Applied Bayesian Nonparametrics
3. Infinite Hidden Markov Models

Tutorial at CVPR 2012

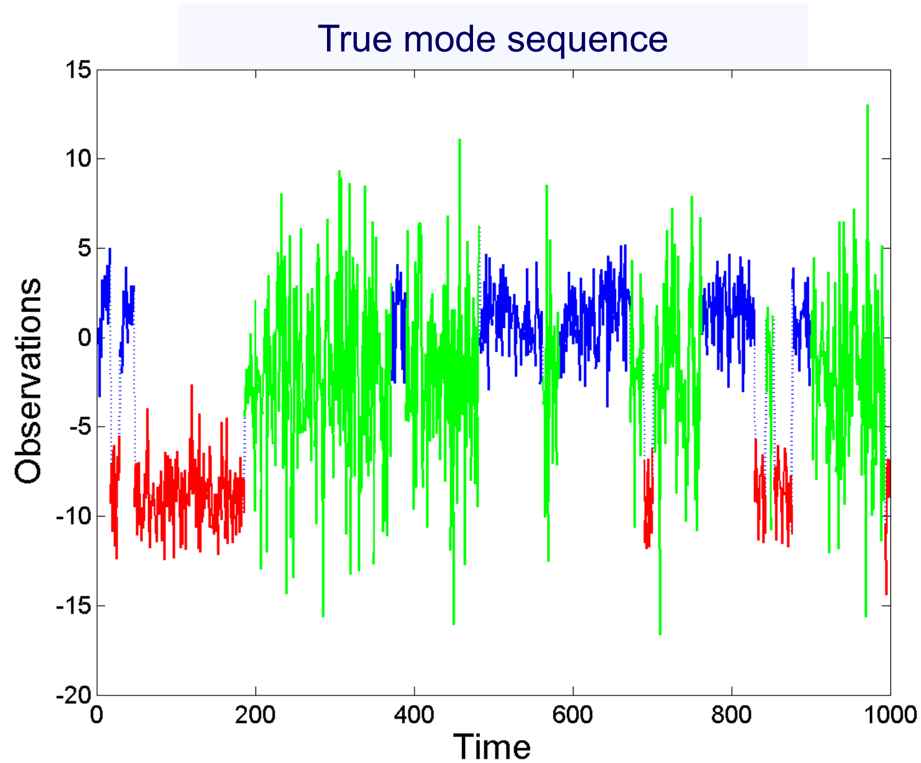
Erik Sudderth

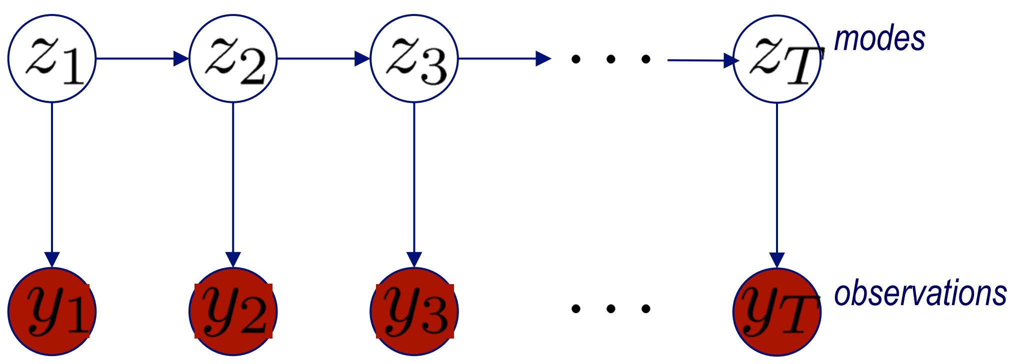
Brown University

Work by E. Fox, E. Sudderth, M. Jordan, & A. Willsky
AOAS 2011: *A Sticky HDP-HMM with Application to Speaker Diarization*
IEEE TSP 2011 & NIPS 2008: *Bayesian Nonparametric Inference of Switching Dynamic Linear Models*
NIPS 2009: *Sharing Features among Dynamical Systems with Beta Processes*

Temporal Segmentation

- Markov switching models for time series data
- Cluster based on underlying *mode dynamics*



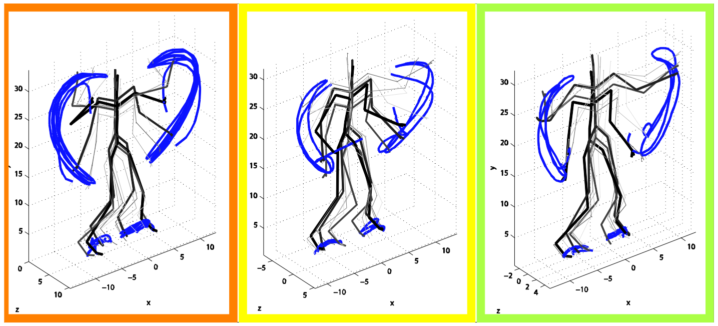


Hidden Markov Model

Outline

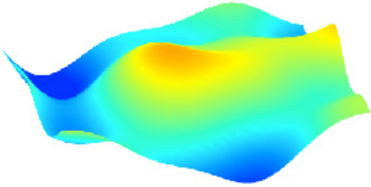
Temporal Segmentation

- How many dynamical modes?
- Mode persistence
- Complex local dynamics
- Multiple time series

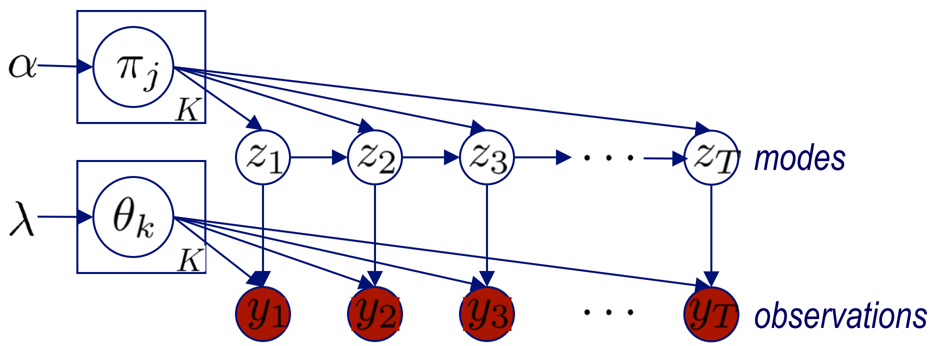


Spatial Segmentation

- Ising and Potts MRFs
- Gaussian processes

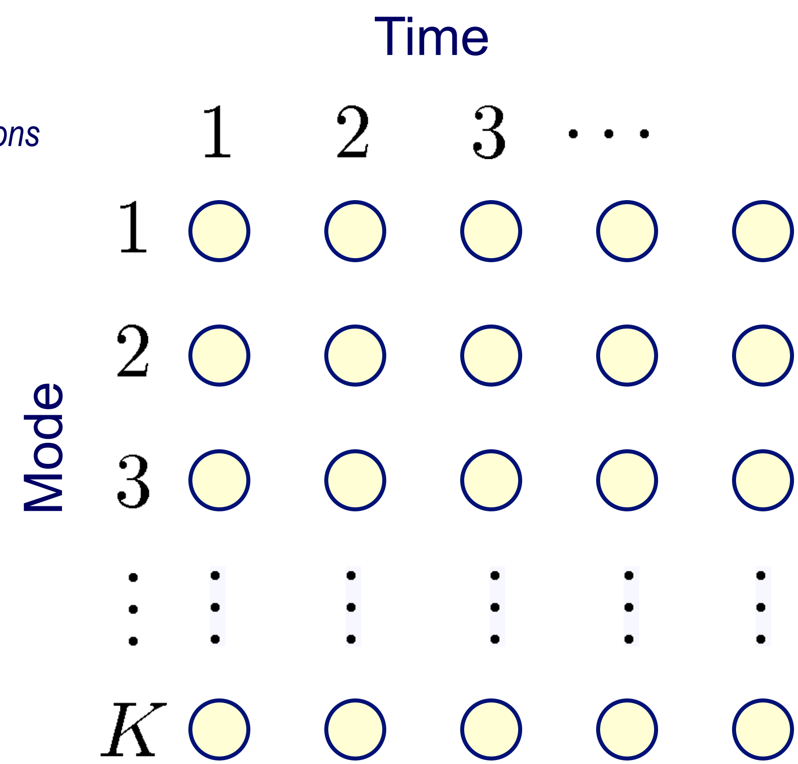


Hidden Markov Models

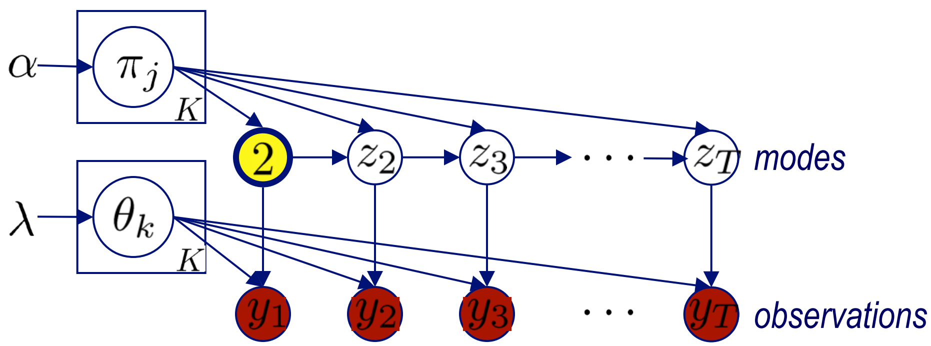


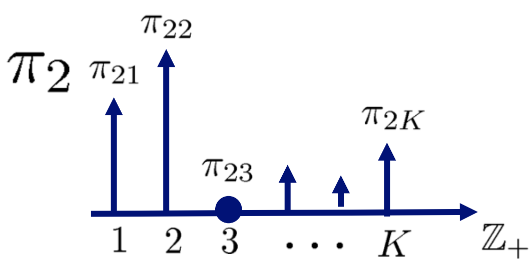
$$
\begin{aligned}
z_t &\sim \pi_{z_{t-1}} \\
y_t &\sim F(\theta_{z_t})
\end{aligned}
$$

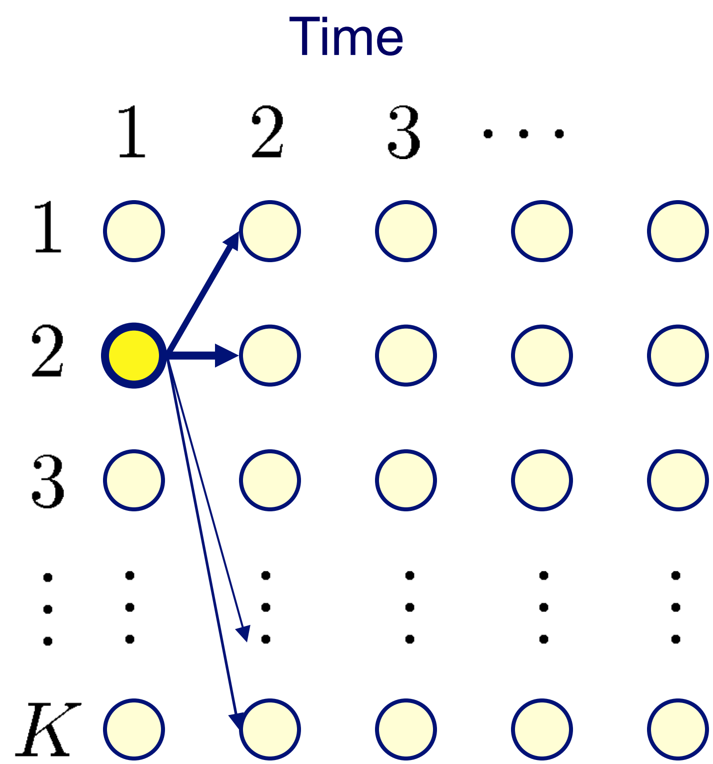
$$
P = \begin{bmatrix} \text{———} & \pi_1 & \text{———} \\ \text{———} & \pi_2 & \text{———} \\ & \vdots & \\ \text{———} & \pi_K & \text{———} \end{bmatrix}
$$



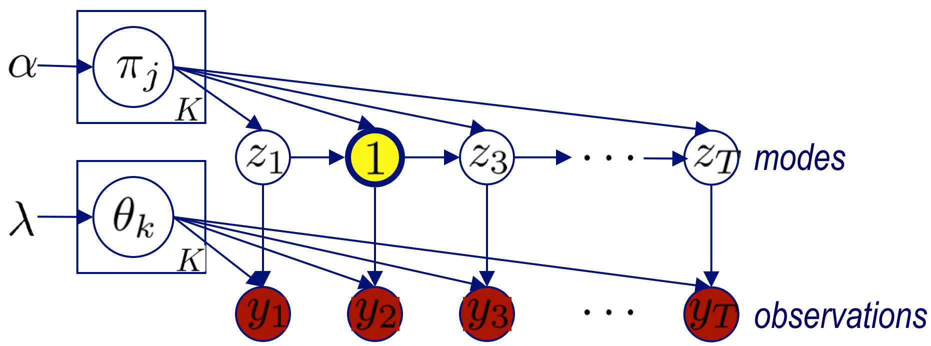
Hidden Markov Models

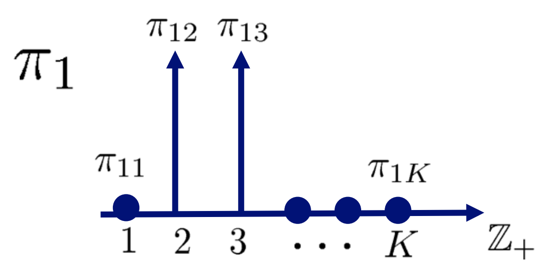


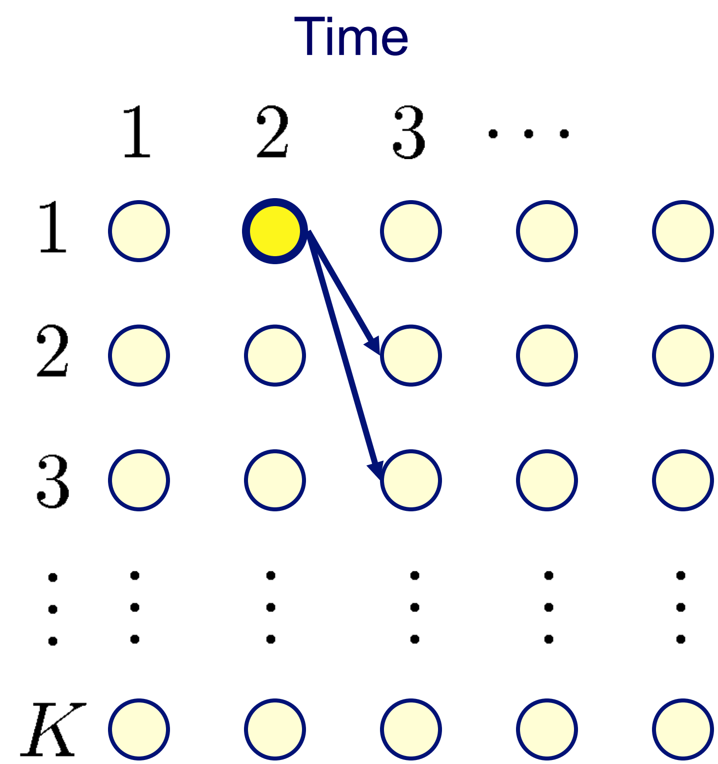




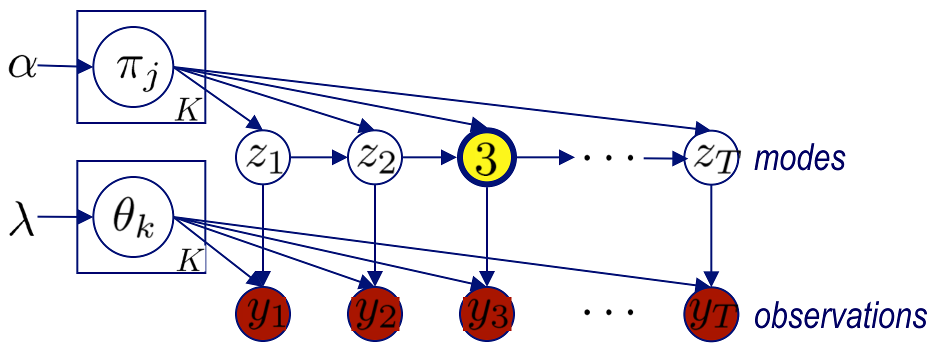
Hidden Markov Models

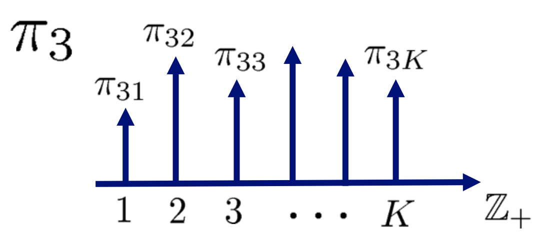


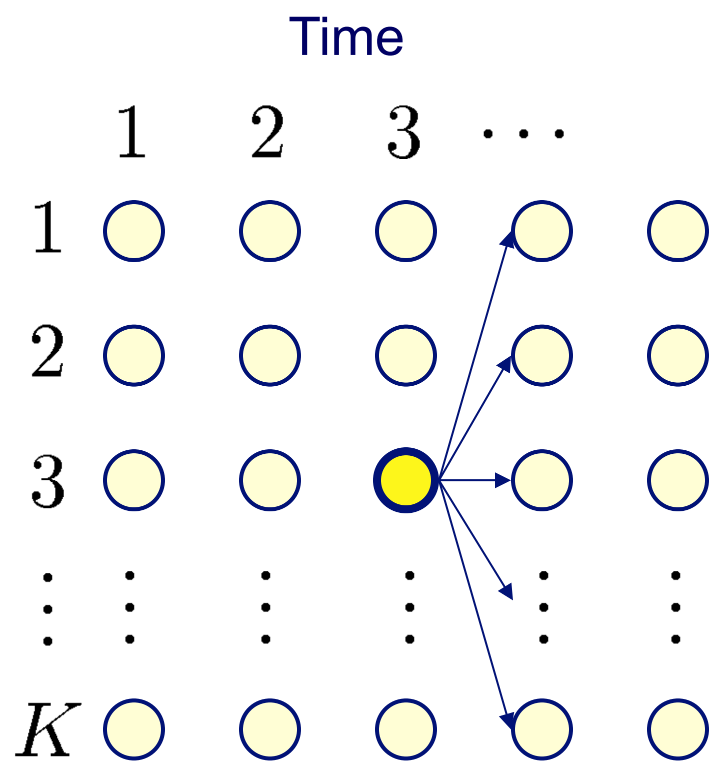




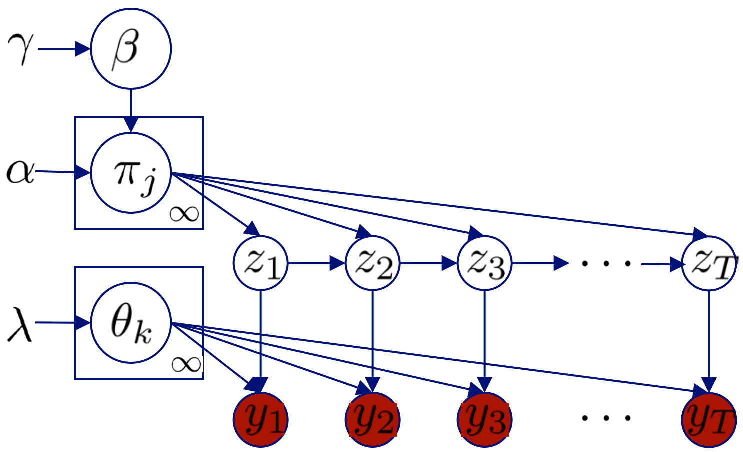
Hidden Markov Models

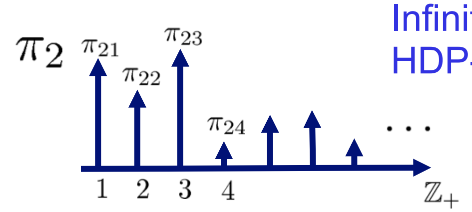




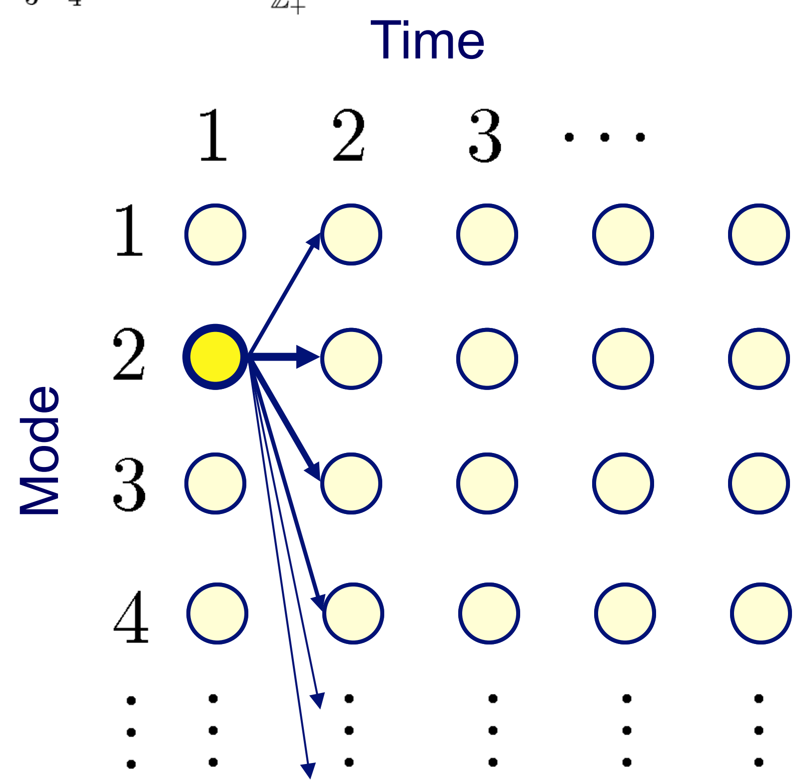


Issue 1: How many modes?





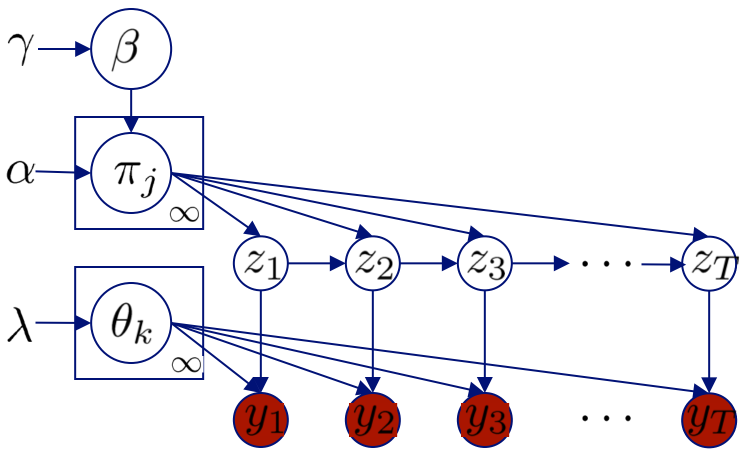
Infinite HMM: Beal, et.al., *NIPS* 2002
HDP-HMM: Teh, et. al., *JASA* 2006

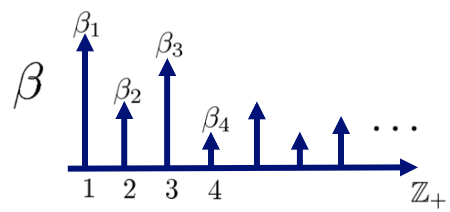


Hierarchical Dirichlet Process HMM

- Dirichlet process (DP):
  - Mode space of unbounded size
  - Model complexity adapts to observations
- Hierarchical:
  - Ties mode transition distributions
  - *Shared* sparsity

HDP-HMM

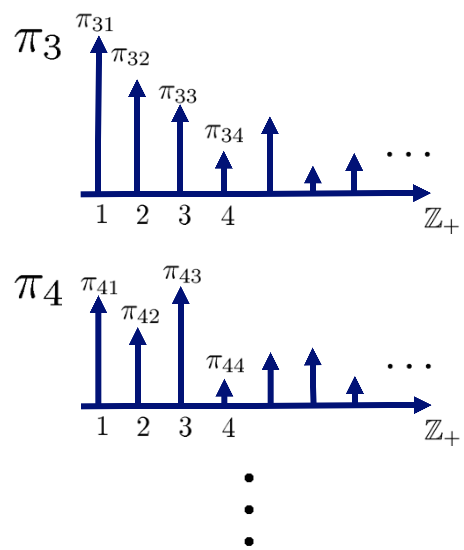




Hierarchical Dirichlet Process HMM

- Global transition distribution:

$$\beta \sim \text{Stick}(\gamma)$$

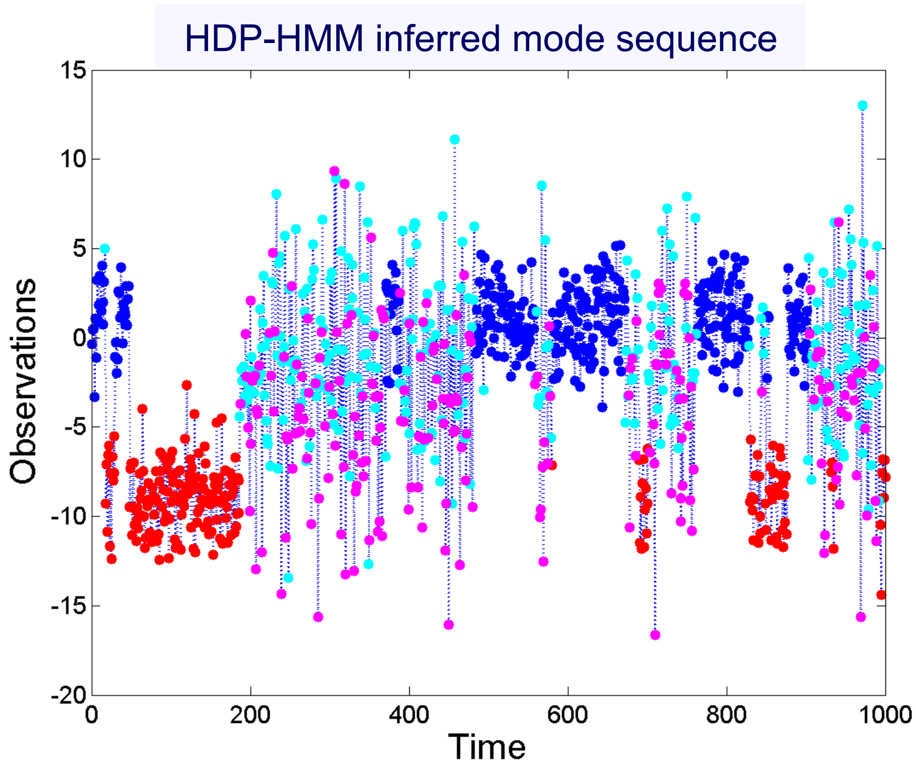


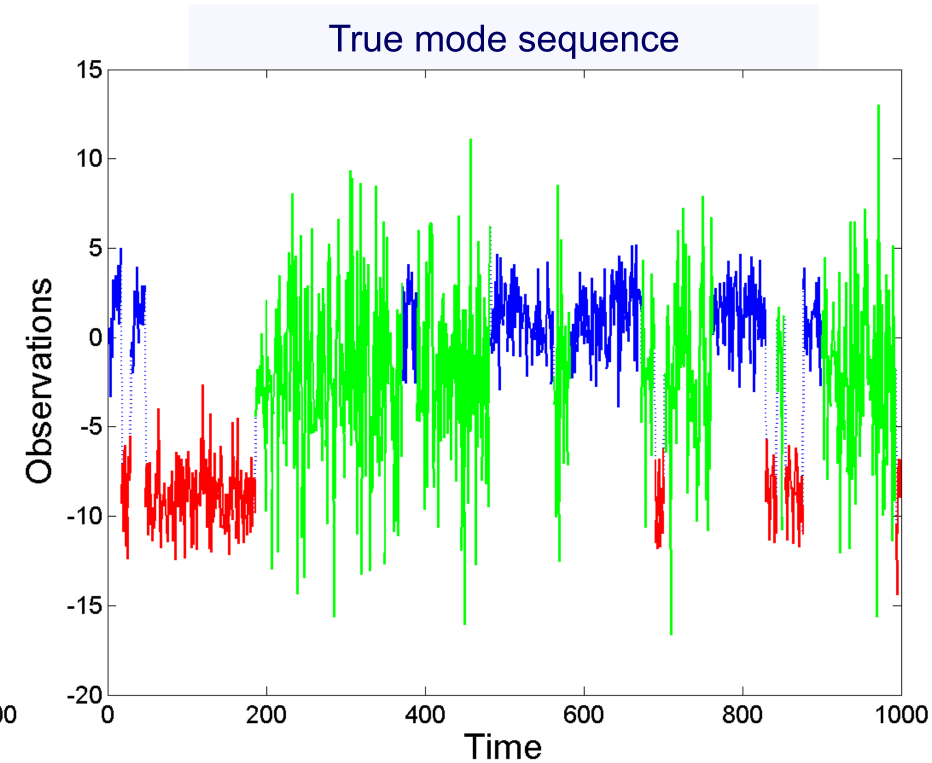
- Mode-specific transition distributions:

$$\pi_j \sim \text{DP}(\alpha\beta) \quad j = 1, 2, 3, \ldots$$

sparsity of β is shared $\longrightarrow$ $\boxed{E[\pi_{jk}] = \beta_k}$

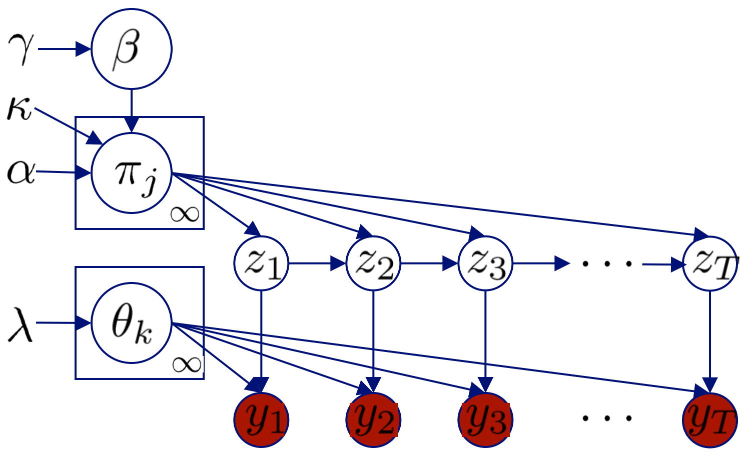
Issue 2: Temporal Persistence

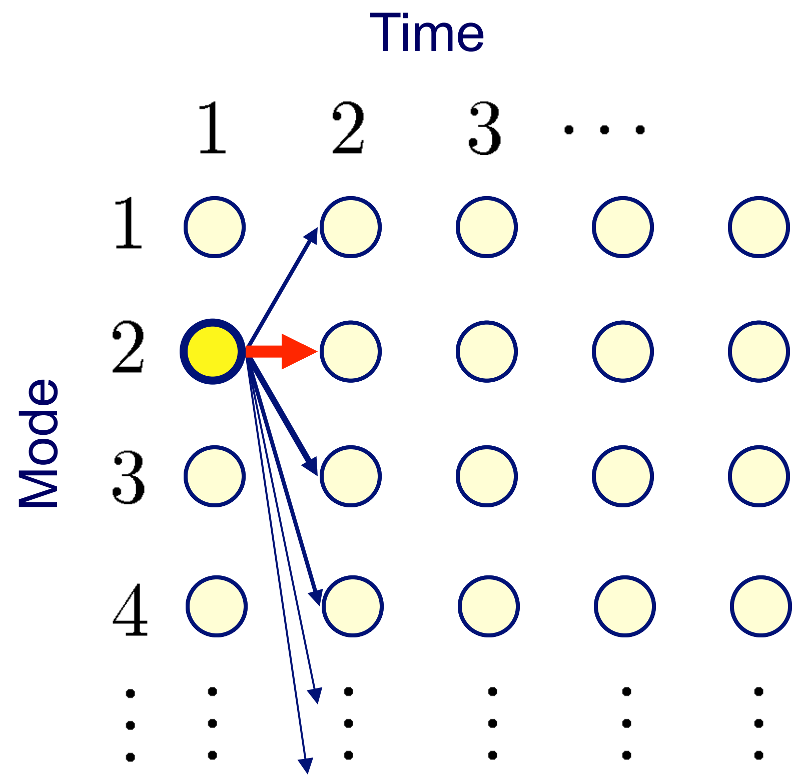




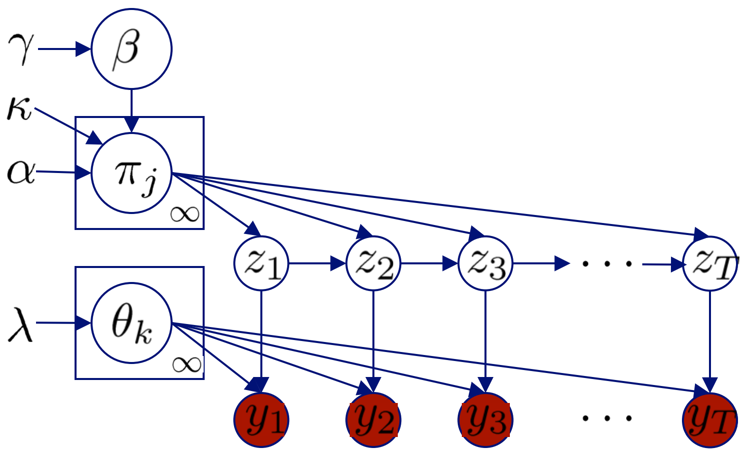
Hidden Markov Model

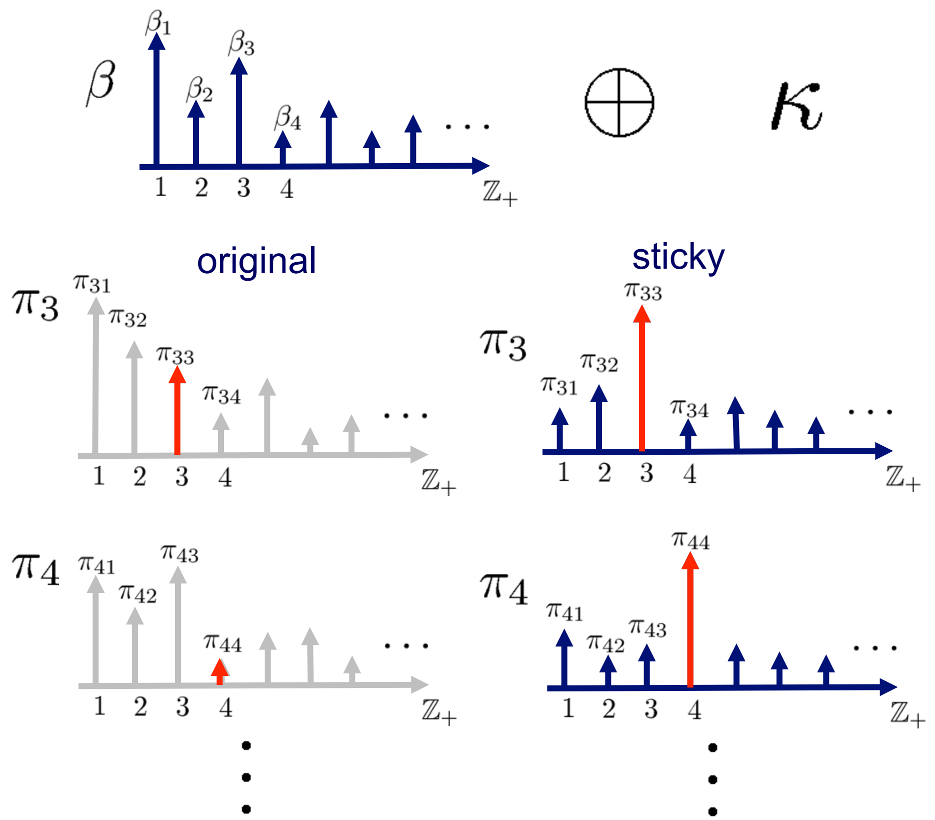
"Sticky" HDP-HMM





"Sticky" HDP-HMM





$$\beta \sim \mathrm{Stick}(\gamma)$$
$$\pi_j \sim \mathrm{DP}\left(\alpha\beta + \kappa\delta_j\right)$$

mode-specific base measure

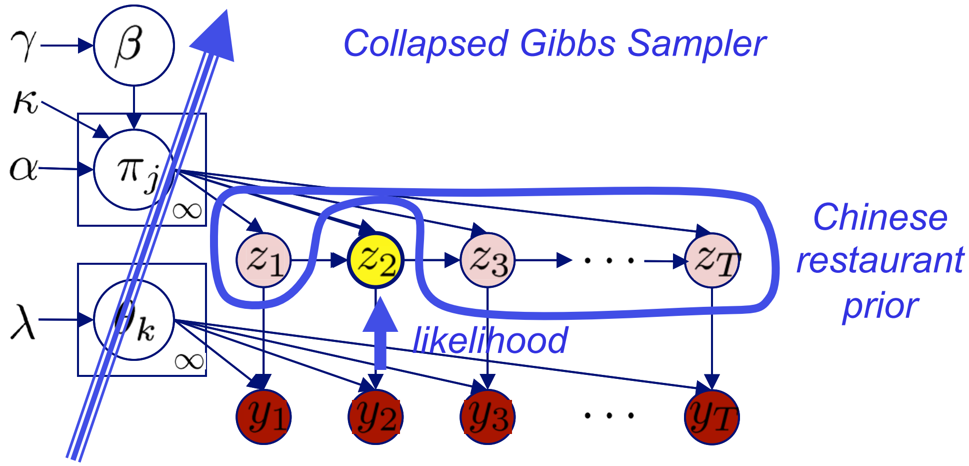
$$\boxed{E[\pi_{jk}] = \beta_k}$$

Increased probability of self-transition

$$\boxed{E[\pi_{jk}] = \frac{\alpha\beta_k + \kappa\delta(j,k)}{\alpha + \kappa}}$$

Infinite HMM: Beal, et.al., *NIPS* 2002

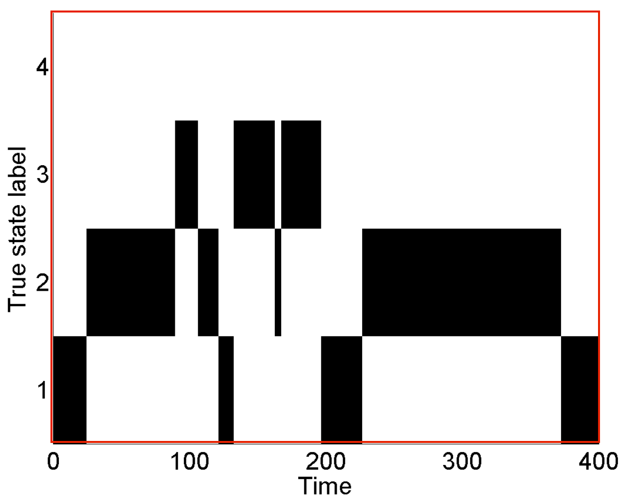
Direct Assignment Sampler

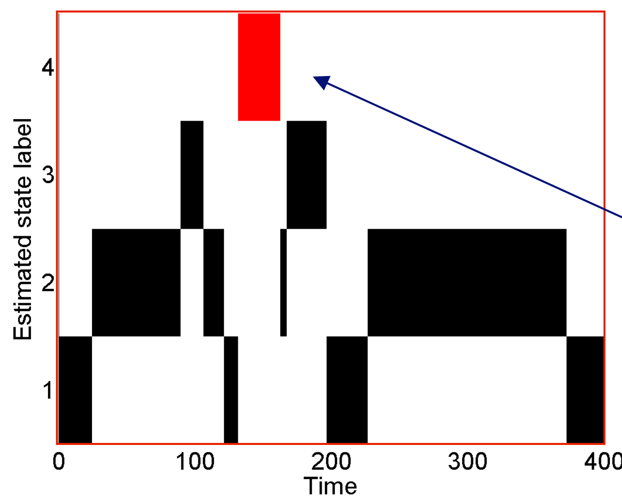


- Marginalize:
  - ➢ Transition densities
  - ➢ Emission parameters
- Sequentially sample:

$$z_t^{(n)} \sim p(z_t \mid z_{\setminus t}^{(n-1)}, \alpha, \kappa) p(y_t \mid z, y_{\setminus t}, \lambda)$$

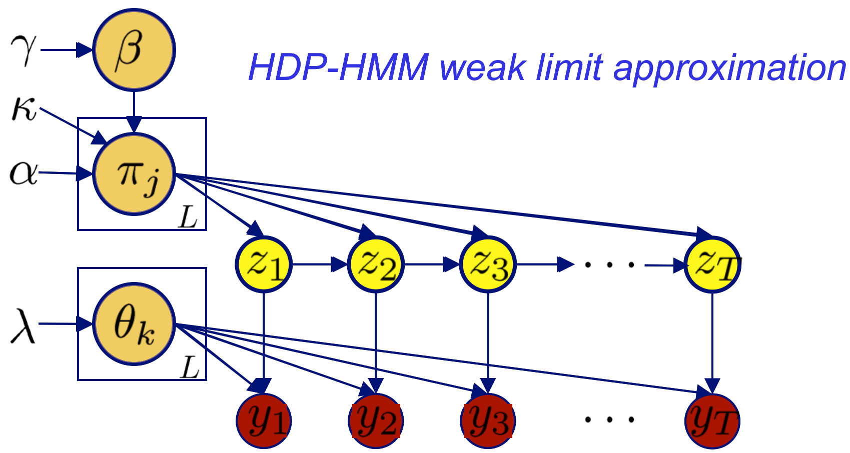
Conjugate base measure ⇒ closed form





Splits true mode, hard to merge

Blocked Resampling



$$\beta \sim \mathrm{Dir}(\gamma/L, \ldots, \gamma/L)$$

$$\pi_j \sim \mathrm{Dir}(\alpha\beta_1, \ldots, \alpha\beta_j + \kappa, \ldots, \alpha\beta_L)$$

- Compute backwards messages:
- Approximate HDP:
  - Average transition density
  - $$m^{(n)}_{t,t-1}(z_{t-1}) \propto \sum_{z_t} p(z_t | \pi^{(n)}_{z_{t-1}}) p(y_t | \theta^{(n)}_{z_t}) m^{(n)}_{t+1,t}(z_t)$$
  - ($\Rightarrow$ transition densities)
- Sample:
- Block sample $z^{(n)}_{1:T}$ as:

$$\beta^{(n)}, \quad \pi^{(n)}_j, \quad \theta^{(n)}_j \quad j = 1, \ldots, L$$

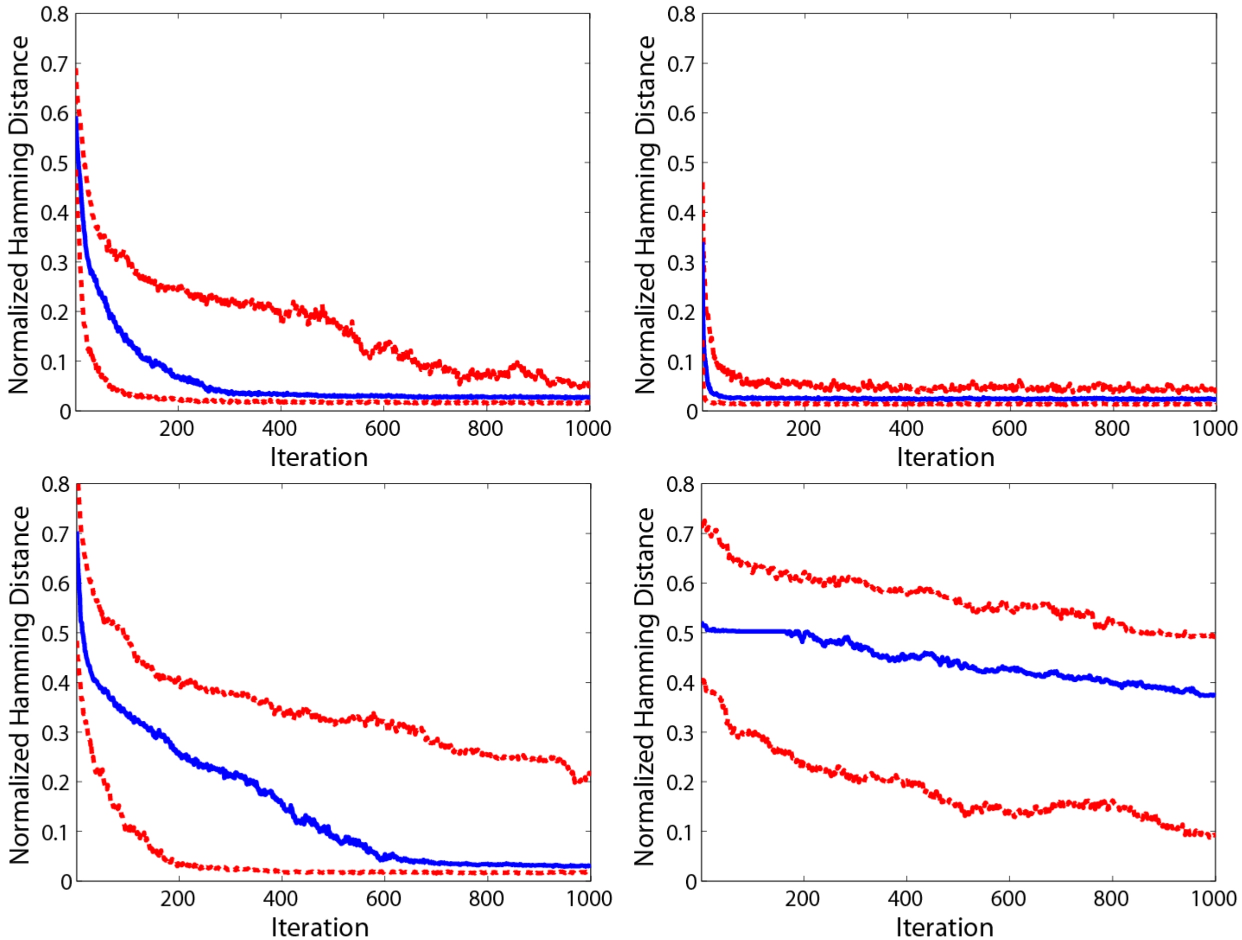
$$z^{(n)}_t \sim p(z_t | \pi^{(n)}_{z^{(n)}_{t-1}}) p(y_t | \theta^{(n)}_{z_t}) m^{(n)}_{t+1,t}(z_t)$$

Results: Gaussian Emissions

HDP-HMM | Sticky HDP-HMM

Blocked sampler

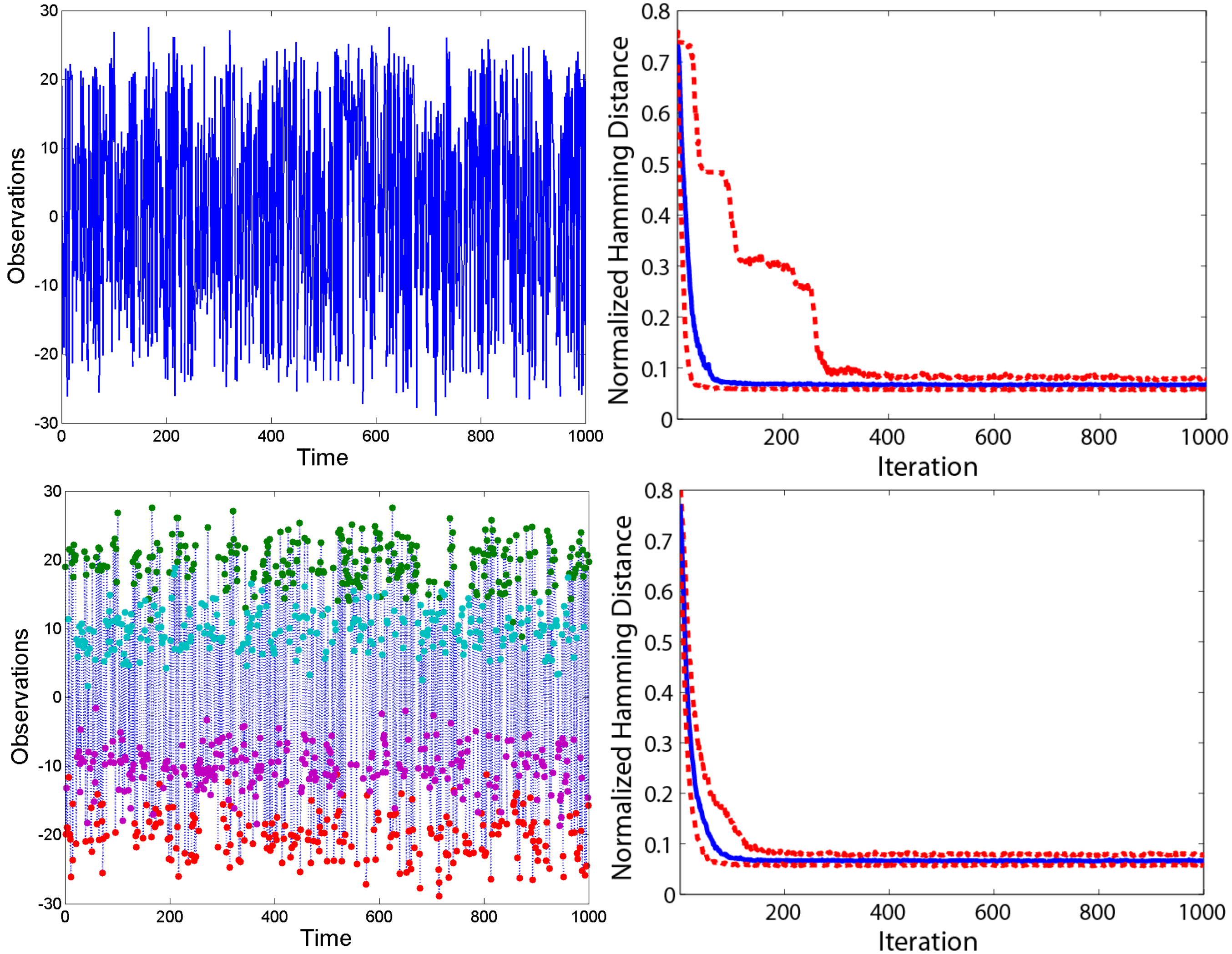
Sequential sampler



Results: Fast Switching

Observations

True mode sequence

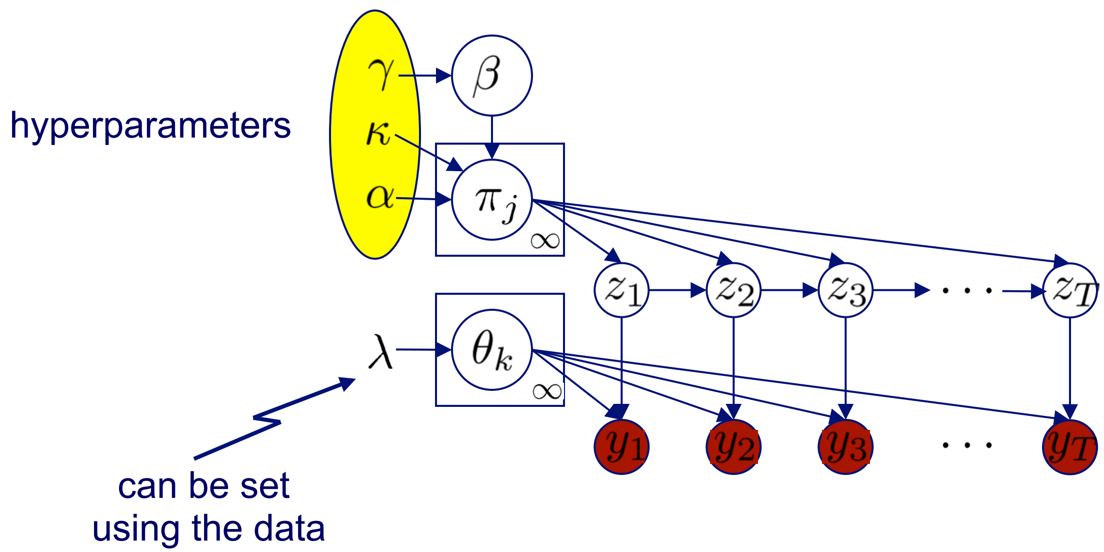


Sticky HDP-HMM

HDP-HMM

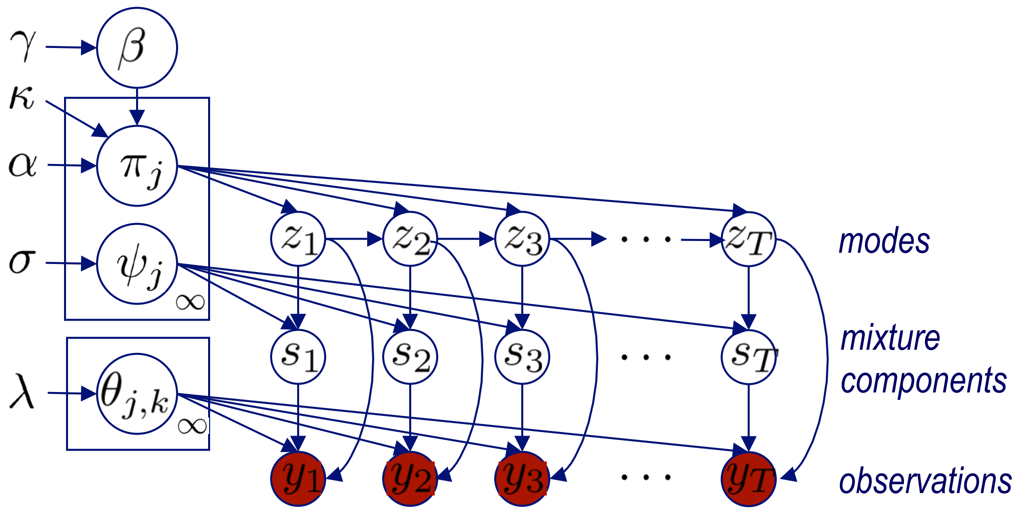
Hyperparameters

- Place priors on hyperparameters and infer them from data
- Weakly informative priors
- All results use the same settings

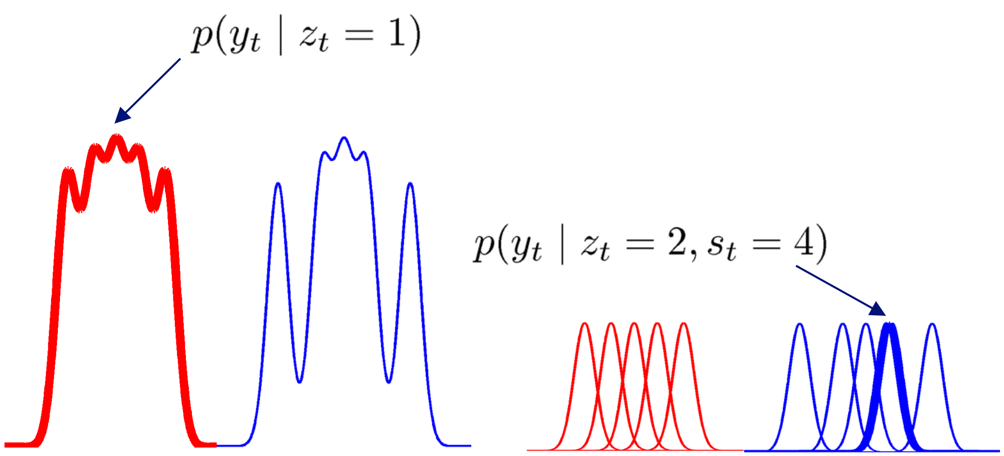


Related self-transition parameter:
Beal, et.al., *NIPS* 2002

HDP-HMM: Multimodal Emissions

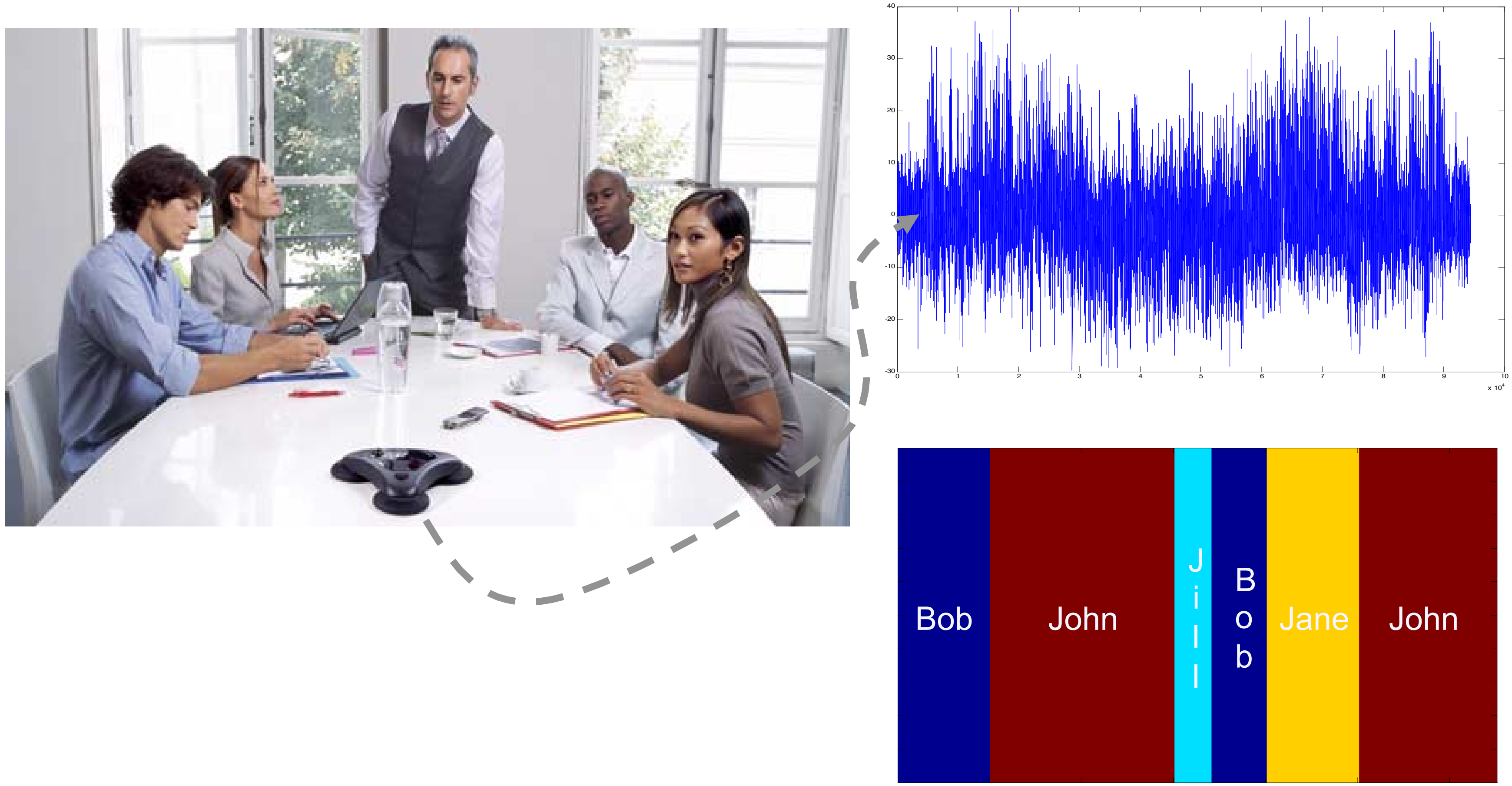


$$
\begin{aligned}
\beta &\sim \text{Stick}(\gamma) \\
\pi_j &\sim \text{DP}\left(\alpha\beta + \kappa\delta_j\right) \\
\psi_j &\sim \text{Stick}(\sigma) \\
z_t &\sim \pi_{z_{t-1}} \\
s_t &\sim \psi_{z_t} \\
y_t &\sim F(\theta_{z_t,s_t})
\end{aligned}
$$

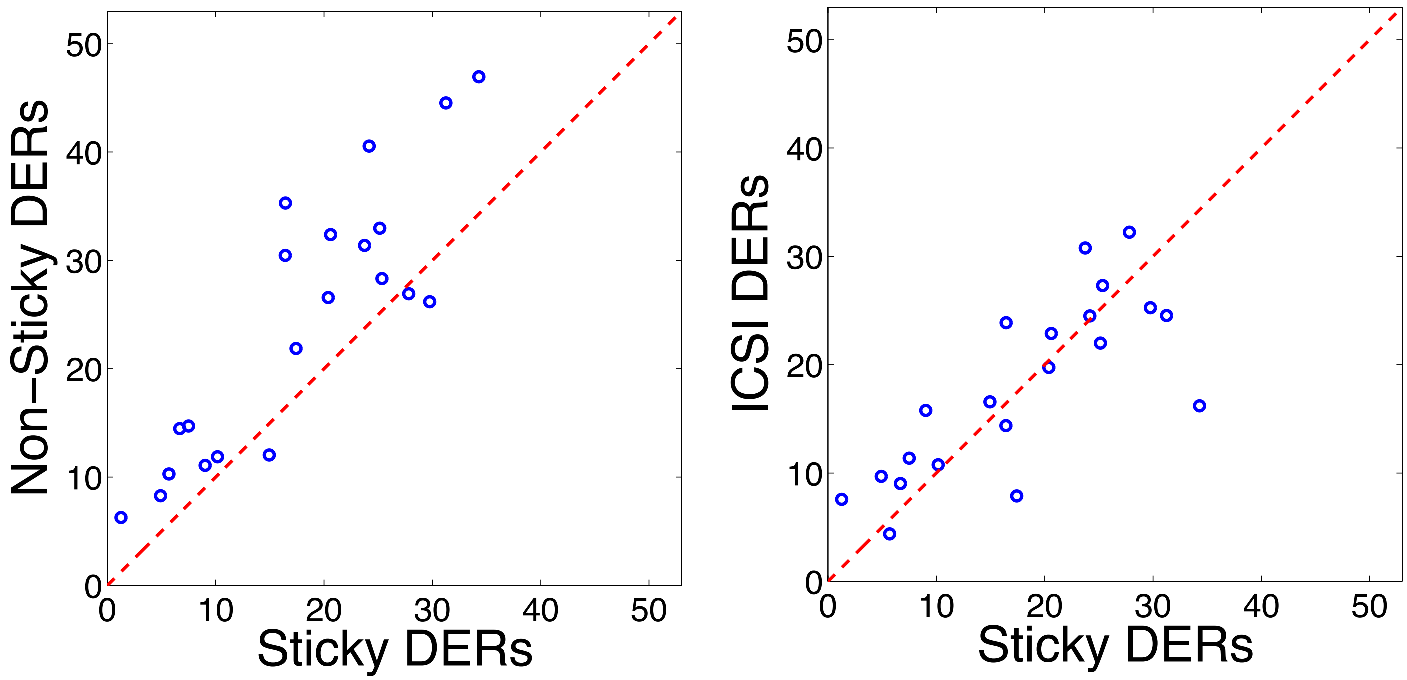


- Approximate multimodal emissions with DP mixture
- Temporal mode persistence disambiguates model

Speaker Diarization

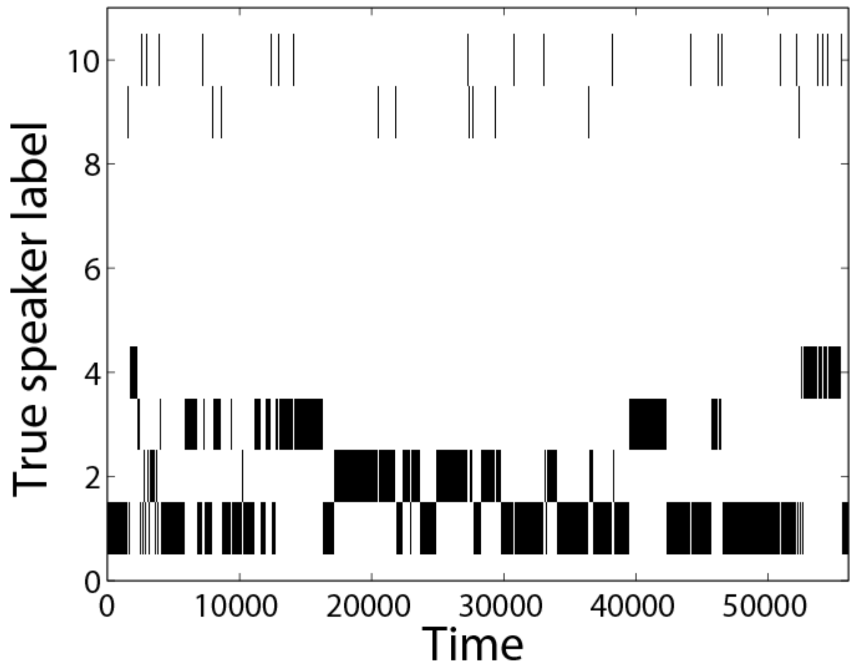


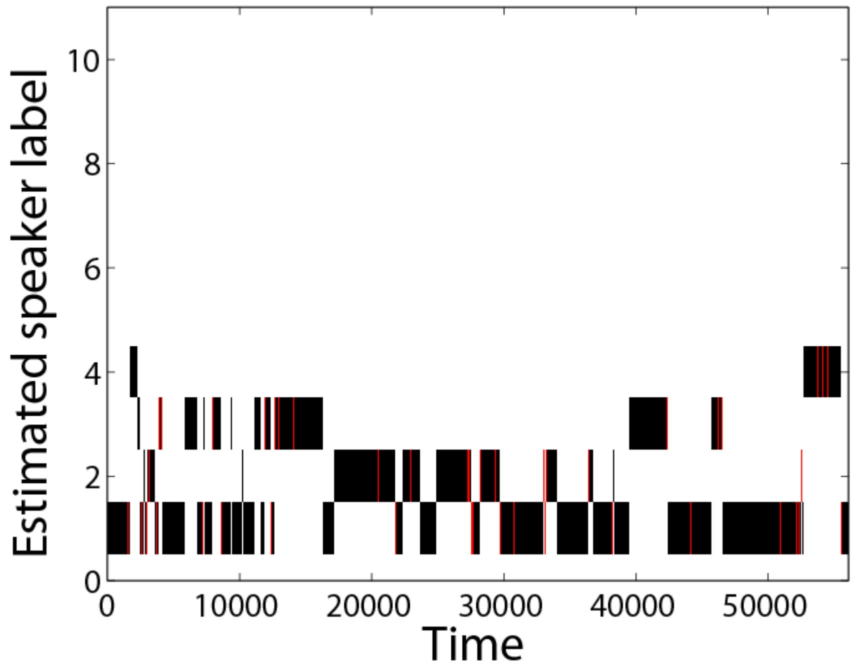
Results: 21 meetings



	Overall DER	Best DER	Worst DER
Sticky HDP-HMM	17.84%	1.26%	34.29%
Non-Sticky HDP-HMM	23.91%	6.26%	46.95%
ICSI	18.37%	4.39%	32.23%

Results: Meeting 1

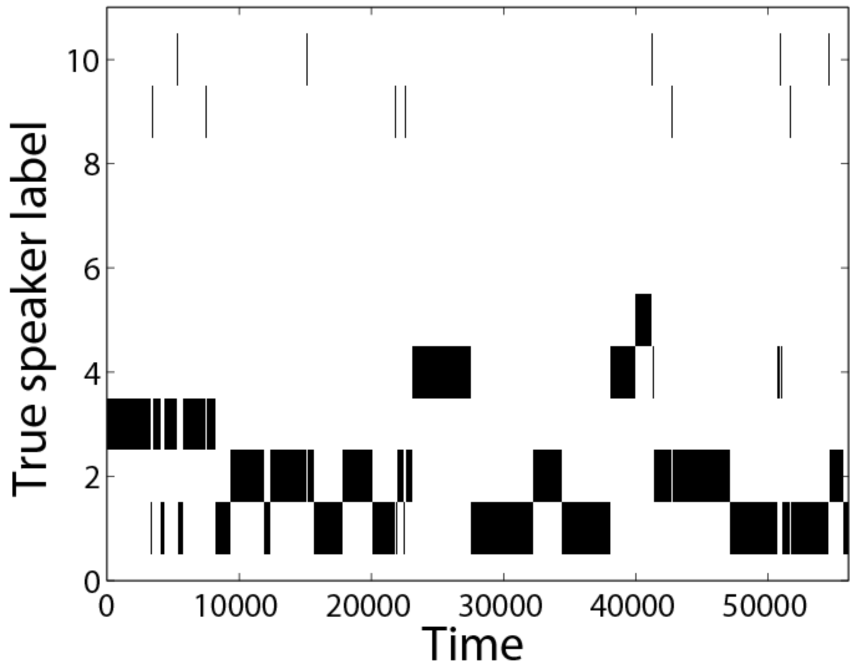


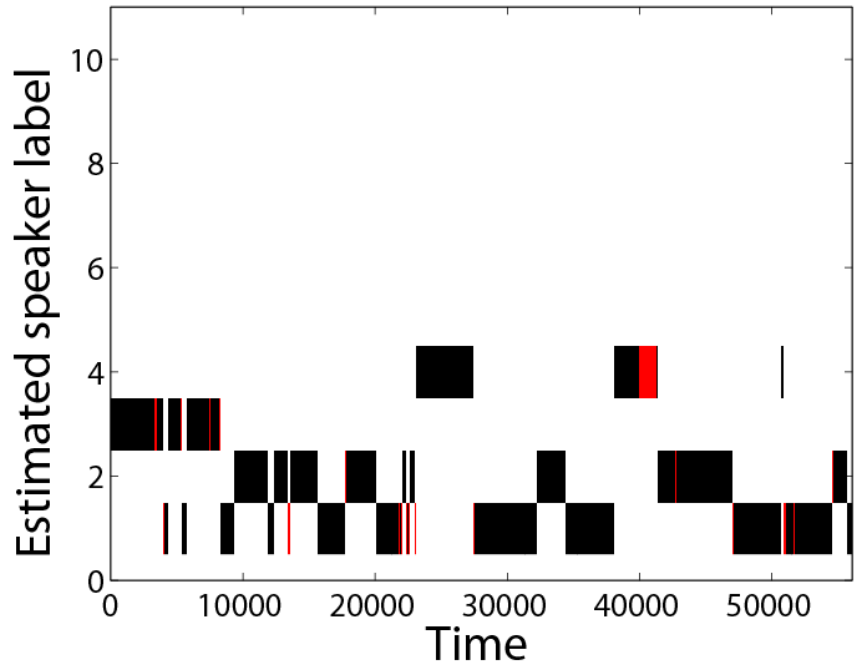


Sticky DER = 1.26%

ICSI DER = 7.56%

Results: Meeting 18



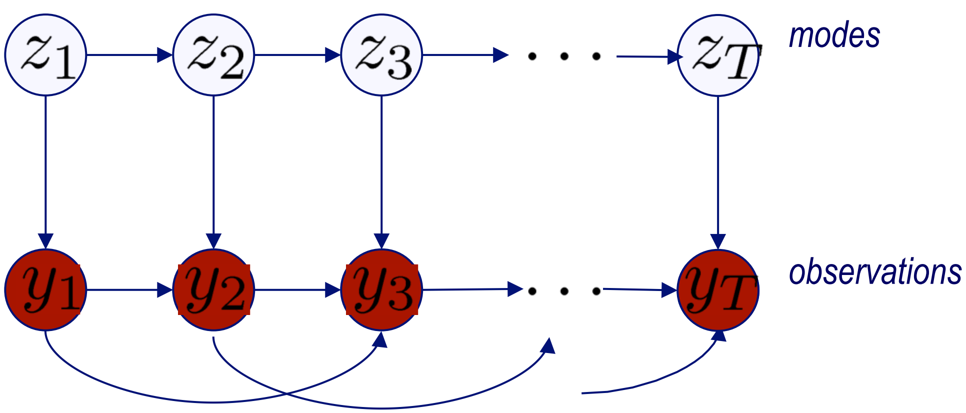


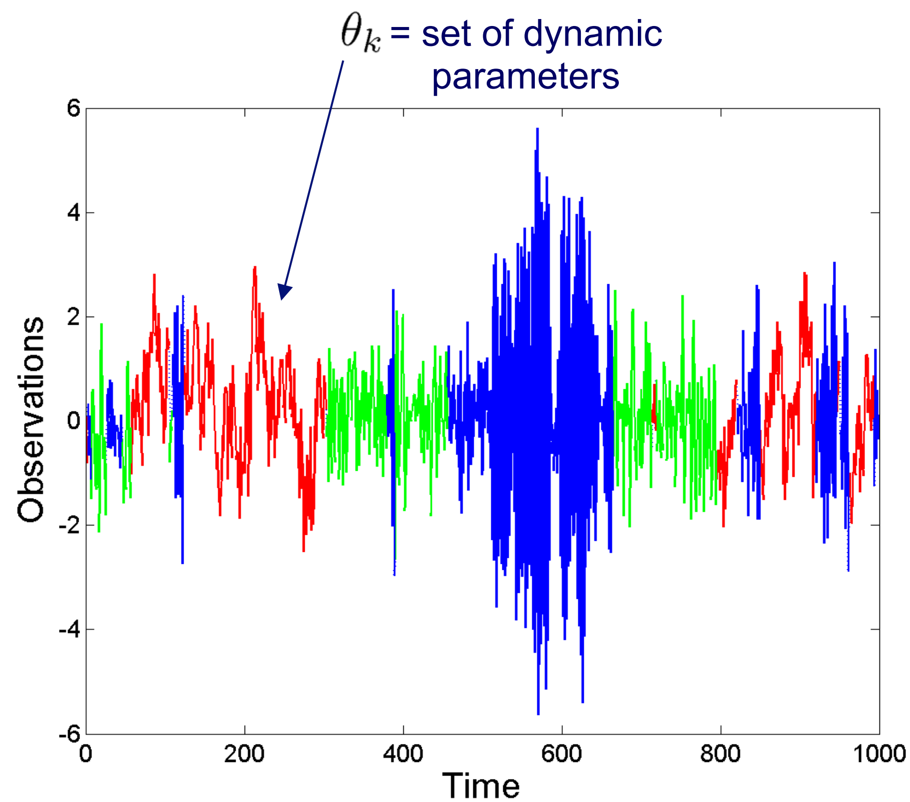
Sticky DER = ~~20.48%~~ 4.81%

ICSI DER = 22.00%

Issue 3: Complex Local Dynamics

- Discrete clusters may not accurately capture high-dimensional data
- Autoregressive HMM: Discrete-mode switching of *smooth* observation dynamics





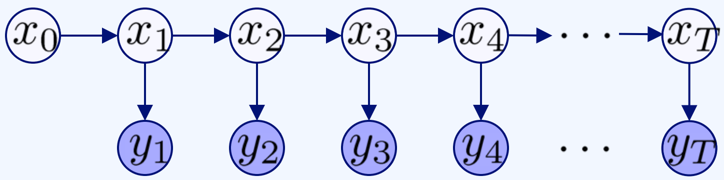
Switching Dynamical Processes

Linear Dynamical Systems

- State space LTI model:

$$x_t = Ax_{t-1} + e_t$$
$$y_t = Cx_t + w_t$$

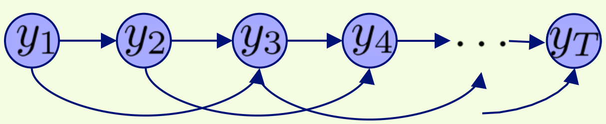
$$e_t \sim \mathcal{N}(0, \Sigma) \quad w_t \sim \mathcal{N}(0, R)$$



- Vector autoregressive (VAR) process:

$$y_t = \sum_{i=1}^{r} A_i y_{t-i} + e_t$$

$$e_t \sim \mathcal{N}(0, \Sigma)$$

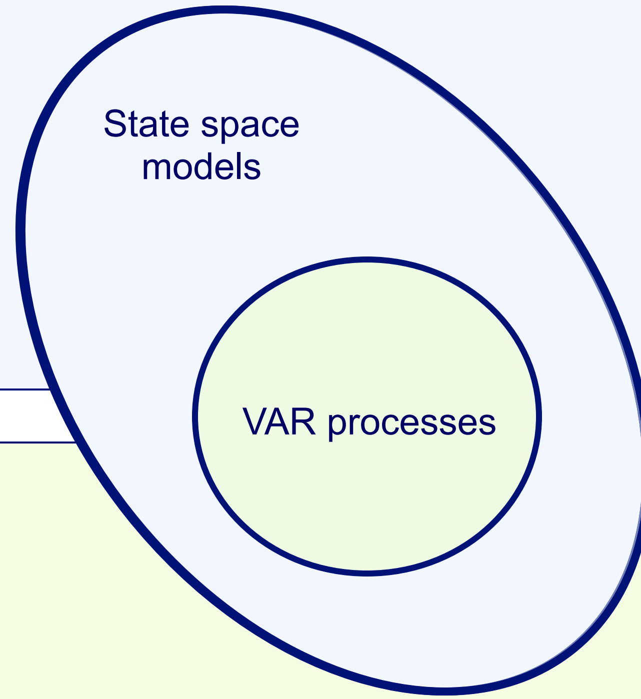


Linear Dynamical Systems

- State space LTI model:

$$x_t = Ax_{t-1} + e_t$$
$$y_t = Cx_t + w_t$$

$$e_t \sim \mathcal{N}(0, \Sigma) \quad w_t \sim \mathcal{N}(0, R)$$



- Vector autoregressive (VAR) process:

$$\begin{aligned} x_t &= \begin{bmatrix} A_1 & A_2 & \dots & A_r \\ I & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & I & 0 \end{bmatrix} x_{t-1} + \begin{bmatrix} I \\ 0 \\ \vdots \\ 0 \end{bmatrix} e_t \\ y_t &= \begin{bmatrix} I & 0 & \dots & 0 \end{bmatrix} x_t. \end{aligned}$$

Switching Dynamical Systems

Switching linear dynamical system (SLDS):

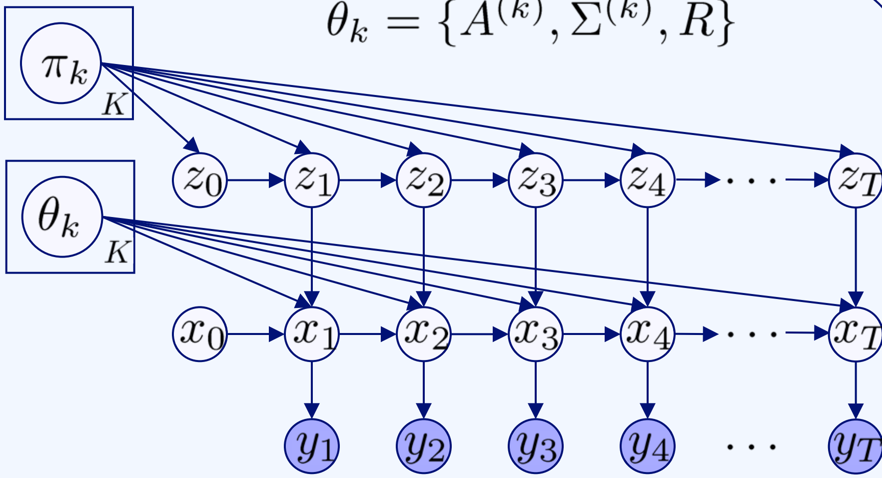
$$z_t \sim \pi_{z_{t-1}}$$

$$x_t = A^{(z_t)} x_{t-1} + e_t(z_t)$$

$$y_t = C x_t + w_t$$

$$e_t \sim \mathcal{N}(0, \Sigma^{(z_t)}) \quad w_t \sim \mathcal{N}(0, R)$$

$$\theta_k = \{A^{(k)}, \Sigma^{(k)}, R\}$$



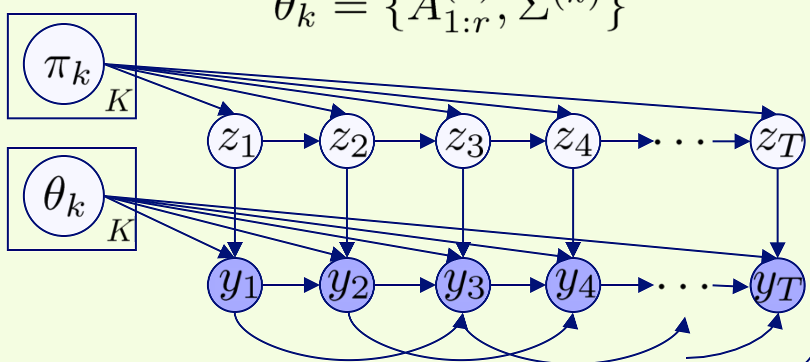
Switching VAR process:

$$z_t \sim \pi_{z_{t-1}}$$

$$y_t = \sum_{i=1}^{r} A_i^{(z_t)} y_{t-i} + e_t(z_t)$$

$$e_t \sim \mathcal{N}(0, \Sigma^{(z_t)})$$

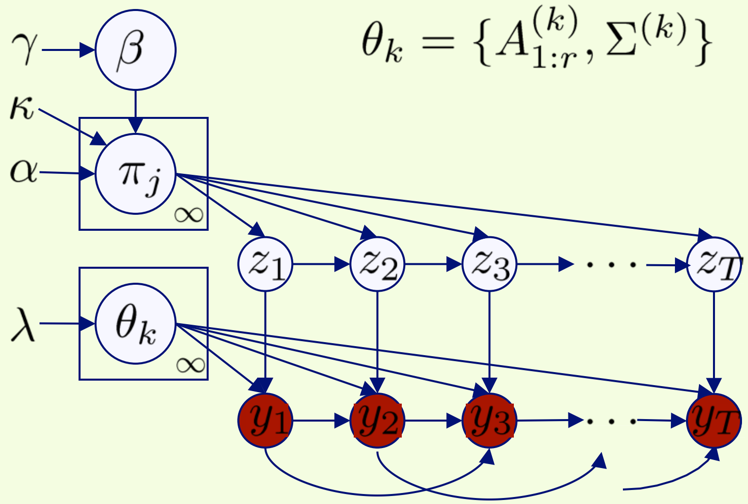
$$\theta_k = \{A_{1:r}^{(k)}, \Sigma^{(k)}\}$$



HDP-AR-HMM and HDP-SLDS

HDP-AR-HMM

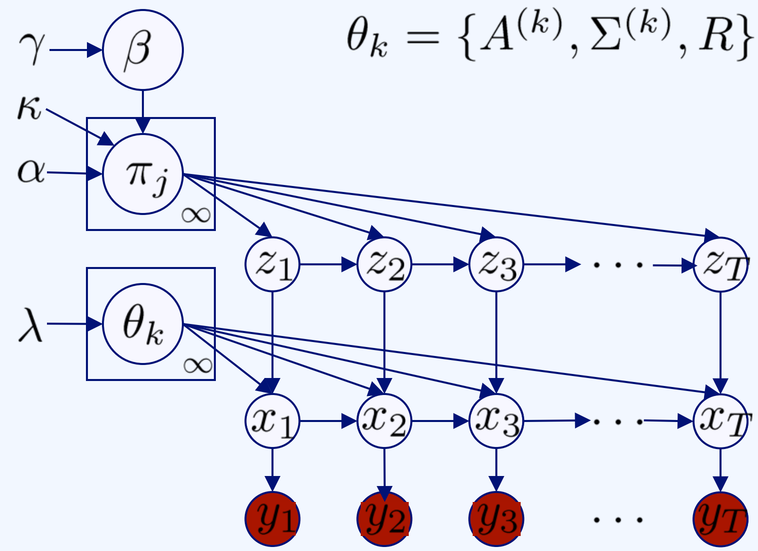
$$\theta_k = \{A^{(k)}_{1:r}, \Sigma^{(k)}\}$$



$$\begin{aligned} z_t &\sim \pi_{z_{t-1}} \\ y_t &= \sum_{i=1}^{r} A_i^{(z_t)} y_{t-i} + e_t(z_t) \end{aligned}$$

HDP-SLDS

$$\theta_k = \{A^{(k)}, \Sigma^{(k)}, R\}$$



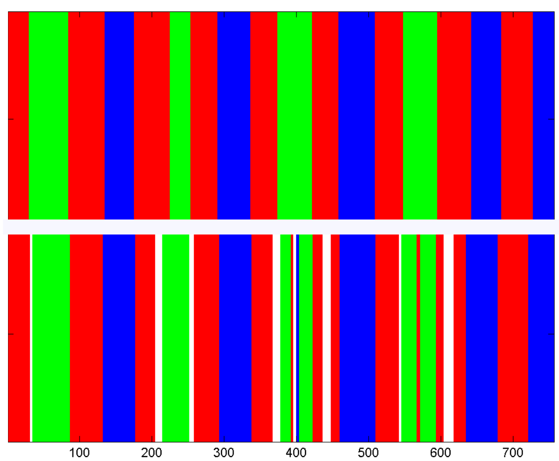
$$\begin{aligned} z_t &\sim \pi_{z_{t-1}} \\ x_t &= A^{(z_t)} x_{t-1} + e_t(z_t) \\ y_t &= C x_t + w_t \end{aligned}$$

$$C = [I \;\; 0]$$

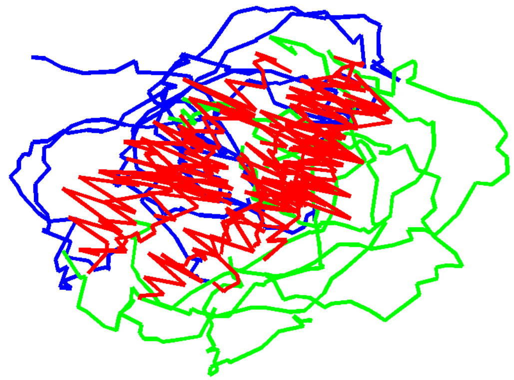
Dancing Honey Bees

Honey Bee Results: HDP-AR(1)-HMM

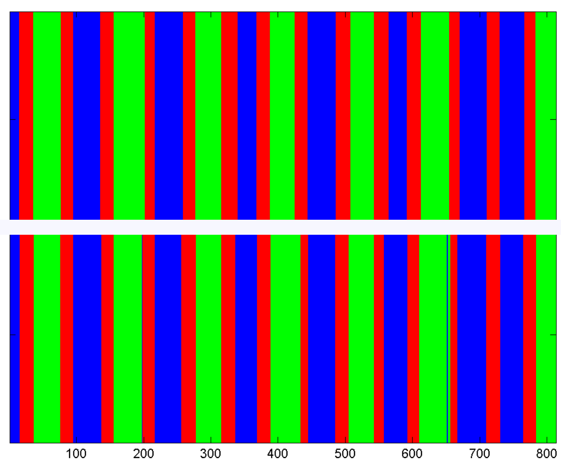
Sequence 1



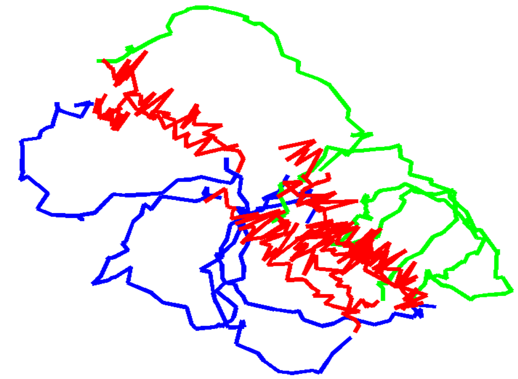
HDP-AR-HMM: 88.1%
SLDS [Oh]: 93.4%



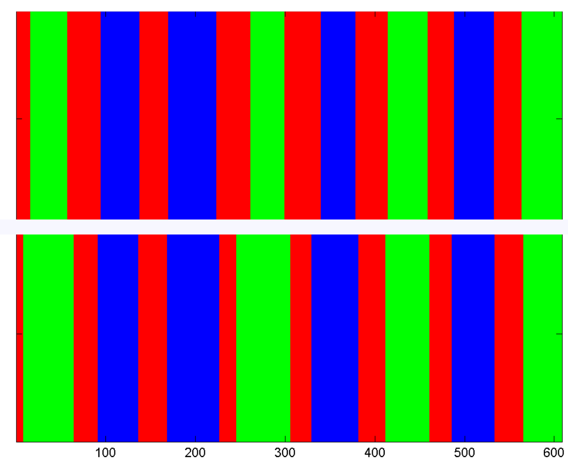
Sequence 2



HDP-AR-HMM: 92.5%
SLDS [Oh]: 90.2%

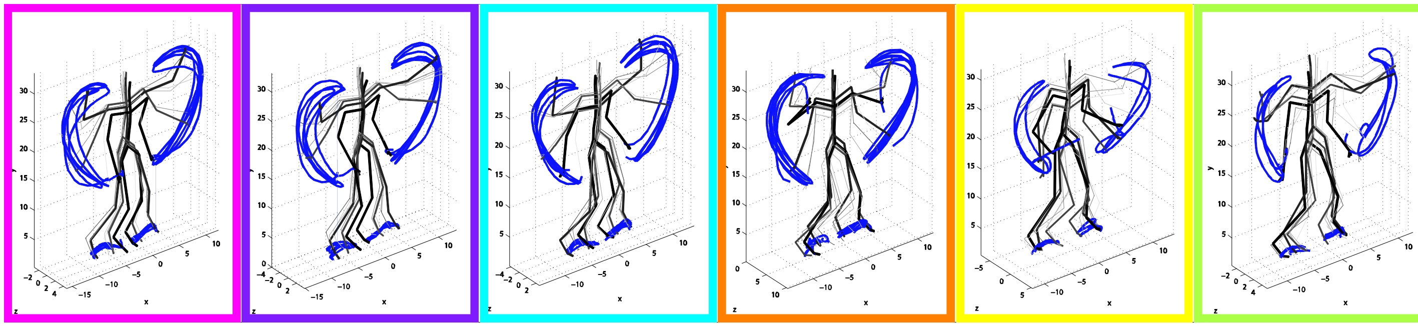


Sequence 3



HDP-AR-HMM: 88.2%
SLDS [Oh]: 90.4%

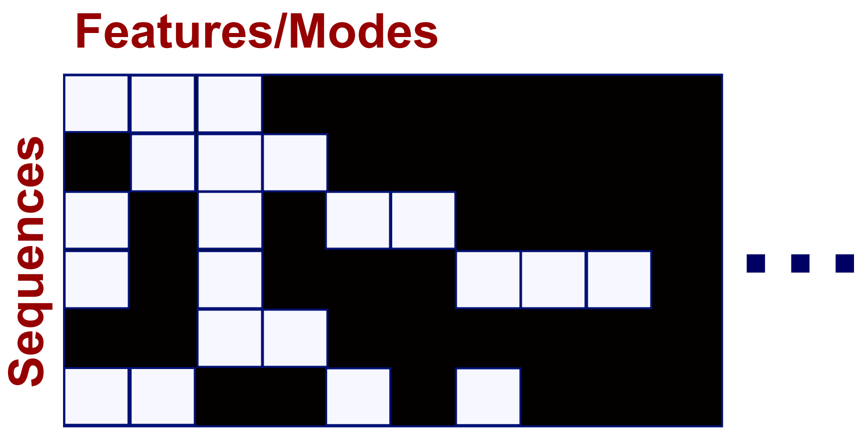
Issue 4: Multiple Time Series

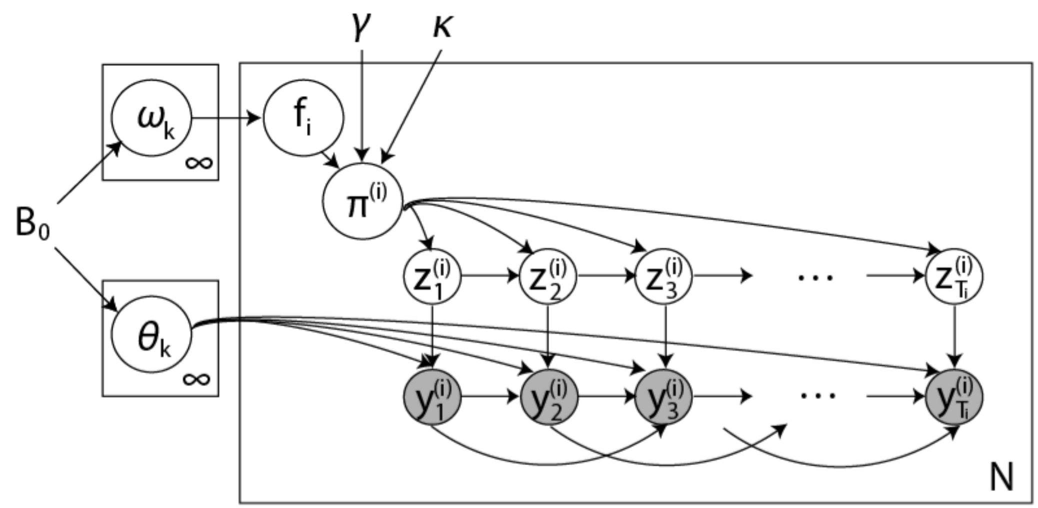


- Goal:
  - ➢ Transfer knowledge between related time series
  - ➢ Allow each system to switch between an arbitrarily large set of dynamical modes
- Method:
  - ➢ Beta process prior
  - ➢ Predictive distribution: Indian buffet process

IBP-AR-HMM

- Latent features determine which dynamical modes are used

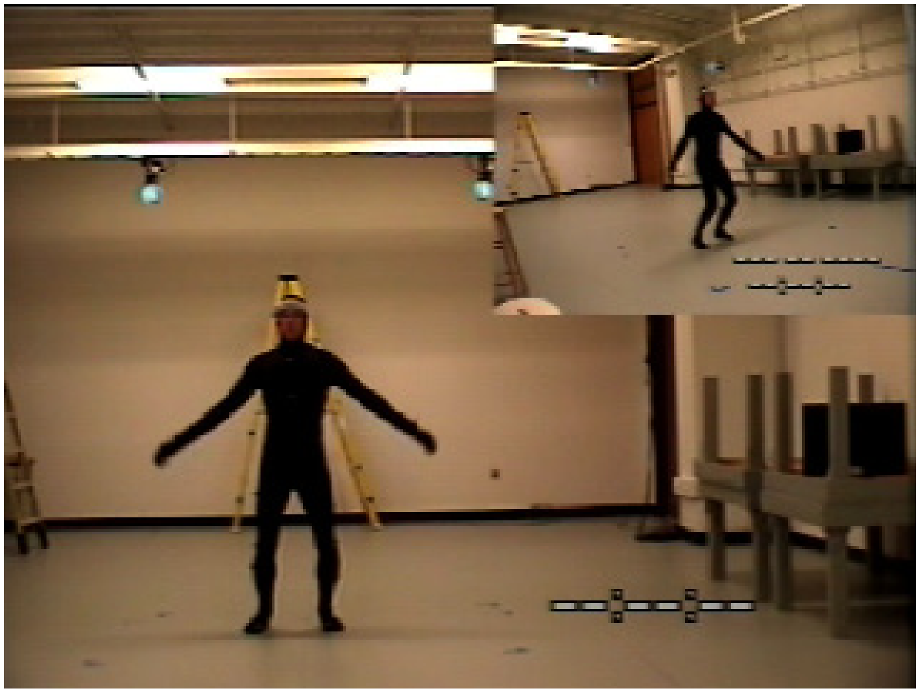




$$
\begin{aligned}
\pi_j^{(i)} \mid \mathbf{f}_i, \gamma, \kappa &\sim \mathrm{Dir}([\gamma, \ldots, \gamma, \gamma + \kappa, \gamma, \ldots] \otimes \mathbf{f}_i) \\
z_t^{(i)} &\sim \pi_{z_{t-1}^{(i)}} \\
\mathbf{y}_t^{(i)} &= \sum_{j=1}^{r} A_{j, z_t^{(i)}} \mathbf{y}_{t-j}^{(i)} + \mathbf{e}_t^{(i)}(z_t^{(i)})
\end{aligned}
$$

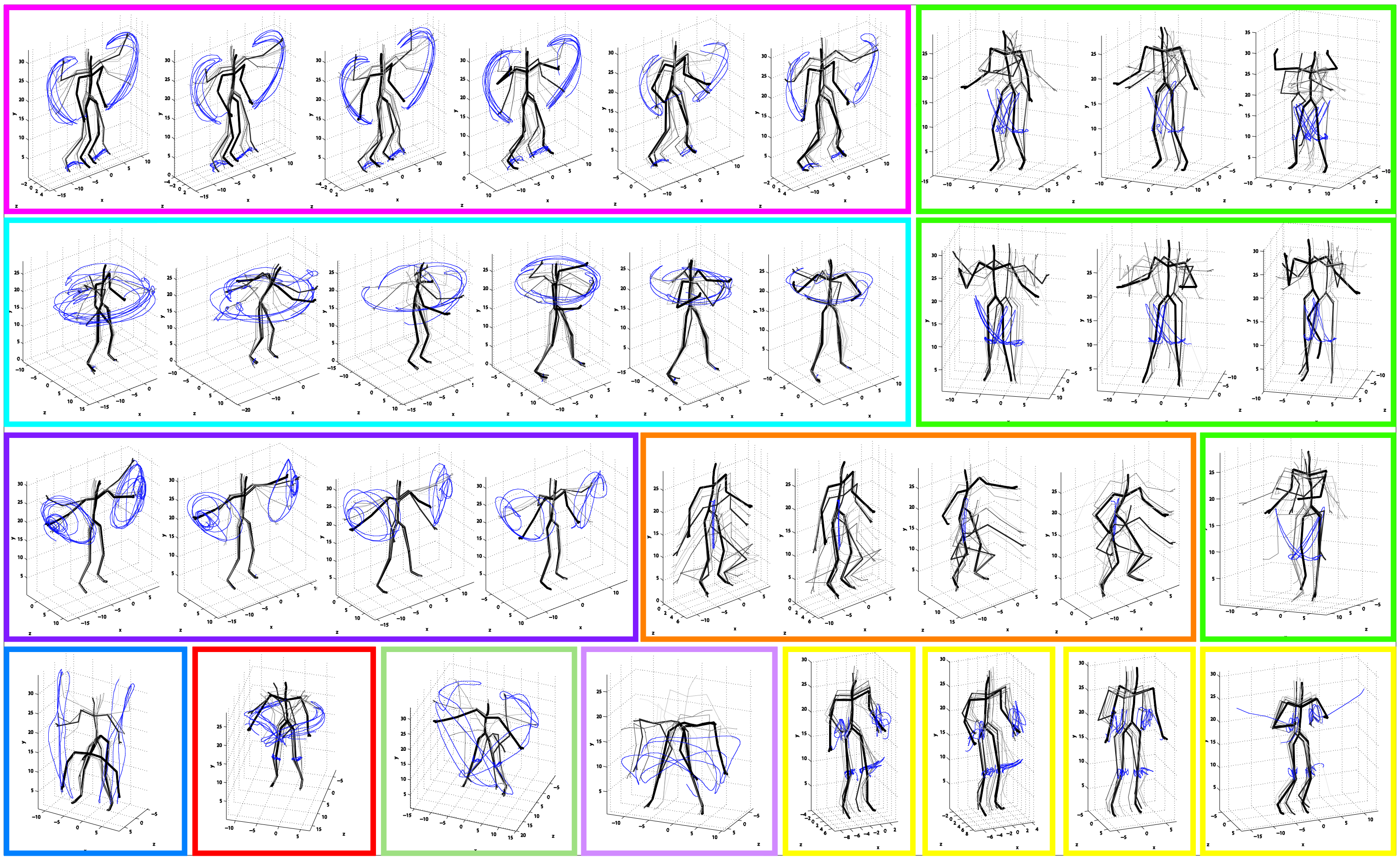
- Beta process prior:
  - ➢ Encourages sharing
  - ➢ Unbounded features

Motion Capture



6 videos of exercise routines: CMU MoCap: http://mocap.cs.cmu.edu/

Library of MoCap Behaviors

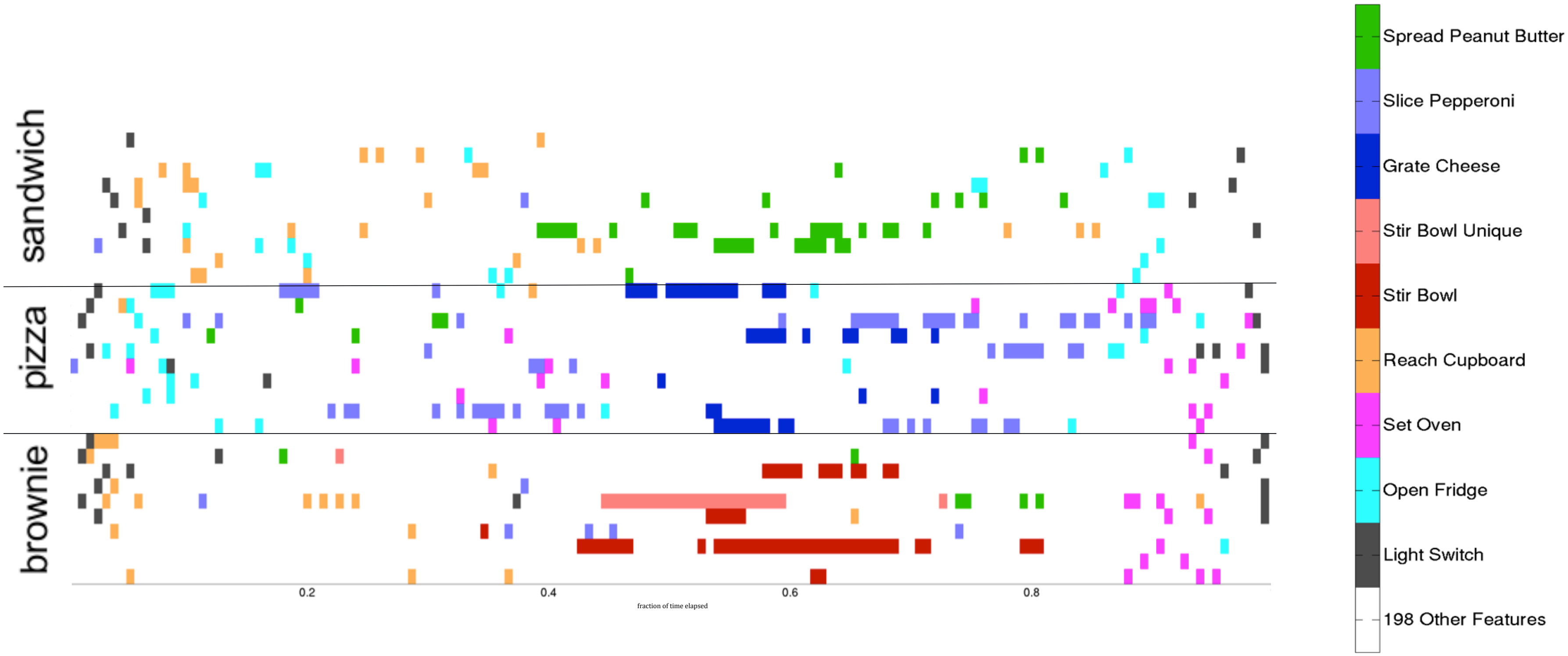


CMU Kitchen Dataset

[Torre et al. TR 10]

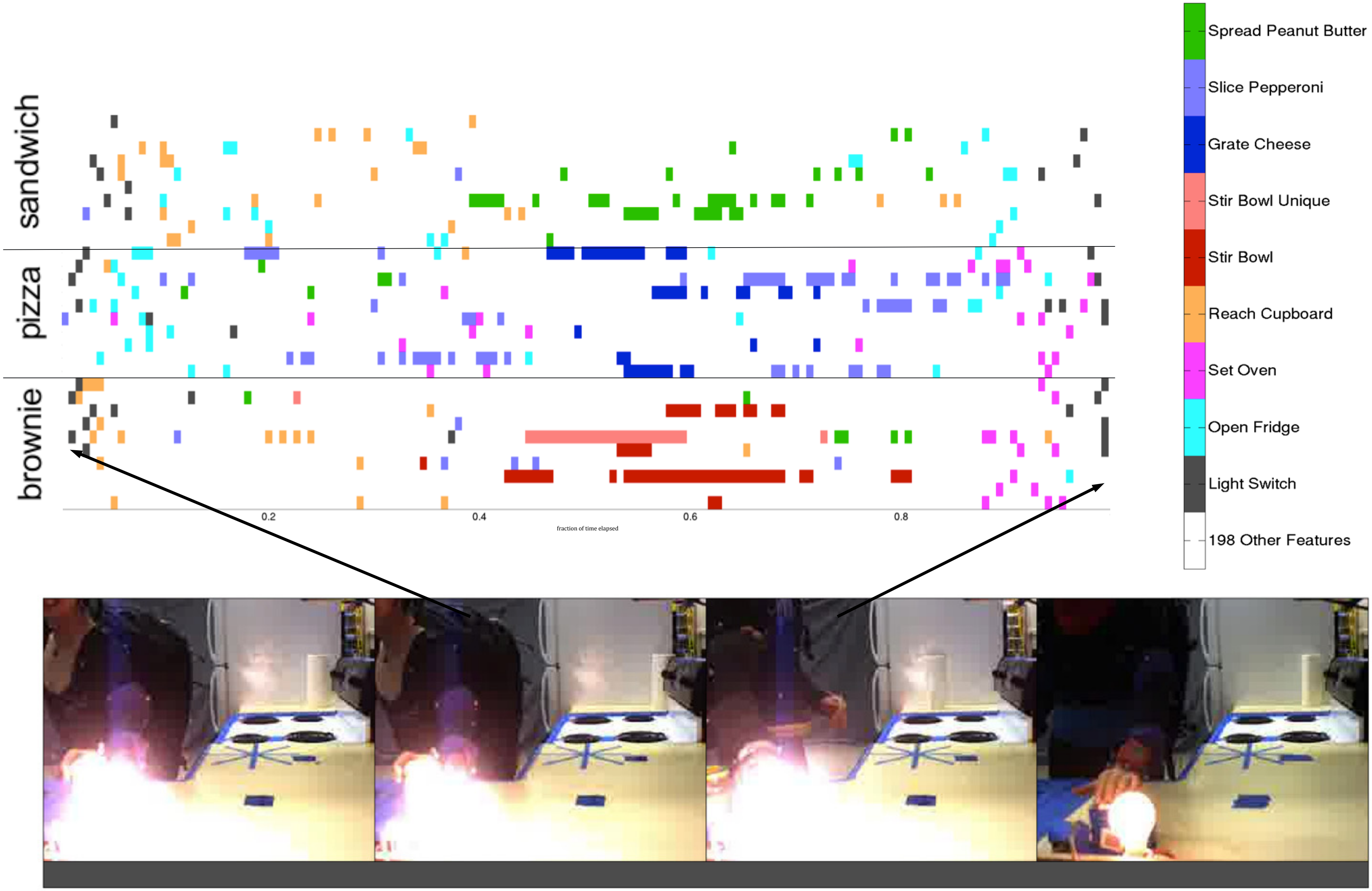
30 videos total, 10 from each recipe category. Category labels not provided to BP-HMM.
Results from Hughes & Sudderth, POCV Workshop, CVPR 2012.

Discovered Kitchen Behaviors



Locations of select behaviors across all videos

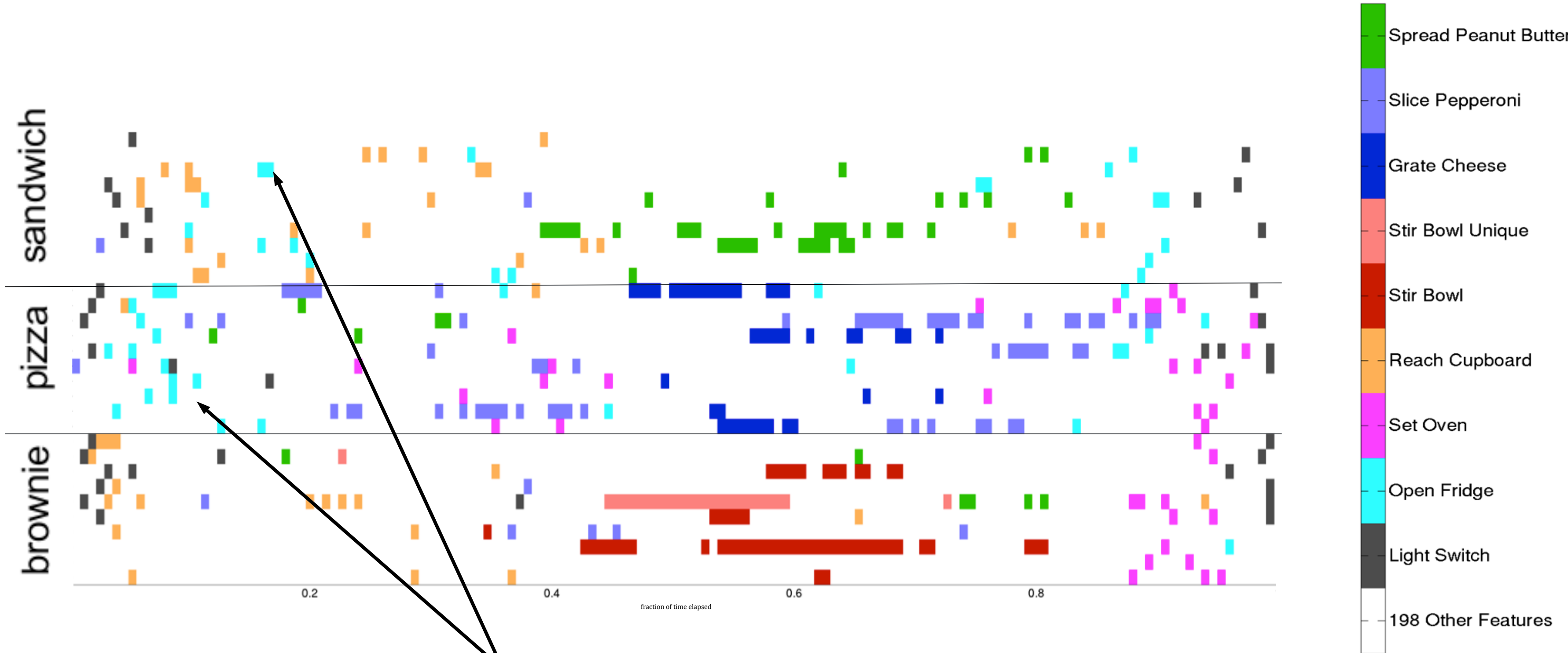
Discovered Kitchen Behaviors



Light Switch

Protocol requires switching light on/off at start and finish

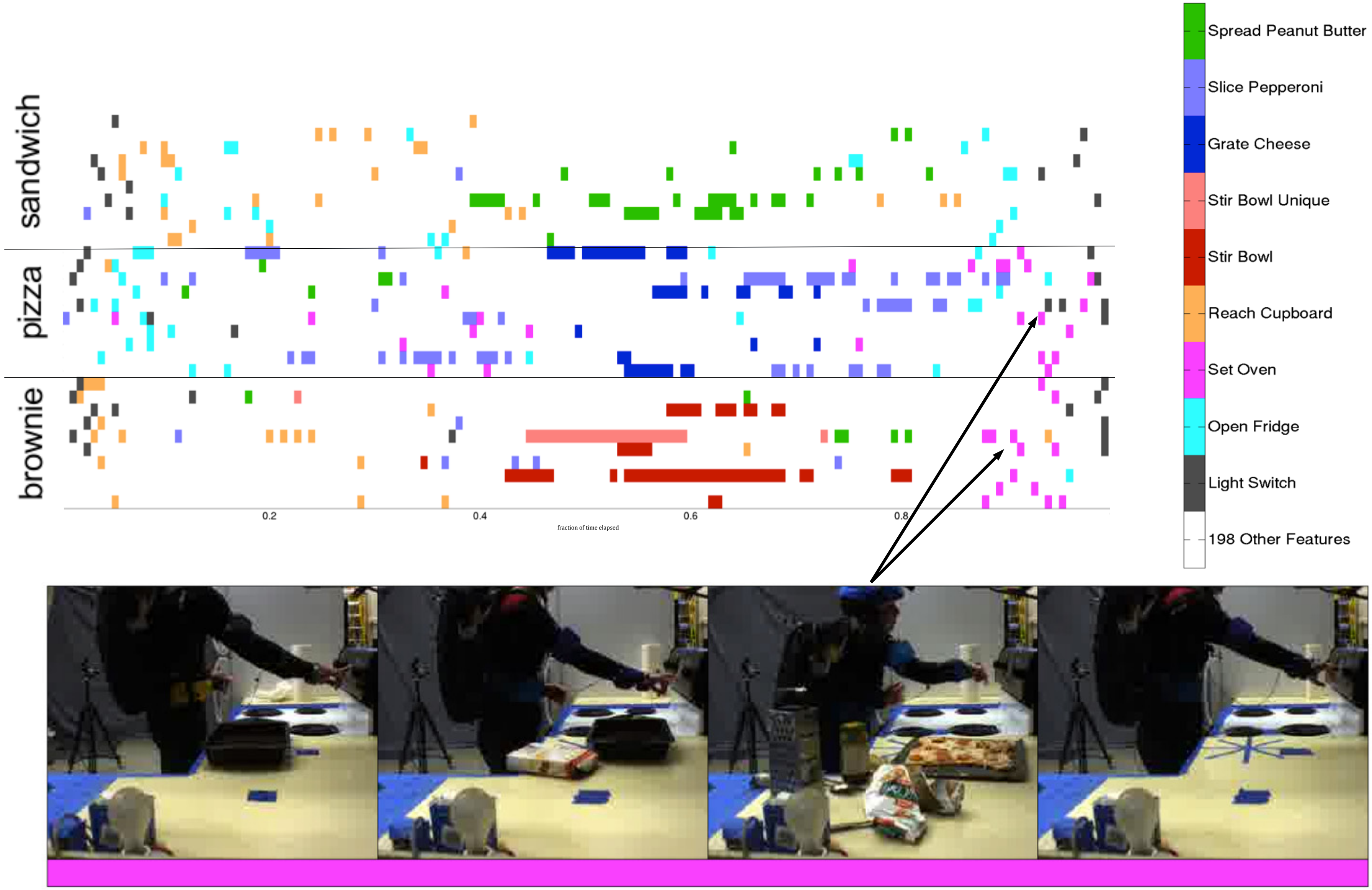
Discovered Kitchen Behaviors



Open Fridge

Pizza needs cheese, Sandwich needs jelly to begin preparation

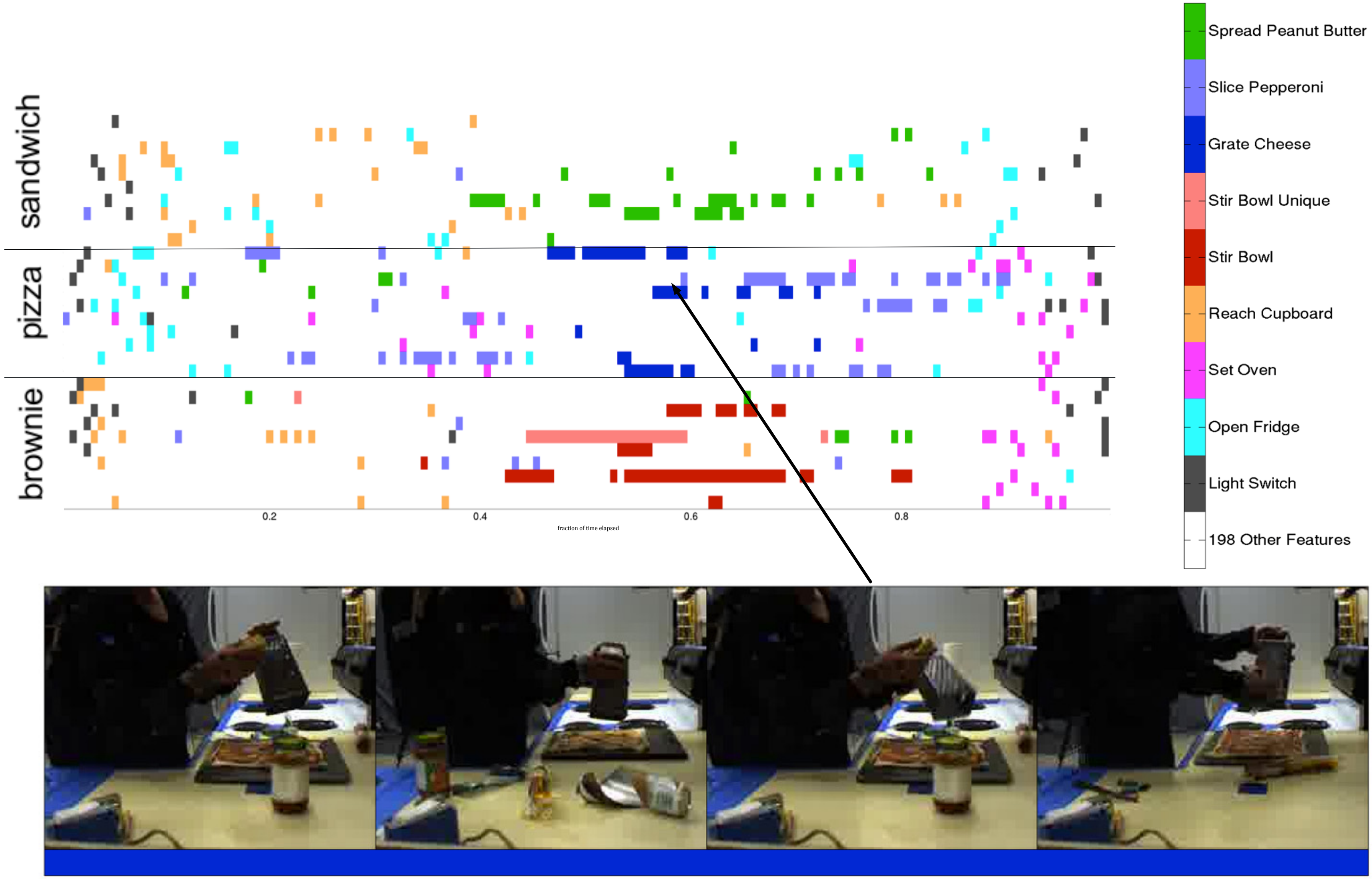
Discovered Kitchen Behaviors



Set Oven

Both Pizzas and Brownies need to be baked to conclude preparation

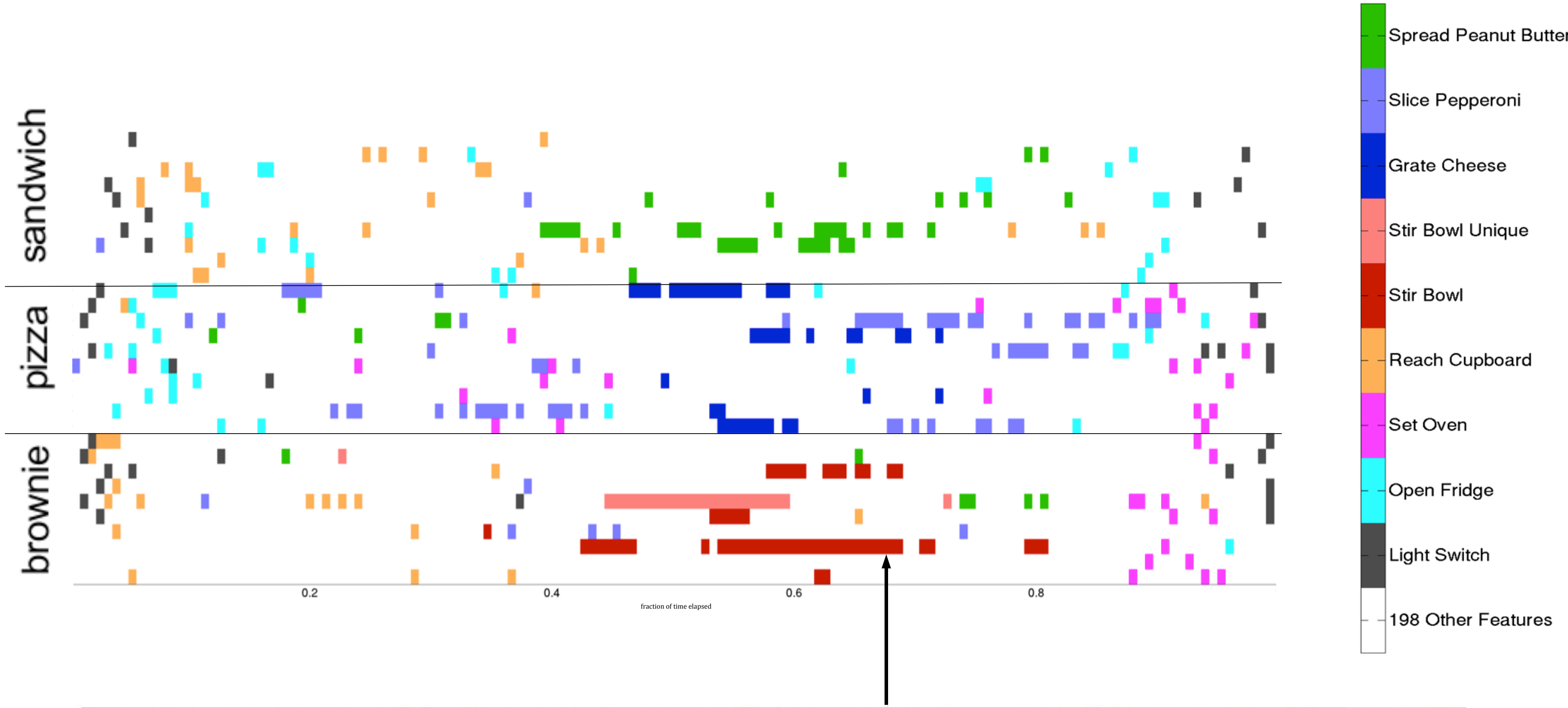
Discovered Kitchen Behaviors



Grate Cheese

Only in Pizza Videos

Discovered Kitchen Behaviors



Stir Bowl

Unique to Brownies, but multiple styles exist! Stabbing vs. swirling, etc.