Spectral Estimation of Hidden Markov Models

Jordan Rodu
Department of Statistics
University of Pennsylvania

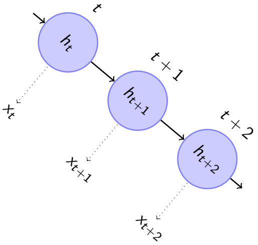
with: Michael Collins[1], Paramveer Dhillon[2], Dean Foster[3], Lyle Ungar[2], and Weichen Wu[2]

[1]Department of Computer Science
Columbia University

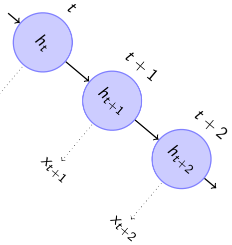
[2]Computer and Information Science
University of Pennsylvania

[3]Department of Statistics
University of Pennsylvania

Spectral Estimation of HMMs

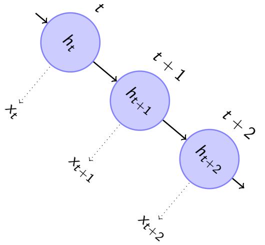


Spectral Estimation of HMMs



$$P(x_1,\ldots,x_t) = \sum_{h_1,\ldots,h_t} [\pi]_{h_1} \prod_{j=2}^{t} [T]_{h_j,h_{j-1}} \prod_{j=1}^{t} [O]_{x_j,h_j}$$

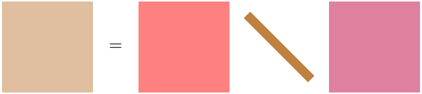
Spectral Estimation of HMMs

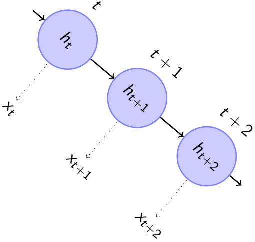


$$P(x_1, \dots, x_t) = \sum_{h_1,\dots,h_t} [\pi]_{h_1} \prod_{j=2}^{t} [T]_{h_j,h_{j-1}} \prod_{j=1}^{t} [O]_{x_j,h_j}$$

$$P(x_1, \dots, x_t) = 1^\top A(x_t) \cdots A(x_1) \pi$$

Spectral Estimation of HMMs

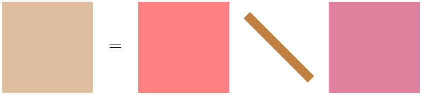


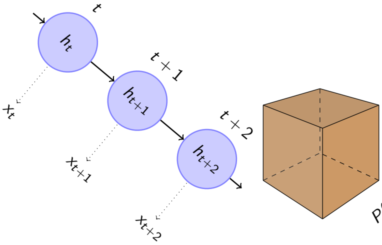


$$P(x_1, \ldots, x_t) = \sum_{h_1,\ldots,h_t} [\pi]_{h_1} \prod_{j=2}^{t} [T]_{h_j,h_{j-1}} \prod_{j=1}^{t} [O]_{x_j,h_j}$$

$$P(x_1, \ldots, x_t) = 1^\top A(x_t) \cdots A(x_1) \pi$$

Spectral Estimation of HMMs





$$P(x_1,\dots,x_t) = \sum_{h_1,\dots,h_t} [\pi]_{h_1} \prod_{j=2}^{t} [T]_{h_j,h_{j-1}} \prod_{j=1}^{t} [O]_{x_j,h_j}$$

$$P(x_1,\dots,x_t) = 1^\top A(x_t)\cdots A(x_1)\pi$$

The basic Hidden Markov Model, in pictures

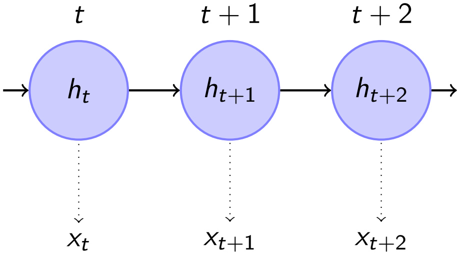


Figure: HMM with states h_t, h_{t+1}, and h_{t+2} which emit observations x_t, x_{t+1}, and x_{t+2} respectively.

Assumptions for Hidden Markov Model

1. Assumption in a Hidden Markov Model: underlying state process is Markovian

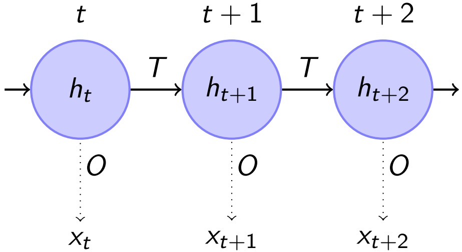
Assumptions for Hidden Markov Model

1. Assumption in a Hidden Markov Model: underlying state process is Markovian
   - $P(h_{t+1}|h_t, \ldots, h_1) = P(h_{t+1}|h_t)$

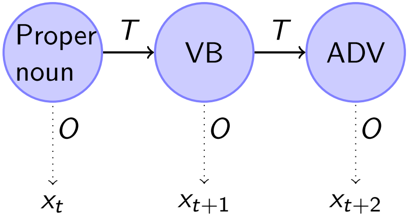
Assumptions for Hidden Markov Model

1. Assumption in a Hidden Markov Model: underlying state process is Markovian
   - $P(h_{t+1}|h_t, \ldots, h_1) = P(h_{t+1}|h_t)$
2. Given the hidden states, the observations are independent

The Hidden Markov Model parameters



The Hidden Markov Model parameters



The Hidden Markov Model parameters

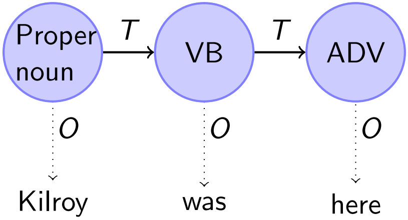


Figure: An example of an HMM with "Kilroy was here" as output

The new users

The new users

The new users

The new users

The new users

The new users

Google

YouTube

The new users

The new users

Google

YouTube

The new users

Google

Neuroscience

YouTube

When spectral methods apply

Let v be the dimension of your observations, and k be the dimension of your hidden state space

When spectral methods apply

Let v be the dimension of your observations, and k be the dimension of your hidden state space

$k << v$

When spectral methods apply

Let v be the dimension of your observations, and k be the dimension of your hidden state space

$k << v$

While observations lie in high-dimensional space v, they distributionally move on a much smaller subspace of dimension k.

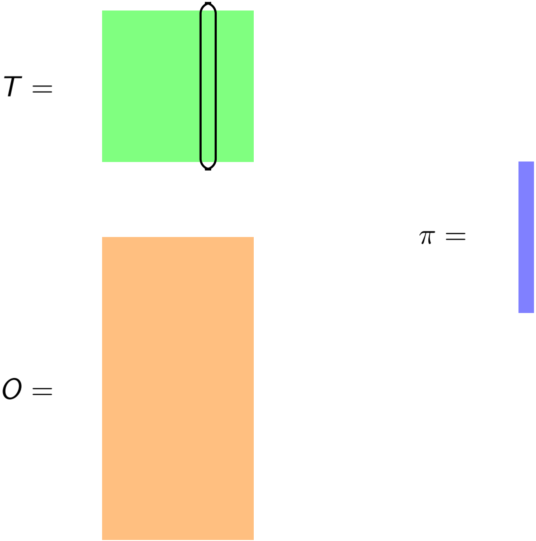
The Hidden Markov Model parameters

$T =$

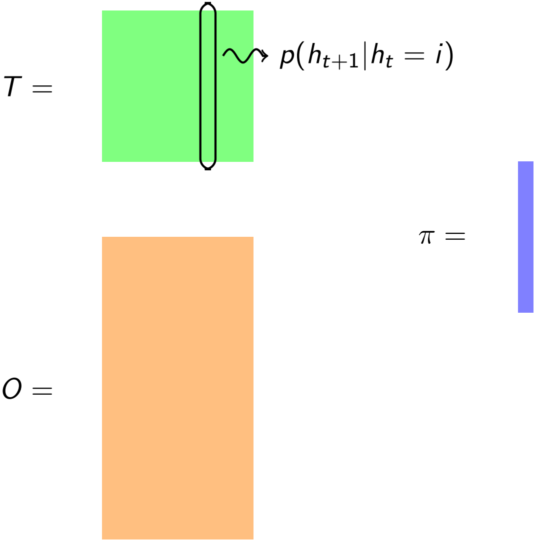
$\pi =$ 

$O =$

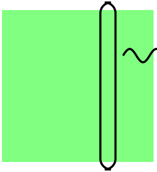
The Hidden Markov Model parameters

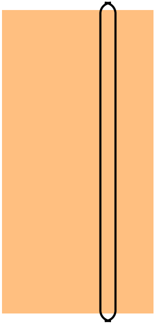


The Hidden Markov Model parameters



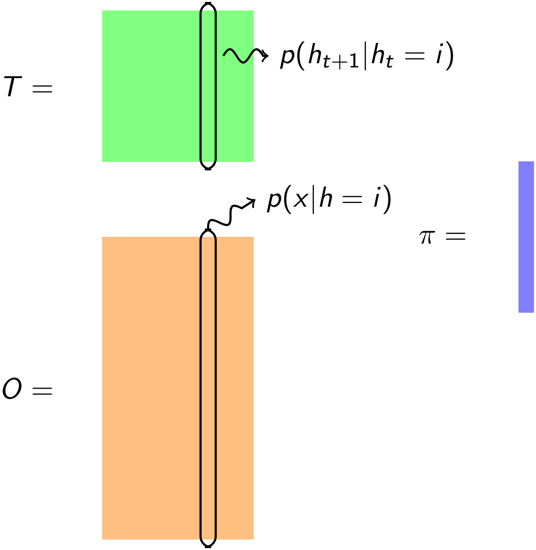
The Hidden Markov Model parameters

$T =$  $\rightsquigarrow p(h_{t+1}|h_t = i)$

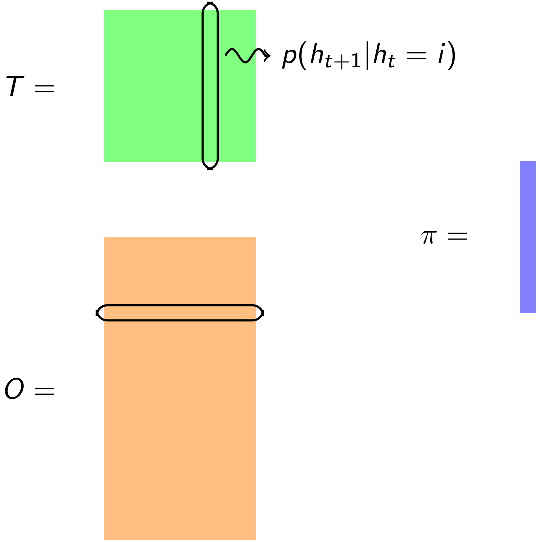
$O =$ 

$\pi =$ 

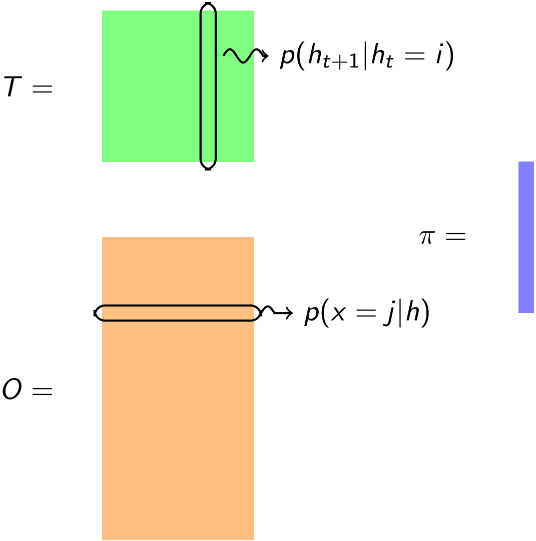
The Hidden Markov Model parameters



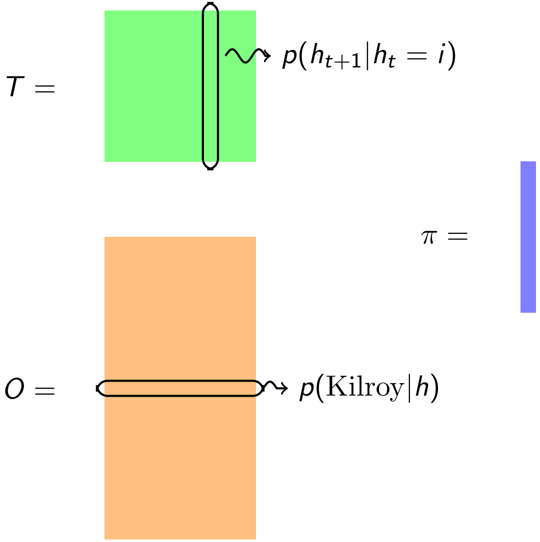
The Hidden Markov Model parameters



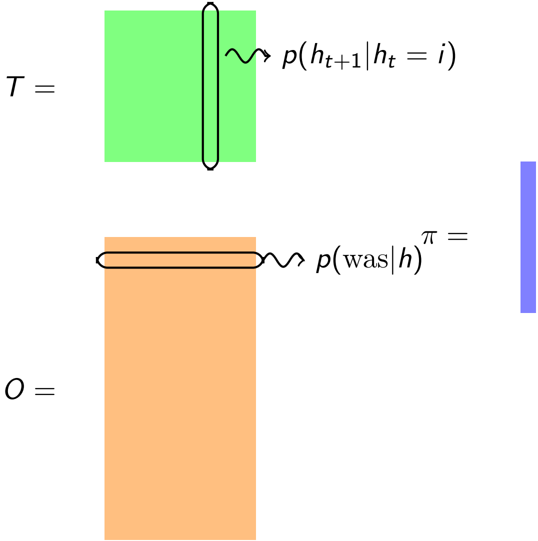
The Hidden Markov Model parameters



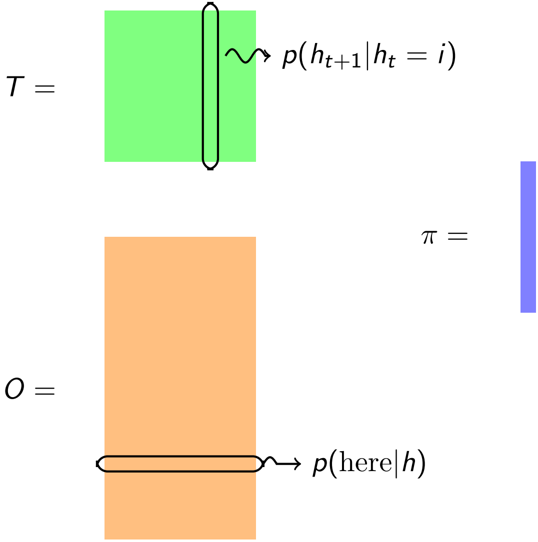
The Hidden Markov Model parameters



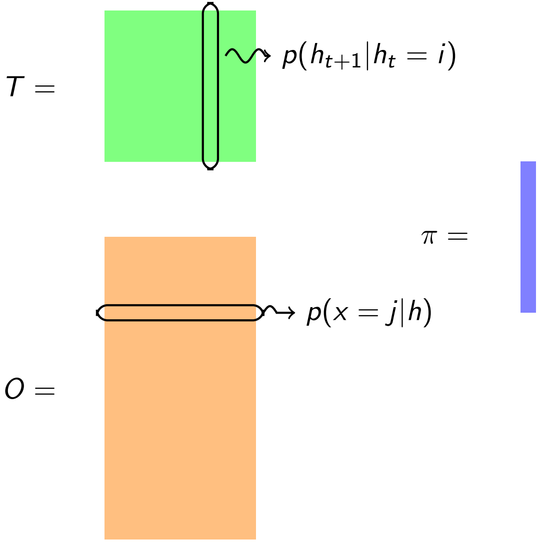
The Hidden Markov Model parameters



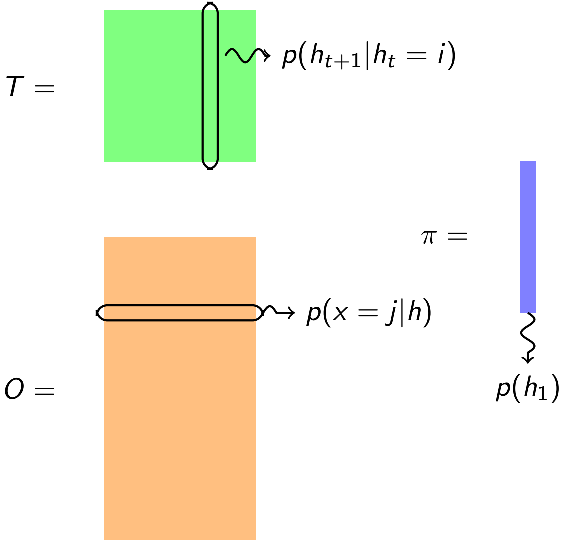
The Hidden Markov Model parameters



The Hidden Markov Model parameters



The Hidden Markov Model parameters



Traditional formula

$$P(x_1, \ldots, x_t) = \sum_{h_1, \ldots, h_t} [\pi]_{h_1} \prod_{j=2}^{t} [T]_{h_j, h_{j-1}} \prod_{j=1}^{t} [O]_{x_j, h_j}$$

The “new” formula

$$P(x_1, \ldots, x_t) = 1^\top A(x_t) \cdots A(x_1) \pi$$

$$\text{where } A(x) = T \operatorname{diag}([O]_{x,\cdot})$$

The "new" formula

$$P(x_1, \ldots, x_t) = 1^\top A(x_t) \cdots A(x_1)\pi$$

$$\text{where } A(x) = T\,\mathrm{diag}([O]_{x,\cdot})$$

In pictures:

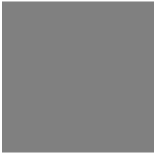
$A(x) =$

The "new" formula

$$P(x_1, \ldots, x_t) = 1^\top A(x_t) \cdots A(x_1)\pi$$

$$\text{where } A(x) = T\,\mathrm{diag}([O]_{x,\cdot})$$

In pictures:

$$A(x) =$$ 

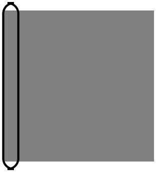
The “new” formula

$$P(x_1, \ldots, x_t) = 1^\top A(x_t) \cdots A(x_1)\pi$$

$$\text{where } A(x) = T\,\mathrm{diag}([O]_{x,\cdot})$$

In pictures:

$$A(x) =$$



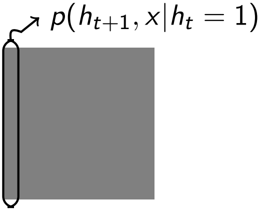
The "new" formula

$$P(x_1, \ldots, x_t) = 1^\top A(x_t) \cdots A(x_1) \pi$$

$$\text{where } A(x) = T \operatorname{diag}([O]_{x,\cdot})$$

In pictures:

$A(x) =$



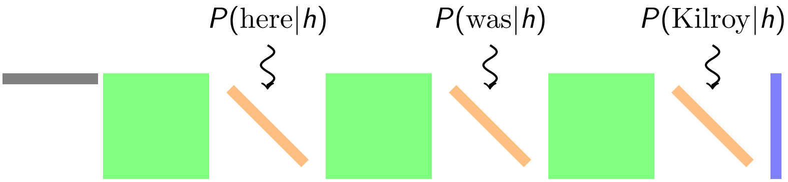
The "new" formula

$$P(\text{Kilroy was here}) = \mathbf{1}^\top\ A(\text{here})\ A(\text{was})\ A(\text{Kilroy})\ \pi$$

The "new" formula

$$P(\text{Kilroy was here}) = 1^\top\ A(\text{here})\ A(\text{was})\ A(\text{Kilroy})\ \pi$$

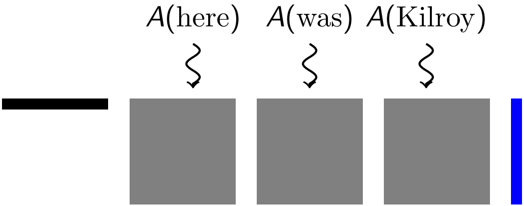
In pictures:



The “new” formula

$$P(\text{Kilroy was here}) = 1^\top \, A(\text{here}) \, A(\text{was}) \, A(\text{Kilroy}) \, \pi$$

In pictures:



Key to estimation

$$1^\top A(x_t) \cdots A(x_1)\pi$$

Main Idea:

For most problems, we don't **need** to recover T and O. We really only need $A(x)$.

Key to estimation

$$P(x_1, \ldots, x_t) = 1^\top A(x_t) \cdots A(x_1)\pi$$

Main Idea:

For most problems, we don't **need** to recover T and O. We really only need $A(x)$.

Key to estimation

$$1^\top A(x_t) \cdots A(x_1)\pi$$

Problem:
$A(x)$ is still not fully observable

Key to estimation

$$1^\top A(x_t) \cdots A(x_1)\pi$$
$$=$$

However:

...

Key to estimation

$$1^\top A(x_t) \cdots A(x_1) \pi$$

$$=$$

$$1^\top \quad \underbrace{S^{-1} S}_{\text{Do nothing}} \quad A(x_t) \ \ S^{-1} S \ \ \cdots \ \ S^{-1} S \ \ A(x_1) \ \ S^{-1} S \ \ \pi$$

However:

...

Key to estimation

$$1^\top A(x_t) \cdots A(x_1)\pi$$

$$=$$

$$1^\top S^{-1} \quad \underbrace{SA(x_t)S^{-1}}_{\text{Similarity Transformation}} \quad S \cdots S^{-1} \; SA(x_1)S^{-1} \; S\pi$$

However:

...

Key to estimation

$$1^\top A(x_t) \cdots A(x_1)\pi$$
$$=$$
$$1^\top S^{-1} \quad \underbrace{SA(x_t)S^{-1}}_{\text{Similarity Transformation}} \quad S \cdots S^{-1} \; SA(x_1)S^{-1} \; S\pi$$

Solution:

Turns out for the right S the similarity transformation of A **is** fully observable

Key to estimation

$$\mathbf{1}^\top A(x_t)\cdots A(x_1)\pi$$
$$=$$
$$b_\infty^\top B(x_t)\cdots B(x_1)b_1$$

Solution:
Turns out for the right S the similarity transformation of A **is** fully observable

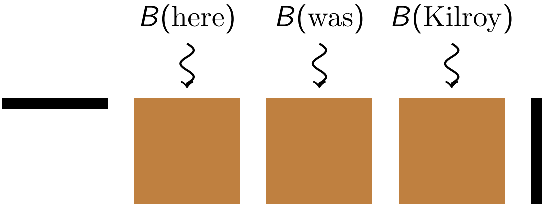
The “new” formula

$$P(\text{Kilroy was here}) = b_\infty^\top \; B(\text{here}) \; B(\text{was}) \; B(\text{Kilroy}) \; b_1$$

The "new" formula

$$P(\text{Kilroy was here}) = b_\infty^\top \; B(\text{here}) \; B(\text{was}) \; B(\text{Kilroy}) \; b_1$$

In pictures:



The HKZ formulation

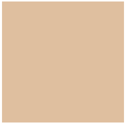
The HKZ formulation

To estimate the $B(x)$ matrices, we need the first three moments...

The HKZ formulation

To estimate the $B(x)$ matrices, we need the first three moments...

$E[x_1] =$

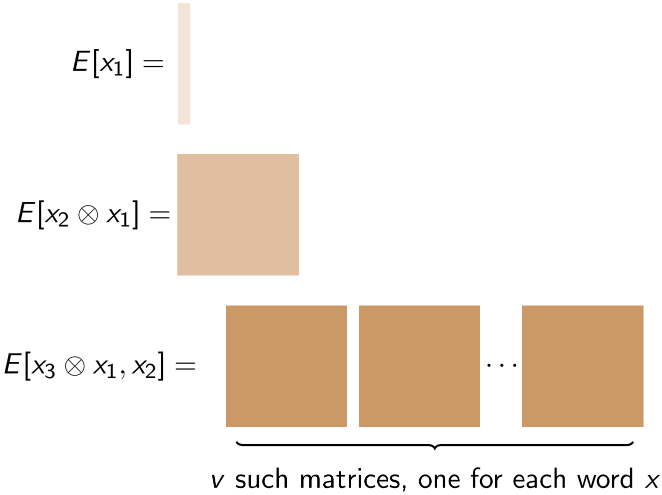
$E[x_2 \otimes x_1] =$ 

$E[x_3 \otimes x_1, x_2] =$ 

v such matrices, one for each word x

The HKZ formulation

To estimate the $B(x)$ matrices, we need the first three moments...



And an eigendictionary U that maps the moments to the lower dimensional subspace...

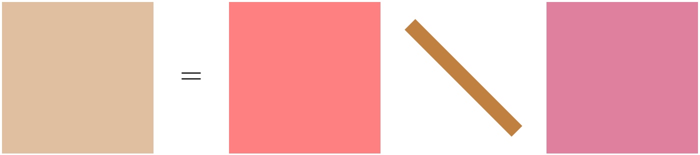
More about U

U maps observations x to a lower dimensional y in a way that preserves the underyling dynamics of x.

More about U

U maps observations x to a lower dimensional y in a way that preserves the underyling dynamics of x.

SVD

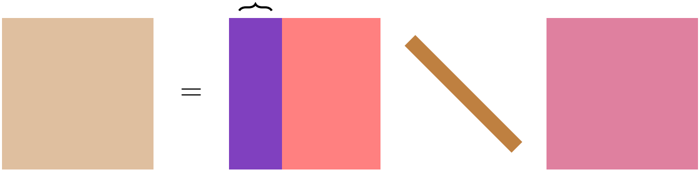


More about U

U maps observations x to a lower dimensional y in a way that preserves the underyling dynamics of x.

SVD

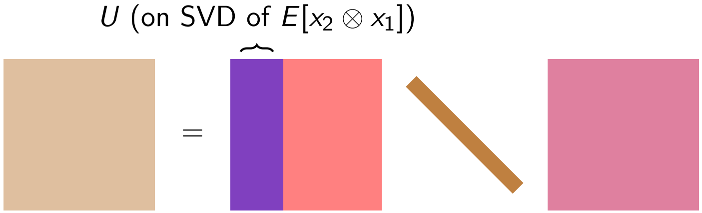
U (on SVD of $E[x_2 \otimes x_1]$)



More about U

U maps observations x to a lower dimensional y in a way that preserves the underlying dynamics of x.

SVD

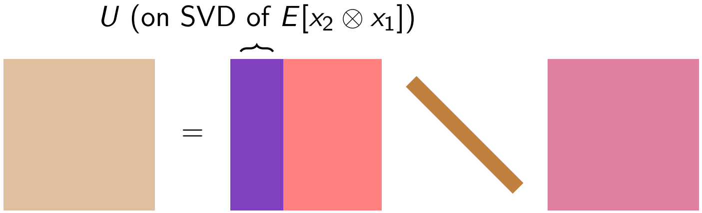


- In the case of words, words that are distributionally similar will map closely together in y-space.

More about U

U maps observations x to a lower dimensional y in a way that preserves the underyling dynamics of x.

SVD



- In the case of words, words that are distributionally similar will map closely together in y-space.
- Example: “I will let **him** know” and “I will let **her** know”, but not “I will let **box** know”

U projection, first two dimension

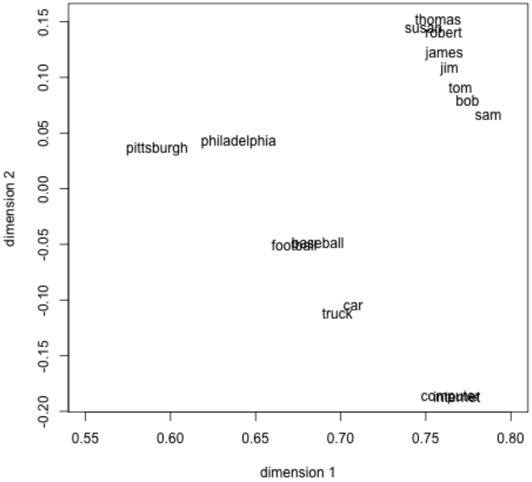


Figure: Projection of words onto the first two dimensions of the U matrix

U projection, second two dimensions

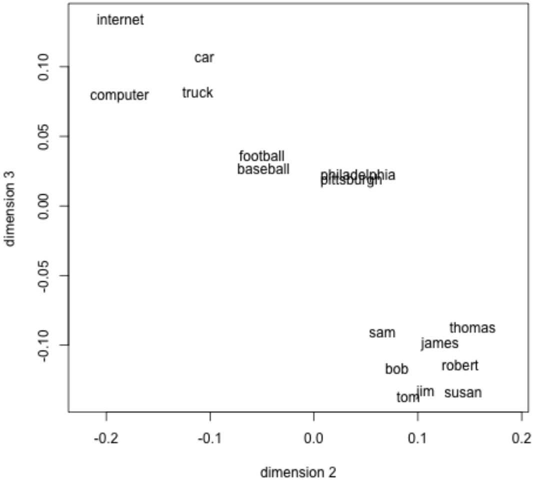


Figure: Projection of words onto the second two dimensions of the U matrix

Moving beyond basic HKZ

We have these v third-moment matrices lying around.

Moving beyond basic HKZ

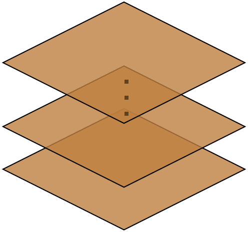
We have these v third-moment matrices lying around.

An alternate way to think of these is to simply stack them

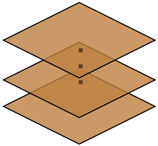
Moving beyond basic HKZ

We have these v third-moment matrices lying around.

An alternate way to think of these is to simply stack them

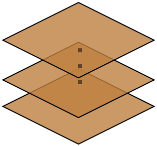


Moving beyond basic HKZ



What changes?

Moving beyond basic HKZ



What changes?

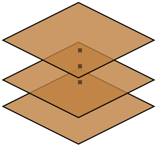
Old way

- Separate matrix for each word
- Select matrix corresponding to the word of interest from a list

Tensor version

- One single tensor for all words
- Have a function that takes a vector (the word) to a matrix

Moving beyond basic HKZ



What changes?

Old way

- Separate matrix for each word
- Select matrix corresponding to the word of interest from a list

Tensor version

- One single tensor for all words
- Have a function that takes a vector (the word) to a matrix

So the natural question is, can we reduce the dimensionality of the third mode of the tensor?

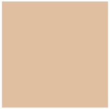
Reduced Dimensional HMM

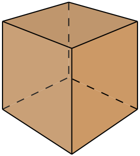
Yes! Calculate the three first moments using the reduced dimensional observations $y = U^\top x$.

Reduced Dimensional HMM

Yes! Calculate the three first moments using the reduced dimensional observations $y = U^\top x$.

$$E[y_1] =$$

$$E[y_2 \otimes y_1] =$$ 

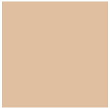
$$E[y_3 \otimes y_1 \otimes y_2] =$$ 

Reduced Dimensional HMM

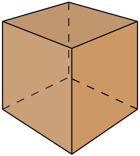
Yes! Calculate the three first moments using the reduced dimensional observations $y = U^\top x$.

$$E[y_1] =$$

$$E[y_2 \otimes y_1] =$$



$$E[y_3 \otimes y_1 \otimes y_2] =$$



Using these moments we construct a function $C(\alpha)$ such that

$$P(x_1, \ldots, x_T) = c_\infty^\top C(y_T) \cdots C(y_1) c_1$$

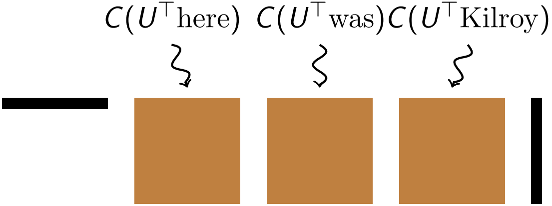
The fully reduced formula

$$P(\text{Kilroy was here}) = c_\infty^\top \; C(U^\top \text{here}) \; C(U^\top \text{was}) \; C(U^\top \text{Kilroy}) \; c_1$$

The fully reduced formula

$$P(\text{Kilroy was here}) = c_\infty^\top \; C(U^\top \text{here}) \; C(U^\top \text{was}) \; C(U^\top \text{Kilroy}) \; c_1$$

In pictures:



Typical theorem

Theorem: Foster, Rodu, Ungar

Let X_t be generated by an $m \geq 2$ state HMM. Suppose we are given a U which has the property that range$(O) \subset$ range(U) and $|U_{ij}| \leq 1$. Using N independent triples, we have

$$N \geq \frac{128 m^2}{(\sqrt[2t+3]{1+\epsilon} - 1)^2 \, \Lambda^2 \sigma_m^4} \log\left(\frac{2m}{\delta}\right) \cdot \overbrace{\frac{\epsilon^2/(2t+3)^2}{(\sqrt[2t+3]{1+\epsilon} - 1)^2}}^{\approx 1}$$

implies that

$$1 - \epsilon \leq \left| \frac{\widehat{\Pr}(x_1, \ldots, x_t)}{\Pr(x_1, \ldots, x_t)} \right| \leq 1 + \epsilon$$

holds with probability at least $1 - \delta$.

Spectral Methods Overview part II

Spectral Methods Overview part II

HKZ

Spectral Methods Overview part II

HKZ

Reduced Dimension HMM

Spectral Methods Overview part II

HKZ

Reduced Rank HMM

Reduced Dimension HMM

Spectral Methods Overview part II

HKZ

Flexible Estimation with Regression
(A topic for another day)

Reduced Rank HMM

Reduced Dimension HMM

Spectral Methods Overview part II

HKZ

Flexible Estimation with Regression
(A topic for another day)

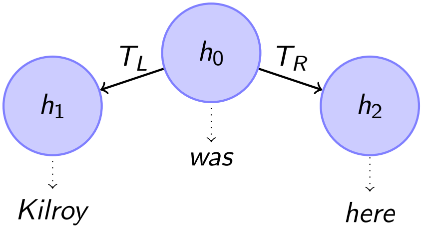
Reduced Rank HMM

Reduced Dimension HMM

Trees

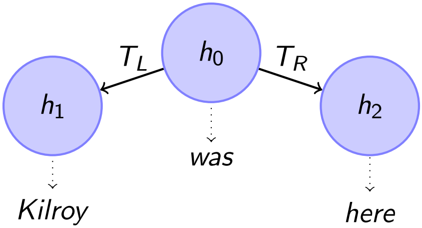
Trees

Extension to hidden variable tree models



Trees

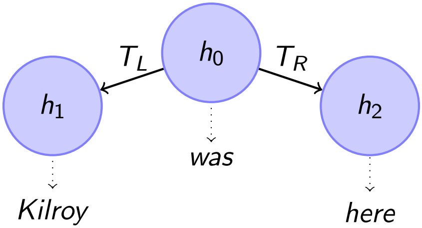
Extension to hidden variable tree models



- Very similar to structure of Hidden Markov Models

Trees

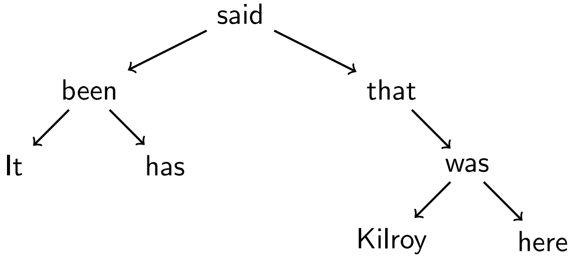
Extension to hidden variable tree models



- Very similar to structure of Hidden Markov Models
- Requires a few modifications, for instance
  1. Defining left and right transition parameters
  2. Estimating additional skip-bigram matrix instead of just bigram.

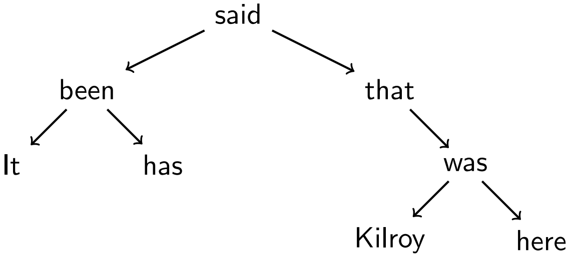
Trees

Extension to hidden variable tree models



Trees

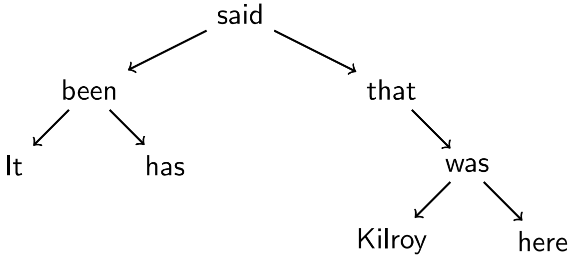
Extension to hidden variable tree models



- Want the marginal probability of a particular tree and tree topology

Trees

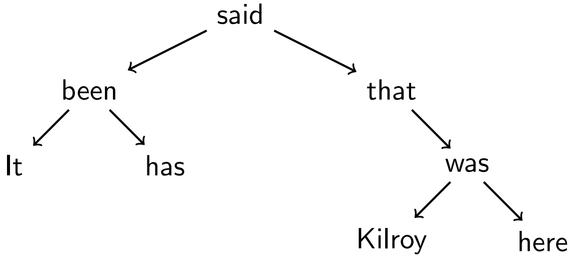
Extension to hidden variable tree models



- Want the marginal probability of a particular tree and tree topology
- Sum over all possible hidden states (not shown on this slide)

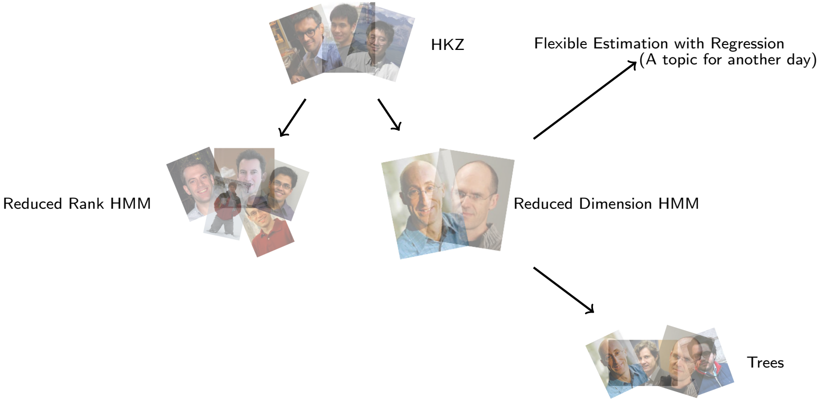
Trees

Extension to hidden variable tree models

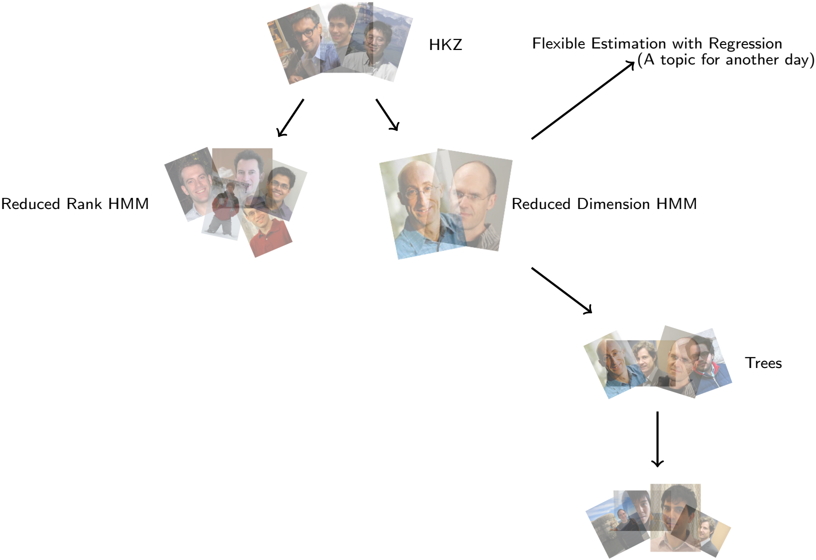


- Want the marginal probability of a particular tree and tree topology
- Sum over all possible hidden states (not shown on this slide)
- Can be used for re-ranking output from a parser

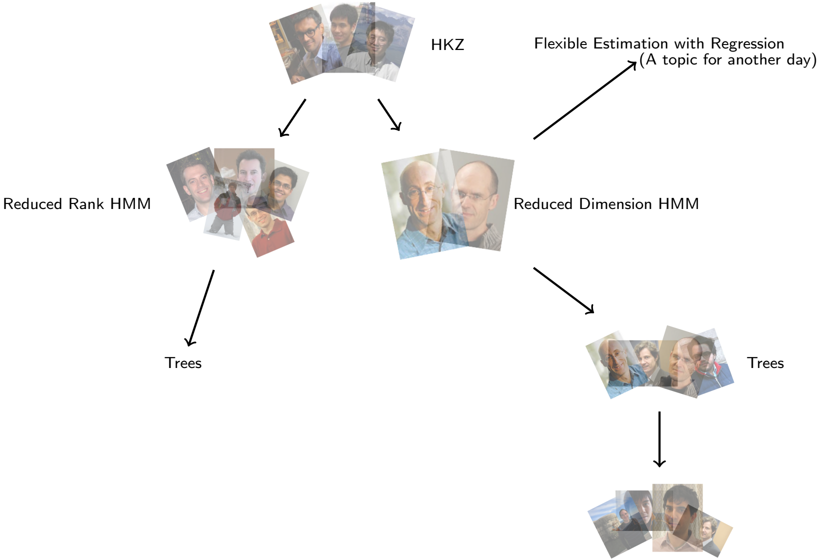
Spectral Methods Overview part II



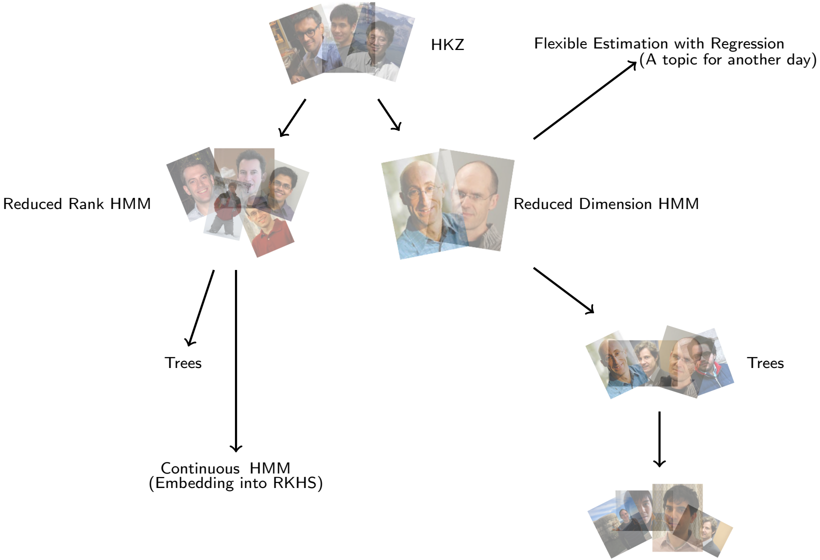
Spectral Methods Overview part II



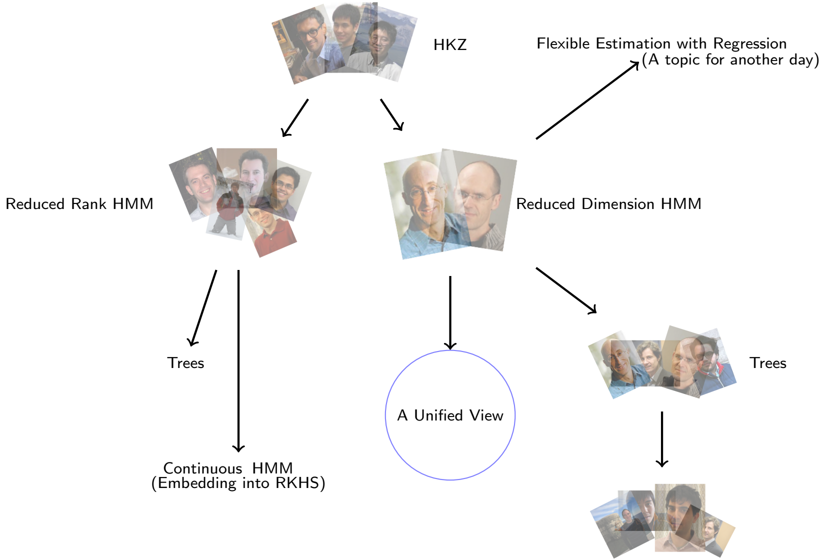
Spectral Methods Overview part II



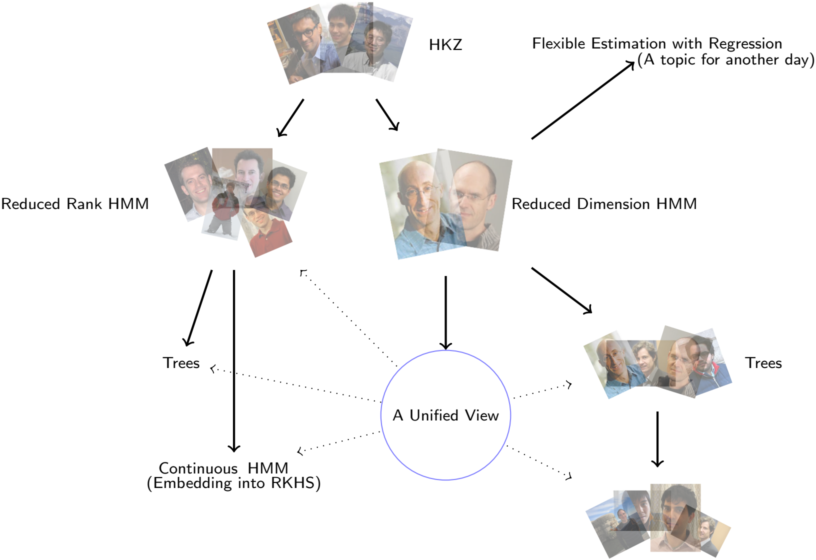
Spectral Methods Overview part II



Spectral Methods Overview part II



Spectral Methods Overview part II



A Unified View

We've learned something valuable from the reduced dimensional spectral HMM.

A Unified View

We've learned something valuable from the reduced dimensional spectral HMM.

Estimation of the hidden state dynamics in spectral estimation is a separate problem from estimation of the output distribution.

A Unified View

We've learned something valuable from the reduced dimensional spectral HMM.

Estimation of the hidden state dynamics in spectral estimation is a separate problem from estimation of the output distribution.

- Any function of the data that preserves the underlying dynamic structure can be used to the first three moments of the data and build the tensor $C(\alpha)$.
- The tensor $C(\alpha)$ fully encodes all of the information for the hidden state dynamics.

A Unified View

We've learned something valuable from the reduced dimensional spectral HMM.

Estimation of the hidden state dynamics in spectral estimation is a separate problem from estimation of the output distribution.

- Any function of the data that preserves the underlying dynamic structure can be used to the first three moments of the data and build the tensor $C(\alpha)$.
- The tensor $C(\alpha)$ fully encodes all of the information for the hidden state dynamics.
- Estimation of the observation distribution lies in the choice of α, and how α *plugs into* the tensor $C(\alpha)$.

A Unified View

We've learned something valuable from the reduced dimensional spectral HMM.

Estimation of the hidden state dynamics in spectral estimation is a separate problem from estimation of the output distribution.

- Any function of the data that preserves the underlying dynamic structure can be used to the first three moments of the data and build the tensor $C(\alpha)$.
- The tensor $C(\alpha)$ fully encodes all of the information for the hidden state dynamics.
- Estimation of the observation distribution lies in the choice of α, and how α *plugs into* the tensor $C(\alpha)$.
  - For HKZ, $\alpha = x$

A Unified View

We've learned something valuable from the reduced dimensional spectral HMM.

Estimation of the hidden state dynamics in spectral estimation is a separate problem from estimation of the output distribution.

- Any function of the data that preserves the underlying dynamic structure can be used to the first three moments of the data and build the tensor $C(\alpha)$.
- The tensor $C(\alpha)$ fully encodes all of the information for the hidden state dynamics.
- Estimation of the observation distribution lies in the choice of α, and how α *plugs into* the tensor $C(\alpha)$.
  - For HKZ, $\alpha = x$
  - For the RDHMM, $\alpha = y$

A Unified View

We've learned something valuable from the reduced dimensional spectral HMM.

Estimation of the hidden state dynamics in spectral estimation is a separate problem from estimation of the output distribution.

- Any function of the data that preserves the underlying dynamic structure can be used to the first three moments of the data and build the tensor $C(\alpha)$.
- The tensor $C(\alpha)$ fully encodes all of the information for the hidden state dynamics.
- Estimation of the observation distribution lies in the choice of α, and how α *plugs into* the tensor $C(\alpha)$.
  - For HKZ, $\alpha = x$
  - For the RDHMM, $\alpha = y$
  - For a general distribution, $\alpha = E[y_{t+1}|x_t] = g(x_t)$.

The fully factored approach: A few example extensions

The fully factored approach: A few example extensions

Factorial HMM

- One model of stock return covariance matrices is that they are generated by an HMM with parameters that vary over time, themselves according to an HMM
- Requires ability to estimate HMM with matrix-valued output, which is possible with the factored approach

The fully factored approach: A few example extensions

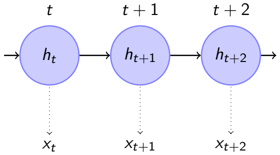
Factorial HMM

- One model of stock return covariance matrices is that they are generated by an HMM with parameters that vary over time, themselves according to an HMM
- Requires ability to estimate HMM with matrix-valued output, which is possible with the factored approach

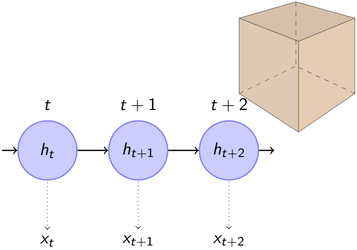
Occasionally available external information

- Consider a system in which external information is occasionally available that might refine our hidden state belief
  - Amazon Mechanical Turk- have people intermittently label a stochastic process (e.g. text or images) as a way to recalibrate an automatic labeling.
- Requires ability to modify the probability of seeing an observation given the hidden states, now possible with the fully factored approach!

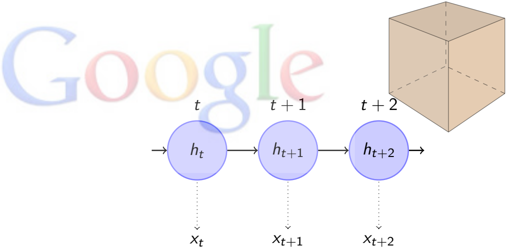
Spectral Estimation of HMMs



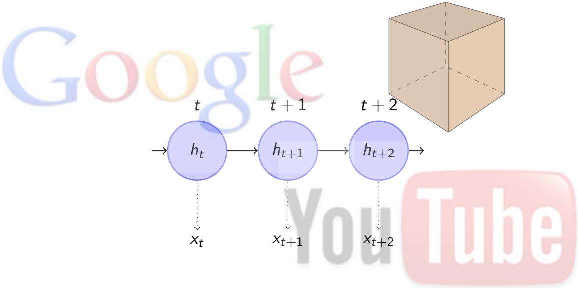
Spectral Estimation of HMMs



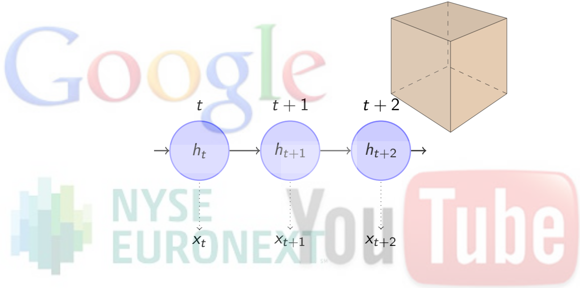
Spectral Estimation of HMMs



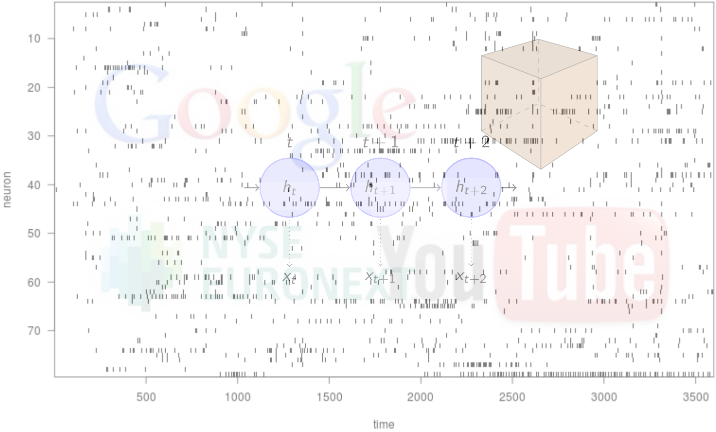
Spectral Estimation of HMMs



Spectral Estimation of HMMs



Spectral Estimation of HMMs



Thanks!!

Papers

Using Regression for Spectral Estimation, Foster, Rodu, Ungar, Wu 2013

Two Step CCA: A new spectral method for estimating vector models of words, Dhillon, Foster, Rodu, Ungar 2013

Spectral Dependency Parsing with Latent Variables, Collins, Dhillon, Foster, Rodu, Ungar 2012

Spectral Demensionality Reduction for HMMs, Foster, Rodu, Ungar 2012

In Progress

Spectral Estimation of HMMs with a continuous output distribution, Foster (in progress)

Spectral Estimation of hierarchical HMMs, Foster, Rodu, Sedoc, Ungar (in progress)

An MDP clustering of neurons by their hidden state paths Jensen, Rodu, Small (in progress)

Thanks!