Hidden Markov models: from the beginning to the state of the art

Frank Wood

Columbia University

November, 2011

Outline

- Overview of hidden Markov models from Rabiner tutorial to now
- EDHMM
  - Gateway to state of the art models
  - Inference
- Tips and tricks for Bayesian inference in general (auxiliary variables and slice sampling)
- Toy examples

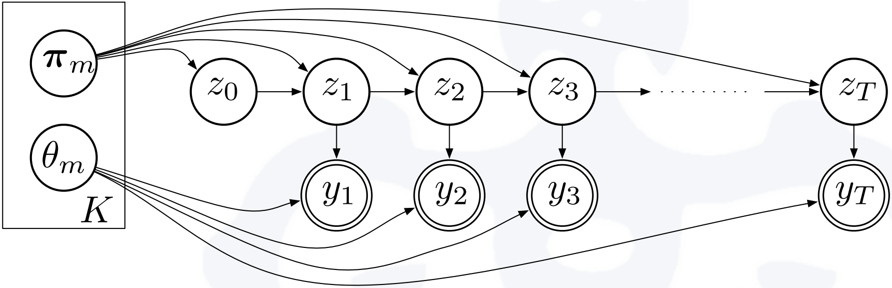
Hidden Markov Models

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

Applications include

- Speech recognition [Jelinek, 1997, Juang and Rabiner, 1985]
- 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]
- Gesture recognition [Tanguay Jr, 1995, Wilson and Bobick, 1999]
- Financial data modeling [Rydén et al., 1998]
- . . . and many more.

Graphical Model: Hidden Markov Model

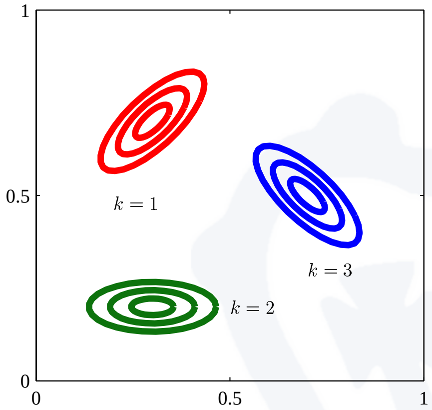


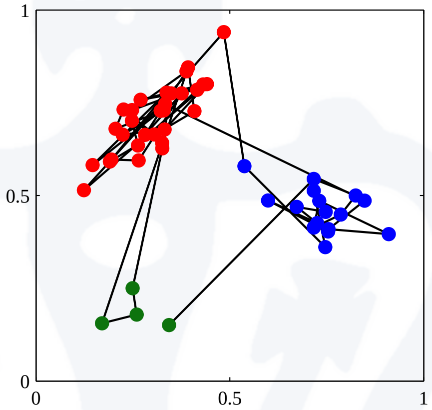
Notation: Hidden Markov Model

$$
\begin{aligned}
z_t | z_{t-1} = m &\sim \text{Discrete}(\boldsymbol{\pi}_m) \\
y_t | z_t = m &\sim F_\theta(\theta_m)
\end{aligned}
$$

$$
\mathbf{A} = \begin{bmatrix} \vdots & & \vdots & & \vdots \\ \boldsymbol{\pi}_1 & \cdots & \boldsymbol{\pi}_m & \cdots & \boldsymbol{\pi}_K \\ \vdots & & \vdots & & \vdots \end{bmatrix}
$$

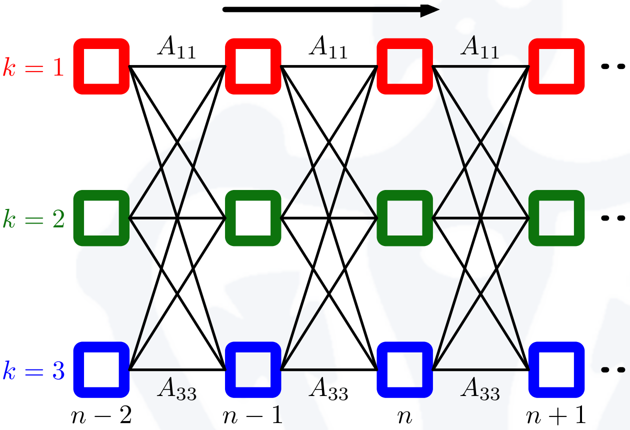
HMM: Dynamic mixture model





Visualization from PRML. [Bishop, 2006]

HMM: Trellis



Visualization from PRML. [Bishop, 2006]

HMM: Typical Usage Scenario (Character Recognition)

- 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: design for correspondence with 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

$$\begin{aligned} & P(z_t = m, \ldots, z_{t+L} = m | A) \\ = & \prod_{\ell=1}^{L} P(z_{t+\ell+1} = m | z_{t+\ell} = m, A) \\ = & \text{Geometric}(L; \boldsymbol{\pi}_m(m)) \end{aligned}$$

- There are often problem-specific structural constraints on allowable transitions, i.e. $A_{i,i} = 0$
- The state cardinality of the latent Markov chain K is usually unknown

Explicit Duration HMM / H. Semi-Markov Model

[Mitchell et al., 1995, Murphy, 2002, Yu and Kobayashi, 2003, Yu, 2010]

- Latent state sequence $\mathbf{z} = (\{s_1, r_1\}, \ldots, \{s_T, r_T\})$
- Latent state id sequence $\mathbf{s} = (s_1, \ldots, s_T)$
- Latent "remaining duration" sequence $\mathbf{r} = (r_1, \ldots, r_T)$
- 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 $\boldsymbol{\pi}_{s_t}$

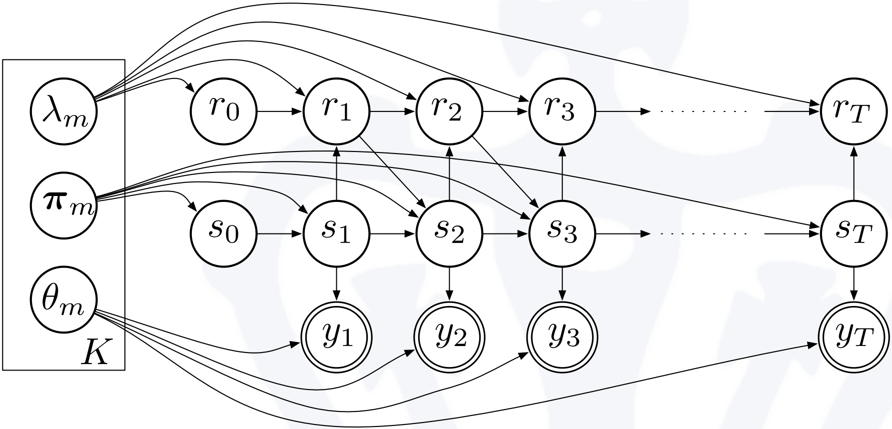
Problem: inference requires enumerating possible durations.

EDHMM notation

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

$$
\begin{aligned}
s_t | s_{t-1}, r_{t-1} &\sim \begin{cases} \mathbb{I}(s_t = s_{t-1}), & r_{t-1} > 0 \\ \text{Discrete}(\boldsymbol{\pi}_{s_{t-1}}), & r_{t-1} = 0 \end{cases} \\
r_t | s_t, r_{t-1} &\sim \begin{cases} \mathbb{I}(r_t = r_{t-1} - 1), & r_{t-1} > 0 \\ F_r(\lambda_{s_t}), & r_{t-1} = 0 \end{cases} \\
y_t | s_t &\sim F_\theta(\theta_{s_t})
\end{aligned}
$$

EDHMM: Graphical Model



Structured HMMs: i.e. left-to-right HMM [Rabiner, 1989]

Example: Chicken pox

Observations vital signs

Latent states pre-infection, infected, post-infection[a]

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. (for a left-to-right HMM)

$$p(s_t = m | s_{t-1} = \ell) = 0 \; \forall \; m < \ell$$

Structured HMM: Trellis: One step at a time, left-to-right HMM

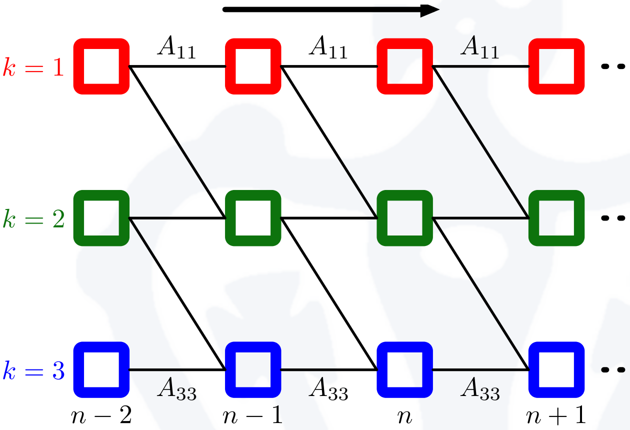
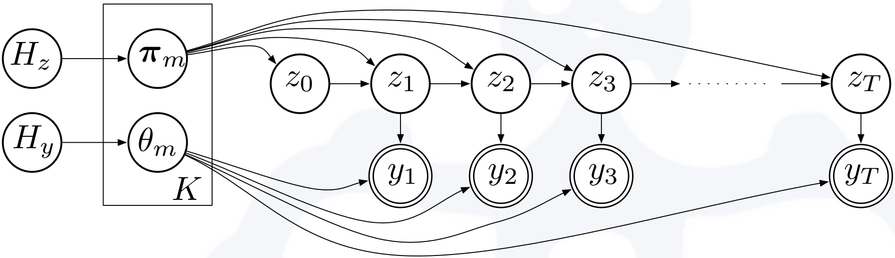


Figure from PRML. [Bishop, 2006]

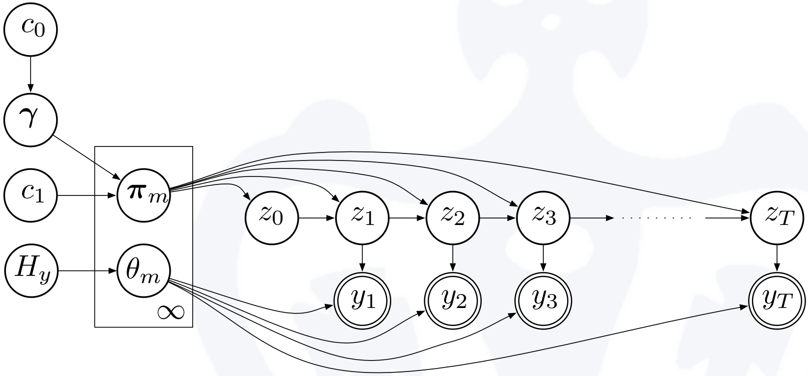
Bayesian HMM

- 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
    - $A_{i,j} = 0$ (hard constraints)
    - $A_{i,j} \approx \sum_j A_{i,j}$ (rich get richer)
- Bayesian *regularization* means that we can specify a model with more parameters than could possibly be needed
  - infinite complexity (i.e. $K \to \infty$) avoids many model selection problems
  - "extra" states can be thought of as auxiliary or nuisance variables
  - inference requires sampling in a model with countably infinite support
- Posterior over latent variables encodes uncertainty about interpretation of data.



$$\pi_m \sim H_z$$
$$\theta_m \sim H_y$$

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



$$K \to \infty,$$

Sticky IHMM [Fox et al., 2011] = IHMM with up-weighted self-transitions

Inference for the Explicit Duration HMM (EDHMM)

Simple Idea

- Infinite HMMs and EDHMMs share fundamental characteristic: countable support
- Inference techniques for Bayesian nonparametric (infinite) HMMs can be used for EDHMM inference

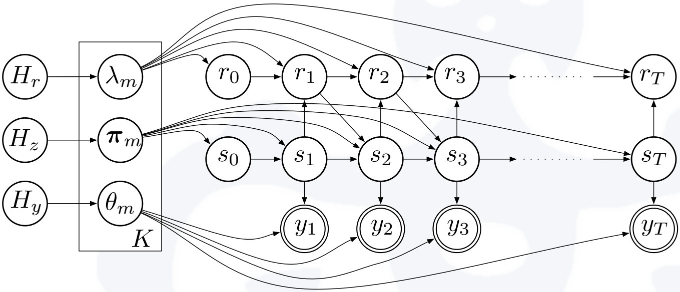
Result

- Approximation-free, efficient inference algorithm for EDHMM inference

Utility

- New HMM for you to try in your applications
- Gateway to understanding and dealing with infinite state cardinality variants

Joint work with Chris Wiggins (Columbia), Mike Dewar (Bitly)



A choice of prior

$$\begin{aligned}\lambda_m|H_r &\sim \mathrm{Gamma}(H_r)\\ \pi_m|H_z &\sim \mathrm{Dir}(1/K, 1/K, \ldots, 1/K, 0, 1/K, \ldots, 1/K, 1/K)\end{aligned}$$

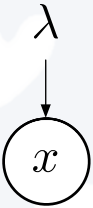
EDHMM Inference: Beam Sampling [Dewar, Wiggins and W, 2011]

We employ the forward-filtering, backward slice-sampling approach for the IHMM of [Van Gael et al., 2008] , 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 and the amount of time spent in them.

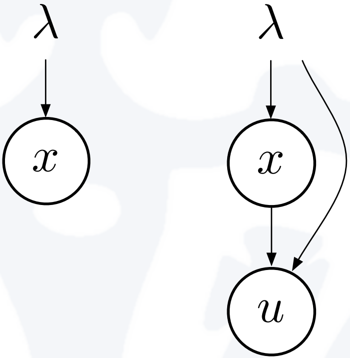
Auxiliary Variables for Sampling

Objective: get samples of x.



Objective: get samples of x.

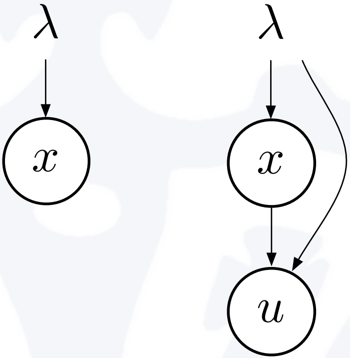
Sometimes it is easier to introduce an auxiliary variable u and to Gibbs sample the joint $P(x, u)$ (i.e. sample from $P(x|u; \lambda)$ then $P(u|x, \lambda)$, etc.) then discard the u values than it is to directly sample from $p(x|\lambda)$.



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.



Slice Sampling: A very useful auxiliary variable sampling trick

Pedagogical Example:

- $x|\lambda \sim \text{Poisson}(\lambda)$ (countable support)
- *enumeration* strategy for sampling x (impossible)[1]
- auxiliary variable u with $P(u|x,\lambda) = \frac{\mathbb{I}(0 \le u \le P(x|\lambda))}{P(x|\lambda)}$

Note: Marginal distribution of x is

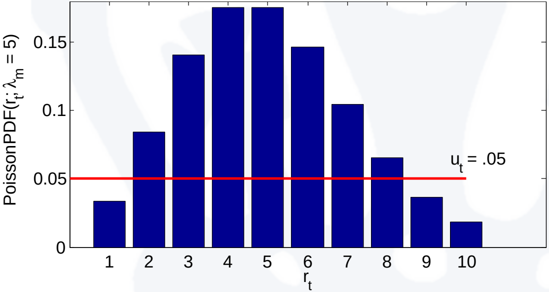
$$
\begin{aligned}
P(x|\lambda) &= \sum_u P(x,u|\lambda) \\
&= \sum_u P(x|\lambda)P(u|x,\lambda) \\
&= \sum_u P(x|\lambda)\frac{\mathbb{I}(0 \le u \le P(x|\lambda))}{P(x|\lambda)} \\
&= \sum_u \mathbb{I}(0 \le u \le P(x|\lambda)) = P(x|\lambda)
\end{aligned}
$$

[1] Necessary in EDHMM

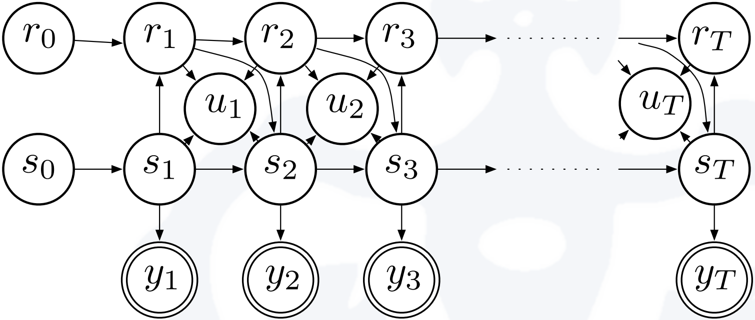
This suggests a Gibbs sampling scheme: alternately sampling from

- $P(x|u,\lambda) \propto \mathbb{I}(u \leq P(x|\lambda))$
  - *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 $P(x|\lambda)$.



EDHMM Graphical Model with Auxiliary Variables



With auxiliary variables defined as

$$p(u_t|z_t, z_{t-1}) = \frac{\mathbb{I}(0 < u_t < p(z_t|z_{t-1}))}{p(z_t|z_{t-1})}$$

and

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

one can run standard forward-backward conditioned on u's.

Forward-backward slice sampling only has to consider a finite number of successor states at each timestep.

Forward recursion

$$
\begin{aligned}
\hat{\alpha}_t(z_t) &= p(z_t, \mathcal{Y}_1^t, \mathcal{U}_1^t) \\
&= \sum_{z_{t-1}} p(z_t, z_{t-1}, \mathcal{Y}_1^t, \mathcal{U}_1^t) \\
&\propto \sum_{z_{t-1}} p(u_t | z_t, z_{t-1}) p(z_t, z_{t-1}, \mathcal{Y}_1^t, \mathcal{U}_1^{t-1}) \\
&= \sum_{z_{t-1}} p(u_t | z_t, z_{t-1}) p(y_t | z_t) p(z_t | z_{t-1}) p(z_{t-1}, \mathcal{Y}_1^{t-1}, \mathcal{U}_1^{t-1}) \\
&= \sum_{z_{t-1}} \mathbb{I}(0 < u_t < p(z_t | z_{t-1})) p(y_t | z_t) \hat{\alpha}_{t-1}(z_{t-1}).
\end{aligned}
$$

Only a finite (small) part of the forward trellis needs to be enumerated (in expectation).

Backward Sampling

Recursively sample a state sequence from the distribution $p(z_{t-1}|z_t, \mathcal{Y}, \mathcal{U})$ which can expressed in terms of the forward variable:

$$\begin{aligned} p(z_{t-1}|z_t, \mathcal{Y}, \mathcal{U}) &\propto p(z_t, z_{t-1}, \mathcal{Y}, \mathcal{U}) \\ &\propto p(u_t|z_t, z_{t-1})p(z_t|z_{t-1})\hat{\alpha}_{t-1}(z_{t-1}) \\ &\propto \mathbb{I}(0 < u_t < p(z_t|z_{t-1}))\hat{\alpha}_{t-1}(z_{t-1}). \end{aligned}$$

Algorithm 1 EDHMM Sampler

Initialise parameters A, λ, θ. Initialize u_t small $\forall$ T
for sweep $\in \{1, 2, 3, \ldots\}$ **do**
  Forward: run FR to get $\hat{\alpha}_t$ given $\mathcal{U}$ and $\mathcal{Y}$ $\forall$ T
  Backward: sample $z_T \sim \hat{\alpha}_T$
  for $t \in \{T, T-1, \ldots, 1\}$ **do**
    sample $z_{t-1} \sim \mathbb{I}(u_{t+1} < p(z_t|z_{t-1}))\hat{\alpha}_{t-1}$
  end for
  Slice:
  for $t \in \{1, 2, \ldots, T\}$ **do**
    evaluate $l = p(d_t|x_t, d_{t-1})p(x_t|x_{t-1}, d_{t-1})$
    sample $u_{t+1} \sim \mathrm{Uniform}(0, l)$
  end for
  sample parameters A, λ, θ
end for

Posterior Utility

For instance, posterior predictive inference

$$
\begin{aligned}
&P(y_{T+1}|\mathbf{y}) \\
&= \int\int\int \sum_{\mathbf{z}} P(y_{T+1}|A, \mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\lambda}, \boldsymbol{\theta}) P(A, \mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\lambda}, \boldsymbol{\theta}|\mathbf{y}) dA d\boldsymbol{\pi} d\boldsymbol{\lambda} d\boldsymbol{\theta} \\
&\approx \frac{1}{L} \sum_{\ell} P(y_{T+1}|A^{(\ell)}, \mathbf{z}^{(\ell)}, \boldsymbol{\pi}^{(\ell)}, \boldsymbol{\lambda}^{(\ell)}, \boldsymbol{\theta}^{(\ell)})
\end{aligned}
$$

where

$$\{A^{(\ell)}, \mathbf{z}^{(\ell)}, \boldsymbol{\pi}^{(\ell)}, \boldsymbol{\lambda}^{(\ell)}, \boldsymbol{\theta}^{(\ell)}\} \sim P(A, \mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\lambda}, \boldsymbol{\theta}|\mathbf{y})$$

Results

To illustrate EDHMM learning on synthetic data, five hundred datapoints were generated using a 4 state EDHMM with Poisson duration distributions

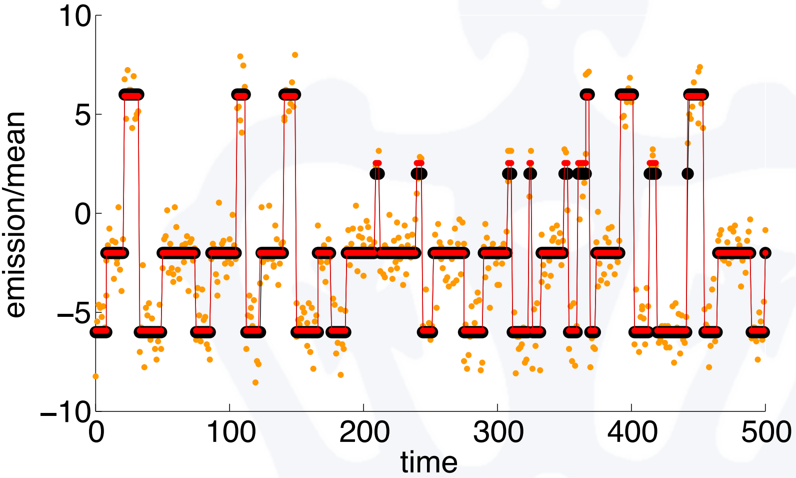
$$\boldsymbol{\lambda} = (10, 20, 3, 7)$$

and Gaussian emission distributions with means

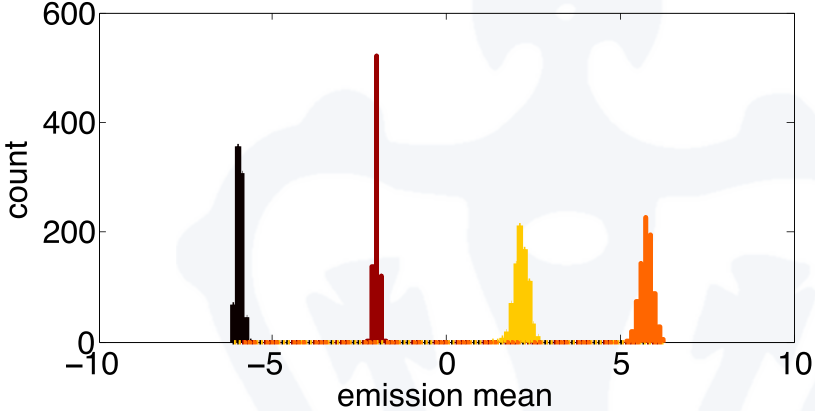
$$\boldsymbol{\mu} = (-6, -2, 2, 6)$$

all unit variance.

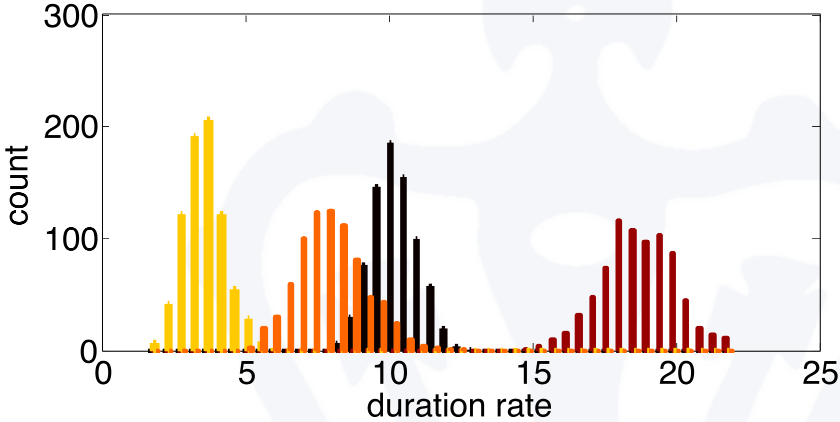
EDHMM: Synthetic Data Results



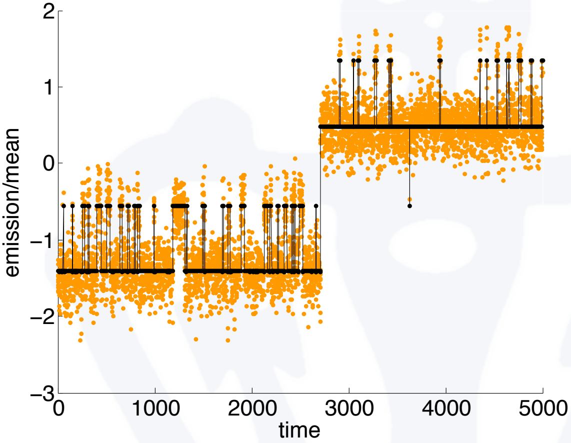
EDHMM: Synthetic Data, State Mean Posterior



EDHMM: Synthetic Data, State Duration Parameter Posterior



EDHMM: Nanoscale Transistor Spontaneous Voltage Fluctuation



[Realov and Shepard, 2010]

Task: understand system with states that have identical observation distributions and differ only in duration distribution.

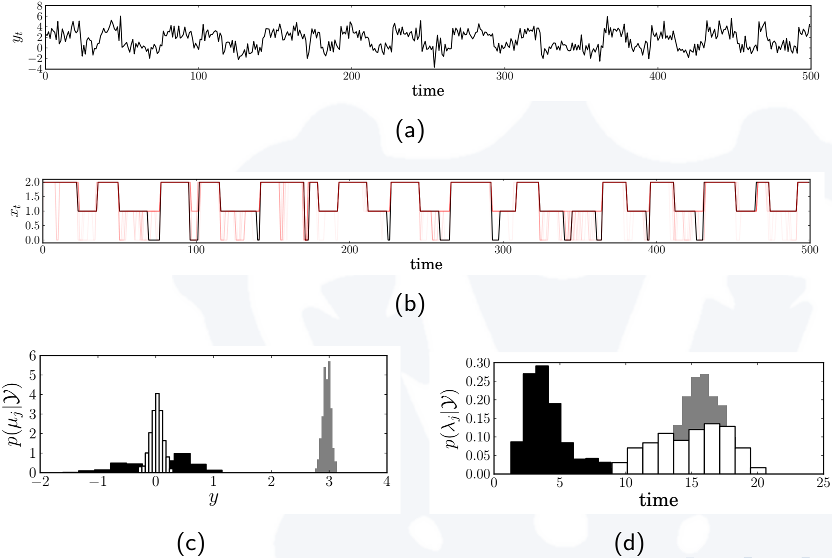
Observation distributions have means $\mu_1 = 0$, $\mu_2 = 0$, and $\mu_3 = 3$ and the duration distributions have rates $\lambda_1 = 5$, $\lambda_2 = 15$, $\lambda_3 = 20$.

(Next slide)

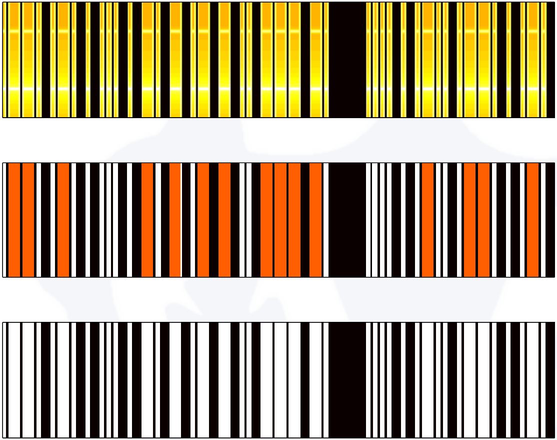
a) observations; b) true states overlaid with 20 randomly selected state traces produced by the sampler.

Samples from the posterior observation distribution mean are shown in c), and samples from the posterior duration distribution rates are shown in d).

EDHMM: States Not Distinguishable By Output



EDHMM vs. IHMM: Modeling the Morse Code Cepstrum



Wrap-Up

- Overview of HMM developments since Rabiner tutorial
- Tools to understand state-of-the art HMMs
- Useful tricks for Bayesian inference in general (sampling)

Extras

- Code: `https://github.com/mikedewar/EDHMM`
- Me: `http://www.stat.columbia.edu/~fwood`
- w: `http://www.stat.columbia.edu/~fwood/w4240/`

Questions?

Thank you!

More Technical Wrap-Up - Small Contribution

- Novel inference procedure for EDHMMs that doesn't require truncation and is more efficient than considering all possible durations.

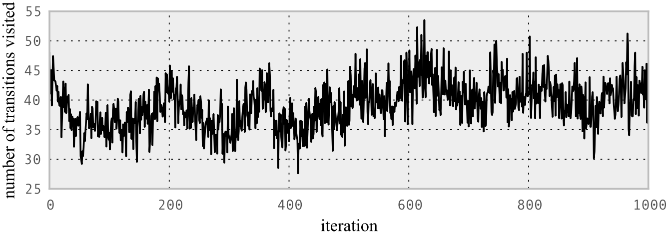


Figure: Mean number of transitions considered per time point by the beam sampler for 1000 post-burn-in iterations. Compare to $(KT)^2 = O(10^6)$ transitions that would need to be considered by standard forward backward without truncation, the only surely safe, truncation-free alternative.

Future Work

- Novel Gamma process construction for dependent, structured, infinite dimensional HMM transition distributions.
- Generalize to spatial prior on HMM states ("location")
  - Simultaneous location and mapping
  - Process diagram modeling for systems biology
- Applications; seeking "users"

M J Beal, Z Ghahramani, and C E Rasmussen. The Infinite Hidden Markov Model. In *Advances in Neural Information Processing Systems*, pages 29–245, March 2002.

C M Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

M Dewar, C Wiggins, and F Wood. Inference in hidden Markov models with explicit state duration distributions. *In Submission*, 2011.

E B Fox, E B Sudderth, M I Jordan, and A S Willsky. A Sticky HDP-HMM with Application to Speaker Diarization. *Annals of Applied Statistics*, 5(2A):1020–1056, 2011.

F. Jelinek. *Statistical Methods for Speech Recognition*. MIT Press, 1997.

B H Juang and L R Rabiner. Mixture Autoregressive Hidden Markov Models for Speech Signals. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(6):1404–1413, 1985.

Bibliography II

A. Krogh, M. Brown, I. S. Mian, K. Sjolander, and D. Haussler. Hidden Markov models in computational biology: Applications to protein modelling . *Journal of Molecular Biology*, 235:1501–1531, 1994.

C. Manning and H. Schütze. *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA, 1999.

C Mitchell, M Harper, and L Jamieson. On the complexity of explicit duration HMM's. *IEEE Transactions on Speech and Audio Processing*, 3 (3):213–217, 1995.

K P Murphy. Hidden semi-markov models (HSMMs). Technical report, MIT, 2002.

R. Nag, K. Wong, and F. Fallside. Script regonition using hidden Markov models. In *ICASSP86*, pages 2071–2074, 1986.

L R Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pages 257–286, 1989.

Bibliography III

Simeon Realov and Kenneth L Shepard. Random Telegraph Noise in 45-nm CMOS : Analysis Using an On- Chip Test and Measurement System. *Analysis*, (212):624–627, 2010. ISSN 01631918. URL `http://dx.doi.org/10.1109/IEDM.2010.5703436`.

T. Rydén, T. Teräsvirta, and S. **r**Asbrink. Stylized facts of daily return series and the hidden markov model. *Journal of Applied Econometrics*, 13(3):217–244, 1998.

D.O. Tanguay Jr. *Hidden Markov models for gesture recognition*. PhD thesis, Massachusetts Institute of Technology, 1995.

Y W Teh, M I Jordan, M J Beal, and D M Blei. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

Bibliography IV

J Van Gael, Y Saatci, Y W Teh, and Z Ghahramani. Beam sampling for the infinite hidden Markov model. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1088–1095. ACM, 2008.

A.D. Wilson and A.F. Bobick. Parametric hidden Markov models for gesture recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 21(9):884–900, 1999.

S Yu and H Kobayashi. An Efficient Forward–Backward Algorithm for an Explicit-Duration Hidden Markov Model. *Signal Processing letters*, 10 (1):11–14, 2003.

S Z Yu. Hidden semi-Markov models. *Artificial Intelligence*, 174(2): 215–243, 2010.