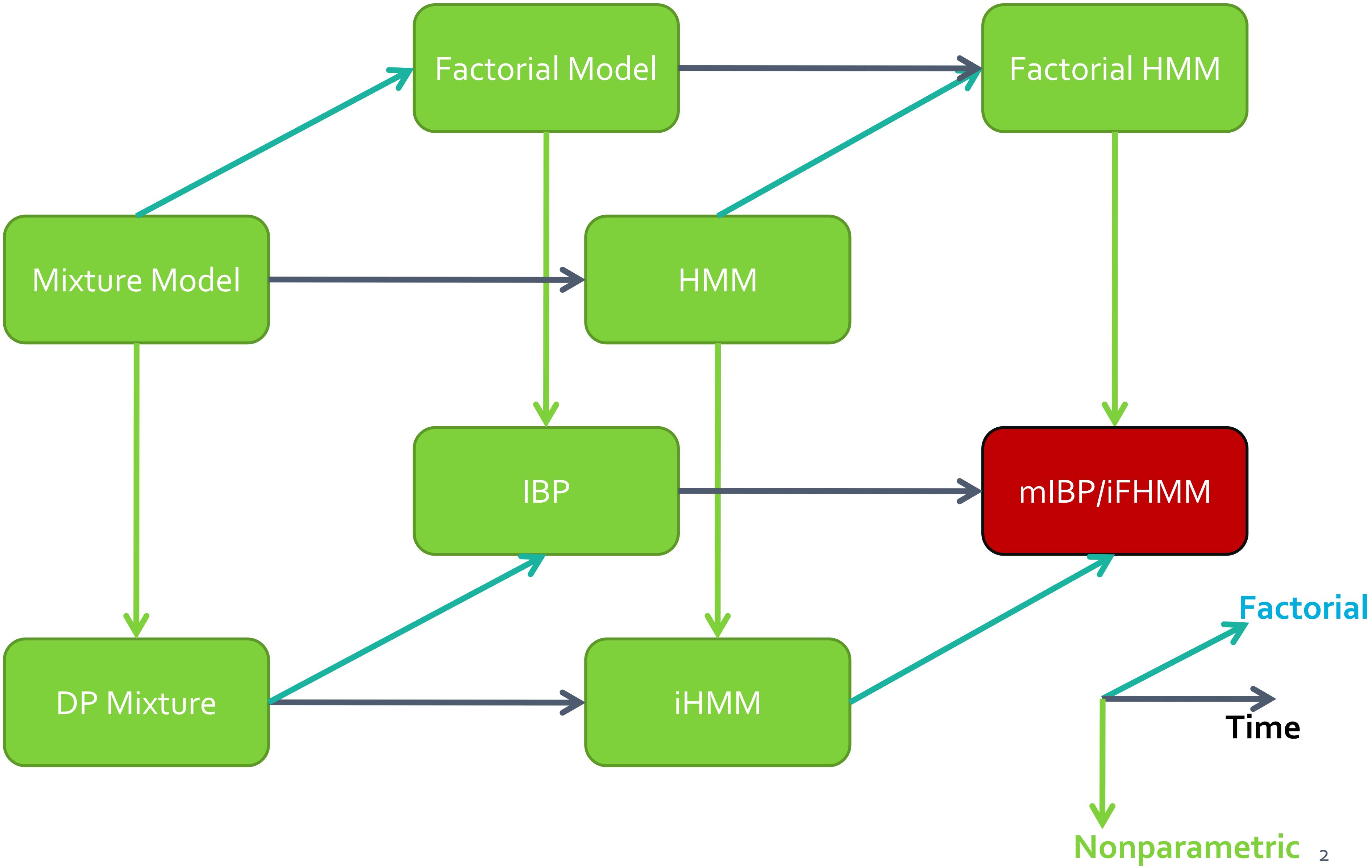
[†]University of Cambridge [*] Gatsby Unit, University College London
Jurgen Van Gael[†], Yee Whye Teh[*], Zoubin Ghahramani[†]

THE MARKOV INDIAN BUFFET PROCESS & INFINITE FACTORIAL HIDDEN MARKOV MODELS

The Landscape



Time Series

- Part-Of-Speech Tagging

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

- Dialogue Segmentation

Speaker	1	1	1	2	2	1	3	3	2	2

Factorial Time Series

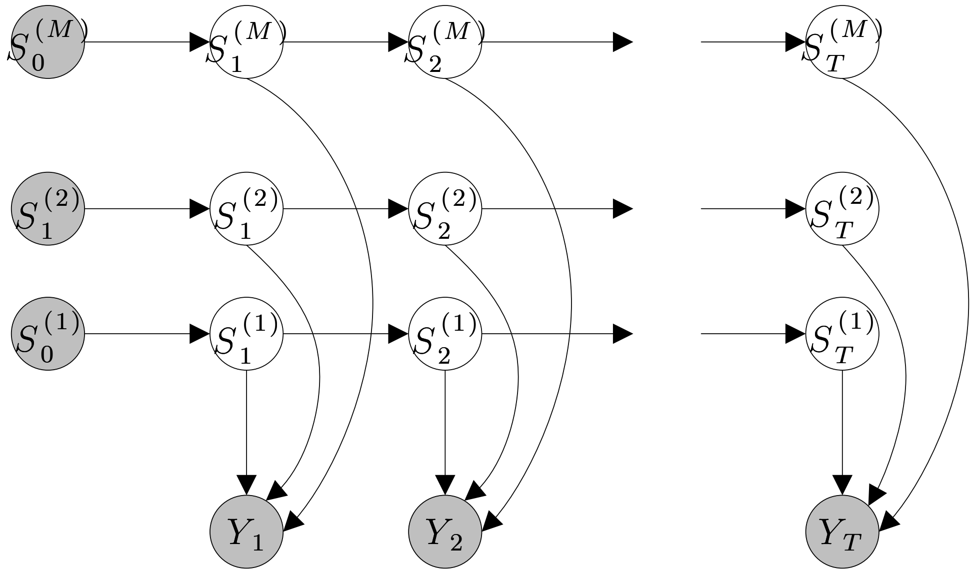
- Part-Of-Speech Tagging + Shallow Parsing

The	representative	put	the	red	chairs	on	the	table.
AT	NN	VBD	AT	ADJ	NNS	IN	AT	NN
B-NP	E-NP		B-NP	I-NP	E-NP		B-NP	E-NP

- Dialogue Segmentation

Speaker								
1	bla	bla		bla	bla			
2		bla				bla	bla	
3		bla	bla	bla		bla	bla	

Factorial Hidden Markov Model



- Core: M hidden binary Markov chains
  - initial distribution $p(s_0^{(m)} = 1) = 1 \quad \forall m \in \{1 \cdots M\}$
  - transition probability $p(s_t^{(m)} = j | s_{t-1}^{(m)} = i) = \pi_{ij} \quad \forall m \in \{1 \cdots M\}$
- Peripheral: observation model $y_t \sim F(\phi, s_t^{(1)}, \cdots, s_t^{(M)})$
- Parameters of the model are $M, \boldsymbol{\pi}, \boldsymbol{\phi}$

The Markov Indian Buffet Process

- We want to take the limit as $M \rightarrow \infty$
- Strategy:

  (Note: focus on binary chains)

  1. Describe a finite model with appropriate priors
  2. Marginalize out the parameters
  3. Take the infinite limit

1. The Finite Model

We define the dynamics of a single binary Markov chain m

Each Markov chain

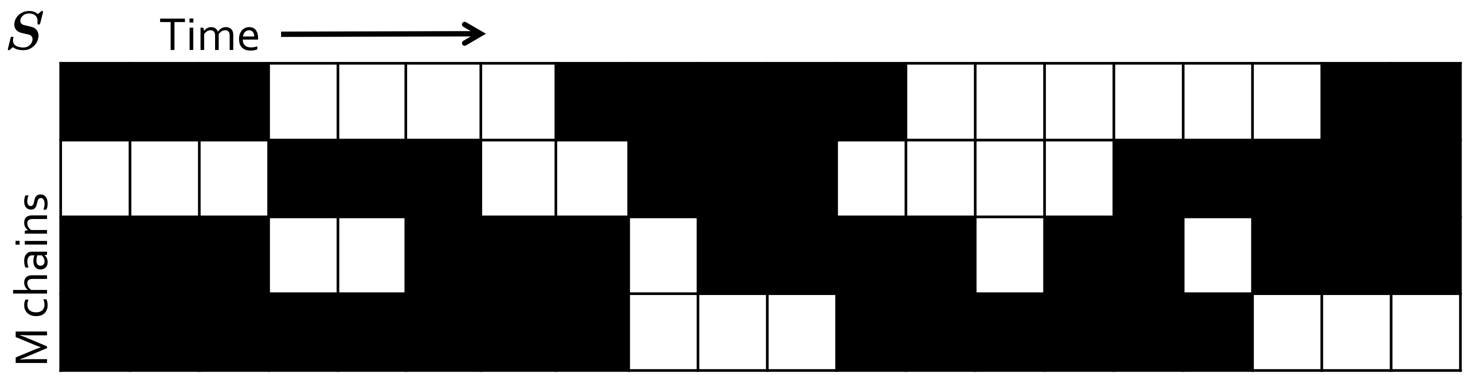
1. Starts in a dummy state $p(s_0 = 1) = 1$
2. Follows dynamics with transition matrix

$$W_m = \begin{pmatrix} 1 - a_m & a_m \\ 1 - b_m & b_m \end{pmatrix}$$

$$a_m \sim \mathrm{Beta}(\alpha / M, 1)$$

$$b_m \sim \mathrm{Beta}(\gamma, \delta)$$

2. Marginalizing



We marginalize out W_m and compute the probability of the latent state sequence

$$p(\boldsymbol{S}|\alpha,\gamma,\delta) = \prod_{m=1}^{M} \frac{\frac{\alpha}{M}\Gamma(\frac{\alpha}{M}+c_m^{01})\Gamma(c_m^{00}+1)\Gamma(\gamma+\delta)\Gamma(\delta+c_m^{10})\Gamma(\gamma+c_m^{11})}{\Gamma(\frac{\alpha}{M}+c_m^{00}+c_m^{01}+1)\Gamma(\gamma)\Gamma(\delta)\Gamma(\gamma+\delta+c_m^{10}+c_m^{11})}$$

where c_m^{01} is the transition count from state 0 to 1, etc.

3. The Infinite Limit

A bit of algebra which essentially follows the argument in [Griffiths & Ghahramani, 2005].

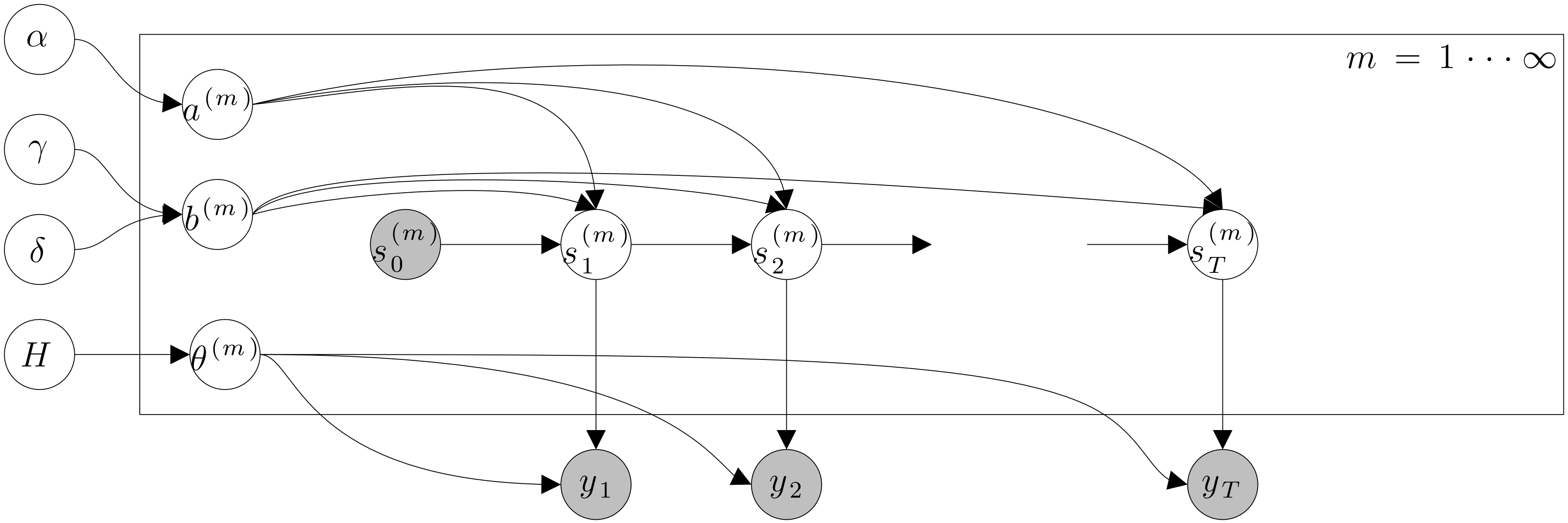
$$\lim_{M\to\infty} p([\boldsymbol{S}]) = \frac{\alpha^{M_+}}{\prod_{h=0}^{2^T-1} M_h!} \exp\{-\alpha H_T\} \prod_{m=1}^{M_+} \frac{(c_m^{01}-1)!c_m^{00}!\Gamma(\gamma+\delta)\Gamma(\delta+c_m^{10})\Gamma(\gamma+c_m^{11})}{(c_m^{00}+c_m^{01})!\Gamma(\gamma)\Gamma(\delta)\Gamma(\gamma+\delta+c_m^{10}+c_m^{11})}$$

Some technical details of the infinite limit:

- We are interested in the distribution over equivalence classes of state sequences

  i.e. *left-ordered form*
- Distribution over S is Markov exchangeable

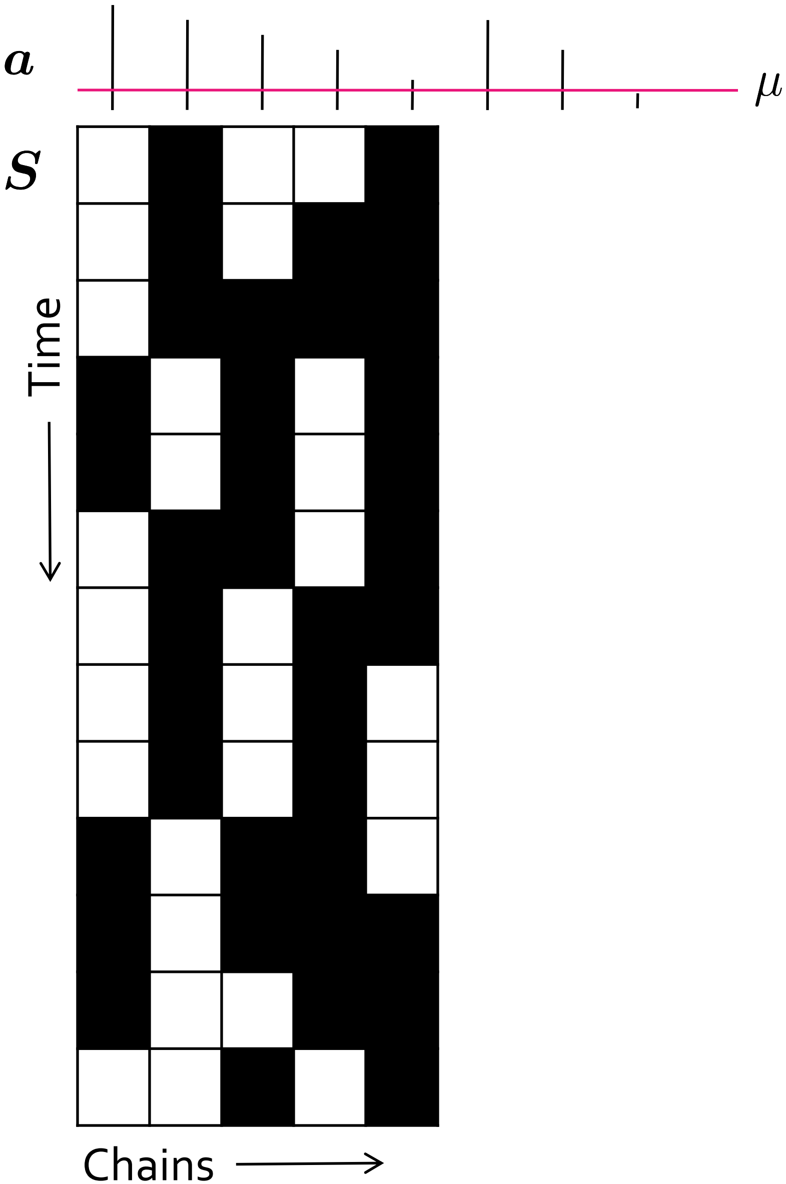
The Infinite Factorial HMM



Start with the state sequence of a Markov Indian Buffet Process and

- Add an observation model: $y_t \sim F(\boldsymbol{\theta}, \boldsymbol{s}_t)$
- Add base distribution H: $\theta \sim H$

Inference



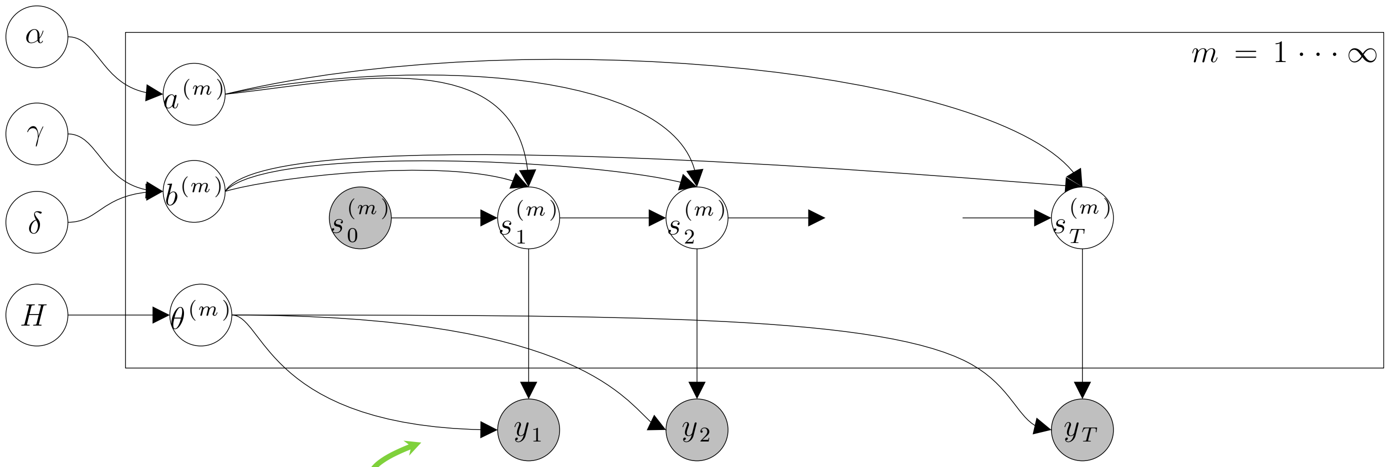
1. Start with current sample of latent states
2. Sample slice variable

$$\mu \sim \text{Uniform}(0, \min_{m:\exists t, s_{tm}=1} a_m)$$

> Key observation: the sticks decrease in size: hence only a finite number of new features need to be considered.

3. Run forward-filtering backward-sampling
   (On one or multiple chains)
4. Sample other model parameters.

Linear Gaussian iFHMM



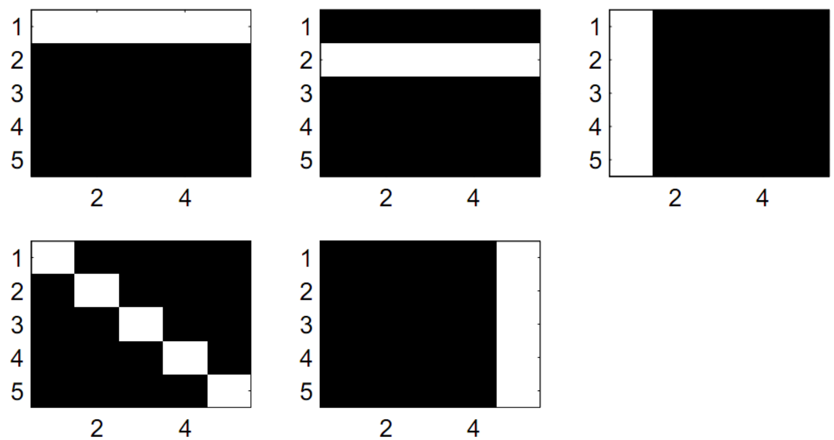
$$H : A \sim \mathrm{Normal}(\cdot, \cdot)$$

$$F : \boldsymbol{y}_t = \boldsymbol{s}_t A + \epsilon \quad \text{with } \epsilon \sim \mathrm{Normal}(\cdot, \cdot)$$

The linear Gaussian iFHMM takes the basic iFHMM construction and adds:

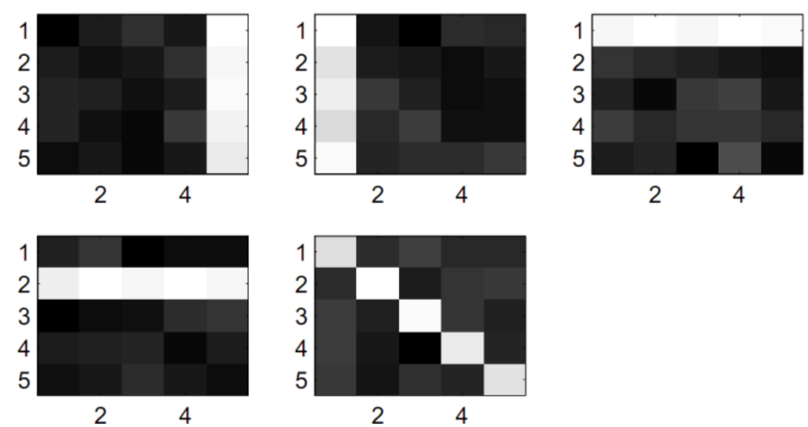
- H: Vector valued features with a Gaussian distributed base distribution
- F: A linear likelihood model with Gaussian distributed noise model

Linear Gaussian iFHMM

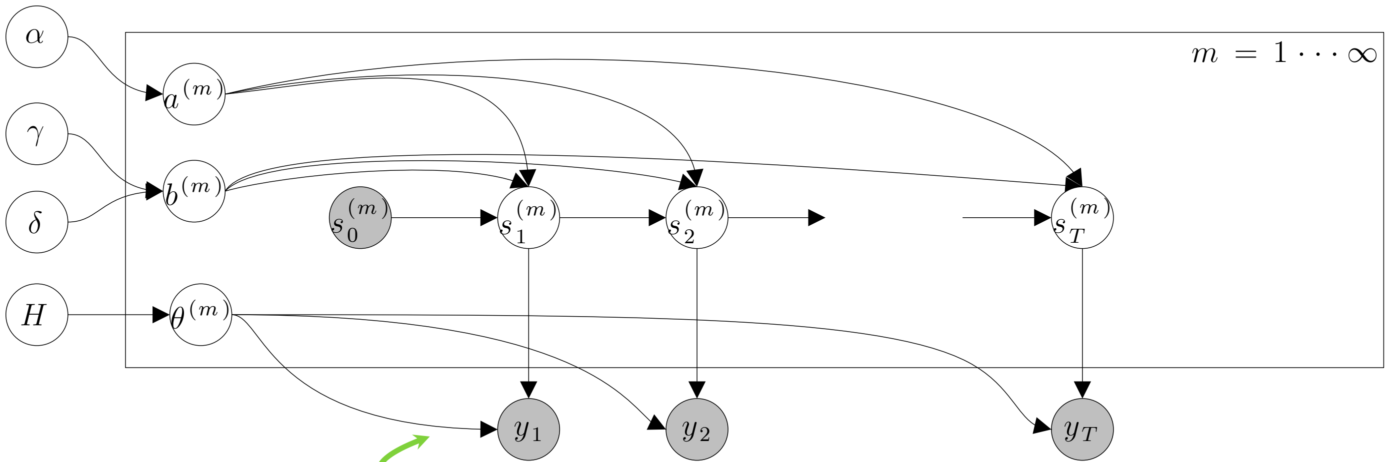


Bars-in-time experiment: generate 25 dimensional datapoints from 5 bars. Use a random 5 dimensional binary Markov chain to turn features on or off.

After inference …



ICA iFHMM



$$H := \begin{cases} X \sim \text{Laplace}(0, 1) \\ W \sim \text{Normal}(\cdot, \cdot) \end{cases}$$

$$F : \boldsymbol{Y} = (\boldsymbol{S} \odot X)W + \epsilon \quad \text{with } \epsilon \sim \text{Normal}(\cdot, \cdot)$$

The Independent Component Analysis iFHMM takes the basic iFHMM construction and adds:

- H: A signal matrix X with IID heavy tailed Laplace prior distribution
- H: A mixing matrix W with a matrix normal distributed prior distribution
- F: An ICA likelihood model with Gaussian noise

ICA iFHMM

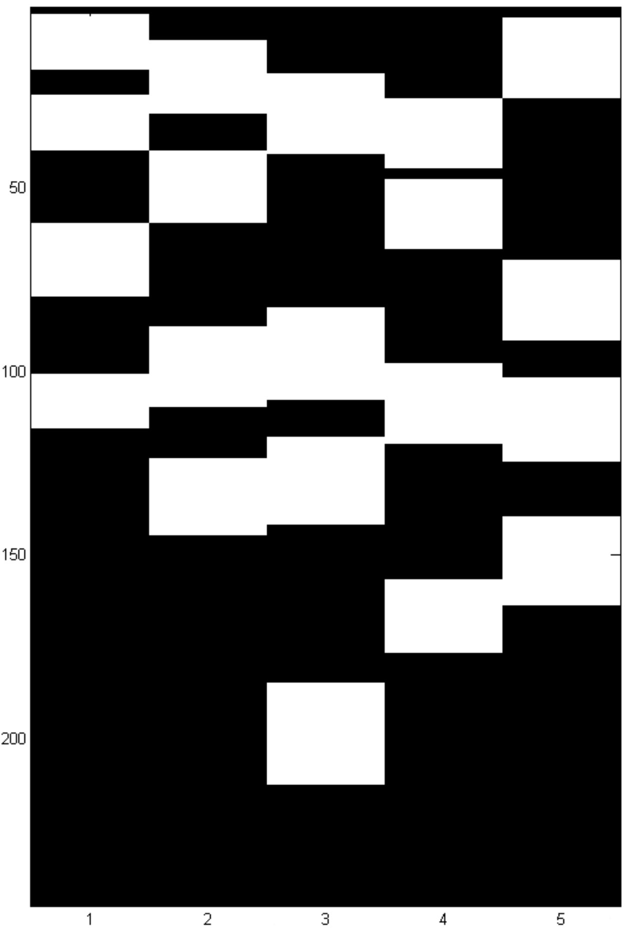
We performed an experiment on blind speaker separation.

- Sentences from Pascal Speech Separation Challenge

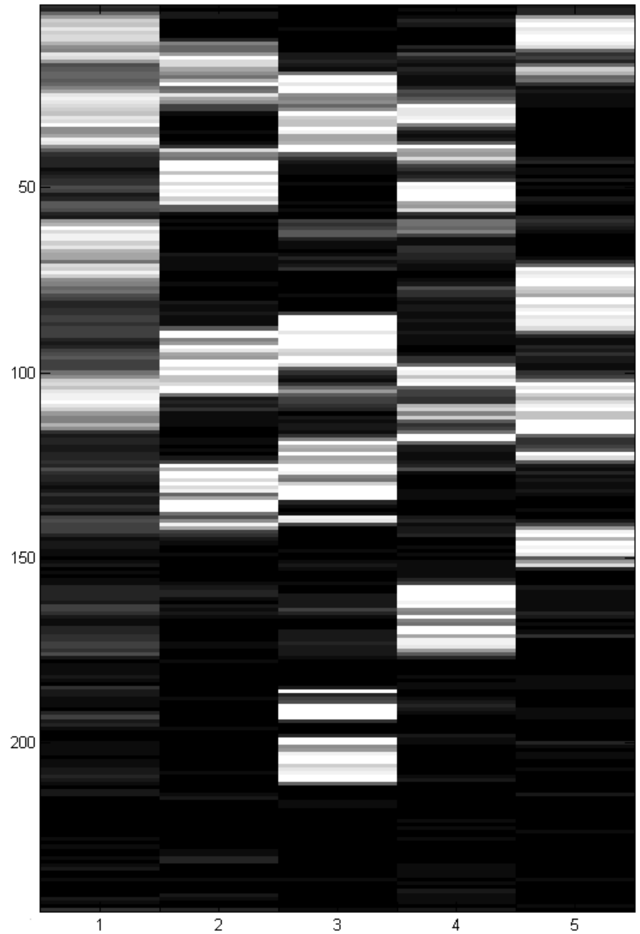
  http://www.dcs.shef.ac.uk/ martin/SpeechSeparationChallenge.htm

- 5 random speakers
- 4 sentences with random intervals
- Experiment 1
  - 10 random microphones
  - Subsampled to 256 datapoints
- Experiment 2
  - 3 random microphones
  - Subsampled to 489 datapoints
- We sample 20 mixing matrices

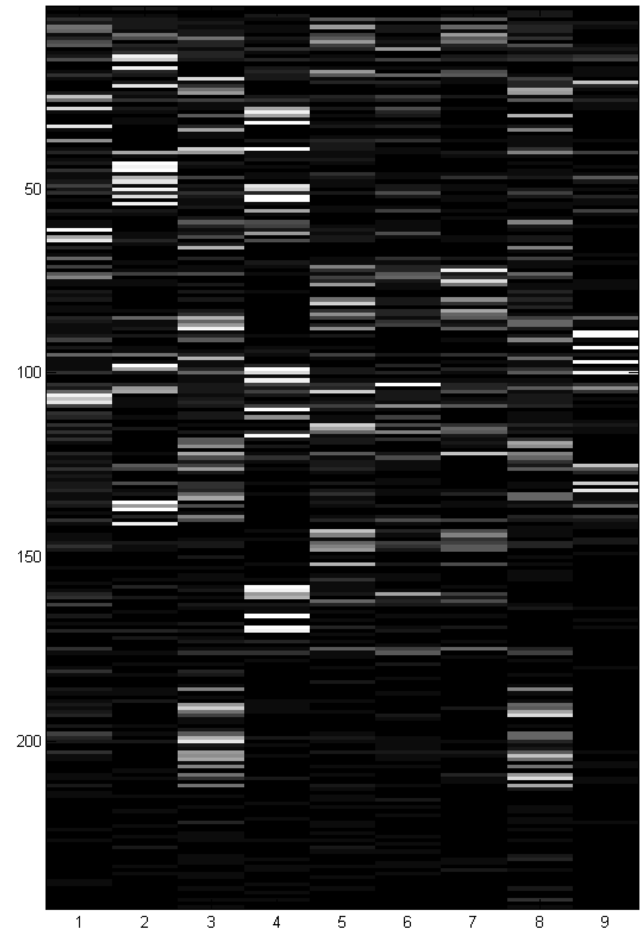
ICA iFHMM (more signals than sources)



True

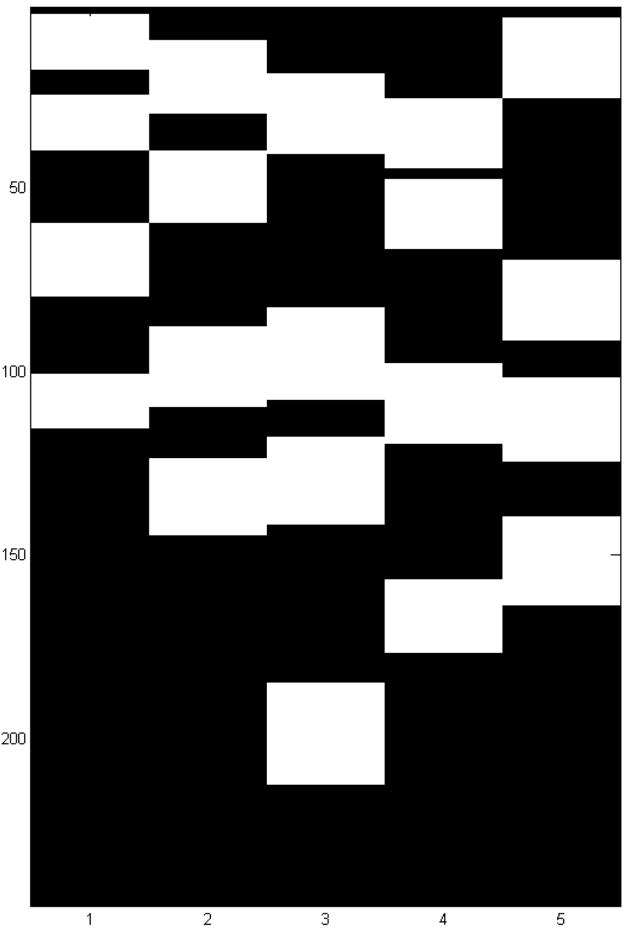


Ordered ICA iFHMM

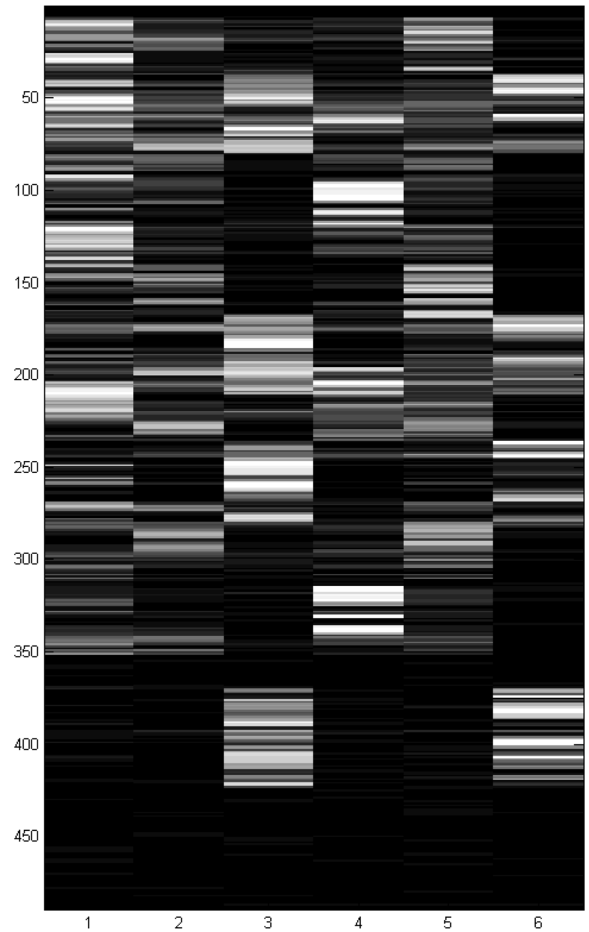


Ordered iICA

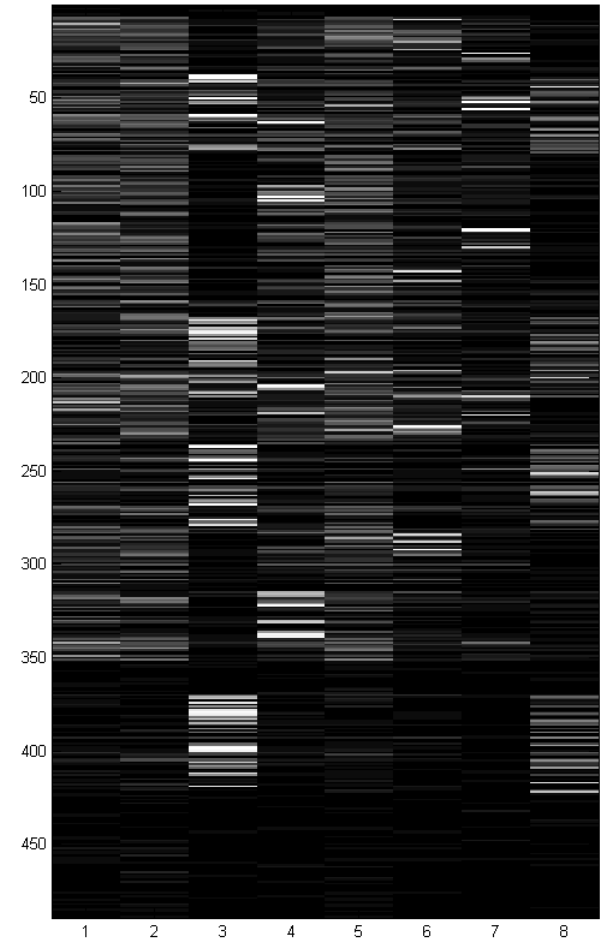
ICA iFHMM (fewer signals than sources)



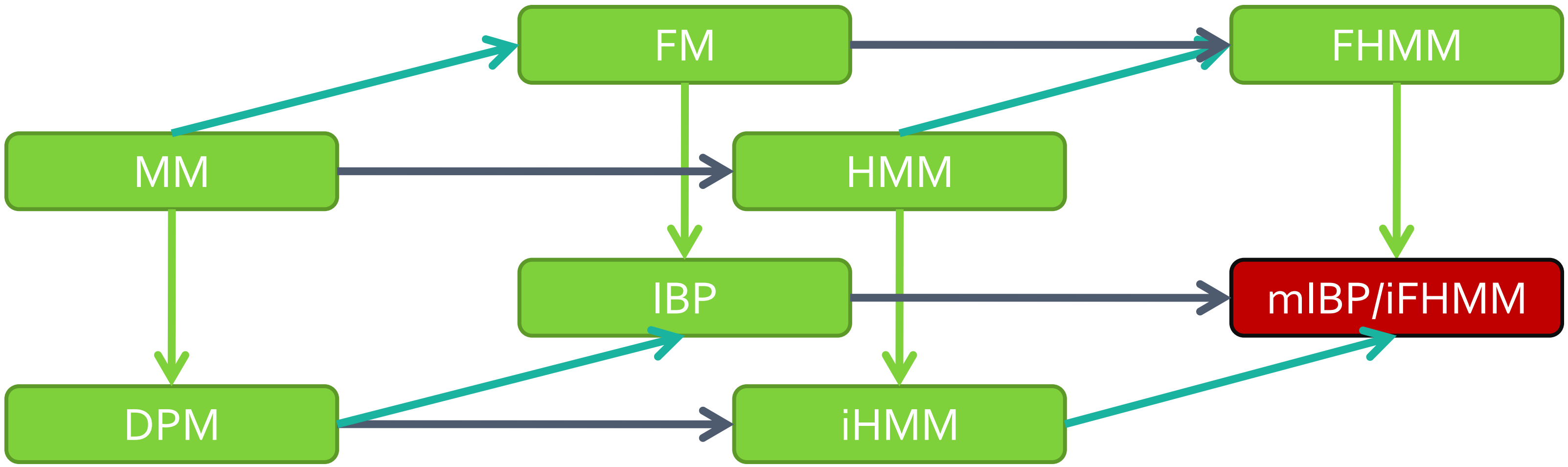
True



Ordered ICA iFHMM



Ordered iICA



Conclusion

mIBP is a new nonparametric building block

Future Work & Notes

- Better inference schemes.
- Time varying model parameters.
- Matlab prototype software available, fast parallel implementation for .NET available too.

Thank You!

Questions?
Suggestions?