## Recap & Look ahead

Emily Fox & Carlos Guestrin Machine Learning Specialization University of Washington

©2016 Emily Fox & Carlos Guestrin

#### What we've learned

#### Module 1: Nearest neighbor search

## 1-NN search

Space of all articles, organized by similarity of text



## k-NN search

Space of all articles, organized by similarity of text



©2016 Emily Fox & Carlos Guestrin

Machine Learning Specialization

## **TF-IDF** document representation

#### Emphasizes important words

- Appears frequently in document (common locally)



#### Scaled Euclidean distance

distance(
$$\mathbf{x}_{i}, \mathbf{x}_{q}$$
) =  
 $\sqrt{a_{1}(\mathbf{x}_{i}[1] - \mathbf{x}_{q}[1])^{2} + ... + a_{d}(\mathbf{x}_{i}[d] - \mathbf{x}_{q}[d])^{2}}$ 

#### weight on each feature

#### Educed M. Ramony

m) dad was working intry hard to prover data. The small not prover it, and us all same back home. Among was no die breit, of your 1 andressend data my two oblied headars was shicking almost velocing.

Note up compa pairs as the prospect of the size Knowly oblices, my loady due to a set by load of a solid or compare for each (Dough F how the load of the solid or compare for each (Dough F how the load set each of the stephenic size. I have some questioned its constant work We depended upon our secretion We second load and some and logders with our seconder. We second load and sample near another, We constanting the second loads. We have a second logder with some standing logd second one second

10.00

monady competence, will an intensity that send more to bu Aut to us upp for Aminano. They rates fored inc. as as the purples of Joseph and Rose Kennedy. They helped as form basely period one peoples, and to develop peo-Rand on door heads, to an errors that pressing to the day and its the down has all as family. We me call. They for as the boost of the areas I wink so will. I was also pours shill in that summer of (1943, the R memory of the handlar world into which I was been. I was or char why up had all come back know from Engl hat I not keppy that we had. I was not proved to fully and-research that my fathers had only and his one I are restrictly any second to compare here I also by it assesses shars the British wight and he republic of Ephring a war again between its would have been prove to me that that had displeand Persident Research with door some tomorha O that when he was some from the Case house that memory is New York and Washington, he was stying to presents when people to join her effort at keying America out of Access Or that designs they differences through Konsult command a segment Franklin Research as president. I just have that on machinals, he and I would take horadack regelier on the Cape, and their new all I really cared We hardly approximing that these house of my father the as a child. If my father over alles

17.41.000

Name (Company) & Street,

#### title abstract main body conclusion

Educat & Farmerij	Part Dorogeneer A. Her	
and the second se		
Contracting and apprendict of the contraction of the		
provide a set of provide back back. Another that is not	and the second se	
the start of the s	have been part of charge and from the	
I submit for first of an for one	hand or down hash in or owned of	
descent on the set of second and a build out had been a	and the second se	
miners of an effect with New of a sector receipt	mail The Scott has done	
and the local sector of the		
the second se	and the local sector would be a set	
strategies in the second descent the second s	and the set of the set of the set of the	
is the metrory working the provides of Jack and they findly.	has I are have due to had I a	
and does to some doorse strand as hold more any commit-	and some likes as following and some	
and the statement with the second for second diversion and the second	I have been a first the second s	
The shall show I have an increased in form of muse them	Income in had allowed over small	
dentities infor the start and arrow would be to see - prime fit.	that the British winds are be reaching	
on Annalase distance or Research time, or makers process	Corners & would have been serve	
endous or entry of even based The use or helper	dashard Propher Research with d	
both one the properties of the last at the station	that also be not store from the Case	
	Non-York and Windowson, Science of	
From an excitant point as the compare of the nine	people to join his offset of heating has	
Knowsh children we hand, dol and as many hirs in the world	that deside they difference deside	
as comprise the world. Through I have long since comprises	separat Frankla Research as preside	
that another time I have some particular in unstand	I not have that on mathematic	
much. We depended upon one another. We second hard and	Installanti sugerbar on the Cape, and di	
many and lengther with our another. We having from and	and some	
mayle one modes. We workgood with one modes, We	b's hardly perpending due they	
keed at another. We save manufa hand area as we saw	his ware asknow the to be a shift.	

10.00

on young to help

on the Dad he

some namedra Or

a first moments in

a ter permanen ber auffahre

and a summer of the

ward Constitutes manual Constructed

eres of my lister's

an falser area alies

114100

©2016 Emily Fox & Carlos Guestrin

Machine Learning Specialization

7

#### Cosine similarity – normalize

Similarity =

$$= \sum_{j=1}^{d} \mathbf{x}_{j}[j] \mathbf{x}_{q}[j]$$

$$\sqrt{\sum_{j=1}^{d} (\mathbf{x}_{j}[j])^{2}} \sqrt{\sum_{j=1}^{d} (\mathbf{x}_{q}[j])^{2}}$$

- Not a proper distance metric

 Efficient to compute for sparse vecs

$$= \mathbf{x}_{i}^{\top} \mathbf{x}_{q} = \cos(\mathbf{\theta})$$

 $||\mathbf{x}_{i}||||\mathbf{x}_{q}||$ 



#### To normalize or not?

short tweet

Educid R. Konsely	Pue Compiler A Remain
	and the second se
my dial was working very hard to preverse disk. The year's not	mentally competence, with an presently that recel more to py
percent it, and we all came back home. America was on the	that is at args for Amenanon. They rates formed into an
fortide of sign 1 and/original Past any Taxa at the Southern work	on the purchase of charge and from furnishing they begind an
Acking alout relating	here feach setting one proflet, and to develop provideline
I submoul dow drop, and yet day even	hand on dran heich, to an atom due tenum to die day
Annumbers, Rantag showers in the fail and, fail dramed	parties approvaled by the choice for all the family. They means
perions of a shall hap's word. Note of an perhaps comprog-	the still. They be at the houst of the ency ( studt to sell,
ny farhes could anticipate what the way much mean to us	5 was time priori old in that summer of (741), the fixed
the terrible anothers is would pract within our family. Not	meteriari of the Samilar month one who it I man here. I man
even my fasher could imagine the commulity of the Rammelys	not clear why we had all come back konsi from England-
a de porte a wild de arregées of Jack, and das faille.	hat I now happy that we had. I was not young to fully
and then to some degree aryoull, to hold agost our country's	and-read that my father had received to and-sead-relep-
addary showy with statutes for world parties and demonstra-	I have contactedly accounted to comparishend show he'd averagened
The plaid chirted figure on hierarback is from of no on those	because he had affended some proph in Kegheid by soving
averting rides was well and arror would be to me - primarily	that the British wight out he capable of Eghning a war optimal
an Antonicus Aghanas, or Emotion lines, or motion pacture	Germany: It would have been arow to me that Dad had
peraduates or matter of entries beyond. The task my Robert	dapleand Persiline Rosarch with dom some romarke. Or
Auch use the perspectice of the boy on the statling	that which he was a wap from the Cape house that moments in
here.	New York and Washington, he was stying to personale other
From my vantage print as the prompter of the mine	people to join his effort at heiging Amorea out of the war. Or
Konsely children, my family dul not so much live in the world	Aut, despite their Alberta on Jony's Kontoly contrand to
as comprise the world, Dhough I have long since compress	sepport Franklin Research as president.
that aimplatic time, I have server spectrosed in unsertional	3 jast have that on unchrade, he used I would rele-
mark. We depended upon one another. We servered land and	bornhack regelest us its Cape, and due nos all I only cared
more and lengther with one another. We harred from and	achare.
angle can another. We wandapped with our another. We	Sr's kardy anyrising due these larges of my failur's
head see another. We need musically legal, over as we were	his wave autoassable to us as a child. If any father over alree
1.010	17.47.00

long document

Normalizing can make dissimilar objects appear more similar

Edward K. Russelly	Pase (sergence 2, Barran
a dad any working one hard to prever this. To could not	metals requires will be break the series are a
promits and or of consthats home. Assertes was on the	that is an area for Aminana. They takes frond inc. a
while of most I and contrast that my tree clobed brachers more	on the margine of Joseph and Rose Kennethe They helped as
histing along coloning	here beach senses are product and to develop percendition
Y subsend day days, and up day one	hand on down hands, to an errors that termine to the day
Assurances, flaring chosenes in the half-coal, half-drammed	maker appreciated by the chronic her of an family. Were monitor
store of a stuff lock wind. Note of an architectory	one will Then for an the board of the many I wink to will.
to fashed could astrongen when the way small mean to us	I may now poors and in that response of (NI), the Read
he workly apprillers it would start within our family. Not	memory of the handler world into which I was been, I was
run my fasher could imagine the controlity of the Kontrolys	not clear ally us had all come back home from England-
a die pusture wuld die eruggies of Jack, and das Bublic	has I now keppy that we had, I was not young to hilly
ad days to some degree struck to hadd agent out conserv's	and-resulting as false had sugged to antonadarity.
address views have been been and assess they exclude	I have containedly acce proving the comparishment alone has it ensigned
The plant chirted figure on Noraelands in front of me on those	Income to had allowed some proph in England to using
scening roles was and - and arror would be to me - primarily	that the British wight aut he republic of Ephring a war optimat
in Associate diploment or Essencial times, or motion protons	Germany: It would have been arow to me that Dad had .
endness or more of easts beyond. He use my father	depleand Persilies Research with down some remarks. Or
Back was the parapetters of the bap on the stalling	that when he was away from the Cape house that memory in
area.	New York and Washington, he was trying to presente other
From my vantage point as the prompter of the nine	people to prin his effort, at herging Amories and of Access (A
formedy children, my family did not as much free in the world	Aut, despite their differences, Joseph Kontech, continued to
a comprise the world. Through I have long since comproves	support Frankla Research as president.
hat applicate stars, I have server questioned its unserioud	3 just have that on wederade, he and I would tale
rath. We depended upon one another. We second load and	bornhack regelest on the Cape, and doe not all I really cared
ness and leighter with one another. We hursted from and	achare.
angle one another. We wandapped with one another. We	S's kardy anyrising due dese larms of my fader's
trad our another. We uses manually legal, over as we cam	his wave and an wall is to see as a child. If any father uses allow
16.01.00	14.00



long document

## long document

#### Common compromise: Just cap maximum word counts

## Complexity of brute-force search

Given a query point, scan through each point

– O(N) distance computations per 1-NN query!

- O(Nlogk) per k-NN query!



Machine Learning Specialization

#### **KD-trees**



## Nearest neighbor with KD-trees



- 1. Start by exploring leaf node containing query point
- 2. Compute distance to each other point at leaf node
- 3. Backtrack and try other branch at each node visited

#### Nearest neighbor with KD-trees



Use distance bound and bounding box of each node to prune parts of tree that cannot include nearest neighbor

#### Approximate k-NN with KD-trees



**Before:** Prune when distance to bounding box > r **Now:** Prune when distance to bounding box >  $r/\alpha$ 

Saves lots of search time at little cost in quality of NN!

## Limitations of KD-trees

- Difficult to implement
- Don't tend to perform well in high dimensions
  - Under some conditions, visit at least 2<sup>d</sup> nodes



## Locality sensitive hashing



©2016 Emily Fox & Carlos Guestrin

Machine Learning Specialization

## LSH for approximate NN search



#### Module 2: k-means and MapReduce

## Module 2: k-means and MapReduce

#### Discover *clusters* of related documents







Cluster 4

## k-means algorithm

- 0. Initialize cluster centers
- 1. Assign observations to closest cluster center
- 2. Revise cluster centers as mean of assigned observations
- 3. Repeat 1.+2. until convergence



#### A coordinate descent algorithm

1. Assign observations to closest cluster center

$$z_i \leftarrow \arg\min_j ||\mu_j - \mathbf{x}_i||_2^2$$

2. Revise cluster centers as mean of assigned observations

$$\mu_j \leftarrow \arg\min_{\mu} \sum_{i:z_i=j} ||\mu - \mathbf{x}_i||_2^2$$

Alternating minimization 1. (z given μ) and 2. (μ given z) = coordinate descent

©2016 Emily Fox & Carlos Guestrin

#### Convergence of k-means to local mode



#### MapReduce framework



Machine Learning Specialization

#### MapReduce abstraction

#### Map:

- Data-parallel over elements, e.g., documents
- Generate (key,value) pairs
  - "value" can be any data type

#### Word count example:

map(doc) for word in doc emit(word,1)

#### **Reduce:**

- Aggregate values for each key
- Must be commutative-associative operation
- Data-parallel over keys
- Generate (key,value) pairs

reduce(word, counts\_list)
c = 0
for i in counts\_list
c += counts\_list[i]
emit(word, c)

MapReduce has long history in functional programming

Popularized by Google, and subsequently by open-source Hadoop implementation from Yahoo!
 ©2016 Emily Fox & Carlos Guestrin

## MapReducing 1 iteration of k-means

**Classify:** Assign observations to closest cluster center

$$z_i \leftarrow \arg\min_j ||\mu_j - \mathbf{x}_i||_2^2$$

**Map:** For each data point, given  $(\{\mu_i\}, \mathbf{x}_i)$ , emit $(z_i, \mathbf{x}_i)$ 

**Recenter:** Revise cluster centers as mean of assigned observations

$$\mu_j = \frac{1}{n_j} \sum_{i: z_i = k} \mathbf{x}_i$$

**Reduce:** Average over all points in cluster j ( $z_i = k$ )

#### Module 3: Mixture models

## Mixture models

#### Probabilistic clustering model



#### captures uncertainty in clustering

#### Failure modes of k-means



#### Jumble of unlabeled images



#### Model of jumble of unlabeled images



Machine Learning Specialization

## Mixture of Gaussians (1D)

Each mixture component represents a unique cluster specified by:



# Mixture of Gaussians for clustering documents

Space of all documents (really lives in  $\mathbf{R}^{\vee}$  for vocab size V)



©2016 Emily Fox & Carlos Guestrin

### Restricting to diagonal covariance

Each cluster has { $\pi_k$ ,  $\mu_k$ ,  $\Sigma_k$  diagonal }



Machine Learning Specialization

#### Inferring cluster labels

Data



## **EM algorithm →** soft assignments



©2016 Emily Fox & Carlos Guestrin

Machine Learning Specialization

#### Expectation maximization (EM): An iterative algorithm

**1. E-step:** <u>e</u>stimate cluster responsibilities given current parameter estimates

$$\hat{r}_{ik} = \frac{\hat{\pi}_k N(x_i \mid \hat{\mu}_k, \hat{\Sigma}_k)}{\sum_{j=1}^K \hat{\pi}_j N(x_i \mid \hat{\mu}_j, \hat{\Sigma}_j)}$$

**2. M-step:** <u>maximize likelihood over</u> parameters given current responsibilities  $\hat{\pi}_k, \hat{\mu}_k, \hat{\Sigma}_k \mid \{\hat{r}_{ik}, x_i\}$ 

## EM for mixtures of Gaussians in pictures - replay



Machine Learning Specialization
## Relationship to k-means

Consider Gaussian mixture model with



and let the variance parameter  $\sigma \rightarrow 0$ 

Datapoint gets fully assigned to nearest center, just as in k-means

### Module 4: Latent Dirichlet allocation

# Topic vocab distributions:

SCIENCE		
experiment	0.1	
test	0.08	
discover	0.05	
hypothesize	0.03	
climate	0.01	

TECH			
develop	0.18		
computer	0.09		
processor	0.032		
user	0.027		
internet	0.02		

SPORTS		
player	0.15	
score	0.07	
team	0.06	
goal	0.03	
injury	0.01	

Drausin F. Wulsin<sup>a</sup>, Emily B. Fox<sup>c</sup>, Brian Litt<sup>a,b</sup>

Modeling the Complex Dynamics and Changing

Correlations of Epileptic Events

<sup>a</sup>Department of Bioengineering, University of Pennsylvania, Philadelphia, PA <sup>b</sup>Department of Neurology, University of Pennsylvania, Philadelphia, PA <sup>c</sup>Department of Statistics, University of Washington, Seattle, WA

Abstract

with epilepsy can manifest short, sub-clinical epileptic "bursts" in addition to tui-blown clinical seizures. We believe the relationship between ese two classes of events—something not previously studied quantitatively ould yield important insights into the nature and intrinsic dynamics of A goal of our work is to parse these complex epileptic events listinct dynamic regimes. A challenge posed by the intracranial EEG EFG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable number of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of EEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

*Keywords:* Bayesian norparametric, EEG, factorial hidden Markov model, graphical model, time series

### 1. Introduction

Despite over three decrdes of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible

### **Clustering**:

One topic indicator  $z_i$  per **document** i

All words come from (get scored under) same topic z<sub>i</sub>

Distribution on prevalence of topics in **corpus**  $\mathbf{\pi} = [\pi_1 \ \pi_2 \dots \pi_K]$ 

## **Comparing and contrasting**

Previously

Prior topic probabilities

$$p(z_i = k) = \pi_k$$

Likelihood under each topic



compute likelihood of **tf-idf** vector under each **Gaussian** 

tf-idf vector

$$p(z_i = k) = \pi_k$$

Now

SCIEN	SCIENCE		TECH		SPO	RTS
experiment	0.1	develop	0.18		player	0.15
test	0.08	computer	0.09		score	0.07
discover	0.05	processor	0.032		team	0.06
hypothesize	0.03	user	0.027		goal	0.03
climate	0.01	internet	0.02		injury	0.01

{modeling, complex, epilepsy, modeling, Bayesian, clinical, epilepsy, EEG, data, dynamic...}

compute likelihood of the collection of words in doc under each topic distribution

©2016 Emily Fox & Carlos Guestrin

		Modeling the Complex Dynamics and Changing
Correlations of Epileptic Events		
_	_	Correlations of Epheptic Events
Same to	pic	Drausin F. Wulsin <sup>a</sup> , Emily B. Fox <sup>c</sup> , Brian Litt <sup>a,b</sup>
distribu	tion	C <sup>a</sup> Department of Biognaingering, University of Penneylyania, Philadelphia, PA
		<sup>b</sup> Department of Neurology, University of Pennsylvania, Philadelphia, PA
	05	<sup>c</sup> Department of Statistics, University of Washington, Seattle, WA
SCIEN	CE	
experiment	0.1	
test	0.08	Abstract
discover	0.05	Abstract
hypothesize	0.03	Faccus with epilepsy can manifest short, sub-clinical epileptic "bursts" in
climate	0.01	addition to full blowing clinical seizures. We believe the relationship between
		these two classes of events—something not previously studied quantitatively—
		solution and a solution of our work is to parse these grapher englantic events
TECH	4	into distinct dynamic regimes. A challenge posed by the intracranial EEG
develop	0.10	(The G) data we then is the fact that the number and placement of electrodes
develop	0.18	The an vary between matients We develop a Bayesian nonparametric Markov
computer	0.09	syntehing process that allows for (i) shared dynamic regimes between a vari-
processor	0.032	able number of channels, (ii) synchronous regime-switching, and (iii) an
internet	0.027	unknown dictionary of dynamic regimes. We encode a sparse and changing
internet	0.02	set of dependencies between the channels using a Markov-switching Gaussian
		graphical model for the innovations process driving the channel dynamics and
		demonstrate the importance of this poder to pushing and out-of-sample pre-
SPOR	TS	assignments that can kelp automate clinical are vsis of seizures and enable
player (	0.15	the comparison of sub-clinical bursts and full clinical seizures.
score (	0.07	Kilmuorie: Bayogian honparametric EEC factorial hidden Markow model
team (	0.06	raphical model time series
goal (	0.03	
injury (	0.01	
		1. Introduction
		Despite over three decades of research, we still have very little idea of
:		what defines beizure. This ignorance stems both from the complexity of
•		epilepsy as a disease and a paucity of quantitative tools that are flexible

### In LDA:

One topic indicator  $z_{iw}$  per **word** in doc i

**Each word** scored under topic z<sub>iw</sub>

Distribution on topics in **document**  $\mathbf{\pi}_{i} = [\mathbf{\pi}_{i1} \ \mathbf{\pi}_{i2} \dots \mathbf{\pi}_{iK}]$ 



## Topic vocab distributions:

TOPIC 1		
Word 1	?	
Word 2	?	
Word 3	?	
Word 4	?	
Word 5	?	

TOPIC 2		
Word 1	?	
Word 2	?	
Word 3	?	
Word 4	?	
Word 5	?	

TOPIC 3		
Word 1	?	
Word 2	?	
Word 3	?	
Word 4	?	
Word 5	?	

Correlations of Epileptic Events

Drausin F. Wulsin<sup>a</sup>, Emily B. Fox<sup>c</sup>, Brian Litt<sup>a,b</sup>

Modeling the Complex Dynamics and Changing

<sup>a</sup>Department of Bioengineering, University of Pennsylvania, Philadelphia, PA <sup>b</sup>Department of Neurology, University of Pennsylvania, Philadelphia, PA <sup>c</sup>Department of Statistics, University of Washington, Seattle, WA

### Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic "bursts" in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively could yield important insights into the nature and intrinsic dynamics of seizures. A goal of our work is to parse these complex epileptic events into distinct dynamic regimes. A challenge posed by the intracranial EEG (iEEG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable number of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures

*Keywords:* Bayesian nonparametric, EEG, factorial hidden Markov model, graphical model, time series

### 1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible

## Document topic proportions: $\boldsymbol{\pi}_{i} = [\boldsymbol{\pi}_{i1} \ \boldsymbol{\pi}_{i2} \ ... \ \boldsymbol{\pi}_{iK}]$



# Gibbs sampling for LDA

TOPIC 1		
experiment	0.1	
test	0.08	
discover	0.05	
hypothesize	0.03	
climate	0.01	

TOPIC 2		
develop	0.18	
computer	0.09	
processor	0.032	
user	0.027	
internet	0.02	

TOPIC 3		
player	0.15	
score	0.07	
team	0.06	
goal	0.03	
injury	0.01	

Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin<sup>a</sup>, Emily B. Fox<sup>c</sup>, Brian Litt<sup>a,b</sup>

<sup>a</sup>Department of Bioengineering, University of Pennsylvania, Philadelphia, PA <sup>b</sup>Department of Neurology, University of Pennsylvania, Philadelphia, PA <sup>c</sup>Department of Statistics, University of Washington, Seattle, WA

#### Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic "bursts" in addition to full-blown clinical seizures. We believe the relationship between these two classes of events-something not previously studied quantitativelycould yield important insights into the nature and intrinsic dynamics of seizures. A goal of our work is to parse these complex epileptic events into distinct dynamic regimes. A challenge posed by the intracranial EEG (iEEG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable number of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures

*Keywords:* Bayesian nonparametric, EEG, factorial hidden Markov model, graphical model, time series

#### 1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible



# **Step 1:** Randomly reassign all z<sub>iw</sub> based on

- doc topic proportions
- topic vocab distributions

Draw randomly from responsibility vector [r<sub>iw1</sub> r<sub>iw2</sub> ... r<sub>iwK</sub>]

# Gibbs sampling for LDA

TOPIC 1		
experiment	0.1	
test	0.08	
discover	0.05	
hypothesize	0.03	
climate	0.01	

TOPIC 2	
develop	0.18
computer	0.09
processor	0.032
user	0.027
internet	0.02

TOPIC 3	
player	0.15
score	0.07
team	0.06
goal	0.03
injury	0.01

Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin<sup>a</sup>, Emily B. Fox<sup>c</sup>, Brian Litt<sup>a,b</sup>

<sup>a</sup>Department of Bioengineering, University of Pennsylvania, Philadelphia, PA <sup>b</sup>Department of Neurology, University of Pennsylvania, Philadelphia, PA <sup>c</sup>Department of Statistics, University of Washington, Seattle, WA

#### Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic "bursts" in addition to full-blown clinical seizures. We believe the relationship between these two classes of events-something not previously studied quantitativelycould yield important insights into the nature and intrinsic dynamics of seizures. A goal of our work is to parse these complex epileptic events into distinct dynamic regimes. A challenge posed by the intracranial EEG (iEEG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable number of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures

*Keywords:* Bayesian nonparametric, EEG, factorial hidden Markov model, graphical model, time series

#### 1. Introduction





**Step 2:** Randomly reassign doc topic proportions based on assignments z<sub>iw</sub> in **current doc** 

**Step 3:** Repeat for all docs

# Gibbs sampling for LDA

<b>TOPIC 1</b>	
Word 1	?
Word 2	?
Word 3	?
Word 4	?
Word 5	?

<b>TOPIC 2</b>	
Word 1	?
Word 2	?
Word 3	?
Word 4	?
Word 5	?

TOPIC 3		
Word 1	?	
Word 2	?	
Word 3	?	
Word 4	?	
Word 5	?	

Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin<sup>a</sup>, Emily B. Fox<sup>c</sup>, Brian Litt<sup>a,b</sup>

<sup>a</sup>Department of Bioengineering, University of Pennsylvania, Philadelphia, PA <sup>b</sup>Department of Neurology, University of Pennsylvania, Philadelphia, PA <sup>c</sup>Department of Statistics, University of Washington, Seattle, WA

#### Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic "bursts" in addition to full-blown clinical seizures. We believe the relationship between these two classes of events-something not previously studied quantitativelycould yield important insights into the nature and intrinsic dynamics of seizures. A goal of our work is to parse these complex epileptic events into distinct dynamic regimes. A challenge posed by the intracranial EEG (iEEG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable number of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures

Keywords: Bayesian nonparametric, EEG, factorial hidden Markov model, graphical model, time series

1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible



**Step 4:** Randomly reassign topic vocab distributions based on assignments z<sub>iw</sub> in **entire corpus** 

## Collapsed Gibbs sampling for LDA

experiment	0.1
test	0.08
discover	0.05
hypothesize	0.03
climate	0.01

TOPIC 2	
develop	0.18
computer	0.09
processor	0.032
user	0.027
internet	0.02

player	0.15
score	0.07
team	0.06
goal	0.03
injury	0.01

Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin<sup>a</sup>, Emily B. Fox<sup>c</sup>, Brian Litt<sup>a,b</sup>

<sup>a</sup>Department of Bioengineering, University of Pennsylvania, Philadelphia, PA <sup>b</sup>Department of Neurology, University of Pennsylvania, Philadelphia, PA <sup>c</sup>Department of Statistics, University of Washington, Seattle, WA

#### Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic "bursts" in addition to full-blown clinical seizures. We believe the relationship between these two classes of events-something not previously studied quantitativelycould yield important insights into the nature and intrinsic dynamics of seizures. A goal of our work is to parse these complex meptic events into distinct dynamic regimes. A challenge posed or one intracranial EEG (iEEG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable number of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures

*Keywords:* Bayesian nonparametric, EEG, factorial hidden Markov model, graphical model, time series

#### 1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible



Randomly reassign z<sub>iw</sub> based on current assignments z<sub>jv</sub> of all other words **in doc and corpus** 

## **Collapsed conditional distribution**





# What to do with sampling output?

### **Predictions:**

- 1. Make prediction for each snapshot of randomly assigned variables/parameters (full iteration)
- 2. Average predictions for final result

### Parameter or assignment estimate:

Look at snapshot of randomly assigned variables/parameters that maximizes
"joint model probability"



## Summary of what we learned



## Bonus content: Hierarchical clustering

# Why hierarchical clustering?

- Avoid choosing # clusters beforehand
- **Dendrograms** help visualize different clustering granularities
  - No need to rerun algorithm



- Most algorithms allow user to choose any distance metric
  - k-means restricted us to Euclidean distance

# Why hierarchical clustering?

Can often find more complex shapes than k-means or Gaussian mixture models

> Gaussian mixtures: ellipsoids

k-means: spherical clusters





# Why hierarchical clustering?

Can often find more complex shapes than k-means or Gaussian mixture models

### What about these?



# Two main types of algorithms

**Divisive**, *a.k.a top-down:* Start with all data in one big cluster and recursively split.

- Example: recursive k-means

Agglomerative *a.k.a.* bottom-up: Start with each data point as its own cluster. Merge clusters until all points are in one big cluster.

– Example: single linkage

## **Divisive clustering**

## **Divisive** in pictures – level 1



## Divisive in pictures – level 2



## **Divisive:** Recursive k-means

Wikipedia



## **Divisive:** Recursive k-means



## **Divisive:** Recursive k-means



## **Divisive** choices to be made

- Which algorithm to recurse
- How many clusters per split
- When to split vs. stop
  - Max cluster size:

number of points in cluster falls below threshold

- Max cluster radius:

distance to furthest point falls below threshold

- Specified # clusters:

split until pre-specified # clusters is reached

### Agglomerative clustering

1. Initialize each point to be its own cluster



1. Initialize each point to be its own cluster



2. Define distance between clusters to be:



3. Merge the two closest clusters













## **Clusters of clusters**

Just like our picture for divisive clustering...


# The dendrogram for agglomerative clustering

# The dendrogram

- x axis shows data points (carefully ordered)
- y-axis shows distance between pair of clusters



# The dendrogram

- x axis shows data points (carefully ordered)
- y-axis shows distance between pair of clusters



## The dendrogram

Path shows all clusters to which a point belongs and the order in which clusters merge



©2016 Emily Fox & Carlos Guestrin

# Extracting a partition

Choose a distance D\* at which to cut dendogram



# Extracting a partition

Every branch that crosses D\* becomes a separate cluster



## Extracting a partition

Every branch that crosses D\* becomes a separate cluster



#### Agglomerative clustering details

### Agglomerative choices to be made

- Distance metric: d(x<sub>i</sub>, x<sub>j</sub>)
- Linkage function: e.g.,  $\min_{\substack{\mathbf{x}_i \text{ in } C_1, \\ \mathbf{x}_j \text{ in } C_2}} d(\mathbf{x}_i, \mathbf{x}_j)$
- Where and how to cut dendrogram



# More on cutting dendrogram

- For visualization, smaller # clusters is preferable
- For tasks like outlier detection, cut based on:
  - Distance threshold
  - Inconsistency coefficient
    - Compare height of merge to average merge heights below
    - If top merge is substantially higher, then it is joining two subsets that are relatively far apart compared to the members of each subset internally
    - Still have to choose a threshold to cut at, but now in terms of "inconsistency" rather than distance
- No cutting method is "incorrect", some are just more useful than others



# **Computational considerations**

- Computing all pairs of distances is expensive
  - Brute force algorithm is O(N<sup>2</sup>log(N))

- Smart implementations use triangle inequality to rule out candidate pairs
- Best known algorithm is O(N<sup>2</sup>)

# datapoints

## Statistical issues

**Chaining**: Distant points clustered together if there is a chain of pairwise close points between



Other linkage functions can be more robust, but restrict the shapes of clusters that can be found

- Complete linkage:

max pairwise distance between clusters

#### - Ward criterion:

min within-cluster variance at each merge

#### Hidden Markov models (HMMs): Another notion of "clustering"

## So far, looked at clustering unordered data



Data index (i.e., when observation was recorded) does not influence clustering

### What if we have time series data?



Would be hard to distinguish red, blue, and green clusters if we ignored order of data

#### Example: Honey bee dances





©2016 Emily Fox & Carlos Guestrin

#### Repeated patterns of dance transitions



#### Similar ideas appear in many applications





Bob John	J I I J	Jane	John
----------	------------------	------	------









©2016 Emily Fox & Carlos Guestrin

# Hidden Markov model (HMM)

#### As in mixture model...

Every observation  $x_t$  is associated with cluster assignment variable  $z_t$ 

Each cluster has a distribution over observed values



## Hidden Markov model (HMM)

#### Difference from mixture model:

Probability of  $(z_t = k)$  depends on previous cluster assignment  $z_{t-1}$ 



## Inference in HMMs

- Learn MLE of HMM parameters using EM algorithm = Baum Welch
- Infer MLE of state sequence given fixed model parameters using dynamic programming = Viterbi algorithm
- Infer soft assignments of state sequence using dynamic programming
  = forward-backward algorithm

#### What we didn't cover

# Other clustering + retrieval topics

#### Retrieval:

- Other distance metrics
- Distance metric learning

#### Clustering:

- Nonparametric clustering
- Spectral clustering

#### Related ideas:

- Density estimation
- Anomaly detection

# What's ahead in this specialization

### This course is a part of the Machine Learning Specialization



## 5. Recommender Systems & Dimensionality Reduction Case study: Recommending Products • Collaborative filtering

- Matrix factorization
- PCA

#### Algorithms

Concepts

Models

- Coordinate descent
- Eigen decomposition
- SVD

 Matrix completion, eigenvalues, cold-start problem, diversity, scaling up

©2016 Emily Fox & Carlos Guestrin

Rating= **R**<sup>′</sup>  $\sim$ Parameters of model Machine Learning Specialization

# 6. Capstone: Build and deploy an intelligent application with deep learning



©2016 Emily Fox & Carlos Guestrin

#### Thank you...