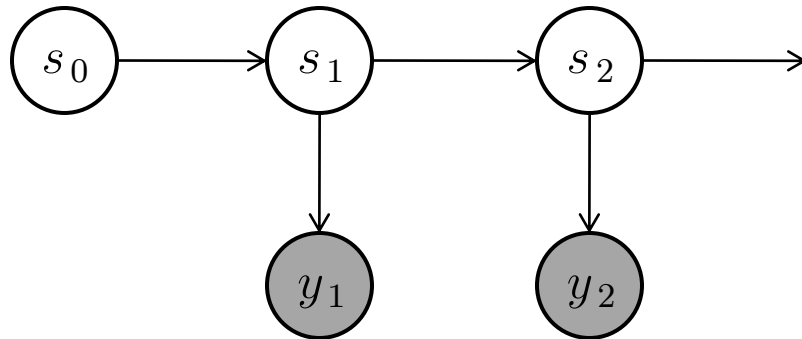
The	representative	put	chairs	on	the	table.
AT	NN	VBD	NNS	IN	AT	NN

- Speech Recognition



- DNA Sequence Alignment
- Machine Translation
- ...

Hidden Markov Model



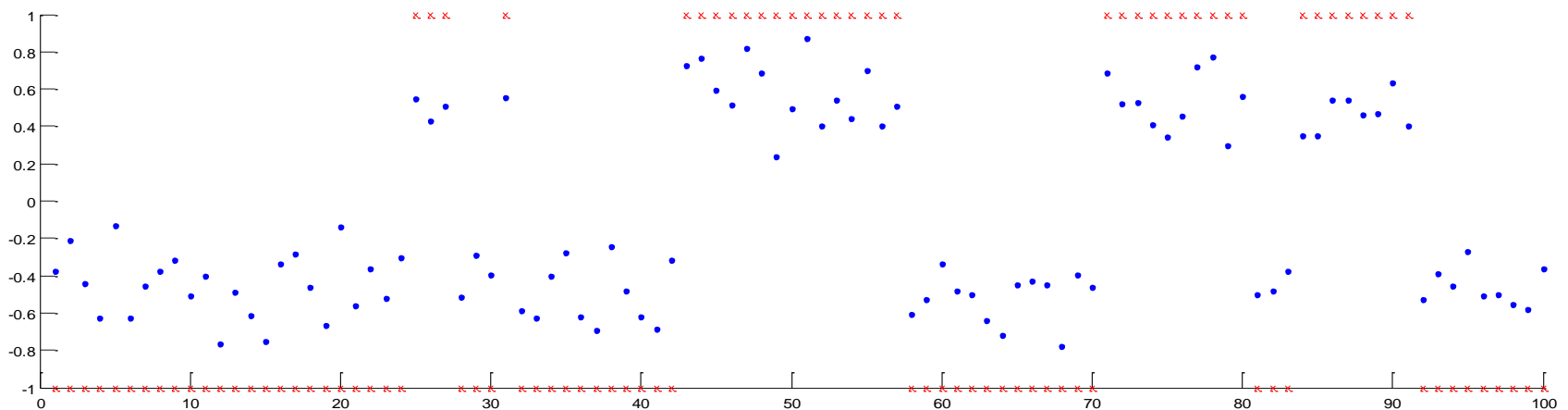
- Core: hidden K -state Markov chain
 - initial distribution: $p(s_0 = 1) = 1$
 - transition probability: $p(s_t = j | s_{t-1} = i) = \pi_{ij}$
- Peripheral: observation model $y_t \sim F(\phi_{s_t})$
 - e.g. $y_t | s_t \sim \mathcal{N}(\mu_{s_t}, \sigma_{s_t}^2)$ or $y_t | s_t \sim \text{Multinomial}(\theta_{s_t})$
 - easy to extend to other observation models
- Parameters of the model are K, π, ϕ

Hidden Markov Model

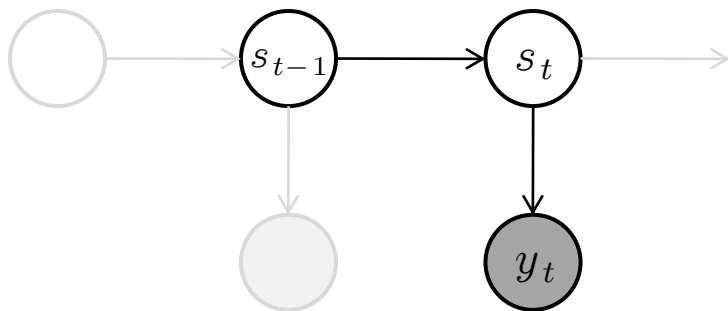
- Likelihood

$$p(y_1, \dots, y_T, s_1, \dots, s_T | \boldsymbol{\pi}, \boldsymbol{\phi}) = \prod_{i=1}^T p(s_t | s_{t-1}) p(y_t | s_t)$$
$$= \prod_{i=1}^T \pi_{s_{t-1}, s_t} F(\phi_{s_t})$$

- Example

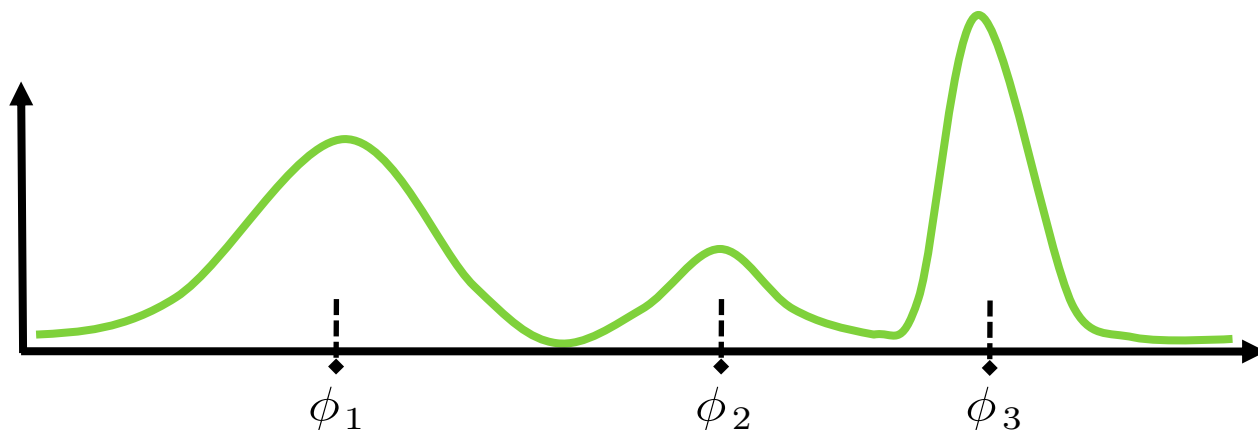


Different Perspective: HMM's as a sequential mixture model



$$\begin{aligned} p(y_t | s_{t-1} = k) &= \sum_{s_t=1}^K p(s_t | s_{t-1} = k) p(y_t | s_t) \\ &= \sum_{s_t=1}^K \pi_{k,s_t} F(\phi_{s_t}) \end{aligned}$$

What is conditional distribution of y_t ?



$p(y_t | s_{t-1} = k)$ is a mixture distribution with K components.

Infinite Hidden Markov Model

- We want HMM in the limit of $K \rightarrow \infty$

Dirichlet Process

- Specifies a distribution over distributions
- We write $G_k \sim \text{DP}(\alpha, H)$ with
 - concentration parameter α
 - base distribution H
- A DP is discrete with probability 1

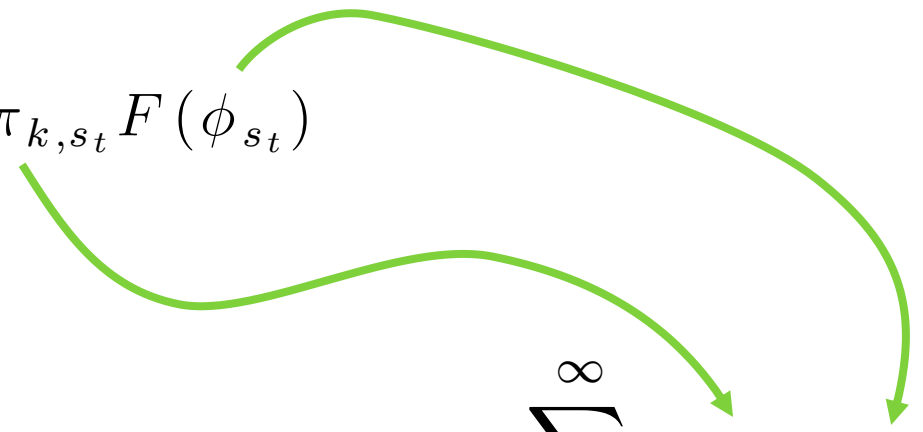
$$G_k(\phi) = \sum_{k'=1}^{\infty} \pi_{k'} \delta_{\phi_{k'}}(\phi) \quad \forall k' : \phi_{k'} \sim H,$$

- A DP specifies both mixture weights and parameters

Infinite Hidden Markov Model

- Idea: introduce DP's
 - identify mixture weights with HMM transitions
 - identify base distribution draws with observation model parameters

$$p(y_t | s_{t-1} = k) = \sum_{s_t=1}^K \pi_{k,s_t} F(\phi_{s_t})$$

$$G_k(\phi) = \sum_{k'=1}^{\infty} \pi_{k,k'} \delta_{\phi_{k,k'}}(\phi)$$


Infinite Hidden Markov Model

- Almost there: if H is continuous, all DP's will have different parameters

→ introduce a DP (G_0) between H and G_k

- Formally $G_0 \sim \text{DP}(\gamma, H)$

$$G_k \sim \text{DP}(\alpha, G_0)$$

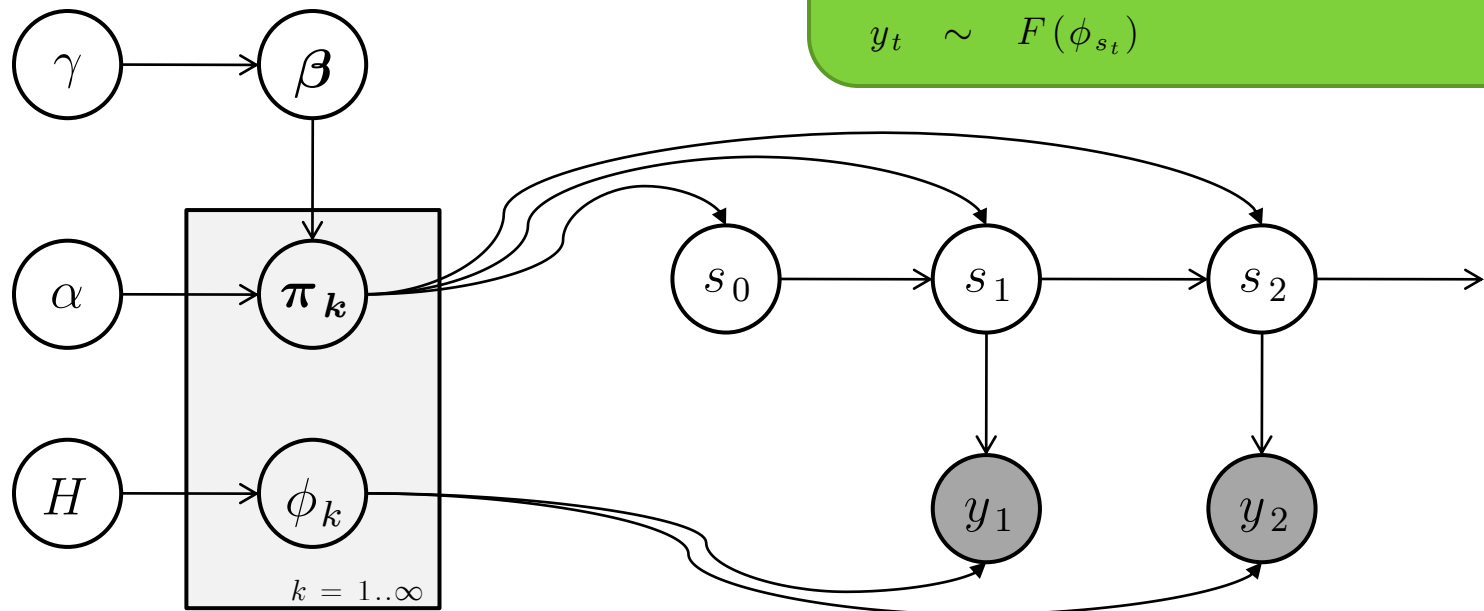
- Hierarchical Dirichlet Process [Teh et al., 2006]

Infinite Hidden Markov Model

- Recall $G_0(\phi) = \sum_{k'=1}^{\infty} \beta_{k'} \delta_{\phi_{k'}}(\phi)$ $\forall k' : \phi_{k'} \sim H, G_k(\phi) = \sum_{k'=1}^{\infty} \pi_{k,k'} \delta_{\phi_{k'}}(\phi)$

- Generative Model for iHMM

$\beta \sim \text{Stick}(\gamma),$
 $\phi_k \sim H,$
 $\pi_k \sim \text{Dirichlet}(\alpha\beta),$
 $s_t \sim \text{Multinomial}(\pi_{s_{t-1}}), \quad (s_0 = 1)$
 $y_t \sim F(\phi_{s_t})$



HMM versus iHMM

HMM is fully specified given

- K parameters
- K by K transition matrix

ϕ	ϕ_1	ϕ_2	ϕ_3	\dots	ϕ_K
π	π_{11}	π_{12}	\dots		
	π_{12}	\dots			
	\vdots				
					π_{KK}

HMM versus iHMM

iHMM is fully specified given an infinite number of DP's ?!?

ϕ	ϕ_1	ϕ_2	ϕ_3
π	π_{11}	π_{12}	...		
	π_{12}	...			
	⋮				
	⋮				
	⋮				

Inference & Learning

- Hidden Markov Model
 - Inference (= hidden states)
 - Dynamic Programming
 - Gibbs Sampling
 - Learning (= parameters)
 - Expectation Maximization
 - Gibbs Sampling
- Infinite Hidden Markov Model (so far)
 - Inference (= hidden states): Gibbs sampling
 - Learning (= parameters): Gibbs sampling
- This is unfortunate: Gibbs sampling for time series?!?

Dynamic Programming

Forward-Backtrack Sampling

1. Compute conditional probabilities

1. Initialize

$$p(s_0 = 1) = 1$$

$$O(TK^2)$$

2. For each $t = 1 \dots T$

$$p(s_t | y_{1:t}) \propto p(y_t | s_t) \sum_{s_{t-1}} p(s_t | s_{t-1}) p(s_{t-1} | y_{1:t-1})$$

2. Sample hidden states

1. Sample for time T

$$p(s_T | y_{1:T})$$

$$O(TK)$$

2. For each $t = T-1 \dots 1$

$$p(s_t | s_{t+1}, y_{1:t}) \propto p(s_{t+1} | s_t) p(s_t | y_{1:t})$$

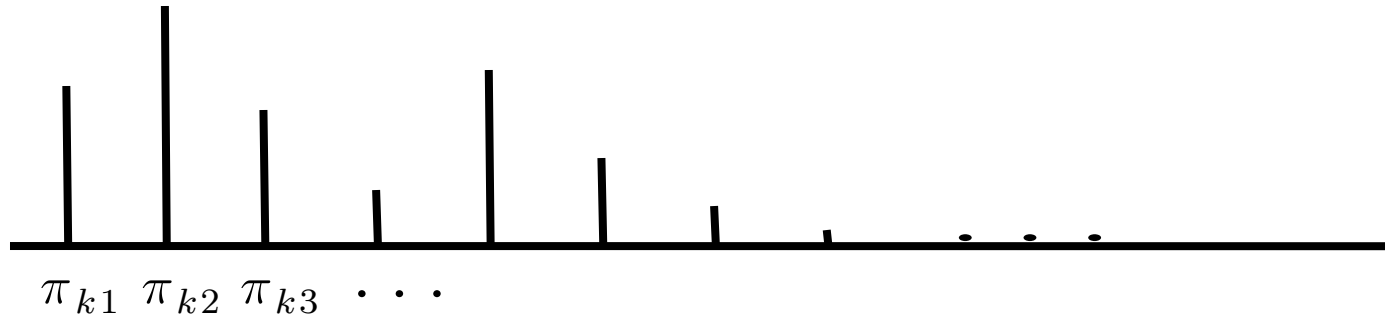
Beam Sampling

- Can we use Forward-Backtrack for iHMM?
 - No, $O(TK^2)$ with $K \rightarrow \infty$ is intractable
- A (bad?) idea:
 - Truncate transition matrix
 - Use dynamic programming to sample \mathbf{s}
- This is only approximately correct.

→ Beam Sampling = Slice Sampling
+
Dynamic Programming

Beam Sampling

- Each G_k can be represented as



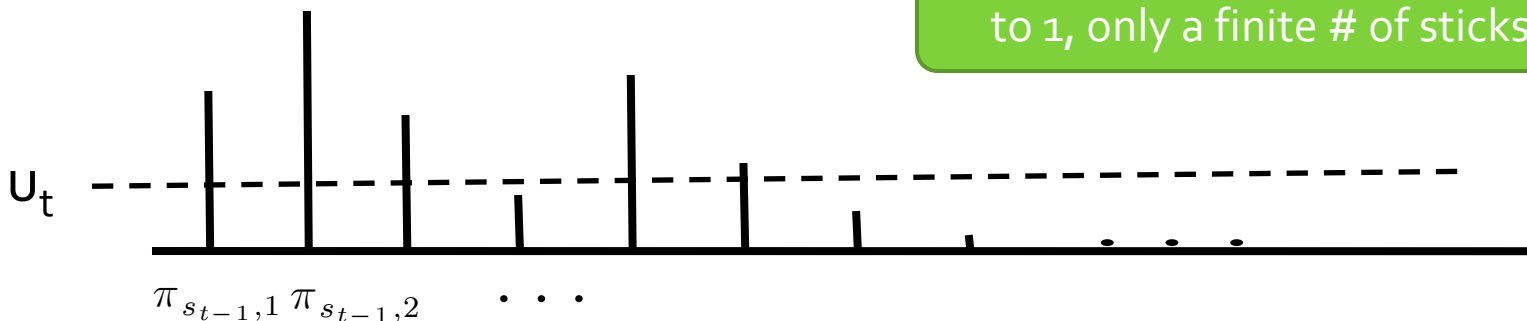
- Let us introduce an auxiliary variable

$$u_t \sim \text{Uniform}(0, \pi_{s_{t-1}, s_t})$$

[Neal, 2003; Walker 2006]

- u_t partitions up $G_{s_{t-1}}$

Key Observation: since π must sum to 1, only a finite # of sticks $> u_t$.



Beam Sampling

■ Algorithm

1. Initialize hidden states + parameters

2. While (enough samples)

1. Sample $p(u | s)$: $u_t \sim \text{Uniform}(0, \pi_{s_{t-1}, s_t})$

2. Sample $p(s | u, y)$ using dynamic programming

1. Initialize DP $p(s_0 = 1) = 1$

2. For each $t = 1 \dots T$

$$p(s_t | y_{1:t}, u_{1:t}) \propto p(y_t | s_t) \sum_{s_{t-1}: u_t \leq \pi_{s_{t-1}, s_t}} p(s_{t-1} | y_{1:t-1}, u_{1:t-1})$$

3. Sample T $p(s_T | y_{1:T})$

4. Sample $t = T-1 \dots 1$ $p(s_t | s_{t+1}, y_{1:t}) \propto p(s_{t+1} | s_t) p(s_t | y_{1:t})$

3. Resample $\pi, \phi, \beta, \gamma, \alpha | s$

Beam Sampling Properties

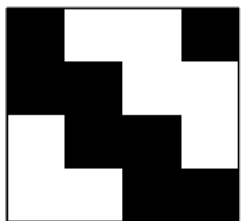
- The slice sampler adaptively truncates the infinitely large transition matrix
- Dynamic program allows us to resample the whole sequence \mathbf{s}
 - Gibbs sampler only changes one hidden state conditioned on all other states
- The dynamic program needs all parameters to be instantiated
 - Gibbs sampler can collapse variables
 - Beam sampler can do inference for non-conjugate models
- (Hyper)parameter sampling is identical to Gibbs sampling

Experiment I - HMM Data

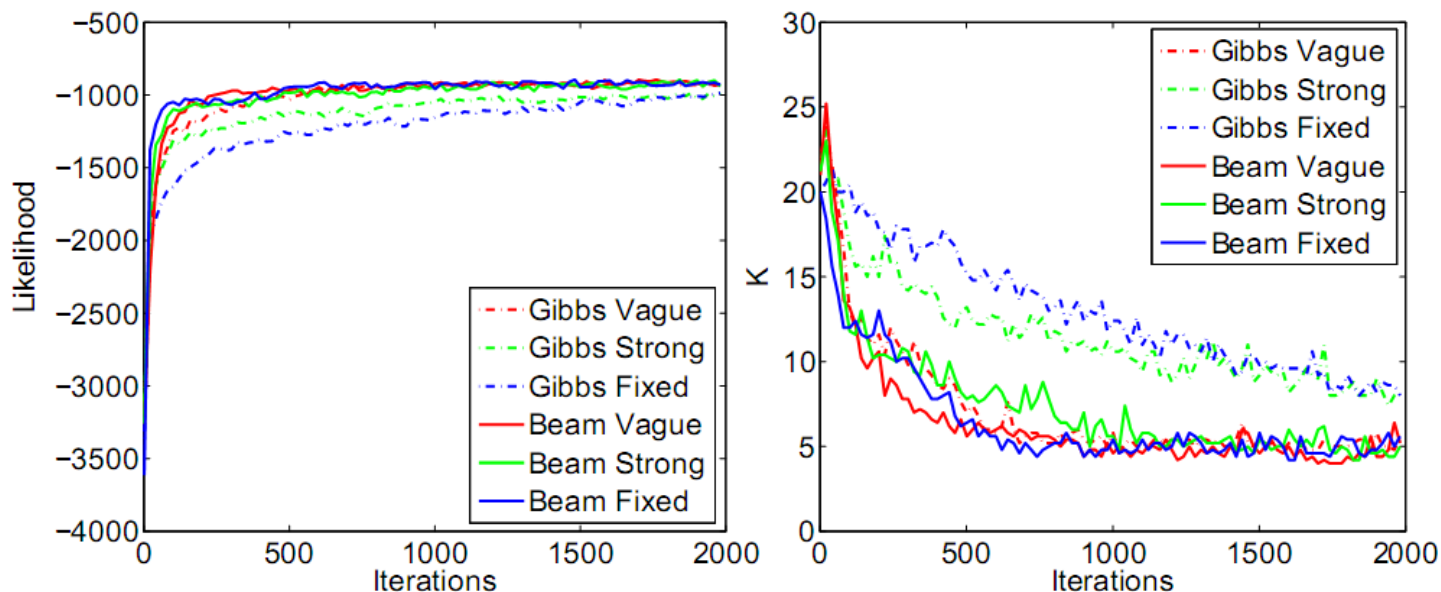
Synthetic data generated by HMM with $K=4$

- Vague : $\alpha \sim \text{Gamma}(1,1)$; $\gamma \sim \text{Gamma}(2,1)$
- Strong: $\alpha \sim \text{Gamma}(6,15)$; $\gamma \sim \text{Gamma}(16,4)$
- Fixed : $\alpha = 0.4$; $\gamma = 3.8$

Transition Matrix



Emission Matrix

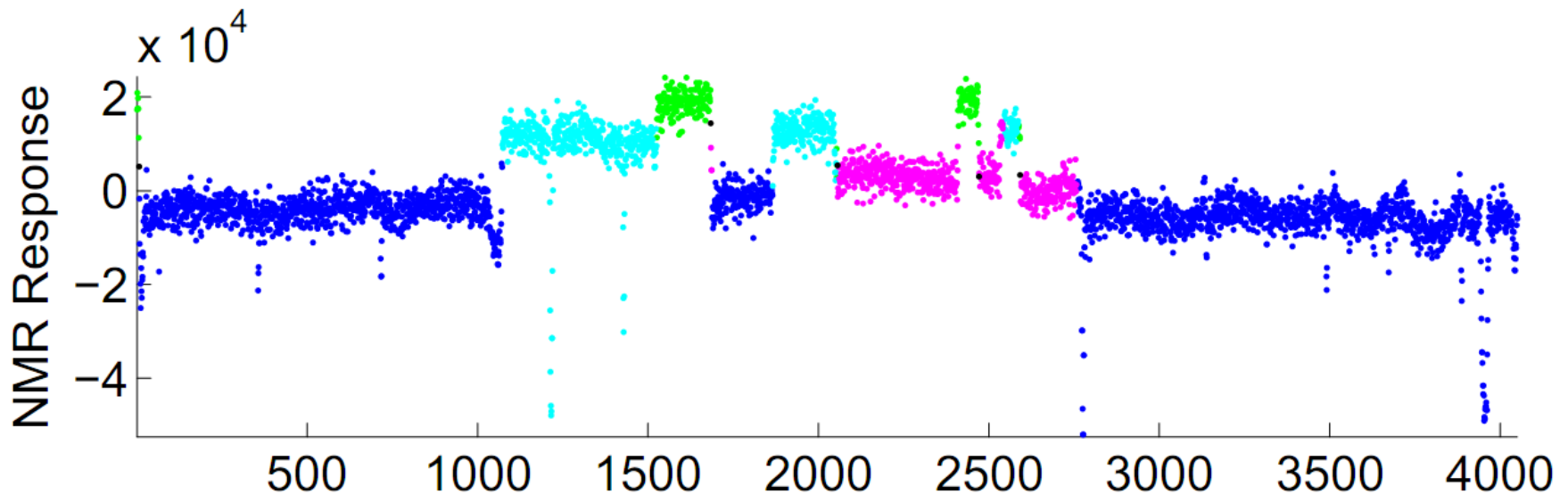


Experiment II – Changepoint Detection

Well Log (NMR Response) – Change point Detection

- 4050 noisy NMR response measurements
- Output model is Student-t with known scale

Beam sampler output of iHMM after 8000 iterations:



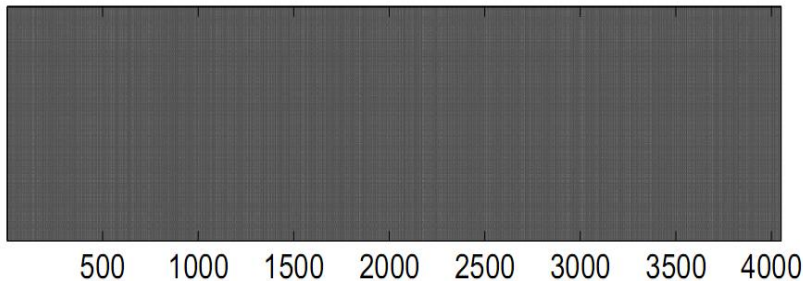
Experiment II – Changepoint Detection

What is probability of two data points in same cluster?

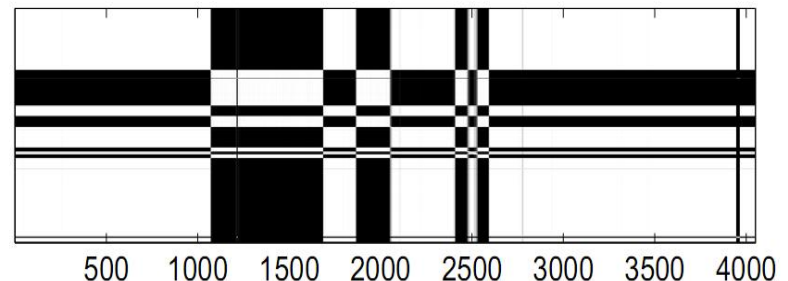
- Left: average over first 5 samples
- Right: average over last 30 samples datapoints

Note: 1) gray areas for beam; 2) slower mixing for Gibbs

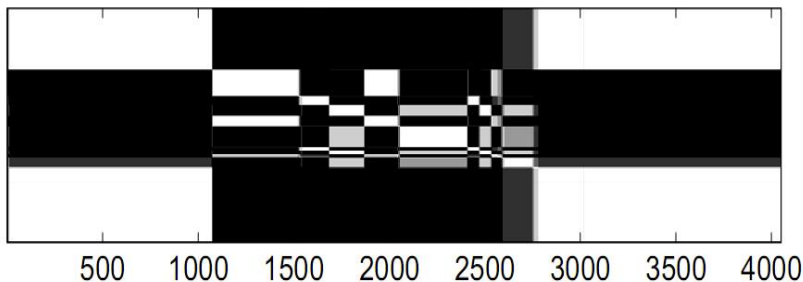
Gibbs Sampler



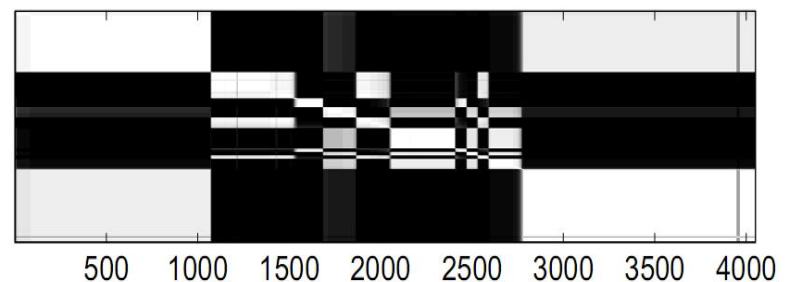
Gibbs Sampler



Beam Sampler



Beam Sampler



Conclusion

- iHMM could be good alternative for HMM
- Beam sampler is algorithm of choice for iHMM
 - at least as good mixing properties
 - accommodates non-conjugate models

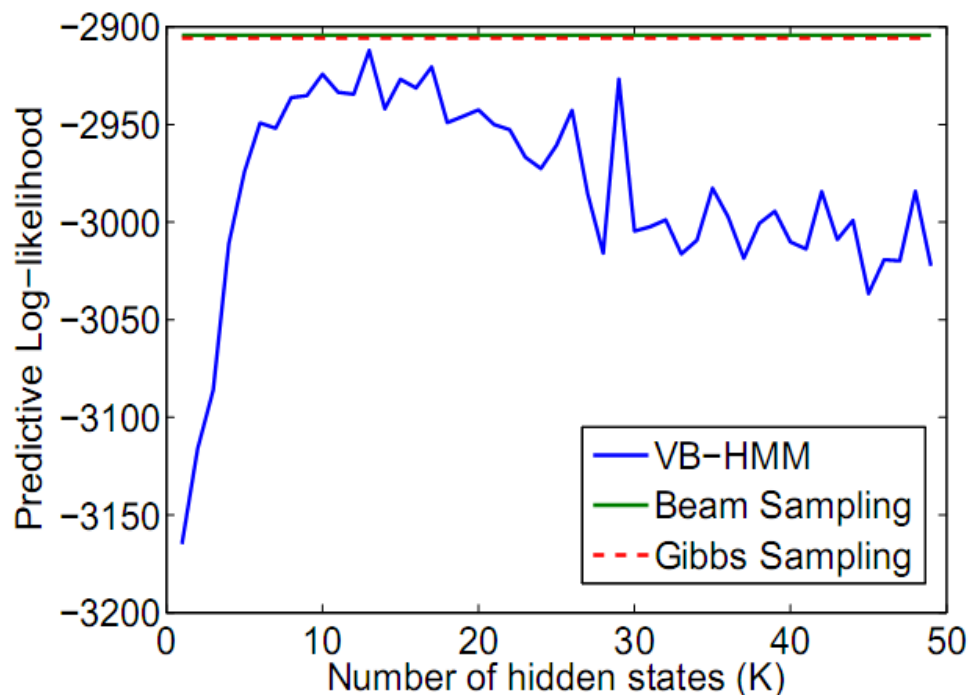
Future Work

- Extend models
 - IO-iHMM
 - AR(MA)-iHMM
 - infinite Switching State Space Models
 - infinite Factorial HMM
- Challenge: automatically generate inference algorithms?

Experiment III – Text Prediction

Alice in Wonderland

- training data: 1000 characters from 1st chapter
- 35 possible output characters
- testing data: 1000 subsequent characters



VB-HMM:

- Transition matrix: Dirichlet($4/K, \dots, 4/K$)
- Emission matrix: Dirichlet(0.3)

iHMM:

- $\alpha \sim \text{Gamma}(4, 1)$
- $\gamma \sim \text{Gamma}(1, 1)$
- $H \sim \text{Dirichlet}(0.3)$