Collapsed Gibbs sampling in LDA

“Collapsed” Gibbs sampling for LDA

Based on special structure of LDA model, can sample **just** indicator variables z_{iw}

- No need to sample other parameters
  - corpus-wide topic vocab distributions
  - per-doc topic proportions

Often leads to much better performance because examining uncertainty in smaller space

Collapsed Gibbs sampling for LDA

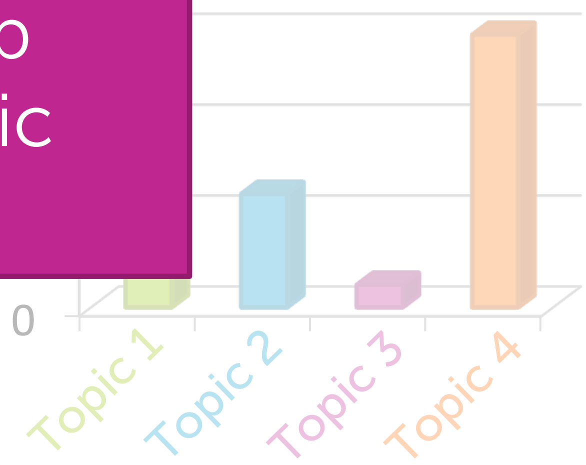
Never draw topic vocab distributions or doc topic proportions

TOPIC 1	
experiment	0.
test	0.
discover	0.
hypothesize	0.
climate	0.
…	…

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
…	…

⋮



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

Randomly reassign z_{iw} based on current assignments z_{jv} of all other words **in document and corpus**

Select a document

epilepsy	dynamic	Bayesian	EEG	model

5 word document

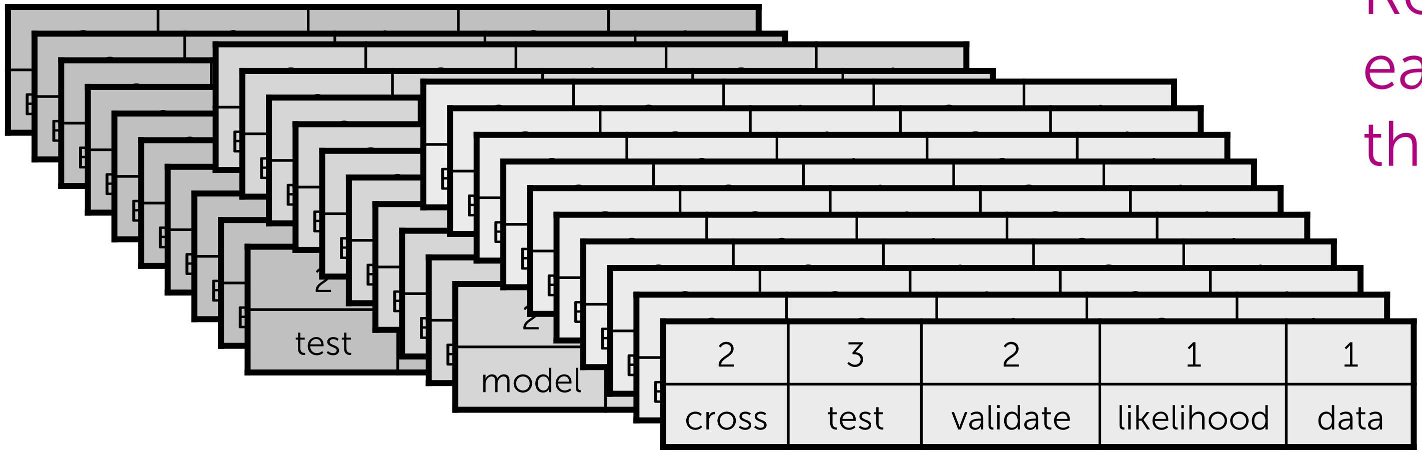
Randomly assign topics

3	2	1	3	1
epilepsy	dynamic	Bayesian	EEG	model

(one possible approach)

Randomly assign topics

3	2	1	3	1
epilepsy	dynamic	Bayesian	EEG	model



Repeat for each doc in the corpus

Maintain local statistics

3	2	1	3	1
epilepsy	dynamic	Bayesian	EEG	model

	Topic 1	Topic 2	Topic 3
Doc i	2	1	2

Maintain global statistics

3	2	1	3	1
epilepsy	dynamic	Bayesian	EEG	model

	Topic 1	Topic 2	Topic 3
Doc i	2	1	2

	Topic 1	Topic 2	Topic 3
epilepsy	1	0	35
Bayesian	50	0	1
model	42	1	0
EEG	0	0	20
dynamic	10	8	1
…			

Total counts from **all** docs

Randomly reassign topics

3	~~2~~	1	3	1
epilepsy	dynamic	Bayesian	EEG	model

	Topic 1	Topic 2	Topic 3
epilepsy	1	0	35
Bayesian	50	0	1
model	42	1	0
EEG	0	0	20
dynamic	10	7 ~~8~~	1
…			

	Topic 1	Topic 2	Topic 3
Doc i	2	0 ~~1~~	2

decrementing counts after removing current assignment $z_{iw}=2$

Probability of new assignment

3	?	1	3	1
epilepsy	dynamic	Bayesian	EEG	model

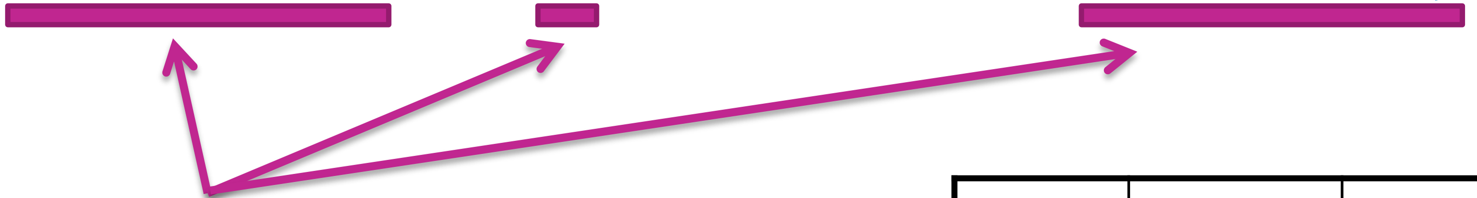
reassign with probability

$p(z_{iw} \mid$ every other z_{jv} in corpus, words in corpus$)$

Probability of new assignment

3	?	1	3	1
epilepsy	dynamic	Bayesian	EEG	model

Topic 1 Topic 2 Topic 3



How much doc "likes" each topic based on other assignments in doc

	Topic 1	Topic 2	Topic 3
Doc i	2	0	2

$$\frac{n_{ik} + \alpha}{N_i - 1 + K\alpha}$$

current assignments to topic k in doc i

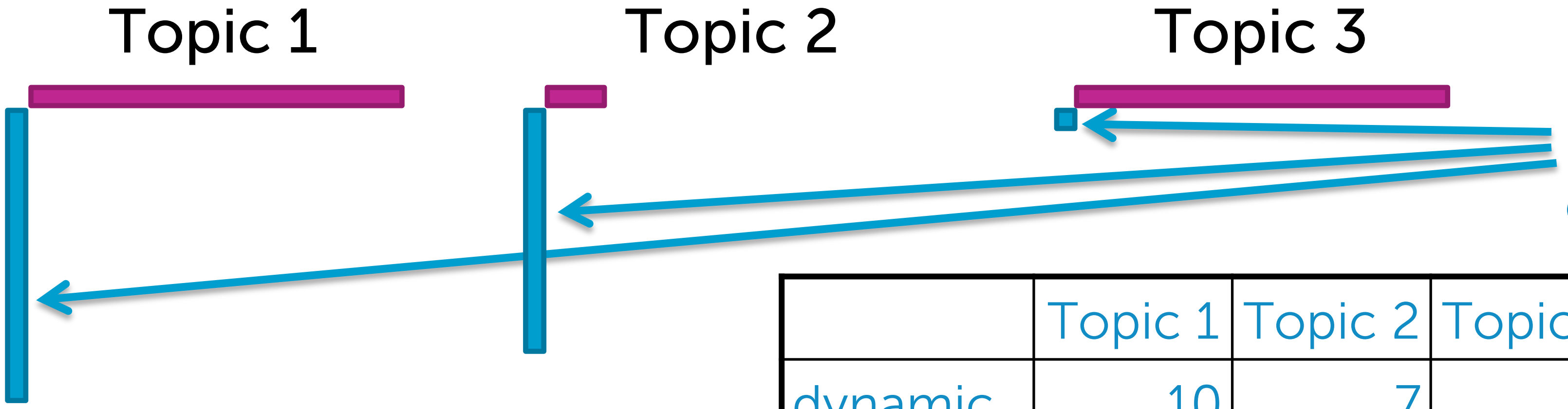
smoothing param from Bayes prior

words in doc i

ignore current word

Probability of new assignment

3	?	1	3	1
epilepsy	dynamic	Bayesian	EEG	model



How much each topic likes the word "dynamic" based on assignments in other docs in corpus

	Topic 1	Topic 2	Topic 3
dynamic	10	7	1

assignments **corpus-wide** of word "dynamic" to topic k

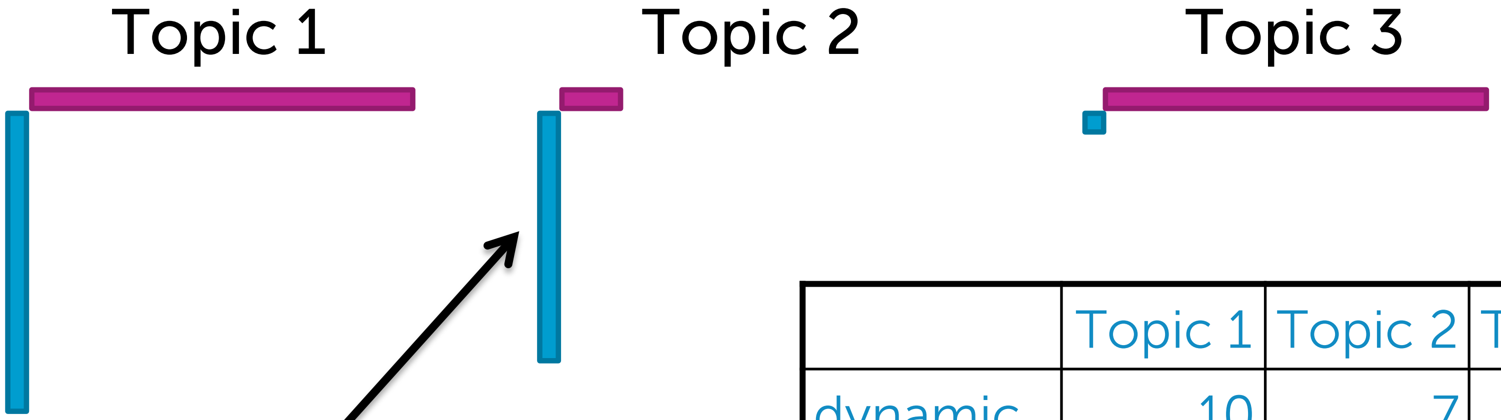
$$\frac{m_{\text{dynamic},k} + \gamma}{\sum_{w \in V} m_{w,k} + V\gamma}$$

smoothing param — from Bayes prior

size of vocab

Probability of new assignment

3	?	1	3	1
epilepsy	dynamic	Bayesian	EEG	model



	Topic 1	Topic 2	Topic 3
dynamic	10	7	1

Topic 2 also really likes "dynamic", but in a different context…
e.g., a topic on fluid dynamics

Probability of new assignment

3	?	1	3	1
epilepsy	dynamic	Bayesian	EEG	model

Topic 1 | Topic 2 | Topic 3

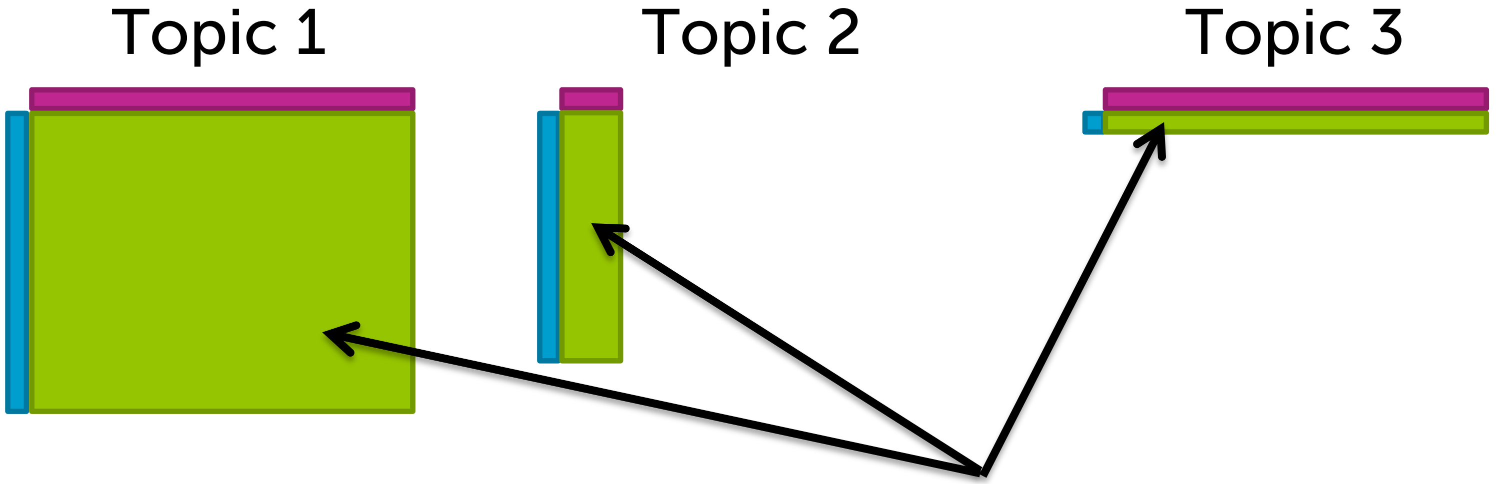
Topic fits word **and** document

Topic fits word, but not doc

Topic fits doc, but not word

Probability of new assignment

3	?	1	3	1
epilepsy	dynamic	Bayesian	EEG	model



How much doc likes topic

$$\frac{n_{ik} + \alpha}{N_i - 1 + K\alpha} \; \frac{m_{\text{dynamic},k} + \gamma}{\sum_{w \in V} m_{w,k} + V\gamma}$$

How much topic likes word

Randomly draw a new topic indicator

3	?	1	3	1
epilepsy	dynamic	Bayesian	EEG	model

Topic 1 **Topic 2** **Topic 3**

To draw new topic assignment (equivalently):

- roll K-sided die with these probabilities
- throw dart at these regions

Normalize this product of terms over K possible topics!

How much doc likes topic

$$\boxed{\frac{n_{ik}+\alpha}{N_i-1+K\alpha}\;\frac{m_{\text{dynamic},k}+\gamma}{\sum_{w\in V} m_{w,k}+V\gamma}}$$

How much topic likes word

Update counts

3	1	1	3	1
epilepsy	dynamic	Bayesian	EEG	model

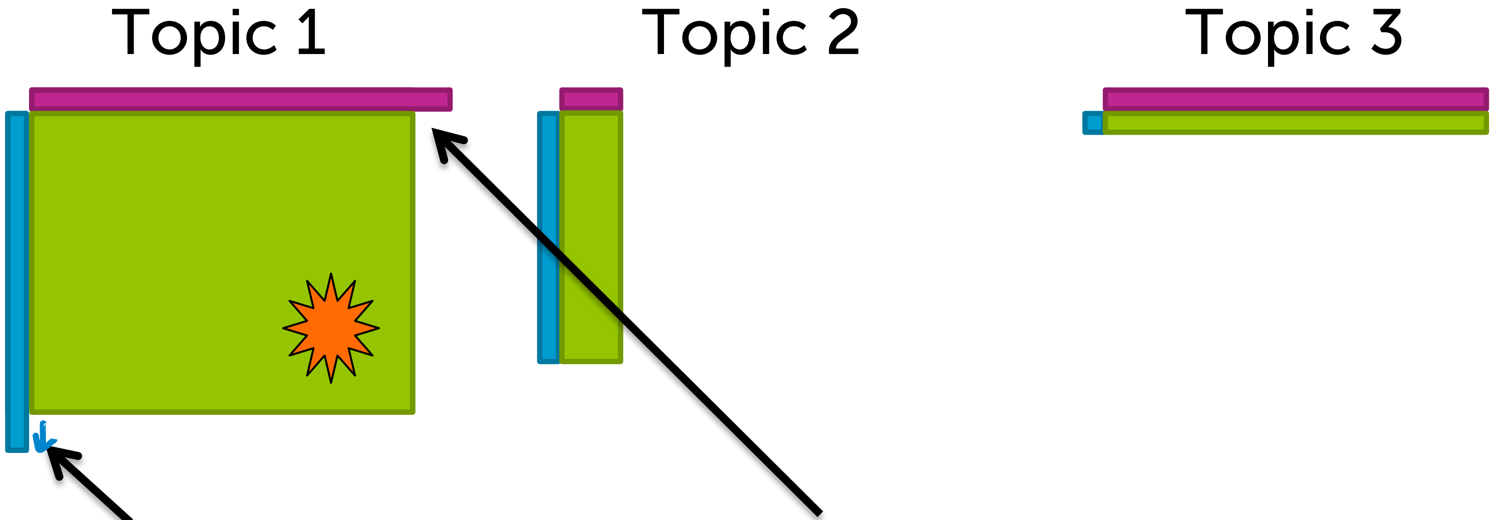
	Topic 1	Topic 2	Topic 3
Doc i	3 ~~2~~	0	2

	Topic 1	Topic 2	Topic 3
epilepsy	1	0	35
Bayesian	50	0	1
model	42	1	0
EEG	0	0	20
dynamic	11 ~~10~~	7	1
…			

increment counts based on new assignment of $z_{iw}=1$

Geometrically…

3	1	1	3	1
epilepsy	dynamic	Bayesian	EEG	model



Increase popularity of "dynamic" in topic 1 (**corpus-wide**)

Increase popularity of topic 1 **in doc i**

Iterate through all words/docs

3	1	1	3	1
epilepsy	dynamic	Bayesian	EEG	model

	Topic 1	Topic 2	Topic 3
Doc i	2	0	2

	Topic 1	Topic 2	Topic 3
epilepsy	1	0	35
Bayesian	50	0	1
model	42	1	0
EEG	0	0	20
dynamic	10	7	1
…			

Using samples from collapsed Gibbs

What to do with the collapsed samples?

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[a], Emily B. Fox[c], Brian Litt[a,b]

[a]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA
[b]Department of Neurology, University of Pennsylvania, Philadelphia, PA
[c]Department of Statistics, University of Washington, Seattle, WA

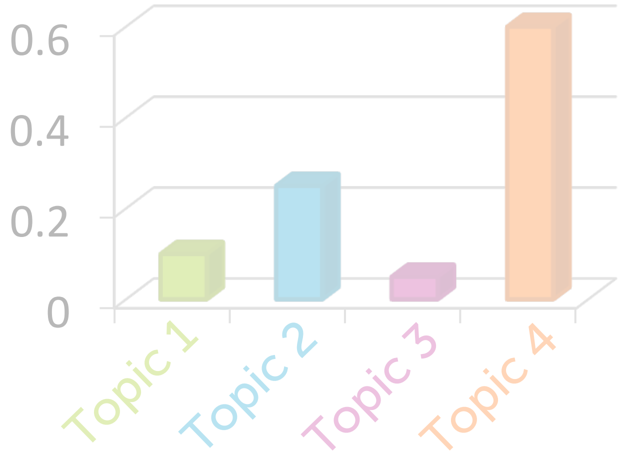
Abstract

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



From "best" sample of $\{z_{iw}\}$, can infer:

What to do with the collapsed samples?

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[a], Emily B. Fox[c], Brian Litt[a,b]

[a]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA
[b]Department of Neurology, University of Pennsylvania, Philadelphia, PA
[c]Department of Statistics, University of Washington, Seattle, WA

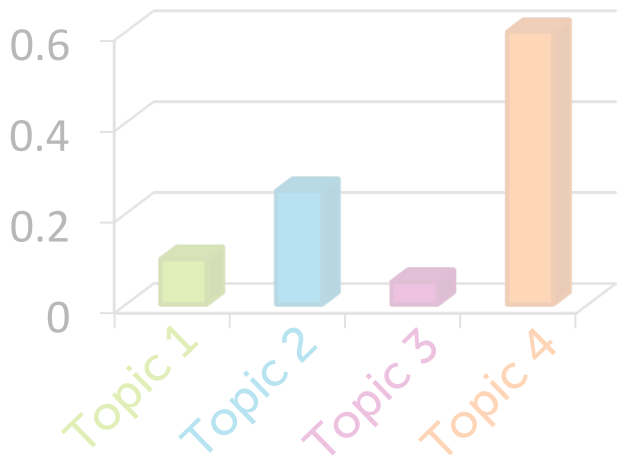
Abstract

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



From "best" sample of $\{z_{iw}\}$, can infer:

1. Topics from conditional distribution…
   need corpus-wide info

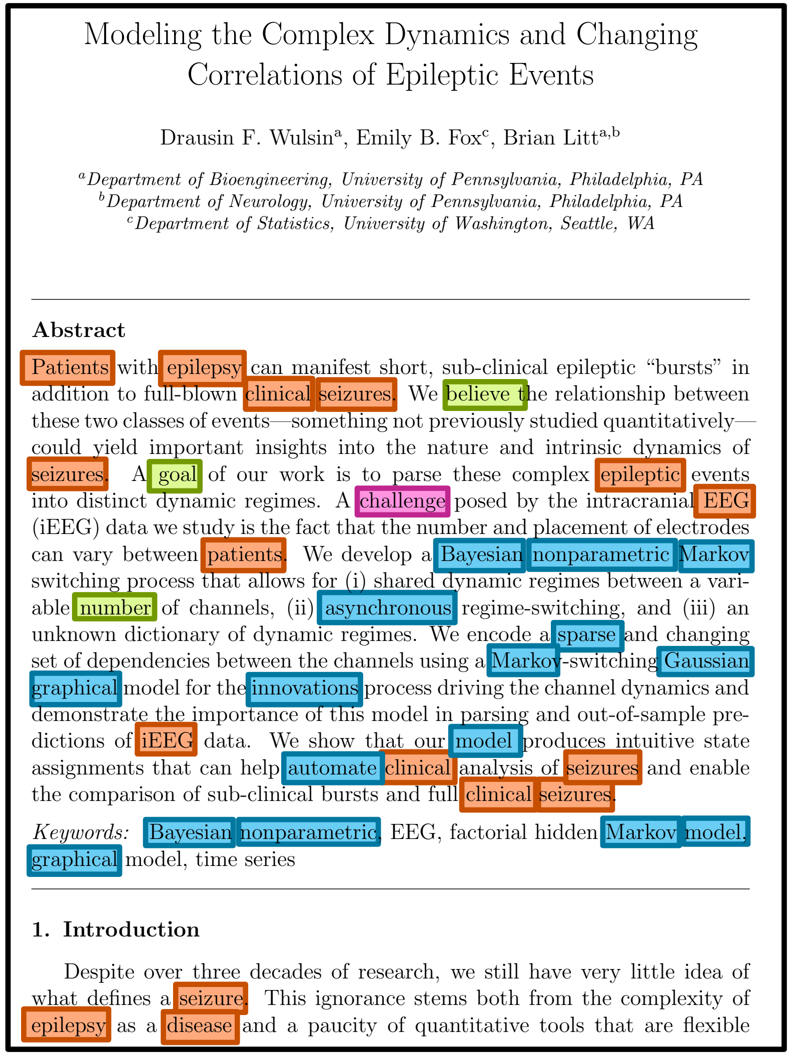
What to do with the collapsed samples?

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[a], Emily B. Fox[c], Brian Litt[a,b]

[a]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA
[b]Department of Neurology, University of Pennsylvania, Philadelphia, PA
[c]Department of Statistics, University of Washington, Seattle, WA

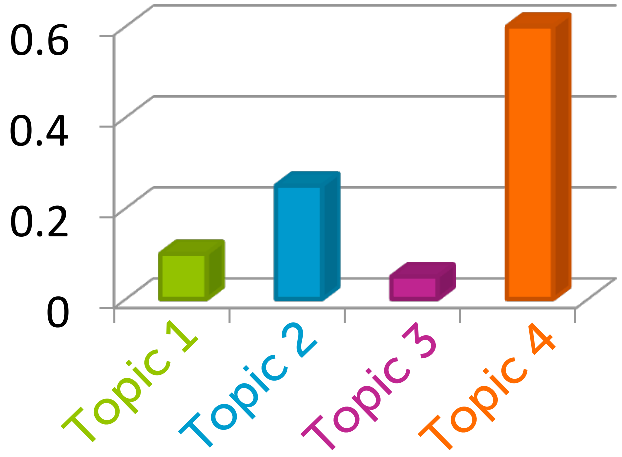
Abstract

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



From "best" sample of $\{z_{iw}\}$, can infer:

1. Topics from conditional distribution…
   need corpus-wide info
2. Document "embedding"…
   need doc info only

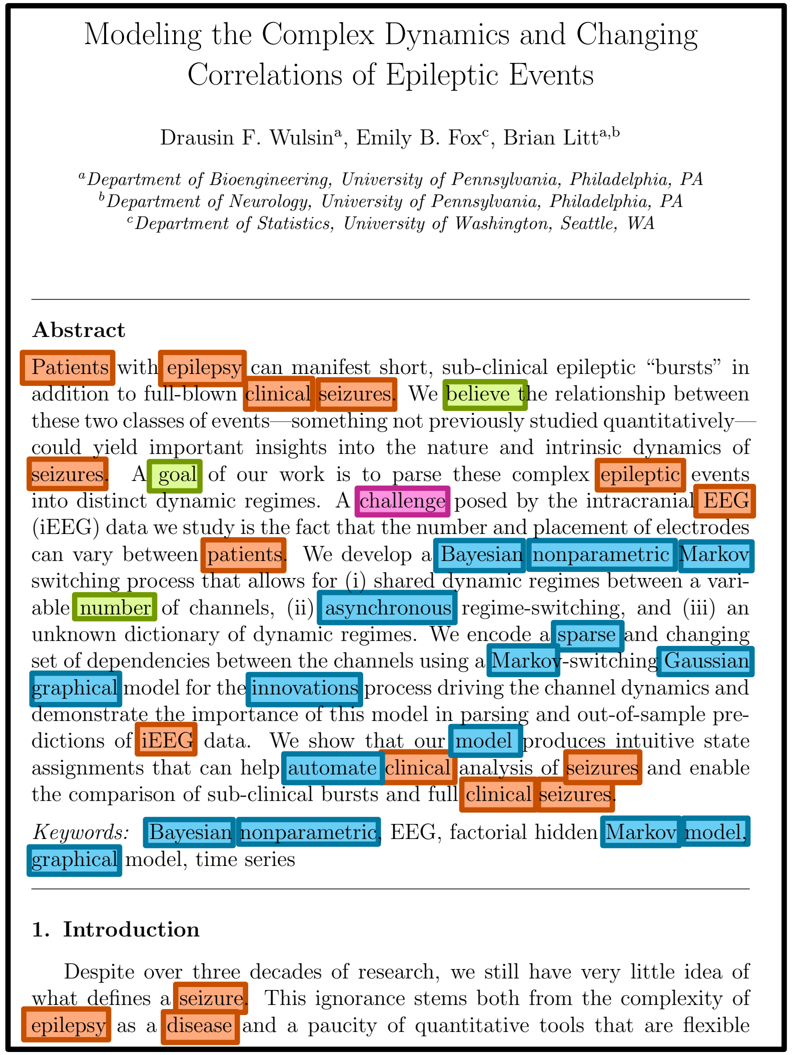
Embedding new documents

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[a], Emily B. Fox[c], Brian Litt[a,b]

[a]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA
[b]Department of Neurology, University of Pennsylvania, Philadelphia, PA
[c]Department of Statistics, University of Washington, Seattle, WA

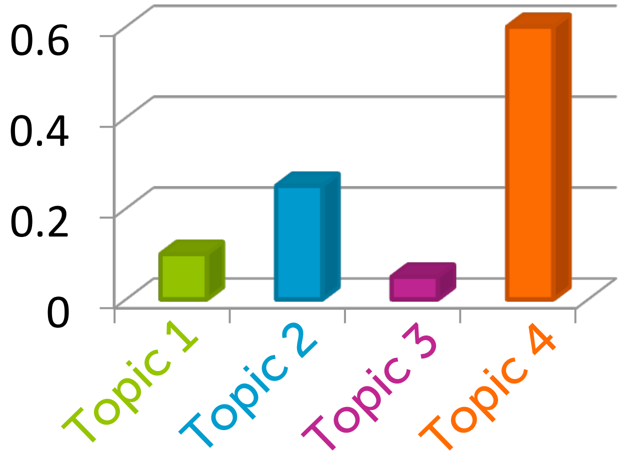
Abstract

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



Simple approach:

1. Fix topics based on training set collapsed sampling
2. Run uncollapsed sampler on new doc(s) only