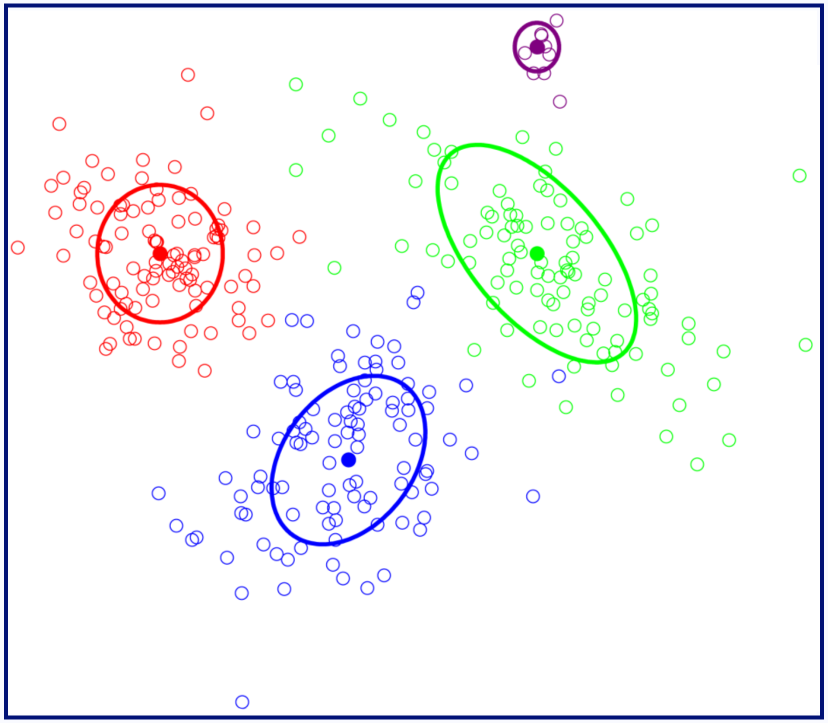
# Applied Bayesian Nonparametrics

Special Topics in Machine Learning
Brown University CSCI 2950-P, Fall 2011

November 1: Hierarchical Dirichlet Process
Hidden Markov Models & Hidden Markov Trees

# Static Clustering

- How many clusters are there?
- How should model complexity grow as more data is observed?

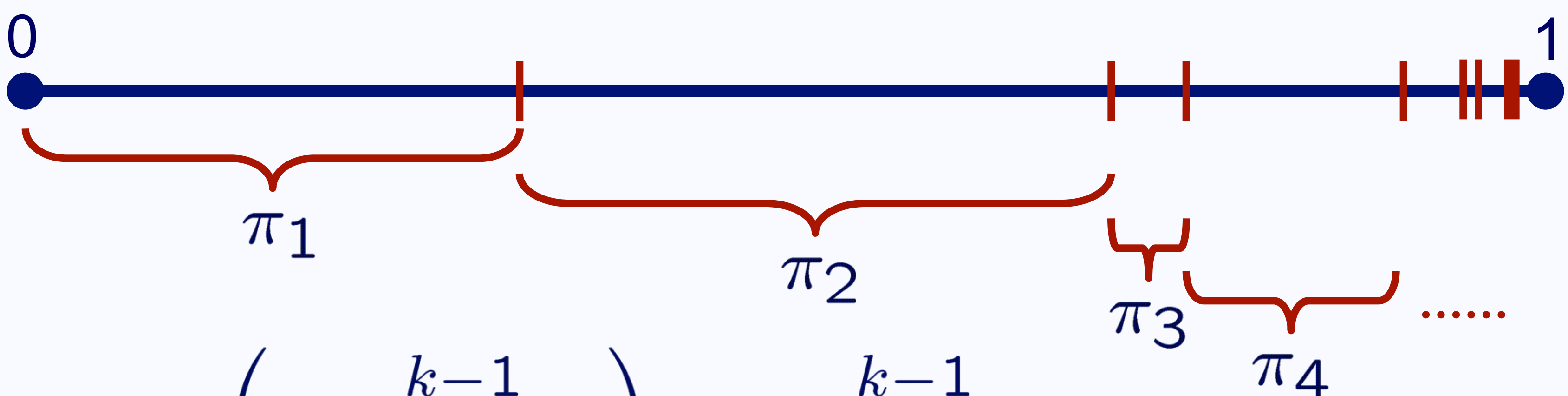


Mixture of Gaussians

# Dirichlet Process (DP) Mixtures

$$p(y) = \sum_{k=1}^{\infty} \pi_k \mathcal{N}(y \mid \mu_k, \Lambda_k)$$

- Dirichlet processes define a prior distribution on weights assigned to mixture components:



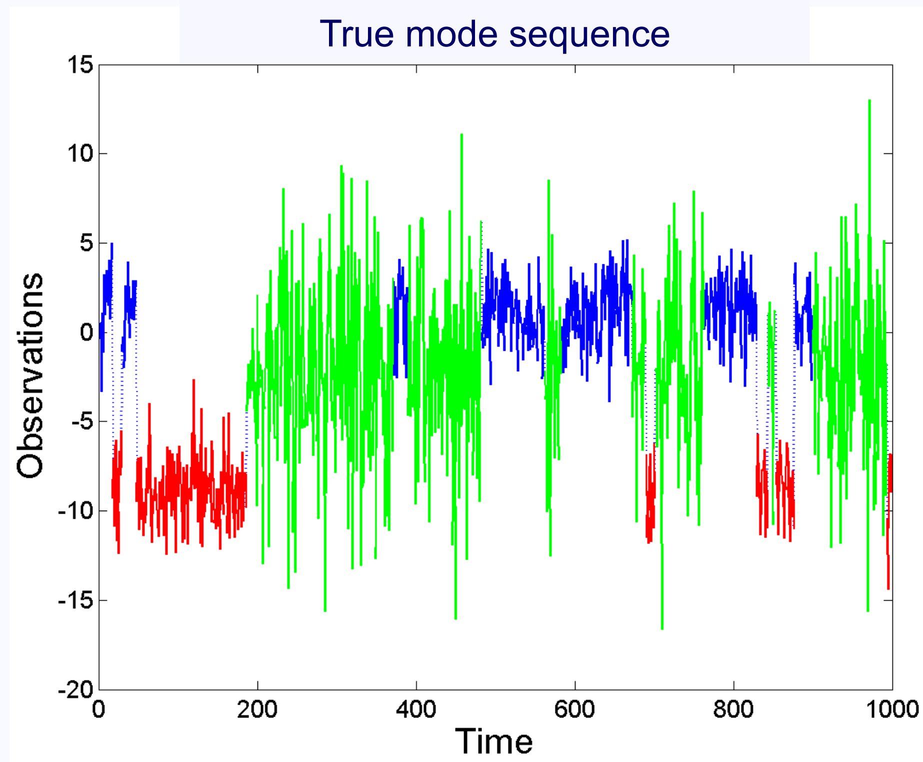

$$\pi_k = v_k \left( 1 - \sum_{\ell=1}^{k-1} \pi_\ell \right) = v_k \prod_{\ell=1}^{k-1} (1 - v_\ell)$$

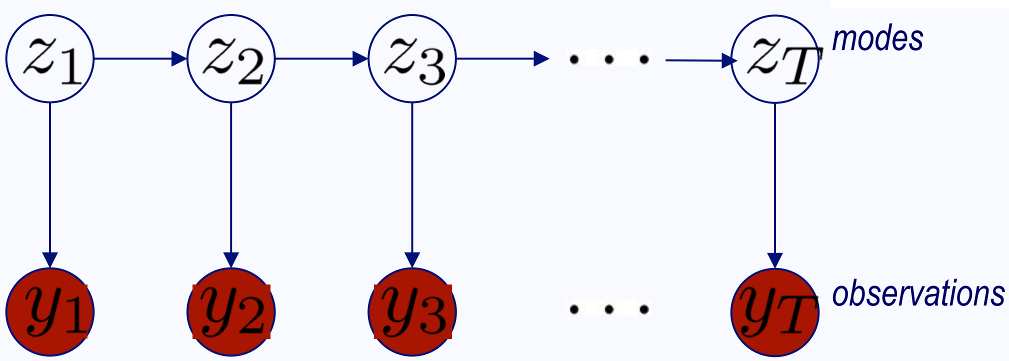
$$v_k \sim \mathrm{Beta}(1, \alpha)$$

*Ferguson 1973, Sethuraman 1994*

# Temporal Segmentation

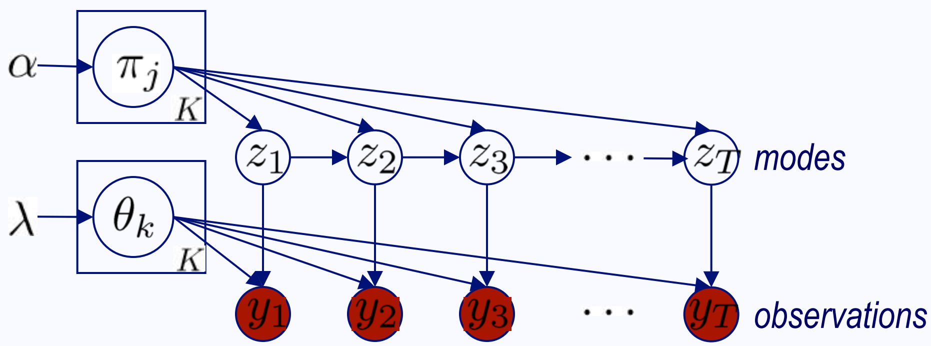
- Markov switching models for time series data
- Cluster based on underlying *mode dynamics*



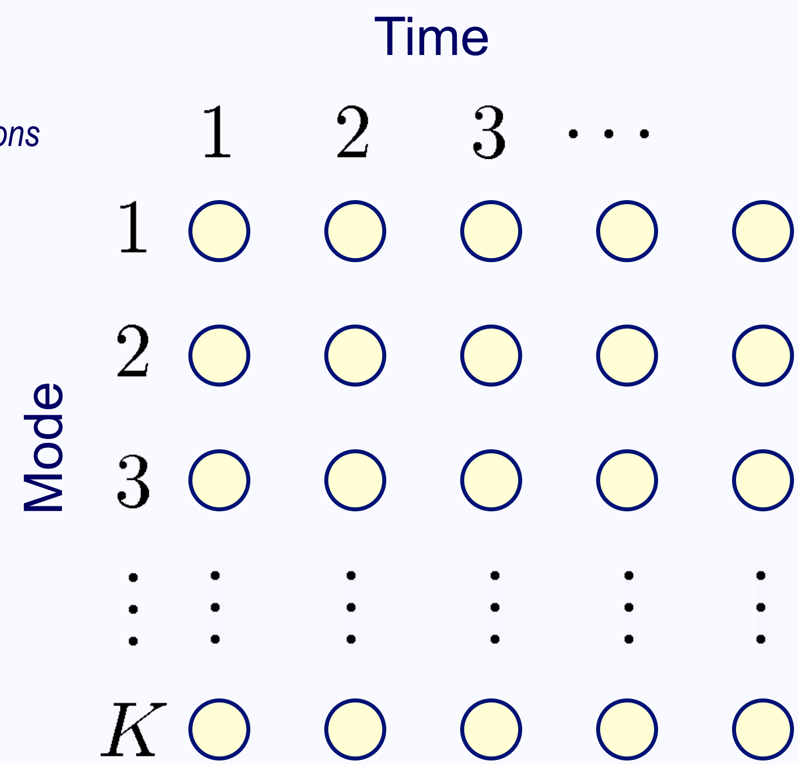

Hidden Markov Model

# Hidden Markov Models

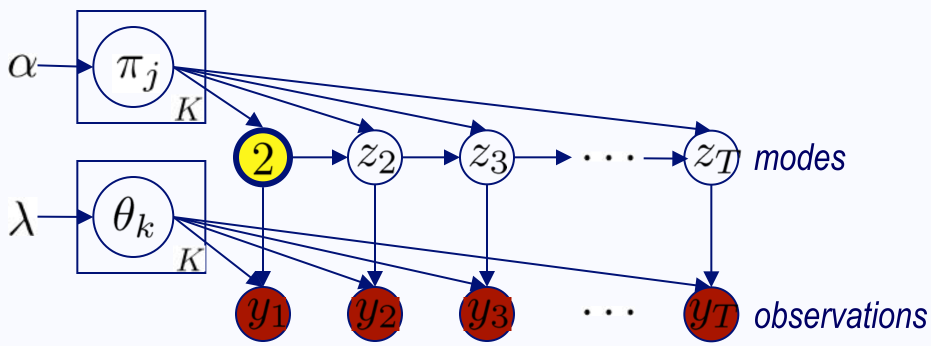


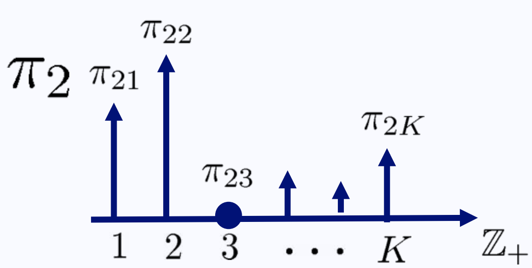

$$\begin{aligned} z_t &\sim \pi_{z_{t-1}} \\ y_t &\sim F(\theta_{z_t}) \end{aligned}$$

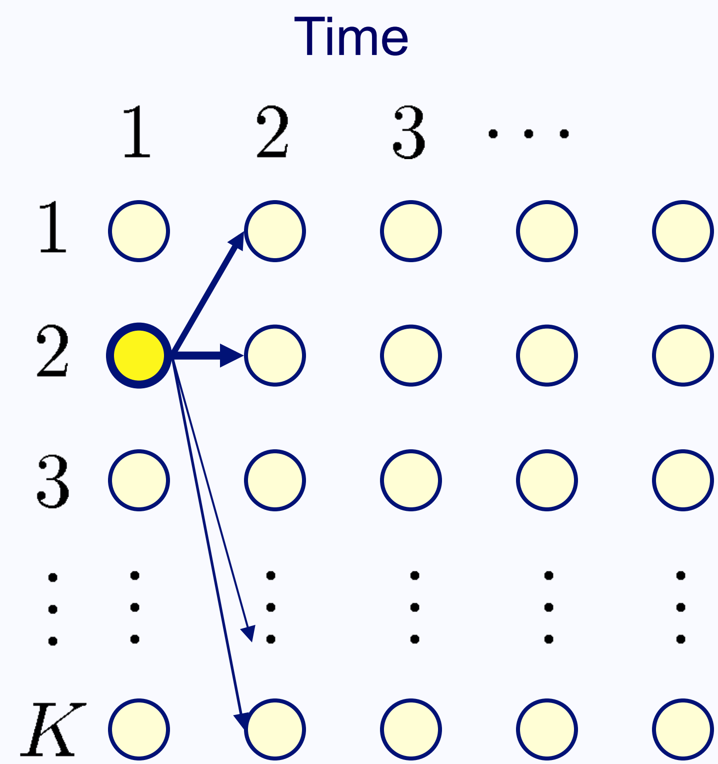
$$P = \begin{bmatrix} \text{——} & \pi_1 & \text{——} \\ \text{——} & \pi_2 & \text{——} \\ & \vdots & \\ \text{——} & \pi_K & \text{——} \end{bmatrix}$$



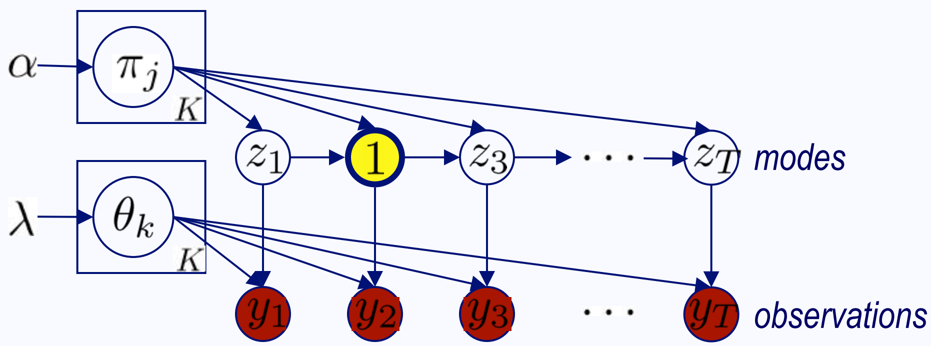
# Hidden Markov Models

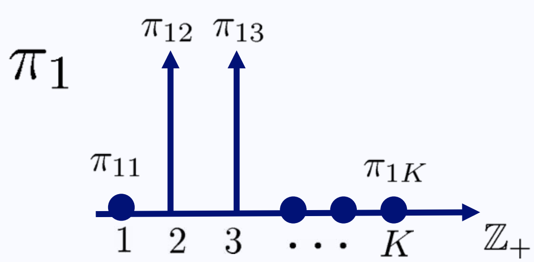


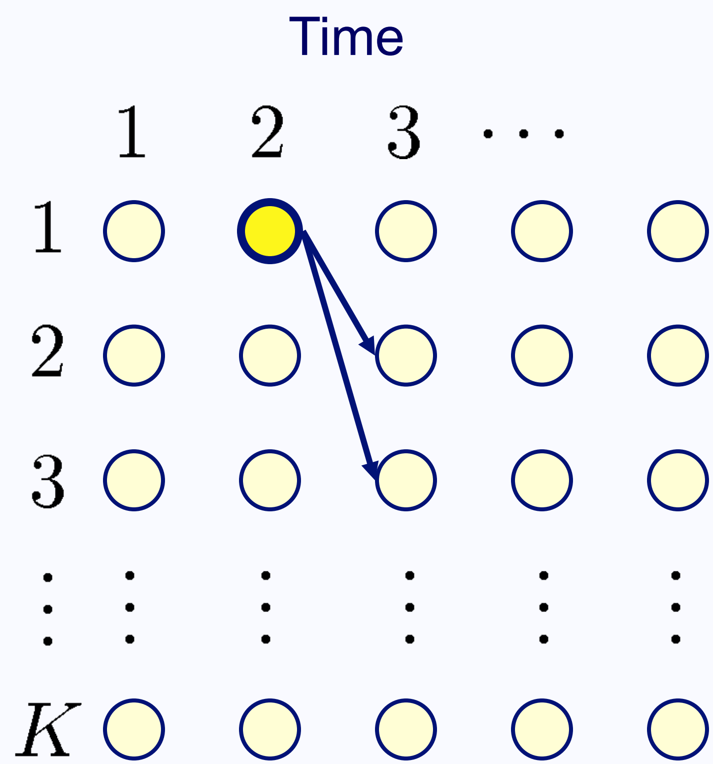



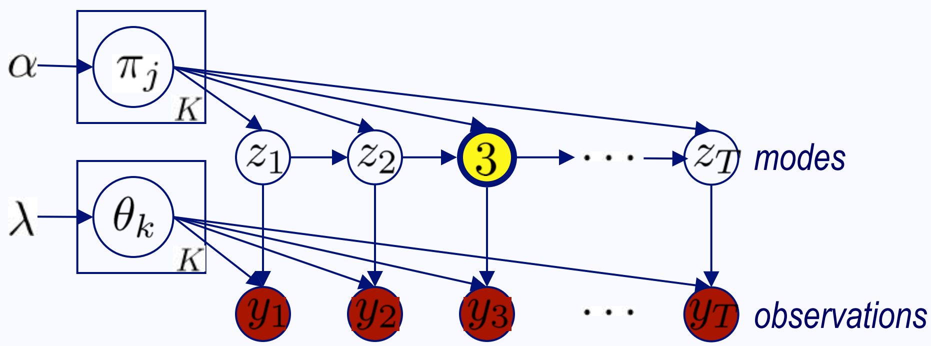
# Hidden Markov Models

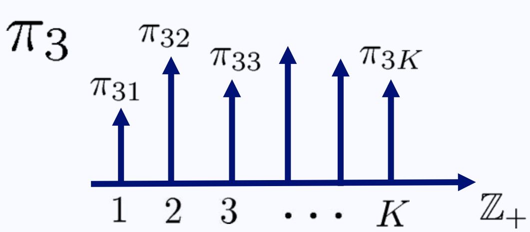


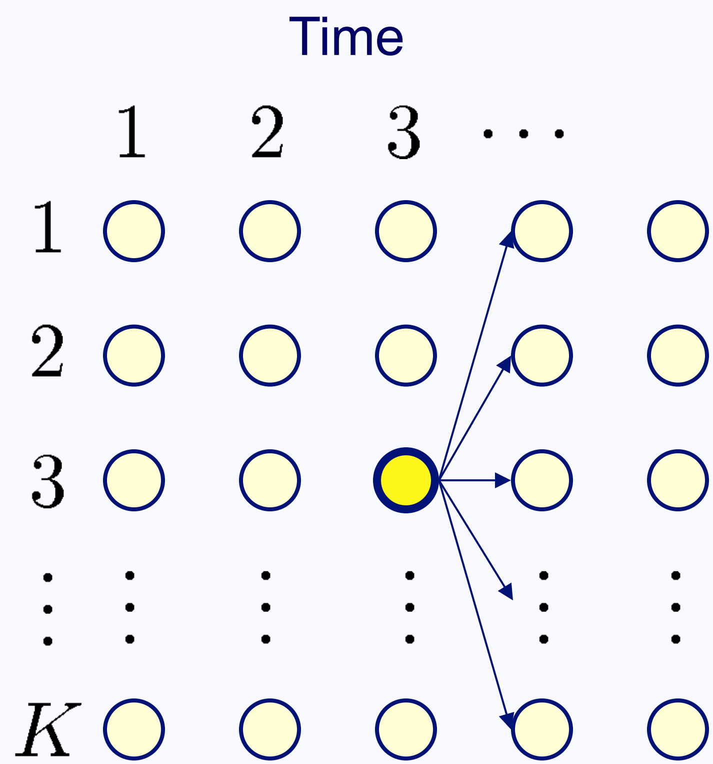



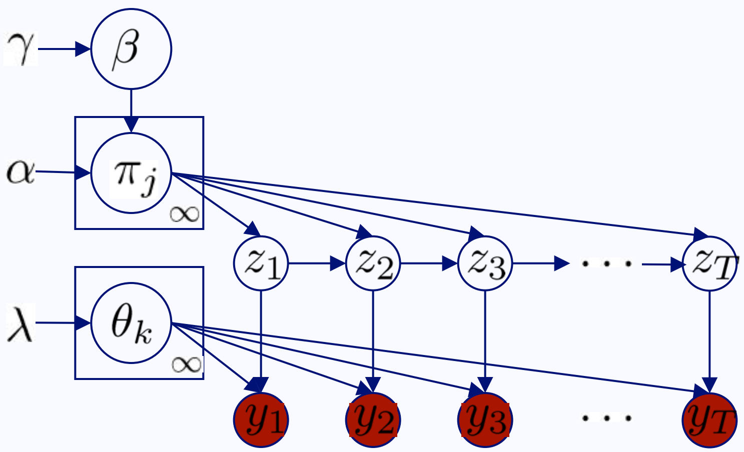
# Hidden Markov Models

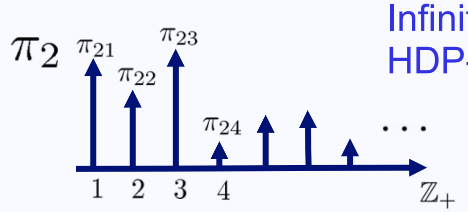


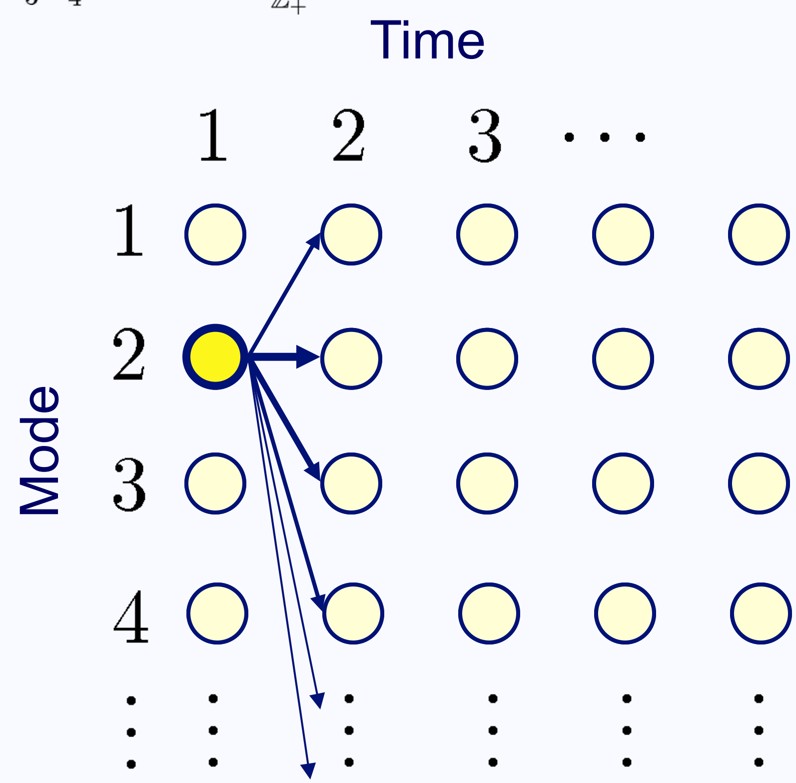



# Issue 1: How many modes?



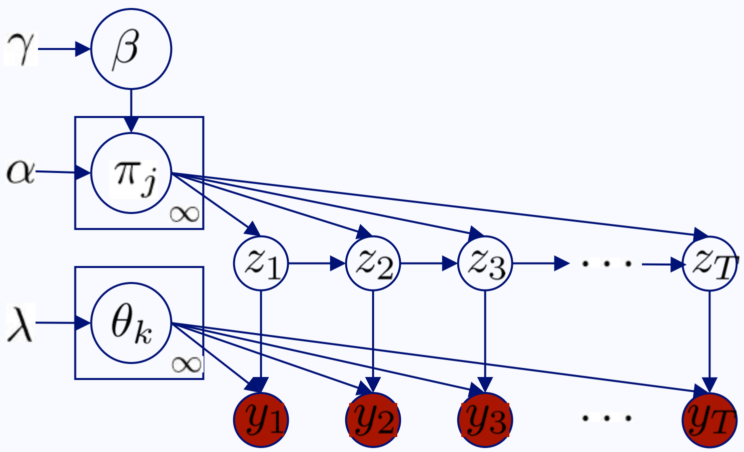

Infinite HMM: Beal, et.al., *NIPS* 2002
HDP-HMM: Teh, et. al., *JASA* 2006

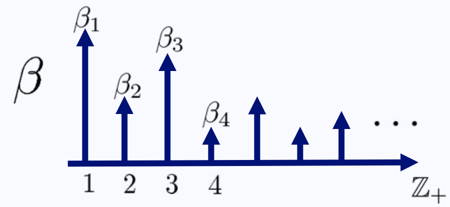


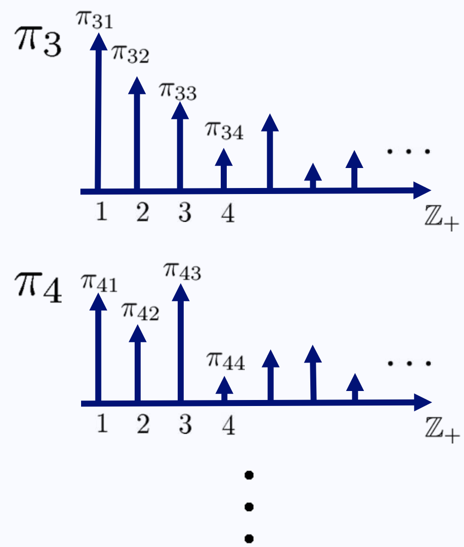

***Hierarchical Dirichlet Process HMM***

- 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:

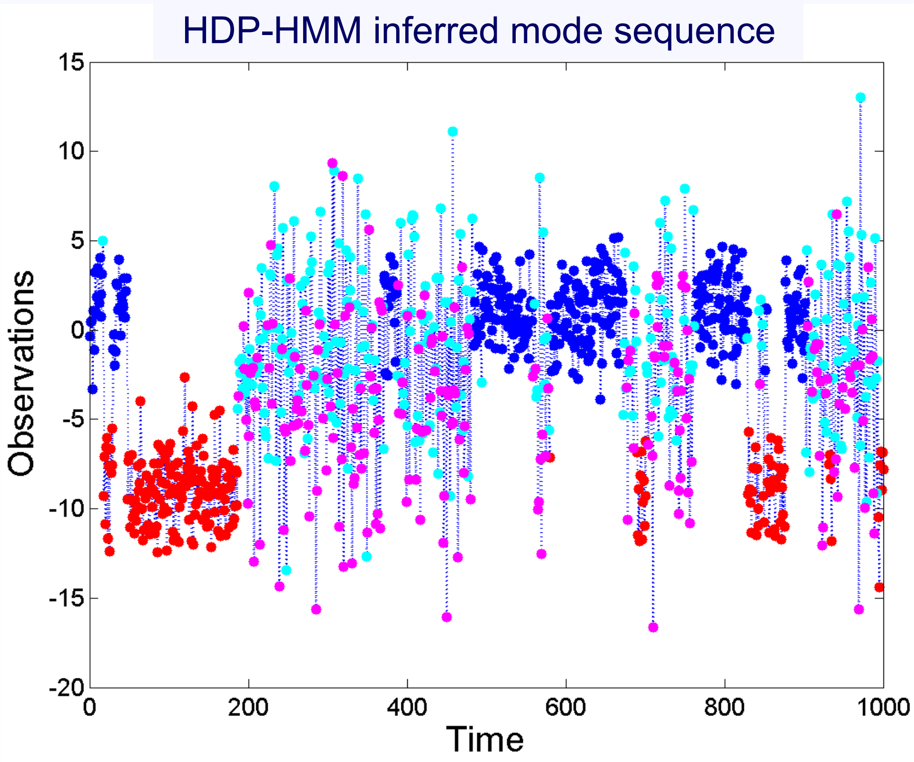
$$\beta \sim \mathrm{Stick}(\gamma)$$

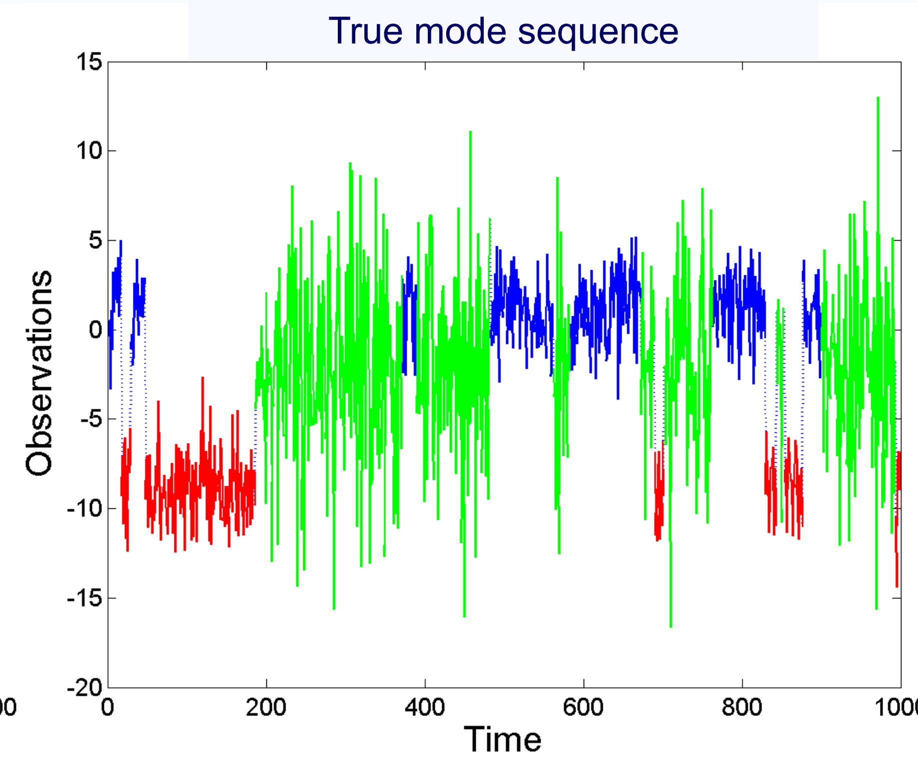
- Mode-specific transition distributions:

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

***sparsity of $\beta$ is shared*** → $E[\pi_{jk}] = \beta_k$

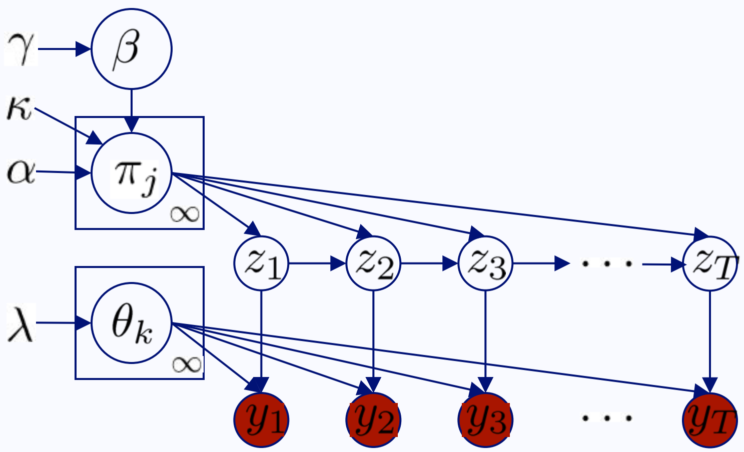
# Issue 2: Temporal Persistence

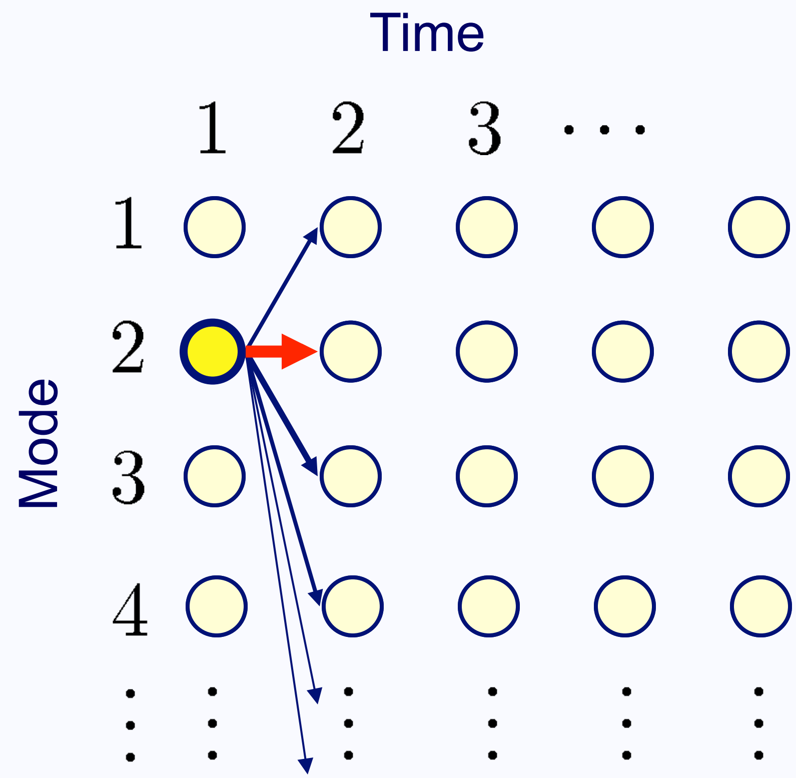


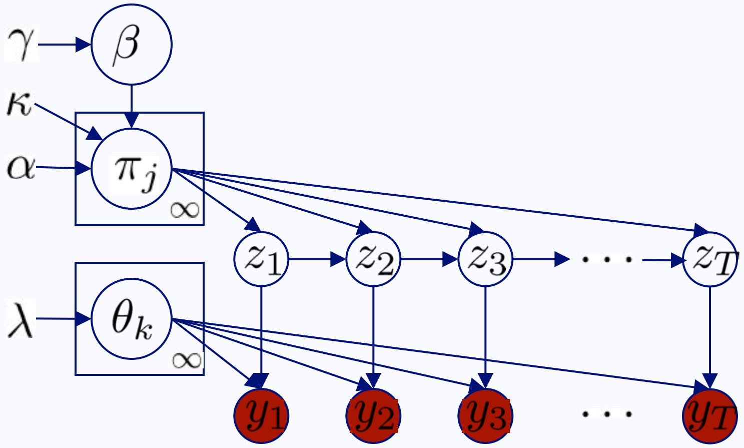

Hidden Markov Model

# "Sticky" HDP-HMM



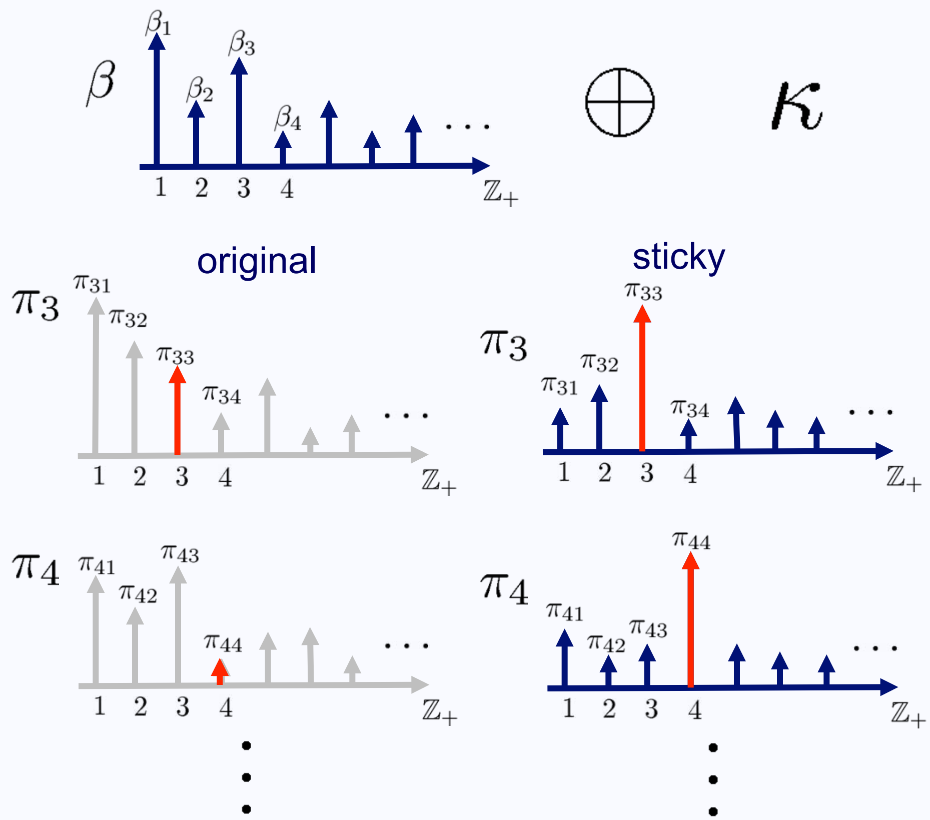



# “Sticky” HDP-HMM




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

mode-specific base measure



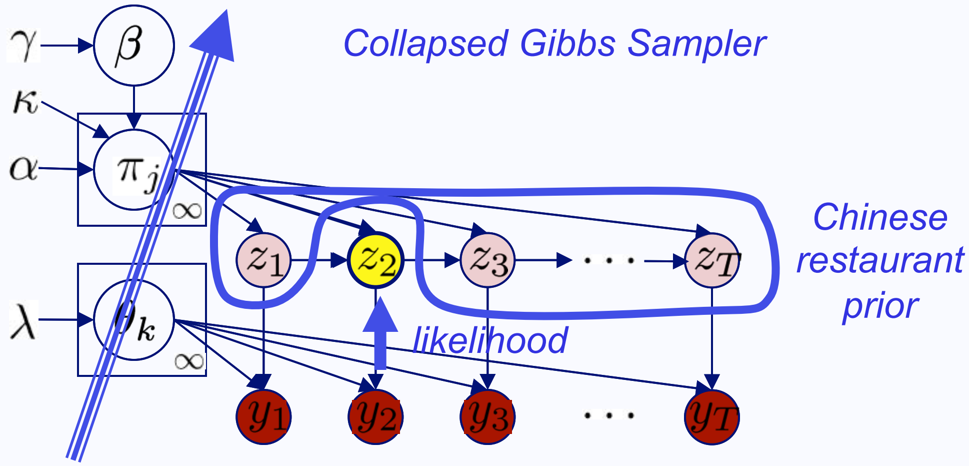

$$E[\pi_{jk}] = \beta_k$$

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

Increased probability of self-transition

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

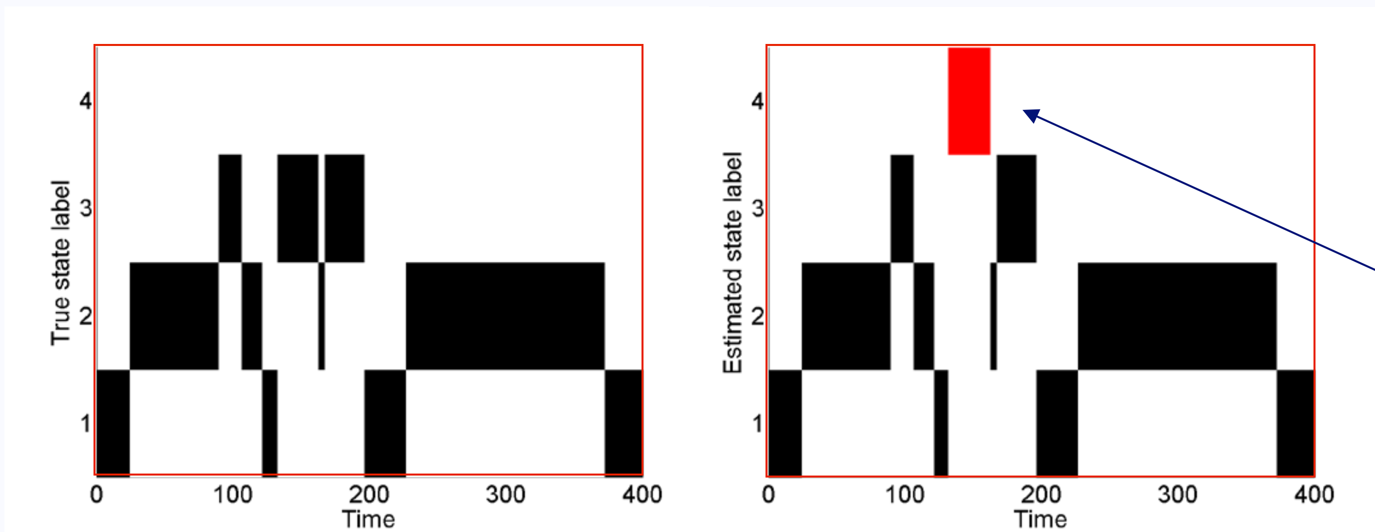
# Direct Assignment Sampler




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

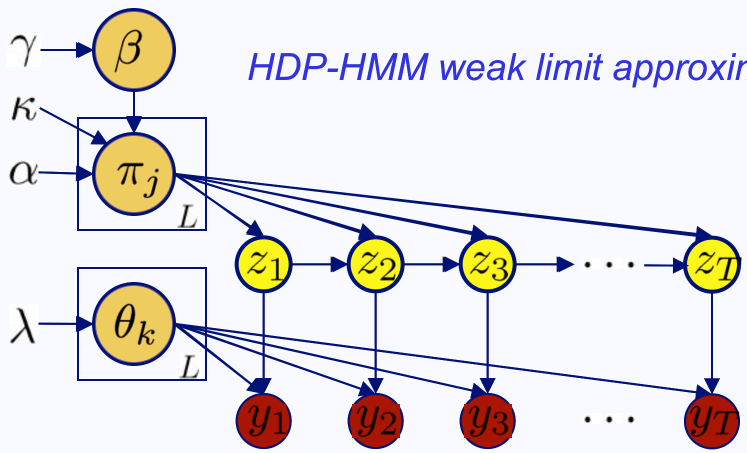
$$z_t^{(n)} \sim p(z_t | z_{\setminus t}^{(n-1)}, \alpha, \kappa) p(y_t | z, y_{\setminus t}, \lambda)$$

Conjugate base measure ⇒ closed form




Splits true mode, hard to merge

# Blocked Resampling



*HDP-HMM weak limit approximation*

- Approximate HDP:
- Compute backwards messages:
  - Average transition density
  - ($\Rightarrow$ transition densities)

$$m_{t,t-1}^{(n)}(z_{t-1}) \propto \sum_{z_t} p(z_t \mid \pi_{z_{t-1}}^{(n)}) p(y_t \mid \theta_{z_t}^{(n)}) m_{t+1,t}^{(n)}(z_t)$$

- Sample:
- Block sample $z_{1:T}^{(n)}$ as:

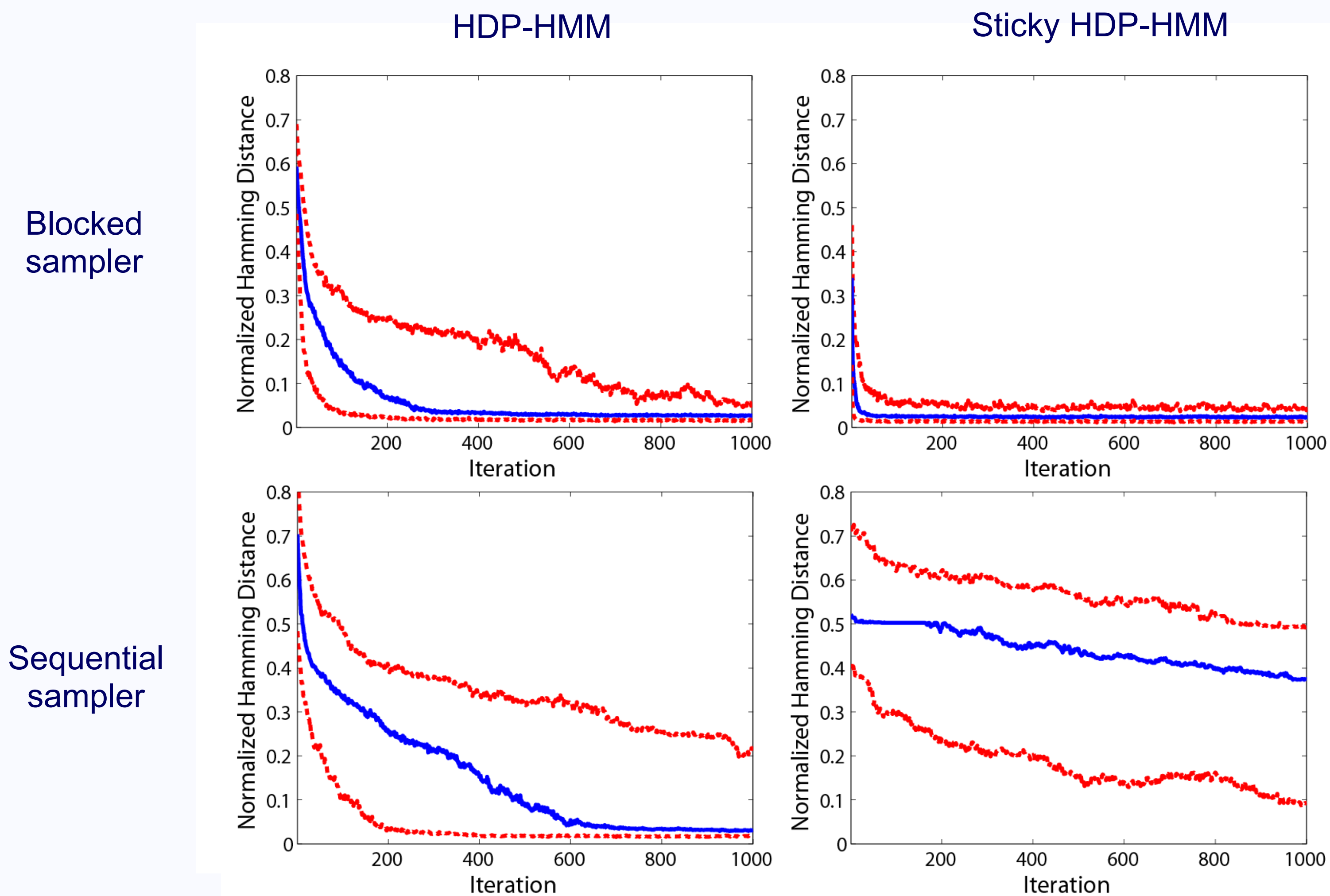
$$\beta^{(n)}, \quad \pi_j^{(n)}, \quad \theta_j^{(n)} \quad j = 1, \ldots, L$$

$$z_t^{(n)} \sim p(z_t \mid \pi_{z_{t-1}^{(n)}}^{(n)}) p(y_t \mid \theta_{z_t}^{(n)}) m_{t+1,t}^{(n)}(z_t)$$

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

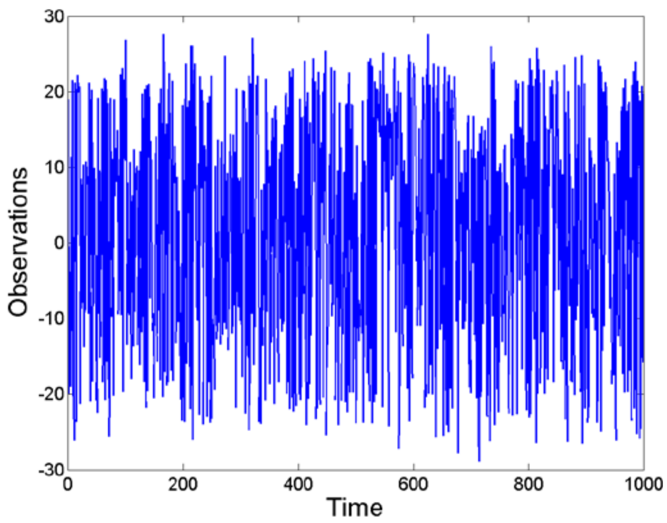
$$\pi_j \sim \mathrm{Dir}(\alpha\beta_1, \ldots, \alpha\beta_j + \kappa, \ldots, \alpha\beta_L)$$

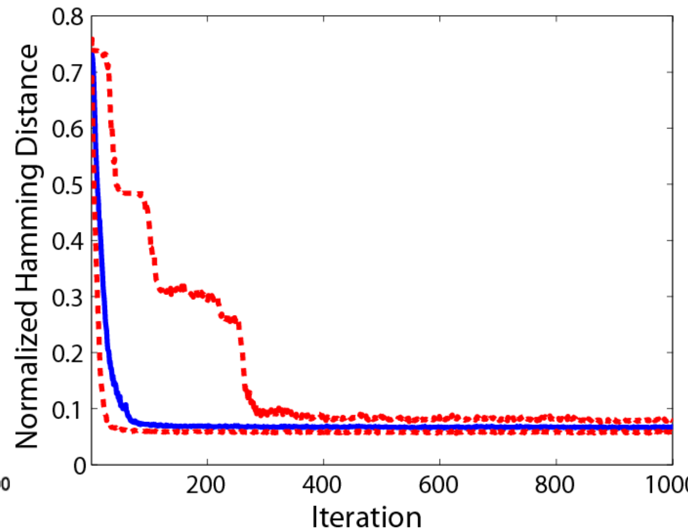
# Results: Gaussian Emissions



# Results: Fast Switching

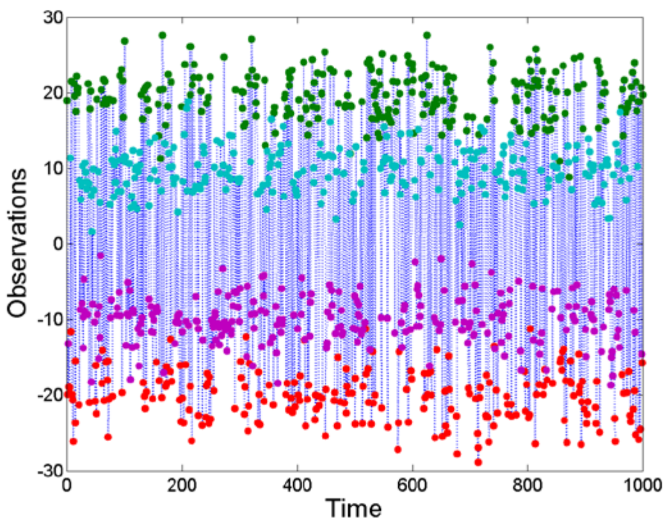
Observations

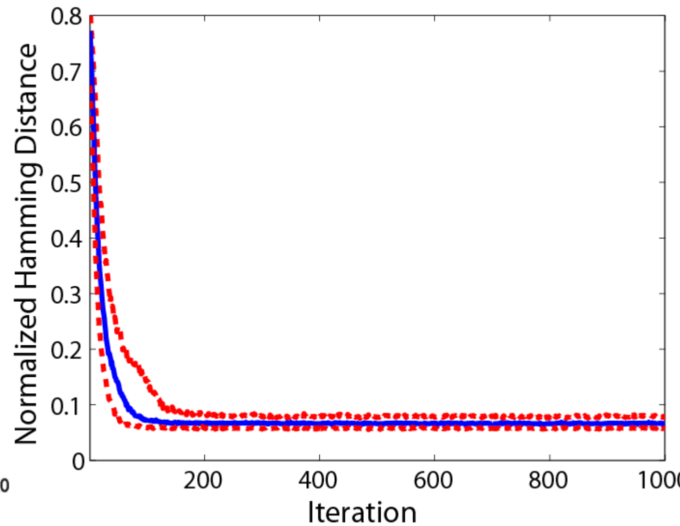



Sticky
HDP-HMM

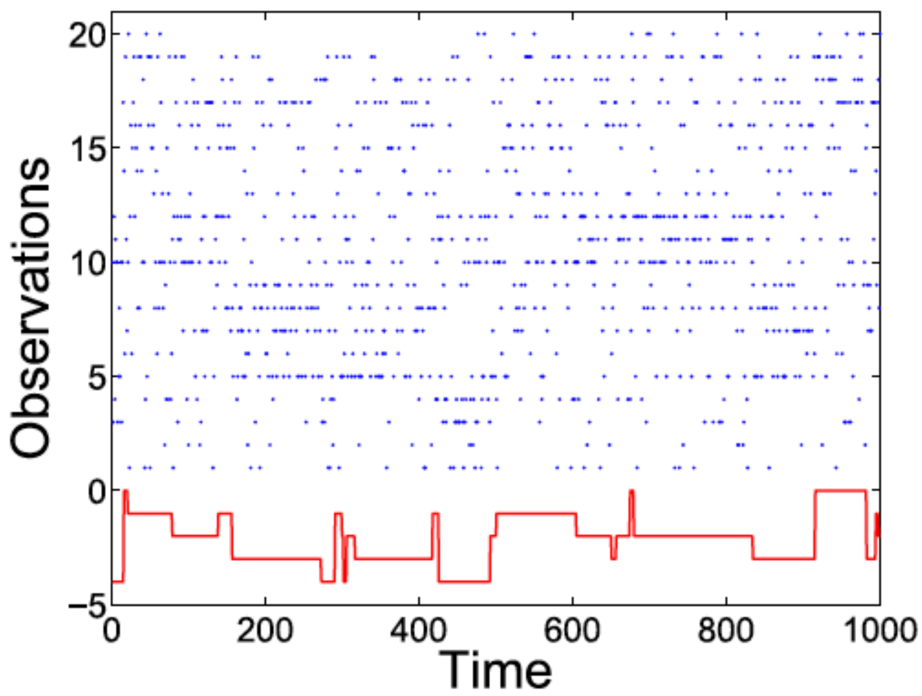
True mode
sequence

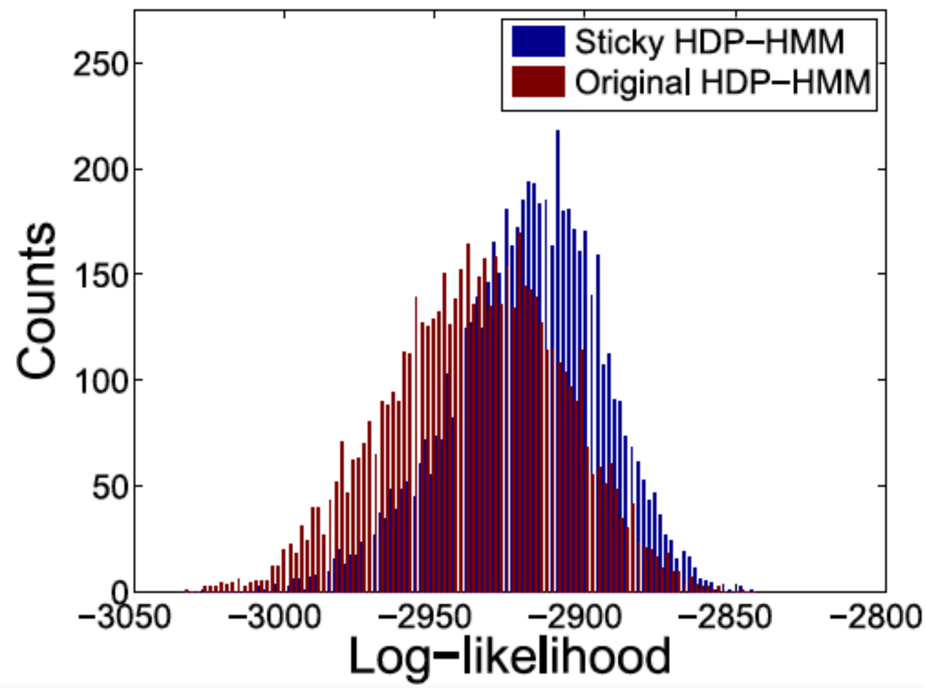



HDP-HMM

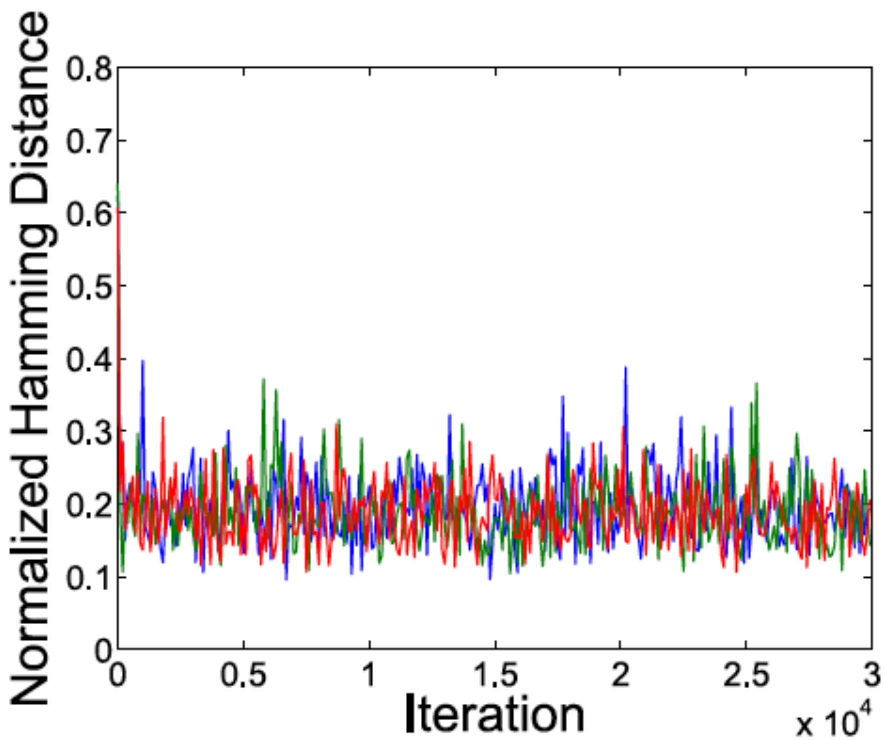
# Results: Discrete Data



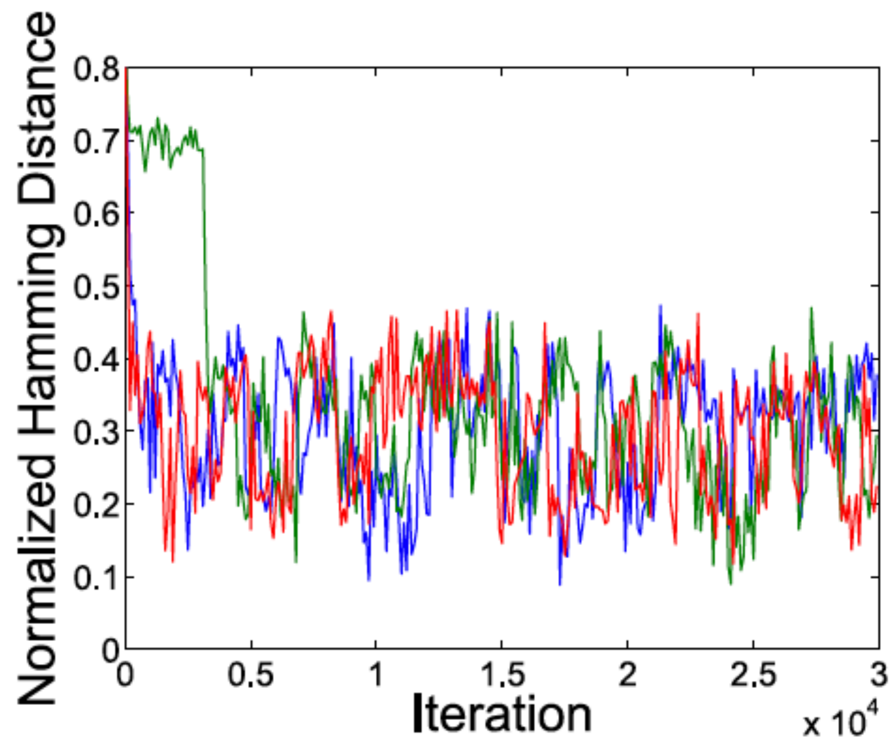




Test Predictions

# Results: Discrete Data

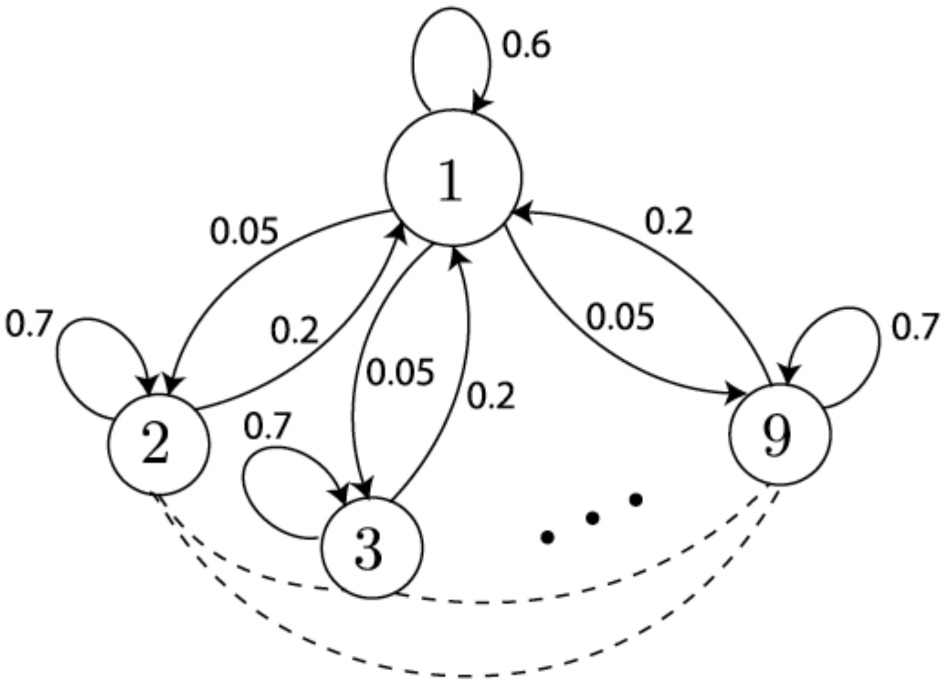


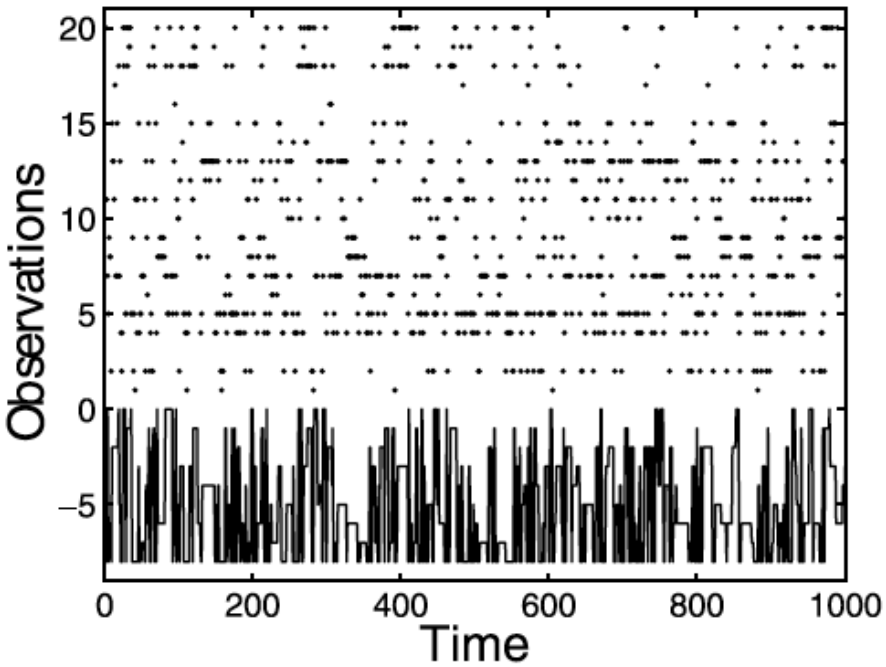

Sticky HDP-HMM



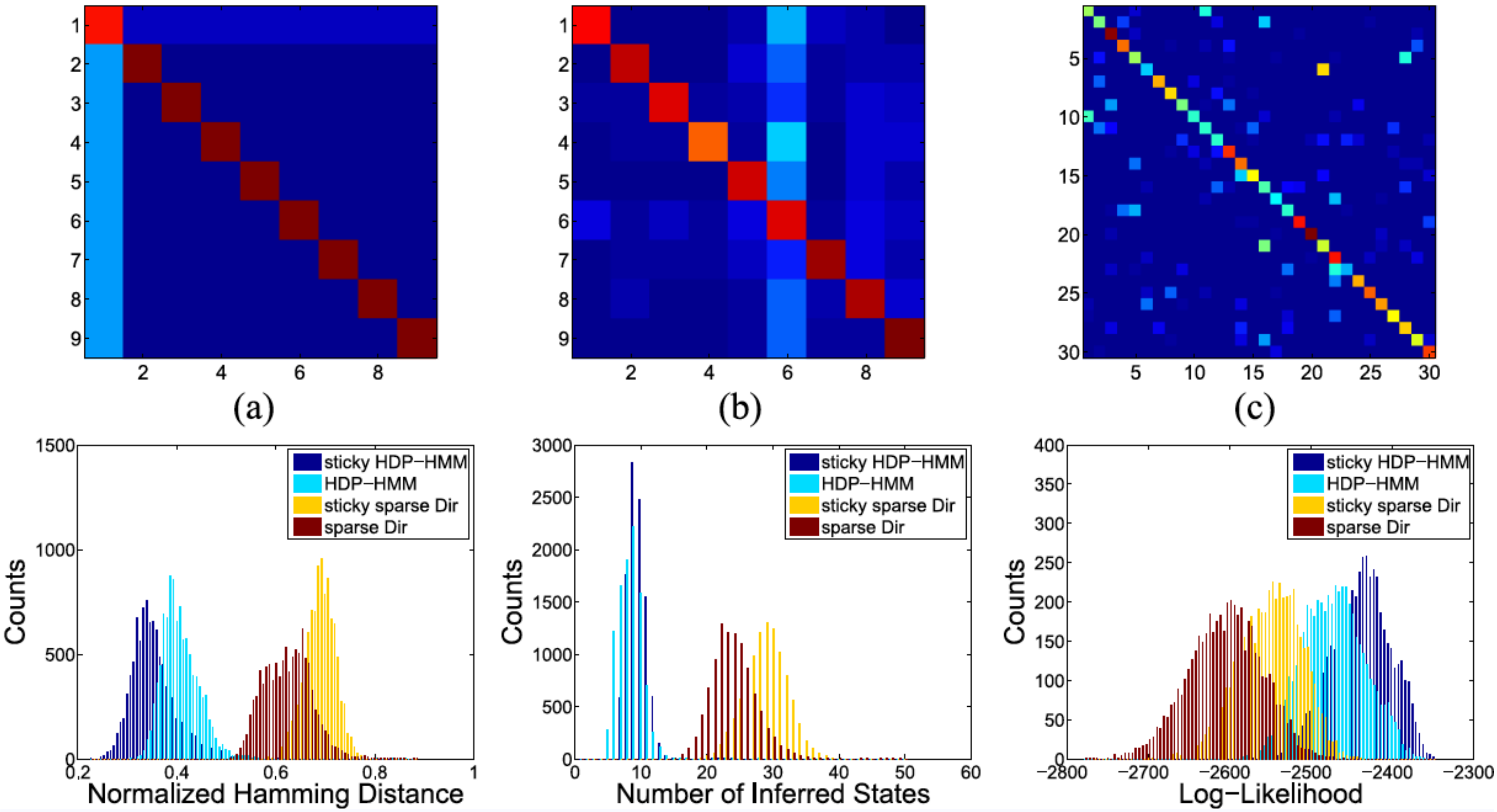

Non-sticky HDP-HMM

# Why a Global Base Measure?



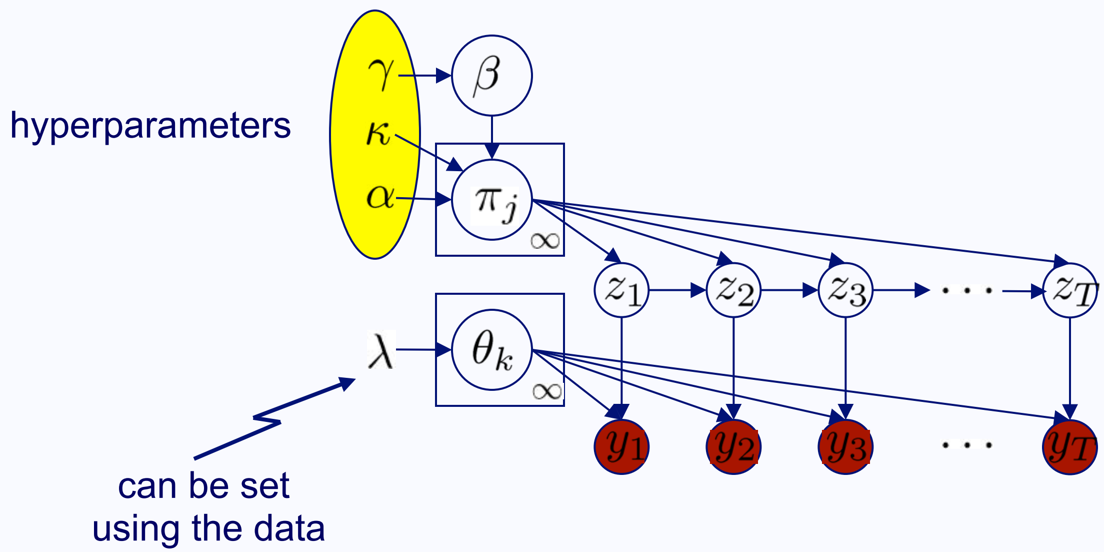



# Why a Global Base Measure?



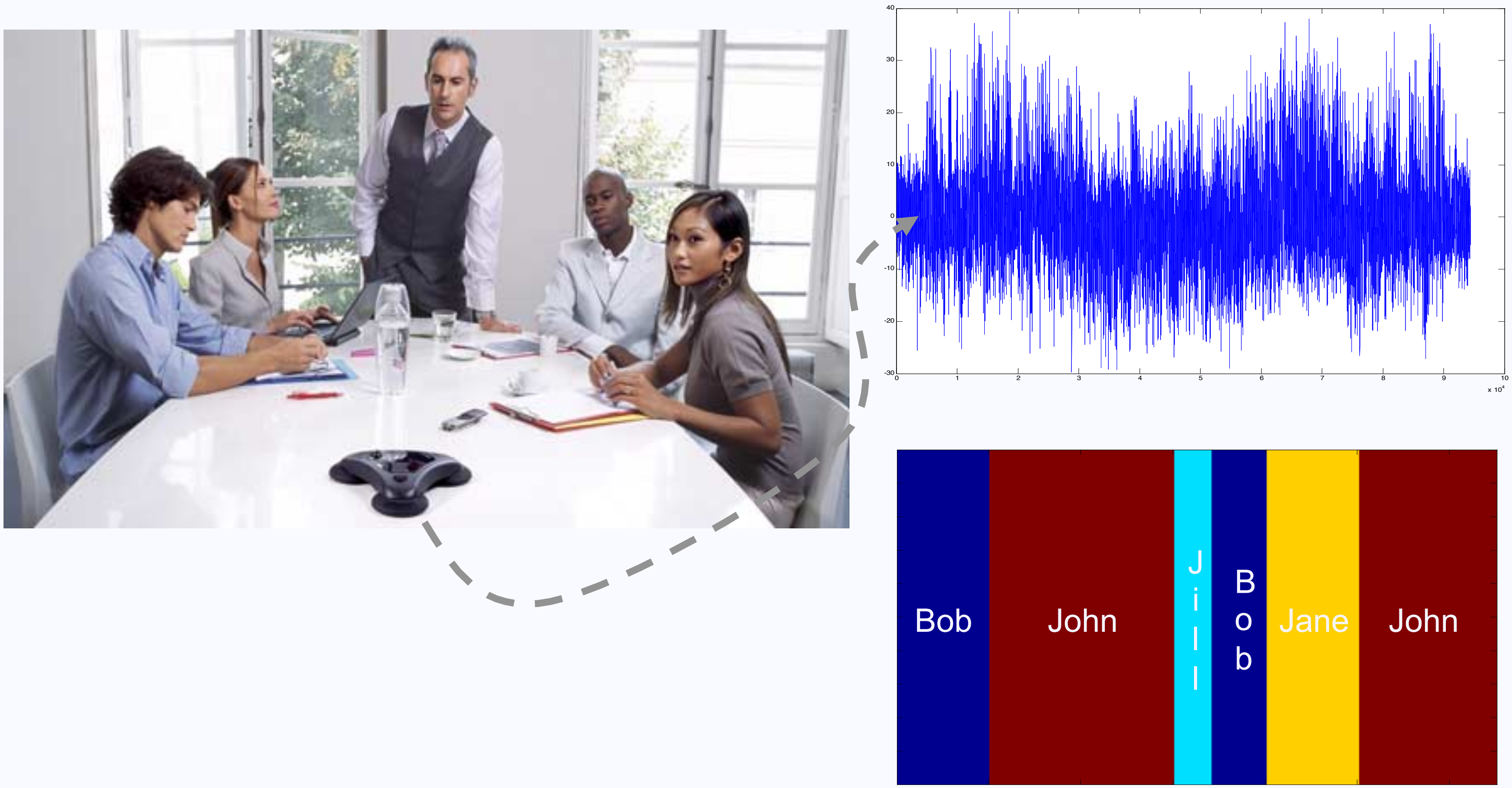
# Hyperparameters

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

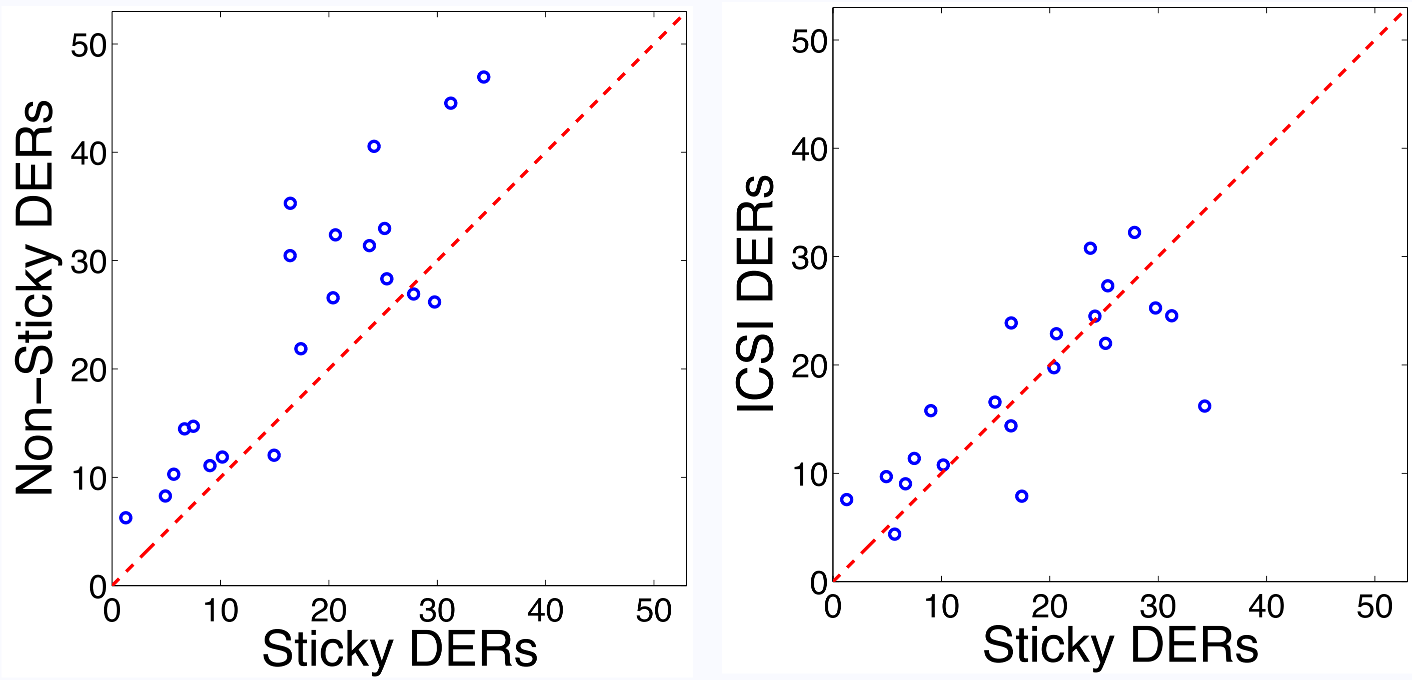



Related self-transition parameter:
Beal, et.al., *NIPS* 2002

# Speaker Diarization

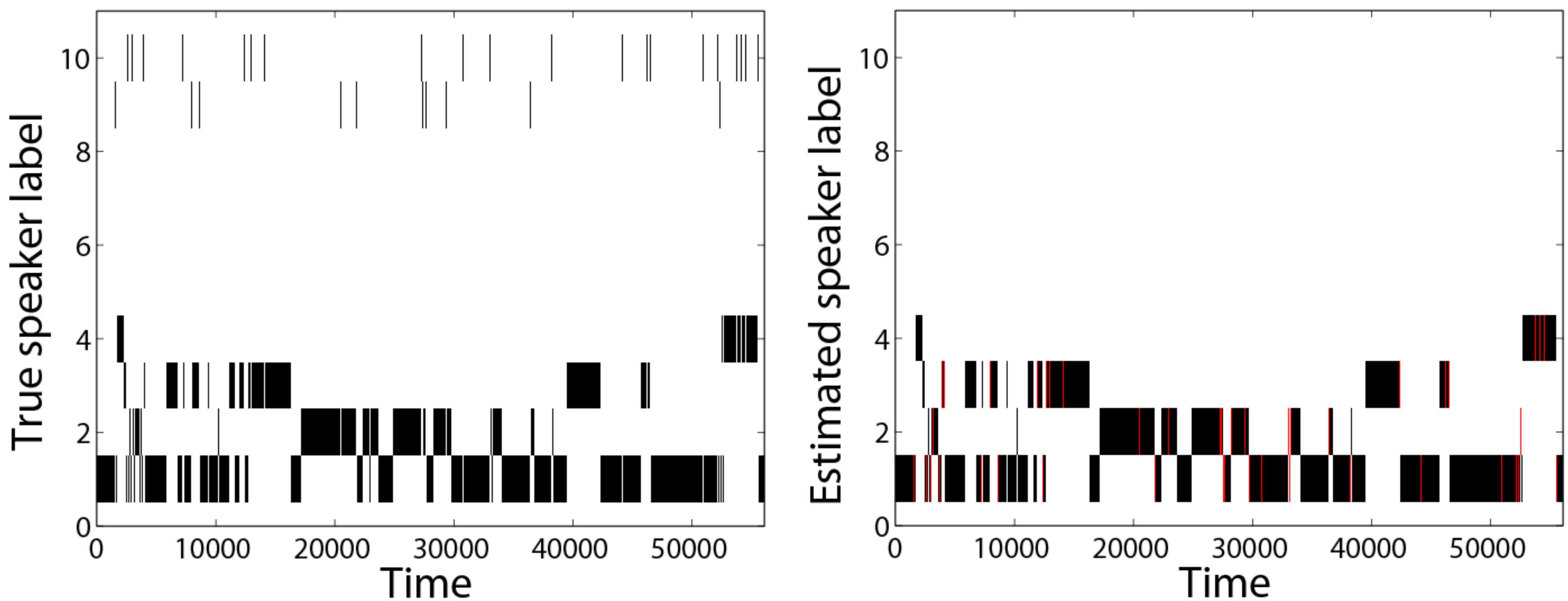


# Results: 21 meetings




| | 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% |

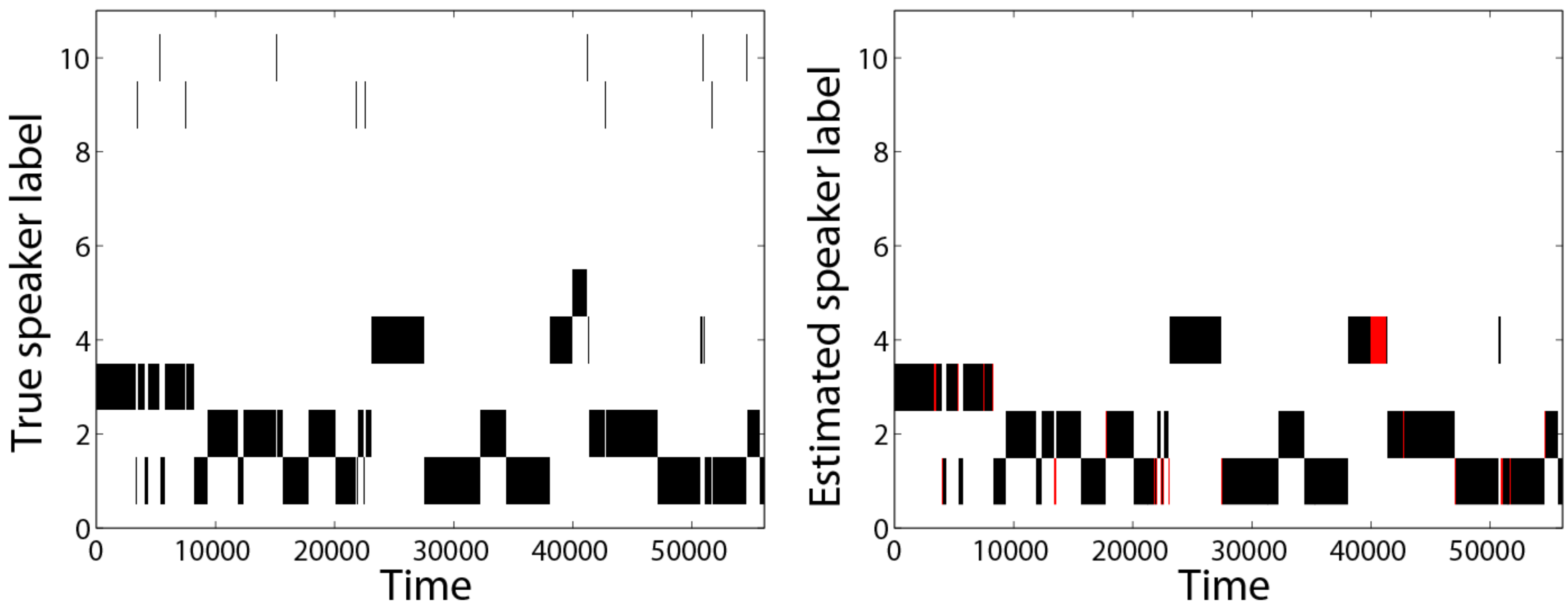
# Results: Meeting 1




Sticky DER = 1.26%

ICSI DER = 7.56%

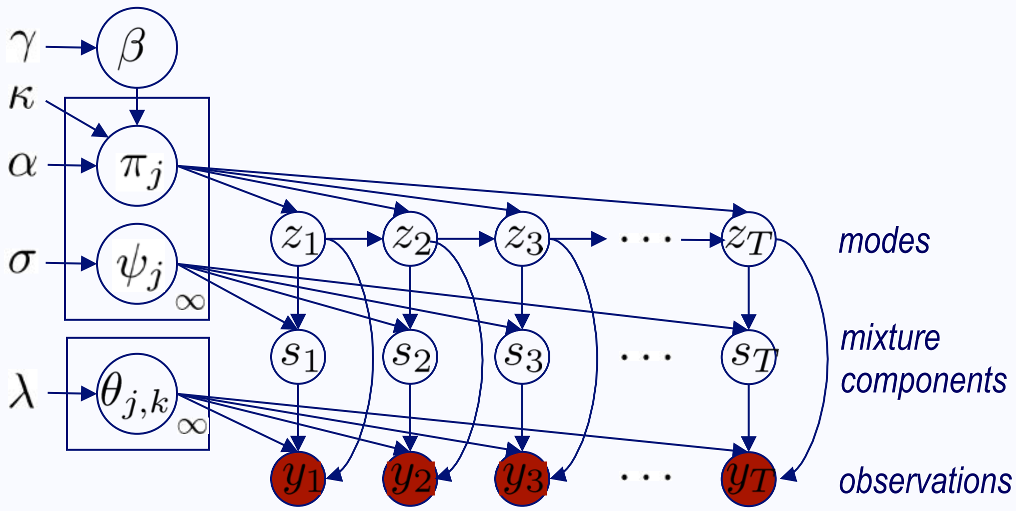
# Results: Meeting 18



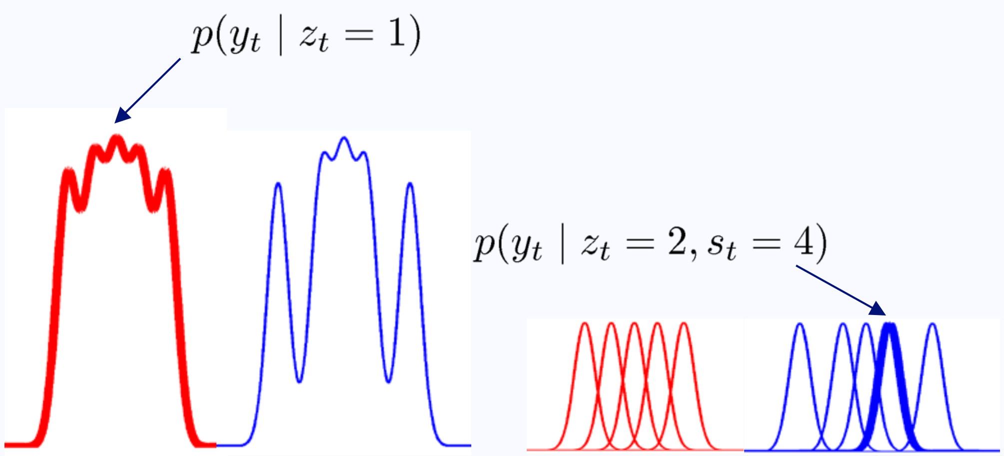

Sticky DER = ~~20.48%~~ 4.81%

ICSI DER = 22.00%

# HDP-HMM: Multimodal Emissions

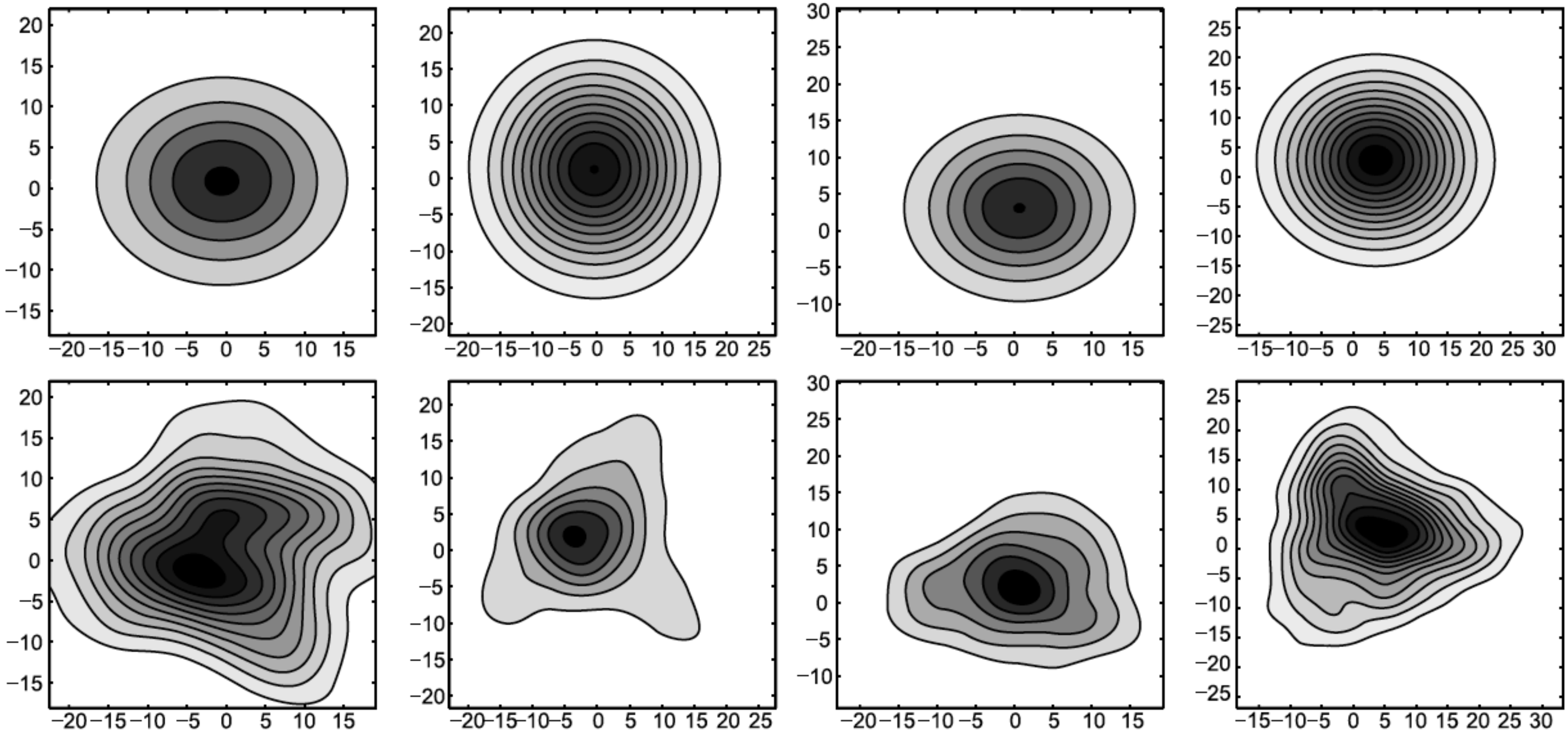



$$
\begin{aligned}
\beta &\sim \mathrm{Stick}(\gamma) \\
\pi_j &\sim \mathrm{DP}\left(\alpha\beta + \kappa\delta_j\right) \\
\psi_j &\sim \mathrm{Stick}(\sigma) \\
z_t &\sim \pi_{z_{t-1}} \\
s_t &\sim \psi_{z_t} \\
y_t &\sim F(\theta_{z_t, s_t})
\end{aligned}
$$

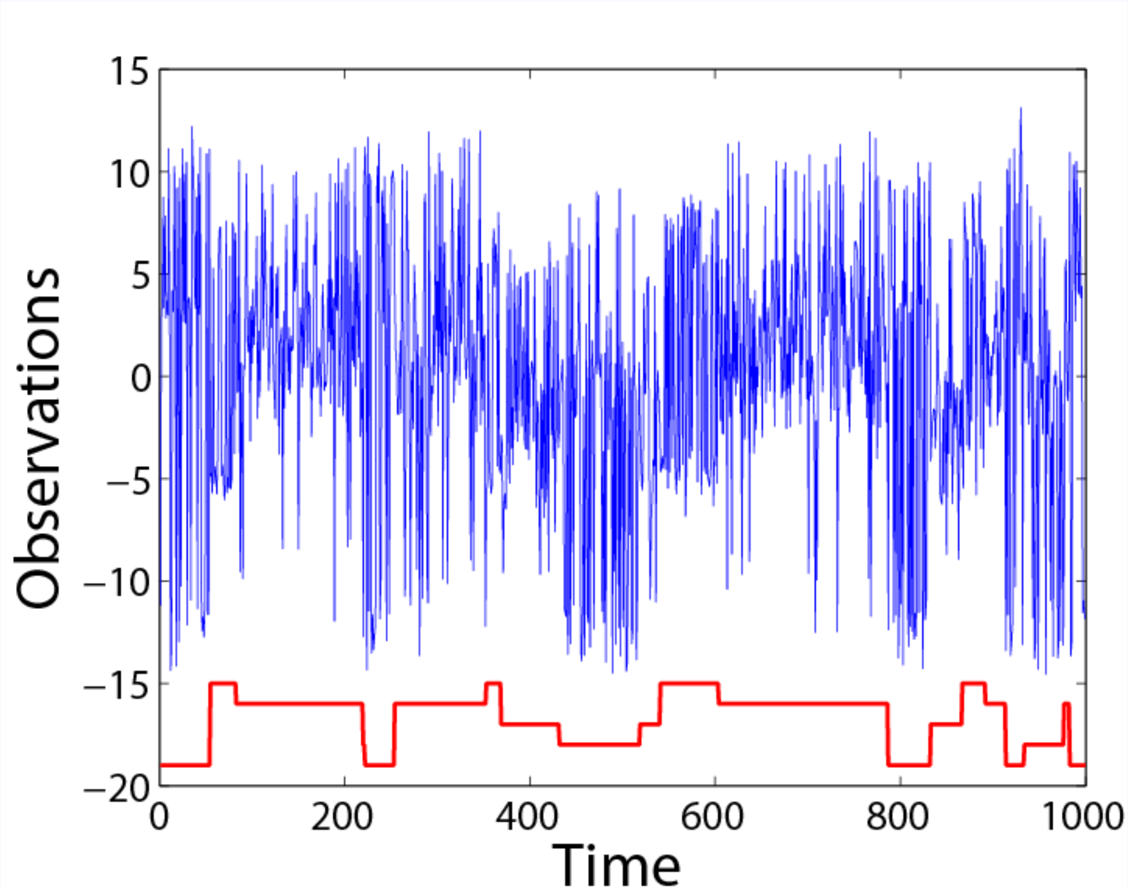



- Approximate multimodal emissions with DP mixture
- Temporal mode persistence disambiguates model

# Why Complex Emissions?

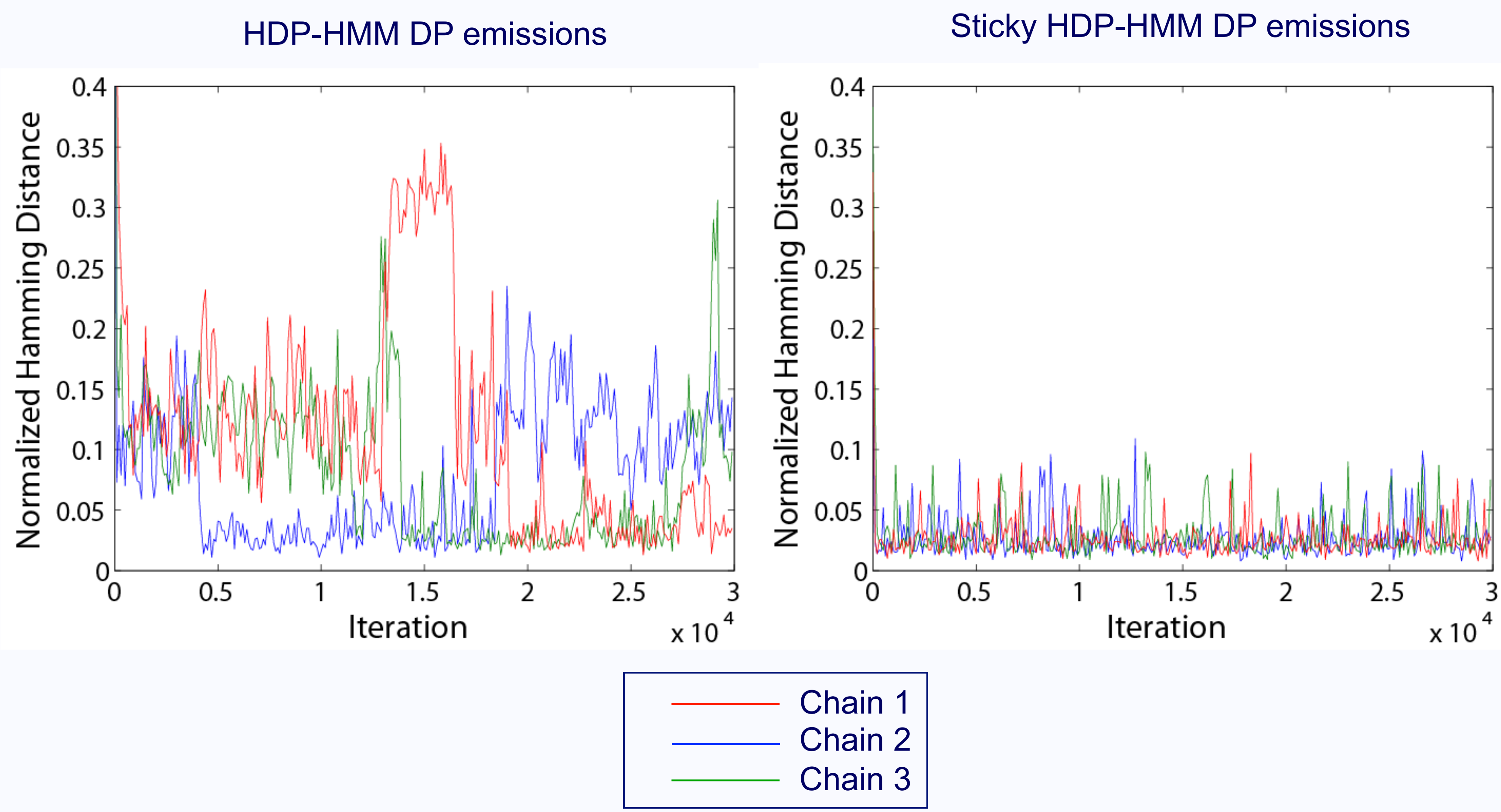


# Results: Mixture Emissions

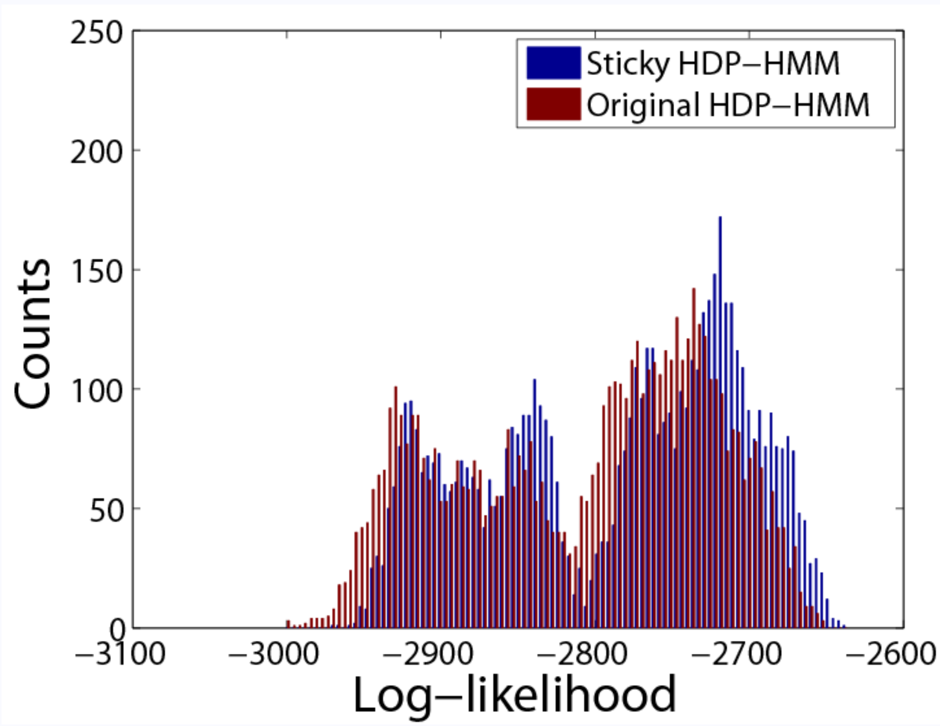



- 5-mode HMM
  - # emission components $n_k \sim \text{Uniform}[1, 10]$
  - Equal mixture weights
- Distance between observations not direct factor in grouping observations within mode

# Results: Mixture Emissions

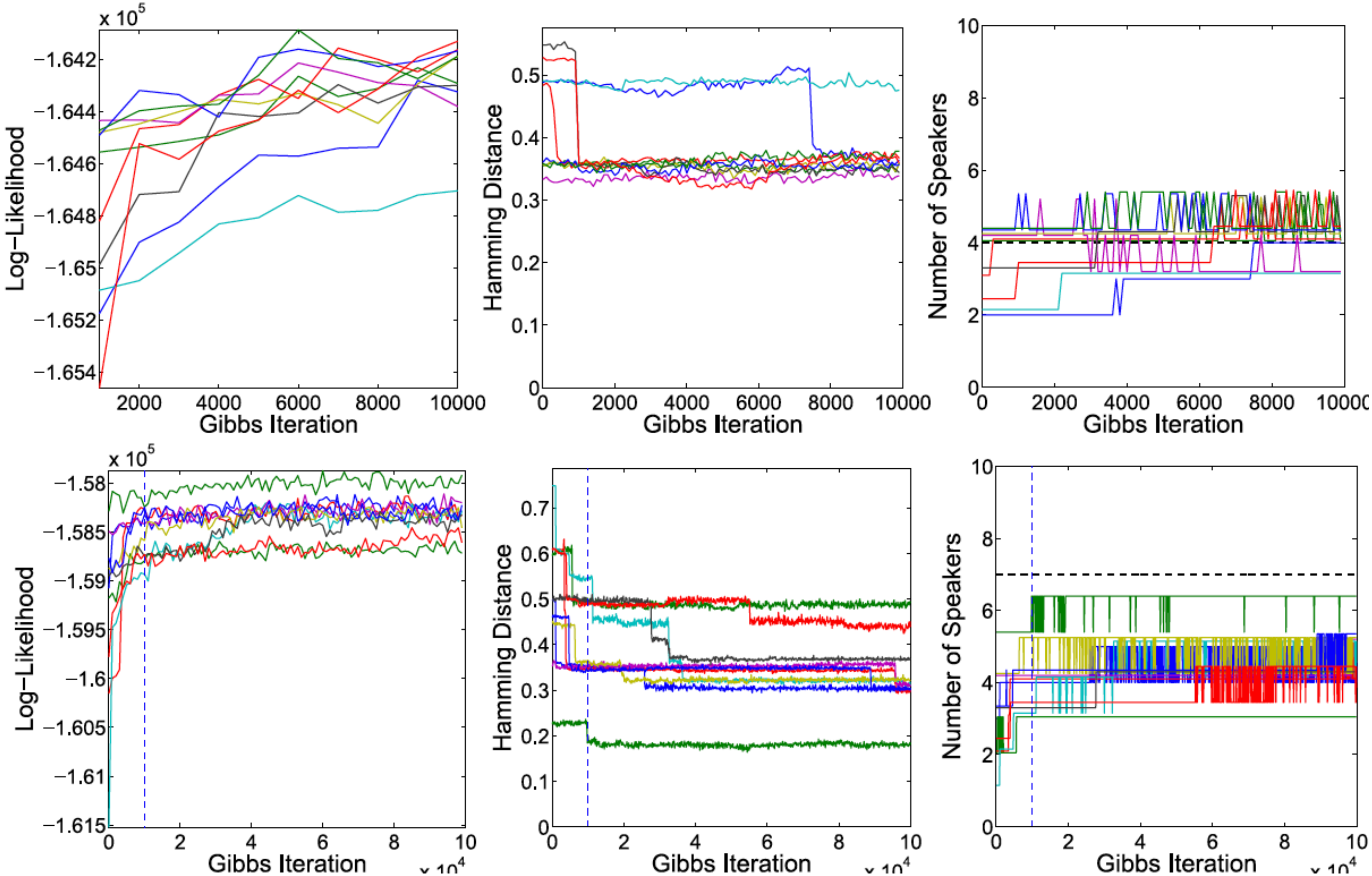


# Results: Mixture Emissions




- Improves *predictive probability* of test sequences
- Likely to see larger improvement in higher dimensions

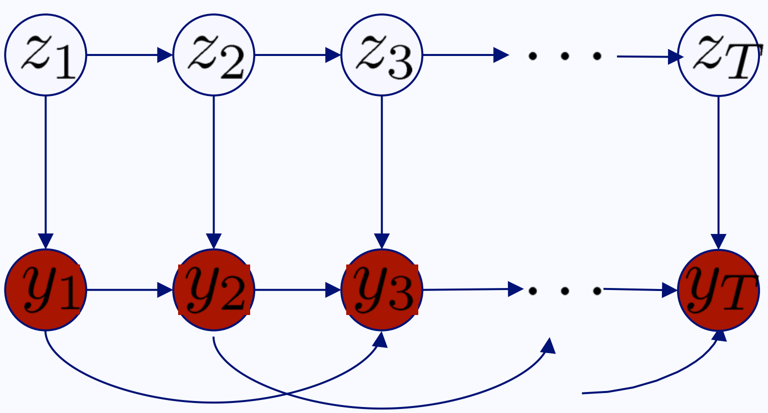
# Is it Mixing?

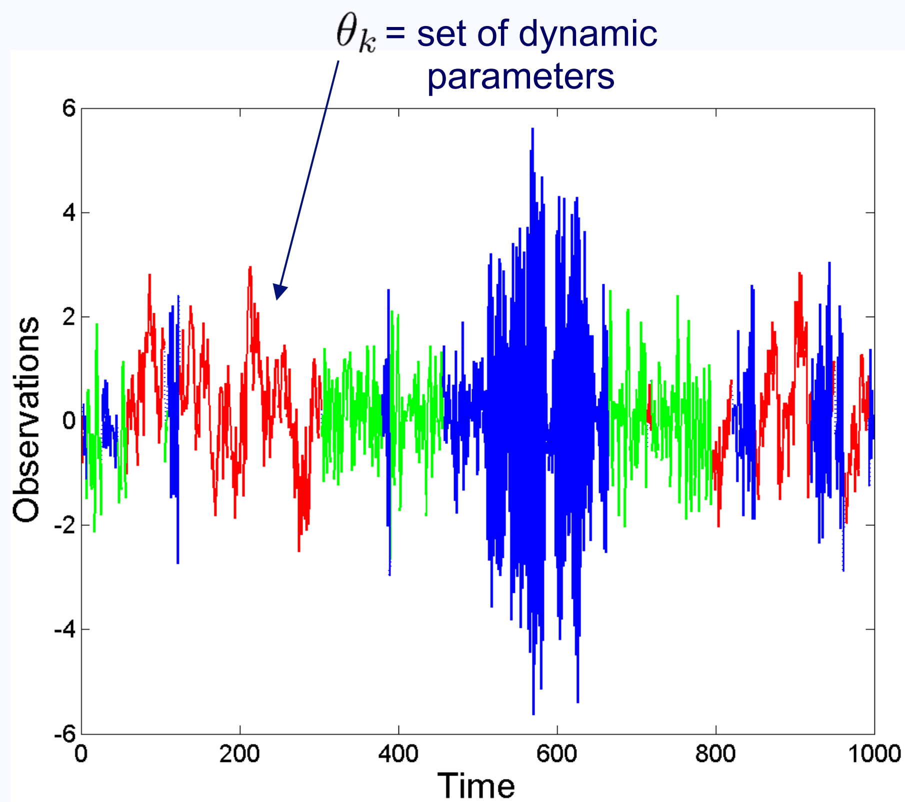


x 10^5
Log−Likelihood
Gibbs Iteration
Hamming Distance
Gibbs Iteration
Number of Speakers
Gibbs Iteration
x 10^5
Log−Likelihood
Gibbs Iteration
x 10^4
Hamming Distance
Gibbs Iteration
x 10^4
Number of Speakers
Gibbs Iteration
x 10^4

# Issue 3: Complex Local Dynamics

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



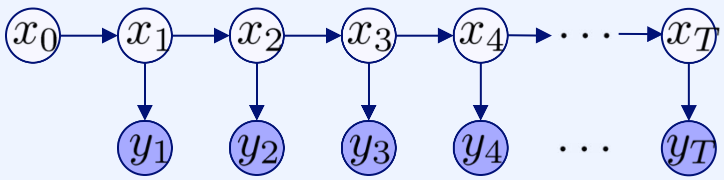

Switching Dynamical Processes

# Linear Dynamical Systems

- State space LTI model:

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

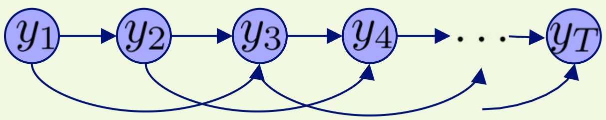
$$e_t \sim \mathcal{N}(0, \Sigma) \quad w_t \sim \mathcal{N}(0, R)$$




- Vector autoregressive (VAR) process:

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

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

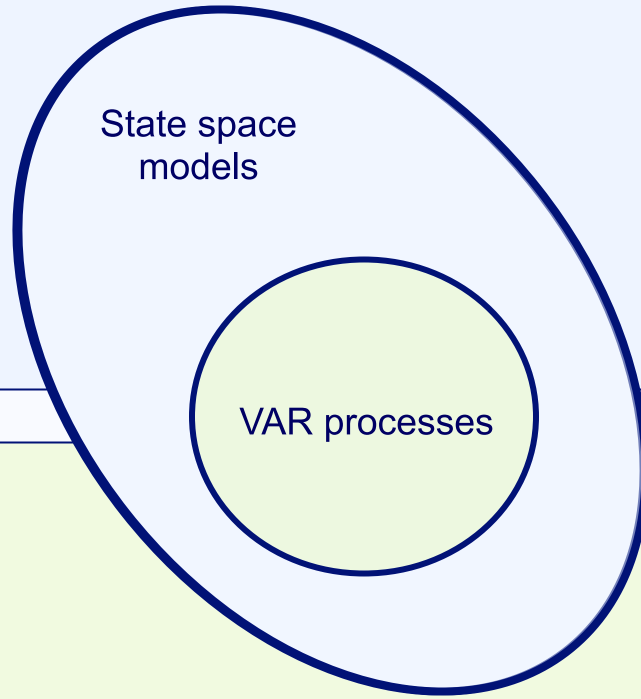


# Linear Dynamical Systems

- 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:

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

$$y_t = \begin{bmatrix} I & 0 & \dots & 0 \end{bmatrix} x_t.$$

# Switching Dynamical Systems

Switching linear dynamical system (SLDS):

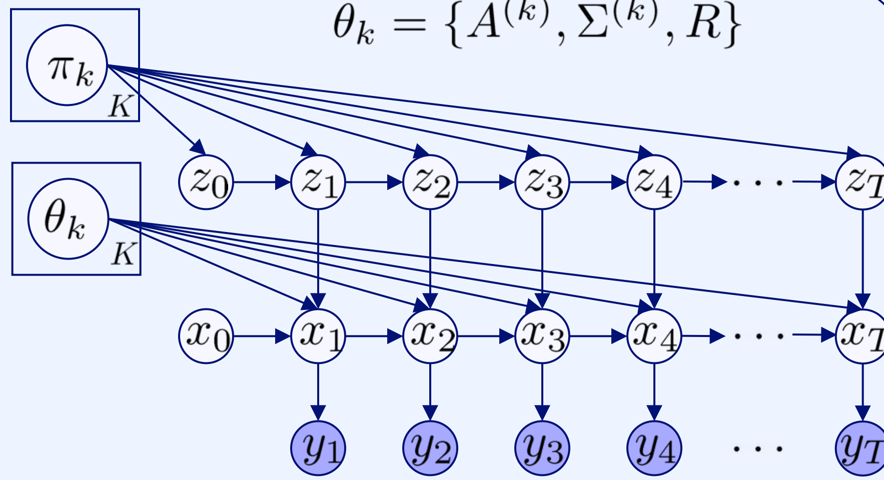
$$z_t \sim \pi_{z_{t-1}}$$

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

$$y_t = C x_t + w_t$$

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

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



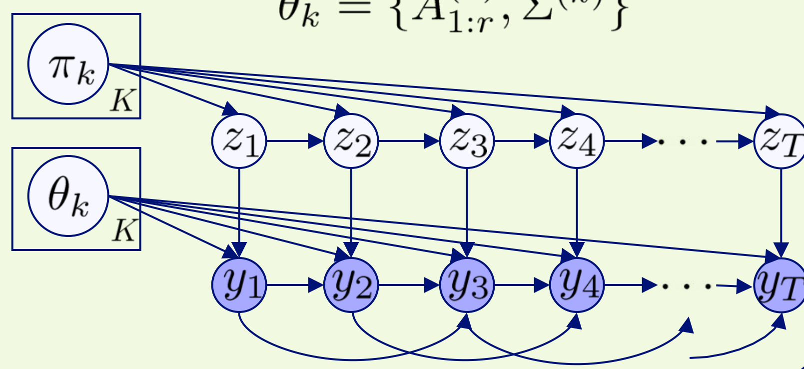

Switching VAR process:

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

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

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

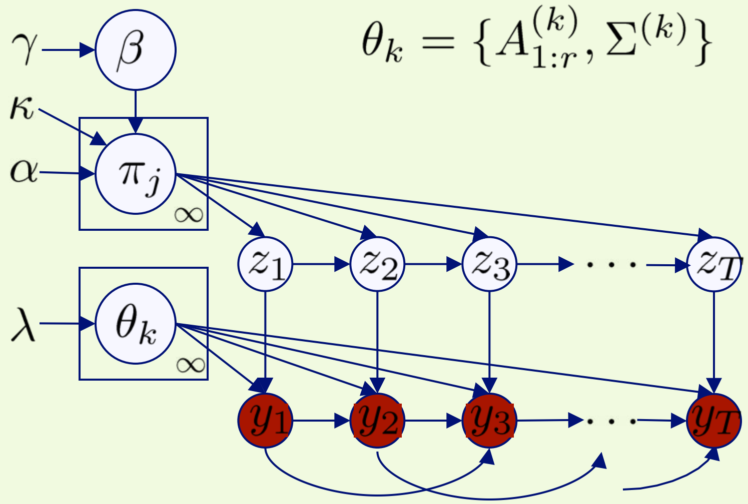
$$\theta_k = \{A_{1:r}^{(k)}, \Sigma^{(k)}\}$$



# HDP-AR-HMM and HDP-SLDS

## HDP-AR-HMM

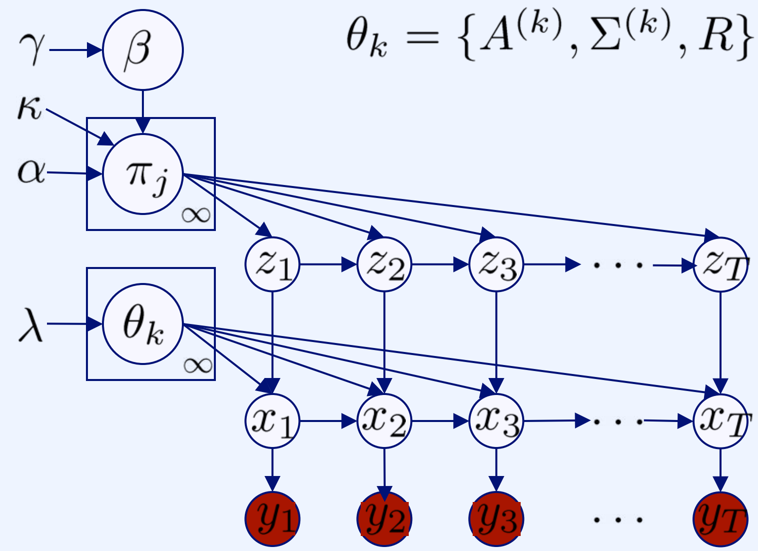
$$\theta_k = \{A_{1:r}^{(k)}, \Sigma^{(k)}\}$$



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

## HDP-SLDS

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



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

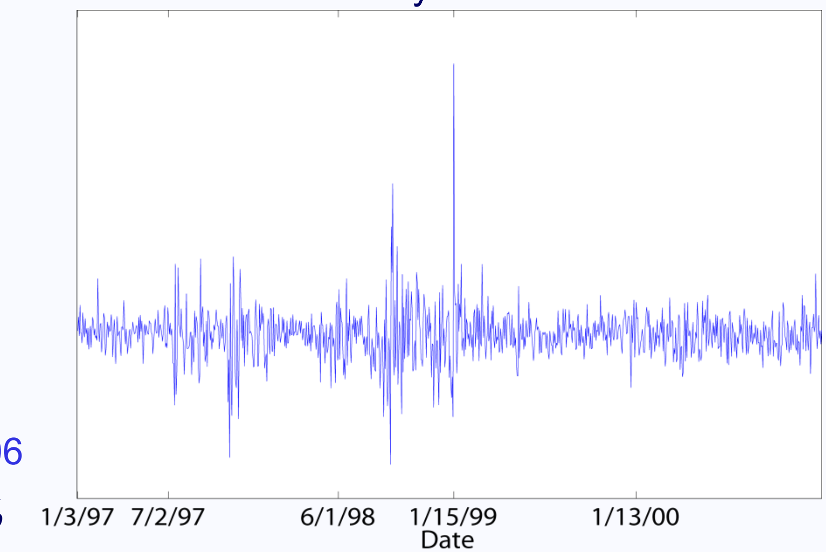
$$C = [I \;\; 0]$$

# Results: IBOVESPA

- Data: Sao Paolo stock index
- Goal: detect changes in volatility
- Compare inferred change-points to 10 cited world events

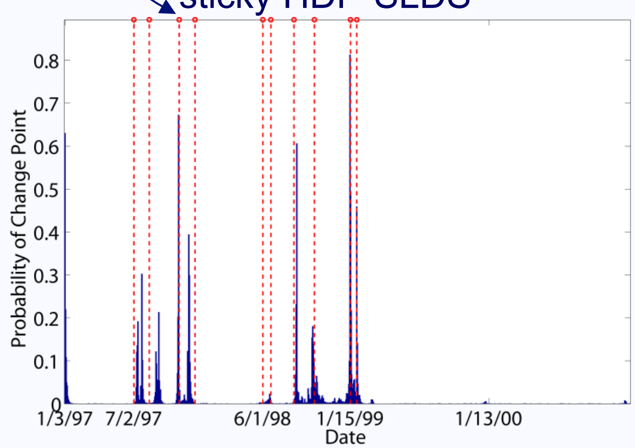
Carvalho and Lopes, *Comp. Stat. & Data Anal.*, 2006

Daily Returns

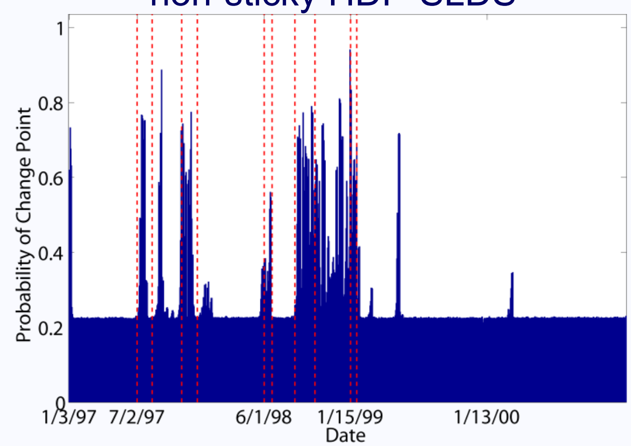



*Hong Kong stock index falls 10.4%*

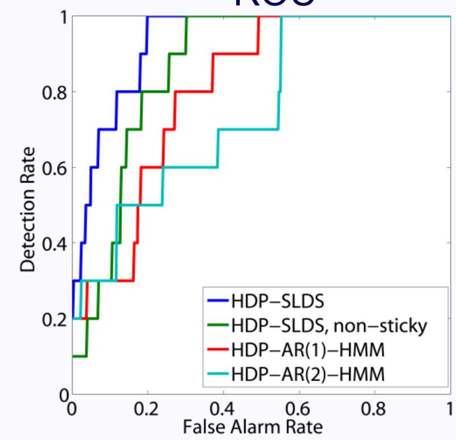
sticky HDP-SLDS




non-sticky HDP-SLDS

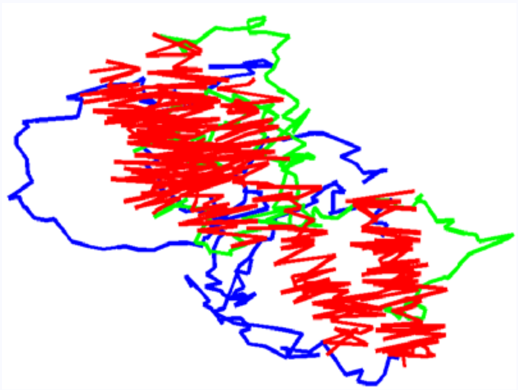



ROC

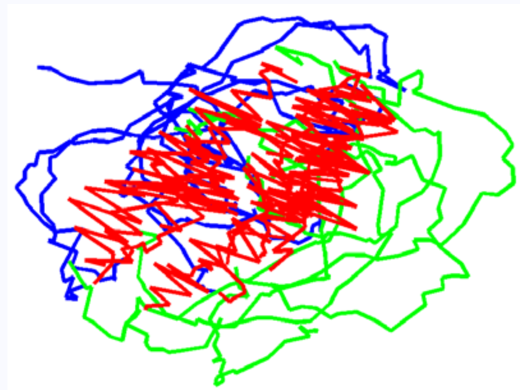


# Dancing Honey Bees

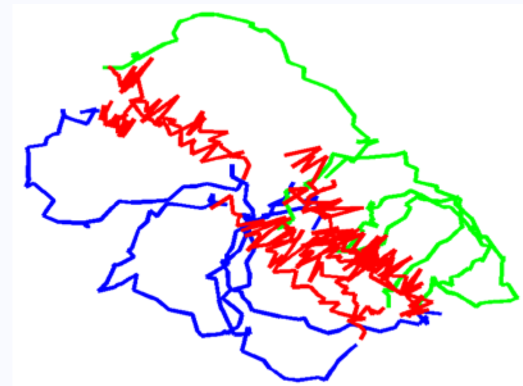
# Honey Bee Results: HDP-AR(1)-HMM



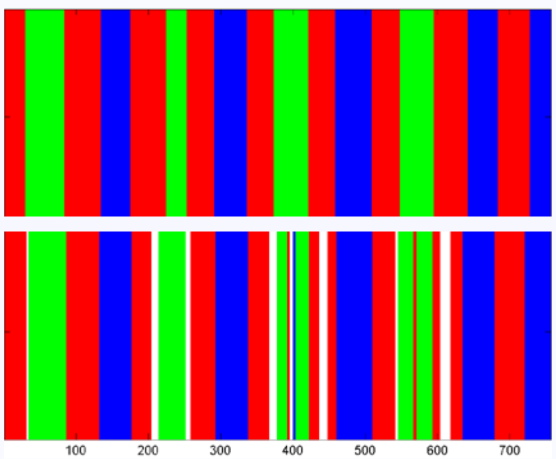
Sequence 1



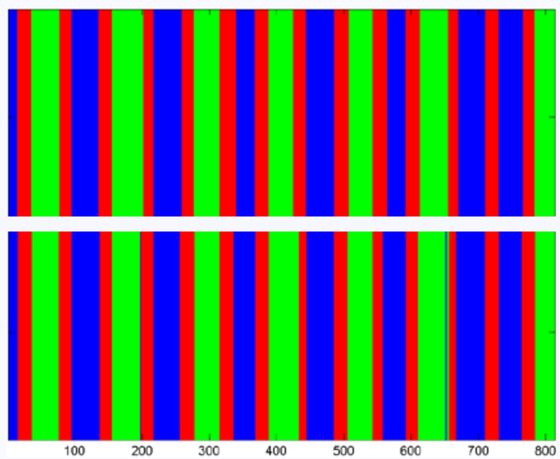
Sequence 2



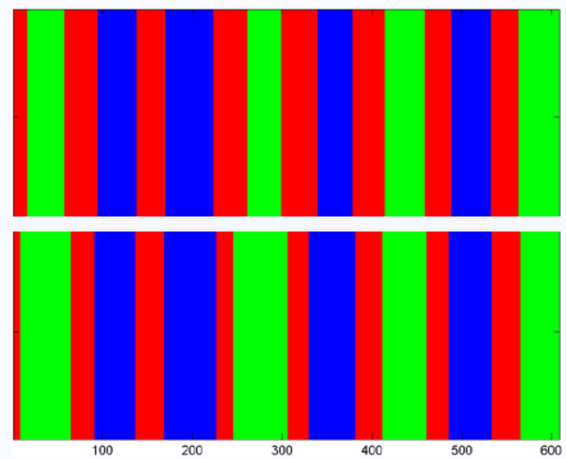
Sequence 3




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

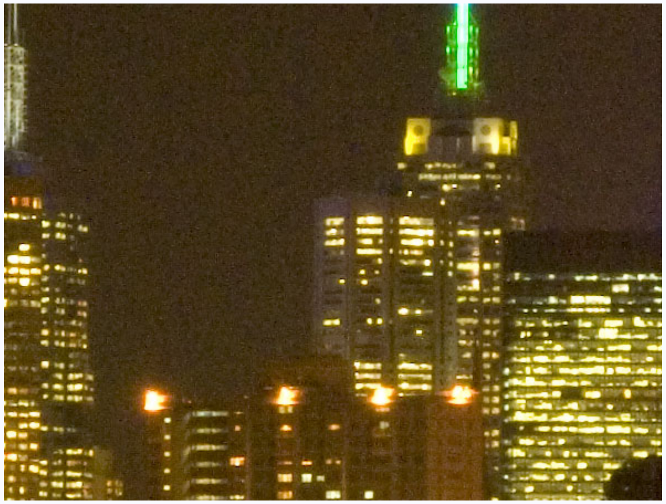



HDP-AR-HMM: 92.5%
SLDS [Oh]: 90.2%




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

# Low-level Image Analysis



Noise Removal

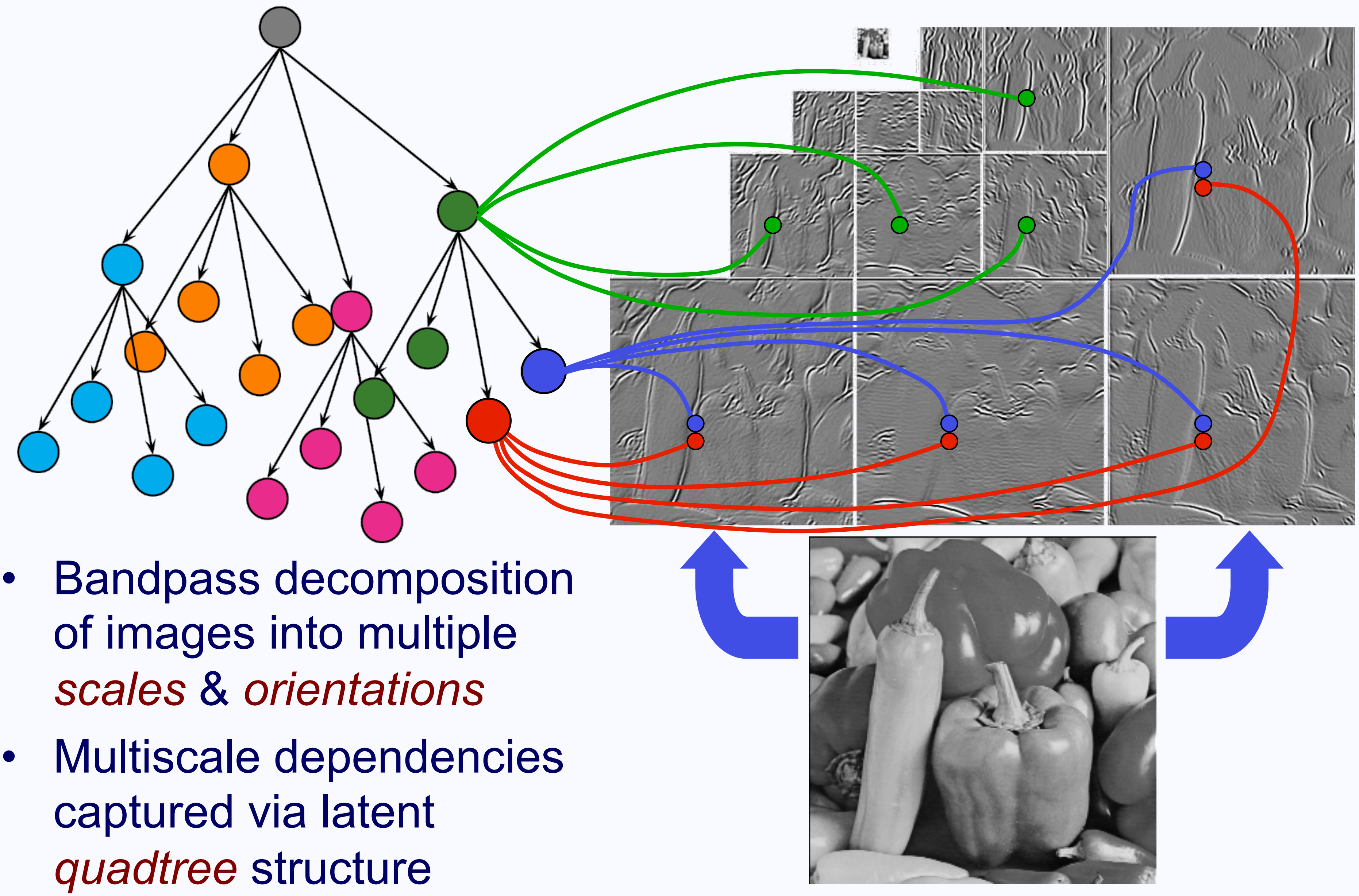
Deblurring

Inpainting & Restoration

*Goals:*

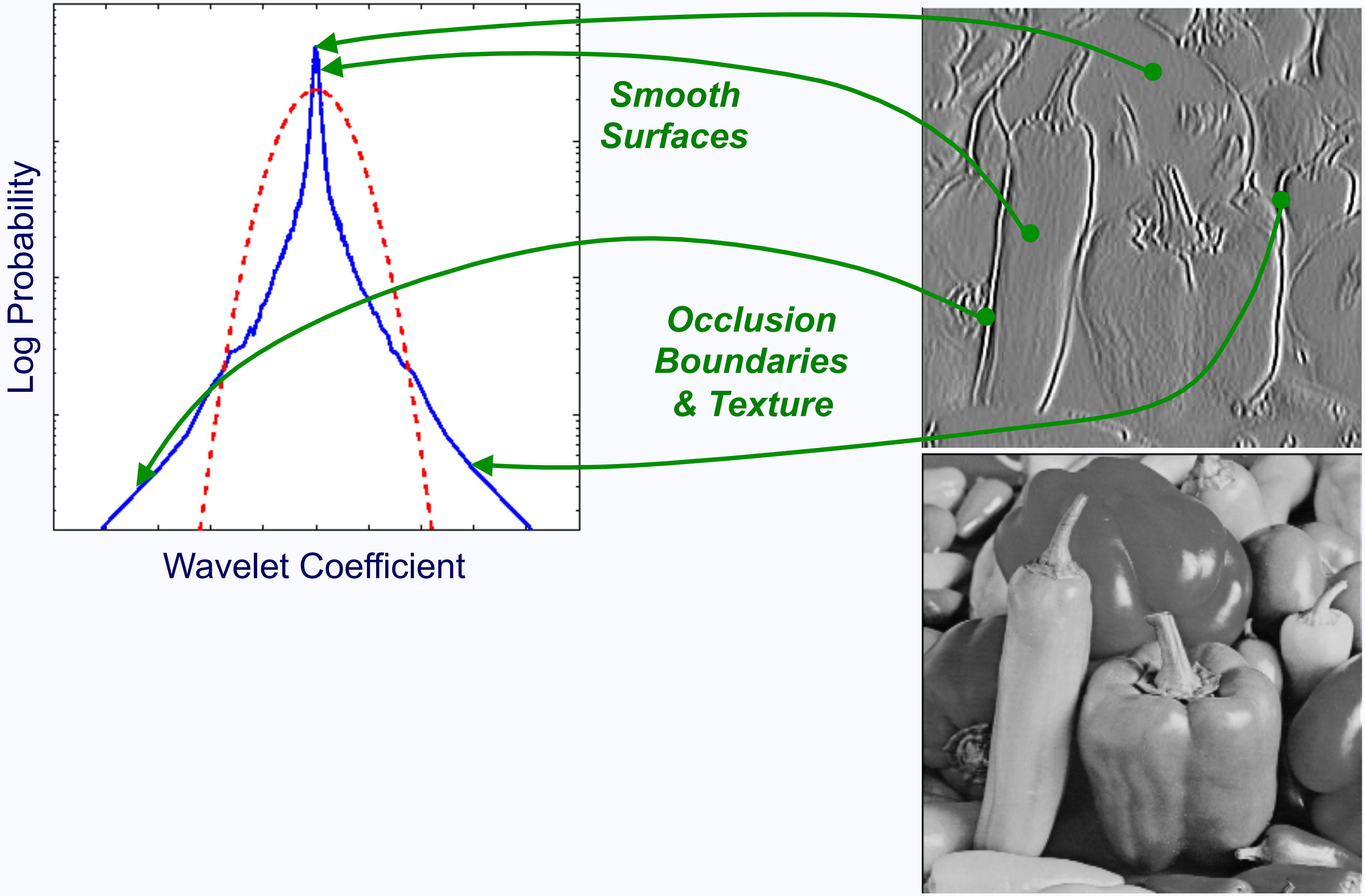
- Accurately model the statistics of *natural images*
- Exploit the availability of large digital *image collections*

# Wavelet Decompositions

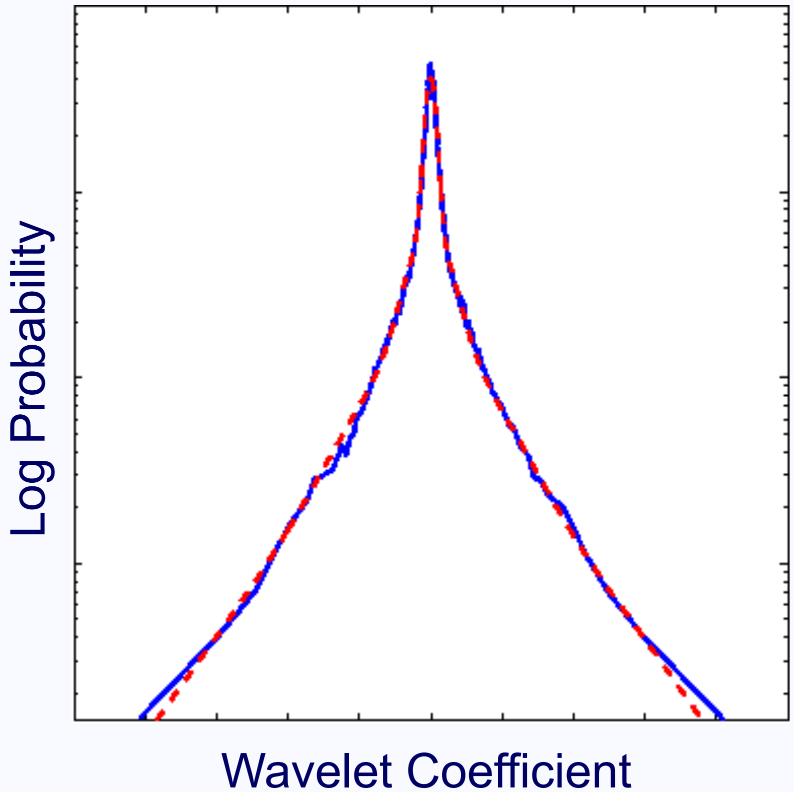


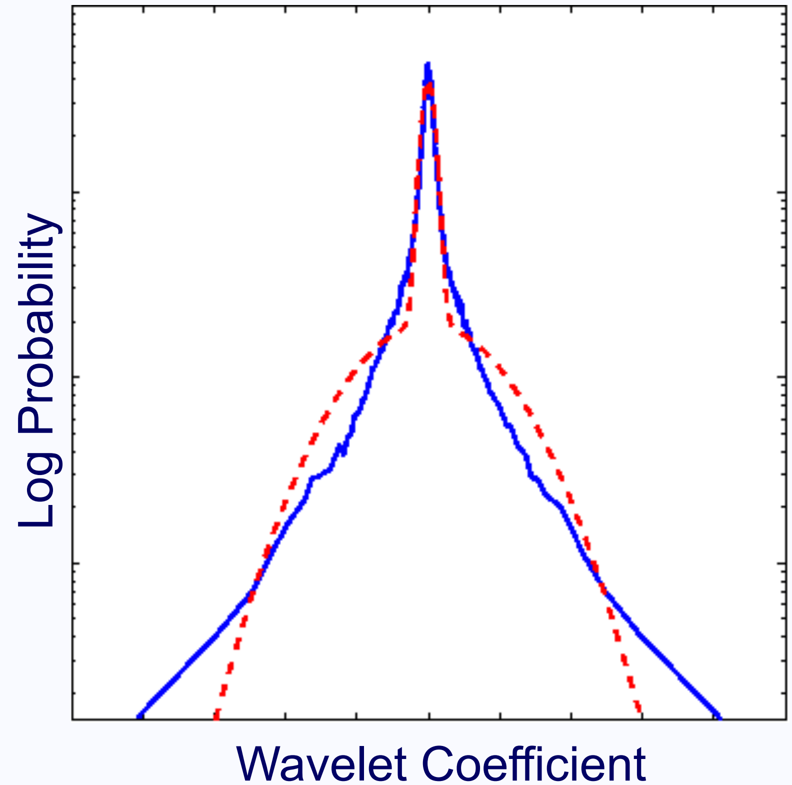
- Bandpass decomposition of images into multiple *scales* & *orientations*
- Multiscale dependencies captured via latent *quadtree* structure

# Wavelets: Marginal Statistics



# Gaussian Mixture Models




$$x_i = v_i u_i$$

$$v_i \geq 0 \qquad u_i \sim \mathcal{N}(0, \Lambda)$$

**Gaussian Scale Mixture (GSM)**

*Wainwright & Simoncelli, 2000*

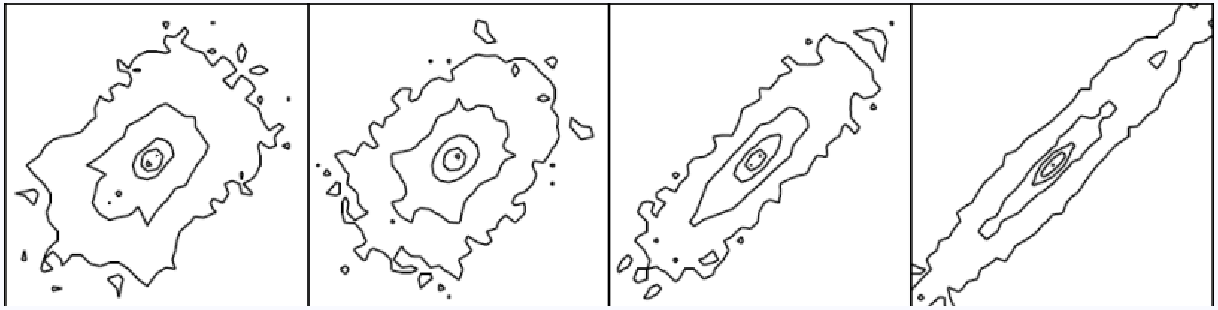
$$x_i \sim \pi\, \mathcal{N}(0, \Lambda_0) + (1-\pi)\mathcal{N}(0, \Lambda_1)$$

**Binary Gaussian Mixture**

*Computational advantages…*

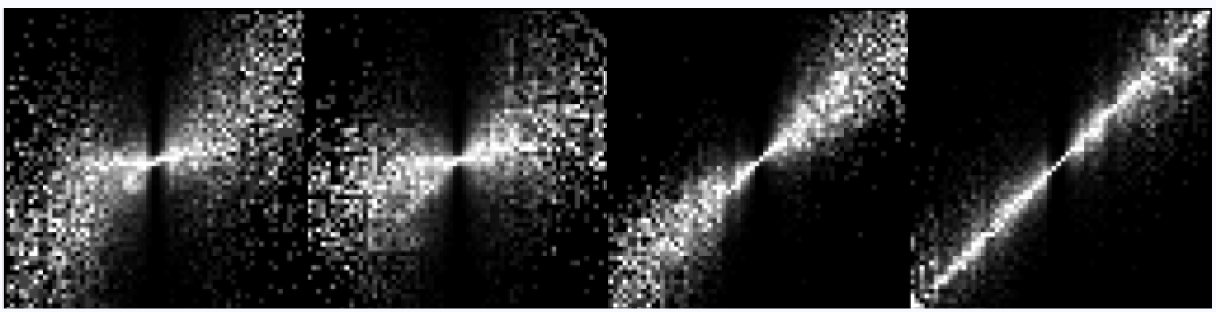
# Wavelets: Joint Statistics

**Pairwise Joint Histograms:**

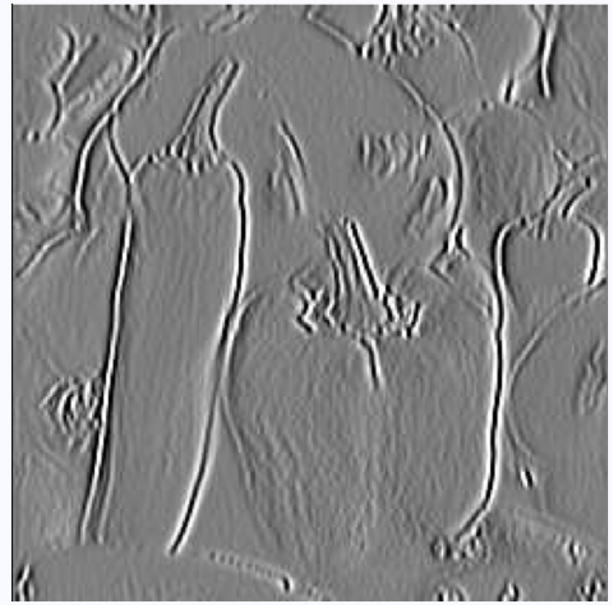


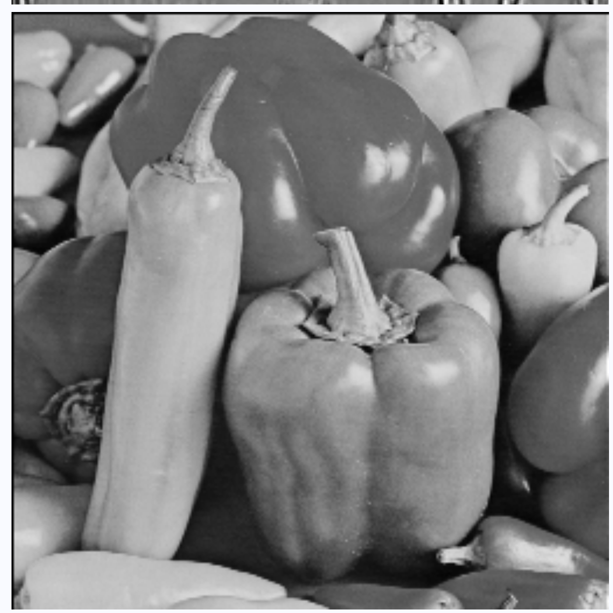
*Orientation* *Scale* *Vertical* *Horizontal*

**Pairwise Conditional Histograms:**



*Orientation* *Scale* *Vertical* *Horizontal*



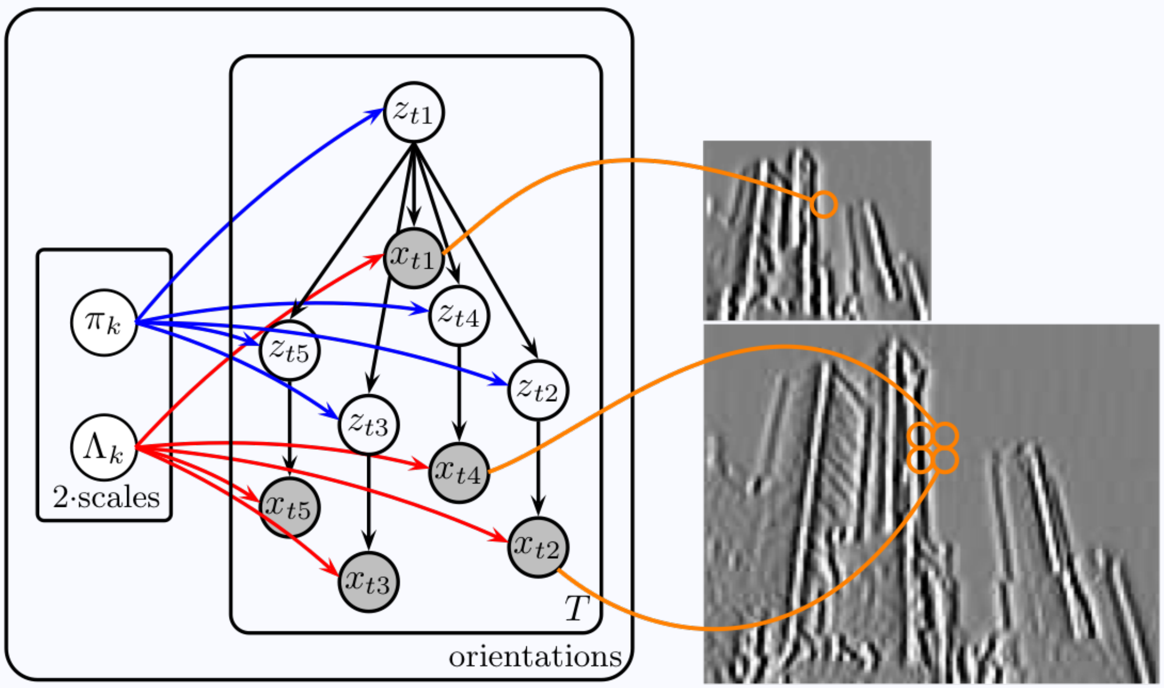


## Large magnitude wavelet coefficients…

- *Persist* across multiple scales
- *Cluster* at adjacent spatial locations

# Binary Hidden Markov Trees

*Crouse, Nowak, & Baraniuk, 1998*




$\pi_k \rightarrow$ state *transition* distributions

$$z_{ti} \sim \pi_{z_{\mathrm{Pa}(ti)}}$$

$\Lambda_k \rightarrow$ state-specific *emission* covariances

$$x_{ti} \sim \mathcal{N}\left(0, \Lambda_{z_{ti}}\right)$$

$z_{ti} \rightarrow$ hidden *state* or cluster assignment

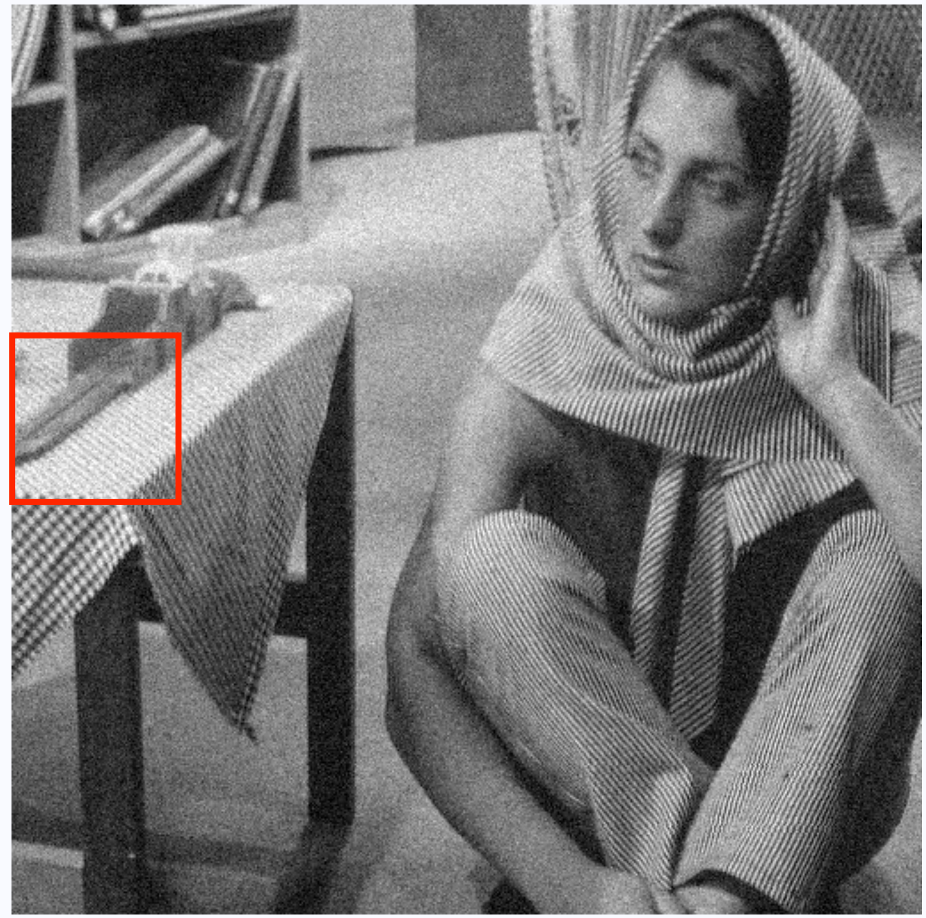
$$z_{ti} \in \{0, 1\}$$

$x_{ti} \rightarrow$ *observed* wavelet coefficient

- Coefficients marginally distributed as mixtures of two Gaussians
- Markov dependencies between hidden states capture persistence of image contours across locations and scales
- Each orientation is modeled independently

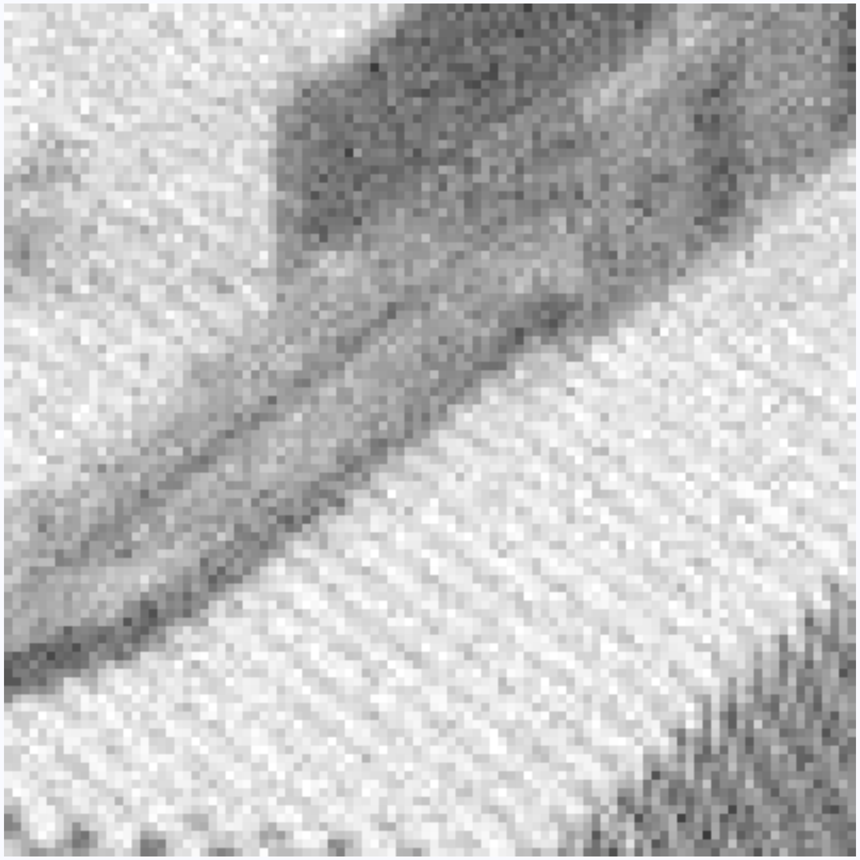
# Validation : Image Denoising

Original

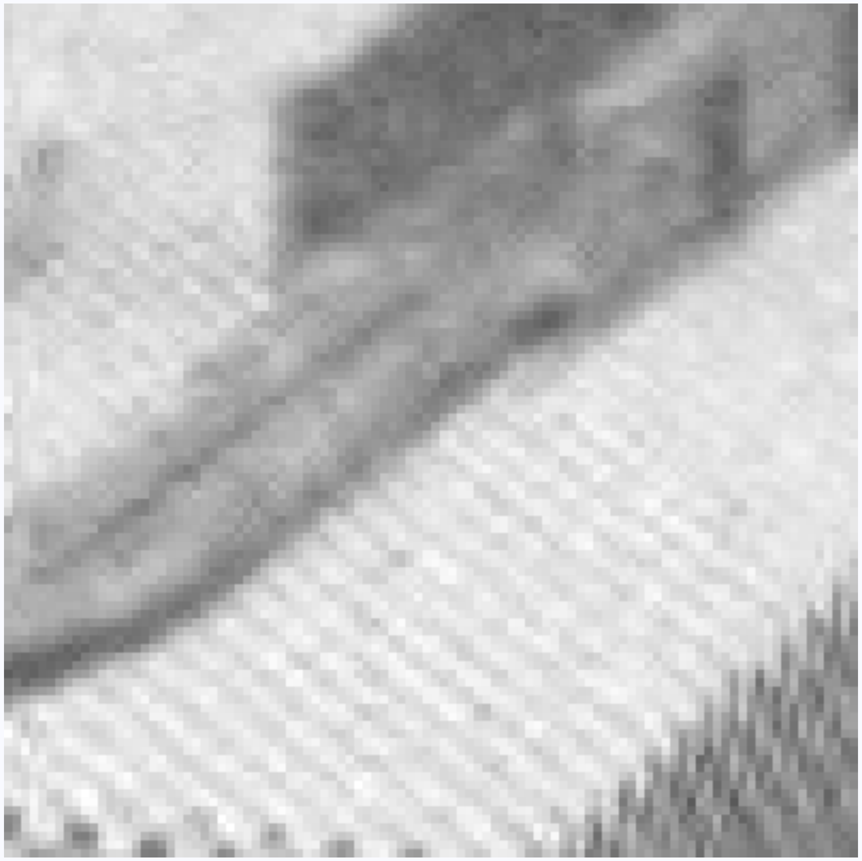


Corrupted by Additive White Gaussian Noise (PSNR = 24.61 dB)

# Denoising with Binary HMTs



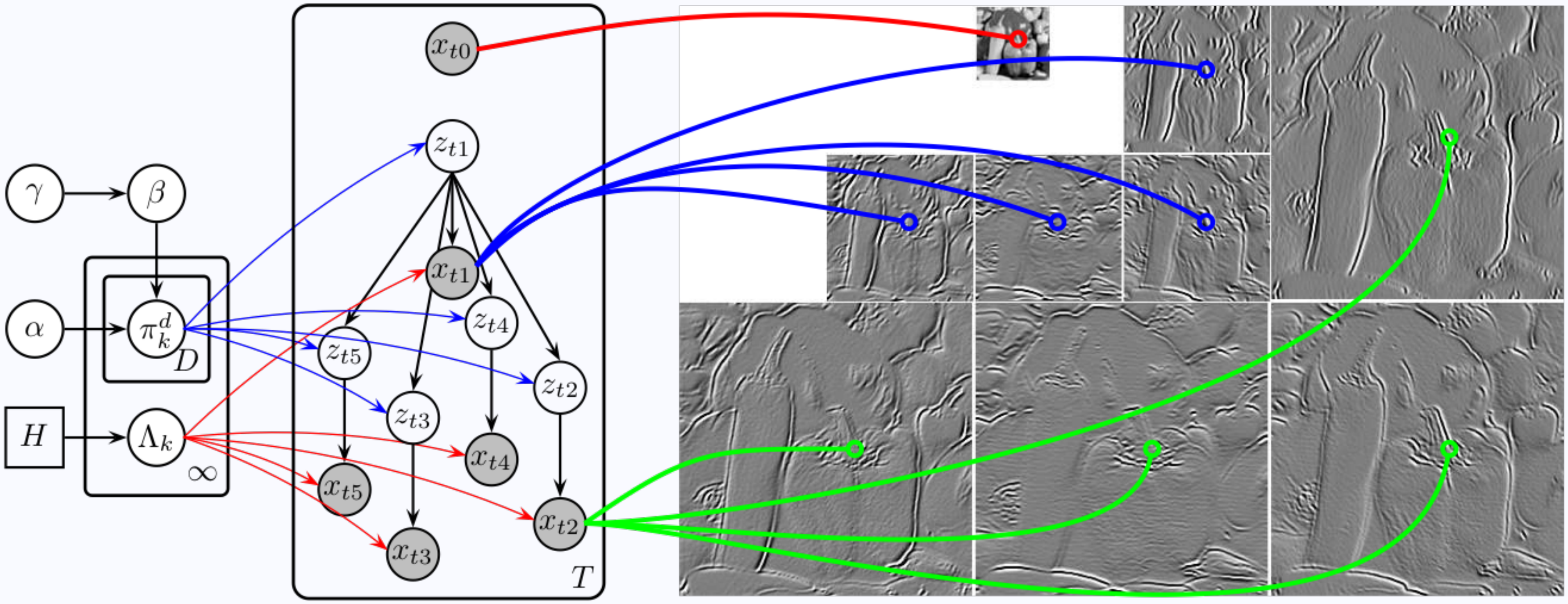
Noisy Input



Denoised (EM algorithm)

- Is two states per scale sufficient? How many is enough?
- Should states be shared the same way for all images, or for all wavelet decompositions?

# Hierarchical Dirichlet Process Hidden Markov Trees




$z_{ti} \longrightarrow$ indexes *infinite* set of hidden states

$$z_{ti} \in \{1, 2, 3, \ldots\}$$

$\pi_k \longrightarrow$ infinite set of state *transition* distributions

$$z_{ti} \sim \pi_{z_{\mathrm{Pa(ti)}}}^{d_{ti}}$$

$x_{ti} \longrightarrow$ observed *vector* of wavelet coefficients

$\Lambda_k \longrightarrow$ state-specific *emission* covariances

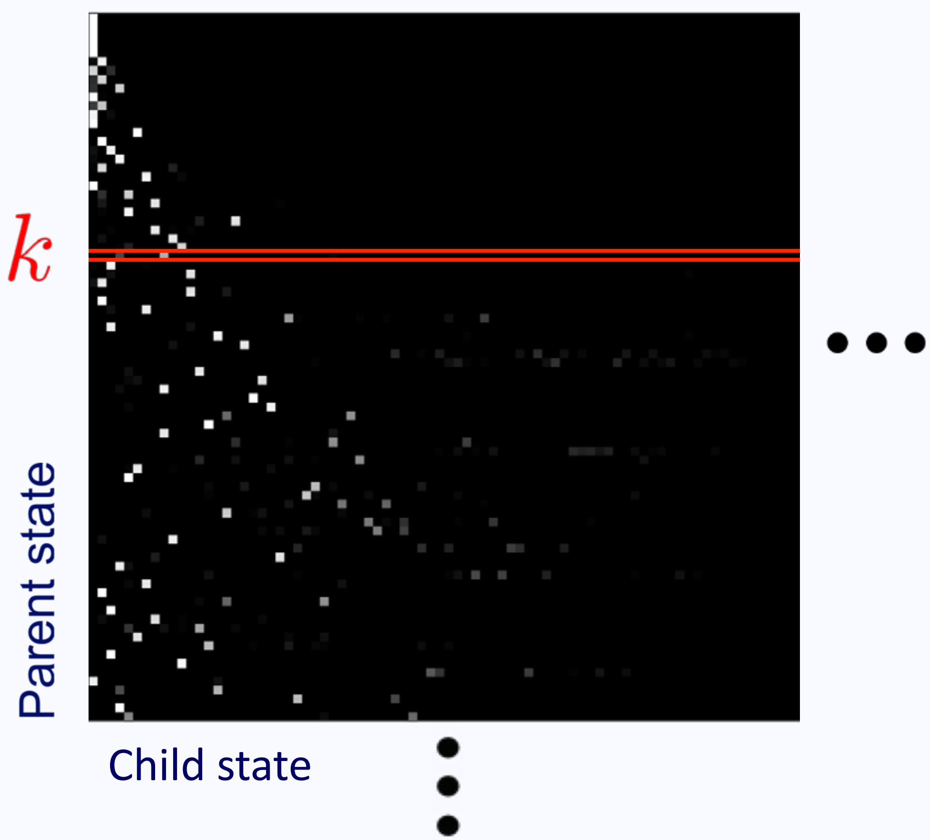
$$x_{ti} \sim \mathcal{N}\left(0, \Lambda_{z_{ti}}\right)$$

$$\Lambda_k \sim H$$

# Why a Hierarchical DP ? *(Teh et. al. 2004)*

- Hierarchical DP prior allows us to learn a potentially infinite set of *appearance patterns* from natural images
- Hierarchical coupling ensures, with high probability, that a common set of *child* states are reachable from each *parent*

$$\pi_k^{d_{ti}}(\ell) = \Pr\left[ z_{ti} = \ell \mid z_{\mathrm{Pa}(ti)} \right]$$




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

*Average state frequencies*

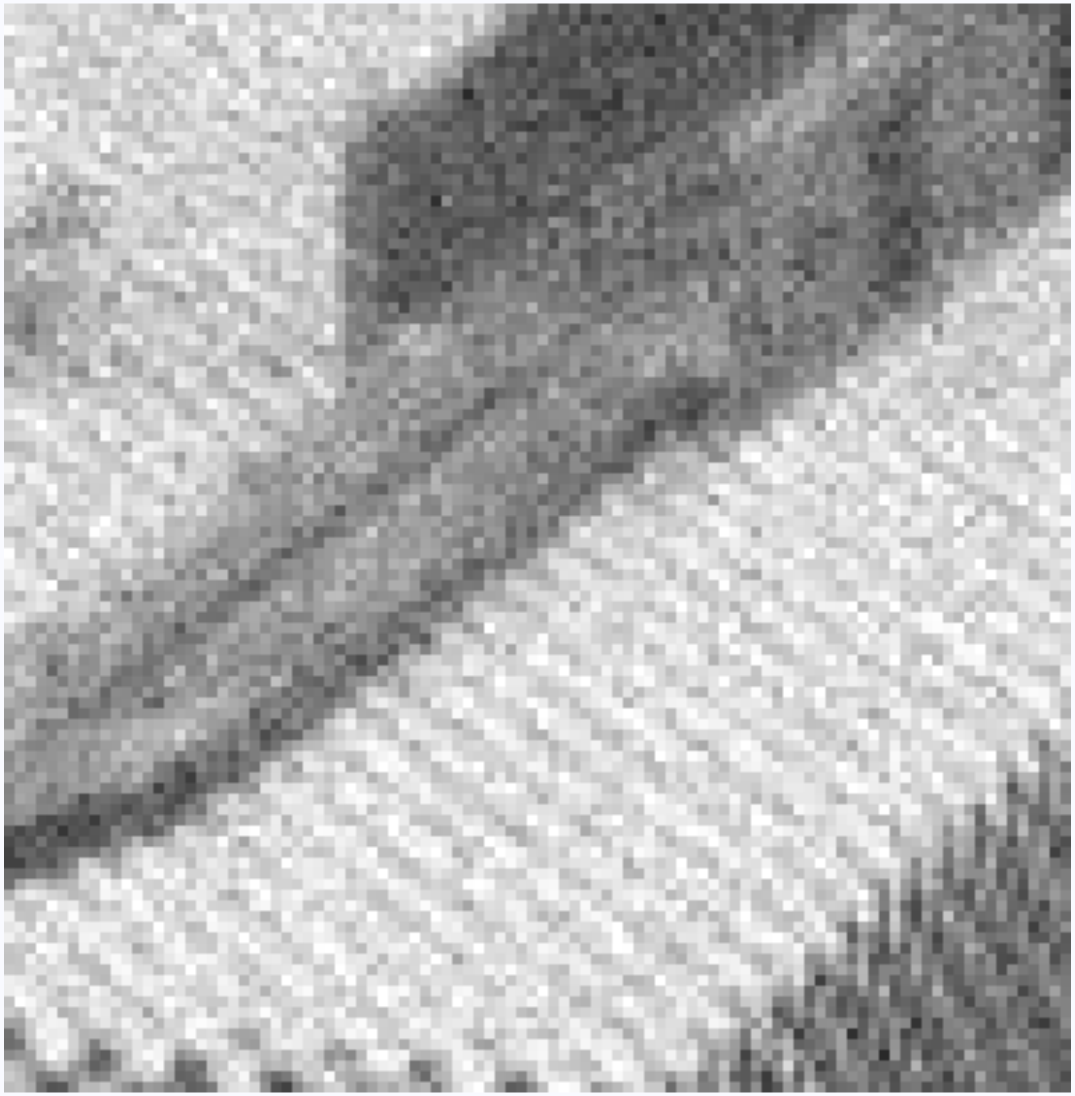

$$\pi_k^d \sim \mathrm{DP}(\alpha, \beta)$$

*Transition distributions*

$$\mathbb{E}\left[\pi_k^d\right] = \beta$$

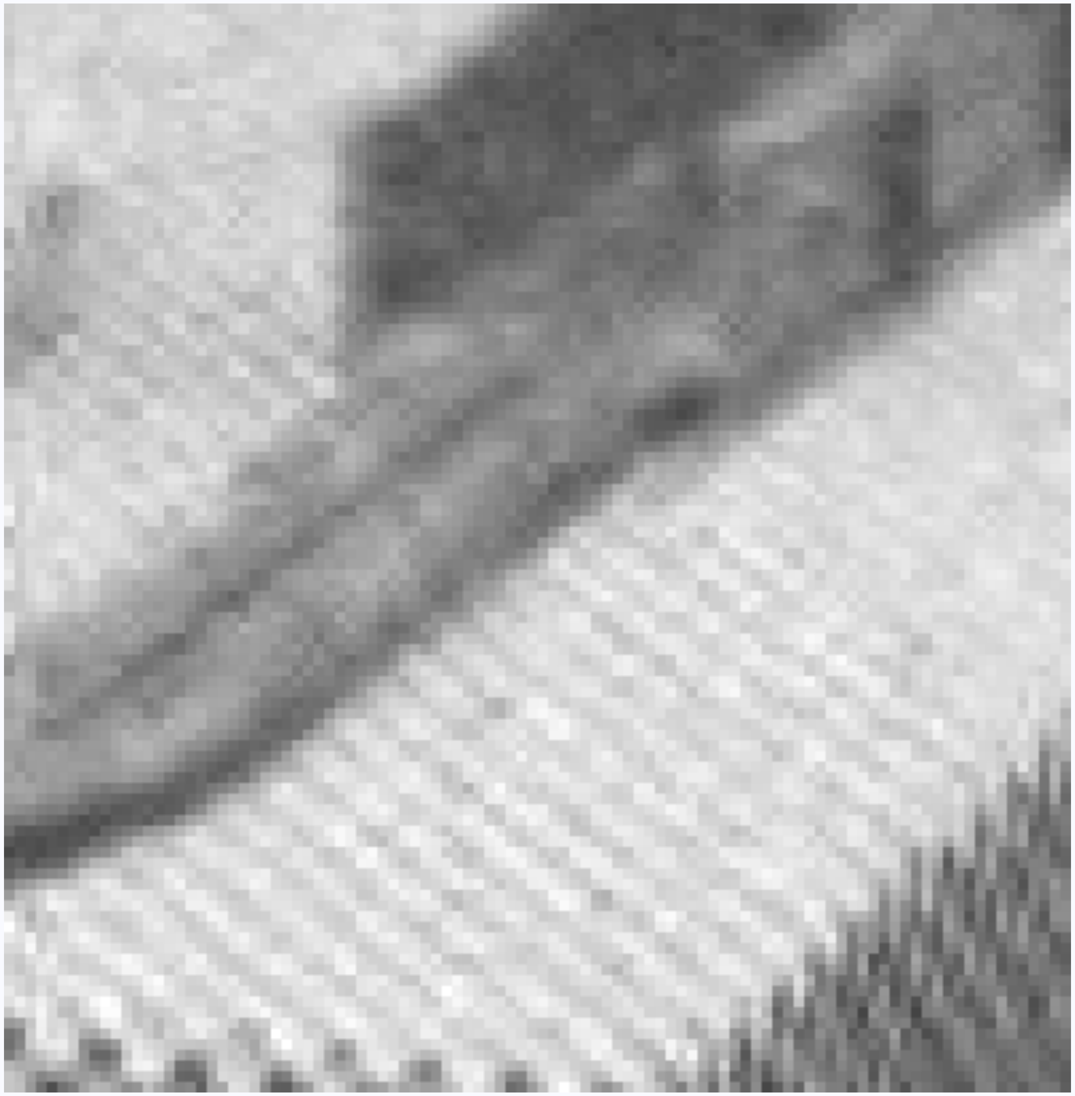
$\alpha \longrightarrow$ *Sparsity & variability of transition distributions*

# Denoising: Input



24.61 dB

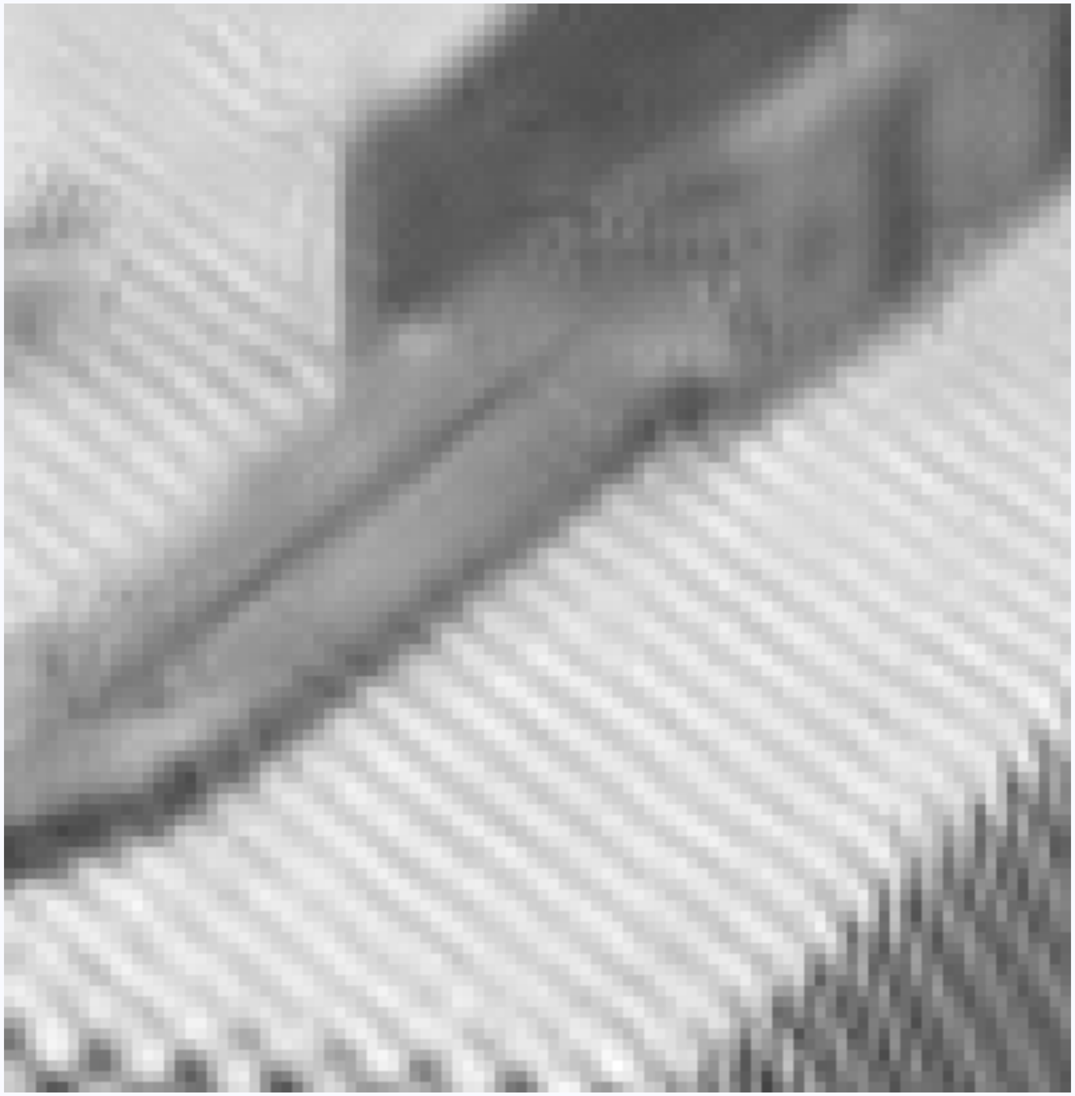
# Denoising: Binary HMT



**29.35 dB**

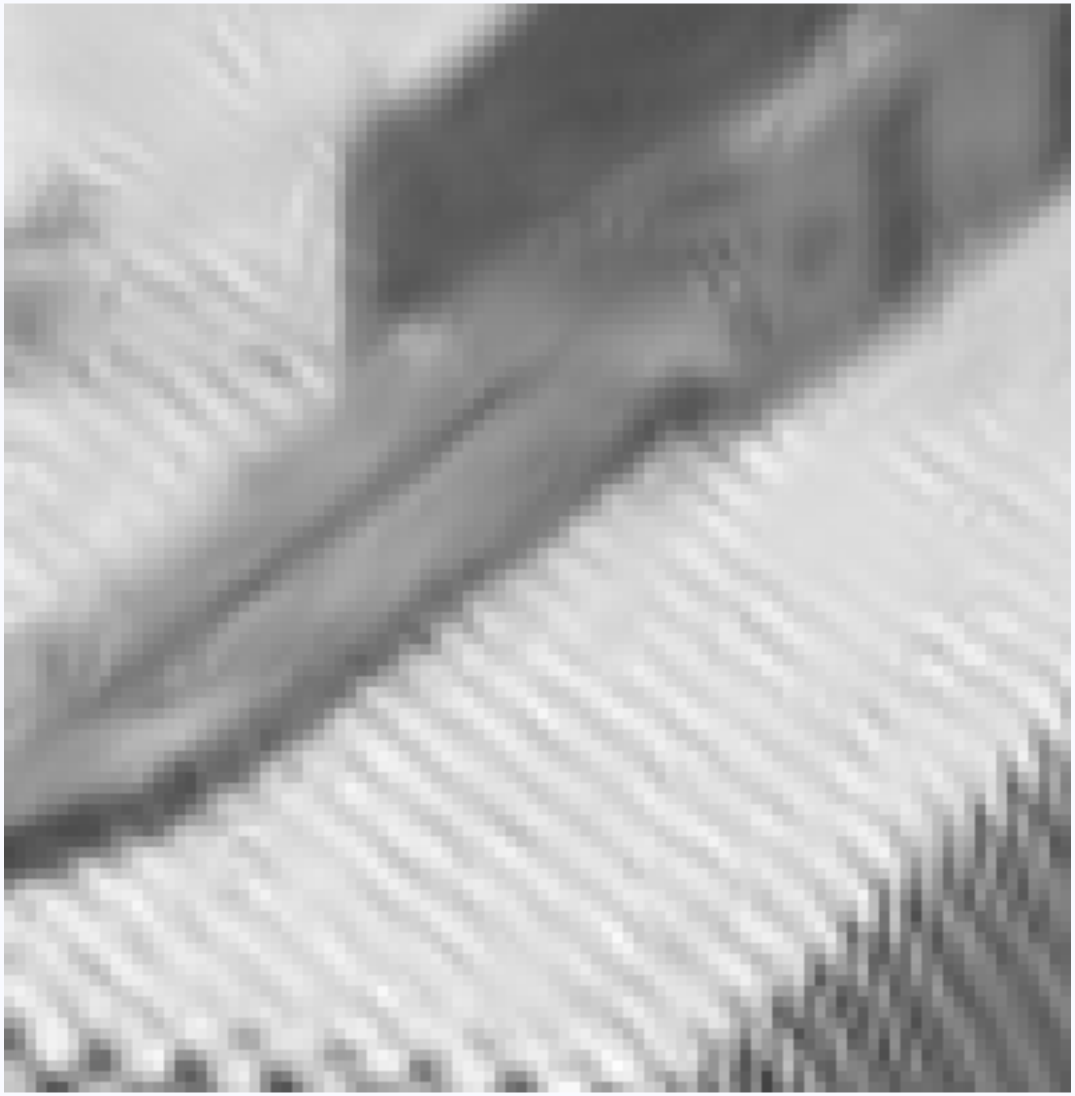
*Crouse, Nowak, & Baraniuk, 1998*

# Denoising: HDP-HMT



**32.10 dB**

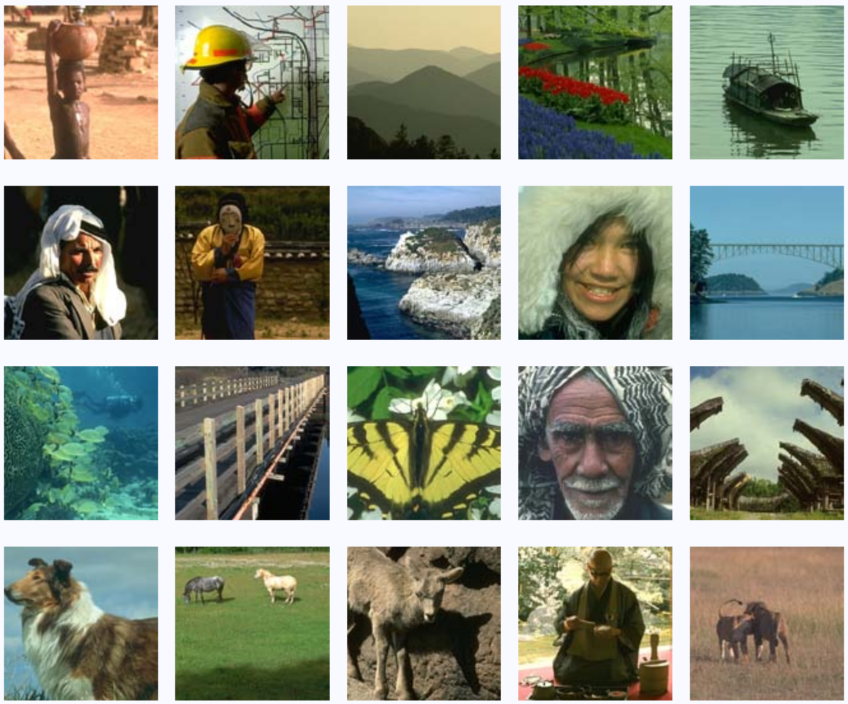
# Denoising: Local GSM



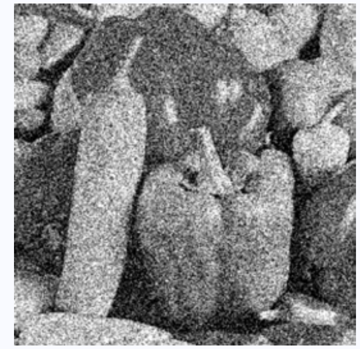
**31.84 dB**

*Portilla et. al., 2003*

# Estimating Clean Images



Empirical Bayesian approach estimates model parameters from the noisy image



Transfer denoising approach reuses multiscale hidden state patterns of clean images for making robust predictions

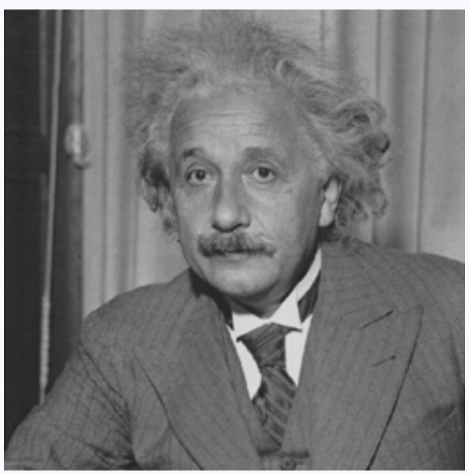
# Denoising Einstein

*Noisy*
*10.60 dB, 0.057*

*HDP-HMT*
*(Emp. Bayes)*
*25.64 dB, 0.564*

*HDP-HMT*
*(Transfer)*
*26.80 dB, 0.664*

*Original*



*BLS-GSM*
*26.38 dB, 0.647*

*BM3D*
*26.49 dB, 0.659*

# Natural Scene Denoising

*Noisy*
*8.14 dB, 0.033*

*HDP-HMT*
*(Emp. Bayes)*
*24.24 dB, 0.519*

*HDP-HMT*
*(Transfer)*
*26.50 dB, 0.794*

*Original*

*BLS-GSM*
*25.59 dB, 0.726*

*BM3D*
*25.74 dB, 0.751*

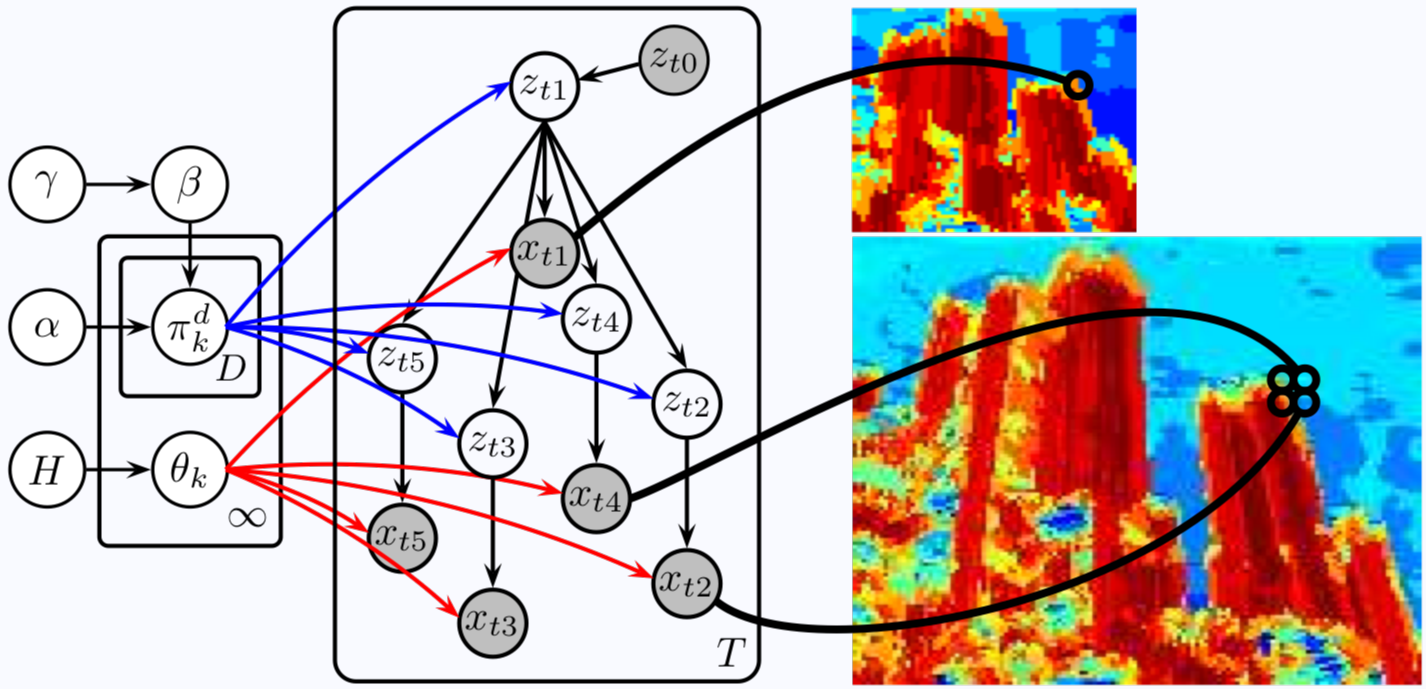
# Natural Scene Categorization

**Coast** **Forest** **Open Country** **Street** **Tall Building**

*Goals:*

- Visually *recognize* natural scene categories
- Accurately model the statistics of *natural scene categories*

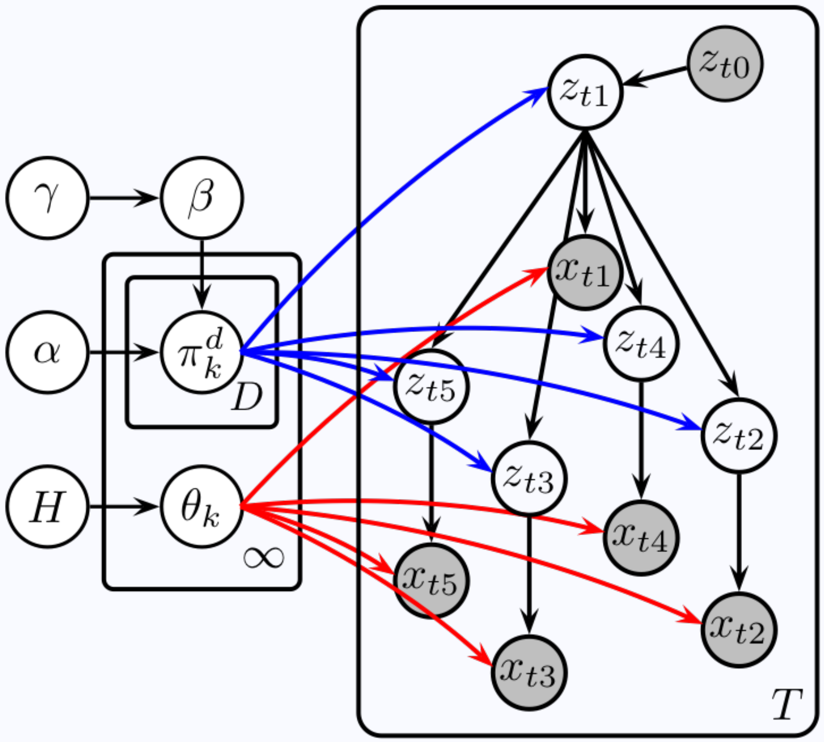
# HDP-HMT Scene Model



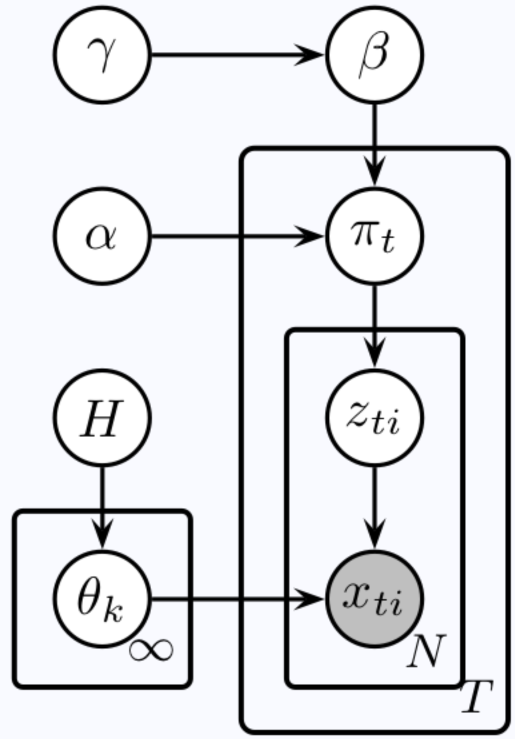

- Hidden states $z_{ti}$ generate vectors of clean wavelet coefficients $x_{ti}$ at multiple orientations, or dense multiscale SIFT descriptors

# … versus baseline HDP-BOF

*HDP-HMT*



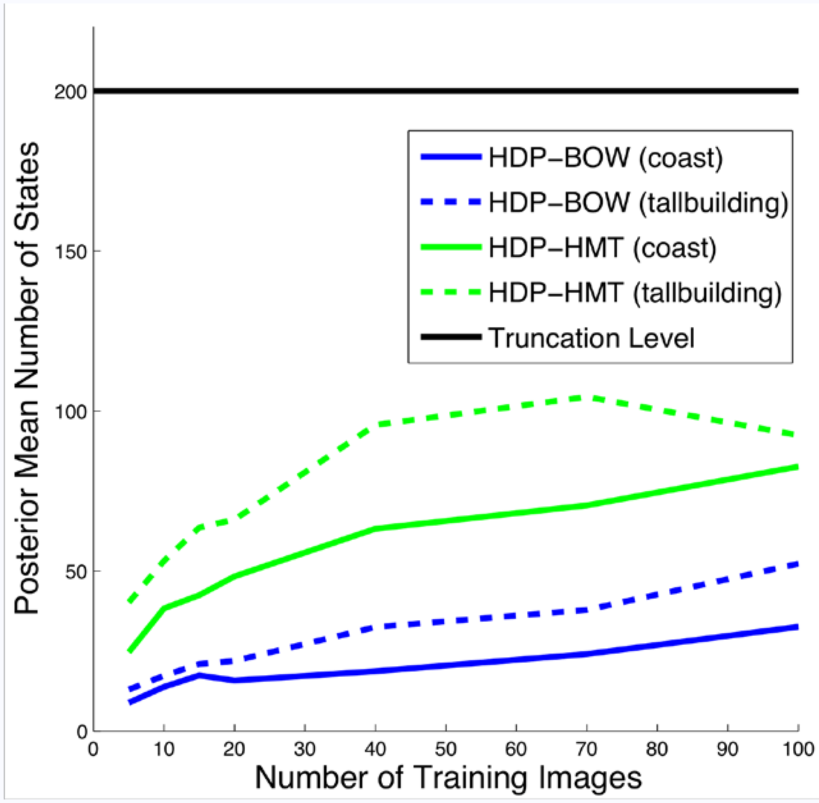

*HDP-BOF*



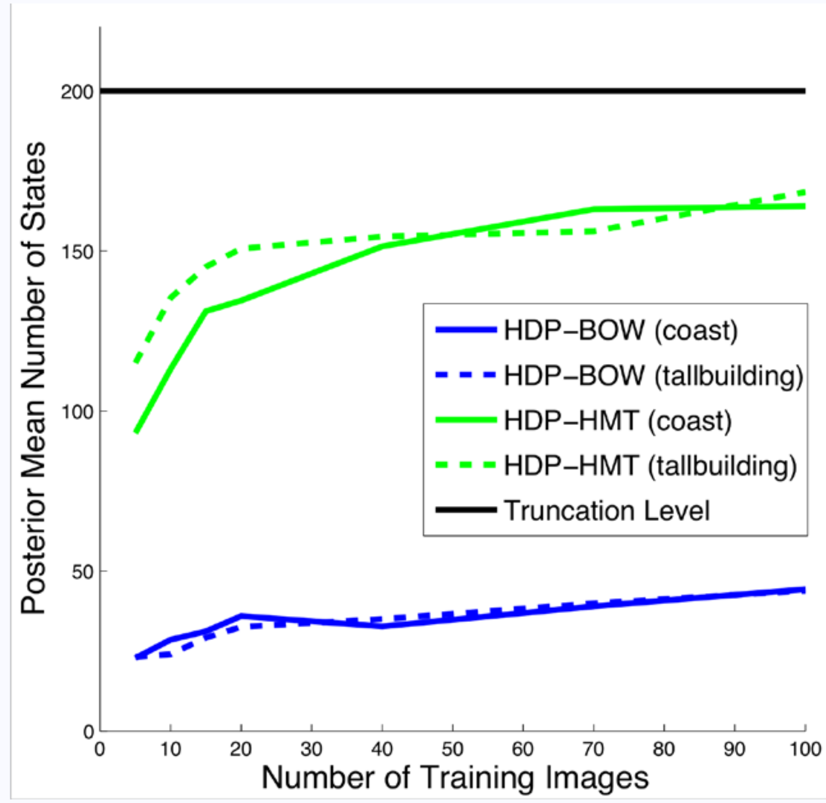

Nonparametric Bayesian extension of LDA scene models (Fei-Fei & Perona, 2005) which ignore spatial locations of locally extracted image features

# Number of States

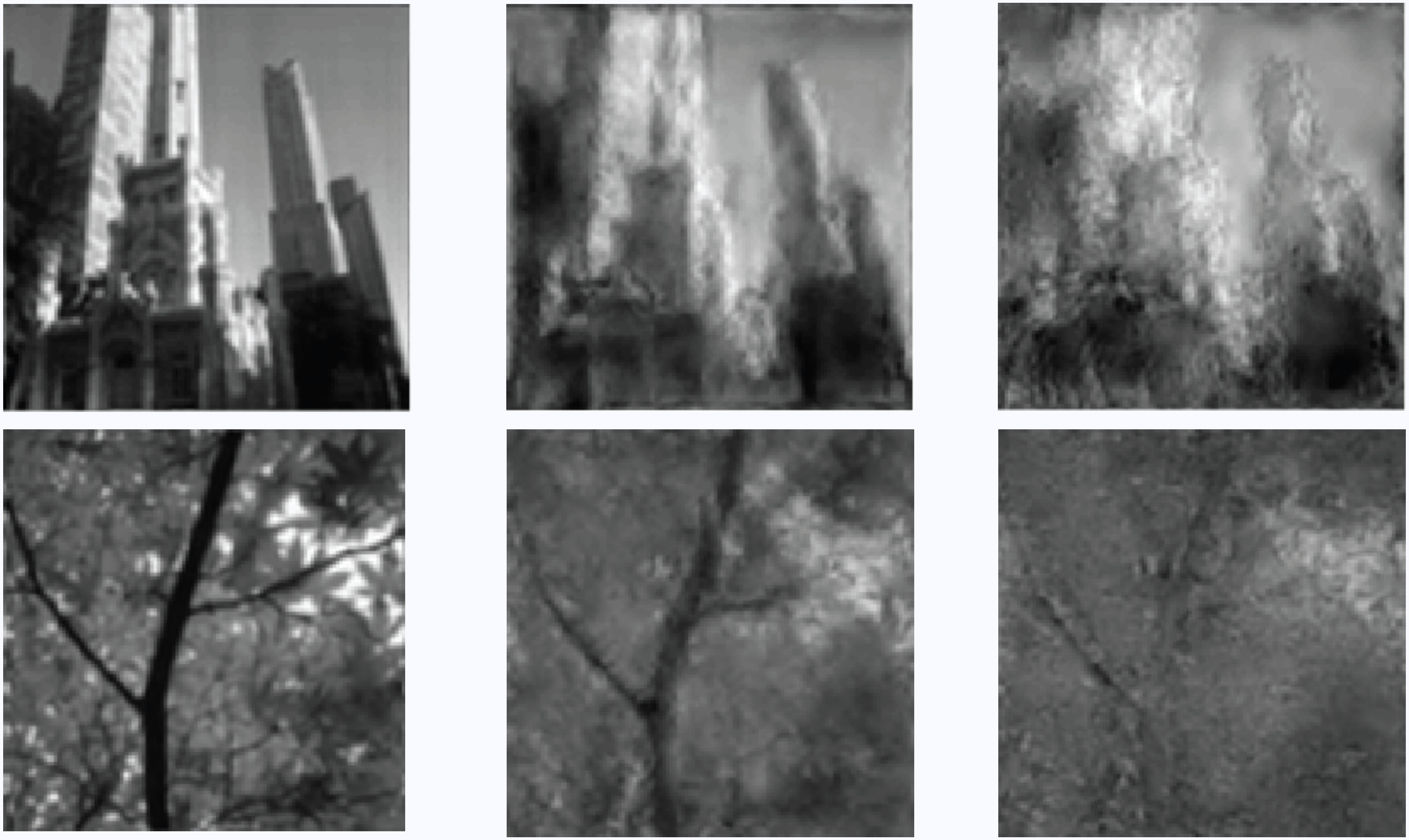
*Wavelet (sp5)*




*SIFT*



# Samples given MAP states



**Input Image** | **HDP Hidden Markov Tree** | **HDP Bag of Features**

# Categorizing Natural Scenes

## *Wavelet (sfp7)*

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| coast | 77.5 | 0.6 | 10.0 | 0.0 | 0.6 | 10.6 | 0.6 | 0.0 |
| forest | 0.0 | 91.4 | 0.0 | 0.0 | 5.5 | 0.8 | 2.3 | 0.0 |
| highway | 3.3 | 0.0 | 75.0 | 0.0 | 10.0 | 10.0 | 1.7 | 0.0 |
| inside city | 0.9 | 0.9 | 2.8 | 77.8 | 0.0 | 3.7 | 9.3 | 4.6 |
| mountain | 0.6 | 13.8 | 4.6 | 0.6 | 63.2 | 9.2 | 8.0 | 0.0 |
| open country | 8.6 | 10.0 | 3.3 | 0.5 | 11.0 | 61.9 | 4.8 | 0.0 |
| street | 0.0 | 1.1 | 5.4 | 2.2 | 7.6 | 0.0 | 81.5 | 2.2 |
| tall building | 0.0 | 0.0 | 2.6 | 13.5 | 0.6 | 0.6 | 8.3 | 74.4 |

*HDP-BOF [75.3 %]*

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| coast | 90.6 | 0.6 | 4.4 | 0.0 | 1.9 | 2.5 | 0.0 | 0.0 |
| forest | 0.0 | 85.2 | 0.8 | 0.0 | 8.6 | 3.1 | 2.3 | 0.0 |
| highway | 8.3 | 0.0 | 80.0 | 0.0 | 6.7 | 0.0 | 3.3 | 1.7 |
| inside city | 1.9 | 0.9 | 5.6 | 75.0 | 0.9 | 0.9 | 10.2 | 4.6 |
| mountain | 1.7 | 1.1 | 2.9 | 0.0 | 91.4 | 1.1 | 1.7 | 0.0 |
| open country | 18.1 | 4.3 | 3.3 | 1.0 | 13.3 | 59.5 | 0.5 | 0.0 |
| street | 0.0 | 0.0 | 8.7 | 1.1 | 7.6 | 0.0 | 81.5 | 1.1 |
| tall building | 0.0 | 0.0 | 1.9 | 12.2 | 0.0 | 0.0 | 3.8 | 82.1 |

*HDP-HMT [80.7 %]*

## *SIFT*

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 90.0 | 0.6 | 1.2 | 0.0 | 1.9 | 6.2 | 0.0 | 0.0 | coast |
| 0.0 | 87.5 | 0.0 | 0.0 | 7.8 | 4.7 | 0.0 | 0.0 | forest |
| 6.7 | 0.0 | 80.0 | 1.7 | 1.7 | 5.0 | 5.0 | 0.0 | highway |
| 0.0 | 0.0 | 1.9 | 87.0 | 0.0 | 0.0 | 9.3 | 1.9 | inside city |
| 1.1 | 0.6 | 0.6 | 0.0 | 90.2 | 5.7 | 0.6 | 1.1 | mountain |
| 11.0 | 1.9 | 1.0 | 0.0 | 5.7 | 80.0 | 0.5 | 0.0 | open country |
| 0.0 | 0.0 | 4.3 | 2.2 | 2.2 | 0.0 | 91.3 | 0.0 | street |
| 0.0 | 0.0 | 0.0 | 9.0 | 0.6 | 0.0 | 4.5 | 85.9 | tall building |

*HDP-BOF [82.4 %]*

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 86.2 | 1.2 | 4.4 | 0.0 | 0.6 | 7.5 | 0.0 | 0.0 | coast |
| 0.0 | 91.4 | 0.0 | 0.0 | 4.7 | 3.1 | 0.8 | 0.0 | forest |
| 6.7 | 0.0 | 75.0 | 1.7 | 3.3 | 6.7 | 6.7 | 0.0 | highway |
| 0.0 | 0.9 | 3.7 | 82.4 | 0.0 | 0.9 | 10.2 | 1.9 | inside city |
| 0.6 | 4.0 | 3.4 | 0.0 | 81.0 | 8.0 | 2.3 | 0.6 | mountain |
| 11.0 | 5.2 | 2.9 | 0.0 | 7.6 | 72.9 | 0.5 | 0.0 | open country |
| 0.0 | 0.0 | 6.5 | 2.2 | 1.1 | 0.0 | 89.1 | 1.1 | street |
| 0.0 | 0.0 | 0.6 | 7.1 | 1.3 | 0.0 | 10.3 | 80.8 | tall building |

*HDP-HMT [86.5 %]*