Bayesian Hidden Markov Models and Extensions

Zoubin Ghahramani
Department of Engineering
University of Cambridge

joint work with

Matt Beal, Jurgen van Gael,
Yunus Saatci, Tom Stepleton, Yee Whye Teh

Modeling time series

Sequence of observations:

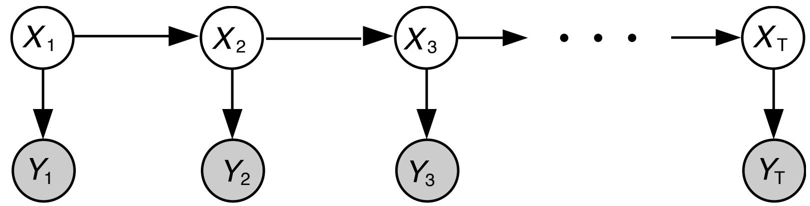
$$\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \ldots, \mathbf{y}_t$$

For example:

- Sequence of images
- Speech signals
- Stock prices
- Kinematic variables in a robot
- Sensor readings from an industrial process
- Amino acids, etc...

Goal: To build a probilistic model of the data:
something that can predict $p(\mathbf{y}_t | \mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \mathbf{y}_{t-3} \ldots)$

Causal structure and "hidden variables"



Speech recognition:

- $\mathbf{x}$ - underlying phonemes or words
- $\mathbf{y}$ - acoustic waveform

Vision:

- $\mathbf{x}$ - object identities, poses, illumination
- $\mathbf{y}$ - image pixel values

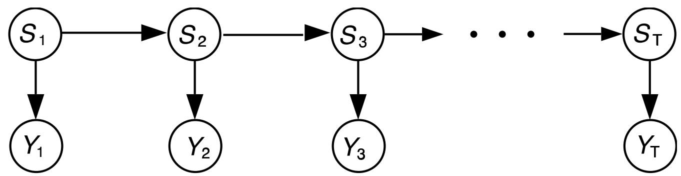
Industrial Monitoring:

- $\mathbf{x}$ - current state of molten steel in caster
- $\mathbf{y}$ - temperature and pressure sensor readings

Two frequently-used tractable models:

- Linear-Gaussian state-space models
- Hidden Markov models

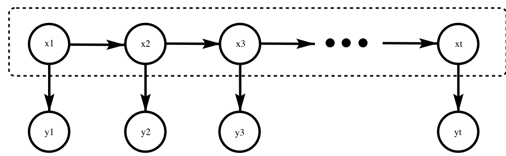
Graphical Model for HMM



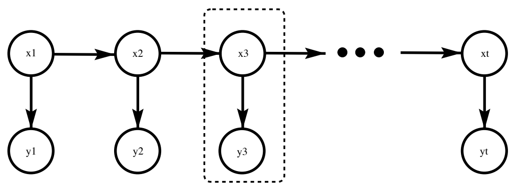
- Discrete hidden states $s_t \in \{1 \ldots, K\}$, and outputs $\mathbf{y}_t$ (discrete or continuous). Joint probability factorizes:

$$\mathsf{P}(s_1, \ldots, s_\tau, \mathbf{y}_1 \ldots, \mathbf{y}_\tau) = \mathsf{P}(s_1)\mathsf{P}(\mathbf{y}_1|s_1) \prod_{t=2}^{\tau} \mathsf{P}(s_t|s_{t-1})\mathsf{P}(\mathbf{y}_t|s_t)$$

- a Markov chain with stochastic measurements:



- or a mixture model with states coupled across time:



Hidden Markov Models

- Hidden Markov models (HMMs) are widely used, but how do we choose the number of hidden states?
  - Variational Bayesian learning of HMMs
  - A non-parametric Bayesian approach: infinite HMMs.
- Can we extract richer structure from sequences by grouping together states in an HMM?
  - Block-diagonal iHMMs.
- A single discrete state variable is a poor representation of the history. Can we do better?
  - Factorial HMMs
- Can we make Factorial HMMs non-parametric?
  - infinite factorial HMMs and the Markov Indian Buffet Process

Part I:

Variational Bayesian learning of Hidden Markov Models

Bayesian Learning

Apply the basic rules of probability to learning from data.

Data set: $\mathcal{D} = \{x_1, \ldots, x_n\}$ Models: m, m' etc. Model parameters: θ

Prior probability of models: $P(m)$, $P(m')$ etc.
Prior probabilities of model parameters: $P(\theta|m)$
Model of data given parameters (likelihood model): $P(x|\theta, m)$

If the data are independently and identically distributed then:

$$P(\mathcal{D}|\theta, m) = \prod_{i=1}^{n} P(x_i|\theta, m)$$

Posterior probability of model parameters:

$$P(\theta|\mathcal{D}, m) = \frac{P(\mathcal{D}|\theta, m)P(\theta|m)}{P(\mathcal{D}|m)}$$

Posterior probability of models:

$$P(m|\mathcal{D}) = \frac{P(m)P(\mathcal{D}|m)}{P(\mathcal{D})}$$

Bayesian Occam's Razor and Model Comparison

Compare model classes, e.g. m and m', using posterior probabilities given $\mathcal{D}$:

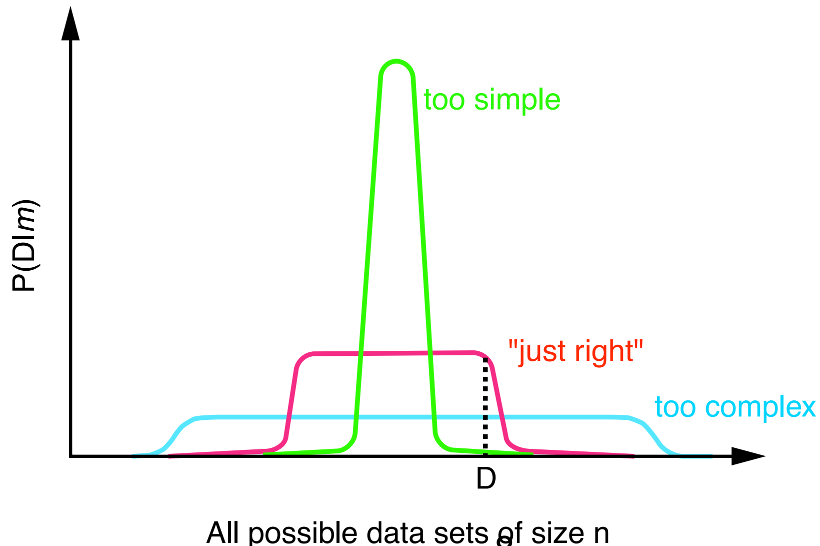
$$p(m|\mathcal{D}) = \frac{p(\mathcal{D}|m)\, p(m)}{p(\mathcal{D})}, \qquad p(\mathcal{D}|m) = \int p(\mathcal{D}|\boldsymbol{\theta}, m)\, p(\boldsymbol{\theta}|m)\, d\boldsymbol{\theta}$$

Interpretations of the Marginal Likelihood ("model evidence"):

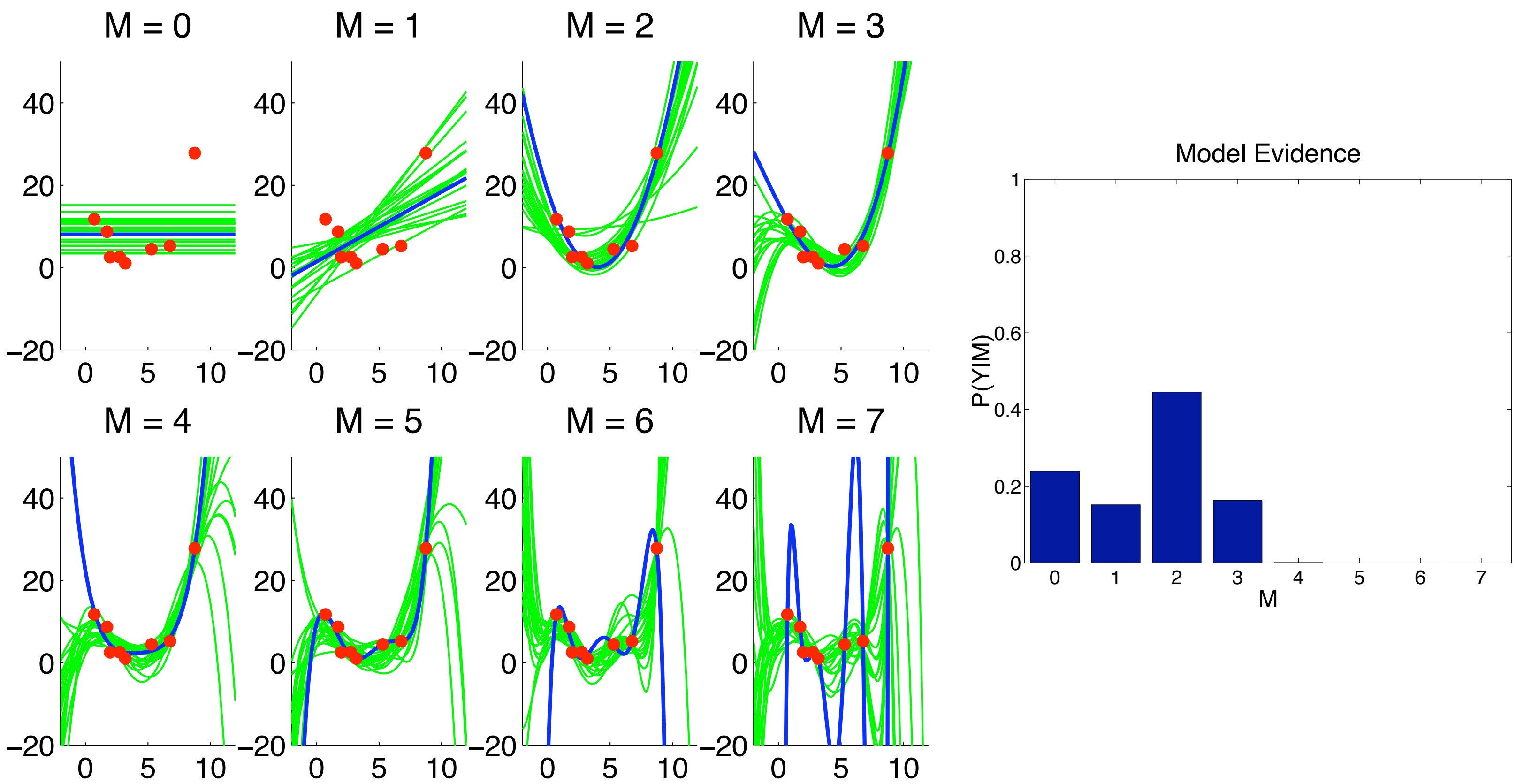
- The probability that *randomly selected* parameters from the prior would generate $\mathcal{D}$.
- Probability of the data under the model, *averaging* over all possible parameter values.
- $\log_2 \left(\frac{1}{p(\mathcal{D}|m)} \right)$ is the number of *bits of surprise* at observing data $\mathcal{D}$ under model m.

Model classes that are too simple are unlikely to generate the data set.

Model classes that are too complex can generate many possible data sets, so again, they are unlikely to generate that particular data set at random.



Bayesian Model Comparison: Occam's Razor at Work

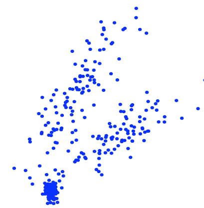


For example, for quadratic polynomials ($m = 2$): $y = a_0 + a_1 x + a_2 x^2 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and parameters $\boldsymbol{\theta} = (a_0\ a_1\ a_2\ \sigma)$

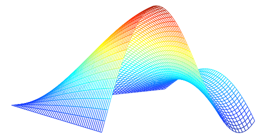
`demo: polybayes`

Learning Model Structure

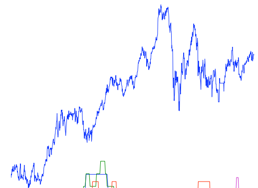
How many clusters in the data?



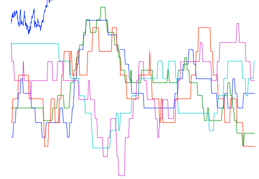
What is the intrinsic dimensionality of the data?



Is this input relevant to predicting that output?



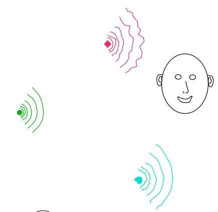
What is the order of a dynamical system?



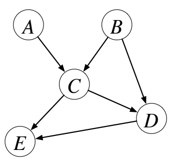
How many states in a hidden Markov model?

`SVYDAAAQLTADVKKDLRDSWKVIGSDKKGNGVALMTTY`

How many auditory sources in the input?



Which graph structure best models the data?



`demo: run_simple`

Variational Bayesian Learning

Lower Bounding the Marginal Likelihood

Let the observed data be $\mathcal{D}$, the hidden state variables be s, and the parameters be $\boldsymbol{\theta}$.

Lower bound the marginal likelihood (Bayesian model evidence) using Jensen's inequality:

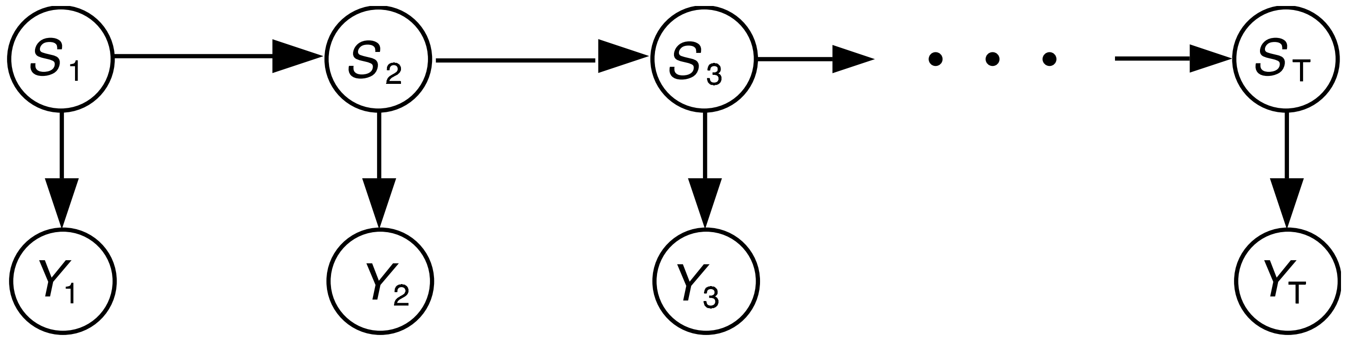
$$\begin{aligned}
\log P(\mathcal{D}|m) &= \log \int \sum_{\mathbf{s}} P(\mathcal{D}, \mathbf{s}, \boldsymbol{\theta}|m)\, d\boldsymbol{\theta} \\
&= \log \int \sum_{\mathbf{s}} Q(\mathbf{s}, \boldsymbol{\theta}) \frac{P(\mathcal{D}, \mathbf{s}, \boldsymbol{\theta}|m)}{Q(\mathbf{s}, \boldsymbol{\theta})}\, d\boldsymbol{\theta} \\
&\geq \int \sum_{\mathbf{s}} Q(\mathbf{s}, \boldsymbol{\theta}) \log \frac{P(\mathcal{D}, \mathbf{s}, \boldsymbol{\theta}|m)}{Q(\mathbf{s}, \boldsymbol{\theta})}\, d\boldsymbol{\theta}.
\end{aligned}$$

Here $Q(\mathbf{s}, \boldsymbol{\theta})$ is an approximation to the posterior $P(\mathbf{s}, \boldsymbol{\theta}|\mathcal{D}, m)$.
Assume $Q(\mathrm{s}, \boldsymbol{\theta})$ is a simpler factorised distribution:

$$\log P(\mathcal{D}|m) \geq \int \sum_{\mathbf{s}} Q_{\mathrm{s}}(\mathbf{s}) Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \log \frac{P(\mathcal{D}, \mathbf{s}, \boldsymbol{\theta}|m)}{Q_{\mathrm{s}}(\mathbf{s}) Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}\, d\boldsymbol{\theta} = \mathcal{F}(Q_{\mathrm{s}}(\mathbf{s}), Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}), \mathcal{D}).$$

Maximize this lower bound with respect to Q leads to generalization of the EM algorithm.

Hidden Markov Models



Discrete hidden states, $\mathbf{s}_t$.

Observations $\mathbf{y}_t$.

How many hidden states?
What structure state-transition matrix?

Variational Bayesian HMMs (MacKay 1997; Beal PhD thesis 2003):

`demo:   vbhmm_demo`

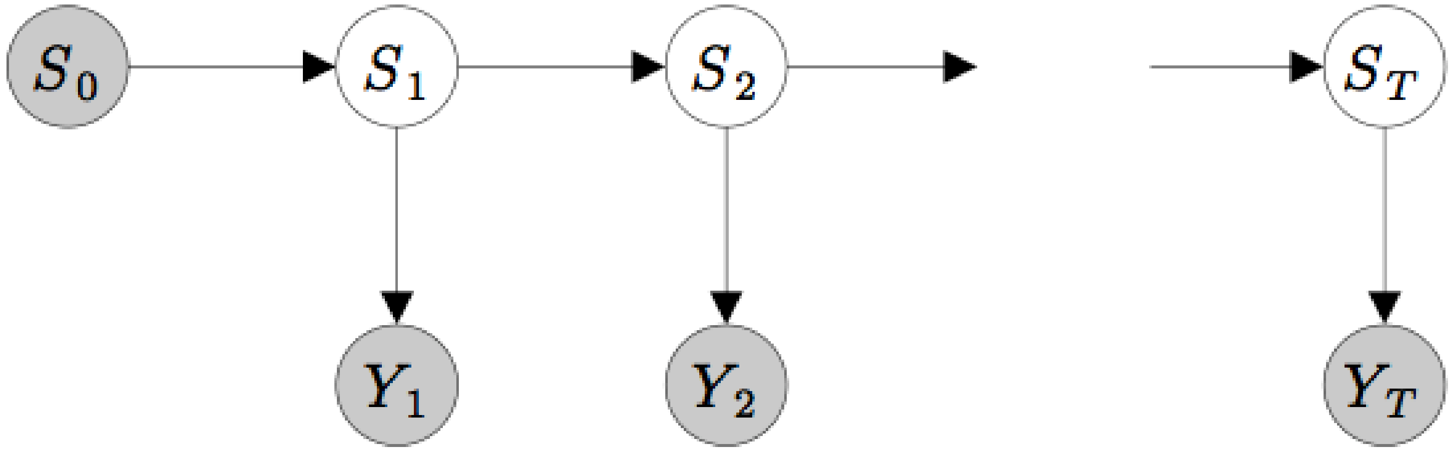
Summary of Part I

- Bayesian machine learning
- Marginal likelihoods and Occam's Razor
- Variational Bayesian lower bounds
- Application to learning the number of hidden states and structure of an HMM

Part II

The Infinite Hidden Markov Model

Hidden Markov Models



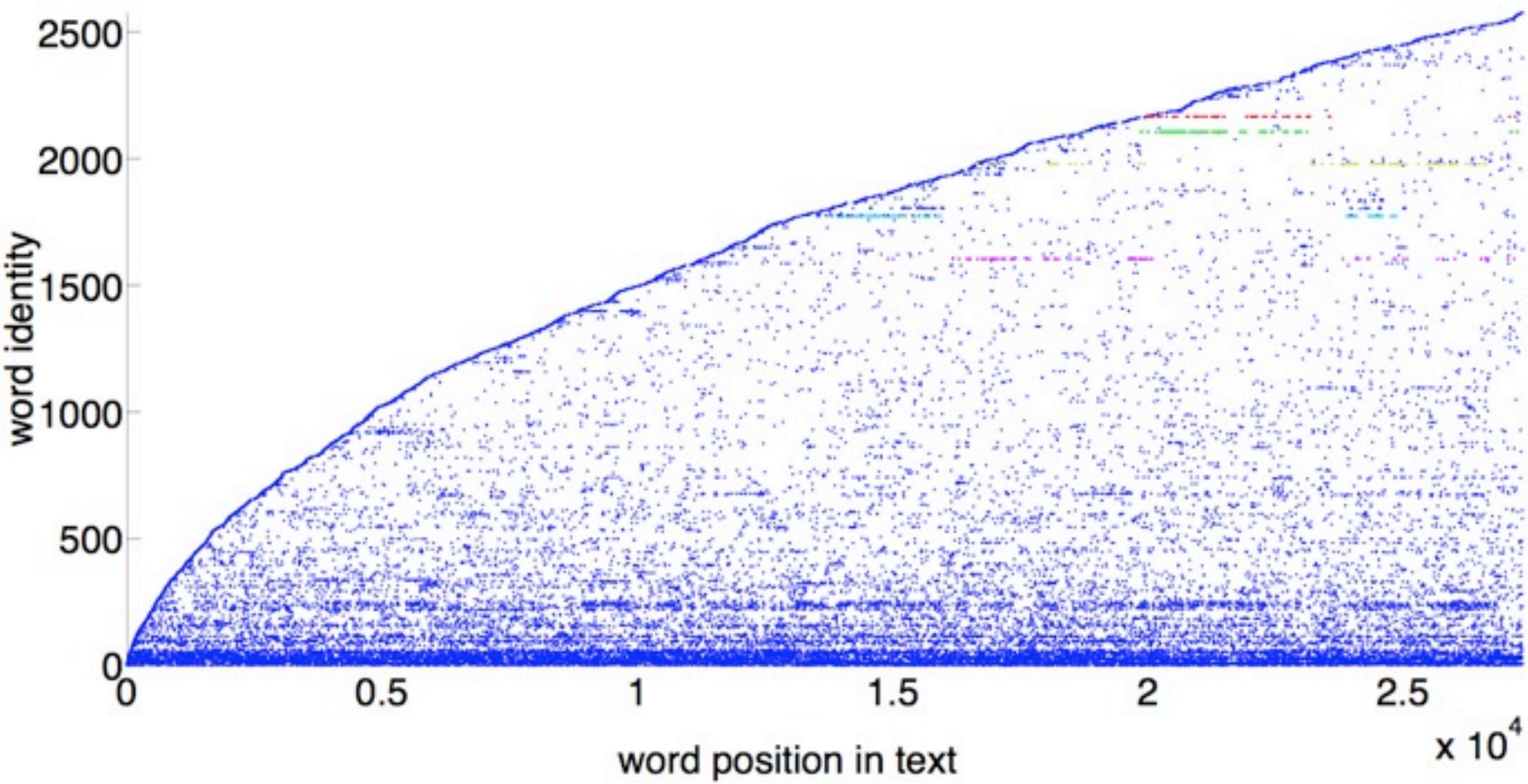
- Core: hidden K-state Markov chain
  - initial distribution $p(s_0 = 1) = 1$
  - transition probability $p(s_t = j | s_{t-1} = i) = \pi_{ij}$
- Peripheral: observation model $y_t \sim F(\phi_{s_t})$
- Parameters of the model are $K, \boldsymbol{\pi}, \boldsymbol{\phi}$

Choosing the number of hidden states

- How do we choose K, the number of hidden states, in an HMM?

- Can we define a model with an *unbounded* number of hidden states, and a suitable inference algorithm?

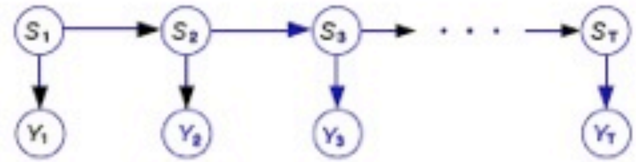
Alice in Wonderland



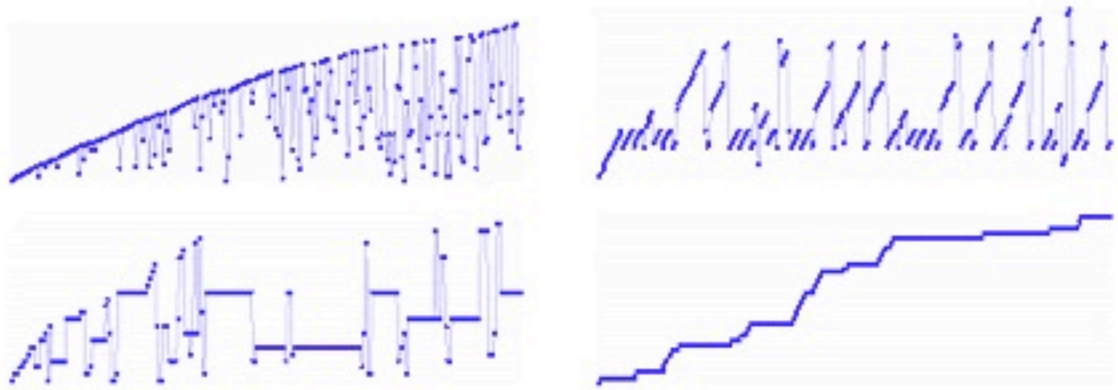
Infinite Hidden Markov models

Hidden Markov models (HMMs) can be thought of as time-dependent mixtures.

In an HMM with K states, the transition matrix has $K \times K$ elements.

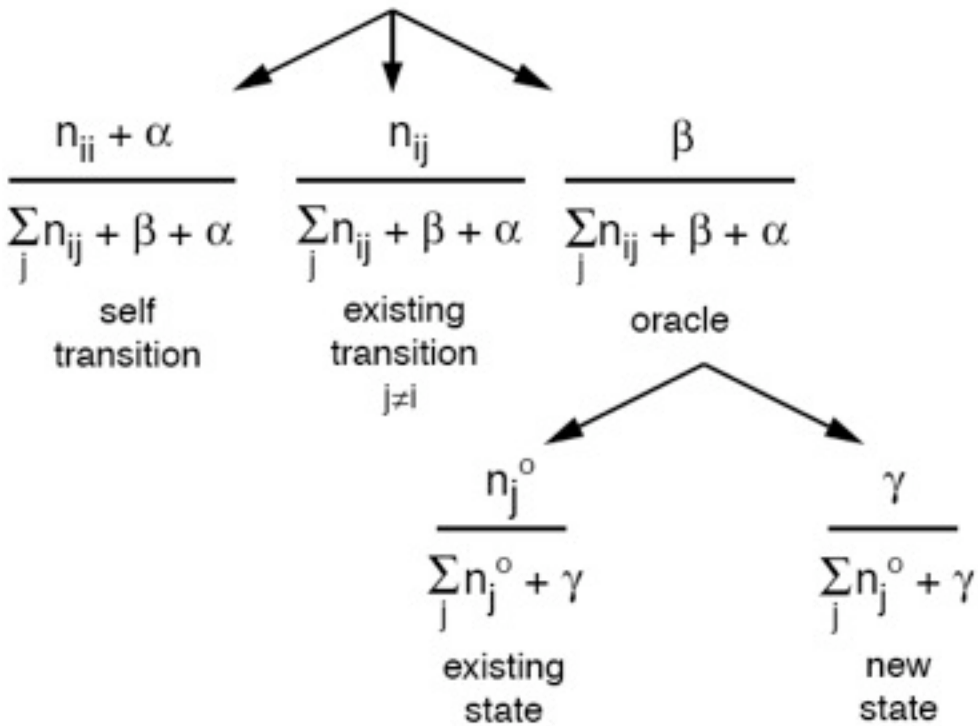


We let $K \to \infty$, this results in an iHMM.



- Introduced in (Beal, Ghahramani and Rasmussen, 2002).
- Teh, Jordan, Beal and Blei (2005) showed that iHMMs can be derived from hierarchical Dirichlet processes, and provided a more efficient Gibbs sampler.
- We have recently derived a much more efficient sampler based on Dynamic Programming (Van Gael, Saatci, Teh, and Ghahramani, 2008).

Hierarchical Urn Scheme for generating transitions in the iHMM (2002)

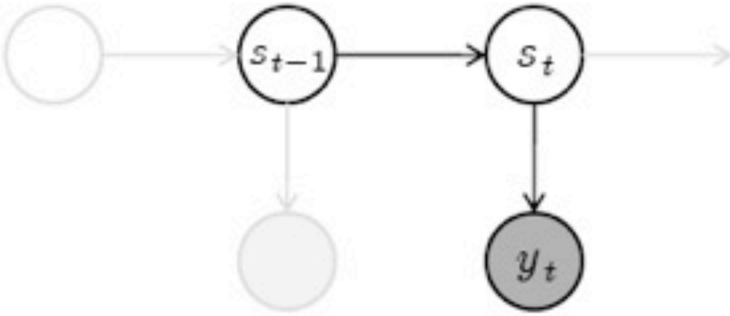


- n_{ij} is the number of previous transitions from *i* to *j*
- α, β, and γ are hyperparameters
- prob. of transition from *i* to *j* proportional to n_{ij}
- with prob. proportional to $\beta\gamma$ jump to a new state

Relating iHMMs to DPMs

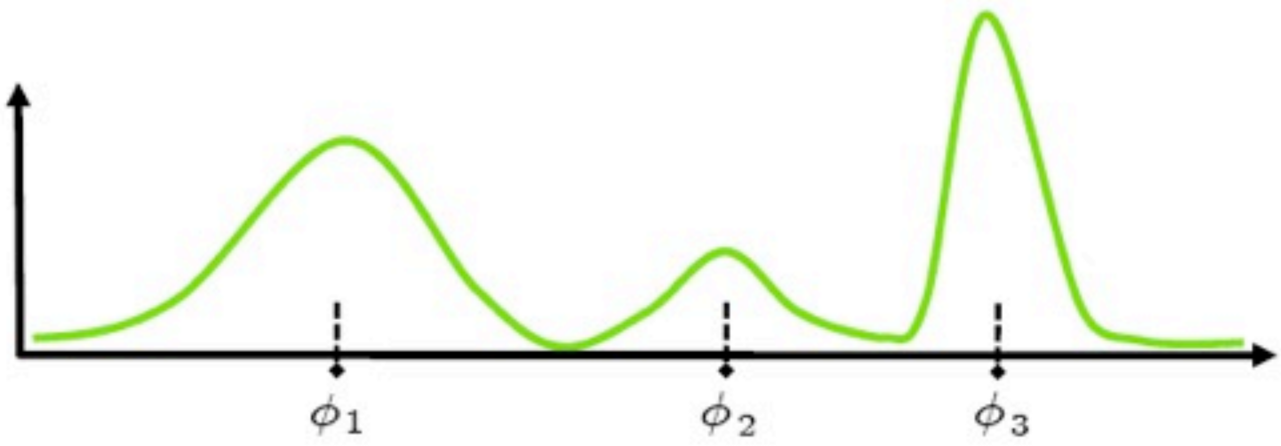
- The infinite Hidden Markov Model is closely related to Dirichlet Process Mixture (DPM) models
- This makes sense:
  - HMMs are time series generalisations of mixture models.
  - DPMs are a way of defining mixture models with countably infinitely many components.
  - iHMMs are HMMs with countably infinitely many states.

HMMs as sequential mixtures



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

What is conditional distribution of y_t?



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

Infinite Hidden Markov Models

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

Dirichlet Process

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

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

- A DP specifies both mixture weights and parameters

Infinite Hidden Markov Models

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

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

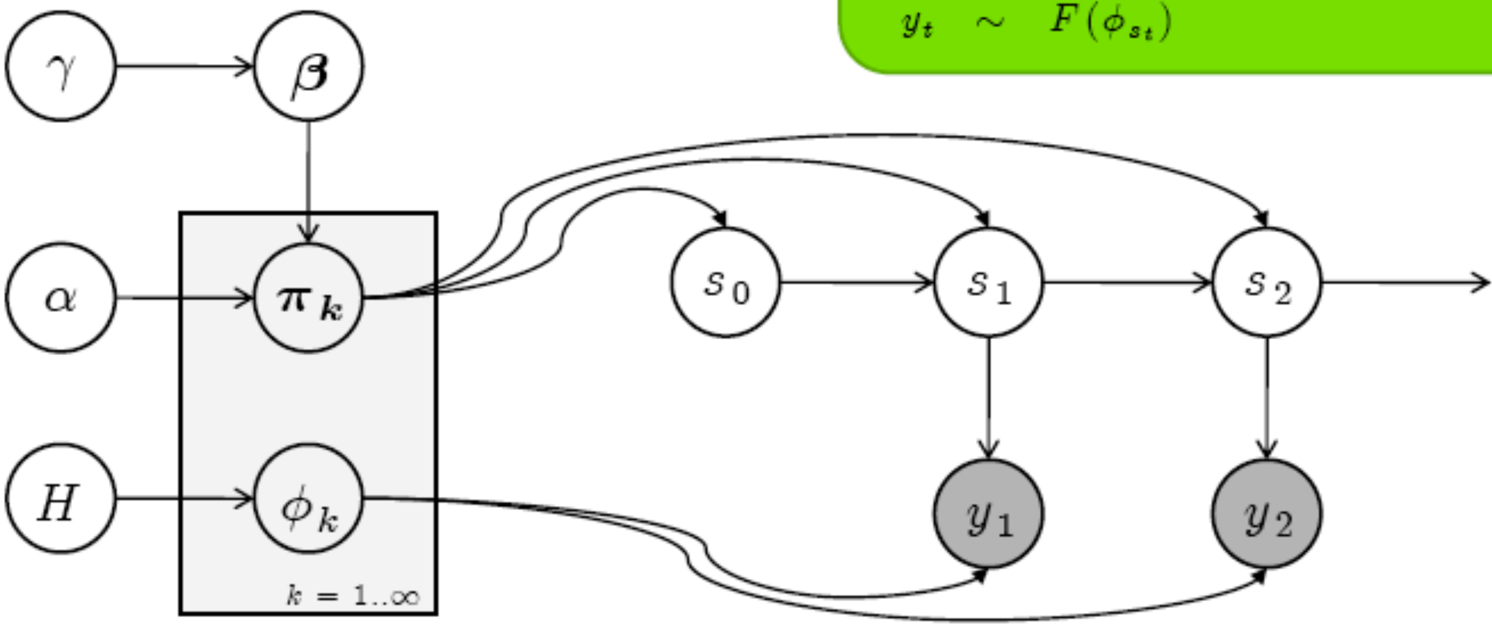
ϕ	ϕ_1	ϕ_2	ϕ_3	$\cdots$	ϕ_K
π	π_{11}	π_{12}	$\cdots$		
	π_{12}	$\ddots$			
	$\vdots$				
					π_{KK}

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

Infinite Hidden Markov Models

- Generative Model for iHMM

$$
\begin{aligned}
\boldsymbol{\beta} &\sim \mathrm{Stick}(\gamma), \\
\boldsymbol{\phi}_k &\sim H, \\
\boldsymbol{\pi}_k &\sim \mathrm{Dirichlet}(\alpha\boldsymbol{\beta}), \\
s_t &\sim \mathrm{Multinomial}(\boldsymbol{\pi}_{s_{t-1}}), \qquad (s_0 = 1) \\
y_t &\sim F(\phi_{s_t})
\end{aligned}
$$



Teh, Jordan, Beal and Blei (2005) derived iHMMs in terms of Hierarchical Dirichlet Processes.

Efficient inference in iHMMs?

Inference and Learning in HMMs and iHMMs

- HMM inference of hidden states $p(s_t|y_1 \ldots y_T, \theta)$:
  - forward backward = dynamic programming = belief propagation
- HMM parameter learning:
  - Baum Welch = expectation maximization (EM), or Gibbs sampling (Bayesian)
- iHMM inference and learning, $p(s_t, \theta \,|y_1 \ldots y_T)$:
  - Gibbs Sampling

- This is unfortunate: Gibbs can be very slow for time series!
- Can we use dynamic programming?

Dynamic Programming in HMMs
Forward Backtrack Sampling

1. Compute conditional probabilities
   1. Initialize

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

   2. For each t = 1 .. T

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

O(TK²)

2. Sample hidden states
   1. Sample for time T

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

   2. For each t = T-1 .. 1

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

O(TK)

Beam Sampling

- Can we use Forward-Backtrack for iHMM?

  ➔ No, $O(TK^2)$ with $K \rightarrow$ infinity is intractable
- A (bad?) idea:
  - Truncate transition matrix
  - Use dynamic programming to sample **s**
- This is only approximately correct.

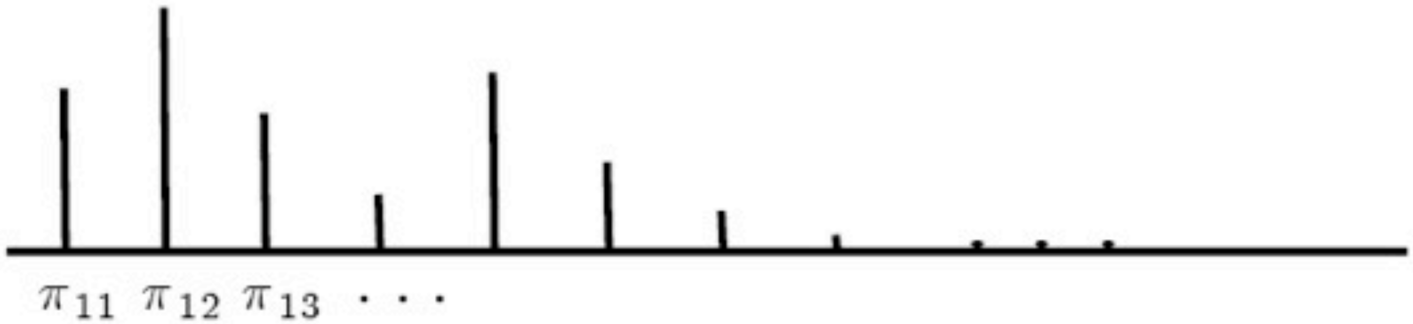
➔ Beam Sampling = Slice Sampling

+

Dynamic Programming

Beam Sampling

- Each G_k can be represented as

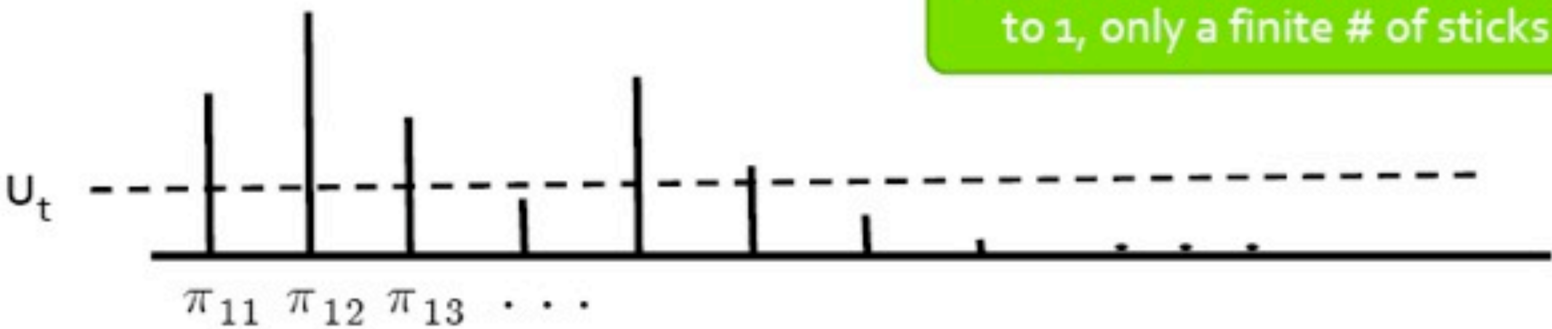


- Let us introduce an auxiliary variable

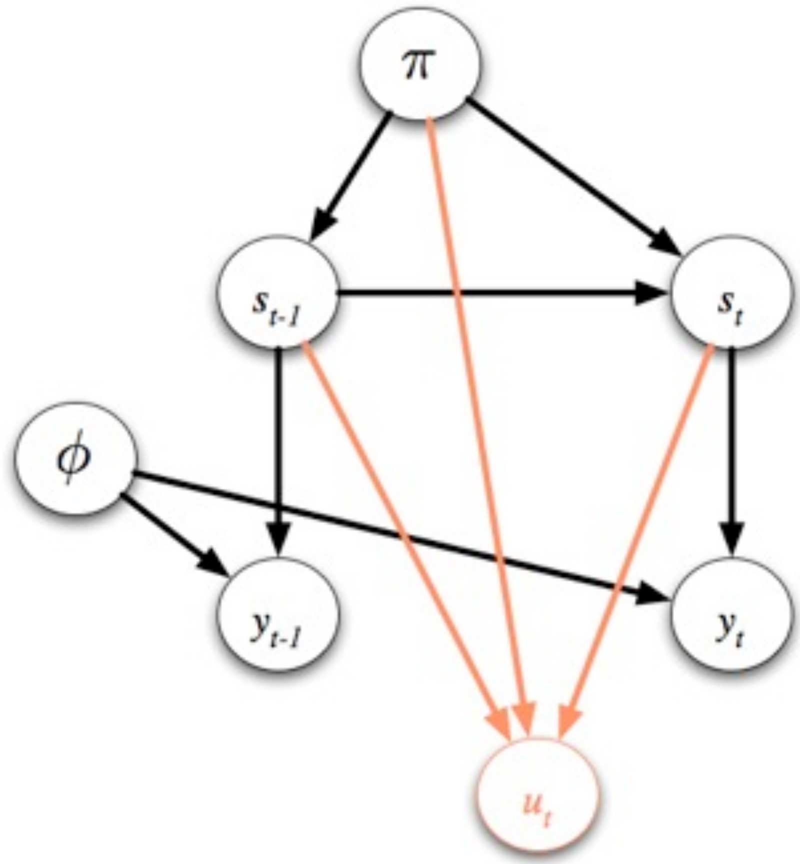
$$u_t \sim \text{Uniform}(0, \pi_{s_{t-1}, s_t})$$

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

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



Auxiliary variables



Note: adding *u* variables, does not change distribution over other vars.

Beam Sampling

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 .. T

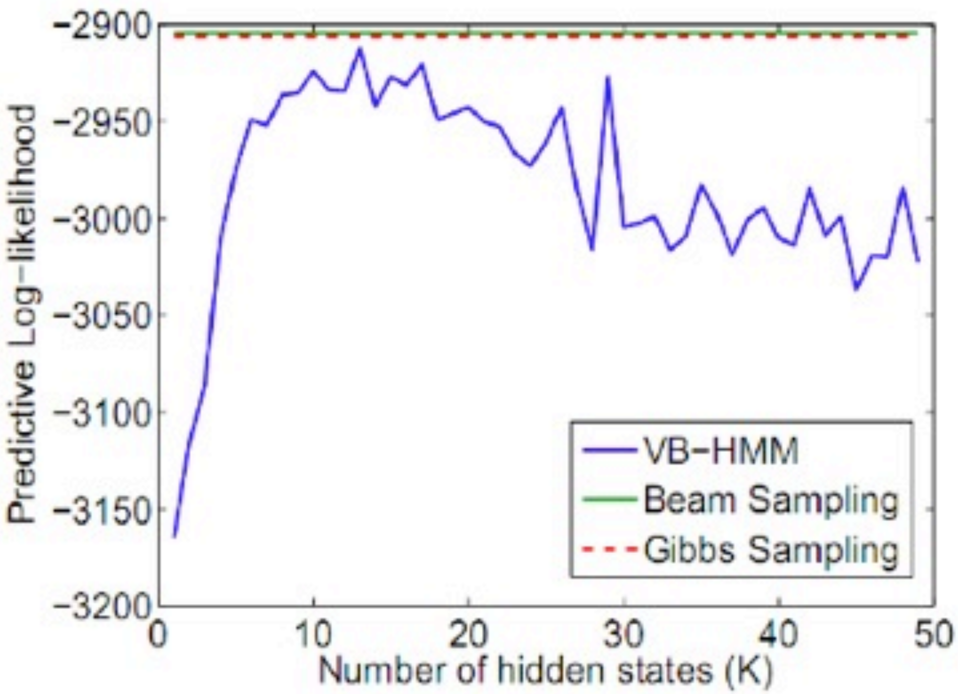
         $$p(s_t | y_{1:t}, u_{1:t}) \propto p(y_t | s_t) \sum_{s_{t-1}: u_t \leq \pi_{s_{t-1}, s_t}} p(s_{t-1} | y_{1:t-1}, u_{1:t-1})$$

      3. Sample T $p(s_T | y_{1:T})$
      4. Sample t = T-1 .. 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 $\boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{beta}, \gamma, \alpha$ | s

Experiment: 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, …, 4/K)
- Emission matrix: Dirichlet(0.3)

iHMM:

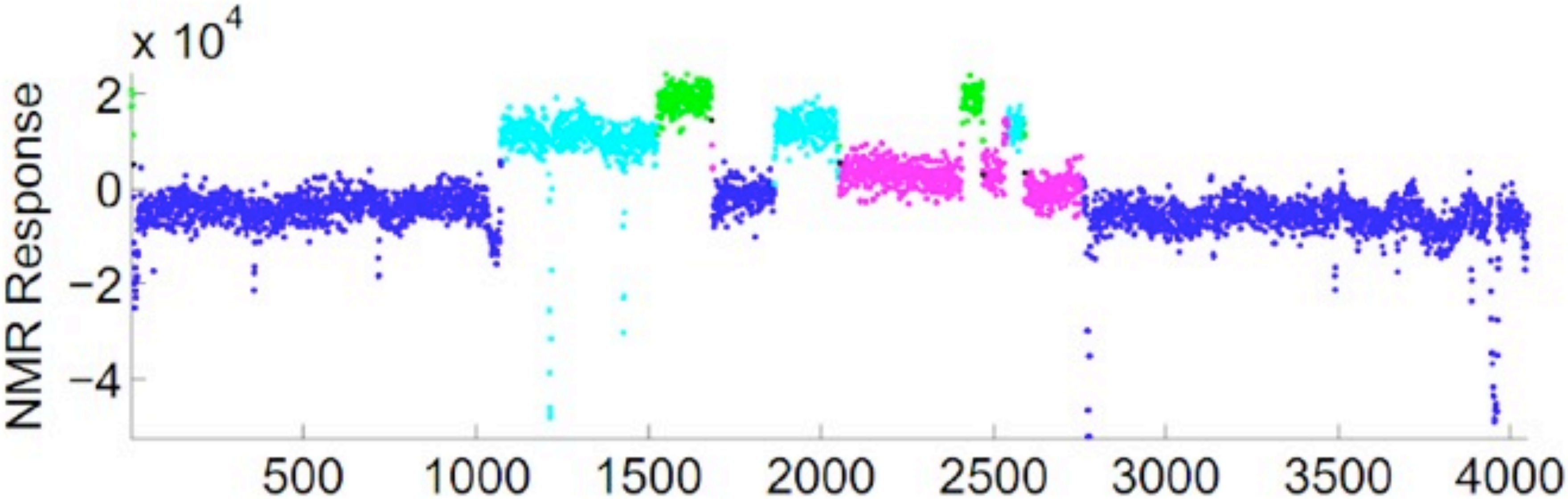
- $\alpha \sim$ Gamma(4,1)
- $\gamma \sim$ Gamma(1,1)
- H $\sim$ Dirichlet(0.3)

Experiment: 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:

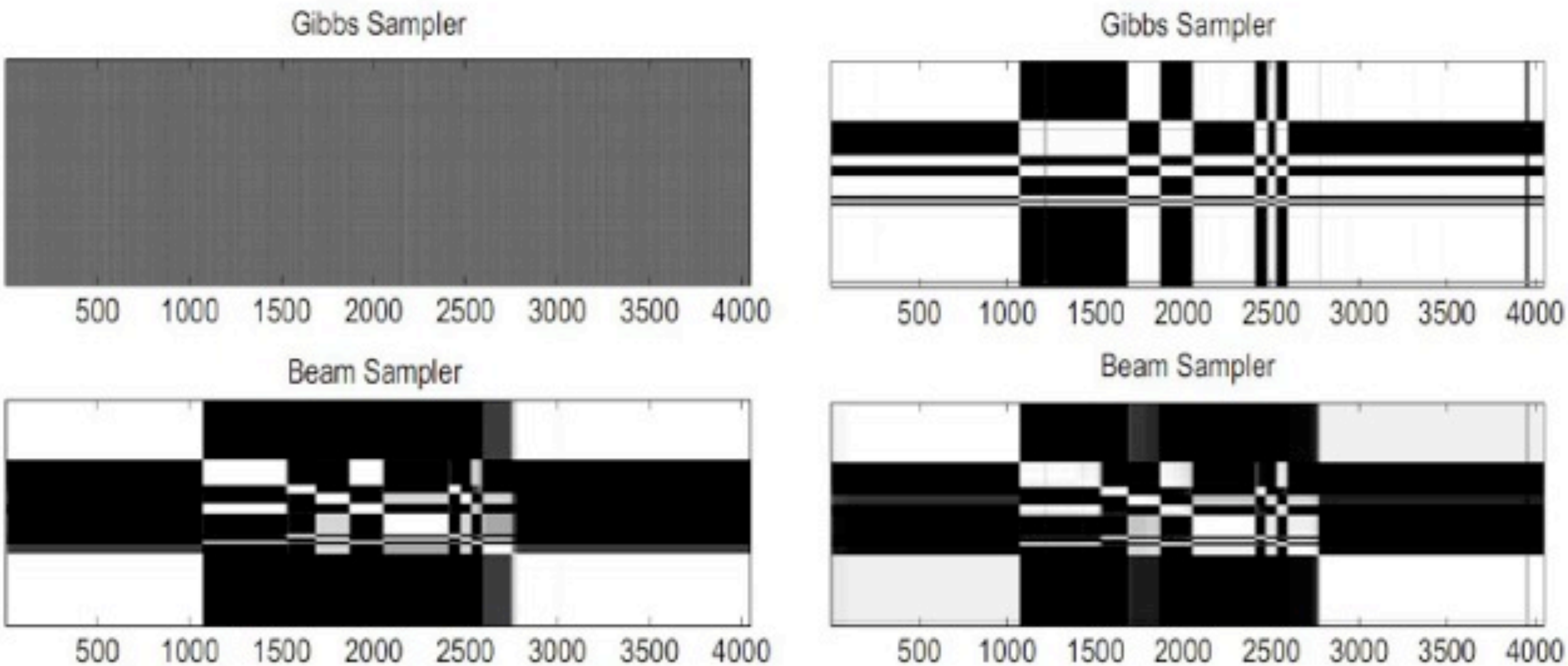


Experiment: Changepoint Detection

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



Parallel and Distributed Implementations of iHMMs

- Recent work on parallel (.NET) and distributed (Hadoop) implementations of beam-sampling for iHMMs (Bratieres, Van Gael, Vlachos and Ghahramani, 2010).
- Applied to unsupervised learning of part-of-speech tags from Newswire text (10 million word sequences).
- Promising results; open source code available for beam sampling iHMM: http://mloss.org/software/view/205/

Part III: iHMMs with clustered states

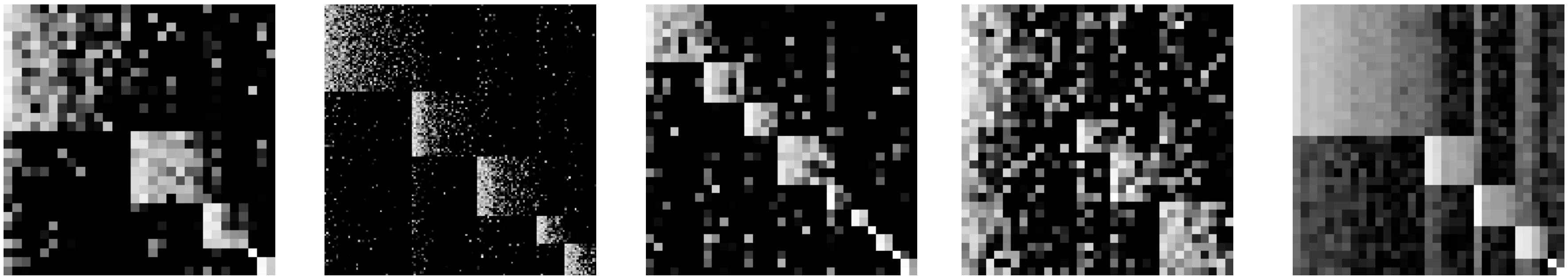
- We would like HMM models that can automatically group or cluster states.

- States within a group are more likely to transition to other states within the same group.

- This implies a ***block-diagonal*** transition matrix.

The Block-Diagonal iHMM

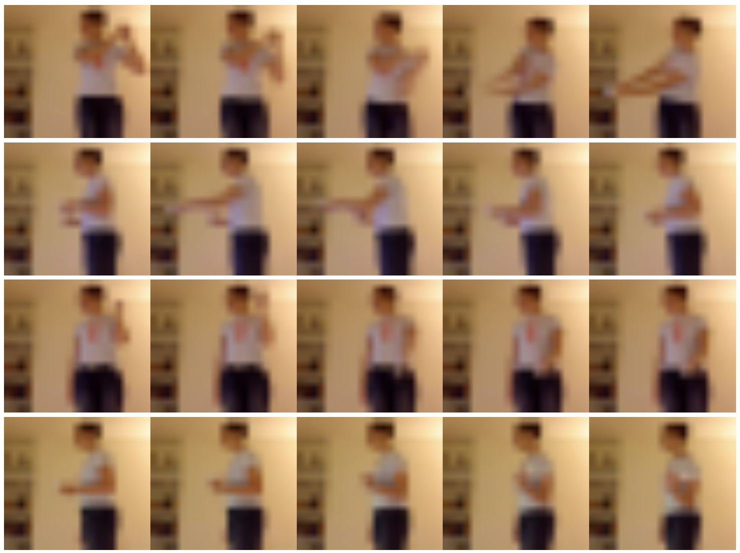
(Stepleton, Ghahramani, Gordon & Lee, 2009)



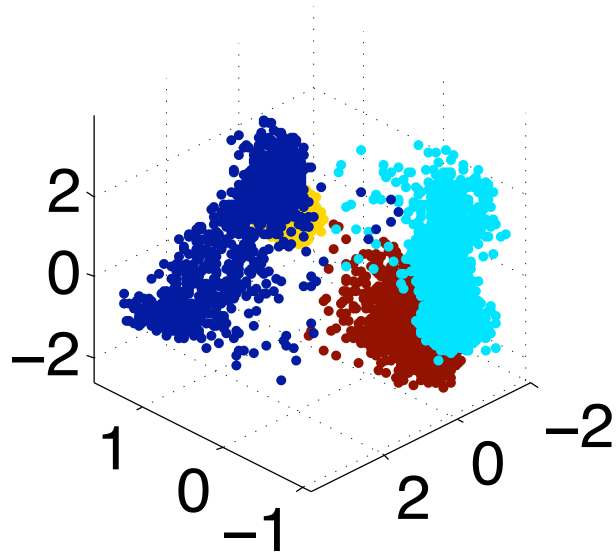
$\zeta = 1$, $\gamma = 10$	$\zeta = 1$, $\gamma = 50$	$\zeta = 5$, $\gamma = 10$	$\zeta = 1$, $\gamma = 10$	$\zeta = 10$, $\gamma = 10$
$\alpha_0 = 10$, $\xi = 10$	$\alpha_0 = 10$, $\xi = 10$	$\alpha_0 = 10$, $\xi = 10$	$\alpha_0 = 1$, $\xi = 1$	$\alpha_0 = 1000$, $\xi = 10$

BD-iHMM finding sub-behaviours in video gestures (Nintendo Wii)

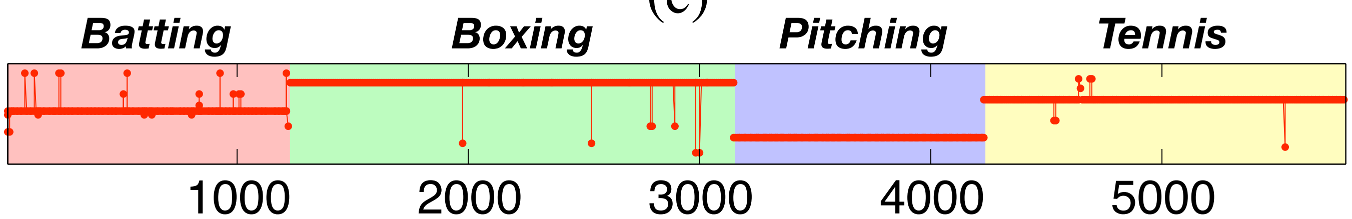
(a)



(b)



(c)



Part IV

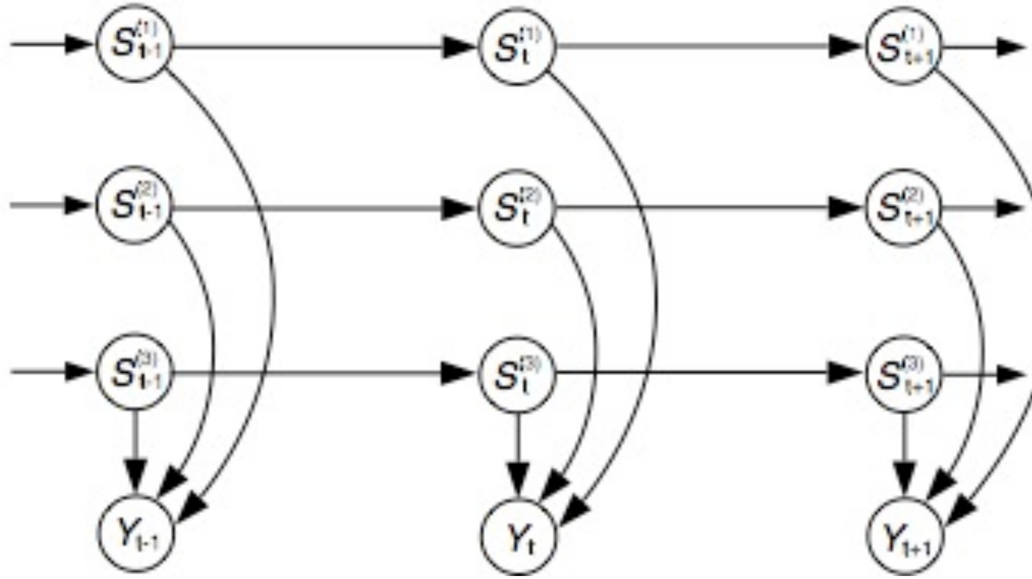
- Hidden Markov models represent the entire history of a sequence using a single state variable s_t

- This seems restrictive...

- It seems more natural to allow many hidden state variables, a “distributed representation” of state.

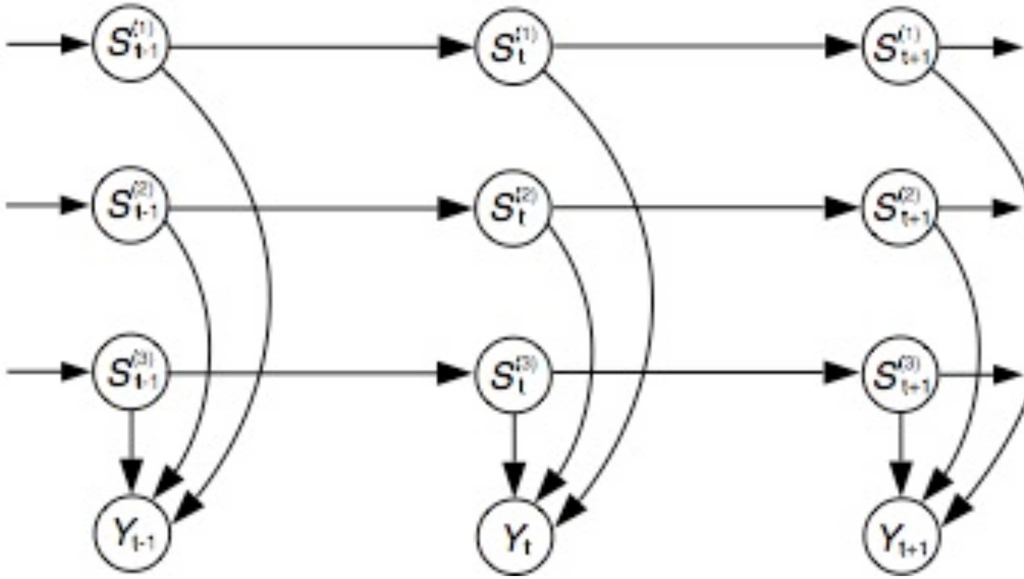
- …the *Factorial Hidden Markov Model*

Factorial HMMs



- Factorial HMMs (Ghahramani and Jordan, 1997)
- A kind of dynamic Bayesian network.
- Inference using variational methods or sampling.
- Have been used in a variety of applications (e.g. condition monitoring, biological sequences, speech recognition).

From factorial HMMs to infinite factorial HMMs?



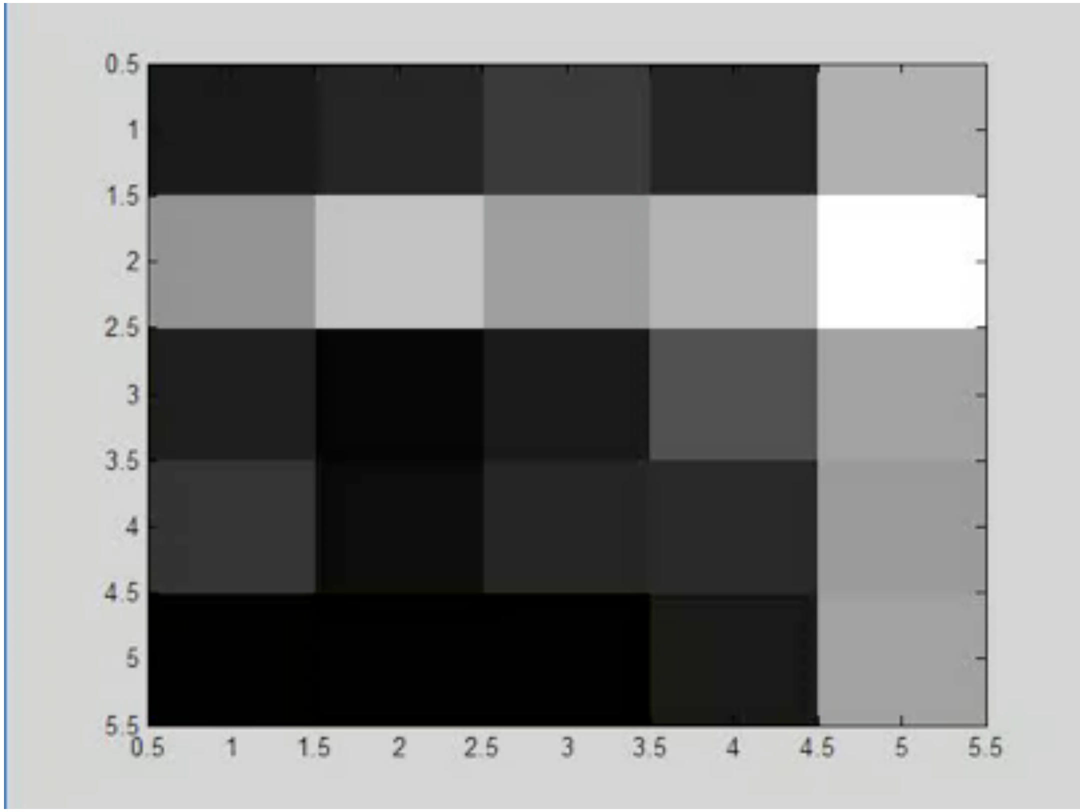
• A non-parametric version where the number of chains is unbounded?

• In infinite factorial HMM (ifHMM) each chain is binary (van Gael, Teh, and Ghahramani, 2008).

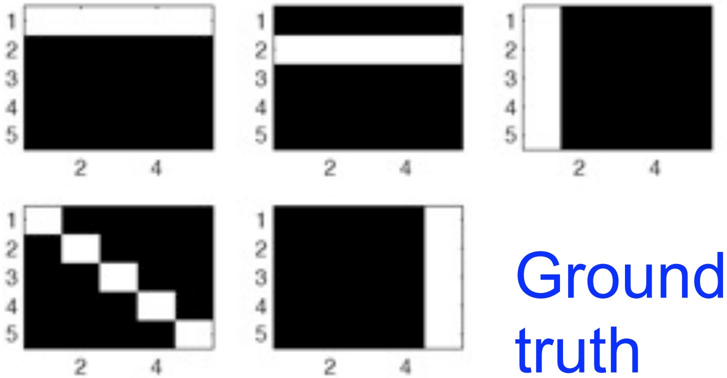
• Based on the Markov extension of the Indian Buffet Process (IBP).

Bars-in-time data

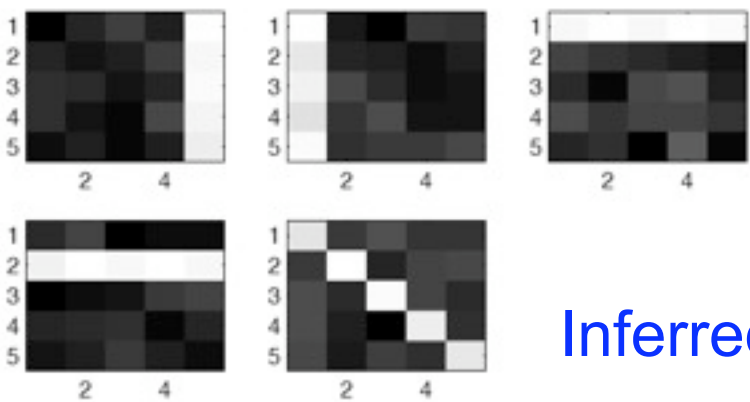
Bars-in-time data



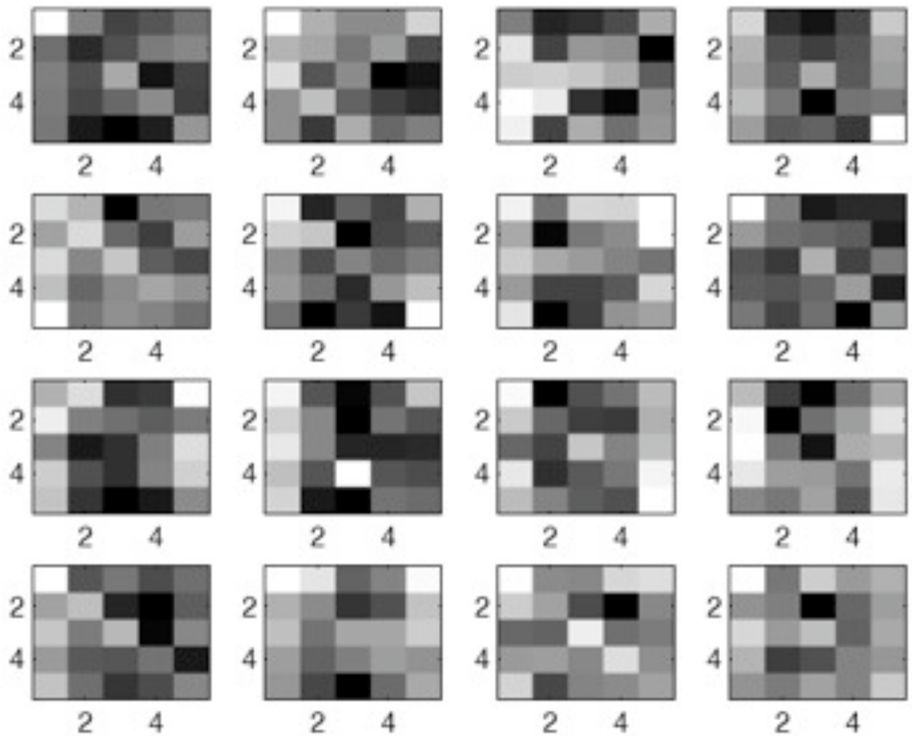
ifHMM Toy Experiment: Bars-in-time



Ground truth

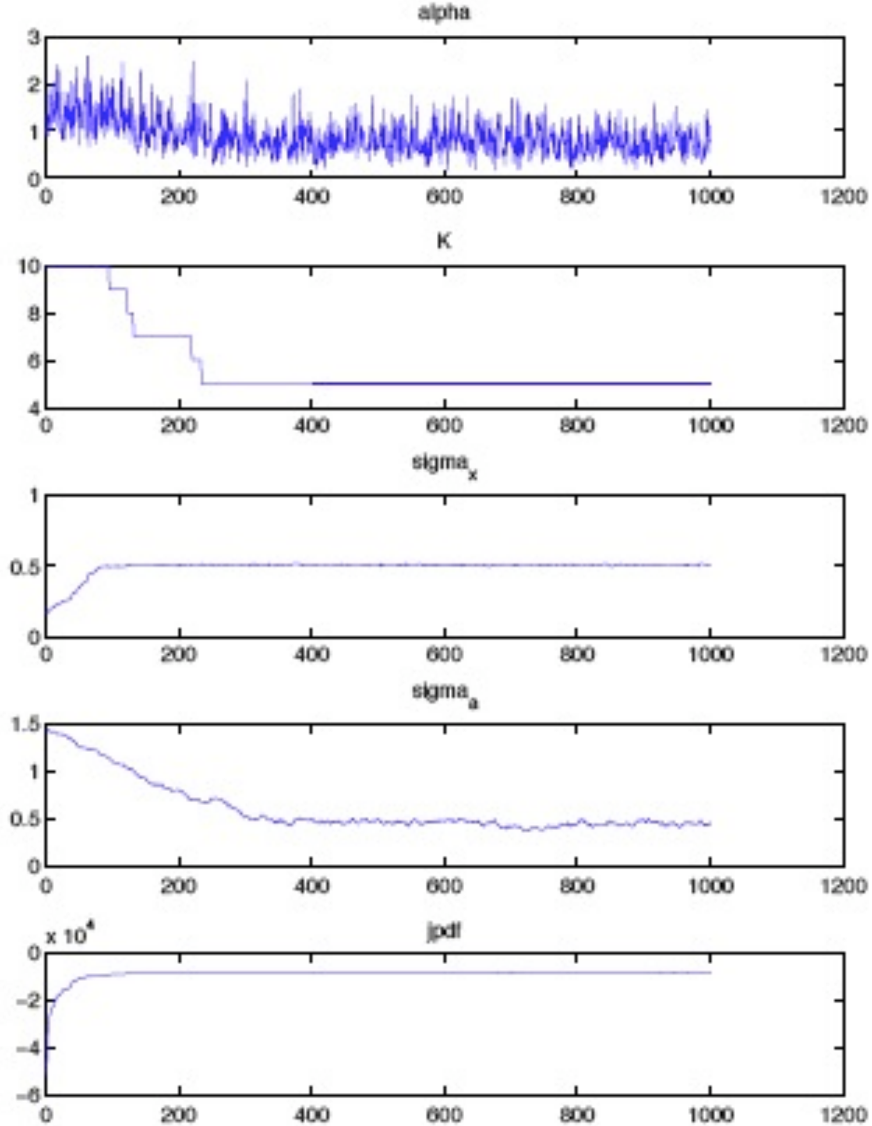


Inferred

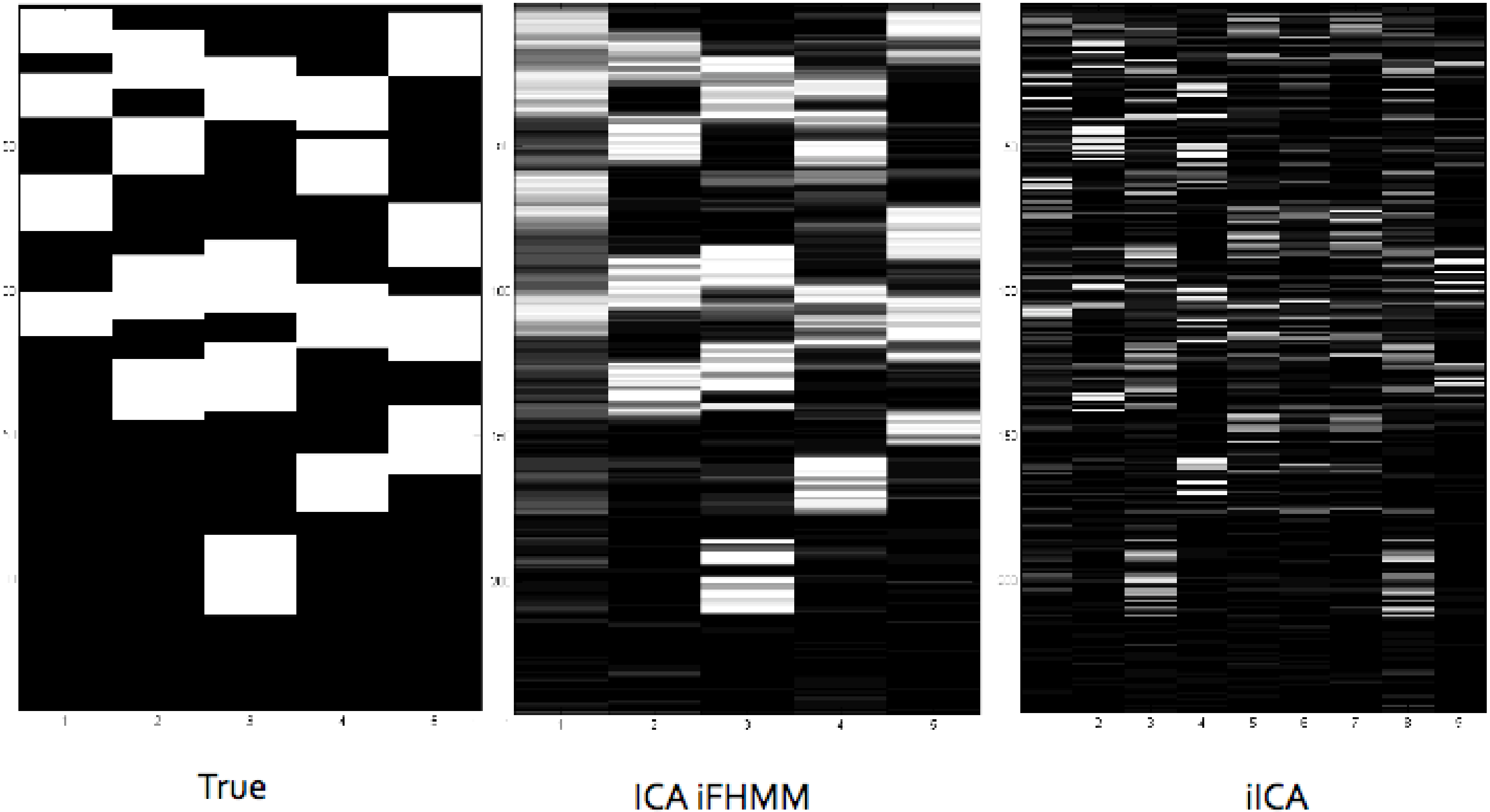


Data

ifHMM Experiment: Bars-in-time

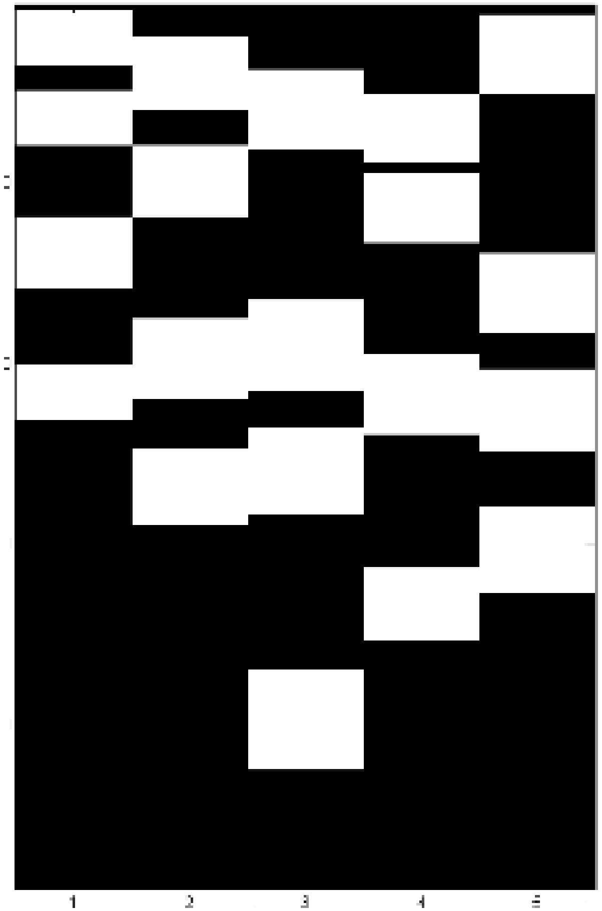


ICA iFHMM (more signals than sources)

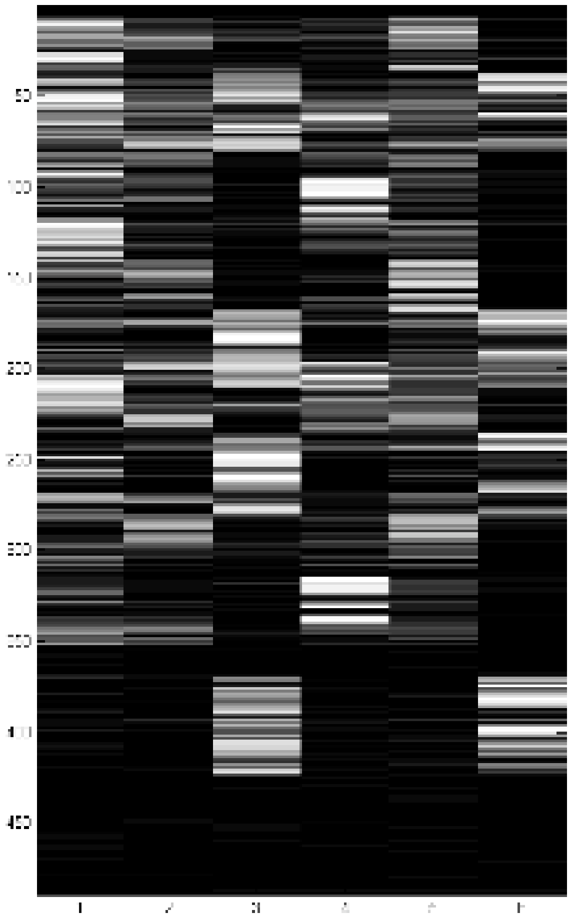


separating speech audio of multiple speakers in time

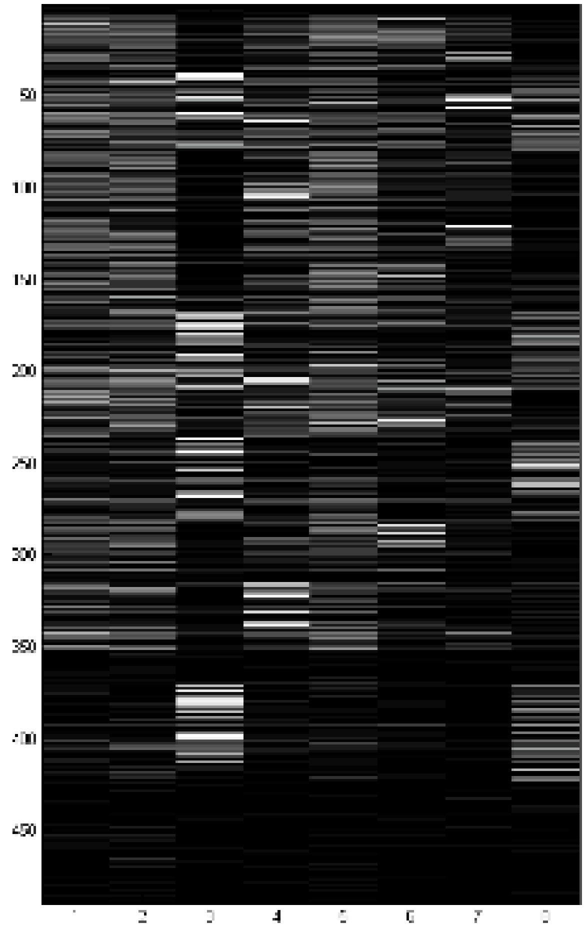
ICA iFHMM (fewer signals than sources)



True

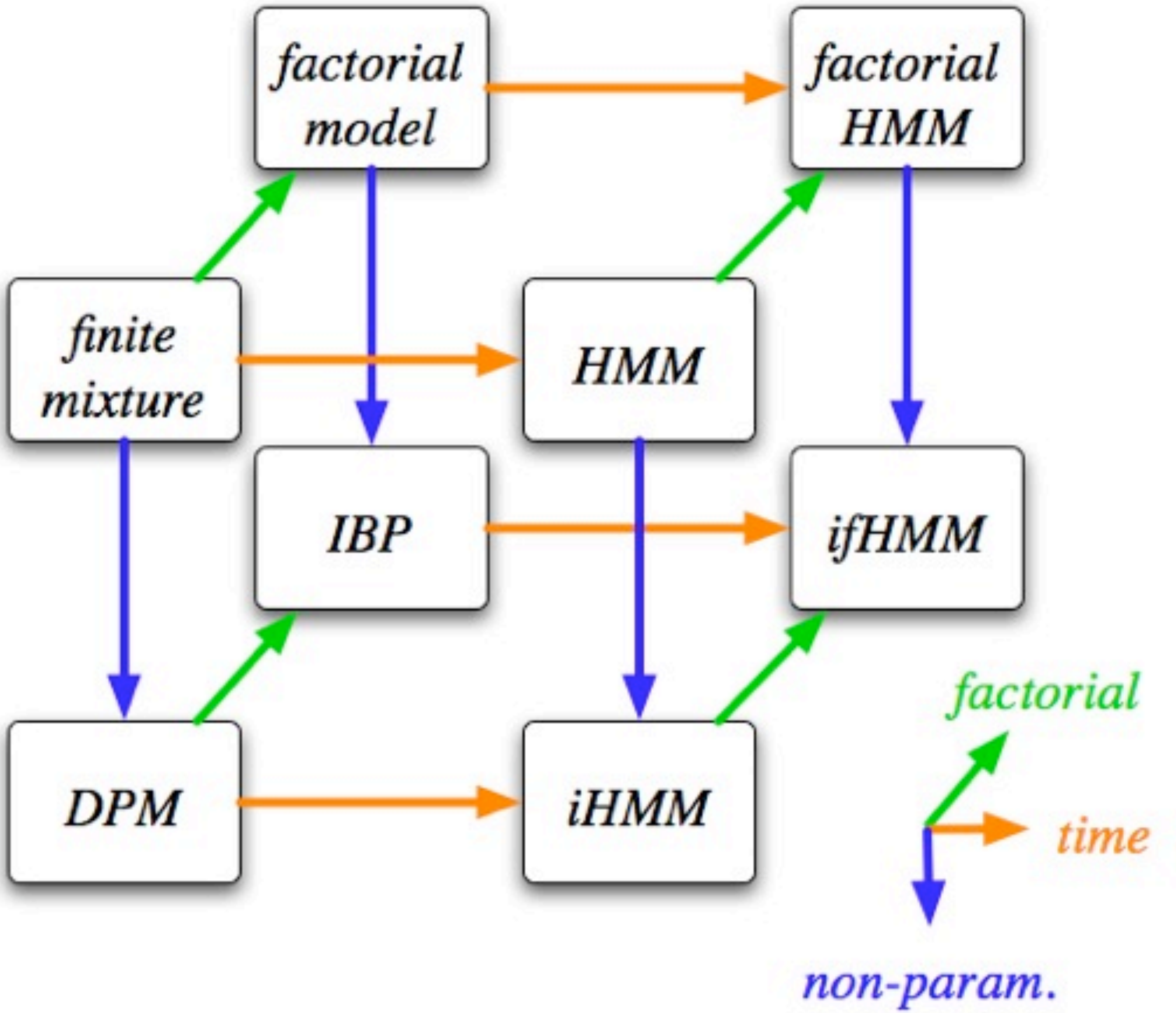


ICA iFHMM



iICA

The Big Picture



Summary

- Bayesian methods provide a flexible framework for modelling.
- HMMs can be learned using variational Bayesian methods. This should always be preferable to EM.
- iHMMs provide a non-parametric sequence model where the number of states is not bounded a priori.
- Beam sampling provides an efficient exact dynamic programming-based MCMC method for iHMMs.
- Block-Diagonal iHMMs learn to cluster states into sub-behaviours.
- ifHMMs extend iHMMs to have multiple state variables in parallel.
- **Future directions:** new models, fast algorithms, and other compelling applications.

References

Beal, M.J. (2003) Variational Algorithms for Approximate Bayesian Inference. PhD Thesis. University College London, UK.

Beal, M.J., Ghahramani, Z. and Rasmussen, C.E. (2002) The infinite hidden Markov model. Advances in Neural Information Processing Systems 14:577–585. Cambridge, MA: MIT Press.

Bratieres, S., van Gael, J., Vlachos, A., and Ghahramani, Z. (2010) Scaling the iHMM: Parallelization versus Hadoop. International Workshop on Scalable Machine Learning and Applications (SMLA-10).

Ghahramani, Z. and Jordan, M.I. (1997) Factorial Hidden Markov Models. Machine Learning, 29: 245–273.

Griffiths, T.L., and Ghahramani, Z. (2006) Infinite Latent Feature Models and the Indian Buffet Process. In Advances in Neural Information Processing Systems 18:475–482. Cambridge, MA: MIT Press.

MacKay, D.J.C. (1997) Ensemble learning for hidden Markov models. Technical Report.

Stepleton, T., Ghahramani, Z., Gordon, G., and Lee, T.-S. (2009) The Block Diagonal Infinite Hidden Markov Model. AISTATS 2009.

Teh, Y.W., Jordan, M.I., Beal, M.J. and Blei, D.M. (2006) Hierarchical Dirichlet processes. Journal of the American Statistical Association. 101(476):1566-1581.

van Gael, J., Teh, Y.-W., and Ghahramani, Z. (2009) The Infinite Factorial Hidden Markov Model. In Advances in Neural Information Processing Systems 21. Cambridge, MA: MIT Press.

van Gael, J., Saatci, Y., Teh, Y.-W., and Ghahramani, Z. (2008) Beam sampling for the infinite Hidden Markov Model. International Conference on Machine Learning (ICML 2008).

van Gael, J., Vlachos, A. and Ghahramani, Z. (2009) The Infinite HMM for Unsupervised POS Tagging. EMNLP 2009. Singapore.