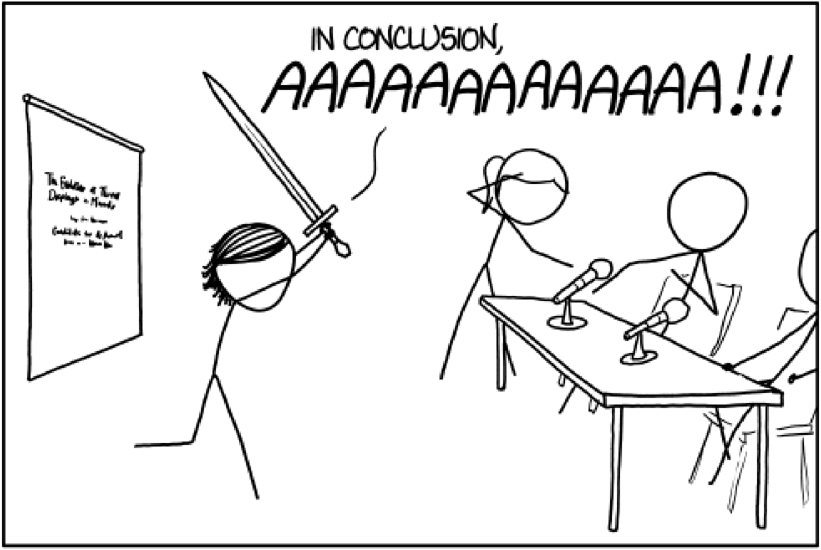
HaMMLeT: An Infinite-State Hidden Markov Model With Local Transitions

Colin Reimer Dawson

8 August 2016



THE BEST THESIS DEFENSE IS A GOOD THESIS OFFENSE.

Outline

Some Motivating Applications

Hidden Markov Models (HMMs)
- Mixture Models

"Nonparametric" Bayesian Models
- Infinite Mixture Models
- An Infinite HMM?
- Hierarchical Dirichlet Process Mixtures

HDP HMM with Local Transitions
- Adding the Concept of "Local" Transitions
- Posterior Inference

Experiments
- Data With Block Diagonal Transitions

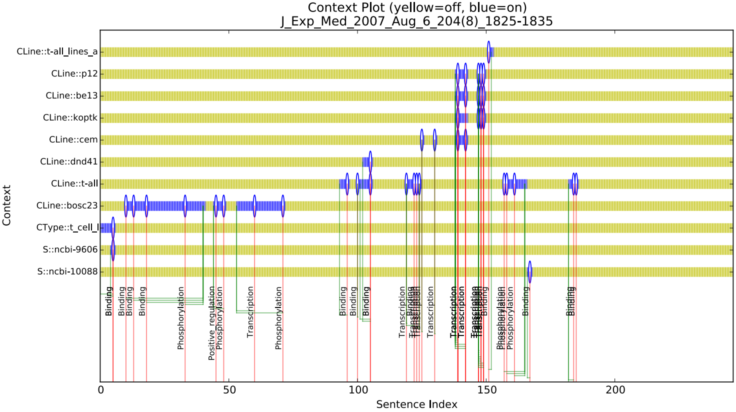
Future Work

Biological Context in Text

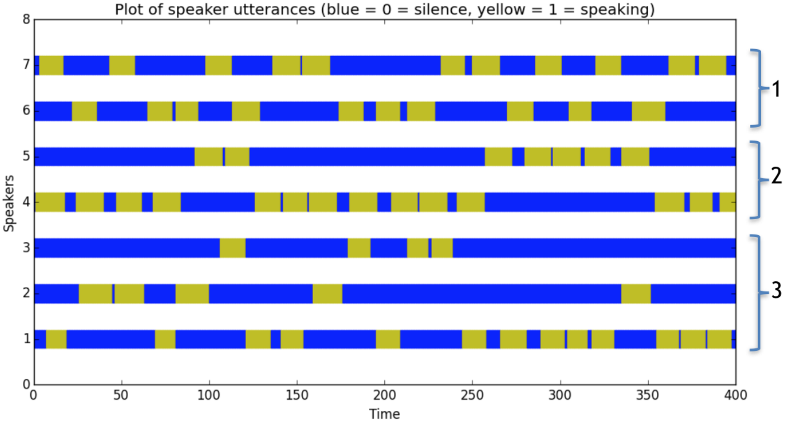
Mutations in oncogenes are much more likely to lead to cancer in some tissue types than others, because some tissues express other proteins that counteract the oncogene. For example, in MICE the G12D activating mutation in K-ras causes lung tumors but not muscle-derived sarcomas, because muscle cells express two proteins (Arf and Ink4a) that cause cell division to halt when Ras is overactive.

(Young and Jacks, PNAS, 2010)

Biological Context in Text



Blind Source Separation: The "Cocktail Party" Problem

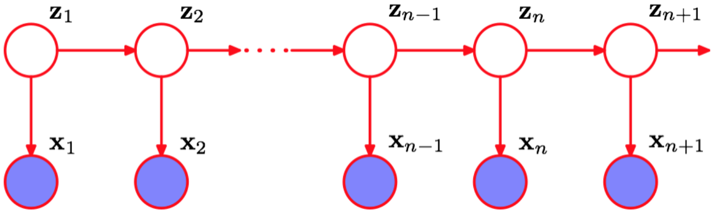


Cocktail Party Data

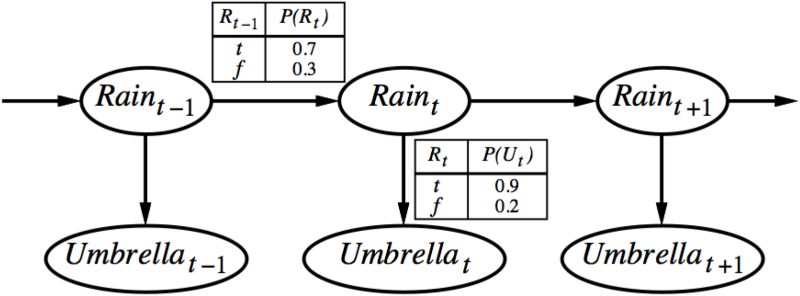
A “Cocktail Party”

Hidden Markov Model

Data depends at each t on an *unobserved* state, z_t, which evolve as a Markov chain.



HMM Example



R_{t-1}	$P(R_t)$
t	0.7
f	0.3

R_t	$P(U_t)$
t	0.9
f	0.2

The HMM is defined by:

Initial distribution	π_0
Transition matrix	π
Emission distributions	$\{F_k\}$

Applications of HMMs

Speech recognition

- Y_t: acoustic signal
- Z_t: word segments

Machine translation

- Y_t: words in language A
- Z_t: words in language B

Robot tracking

- Y_t: sensor data
- Z_t: real position in space

Some Inference Objectives for HMMs

1. Given data, find good parameters (estimation)
2. Given parameters and data, find the distribution over states at a particular t (marginalization).
3. Given parameters and data, find the most likely joint state sequence (maximization).

HMMs as Mixture Models

HMMs can be viewed as a dynamic version of a **mixture model**

- Each data point, $y \in \mathcal{Y}$, arises from one of several *classes*, each with its own distribution.
- Goal: Estimate unknown density, f, of the form

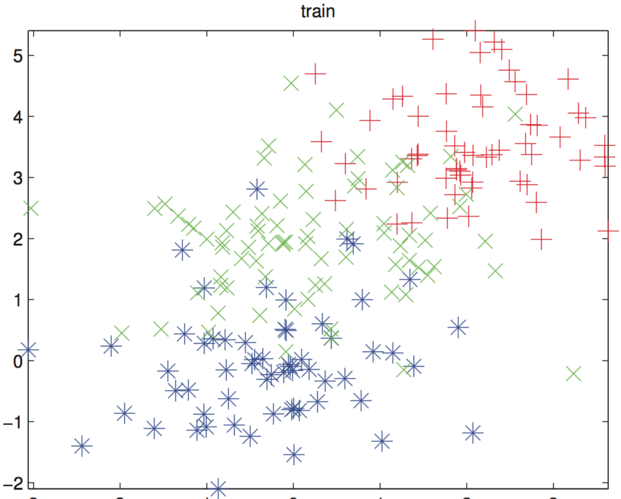
$$f(y) = \sum_{j} \pi_j f_k(y)$$

- Traditionally, number of components is fixed and f_k have parametric form:

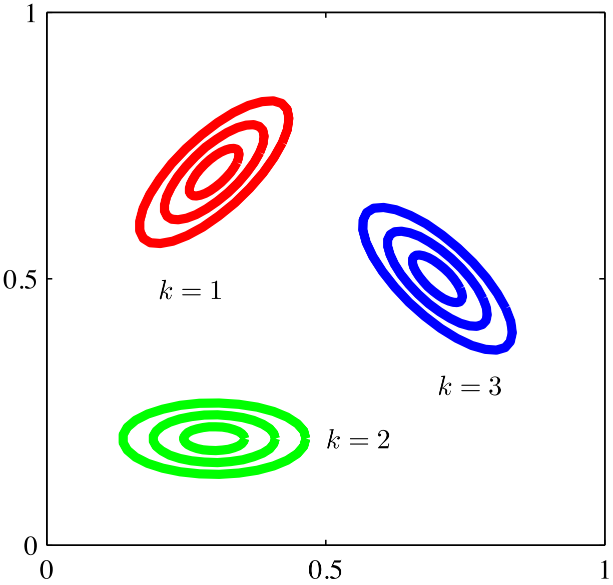
$$f(y; \pi, \theta) = \sum_{k=1}^{K} \pi_k f(y; \theta_k) \tag{1}$$

- Goal reduces to estimating π and θ.

Data from a Mixture Model



A Mixture Density

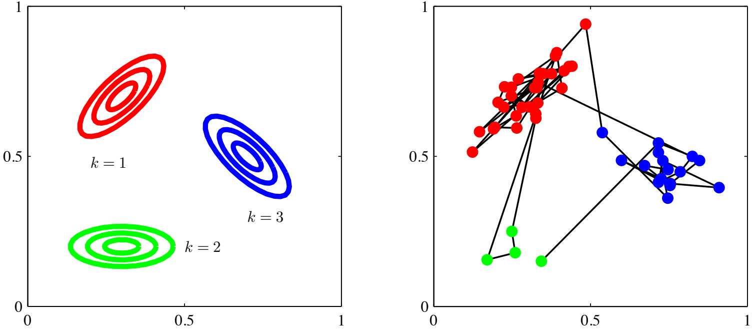


A Bayesian Approach to Mixture Models

$$\begin{aligned} f(\theta, \pi \mid x) &= f(\theta, \pi) \frac{f(x \mid \theta, \pi)}{f(x)} \\ &\propto f(\theta, \pi) f(x \mid \theta, \pi) \end{aligned}$$

Standard Bayesian version uses *conjugate priors*: the family is closed under Bayesian updating (for a particular likelihood family). **Makes inference simple!**

A Dynamic Mixture



A Bayesian HMM

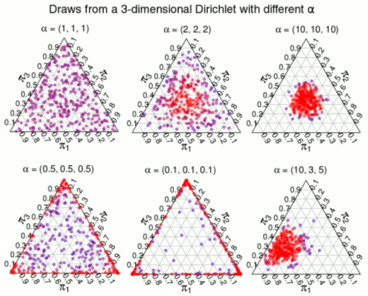
An HMM allows the mixing weights to change depending on the previous state.

Standard: separate conjugate (Dirichlet) priors for each transition

row π_k

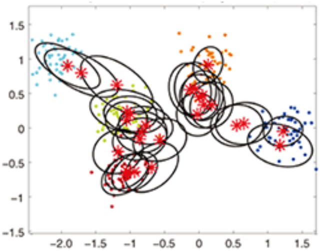
$$\pi_k \overset{i.i.d}{\sim} \text{Dirichlet}(\alpha \mathbf{1}_K)$$

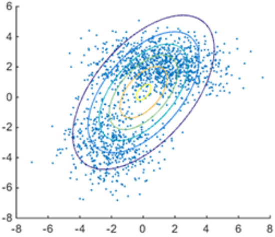
$$\theta_k \overset{i.i.d.}{\sim} f\text{-Conjugate}(\xi)$$



Unbounded number of components

- Having to specify K in advance is limiting. Too high $\rightarrow$ overfitting. Too low $\rightarrow$ underfitting.





- We can instead use an *infinite mixture model*, with a prior to guard against overfitting.

Dirichlet Processes

Definition: Dirichlet Process (Ferguson, 1973)

A **Dirichlet Process** with **base probability measure** G_0 and **concentration parameter** $\alpha > 0$ is a random measure, μ on a measure space $(\mathcal{X}, \Sigma)$ with the property that, for any finite partition, $\{A_1, \dots, A_n\}$ of $\mathcal{X}$,

$$(\mu(A_1), \dots, \mu(A_n)) \sim \mathrm{Dirichlet}(\alpha G_0(A_1), \dots, \alpha G_0(A_n)) \tag{2}$$

- Note 1: μ is atomic $a.s.$.
- Note 2: If G_0 has finite support, μ is just a Dirichlet distribution over those atoms.

A Constructive Definition of the DP

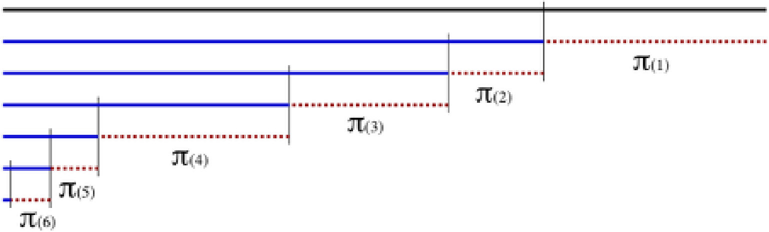
The Stick-Breaking Construction (Sethuraman, 1991) Define

$$\{\tilde{\pi}_k\}_{k=1}^{\infty} \overset{i.i.d}{\sim} \mathcal{B}\text{eta}(1,\alpha)$$

$$\pi_k = \tilde{\pi}_k \prod_{k=1}^{k-1}(1-\tilde{\pi_k})$$

$$\{\theta_k\}_{k=1}^{\infty} \overset{i.i.d.}{\sim} G_0$$

Then $\mu := \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$ is distributed DP(αG_0)



An Infinite Mixture Model

DP Mixture Model

If we let

$$\pi \sim \mathrm{Stick}(\alpha) \tag{3}$$

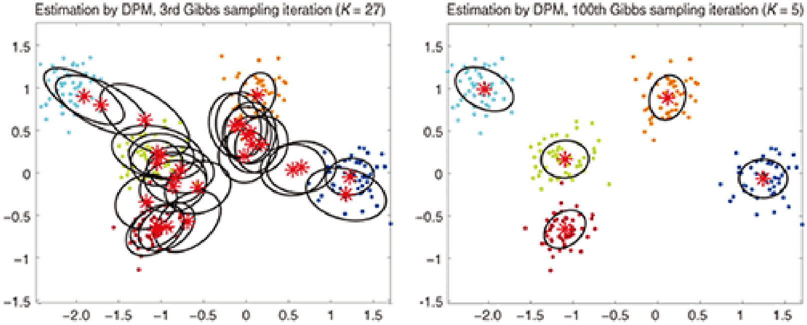
$$\{\theta_k\} \overset{i.i.d}{\sim} G_0 \tag{4}$$

$$f(x \,|\, \pi, \theta) = \sum_{k=1}^{\infty} \pi_k f(x \,|\, \theta_k) \tag{5}$$

then x are distributed according to a **Dirichlet Process Mixture Model**

Infinite Gaussian Mixture Model

For example, if $f(x \mid \theta_k)$ is a Normal density, we have the **Infinite Gaussian Mixture model** (Rasmussen, 2000)



Estimation by DPM, 3rd Gibbs sampling iteration ($K = 27$)

Estimation by DPM, 100th Gibbs sampling iteration ($K = 5$)

(a)

(b)

Infinite State HMM

We want to allow HMM to have (countably) infinite state space.

Could put separate DP prior on each row of the transition matrix.
Why might this not be ideal?

We would never visit the same state twice!

Hierarchical Dirichlet Processes

- Data from multiple sources, $j = 1, \ldots, J$ (such as temporal contexts), whose generating distributions, $\{G_j\}$ are distinct but related.
- Can use a hierarchical prior to couple them, e.g.,

$$G_j \overset{i.i.d}{\sim} \mathsf{DP}(\alpha G_0), \tag{6}$$

- But, if G_0 is absolutely continuous, sets of atoms in the G_j will be disjoint.
- Solution: Let G_0 itself have a DP prior.

Hierarchical Dirichlet Processes

The Hierarchical Dirichlet Process Mixture Model (Teh, 2006)

Define

$$G_0 \sim \mathrm{DP}(\gamma H) \tag{7}$$

$$G_j \,|\, G_0 \overset{i.i.d}{\sim} \mathrm{DP}(\alpha G_0) \qquad j = 1, \ldots, J \tag{8}$$

$$y_{jn} \,|\, G_j \overset{i.i.d}{\sim} \sum_{k=1}^{\infty} \pi_{jk} f(\mathbf{y} \,|\, \theta_{jk}) \tag{9}$$

This defines a Hierarchical Dirichlet Process (HDP) Mixture. Atoms are shared among contexts by virtue of the discreteness of G_0.

- γ: how close to "uniform" is the overall distribution of components?
- α: how similar are the mixture weights across contexts?

Examples of HDP Mixture Models

- f Normal: hierarchical Infinite Gaussian Mixture.
- f Multinomial: **hierarchical infinite topic model**
  - y words
  - θ_{jk} word distributions corresponding to "topics"

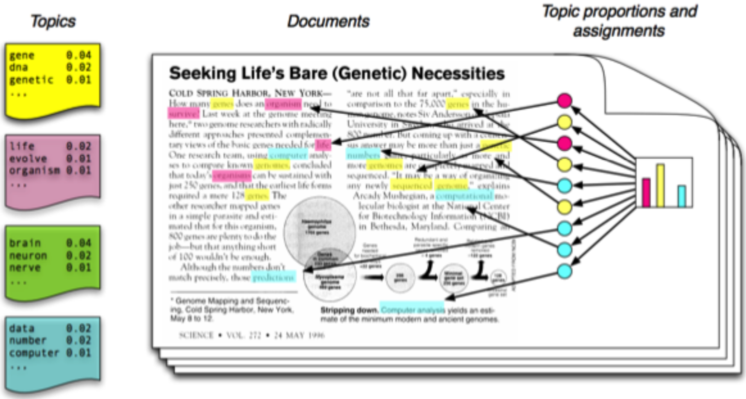

Topics
gene 0.04
dna 0.02
genetic 0.01
...
life 0.02
evolve 0.01
organism 0.01
...
brain 0.04
neuron 0.02
nerve 0.01
...
data 0.02
number 0.02
computer 0.01
...
Documents
Topic proportions and assignments
Seeking Life's Bare (Genetic) Necessities
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.
SCIENCE • VOL. 272 • 24 MAY 1996

Figure source: Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77-84.

An Infinite (HDP) HMM

We can allow infinitely many states by putting a Dirichlet Process prior on each row distribution, π_j.

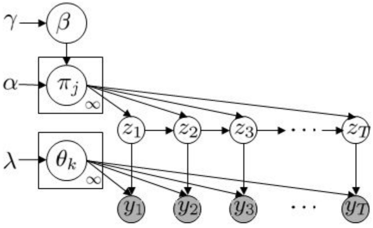
$$(\pi_j, \theta_j) \overset{i.i.d}{\sim} \mathrm{DP}(\alpha G_0) \tag{10}$$

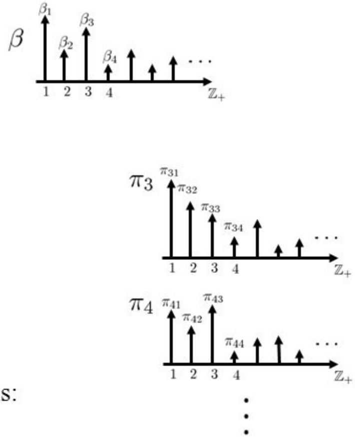
θ_j: emission parameters for states reachable from state j.

To ensure that the θ_j contain overlapping values for different j, G_0 must be discrete! Solution: A hierarchical prior: $G_0 \sim \mathrm{DP}(\gamma H)$.

This is the Infinite or HDP HMM (Beal, 2001; Teh, 2006).

The HDP-HMM





- Average transition distribution:

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

- State-specific transition distributions:

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

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

Limitations of HDP-HMM

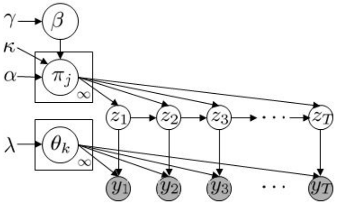
Two properties of HDP-HMM not shared with non-temporal HDP:

1. Contexts (except the first) are random
2. Set of contexts identified with set of states

Limitations:

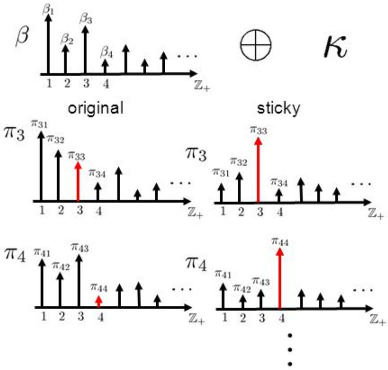
- Self-transitions are not special
- No notion of a state geometry

The "Sticky" HDP-HMM (Fox, 2008)



$$\begin{aligned} \beta &\sim \mathrm{GEM}(\gamma) \\ \pi_j &\sim \mathrm{DP}\left(\alpha\beta + \kappa\delta_j\right) \end{aligned}$$

state-specific base measure



Increased probability of self-transition →
$$\boxed{E[\pi_{jk}] = \frac{\alpha\beta_k + \kappa\delta(j,k)}{\alpha + \kappa}}$$

See also the HDP-HSMM (Johnson, 2013): rules out self-transitions, and models durations separately

Local Transitions: The HDP-HMM-LT

- Generalize preference for self-transitions to "local" transitions: HDP-HMM-LT.
- Latent states located in a space with a symmetric similarity kernel, ϕ:

$$0 \leq \phi(\ell_j, \ell_{j'}) = \phi(\ell_{j'}, \ell_j) \leq \phi(\ell_j, \ell_j) \equiv 1 \tag{11}$$

- $P(j \to j')$ from HDP prior is rescaled by $\phi_{jj'}$.

The HDP-HMM-LT

Definition: HDP-HMM-LT

Assume we have a sequence of location pairs $\{(\theta_j, \ell_j)\}_{j=1}^{\infty}$ from some distribution, and a similarity kernel ϕ. We define

$$\beta \sim \mathsf{Stick}(\gamma) \tag{12}$$

$$\tilde{\pi}_j \sim \mathsf{DP}(\alpha\beta) \tag{13}$$

$$a_{jj'} = \tilde{\pi}_{jj'}\phi_{jj'} \tag{14}$$

$$z_t \mid z_{t-1} \sim \sum_{j'} \frac{a_{z_{t-1}j'}}{\sum_{j''} a_{z_{t-1}j''}} \delta_{j'} \tag{15}$$

$$y_t \mid z_t \sim F(\theta_{z_t}) \tag{16}$$

The normalization term is finite and positive since it is bounded above by $\sum_{j'} \tilde{\pi}_{jj'} = 1$, and below by $\tilde{\pi}_{jj'}$.

Normalized Gamma Process Representation

The DP can also be obtained by normalizing a *Gamma Process*:

A **Gamma Process** is a Poisson Process on $\mathbb{R}^+ \times \Theta$ with Lévy intensity measure

$$\nu(d\pi, d\theta) = \alpha\pi^{-1}e^{-\pi}d\pi G_0(d\theta) \tag{17}$$

Consider a random collection of point masses $\{\theta_j\}$ on Θ, with respective random masses $\{\pi_j\}$ as points $(\pi_j, \theta_j) \in \mathbb{R}^+ \times \Theta$. The number $n(A)$ of such points in a region $A \subset \mathbb{R}^+ \times \Theta$ is distributed

$$n(A) \sim \mathcal{P}\mathrm{ois}\left(\int_A \nu(d\pi, d\theta)\right) \tag{18}$$

The normalized set of atoms form a probability measure on Θ. This normalized measure is distributed $\mathrm{DP}(\alpha G_0)$.

A Gamma Process representation

By the Gamma Process representation of the DP, we obtain the same model by drawing

$$\begin{aligned} \beta &\sim \mathsf{Stick}(\gamma) \\ \{\pi_{jj'}\}_j &\overset{i.i.d.}{\sim} \mathsf{Gamma}(\alpha\beta_{j'}, 1) \qquad j' \geq 1 \end{aligned}$$

and setting

$$\tilde{\pi}_{jj'} = \frac{\pi_{jj'}}{\sum_{j''} \pi_{jj''}}$$

A Gamma Process representation

Combining the two normalizations yields

HDP-HMM-LT (Gamma Process Representation)

$$
\begin{aligned}
\beta &\sim \mathsf{Stick}(\gamma) \\
\pi_{jj'} &\sim \mathsf{Gamma}(\alpha\beta_{j'}, 1) \\
a_{jj'} &= \pi_{jj'}\phi_{jj'} \\
z_t \mid z_{t-1} &\sim \sum_{j'} \frac{a_{z_{t-1}j'}}{\sum_{j''} a_{z_{t-1}j''}} \delta_{j'} \\
y_t \mid z_t &\sim F(\theta_{z_t})
\end{aligned}
$$

Loss of Conjugacy for $\boldsymbol{\pi}$

However the likelihood for π_j contains a random normalization term, which renders all π_j conditionally dependent

$$p(\mathbf{z} \,|\, \pi, \phi) = \prod_j \left(\sum_{j''} \pi_{jj''} \phi_{jj''} \right)^{-n_{j\cdot}} \prod_{j'} (\pi_{jj'} \phi_{jj'})^{n_{jj'}} \qquad (19)$$

$n_{jj'}$: # of transitions from j to j'
$n_{j\cdot}$: total visits to j

We can simplify with an "augmented data" representation.

The Markov Process With Failed Jumps Representation

Lemma Let $(z_t, u'_t)_{t=1}^T$ be a pure jump Markov process with rate matrix $(a_{jj'})$, z_t the state after the tth jump, u'_t the time between jump $t-1$ and t.

Given $z_{t-1} = j$,

(a) $u'_t \sim \text{Exponential}(\sum_{j'} a_{jj'})$

(b) $P(z_t = j') \propto a_{jj'}$

(c) u'_t and z_t are independent

Hence, in $n_{k\cdot}$ visits to state k, the process spends

$$u_j \sim \text{Gamma}(n_{j\cdot}, \sum_{j'} a_{jj'}) \tag{20}$$

time there.

Introducing failed jumps

Lemma Suppose also that while in state j, unsuccessful attempts to jump to state j' are made at rate $\pi_{jj'} - a_{jj'} = \pi_{jj'}(1 - \phi_{jj'})$. The total number of these is

$$q_{jj'} \sim \text{Poisson}(\pi_{jj'}(1 - \phi_{jj'})u_j) \tag{21}$$

Augmented Likelihood for π

Augmenting the data with u and Q, the likelihood for π becomes

$$\mathcal{L}(\pi \,|\, z, u, Q; \phi) \propto \prod_{j} \prod_{j'} \pi_{jj'}^{n_{jj'} + q_{jj'}} e^{-u_j \pi_{jj'}}$$

which is conjugate to the Gamma prior.

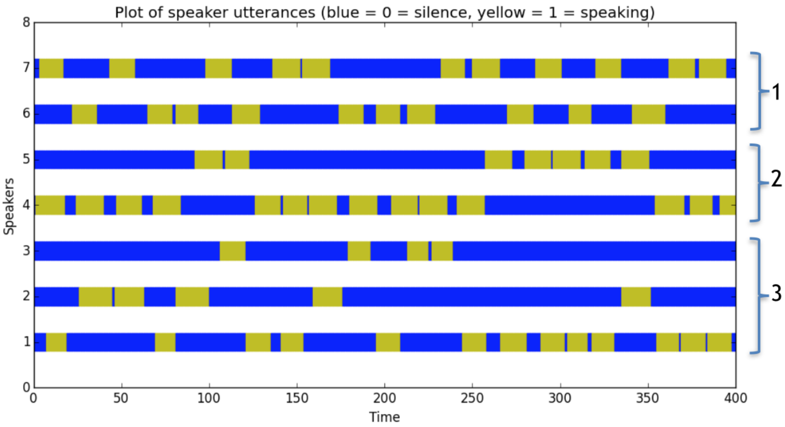
Gibbs Sampling

- Conditioned on A and the observations, sample the state sequence z jointly with a “message passing” algorithm.
- Sample augmented data from Exponential and Poisson distributions.
- Sample $\boldsymbol{\pi}$ and hyperparameters ($\boldsymbol{\beta}$, α, and γ) using the factorization

  $$p(\gamma, \alpha, \boldsymbol{\beta}, \boldsymbol{\pi} \,|\, \mathcal{D}) = p(\gamma \,|\, \mathcal{D}) p(\alpha \,|\, \mathcal{D}) p(\boldsymbol{\beta} \,|\, \gamma, \mathcal{D}) p(\boldsymbol{\pi} \,|\, \boldsymbol{\beta}, \alpha, \mathcal{D}) \tag{22}$$

  where $\mathcal{D}$ is the augmented “data”.
- (We require some additional data augmentation to get the marginal distributions for γ, α and β.)

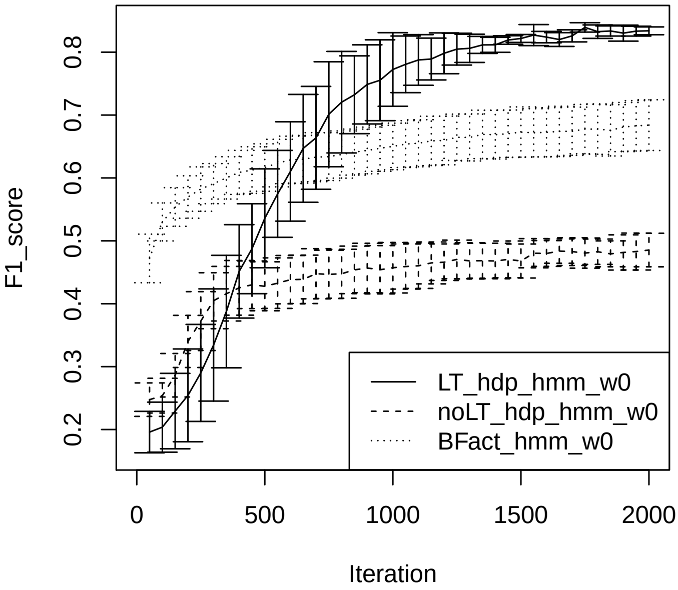
Cocktail Party Data



Similarity and Emission Model

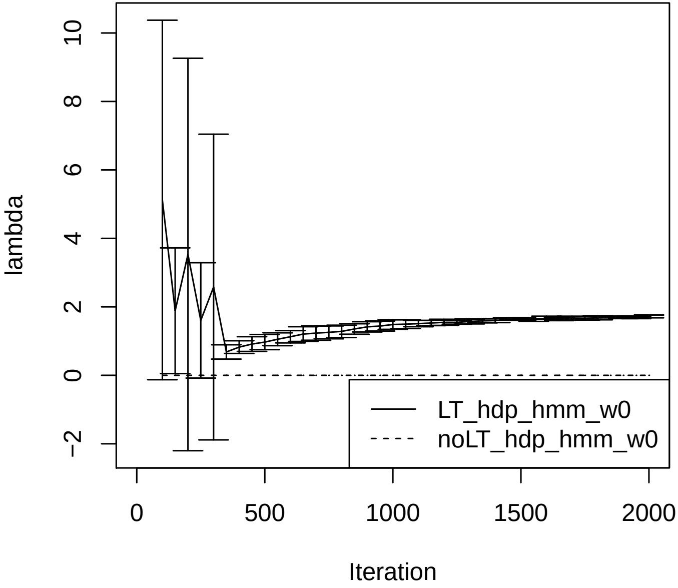
- Latent states: binary vectors, $\{\theta_j\}$
- F: Normal linear model
- $\phi_{jj'} = \exp(-\lambda \left\|\theta_j - \theta_{j'}\right\|_{L_1})$

Recovery of Binary States (F_1 measure)



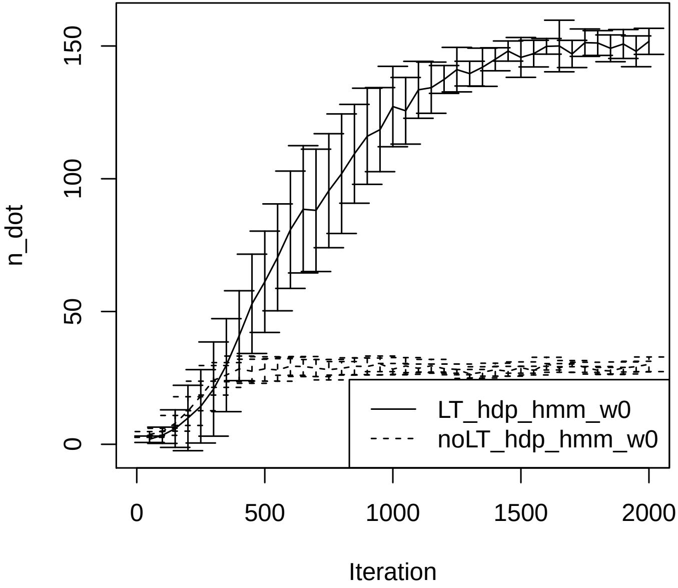
Error

Similarity Parameter



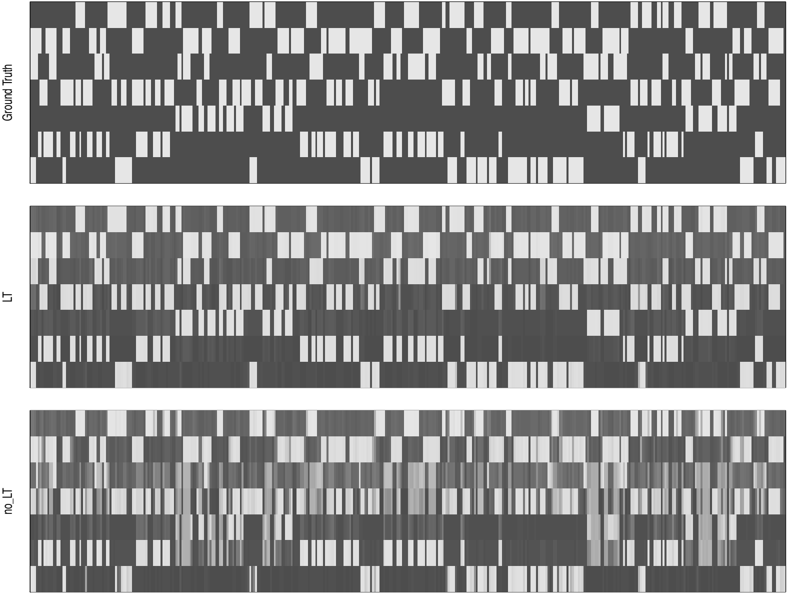
Error

Number of States Found

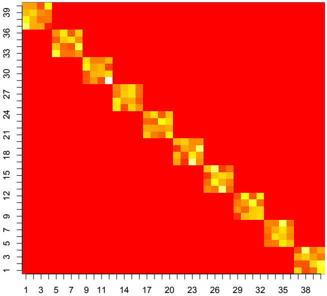


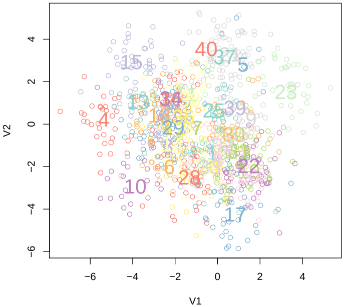
Error

Inferred Speaker Sequence



(Near) Block Diagonal Transition Matrix





Many more transitions “within groups”.

Unlike previous experiment, similarity via ϕ and similarity in emissions are decoupled.

Block Diagonal Experiment

- Data: 95% within group transitions
- Emission model is Normal: $\theta = (\mu, \Sigma)$
- Similarity is Gaussian kernel:

$$\phi(\eta_j, \eta_{j'}) = \exp(-\lambda \left\| \eta_j - \eta_{j'} \right\|_{L_2}^2)$$

- η locations in an abstract latent space; "likelihood" in terms of Bernoulli process of successful and failed transitions

$$p(z, Q \,|\, \eta) = \prod_{j} \prod_{j'} \phi_{jj'}^{n_{jj'}} (1 - \phi_{jj'})^{q_{jj'}}$$

- Infer η separately from θ, using a Hamiltonian Monte Carlo (HMC) step (Duane and Pendleton, 1987)

Results: LT vs no LT

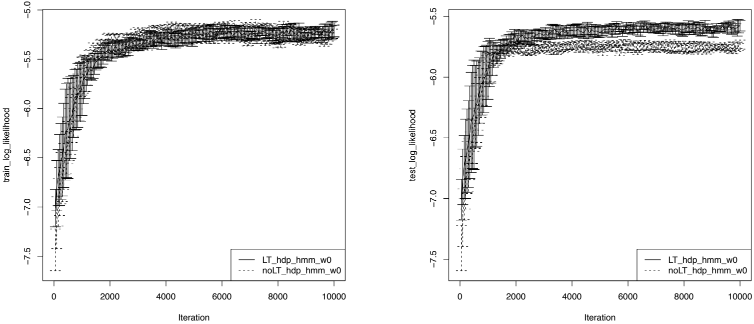


Figure: Left: Log likelihood on the training set by Gibbs iteration (marginalizing out state sequence) for LT and no LT (HDP-HMM) models. Right: Log likelihood on a held out test set.

Inferred Transition Matrices

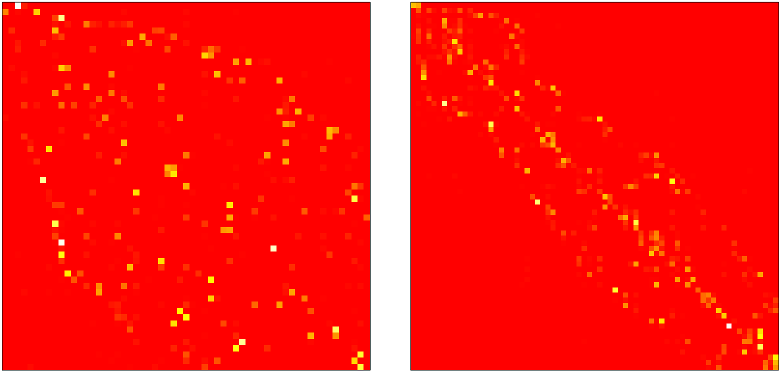
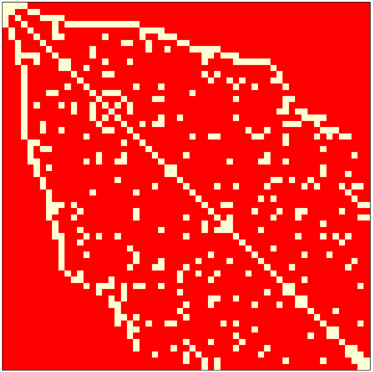


Figure: Left: Inferred transition matrix using the noLT (HDP-HMM) model. Right: Inferred transition matrix using the LT model. State permutation found using Reverse Cuthill-McKee algorithm.

Inferred Transition Matrices



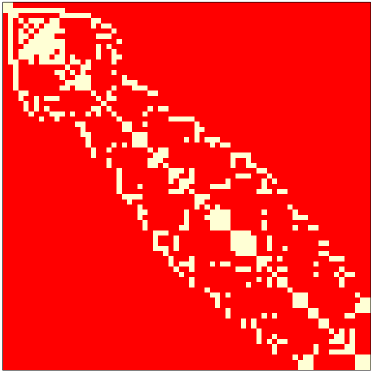


Figure: Left: Inferred transition matrix using the noLT (HDP-HMM) model. Right: Inferred transition matrix using the LT model. State permutation found using Reverse Cuthill-McKee algorithm.

Ongoing: Discovering Chord Classes in Music

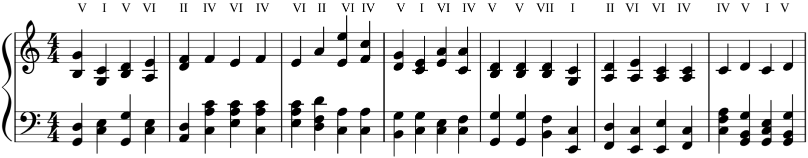


Figure: Transcription of a four-voice chorale, annotated with chord classes

- Encode each unique chord as an integer
- Infer latent states w/ LT vs no-LT, using a categorical emission model

Results: Emission Distributions

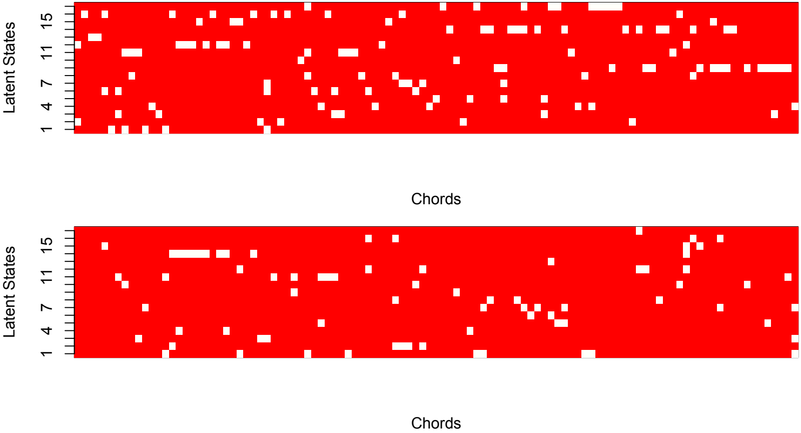
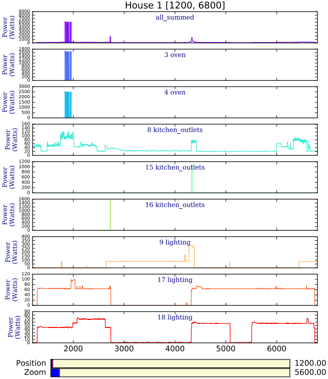


Figure: Emission distributions for states used (normalized by overall chord frequency, and thresholded). Top: no LT, Bottom: LT

Summary of Contributions

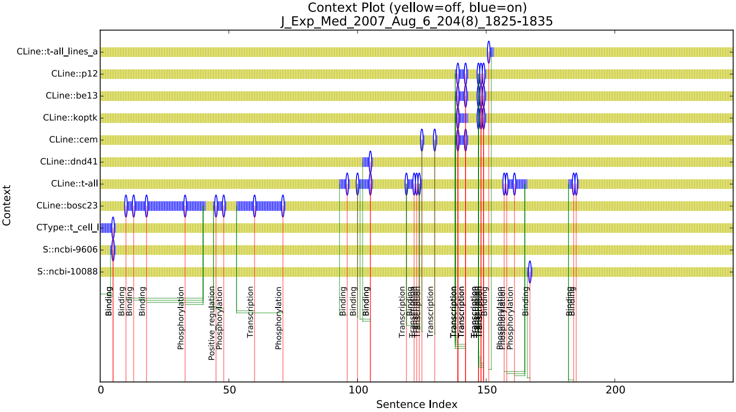
- A new model, the HDP-HMM-LT, in which there is a notion of geometry of the transition state space, such that transitions are more likely between nearby states.
- Formulation of HDP-HMM-LT as the marginalization of the "Markov Process With Failed Jumps".
- A straightforward Gibbs sampler after data augmentation
- Better generalization than existing models on data w/ nearby state changes

Ongoing: Power Disaggregation



- Data: Time series of total power consumptioo
- Latent state: How much power is each channel using at each time
- Different appliances have different discrete set of “modes”

Future: Discovering Biological Context



Theoretical Challenges

- Beam sampling
  - Now: need to bound the number of states considered
  - Beam sampling: adaptively considers new states (slice sampling)
  - But, auxiliary representation of u and Q requires explicit representation
- Adaptive HMC
  - In HMC need to hand-tune step size. Incorporate "adaptive" methods based on Riemannian manifolds.