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
Observations

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

Temporal Segmentation

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

\[
\begin{align*}
\pi_j & \sim P_j \\
\theta_k & \sim \rho_k \\
\alpha & \sim \alpha \\
\lambda & \sim \lambda
\end{align*}
\]

\[
\begin{align*}
z_t & \sim \pi_{z_{t-1}} \\
y_t & \sim F(\theta_{z_t})
\end{align*}
\]

\[
P = \begin{bmatrix}
\pi_1 \\
\pi_2 \\
\vdots \\
\pi_K
\end{bmatrix}
\]

<table>
<thead>
<tr>
<th>Time</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>\ldots</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>\vdots</td>
</tr>
<tr>
<td>K</td>
</tr>
</tbody>
</table>
Hidden Markov Models

\[ \alpha \rightarrow \pi_j \rightarrow \theta_k \rightarrow y_1, y_2, y_3, \ldots, y_T \]

\[ \lambda \]

\[ K \]

\[ \pi_2 \rightarrow \pi_21, \pi_22, \pi_23, \ldots, \pi_2K \]

\[ \mathbb{Z}^+ \]

modes

observations

Time

1 2 3 \ldots

1

2

3

\ldots

K
Hidden Markov Models

\[\alpha, \pi_j, K, \theta_k, K, z_1, z_2, z_3, \ldots, z_T\text{ modes}\]

\[\pi_1, \pi_{12}, \pi_{13}, \ldots, \pi_{1K}\text{ observations}\]

\[\lambda, z_1, z_2, z_3, \ldots, z_T\text{ observations}\]

Time

1 2 3 \ldots

1 2 3 \ldots K
Hidden Markov Models

\[ \pi_j \]

\[ \theta_k \]

\[ z_1 \rightarrow z_2 \rightarrow \ldots \rightarrow z_T \] modes

\[ y_1 \rightarrow y_2 \rightarrow y_3 \rightarrow \ldots \rightarrow y_T \] observations

\[ \pi_3 \]

\[ \pi_{31} \rightarrow \pi_{32} \rightarrow \pi_{33} \rightarrow \pi_{3K} \]

\[ 1 \rightarrow 2 \rightarrow 3 \rightarrow \ldots \rightarrow K \]

Time

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>\ldots</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>\vdots</td>
<td></td>
<td></td>
<td></td>
<td>\vdots</td>
</tr>
<tr>
<td>K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Issue 1: How many modes?

Hierarchical Dirichlet Process HMM

- Dirichlet process (DP):
  - Mode space of unbounded size
  - Model complexity adapts to observations

- Hierarchical:
  - Ties mode transition distributions
  - Shared sparsity

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

Hierarchical Dirichlet Process HMM

- Global transition distribution:
  \( \beta \sim \text{Stick}(\gamma) \)

- Mode-specific transition distributions:
  \( \pi_j \sim \text{DP}(\alpha/\beta) \quad j = 1, 2, 3, \ldots \)

\text{sparsity of } \beta \text{ is shared}
Issue 2: Temporal Persistence

HDP-HMM inferred mode sequence

True mode sequence

Hidden Markov Model
“Sticky” HDP-HMM

\[
\gamma \rightarrow \beta \\
\kappa \rightarrow \pi_j \infty \\
\alpha \rightarrow z_1 \rightarrow z_2 \rightarrow z_3 \rightarrow \ldots \rightarrow z_T \\
\lambda \rightarrow \theta_k \infty \\
y_1 \rightarrow y_2 \rightarrow y_3 \rightarrow \ldots \rightarrow y_T
\]

<table>
<thead>
<tr>
<th>Mode</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>…</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>…</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
“Sticky” HDP-HMM

\[ \beta \sim \text{Stick}(\gamma) \]
\[ \pi_j \sim \text{DP}(\alpha \beta + \kappa \delta_j) \]

**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:
  \[ z_t^{(n)} \sim p(z_t \mid z_{\backslash t}^{(n-1)}, \alpha, \kappa) p(y_t \mid z, y_{\backslash t}, \lambda) \]

Collapsed Gibbs Sampler

Chinese restaurant prior

Splits true mode, hard to merge
Conjugate base measure \( \Rightarrow \) closed form
### Block Resampling

**Approximate HDPs messages:**
- **Average transition density**
  \[ m_{t,t-1}(z_{t-1}) \propto \sum p(z_t | \pi_{z_{t-1}}^{(n)}) p(y_t | \theta_{z_t}^{(n)}) m_{t+1,t}^{(n)}(z_t) \]  
  \( \Rightarrow \) transition densities

**Sample:**
- Block sample \( z_{1:T}^{(n)} \) as:
  \[ z_t^{(n)} \sim p(z_t | \pi_{z_{t-1}}^{(n)}) p(y_t | \theta_{z_t}^{(n)}) m_{t+1,t}^{(n)}(z_t) \]

**HDP-HMM weak limit approximation**

- \( \beta \sim \text{Dir}(\gamma/L, \ldots, \gamma/L) \)
- \( \pi_j \sim \text{Dir}(\alpha \beta_1, \ldots, \alpha \beta_j + \kappa, \ldots, \alpha \beta_L) \)
Results: Gaussian Emissions

Blocked sampler

Sequential sampler

HDP-HMM

Sticky HDP-HMM
Results: Fast Switching

Observations

True mode sequence

Normalized Hamming Distance

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

- Approximate multimodal emissions with DP mixture
- Temporal mode persistence disambiguates model
Speaker Diarization
Results: 21 meetings

<table>
<thead>
<tr>
<th></th>
<th>Overall DER</th>
<th>Best DER</th>
<th>Worst DER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sticky HDP-HMM</td>
<td>17.84%</td>
<td>1.26%</td>
<td>34.29%</td>
</tr>
<tr>
<td>Non-Sticky HDP-HMM</td>
<td>23.91%</td>
<td>6.26%</td>
<td>46.95%</td>
</tr>
<tr>
<td>ICSI</td>
<td>18.37%</td>
<td>4.39%</td>
<td>32.23%</td>
</tr>
</tbody>
</table>
Results: Meeting 1

Sticky DER = 1.26%
ICSII DER = 7.56%
Results: Meeting 18

Sticky DER = 20.48%
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

\[ \theta_k = \text{set of dynamic parameters} \]

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:

\[
\begin{align*}
x_t &= Ax_{t-1} + e_t \\
y_t &= Cx_t + w_t \\
e_t &\sim \mathcal{N}(0, \Sigma) \quad w_t \sim \mathcal{N}(0, R)
\end{align*}
\]

• Vector autoregressive (VAR) process:

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

\[
y_t = \begin{bmatrix} I & 0 & \ldots & 0 \end{bmatrix} x_t.
\]
Switching Dynamical Systems

Switching linear dynamical system (SLDS):

\[ z_t \sim \pi_{z_{t-1}} \]
\[ x_t = A^{(z_t)} x_{t-1} + e_t(z_t) \]
\[ y_t = C x_t + w_t \]
\[ e_t \sim \mathcal{N}(0, \Sigma^{(z_t)}) \quad w_t \sim \mathcal{N}(0, R) \]

Switching VAR process:

\[ z_t \sim \pi_{z_{t-1}} \]
\[ y_t = \sum_{i=1}^{r} A_i^{(z_t)} y_{t-i} + e_t(z_t) \]
\[ e_t \sim \mathcal{N}(0, \Sigma^{(z_t)}) \]
HDP-AR-HMM and HDP-SLDS

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

\[ z_t \sim \pi_{z_{t-1}} \]

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

\[ x_t = A^{(z_t)} x_{t-1} + e_t(z_t) \]

\[ y_t = C x_t + w_t \]

\[ C = [I \ 0] \]
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

Features/Modes

- Beta process prior:
  - Encourages sharing
  - Unbounded features

\[
\pi_j^{(i)} \mid f_i, \gamma, \kappa \sim \text{Dir}([\gamma, \ldots, \gamma, \gamma + \kappa, \gamma, \ldots]) \otimes f_i \\
\gamma \sim \pi^{(i)} \otimes_{z_t^{(i-1)}} \\
y_t^{(i)} = \sum_{j=1}^r A_{j,z_t^{(i)}} y_{t-j}^{(i)} + e_t^{(i)}(z_t^{(i)})
\]
Motion Capture

6 videos of exercise routines: CMU MoCap: http://mocap.cs.cmu.edu/
Library of MoCap Behaviors
CMU Kitchen Dataset

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.