Infinite Structured Explicit Duration
Hidden Markov Models

Frank Wood

Joint work with Chris Wiggins, Mike Dewar, and Jonathan Huggins

Columbia University

November, 2011
Hidden Markov models (HMMs) [Rabiner, 1989] are an important tool for data exploration and engineering applications.

Applications include

- Natural language processing [Manning and Schütze, 1999]
- Hand-writing recognition [Nag et al., 1986]
- DNA and other biological sequence modeling applications [Krogh et al., 1994]
- Financial data modeling [Rydén et al., 1998]
- ... and many more.
Notation: Hidden Markov Model

\[ z_t | z_{t-1} = m \sim \text{Discrete}(\pi_m) \]
\[ y_t | z_t = m \sim F_\theta(\theta_m) \]

\[ A = \begin{bmatrix}
  \vdots & \vdots & \vdots & \vdots \\
  \pi_1 & \cdots & \pi_m & \cdots & \pi_K \\
  \vdots & \vdots & \vdots & \vdots 
\end{bmatrix} \]
Training data: multiple “observed” $y_t = \{v_t, h_t\}$ sequences of stylus positions for each kind of character

Task: train $|\Sigma|$ different models, one for each character

Latent states: sequences of strokes

Usage: classify new stylus position sequences using trained models $\mathcal{M}_\sigma = \{A_\sigma, \Theta_\sigma\}$

\[
P(\mathcal{M}_\sigma | y_1, \ldots, y_T) \propto P(y_1, \ldots, y_T | \mathcal{M}_\sigma)P(\mathcal{M}_\sigma)
\]
Shortcomings of Original HMM Specification

- Latent state dwell times are not usually geometrically distributed

\[
P(z_t = m, \ldots, z_{t+L} = m | A) = \prod_{\ell=1}^{L} P(z_{t+\ell+1} = m | z_{t+\ell} = m, A) = \text{Geometric}(L; \pi_m(m))
\]

- Prior knowledge often suggests structural constraints on allowable transitions

- The state cardinality of the latent Markov chain $K$ is usually unknown

- Latent state sequence \( z = (\{s_1, r_1\}, \ldots, \{s_T, r_T\}) \)
- Latent state id sequence \( s = (s_1, \ldots, s_T) \)
- Latent “remaining duration” sequence \( r = (r_1, \ldots, r_T) \)
- State-specific duration distribution \( F_r(\lambda_m) \)
- Other distributions the same

An EDHMM transitions between states in a different way than does a typical HMM. Unless \( r_t = 0 \) the current remaining duration is decremented and the state does not change. If \( r_t = 0 \) then the EDHMM transitions to a state \( m \neq s_t \) according to the distribution defined by \( \pi_{s_t} \)
EDHMM notation

Latent state $z_t = \{s_t, r_t\}$ is tuple consisting of state identity and time left in state.

\[
s_t \mid s_{t-1}, r_{t-1} \sim \begin{cases} 
\mathbb{I}(s_t = s_{t-1}), & r_{t-1} > 0 \\
\text{Discrete}(\pi_{s_{t-1}}), & r_{t-1} = 0 
\end{cases}
\]

\[
r_t \mid s_t, r_{t-1} \sim \begin{cases} 
\mathbb{I}(r_t = r_{t-1} - 1), & r_{t-1} > 0 \\
F_r(\lambda_{s_t}), & r_{t-1} = 0 
\end{cases}
\]

\[
y_t \mid s_t \sim F_\theta(\theta_{s_t})
\]
**Structured HMMs: i.e. left-to-right HMM [Rabiner, 1989]**

**Example: Chicken pox**

**Observations** vital signs

**Latent states** pre-infection, infected, post-infection

**State transition structure** can’t go from infected to pre-infection

\(^a\)disregarding shingles

Structured transitions imply zeros in the transition matrix \( A \), i.e.

\[
p(s_t = m | s_{t-1} = \ell) = 0 \quad \forall \ m < \ell
\]
We will put a prior on parameters so that we can effect a solution that conforms to our ideas about what the solution should look like.

Structured prior examples

- \( A_{i,j} = 0 \) (hard constraints)
- \( A_{i,j} \approx \sum_j A_{i,j} \) (rich get richer)

Regularization means that we can specify a model with more parameters than could possibly be needed.

- Infinite complexity (i.e. \( K \rightarrow \infty \)) avoids many model selection problems.
- “Extra” states can be thought of as auxiliary or nuisance variables.
Bayesian HMM

\[
\begin{align*}
\pi_m \sim H_z \\
\theta_m \sim H_y
\end{align*}
\]

\[
z_t | z_{t-1} = m \sim \text{Discrete}(\pi_m)
\]

\[
y_t | z_t = m \sim F_\theta(\theta_m)
\]
Other HMM variants

- Sticky Infinite HMMs [Fox et al., 2011]
  - Extra parameter per state used to bias towards self-transition
- Hierarchical HMMs [Murphy, 2001]
  - State is hierarchical (i.e. sequence of letters composed of stroke sequences)
- Factorial HMMs [Ghahramani and Jordan, 1996]
- Infinite explicit duration HMM [Johnson and Willsky, 2010]
  - No generative model. Latent history sampling used to assert the existence of an implicitly defined IEDHMM that can be sampled from by rejecting HDP-HMM samples that violate transition constraints.
Generative framework for HMMs with
- Explicitly parameterized duration distributions
- Structured transition priors
- Countable state cardinality

Fundamental Problems
- How to generate structured, dependent, infinite-dimensional transition distributions.
- How to do inference in HMMs with countable state cardinality and countable duration distribution support.

[Huggins and W, 2011]
Structured, dependent, infinite-dimensional transition distributions?
ISEDHMM: Recipe

- Poisson process [Kingman, 1993]
- Gamma process [Kingman, 1993]
- SNΓPs [Rao and Teh, 2009]
- *Structured, dependent, infinite-dimensional* transition distributions [Huggins and W, 2011]
- Gamma process [Kingman, 1993] as Poisson process over $\Theta \otimes V \otimes [0, \infty)$ with rate / mean measure

$$\mu(\tilde{\Theta}, \tilde{V}, \tilde{S}) = \alpha(\tilde{\Theta}, \tilde{V}) \int_{\tilde{S}} \gamma^{-1} e^{-\gamma}$$

- A draw from a Gamma process with

$$\alpha(\tilde{\Theta}, \tilde{V}) = c_0 H_\theta(\tilde{\Theta}) H_v(\tilde{V}).$$

[Kingman, 1993] has the form

$$G = \sum_{m=1}^{\infty} \gamma_m \delta(\theta_m, v_m)$$

where $(\theta_m, v_m) \sim H_\theta \times H_v.$
Non-disjoint “restricted projections” of Gamma processes are *dependent* Gamma processes (SNΓPs) [Rao and Teh, 2009]

\[ H_v = \text{Geom}(p) \]

\[ G_0 = \sum_{m \neq 0} \cdots, \quad \cdots, \quad G_4 = \sum_{m \neq 4} \gamma_m \delta_{\theta_m}, \quad \cdots \]
Normalized dependent GP draws are dependent Dirichlet process draws. In the ISEDHMM, DP draws are the dependent, structured, infinite-dimensional transition distributions.

\[ D_4 = \frac{G_4}{G_4(\Theta)} \]

\[ = \frac{\sum_{m \neq 4} \gamma_m \delta_{\theta_m}}{\sum_{\theta \in \Theta} \sum_{m' \neq 4} \gamma_{m'} \delta_{\theta_{m'}}} \]

\[ = \frac{\sum_{m \neq 4} \gamma_m \delta_{\theta_m}}{\sum_{m' \neq 4} \gamma_{m'}} \]

\[ = \sum_{m \neq 4} \frac{\gamma_m}{\sum_{m' \neq 4} \gamma_{m'}} \delta_{\theta_m} \]

\[ ^1 \text{a draw from a Dirichlet process [Ferguson, 1973] is an infinite sum of weighted atoms [Sethuraman, 1994] where the weights sum to one.} \]
Structured, dependent, infinite dimensional transition distributions $\pi_m$ can be formed from draws from DDPs [Huggins and W, 2011]
We employ the forward-filtering, backward slice-sampling approach for the IHMM of [Van Gael et al., 2008] and EDHMM of [Dewar, Wiggins and W, 2011], in which the state and duration variables $s$ and $r$ are sampled conditioned on auxiliary slice variables $u$.

Net result: efficient, always finite forward backward procedure for sampling latent states.
Objective: get samples of $x$. 

Sometimes it is easier to introduce an auxiliary variable $u$ and to Gibbs sample the joint $P(x, u)$ (i.e. sample from $P(x | u; \lambda)$ then $P(u | x, \lambda)$), etc. then discard the $u$ values than it is to directly sample from $p(x | \lambda)$. Useful when: $p(x | \lambda)$ does not have a known parametric form but adding $u$ results in a parametric form and when $x$ has countable support and sampling it requires enumerating all values.
Objective: get samples of $x$.

Sometimes it is easier to introduce an auxiliary variable $u$ and to Gibbs sample the joint $P(x, u)$ (i.e. sample from $P(x|u; \lambda)$ then $P(u|x, \lambda)$, etc.) then discard the $u$ values than it is to directly sample from $p(x|\lambda)$. 

Useful when: $p(x|\lambda)$ does not have a known parametric form but adding $u$ results in a parametric form and when $x$ has countable support and sampling it requires enumerating all values.
Auxiliary Variables for Sampling

Objective: get samples of $x$.

Sometimes it is easier to introduce an auxiliary variable $u$ and to Gibbs sample the joint $P(x, u)$ (i.e. sample from $P(x|u; \lambda)$ then $P(u|x, \lambda)$, etc.) then discard the $u$ values than it is to directly sample from $p(x|\lambda)$.

Useful when: $p(x|\lambda)$ does not have a known parametric form but adding $u$ results in a parametric form and when $x$ has countable support and sampling it requires enumerating all values.
Unreasonable Pedagogical Example:

- \( x | \lambda \sim \text{Poisson}(\lambda) \) (countable support)
- *enumeration* strategy for sampling \( x \) (impossible)
- auxiliary variable \( u \) with \( P(u|x, \lambda) = \frac{\mathbb{I}(0 \leq u \leq P(x|\lambda))}{P(x|\lambda)} \)

Note: Marginal distribution of \( x \) is

\[
P(x|\lambda) = \sum_u P(x, u|\lambda)
= \sum_u P(x|\lambda) P(u|x, \lambda)
= \sum_u P(x|\lambda) \frac{\mathbb{I}(0 \leq u \leq P(x|\lambda))}{P(x|\lambda)}
= \sum_u \mathbb{I}(0 \leq u \leq P(x|\lambda)) = P(x|\lambda)
\]
Slice Sampling: A very useful auxiliary variable sampling trick

This suggests a Gibbs sampling scheme: alternately sampling from

- \( P(x|u, \lambda) \propto \mathbb{I}(u \leq P(x|\lambda)) \)
  - finite support, uniform above slice, enumeration possible

- \( P(u|x, \lambda) = \frac{\mathbb{I}(0 \leq u \leq P(x|\lambda))}{P(x|\lambda)} \)
  - uniform between 0 and \( y = P(x|\lambda) \)

then discarding the \( u \) values to arrive at \( x \) samples marginally distributed according to \( P(x|\lambda) \).
Forward-backward slice sampling only has to consider a finite number of successor states at each timestep. With auxiliary variables

\[
p(u_t | z_t, z_{t-1}) = \frac{\mathbb{I}(u_t < p(z_t | z_{t-1}))}{p(z_t | z_{t-1})}
\]

and

\[
p(z_t | z_{t-1}) = p((s_t, r_t) | (s_{t-1}, r_{t-1}))
= \begin{cases} 
  r_{t-1} > 0, & \mathbb{I}(s_t = s_{t-1})\mathbb{I}(r_t = r_{t-1} - 1) \\
  r_{t-1} = 0, & \pi_{s_{t-1}s_t} F_r(r_t; \lambda_{s_t}).
\end{cases}
\]

one can run standard forward-backward conditioned on \( u \)'s.
To illustrate IEDHMM learning on synthetic data, five hundred datapoints were generated using a 4 state EDHMM with Poisson duration distributions

$$\lambda = (10, 20, 3, 7)$$

and Gaussian emission distributions with means

$$\mu = (-6, -2, 2, 6)$$

all unit variance.
IEDHMM: Synthetic Data, State Duration Parameter Posterior

Wood (Columbia University)
Novel Gamma process construction for dependent, structured, infinite dimensional HMM transition distributions.

Other transition distribution structures (left-to-right) can be implemented simply by changing “restricted projection regions.”

ISEDHMM framework generalizes the HMM, Bayesian HMM, infinite HMM, left-to-right HMM, explicit duration HMM, and more.
Future Work

- Generalize to spatial prior on HMM states ("location")
  - Simultaneous location and mapping
  - Process diagram modeling for systems biology

- Applications; seeking "users"
Questions?

Thank you!


Left: data from Poisson duration distribution ISEDHMM, Right: data from IHMM fit with Poisson duration distribution
Poisson process can be defined by the requirement that the random variables defined as the counts of the number of “events” inside each of a number of non-overlapping finite sub-regions of some space should each have a Poisson distribution and should be independent of each other.

\[ N(A) \sim \text{Poisson}(\mu(A)) \]
Define a Poisson process over the product space $\Theta \otimes [0, \infty)$ with mean measure

$$\mu(\tilde{\Theta}, \tilde{S}) = \alpha(\tilde{\Theta}) \int_{\tilde{S}} \gamma^{-1} e^{-\gamma}$$

where $\tilde{\Theta} \in \Omega$ and $\tilde{S} \subset [0, \infty)$. A draw from this Poisson process yields a countably infinite set of pairs $\{(\theta_n, \gamma_n)\}_{n \geq 1}$, which can be used to form an atomic random measure

$$G = \sum_{n \geq 1} \gamma_n \delta_{\theta_n}.$$
Gamma Process

This discrete measure is drawn from $G \sim \Gamma P(\alpha)$ (from previous slide)

$$G = \sum_{n \geq 1} \gamma_n \delta_{\theta_n}.$$ 

A $\Gamma P$ can be thought of as an unnormalized DP.
$D = G/G(\Theta)$ is a sample from a DP with base measure $\alpha$

$$D \sim \text{DP}(\alpha).$$

A draw from a Dirichlet process is an infinite mixture of weighted, discrete atoms.

For the ISEDHMM

- each atom is a next state (there are a countably infinite number of such states)
- each atoms' weight is the probability of transitioning to that state
- there will be a countably infinite number of such transition distributions
One way to define a set of dependent DPs is to construct a base gamma process over an augmented space by taking the union of disjoint independent gamma processes, then define a series of restricted projections of that base process, which are themselves gamma processes.
Structured Transitions Via Dependent Dirichlet Processes

- One way to define a set of dependent DPs is to construct a base gamma process over an augmented space by taking the union of disjoint independent gamma processes, then define a series of restricted projections of that base process, which are themselves gamma processes.

- The normalization of these dependent gamma processes form a set of dependent DPs.
One way to define a set of dependent DPs is to construct a base gamma process over an augmented space by taking the union of disjoint independent gamma processes, then define a series of restricted projections of that base process, which are themselves gamma processes.

The normalization of these dependent gamma processes form a set of dependent DPs.

We will use this procedure to construct a number of dependent DPs (one for each HMM state) which preclude certain transitions. The precluded transitions, a form of dependence, arise from particulars.
SNGP construction of IEDHMM transition distributions

\[ G = \{ G_{R_1}, G_{R_2}, G_{R_3}, G_{R_M}, G_{R_+} \} \]

\[ \tilde{D}_1, \pi_1 \]
\[ \tilde{D}_2, \pi_2 \]
\[ \tilde{D}_3, \pi_3 \]
\[ \tilde{D}_M, \pi_M \]

[Huggins and W, 2011]
Spatial Normalized Gamma Processes [Rao and Teh, 2009]

To review SNΓPs formally, let \( \mathcal{V} \) be an arbitrary auxiliary space. In general, one can think of this as a covariate, time, or an index. For \( V \subset \mathcal{V} \), let

\[
\alpha(\tilde{\Theta}, V) = c_0 H_\theta(\tilde{\Theta}) H_v(V)
\]

be the base measure for a gamma process \( \mathcal{G} = \Gamma P(\alpha) \) defined over the product space \( \Theta \otimes \mathcal{V} \). Here \( c_0 \) is a concentration parameter and \( H_\theta \) and \( H_v \) are probability measures. We will refer to \( \mathcal{G} \) as the “base ΓP”.
Let $\mathbb{T}$ be an index set and define restricted projected measures $\alpha_m$ for all $m \in \mathbb{T}$ such that

$$\alpha_m(\tilde{\Theta}) = \alpha(\tilde{\Theta}, V_m)$$

where $V_m$ is a subset of $\mathbb{V}$ indexed by $m$. 

(Spatial Normalized Gamma Processes [Rao and Teh, 2009])
Let $\mathbb{T}$ be an index set and define restricted projected measures $\alpha_m$ for all $m \in \mathbb{T}$ such that

$$\alpha_m(\tilde{\Theta}) = \alpha(\tilde{\Theta}, V_m)$$

where $V_m$ is a subset of $\mathbb{V}$ indexed by $m$.

The SNΓP gets its name from thinking of $\mathbb{V}$ as a space and $V_m$ as a region of space indexed by $m$. With $G \sim \mathcal{G}$ being a draw from the base gamma process, define the restricted projection $G_m$ by

$$G_m(\tilde{\Theta}) = G(\tilde{\Theta}, V_m).$$
Let $\mathcal{T}$ be an index set and define restricted projected measures $\alpha_m$ for all $m \in \mathcal{T}$ such that

$$\alpha_m(\tilde{\Theta}) = \alpha(\tilde{\Theta}, V_m)$$

where $V_m$ is a subset of $\mathcal{V}$ indexed by $m$.

The SNΓP gets its name from thinking of $\mathcal{V}$ as a space and $V_m$ as a region of space indexed by $m$. With $G \sim \mathcal{G}$ being a draw from the base gamma process, define the restricted projection $G_m$ by

$$G_m(\tilde{\Theta}) = G(\tilde{\Theta}, V_m).$$

Then $G_m(\tilde{\Theta})$ is distributed according to a gamma process with base measure $\alpha_m$

$$G_m \sim \Gamma P(\alpha_m).$$
Normalizing $G_m$ yields a draw $D_m = G_m / G_m(\Theta)$ distributed according to a Dirichlet process with base measure $\alpha_m$

$$D_m \sim \text{DP}(\alpha_m).$$

$D_m(\tilde{\Theta})$ is not in general independent of $D_{m'}(\tilde{\Theta})$ because they can share atoms from $G$. 
Recall the gamma process $\mathcal{G} = \Gamma \mathcal{P}(\alpha)$ with base measure

$$\alpha(\tilde{\Theta}, V) = c_0 H_{\theta}(\tilde{\Theta}) H_{v}(V).$$

Draws $G \sim \mathcal{G}$ are measures over the parameter and covariate product space $\Theta \otimes V$ with the form

$$G = \sum_{m=1}^{\infty} \gamma_m \delta_{(\theta_m, v_m)}$$

where $(\theta_m, v_m) \sim H_{\theta} \times H_{v}$.

In the IEDHMM $\theta_m = \{\lambda_m, \theta_m\}$ is the duration parameter and output distribution parameters for state $m$.

For pedagogical expediency let $T = V = \{0, 1, 2, 3, 4, \ldots\}$ then $v_m \in \mathbb{N}$.
\[ R_m = \{ v_m \} \quad m \in \mathcal{T} \]
\[ R_+ = \mathcal{V} \setminus \mathcal{V} \]
\[ \mathcal{R} = \{ R_m : m \in \mathcal{T} \} \cup \{ R_+ \} \]
\[ \mathcal{R}_m = \mathcal{R} \setminus \{ R_m \} \quad m \in \mathcal{T} \]
$H_v = \text{Geom}(p)$
From before, using the restricted projection $\alpha_R(\tilde{\Theta}) = \alpha(\tilde{\Theta}, R)$ setting $G_R(\tilde{\Theta}) = G(\tilde{\Theta}, R)$ and $D_R = G_R / G_R(\Theta)$ we have

$$
G_R \sim \Gamma P(\alpha_R)
$$

$$
D_R \sim \Delta P(\alpha_R).
$$
From before, using the restricted projection $\alpha_R(\tilde{\Theta}) = \alpha(\tilde{\Theta}, R)$ setting $G_R(\tilde{\Theta}) = G(\tilde{\Theta}, R)$ and $D_R = G_R/G_R(\Theta)$ we have

\[ G_R \sim \Gamma P(\alpha_R) \]
\[ D_R \sim DP(\alpha_R). \]

In the case of the IEDHMM, the base measure corresponding to the point region $R_m \in \mathcal{R}$ is

\[ \alpha_{R_m}(\tilde{\Theta}) = c_0 \delta_{\theta_m}(\tilde{\Theta}) H_v(R_m), \]

while for $R_+$ it is

\[ \alpha_{R_+}(\tilde{\Theta}) = c_0 H_\theta(\tilde{\Theta}) H_v(R_+). \]
A *highly* dependent transition distribution for state $m$ is a DP draw $D_m$

$$D_m = \frac{G_m}{G_m(\Theta)} = \frac{\sum_{R \in \mathcal{R}_m} G_R}{\sum_{R' \in \mathcal{R}_m} G_{R'}(\Theta)}$$

$$= \sum_{R \in \mathcal{R}_m} \frac{G_R(\Theta) G_R}{\sum_{R' \in \mathcal{R}_m} G_{R'}(\Theta) G_{R'}}$$

$$= \sum_{R \in \mathcal{R}_m} \frac{G_R(\Theta)}{\sum_{R' \in \mathcal{R}_m} G_{R'}} D_R$$
\[ D_{R_m} \sim \text{DP}(\alpha_{R_m}) \]
\[ D_{R_+} \sim \text{DP}(\alpha_{R_+}) \]
\[ \gamma_m \sim \text{Gamma}(c_0 H_v(R_m), 1) \]
\[ \gamma_+ \sim \text{Gamma}(c_0 H_v(R_+), 1) \]
\[ \beta_{mk} = \frac{\mathbb{I}(m \neq k) \gamma_k}{\gamma_+ + \sum_{k' \neq m}^{M} \gamma_{k'}} \]
\[ \beta_{m+} = \frac{\gamma_+}{\gamma_+ + \sum_{k' \neq m}^{M} \gamma_{k'}} \]
\[ D_m = \sum_{k \neq m}^{M} \beta_{mk} D_{R_k} + \beta_{m+} D_{R_+}. \]
These dependent DPs are the base dist.'s for conditional state transition distributions \( \tilde{D}_m \sim \text{DP}(c_1 D_m) \). With \( \beta_m = (\beta_{m1}, \ldots, \beta_{mM}, \beta_{m+}) \)

\[
\pi_m \sim \text{Dirichlet}(c_1 \beta_m),
\]

\[
\tilde{D}_m = \sum_{k=1}^{M} \pi_{mk} \delta_{\theta_m} + \pi_{m+} D_{R+}
\]

where \( c_1 \) is a concentration parameter and \( \pi_m = (\pi_{m1}, \ldots, \pi_{mM}, \pi_{m+}) \).

The conditional state transition probability row vector \( \pi_m \) is finite, since probabilities of transitioning to new states have been merged into a single probability \( \pi_{m+} = \sum_{k=M+1}^{\infty} \pi_{mk} \). This “bin” is dynamically split and joined at inference time.