LDA and Beyond: Topic Models in the Social Sciences

Brandon Stewart[1]

Princeton University

June 21, 2017

[1] My sincere thanks to my many collaborators and particularly Justin Grimmer, Molly Roberts and Dustin Tingley from whom many of these slides are derived.

Papers

- Overview of Text Analysis:
  - "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts" (*Political Analysis*, 2013) with Grimmer
- Structural Topic Model:
  - Structural Topic Models for Open-Ended Survey Responses (*American Journal of Political Science, 2014*) with Roberts, Tingley et al.
  - Computer Assisted Text Analysis for Comparative Politics (*Political Analysis 2015*), with Lucas et al.
  - A Model of Text for Experimentation in the Social Sciences (*2016*) with Roberts and Airoldi

Copies at BrandonStewart.org

Big Data, Big Analytics

Big Data, Big Analytics

- Massive increase in unstructured text due to:

Big Data, Big Analytics

- Massive increase in unstructured text due to:
  - new social structures (the internet, email)

Big Data, Big Analytics

- Massive increase in unstructured text due to:
  - new social structures (the internet, email)
  - new/improved data collection (wiki surveys, survey experiments)

Big Data, Big Analytics

- Massive increase in unstructured text due to:
  - new social structures (the internet, email)
  - new/improved data collection (wiki surveys, survey experiments)
  - digitization efforts (government documents, Google Books)

Big Data, Big Analytics

- Massive increase in unstructured text due to:
  - new social structures (the internet, email)
  - new/improved data collection (wiki surveys, survey experiments)
  - digitization efforts (government documents, Google Books)
- Communities leave digitized footprints

Big Data, Big Analytics

- Massive increase in unstructured text due to:
  - new social structures (the internet, email)
  - new/improved data collection (wiki surveys, survey experiments)
  - digitization efforts (government documents, Google Books)
- Communities leave digitized footprints
- Tools to analyze text advancing in parallel

Big Data, Big Analytics

- Massive increase in unstructured text due to:
  - new social structures (the internet, email)
  - new/improved data collection (wiki surveys, survey experiments)
  - digitization efforts (government documents, Google Books)
- Communities leave digitized footprints
- Tools to analyze text advancing in parallel
  - text by itself is useless

Big Data, Big Analytics

- Massive increase in unstructured text due to:
  - new social structures (the internet, email)
  - new/improved data collection (wiki surveys, survey experiments)
  - digitization efforts (government documents, Google Books)
- Communities leave digitized footprints
- Tools to analyze text advancing in parallel
  - text by itself is useless
  - importing methods from many different fields

Big Data, Big Analytics

- Massive increase in unstructured text due to:
  - new social structures (the internet, email)
  - new/improved data collection (wiki surveys, survey experiments)
  - digitization efforts (government documents, Google Books)
- Communities leave digitized footprints
- Tools to analyze text advancing in parallel
  - text by itself is useless
  - importing methods from many different fields
  - new analysis techniques can even drive new data availability

Different Methods for Different Goals

Different Methods for Different Goals

- Supervised: pursuing a known goal

Different Methods for Different Goals

- Supervised: pursuing a known goal
  - human annotates a subset of documents

Different Methods for Different Goals

- Supervised: pursuing a known goal
  - human annotates a subset of documents
  - algorithm annotates the rest

Different Methods for Different Goals

- Supervised: pursuing a known goal
  - human annotates a subset of documents
  - algorithm annotates the rest
  - usually associated with quantitative research

Different Methods for Different Goals

- Supervised: pursuing a known goal
  - human annotates a subset of documents
  - algorithm annotates the rest
  - usually associated with quantitative research
- Unsupervised: goal is to learn the goal

Different Methods for Different Goals

- Supervised: pursuing a known goal
  - human annotates a subset of documents
  - algorithm annotates the rest
  - usually associated with quantitative research
- Unsupervised: goal is to learn the goal
  - algorithm discovers themes/patterns in the texts

Different Methods for Different Goals

- Supervised: pursuing a known goal
  - human annotates a subset of documents
  - algorithm annotates the rest
  - usually associated with quantitative research
- Unsupervised: goal is to learn the goal
  - algorithm discovers themes/patterns in the texts
  - human interprets the results

Different Methods for Different Goals

- Supervised: pursuing a known goal
  - human annotates a subset of documents
  - algorithm annotates the rest
  - usually associated with quantitative research
- Unsupervised: goal is to learn the goal
  - algorithm discovers themes/patterns in the texts
  - human interprets the results
  - usually associated with qualitative research

Different Methods for Different Goals

- Supervised: pursuing a known goal
  - human annotates a subset of documents
  - algorithm annotates the rest
  - usually associated with quantitative research
- Unsupervised: goal is to learn the goal
  - algorithm discovers themes/patterns in the texts
  - human interprets the results
  - usually associated with qualitative research
- Both strategies amplify human effort, each in different ways.

Different Methods for Different Goals

- Supervised: pursuing a known goal
  - human annotates a subset of documents
  - algorithm annotates the rest
  - usually associated with quantitative research
- Unsupervised: goal is to learn the goal
  - algorithm discovers themes/patterns in the texts
  - human interprets the results
  - usually associated with qualitative research
- Both strategies amplify human effort, each in different ways.

Topic Models in Social Science

Topic Models in Social Science

- Core methods developed in computer science and statistics

Topic Models in Social Science

- Core methods developed in computer science and statistics
  - used as a way to summarize unstructured text

Topic Models in Social Science

- Core methods developed in computer science and statistics
  - used as a way to summarize unstructured text
  - use words within document to infer its subject

Topic Models in Social Science

- Core methods developed in computer science and statistics
  - used as a way to summarize unstructured text
  - use words within document to infer its subject
  - introduced as a form of dimension reduction

Topic Models in Social Science

- Core methods developed in computer science and statistics
  - used as a way to summarize unstructured text
  - use words within document to infer its subject
  - introduced as a form of dimension reduction
- A theory of use in the social sciences

Topic Models in Social Science

- Core methods developed in computer science and statistics
  - used as a way to summarize unstructured text
  - use words within document to infer its subject
  - introduced as a form of dimension reduction
- A theory of use in the social sciences
  - social scientists want to use topics as a form of measurement

Topic Models in Social Science

- Core methods developed in computer science and statistics
  - used as a way to summarize unstructured text
  - use words within document to infer its subject
  - introduced as a form of dimension reduction
- A theory of use in the social sciences
  - social scientists want to use topics as a form of measurement
  - we are often interest in how observed covariates drive trends in language

Topic Models in Social Science

- Core methods developed in computer science and statistics
  - used as a way to summarize unstructured text
  - use words within document to infer its subject
  - introduced as a form of dimension reduction
- A theory of use in the social sciences
  - social scientists want to use topics as a form of measurement
  - we are often interest in how observed covariates drive trends in language
  - we want to tell a story not just about what, but how and why

Topic Models in Social Science

- Core methods developed in computer science and statistics
  - used as a way to summarize unstructured text
  - use words within document to infer its subject
  - introduced as a form of dimension reduction
- A theory of use in the social sciences
  - social scientists want to use topics as a form of measurement
  - we are often interest in how observed covariates drive trends in language
  - we want to tell a story not just about what, but how and why
- Different focus in social science brings different concerns and an emphasis on validation

Topic Models in Social Science

- Core methods developed in computer science and statistics
  - used as a way to summarize unstructured text
  - use words within document to infer its subject
  - introduced as a form of dimension reduction
- A theory of use in the social sciences
  - social scientists want to use topics as a form of measurement
  - we are often interest in how observed covariates drive trends in language
  - we want to tell a story not just about what, but how and why
- Different focus in social science brings different concerns and an emphasis on validation

Some Example Questions

Some Example Questions

- How do senators present their work to the public? What explains variation in representational style? (Grimmer 2013)

Some Example Questions

- How do senators present their work to the public? What explains variation in representational style? (Grimmer 2013)
- Did electoral reform change the portfolio of issues addressed by politicians in Japan? (Catalinac 2016)

Some Example Questions

- How do senators present their work to the public? What explains variation in representational style? (Grimmer 2013)
- Did electoral reform change the portfolio of issues addressed by politicians in Japan? (Catalinac 2016)
- What are the propaganda strategies of the Chinese government? (Roberts and Stewart)

Some Example Questions

- How do senators present their work to the public? What explains variation in representational style? (Grimmer 2013)
- Did electoral reform change the portfolio of issues addressed by politicians in Japan? (Catalinac 2016)
- What are the propaganda strategies of the Chinese government? (Roberts and Stewart)
- What types of actions do countries take towards each other? (O'Connor, Stewart and Smith 2013)

Some Example Questions

- How do senators present their work to the public? What explains variation in representational style? (Grimmer 2013)
- Did electoral reform change the portfolio of issues addressed by politicians in Japan? (Catalinac 2016)
- What are the propaganda strategies of the Chinese government? (Roberts and Stewart)
- What types of actions do countries take towards each other? (O'Connor, Stewart and Smith 2013)
- How do Muslim clerics supporting violent Jihad differ from those who do not in terms of writing in fatwas? (Nielsen, Forthcoming)

Some Example Questions

- How do senators present their work to the public? What explains variation in representational style? (Grimmer 2013)
- Did electoral reform change the portfolio of issues addressed by politicians in Japan? (Catalinac 2016)
- What are the propaganda strategies of the Chinese government? (Roberts and Stewart)
- What types of actions do countries take towards each other? (O'Connor, Stewart and Smith 2013)
- How do Muslim clerics supporting violent Jihad differ from those who do not in terms of writing in fatwas? (Nielsen, Forthcoming)
- Do presidential candidates move to the center after the convention? (Gross et al 2013)

Some Example Questions

- How do senators present their work to the public? What explains variation in representational style? (Grimmer 2013)
- Did electoral reform change the portfolio of issues addressed by politicians in Japan? (Catalinac 2016)
- What are the propaganda strategies of the Chinese government? (Roberts and Stewart)
- What types of actions do countries take towards each other? (O'Connor, Stewart and Smith 2013)
- How do Muslim clerics supporting violent Jihad differ from those who do not in terms of writing in fatwas? (Nielsen, Forthcoming)
- Do presidential candidates move to the center after the convention? (Gross et al 2013)
- Do the proposals of NPC deputies in China represent the interests of their constituents? (Truex 2016)

Four Principles of Automated Text Analysis (Grimmer and Stewart)

Principle 1: All Quantitative Models of Language are Wrong—But Some are Useful

Four Principles of Automated Text Analysis (Grimmer and Stewart)

Principle 1: All Quantitative Models of Language are Wrong—But Some are Useful

- Data generation process for text $\rightsquigarrow$ unknown

Four Principles of Automated Text Analysis (Grimmer and Stewart)

Principle 1: All Quantitative Models of Language are Wrong—But Some are Useful

- Data generation process for text $\rightsquigarrow$ unknown
- Complexity of language:

Four Principles of Automated Text Analysis (Grimmer and Stewart)

Principle 1: All Quantitative Models of Language are Wrong—But Some are Useful

- Data generation process for text $\rightsquigarrow$ unknown
- Complexity of language:
    - Time flies like an arrow

Four Principles of Automated Text Analysis (Grimmer and Stewart)

Principle 1: All Quantitative Models of Language are Wrong—But Some are Useful

- Data generation process for text $\rightsquigarrow$ unknown
- Complexity of language:
  - Time flies like an arrow, fruit flies like a banana

Four Principles of Automated Text Analysis (Grimmer and Stewart)

Principle 1: All Quantitative Models of Language are Wrong—But Some are Useful

- Data generation process for text $\rightsquigarrow$ unknown
- Complexity of language:
  - Time flies like an arrow, fruit flies like a banana
  - Make peace, not war

Four Principles of Automated Text Analysis (Grimmer and Stewart)

Principle 1: All Quantitative Models of Language are Wrong—But Some are Useful

- Data generation process for text $\rightsquigarrow$ unknown
- Complexity of language:
  - Time flies like an arrow, fruit flies like a banana
  - Make peace, not war , Make war not peace

Four Principles of Automated Text Analysis (Grimmer and Stewart)

Principle 1: All Quantitative Models of Language are Wrong—But Some are Useful

- Data generation process for text $\rightsquigarrow$ unknown
- Complexity of language:
  - Time flies like an arrow, fruit flies like a banana
  - Make peace, not war , Make war not peace
  - "Years from now, you'll look back and you'll say that this was the moment, this was the place where America remembered what it means to hope. "

Four Principles of Automated Text Analysis (Grimmer and Stewart)

Principle 1: All Quantitative Models of Language are Wrong—But Some are Useful

- Data generation process for text $\rightsquigarrow$ unknown
- Complexity of language:
  - Time flies like an arrow, fruit flies like a banana
  - Make peace, not war , Make war not peace
  - "Years from now, you'll look back and you'll say that this was the moment, this was the place where America remembered what it means to hope. "
- Models necessarily fail to capture language $\rightsquigarrow$ useful for specific tasks

Four Principles of Automated Text Analysis (Grimmer and Stewart)

Principle 1: All Quantitative Models of Language are Wrong—But Some are Useful

- Data generation process for text $\rightsquigarrow$ unknown
- Complexity of language:
  - Time flies like an arrow, fruit flies like a banana
  - Make peace, not war , Make war not peace
  - "Years from now, you'll look back and you'll say that this was the moment, this was the place where America remembered what it means to hope. "
- Models necessarily fail to capture language $\rightsquigarrow$ useful for specific tasks
- Validation $\rightsquigarrow$ demonstrate methods perform task

Four Principles of Automated Text Analysis (Grimmer and Stewart)

Principle 2: Quantitative Methods Augment Humans, Not Replace Them

Four Principles of Automated Text Analysis (Grimmer and Stewart)

Principle 2: Quantitative Methods Augment Humans, Not Replace Them

- Computer-Assisted Reading

Four Principles of Automated Text Analysis (Grimmer and Stewart)

Principle 2: Quantitative Methods Augment Humans, Not Replace Them

- Computer-Assisted Reading
- Quantitative methods organize, direct, and suggest

Four Principles of Automated Text Analysis (Grimmer and Stewart)

Principle 2: Quantitative Methods Augment Humans, Not Replace Them

- Computer-Assisted Reading
- Quantitative methods organize, direct, and suggest
- Humans: read and interpret

Four Principles of Automated Text Analysis (Grimmer and Stewart)

Principle 3: There is no Globally Best Method for Automated Text Analysis

Four Principles of Automated Text Analysis (Grimmer and Stewart)

Principle 3: There is no Globally Best Method for Automated Text Analysis

- Supervised methods$\rightsquigarrow$ known categories

Four Principles of Automated Text Analysis (Grimmer and Stewart)

Principle 3: There is no Globally Best Method for Automated Text Analysis

- Supervised methods $\rightsquigarrow$ known categories
- Unsupervised methods $\rightsquigarrow$ discover categories

Four Principles of Automated Text Analysis (Grimmer and Stewart)

Principle 3: There is no Globally Best Method for Automated Text Analysis

- Supervised methods $\rightsquigarrow$ known categories
- Unsupervised methods $\rightsquigarrow$ discover categories
- Debate $\rightsquigarrow$ acknowledge differences, resolved

Four Principles of Automated Text Analysis (Grimmer and Stewart)

Principle 4: Validate, Validate, Validate

Four Principles of Automated Text Analysis (Grimmer and Stewart)

Principle 4: Validate, Validate, Validate

- Quantitative methods $\rightsquigarrow$ variable performance across tasks

Four Principles of Automated Text Analysis (Grimmer and Stewart)

Principle 4: Validate, Validate, Validate

- Quantitative methods $\rightsquigarrow$ variable performance across tasks
- Few theorems to guarantee performance

Four Principles of Automated Text Analysis (Grimmer and Stewart)

Principle 4: Validate, Validate, Validate

- Quantitative methods $\rightsquigarrow$ variable performance across tasks
- Few theorems to guarantee performance
- Apply methods $\rightsquigarrow$ validate

Four Principles of Automated Text Analysis (Grimmer and Stewart)

Principle 4: Validate, Validate, Validate

- Quantitative methods $\rightsquigarrow$ variable performance across tasks
- Few theorems to guarantee performance
- Apply methods $\rightsquigarrow$ validate
- Avoid: blind application of methods

Bag of Words

Bag of Words

- Throughout we will use a representation called bag of words because we will discard word order.

Bag of Words

- Throughout we will use a representation called bag of words because we will discard word order.
- Generally instantiated (at least conceptually) as a document-term matrix.

Bag of Words

- Throughout we will use a representation called bag of words because we will discard word order.
- Generally instantiated (at least conceptually) as a document-term matrix.
- This representation is good at capturing subject matter of documents but not nuance.

Bag of Words

- Throughout we will use a representation called bag of words because we will discard word order.
- Generally instantiated (at least conceptually) as a document-term matrix.
- This representation is good at capturing subject matter of documents but not nuance.
- We will breeze through this but these choices are consequential (see for example Denny and Spirling 2017, Schofield and Mimno 2016)

Preprocessing

"Political power grows out of the barrel of a gun" - Mao

Preprocessing

"Political power grows out of the barrel of a gun" - Mao

Compound Words: With substantive justification, words can be combined or split to improve inference.

Preprocessing

"Political power grows out of the barrel of a gun" - Mao

Compound Words: An analyst may want to combine words into a single term that can be analyzed.

Preprocessing

"Political power grows out of the barrel of a gun" - Mao

Compound Words: An analyst may want to combine words into a single term that can be analyzed.

Preprocessing

[Political], [power], [grows], [out], [of], [the], [barrel of a gun]

Compound Words: An analyst may want to combine words into a single term that can be analyzed.

Preprocessing

[Political], [power], [grows], [out], [of], [the], [barrel of a gun]

Stopword Removal: Removing terms that are not related to what the author is studying from the text.

Preprocessing

[Political], [power], [grows], [out], [of], [the], [barrel of a gun]

Stopword Removal: Removing terms that are not related to what the author is studying from the text.

Preprocessing

[Political], [power], [grows], [out], [barrel of a gun]

Stopword Removal: Removing terms that are not related to what the author is studying from the text.

Preprocessing

[Political], [power], [grows], [out], [barrel of a gun]

Stemming: Takes the ends off conjugated verbs or plural nouns, leaving just the "stem."

Preprocessing

[Political], [power], [grows], [out], [barrel of a gun]

Stemming: Takes the ends off conjugated verbs or plural nouns, leaving just the "stem."

Preprocessing

[Polit], [power], [grow], [out], [barrel of a gun]

Stemming: Takes the ends off conjugated verbs or plural nouns, leaving just the "stem."

Preprocessing

Finally, we can turn tokens and documents into a "document-term matrix."

Imagine we have a second document in addition to the Mao quote, which tokenizes as follows.

Document #1: [polit], [power], [grow], [out], [barrel of a gun]
Document #2: [wessi], [compar], [polit], [wessi]

Output: Term-Document Matrix

$$
\begin{array}{c|cc}
 & Doc1 & Doc2 \\
\hline
polit & 1 & 1 \\
power & 1 & 0 \\
grow & 1 & 0 \\
out & 1 & 0 \\
barrel\_of\_a\_gun & 1 & 0 \\
wessi & 0 & 2 \\
compar & 0 & 1
\end{array}
$$

Latent Dirichlet Allocation

- Idea: documents exhibit each topic in some proportion. This is an admixture.

Latent Dirichlet Allocation

- Idea: documents exhibit each topic in some proportion. This is an admixture.
- Each document is a mixture over topics. Each topic is a mixture over words.

Latent Dirichlet Allocation

- Idea: documents exhibit each topic in some proportion. This is an admixture.
- Each document is a mixture over topics. Each topic is a mixture over words.
- Latent Dirichlet Allocation estimates:

Latent Dirichlet Allocation

- Idea: documents exhibit each topic in some proportion. This is an admixture.
- Each document is a mixture over topics. Each topic is a mixture over words.
- Latent Dirichlet Allocation estimates:
  - The distribution over words for each topic.

Latent Dirichlet Allocation

- Idea: documents exhibit each topic in some proportion. This is an admixture.
- Each document is a mixture over topics. Each topic is a mixture over words.

- Latent Dirichlet Allocation estimates:
  - The distribution over words for each topic.
  - The proportion of a document in each topic, for each document.

Latent Dirichlet Allocation

- Idea: documents exhibit each topic in some proportion. This is an admixture.
- Each document is a mixture over topics. Each topic is a mixture over words.
- Latent Dirichlet Allocation estimates:
  - The distribution over words for each topic.
  - The proportion of a document in each topic, for each document.

Latent Dirichlet Allocation

- Idea: documents exhibit each topic in some proportion. This is an admixture.
- Each document is a mixture over topics. Each topic is a mixture over words.
- Latent Dirichlet Allocation estimates:
  - The distribution over words for each topic.
  - The proportion of a document in each topic, for each document.

Maintained assumptions: Bag of words/fix number of topics ex ante.

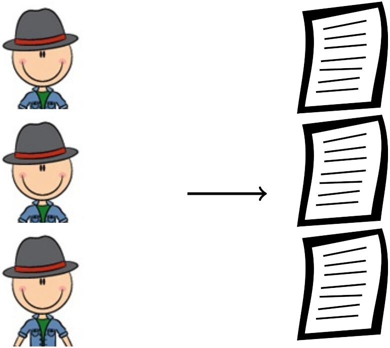
What this means in pictures

Say you have
a lot of people.

What this means in pictures

Say you have
a lot of people.

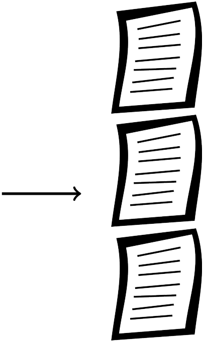
Each writes
some texts



What this means in pictures

Say you have
a lot of people.

Each writes
some texts



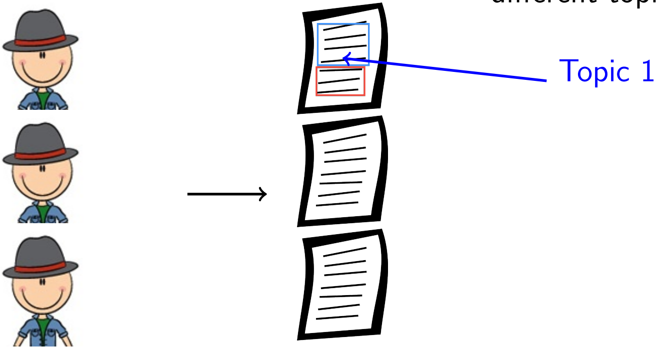
that discuss a few
different topics

What this means in pictures

Say you have
a lot of people.

Each writes
some texts

that discuss a few
different topics

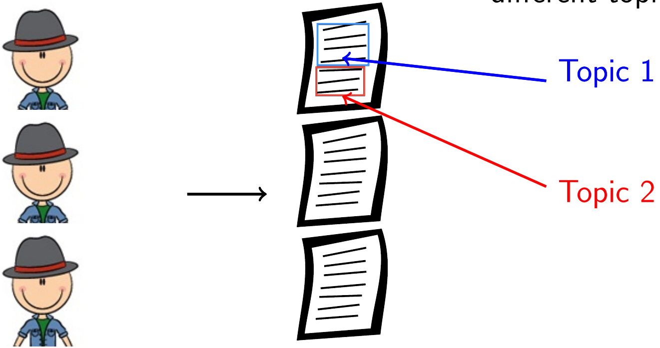


What this means in pictures

Say you have
a lot of people.

Each writes
some texts

that discuss a few
different topics

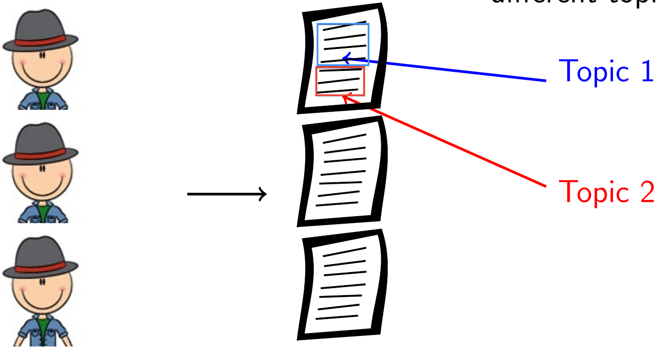


What this means in pictures

Say you have
a lot of people.

Each writes
some texts

that discuss a few
different topics



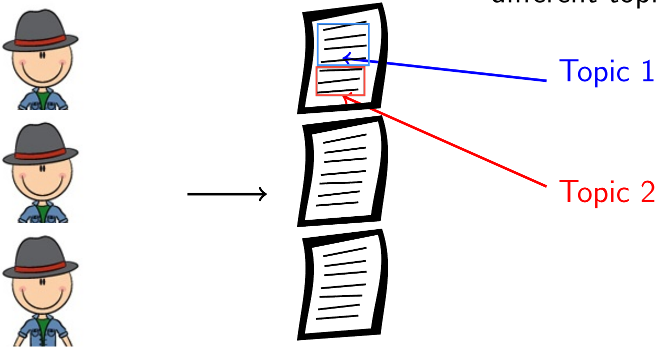
The Latent Dirichlet Allocation estimates:

What this means in pictures

Say you have
a lot of people.

Each writes
some texts

that discuss a few
different topics



The Latent Dirichlet Allocation estimates:

1. The topics- each is a distribution over words

What this means in pictures

Say you have
a lot of people.

Each writes
some texts

that discuss a few
different topics

Politics

congress, nations, power, votes, agreement, bargaining

Statistics

estimator, data, analysis, variance, model, inference

The Latent Dirichlet Allocation estimates:

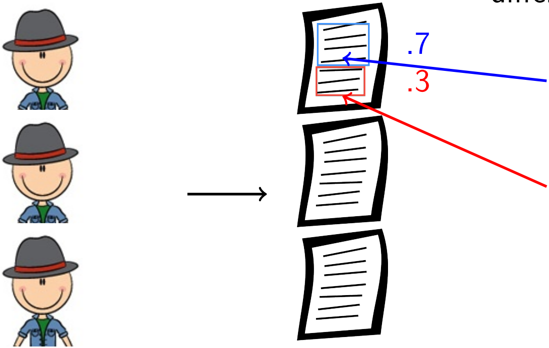
1. The topics- each is a distribution over words

What this means in pictures

Say you have a lot of people.

Each writes some texts

that discuss a few different topics



Politics

congress, nations, power, votes, agreement, bargaining

Statistics

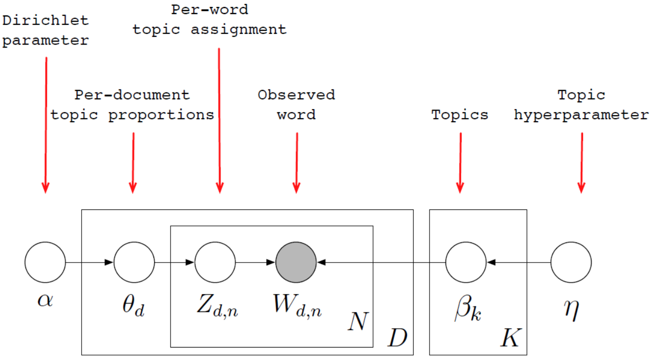
estimator, data, analysis, variance, model, inference

The Latent Dirichlet Allocation estimates:

1. The topics- each is a distribution over words
2. The proportion of each document in each topic

This is a Bayesian Model

Figure: Plate Notation of Latent Dirichlet Allocation



Graphic from David Blei's Website

LDA as a Bayesian Model

$$
\begin{aligned}
\boldsymbol{\beta}_k|\boldsymbol{\eta} &\sim \text{Dirichlet}(\boldsymbol{\eta}) \\
\boldsymbol{\theta}_i|\boldsymbol{\alpha} &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
\boldsymbol{z}_{im}|\boldsymbol{\theta}_i &\sim \text{Multinomial}(1, \boldsymbol{\theta}_i) \\
w_{im}|\boldsymbol{\beta}_k, \boldsymbol{z}_{imk} = 1 &\sim \text{Multinomial}(1, \boldsymbol{\beta}_k)
\end{aligned}
$$

LDA as a Bayesian Model

$$\begin{aligned}
\textbf{Unigram Model}_k &\sim \text{Dirichlet}(\eta) \\
\textbf{Doc. Prop}_i &\sim \text{Dirichlet}(\textbf{Pop. Proportion}) \\
\textbf{Word Topic}_{im} &\sim \text{Multinomial}(1, \textbf{Doc. Prop}_i) \\
\text{Word}_{im} &\sim \text{Multinomial}(1, \textbf{Unigram Model}_k)
\end{aligned}$$

“Vanilla” Latent Dirichlet Allocation

1) Task:

“Vanilla” Latent Dirichlet Allocation

1) Task:
   - Discover thematic content of documents

"Vanilla" Latent Dirichlet Allocation

1) Task:
   - Discover thematic content of documents
   - Quickly explore documents

"Vanilla" Latent Dirichlet Allocation

1) Task:
    - Discover thematic content of documents
    - Quickly explore documents
2) Objective Function

"Vanilla" Latent Dirichlet Allocation

1) Task:
   - Discover thematic content of documents
   - Quickly explore documents
2) Objective Function

$$f(\boldsymbol{W}, \boldsymbol{\beta}, \boldsymbol{\Theta}, \boldsymbol{\alpha})$$

Where:

"Vanilla" Latent Dirichlet Allocation

1) Task:
   - Discover thematic content of documents
   - Quickly explore documents
2) Objective Function

$$f(\boldsymbol{W}, \boldsymbol{\beta}, \boldsymbol{\Theta}, \boldsymbol{\alpha})$$

Where:

- $\boldsymbol{\Theta} = N \times K$ matrix with row $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{iK}) \rightsquigarrow$ proportion of a document allocated to each topic

"Vanilla" Latent Dirichlet Allocation

1) Task:
   - Discover thematic content of documents
   - Quickly explore documents
2) Objective Function

$$f(\boldsymbol{W}, \boldsymbol{\beta}, \boldsymbol{\Theta}, \boldsymbol{\alpha})$$

Where:

- $\boldsymbol{\Theta} = N \times K$ matrix with row $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{iK}) \rightsquigarrow$ proportion of a document allocated to each topic
- $\boldsymbol{\beta} = K \times J$ matrix, with row $\boldsymbol{\beta}_k = (\beta_{1k}, \beta_{2k}, \ldots, \beta_{kJ}) \rightsquigarrow$topics

"Vanilla" Latent Dirichlet Allocation

1) Task:
   - Discover thematic content of documents
   - Quickly explore documents
2) Objective Function

$$f(\boldsymbol{W}, \boldsymbol{\beta}, \boldsymbol{\Theta}, \boldsymbol{\alpha})$$

Where:
   - $\boldsymbol{\Theta} = N \times K$ matrix with row $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{iK}) \rightsquigarrow$ proportion of a document allocated to each topic
   - $\boldsymbol{\beta} = K \times J$ matrix, with row $\boldsymbol{\beta}_k = (\beta_{1k}, \beta_{2k}, \ldots, \beta_{kJ}) \rightsquigarrow$topics
   - $\boldsymbol{\alpha} = K$ element long vector, population prior for $\boldsymbol{\Theta}$.

"Vanilla" Latent Dirichlet Allocation

1) Task:
   - Discover thematic content of documents
   - Quickly explore documents
2) Objective Function

$$f(\boldsymbol{W}, \boldsymbol{\beta}, \boldsymbol{\Theta}, \boldsymbol{\alpha})$$

Where:
   - $\boldsymbol{\Theta} = N \times K$ matrix with row $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{iK}) \rightsquigarrow$ proportion of a document allocated to each topic
   - $\boldsymbol{\beta} = K \times J$ matrix, with row $\boldsymbol{\beta}_k = (\beta_{1k}, \beta_{2k}, \ldots, \beta_{kJ}) \rightsquigarrow$topics
   - $\boldsymbol{\alpha} = K$ element long vector, population prior for $\boldsymbol{\Theta}$.
3) Optimization

"Vanilla" Latent Dirichlet Allocation

1) Task:
   - Discover thematic content of documents
   - Quickly explore documents
2) Objective Function

$$f(\boldsymbol{W}, \boldsymbol{\beta}, \boldsymbol{\Theta}, \boldsymbol{\alpha})$$

Where:

- $\boldsymbol{\Theta} = N \times K$ matrix with row $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{iK}) \rightsquigarrow$ proportion of a document allocated to each topic
- $\boldsymbol{\beta} = K \times J$ matrix, with row $\boldsymbol{\beta}_k = (\beta_{1k}, \beta_{2k}, \ldots, \beta_{kJ}) \rightsquigarrow$topics
- $\boldsymbol{\alpha} = K$ element long vector, population prior for $\boldsymbol{\Theta}$.

3) Optimization
   - Variational Inference $\rightsquigarrow$ deterministic approximation

"Vanilla" Latent Dirichlet Allocation

1) Task:
   - Discover thematic content of documents
   - Quickly explore documents
2) Objective Function

$$f(\boldsymbol{W}, \boldsymbol{\beta}, \boldsymbol{\Theta}, \boldsymbol{\alpha})$$

   Where:
   - $\boldsymbol{\Theta} = N \times K$ matrix with row $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{iK}) \rightsquigarrow$ proportion of a document allocated to each topic
   - $\boldsymbol{\beta} = K \times J$ matrix, with row $\boldsymbol{\beta}_k = (\beta_{1k}, \beta_{2k}, \ldots, \beta_{kJ}) \rightsquigarrow$topics
   - $\boldsymbol{\alpha} = K$ element long vector, population prior for $\boldsymbol{\Theta}$.
3) Optimization
   - Variational Inference $\rightsquigarrow$ deterministic approximation
   - Collapsed Gibbs Sampling$\rightsquigarrow$ MCMC algorithm

"Vanilla" Latent Dirichlet Allocation

1) Task:
   - Discover thematic content of documents
   - Quickly explore documents
2) Objective Function

$$f(\boldsymbol{W}, \boldsymbol{\beta}, \boldsymbol{\Theta}, \boldsymbol{\alpha})$$

Where:
   - $\boldsymbol{\Theta} = N \times K$ matrix with row $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{iK}) \rightsquigarrow$ proportion of a document allocated to each topic
   - $\boldsymbol{\beta} = K \times J$ matrix, with row $\boldsymbol{\beta}_k = (\beta_{1k}, \beta_{2k}, \ldots, \beta_{kJ}) \rightsquigarrow$topics
   - $\boldsymbol{\alpha} = K$ element long vector, population prior for $\boldsymbol{\Theta}$.
3) Optimization
   - Variational Inference $\rightsquigarrow$ deterministic approximation
   - Collapsed Gibbs Sampling$\rightsquigarrow$ MCMC algorithm
   - Spectral/Factorization Methods

"Vanilla" Latent Dirichlet Allocation

1) Task:
   - Discover thematic content of documents
   - Quickly explore documents
2) Objective Function

$$f(\boldsymbol{W}, \boldsymbol{\beta}, \boldsymbol{\Theta}, \boldsymbol{\alpha})$$

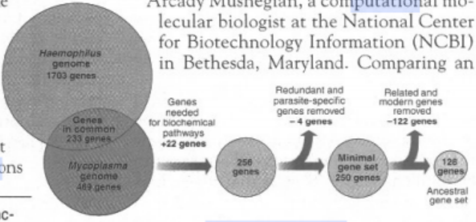
   Where:
   - $\boldsymbol{\Theta} = N \times K$ matrix with row $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{iK}) \rightsquigarrow$ proportion of a document allocated to each topic
   - $\boldsymbol{\beta} = K \times J$ matrix, with row $\boldsymbol{\beta}_k = (\beta_{1k}, \beta_{2k}, \ldots, \beta_{kJ}) \rightsquigarrow$topics
   - $\boldsymbol{\alpha} = K$ element long vector, population prior for $\boldsymbol{\Theta}$.
3) Optimization
   - Variational Inference $\rightsquigarrow$ deterministic approximation
   - Collapsed Gibbs Sampling$\rightsquigarrow$ MCMC algorithm
   - Spectral/Factorization Methods
4) Validation$\rightsquigarrow$ application-specific

A Statistical Highlighter (With Many Colors)

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

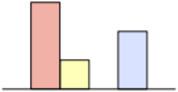


Image from Hanna Wallach

Why does this work $\rightsquigarrow$ Co-occurrence

Where's the information for each word's topic?

Why does this work $\rightsquigarrow$ Co-occurrence

Where's the information for each word's topic?
Reconsider document-term matrix

Why does this work $\rightsquigarrow$ Co-occurrence

Where's the information for each word's topic?
Reconsider document-term matrix

	Word_1	Word_2	$\ldots$	Word_J
Doc_1	0	1	$\ldots$	0
Doc_2	2	0	$\ldots$	3
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Doc_N	0	1	$\ldots$	1

Why does this work $\rightsquigarrow$ Co-occurrence

Where's the information for each word's topic?
Reconsider document-term matrix

	Word_1	Word_2	$\ldots$	Word_J
Doc_1	0	1	$\ldots$	0
Doc_2	2	0	$\ldots$	3
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Doc_N	0	1	$\ldots$	1

We are learning the pattern of what words occur together.

Why does this work $\rightsquigarrow$ Co-occurrence

Where's the information for each word's topic?
Reconsider document-term matrix

	Word_1	Word_2	$\ldots$	Word_J
Doc_1	0	1	$\ldots$	0
Doc_2	2	0	$\ldots$	3
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Doc_N	0	1	$\ldots$	1

We are learning the pattern of what words occur together.

The model wants a topic to contain as few words as possible, but a document to contain as few topics as possible. This tension is what makes the model work.

Extensions to LDA

Extensions to LDA

Have there been extensions to LDA proposed?

Extensions to LDA

Have there been extensions to LDA proposed?
Yes.

Extensions to LDA

Have there been extensions to LDA proposed?
Yes.
correlated topic models,

Extensions to LDA

Have there been extensions to LDA proposed?
Yes.
correlated topic models, dynamic topic models,

Extensions to LDA

Have there been extensions to LDA proposed?
Yes.
correlated topic models, dynamic topic models, hierarchical LDA,

Extensions to LDA

Have there been extensions to LDA proposed?
Yes.
correlated topic models, dynamic topic models, hierarchical LDA, pachinko allocation, nonparametric pachinko allocation,

Extensions to LDA

Have there been extensions to LDA proposed?
Yes.
correlated topic models, dynamic topic models, hierarchical LDA, pachinko allocation, nonparametric pachinko allocation,factorial LDA, gamma-poisson factorization, shared component topic models, dirichlet multinomial regresson topic models,

Extensions to LDA

Have there been extensions to LDA proposed?
Yes.
correlated topic models, dynamic topic models, hierarchical LDA, pachinko allocation, nonparametric pachinko allocation,factorial LDA, gamma-poisson factorization, shared component topic models, dirichlet multinomial regresson topic models, expressed agenda model, structured topic model, nested hierarchical dirichlet process topic model, focused topic model, inverse regression topic model, ideal point topic model, discrete infinite logistic normal topic model

Extensions to LDA

Have there been extensions to LDA proposed?
Yes.
correlated topic models, dynamic topic models, hierarchical LDA, pachinko allocation, nonparametric pachinko allocation,factorial LDA, gamma-poisson factorization, shared component topic models, dirichlet multinomial regresson topic models, expressed agenda model, structured topic model, nested hierarchical dirichlet process topic model, focused topic model, inverse regression topic model, ideal point topic model, discrete infinite logistic normal topic model multilingual topic model, markov topic model, relational topic model, syntactic topic model, supervised latent dirichlet allocation

Extensions to LDA

Have there been extensions to LDA proposed?
Yes.
correlated topic models, dynamic topic models, hierarchical LDA, pachinko allocation, nonparametric pachinko allocation,factorial LDA, gamma-poisson factorization, shared component topic models, dirichlet multinomial regresson topic models, expressed agenda model, structured topic model, nested hierarchical dirichlet process topic model, focused topic model, inverse regression topic model, ideal point topic model, discrete infinite logistic normal topic model multilingual topic model, markov topic model, relational topic model, syntactic topic model, supervised latent dirichlet allocation

different methods for every problem

Extensions to LDA

Have there been extensions to LDA proposed?
Yes.
correlated topic models, dynamic topic models, hierarchical LDA, pachinko allocation, nonparametric pachinko allocation,factorial LDA, gamma-poisson factorization, shared component topic models, dirichlet multinomial regresson topic models, expressed agenda model, structured topic model, nested hierarchical dirichlet process topic model, focused topic model, inverse regression topic model, ideal point topic model, discrete infinite logistic normal topic model multilingual topic model, markov topic model, relational topic model, syntactic topic model, supervised latent dirichlet allocation

different methods for every problem
What is going on with all of these extensions?

Correlated Topic Models (Blei and Lafferty 2007)

Dirichlet distribution$\rightsquigarrow$ Assumes negative covariance between topics

Correlated Topic Models (Blei and Lafferty 2007)

Dirichlet distribution $\rightsquigarrow$ Assumes negative covariance between topics
Logistic Normal Distribution $\rightsquigarrow$ Allows some positive covariance between topics

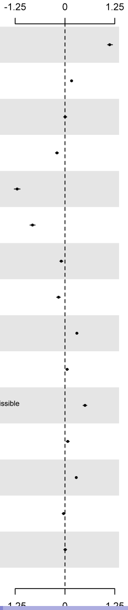
Correlated Topic Models (Blei and Lafferty 2007)

Dirichlet distribution $\rightsquigarrow$ Assumes negative covariance between topics
Logistic Normal Distribution $\rightsquigarrow$ Allows some positive covariance between topics

$$
\begin{aligned}
\boldsymbol{\beta}_k &\sim \text{Dirichlet}(\mathbf{1}) \\
\boldsymbol{\eta}_i | \boldsymbol{\mu}, \boldsymbol{\Sigma} &\sim \text{Multivariate Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
\boldsymbol{\theta}_i &= \frac{\exp\left(\boldsymbol{\eta}_i\right)}{\sum_{k=1}^{K} \exp\left(\eta_{ik}\right)} \\
\boldsymbol{z}_{im} | \boldsymbol{\theta}_i &\sim \text{Multinomial}(1, \boldsymbol{\eta}_i) \\
w_{im} | \boldsymbol{\beta}_k, z_{imk} = 1 &\sim \text{Multinomial}(1, \boldsymbol{\beta}_k)
\end{aligned}
$$

Jihad Example

1.	**Fighting**	F: Muslim, Jihad, Islam, fight, Jihadi fighters, pathway, almighty, that FREX: jihad, fighting, jihadist fighters, pulpit, approves of us, annotated, to fight, vicinity F: مسلم, جهاد, اسلام, قتل, مجاهد, سبيل, تعال, ذين FREX: جهاد, قتال, مجاهد, منبر, يوافقنا, مذيل, يقاتل, بجوار
2.	**Social theory**	F: person, life, soul/self, knowledge/science, society, work, image, material/physical FREX: imagine, morals, develop, society, product,necessarily, environment, traditions, activity F: انس, حيا, نفس, علم, مجتمع, عمل, صور, ماد FREX: تصور, اخلاق, تطور, مجتمع, انتاج, حتم, بيئ, تقاليد
3.	**Politics**	F: Arab, Jews, country, Islam, A.D., year, West, Muslim FREX: capitol, Asia, Iran, South, Washington, A.D., Russia, Turkey F: عرب, يهود, دول, اسلام, م, سن, غرب, مسلم FREX: عاصمت, اسيا, اير, جنوب, اشنطن, م, روسيا, تركيا
4.	**The Prophet**	F: said, prayers (be upon him), peace (be upon him), almighty, messenger, glory, prophet, that FREX: almighty, almighty, glory, bless you, magic, punishment, hypocrisy, sins F: قال, صل, سلم, تعال, رسول, سبح, نب, ذين FREX: وجل, عز, سبح, تبارك, سحر, عذاب, رياء, ذنوب
5.	**Prayer**	F: prayer, pray, son, prophet, sheikh, mosque, fatwas, group FREX: prostration, prostrated, Abd al-Aziz, supplicant, Baz, prayer space, omission, prostration F: صلا, صل, سلم, بن, نب, شيخ, مسجد, فتاو FREX: ركع, ركعت, عبدالعزيز, ماموم, باز, مصل, سهو, ركوع
6.	**Ramadan**	F: day, fasting, Ashura, Ramadan, sheikh, group, fatwas, Uthaymeen FREX: wash, one who fasts, fasting, fasting, to break fast, Ramadan, travel, dirty F: يوم, صيام, عشر, رمض, شيخ, مجموع, فتاو, عثيم FREX: غسل, صائم, صيام, صوم, يفطر, رمض, مسافر, نجاس
7.	**Family and Women**	F: woman, O, man, girl, one, says, men, people FREX: veil, youth, (sheikh) Tamim, Azzam, tanks, finery, wear, r(typo) F: مرا, يا, رجل, نساء, احد, يقول, رجال, ناس FREX: حجاب, شاب, تميم, عزام, دباب, تبرج, لباس, ر
8.	**Money, Pilgrimage, and Marriage**	F: tithing, money, pilgrimage, permitted, religion, marriage, believe/ratify, divorce FREX: tithing, divorce, banks, divorce, card, banks, to perform pilgrimage, poor F: زكا, مال, حج, يجوز, دين, زوج, صدق, طلاق FREX: زكا, طلاق, بنك, طلق, بطاق, بنوك, يحج, فقراء
9.	**Islam and Modernity**	F: Islam, land, mankind, people, religion, life, other, God FREX: Europe, civilization, European, mankind, church, goods, generations, their lives F: اسلام, ارض, بشر, ناس, دين, حيا, اخر, ال FREX: اوربا, حضار, اورب, بشر, كنيس, متاع, اجيال, حياتهم
10.	**Hadith**	F: Saying, hadith, said, prayers (be upon him), peace (be upon him), Muslim, legally, not FREX: to forbid, analogy, permission, general, evidence, forbid, text, absolutely F: قول, حديث, قال, صل, سلم, مسلم, شرع, ليس FREX: تحريم, قياس, جواز, عموم, ادل, منع, نص, مطلقا
11.	**Excommunication**	F: Apostasy, said, almighty, polytheism, Islam, Apostate, saying, people FREX: excommunicate, apostate, apostasy, sponsorship, idolatry, excommunication, idols, to make permissible F: كفر, قال, تعال, شرك, اسلام, كافر, قول, اهل FREX: يكفر, كفرا, كفر, موالا, شرك, تكفير, اصنام, استحلال
12.	**Salafism**	F: Sunna, sheikh, son, people, book, knowledge, Salafi, Muhammad FREX: heterodoxy, innovator, Sufi, Salafi, to draw near to, distinguish, (the) saved (group), to undertake F: سن, شيخ, بن, اهل, كتاب, علم, سلف, محمد FREX: ابتداع, مبتدع, صوف, سلف, جانبوا, فتميزوا, ناج, استقاموا
13.	**Shari'a and Law**	F: Islam, wisdom, right, people, thing, legally, Shari'a, religion FREX: Shari'a, to legislate, to send down, to judge, judgment, justice, parliament, court F: اسلام, حكم, حق, ناس, امر, شرع, شريع, دين FREX: شريع, تشريع, انزل, تحاكم, تحكيم, عدل, برلم, محاكم
14.	**Creed**	F: knowledge, qualities, saying, meaning, Quran, to be, people, said FREX: creatures, characteristic/quality, names, throne, al-Tahawi, proof, mean, question (abbreviation) F: علم, صف, قول, يعن, قر, يك, اهل, قال FREX: مخلوق, صف, اسماء, عرش, طحاو, اثب, يعن, س
15.	**Hadith Narration**	F: Said, son, son, father, hadith, narrated, Abd, (may God be) pleased (with him) FREX: narrate, Daoud, chain of narration, narrate, al-Bayhaqi, al-Tabrani, our saying, al-Tirmidhi F: قال, بن, ابن, اب, حديث, روا, عبد, رض FREX: روا, داود, اسناد, رو, بيهق, طبر, حدثنا, ترمذ



Fighting

F: Muslim, Jihad, Islam, fight, Jihadi fighters, pathway, almighty, that
FREX: jihad, fighting, jihadist fighters, pulpit, approves of us, annotated, to fight, vicinity
F: مسلم, جهاد, اسلام, قتل, مجاهد, سبيل, تعال, ذين FREX: جهاد, قتال, مجاهد, منبر, يوافقنا, مذيل, يقاتل, بجوار

Social theory

F: person, life, soul/self, knowledge/science, society, work, image, material/physical
FREX: imagine, morals, develop, society, product,necessarily, environment, traditions, activity
F: انس, حيا, نفس, علم, مجتمع, عمل, صور, ماد FREX: تصور, اخلاق, تطور, مجتمع, انتاج, حتم, بيئ, تقاليد

Politics

F: Arab, Jews, country, Islam, A.D., year, West, Muslim
FREX: capitol, Asia, Iran, South, Washington, A.D., Russia, Turkey
F: عرب, يهود, دول, اسلام, م, سن, غرب, مسلم FREX: عاصمت, اسيا, اير, جنوب, اشنطن, م, روسيا, تركيا

The Prophet

F: said, prayers (be upon him), peace (be upon him), almighty, messenger, glory, prophet, that
FREX: almighty, almighty, glory, bless you, magic, punishment, hypocrisy, sins
F: قال, صل, سلم, تعال, رسول, سبح, نب, ذين FREX: وجل, عز, سبح, تبارك, سحر, عذاب, رياء, ذنوب

Prayer

F: prayer, pray, son, prophet, sheikh, mosque, fatwas, group
FREX: prostration, prostrated, Abd al-Aziz, supplicant, Baz, prayer space, omission, prostration
F: صلا, صل, سلم, بن, نب, شيخ, مسجد, فتاو FREX: ركع, ركعت, عبدالعزيز, ماموم, باز, مصل, سهو, ركوع

Ramadan

F: day, fasting, Ashura, Ramadan, sheikh, group, fatwas, Uthaymeen
FREX: wash, one who fasts, fasting, fasting, to break fast, Ramadan, travel, dirty
F: يوم, صيام, عشر, رمض, شيخ, مجموع, فتاو, عثيم FREX: غسل, صائم, صيام, صوم, يفطر, رمض, مسافر, نجاس

Family and Women

F: woman, O, man, girl, one, says, men, people
FREX: veil, youth, (sheikh) Tamim, Azzam, tanks, finery, wear, r(typo)
F: مرا, يا, رجل, نساء, احد, يقول, رجال, ناس FREX: حجاب, شاب, تميم, عزام, دباب, تبرج, لباس, ر

Money, Pilgrimage, and Marriage

F: tithing, money, pilgrimage, permitted, religion, marriage, believe/ratify, divorce
FREX: tithing, divorce, banks, divorce, card, banks, to perform pilgrimage, poor
F: زكا, مال, حج, يجوز, دين, زوج, صدق, طلاق FREX: زكا, طلاق, بنك, طلق, بطاق, بنوك, يحج, فقراء

Islam and Modernity

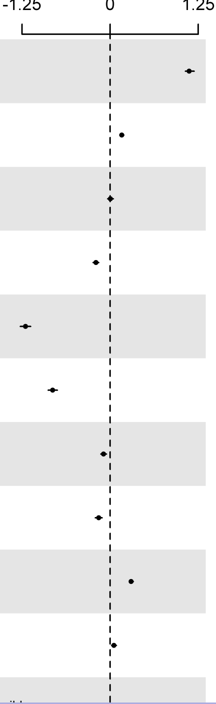
F: Islam, land, mankind, people, religion, life, other, God
FREX: Europe, civilization, European, mankind, church, goods, generations, their lives
F: اسلام, ارض, بشر, ناس, دين, حيا, اخر, ال FREX: اوربا, حضار, اورب, بشر, كنيس, متاع, اجيال, حياتهم

Hadith

F: Saying, hadith, said, prayers (be upon him), peace (be upon him), Muslim, legally, not
FREX: to forbid, analogy, permission, general, evidence, forbid, text, absolutely
F: قول, حديث, قال, صل, سلم, مسلم, شرع, ليس FREX: تحريم, قياس, جواز, عموم, ادل, منع, نص, مطلقا

Excommunication

F: Apostasy, said, almighty, polytheism, Islam, Apostate, saying, people



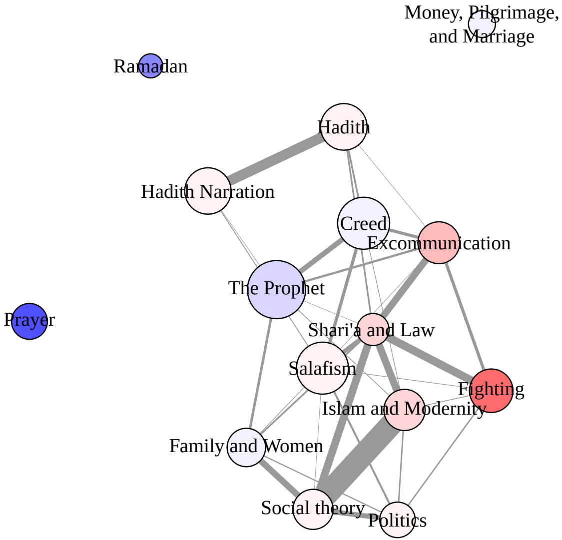
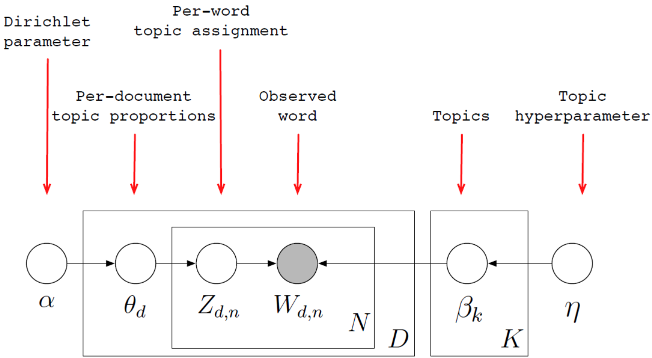


Figure: The network of correlated topics for a 15-topic Structural Topic Model with Jihadi/not-Jihadi as the predictor of topics in Arab Muslim cleric writings.

LDA $\rightsquigarrow$ Dynamic Topic Model (Blei and Lafferty 2007)

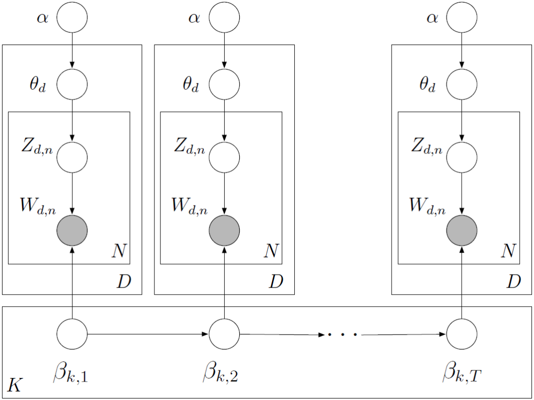
Figure: Plate Notation of Latent Dirichlet Allocation



Graphic from David Blei

LDA $\rightsquigarrow$ Dynamic Topic Model

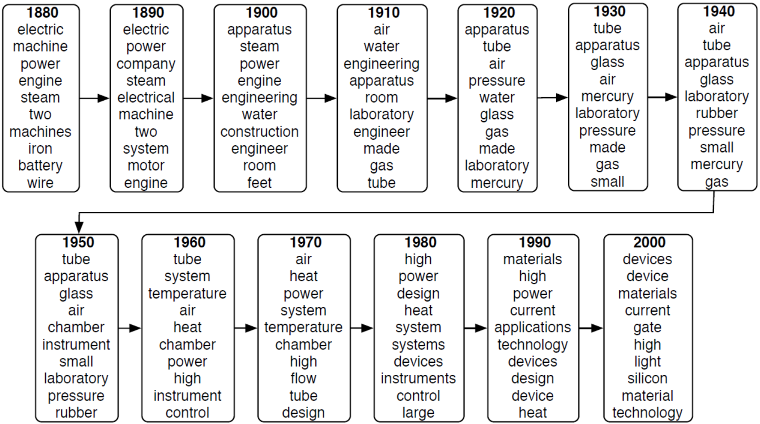
Figure: Dynamic Topic Model



Graphic from David Blei

LDA $\rightsquigarrow$ Dynamic Topic Model

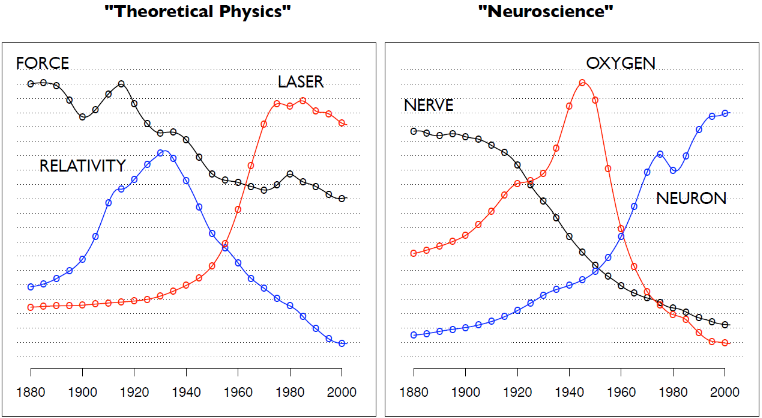
Figure: Topic Evolution over Time



Graphic from David Blei

LDA ⇝ Dynamic Topic Model

Figure: Word Use in Topics Over Time



Graphic from David Blei

Expressed Agenda Model (Grimmer 2010)

Expressed Agenda Model (Grimmer 2010)

1. Assumes:

Expressed Agenda Model (Grimmer 2010)

1. Assumes:
   1. Each document is assigned to one topic

Expressed Agenda Model (Grimmer 2010)

1. Assumes:
    1. Each document is assigned to one topic
    2. Each author allocates some hidden proportion of time to each topic

Expressed Agenda Model (Grimmer 2010)

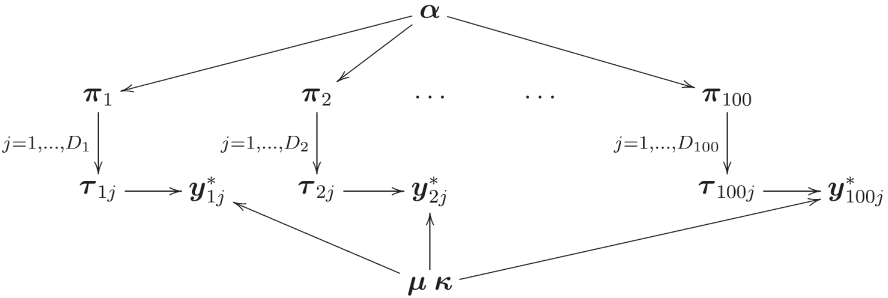
1. Assumes:
   1. Each document is assigned to one topic
   2. Each author allocates some hidden proportion of time to each topic
2. Grimmer's project seeks to quantitatively represent the content of senators' press releases.

Expressed Agenda Model (Grimmer 2010)

1. Assumes:
   1. Each document is assigned to one topic
   2. Each author allocates some hidden proportion of time to each topic
2. Grimmer's project seeks to quantitatively represent the content of senators' press releases.
3. It is called the Expressed Agenda Model because it captures the way they communicate that agenda to constituents.

Expressed Agenda Model

Figure: Expressed Agenda Model



Graphic from Grimmer 2010

STM = LDA + Contextual Information

STM = LDA + Contextual Information

- STM provides two ways to include contextual information

STM = LDA + Contextual Information

- STM provides two ways to include contextual information
  - Topic prevalence can vary by metadata

STM $=$ LDA $+$ Contextual Information

- STM provides two ways to include contextual information
  - Topic prevalence can vary by metadata
    - e.g. city papers cover protests more than provincial papers

STM = LDA + Contextual Information

- STM provides two ways to include contextual information
  - Topic prevalence can vary by metadata
    - e.g. city papers cover protests more than provincial papers
  - Topic content can vary by metadata

STM = LDA + Contextual Information

- STM provides two ways to include contextual information
  - Topic prevalence can vary by metadata
    - e.g. city papers cover protests more than provincial papers
  - Topic content can vary by metadata
    - e.g. city papers talk about protests differently

STM = LDA + Contextual Information

- STM provides two ways to include contextual information
  - Topic prevalence can vary by metadata
    - e.g. city papers cover protests more than provincial papers
  - Topic content can vary by metadata
    - e.g. city papers talk about protests differently
- Including context improves the model:

STM = LDA + Contextual Information

- STM provides two ways to include contextual information
  - Topic prevalence can vary by metadata
    - e.g. city papers cover protests more than provincial papers
  - Topic content can vary by metadata
    - e.g. city papers talk about protests differently
- Including context improves the model:
  - more accurate estimation

STM = LDA + Contextual Information

- STM provides two ways to include contextual information
  - Topic prevalence can vary by metadata
    - e.g. city papers cover protests more than provincial papers
  - Topic content can vary by metadata
    - e.g. city papers talk about protests differently
- Including context improves the model:
  - more accurate estimation
  - better qualitative interpretability

STM: What this means in pictures

Say you have
a lot of people.

Each writes
some text

that discuss a few
different topics

.7

.3

Politics

congress, nations, power, votes, agreement, bargaining

Statistics

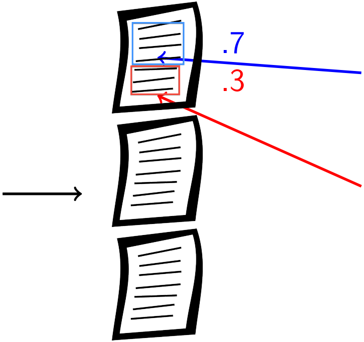
estimator, data, analysis, variance, model, inference

The STM Allows for:

STM: What this means in pictures

Say you have
a lot of people.

Each writes
some text



that discuss a few
different topics

Politics

congress, nations, power, votes, agreement, bargaining

Statistics

estimator, data, analysis, variance, model, inference

The STM Allows for:

1. The words in each topic to vary by gender

STM: What this means in pictures

Say you have a lot of people.

Each writes some text

.7

.3

that discuss a few different topics

Politics

congress, nations, power, votes, agreement, bargaining

Statistics

estimator, data, analysis, variance, model, inference

The STM Allows for:

1. The words in each topic to vary by gender

STM: What this means in pictures

Say you have
a lot of people.

Each writes
some text

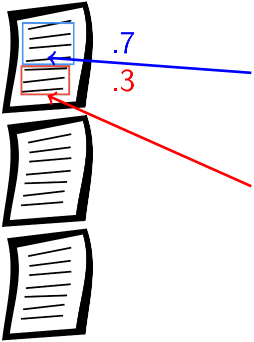
that discuss a few
different topics

Group A

Group B

Group A

→



Politics

congress, nations, power, votes, agree-ment, bargaining

Statistics

estimator, data, analysis, variance, model, inference

The STM Allows for:

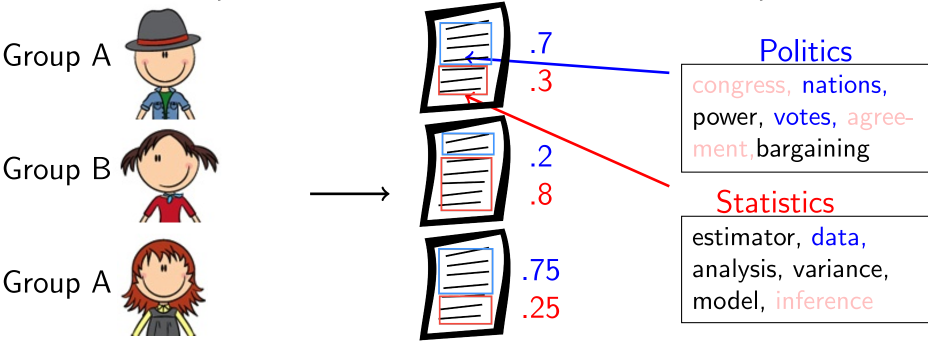
1. The words in each topic to vary by gender
2. The topic proportions to vary by group

STM: What this means in pictures

Say you have a lot of people.

Each writes some text

that discuss a few different topics



The STM Allows for:

1. The words in each topic to vary by gender
2. The topic proportions to vary by group

Mixed-Membership Topic Models

More formal terminology:

Mixed-Membership Topic Models

More formal terminology:

- User specifies the number of topics: K

Mixed-Membership Topic Models

More formal terminology:

- User specifies the number of topics: K
- Observed data for standard topic models

Mixed-Membership Topic Models

More formal terminology:

- User specifies the number of topics: K
- Observed data for standard topic models
  - Each document ($d \in 1 \dots D$) is a collection of N_d tokens

Mixed-Membership Topic Models

More formal terminology:

- User specifies the number of topics: K
- Observed data for standard topic models
  - Each document ($d \in 1 \ldots D$) is a collection of N_d tokens
  - Each token is a particular word from a dictionary of V entries

Mixed-Membership Topic Models

More formal terminology:

- User specifies the number of topics: K
- Observed data for standard topic models
  - Each document ($d \in 1 \ldots D$) is a collection of N_d tokens
  - Each token is a particular word from a dictionary of V entries
  - Data summarized in a single matrix $D \times V$ matrix $\mathbf{W}$

Mixed-Membership Topic Models

More formal terminology:

- User specifies the number of topics: K
- Observed data for standard topic models
  - Each document ($d \in 1 \ldots D$) is a collection of N_d tokens
  - Each token is a particular word from a dictionary of V entries
  - Data summarized in a single matrix $D \times V$ matrix $\mathbf{W}$
- Additional data for STM

Mixed-Membership Topic Models

More formal terminology:

- User specifies the number of topics: K
- Observed data for standard topic models
  - Each document ($d \in 1 \ldots D$) is a collection of N_d tokens
  - Each token is a particular word from a dictionary of V entries
  - Data summarized in a single matrix $D \times V$ matrix $\mathbf{W}$
- Additional data for STM
  - Topic prevalence covariates: $D \times P$ matrix $\mathbf{X}$

Mixed-Membership Topic Models

More formal terminology:

- User specifies the number of topics: K
- Observed data for standard topic models
  - Each document ($d \in 1 \ldots D$) is a collection of N_d tokens
  - Each token is a particular word from a dictionary of V entries
  - Data summarized in a single matrix $D \times V$ matrix $\mathbf{W}$
- Additional data for STM
  - Topic prevalence covariates: $D \times P$ matrix $\mathbf{X}$
  - Topical content groups: D length vector $\mathbf{Y}$

Mixed-Membership Topic Models

More formal terminology:

- User specifies the number of topics: K
- Observed data for standard topic models
  - Each document ($d \in 1 \dots D$) is a collection of N_d tokens
  - Each token is a particular word from a dictionary of V entries
  - Data summarized in a single matrix $D \times V$ matrix $\mathbf{W}$
- Additional data for STM
  - Topic prevalence covariates: $D \times P$ matrix $\mathbf{X}$
  - Topical content groups: D length vector $\mathbf{Y}$
- Latent variables

Mixed-Membership Topic Models

More formal terminology:

- User specifies the number of topics: K
- Observed data for standard topic models
  - Each document ($d \in 1 \ldots D$) is a collection of N_d tokens
  - Each token is a particular word from a dictionary of V entries
  - Data summarized in a single matrix $D \times V$ matrix $\mathbf{W}$
- Additional data for STM
  - Topic prevalence covariates: $D \times P$ matrix $\mathbf{X}$
  - Topical content groups: D length vector $\mathbf{Y}$
- Latent variables
  - $D \times K$ matrix θ: proportion of document on each topic.

Mixed-Membership Topic Models

More formal terminology:

- User specifies the number of topics: K
- Observed data for standard topic models
  - Each document ($d \in 1 \ldots D$) is a collection of N_d tokens
  - Each token is a particular word from a dictionary of V entries
  - Data summarized in a single matrix $D \times V$ matrix $\mathbf{W}$
- Additional data for STM
  - Topic prevalence covariates: $D \times P$ matrix $\mathbf{X}$
  - Topical content groups: D length vector $\mathbf{Y}$
- Latent variables
  - $D \times K$ matrix θ: proportion of document on each topic.
  - $K \times V$ matrix β: probability of drawing a word conditional on topic.

Mixed-Membership Topic Models

More formal terminology:

- User specifies the number of topics: K
- Observed data for standard topic models
  - Each document ($d \in 1 \ldots D$) is a collection of N_d tokens
  - Each token is a particular word from a dictionary of V entries
  - Data summarized in a single matrix $D \times V$ matrix $\mathbf{W}$
- Additional data for STM
  - Topic prevalence covariates: $D \times P$ matrix $\mathbf{X}$
  - Topical content groups: D length vector $\mathbf{Y}$
- Latent variables
  - $D \times K$ matrix θ: proportion of document on each topic.
  - $K \times V$ matrix β: probability of drawing a word conditional on topic.
  - Low rank approximation to expected counts: $\underset{D\times V}{\tilde{W}} \approx \underset{D\times K}{\theta} \underset{K\times V}{\beta}$

Technical Details: The Structural Topic Model

- Low rank approximation to expected counts: $\underset{D \times V}{\tilde{W}} \approx \underset{D \times K}{\theta} \underset{K \times V}{\beta}$
- θ, $D \times K$ document-topic matrix
- β, $K \times V$ topic-word matrix
- Each token has a topic drawn from the document mixture
  - Draw token topic $z_{d,n}$ from Discrete(θ_d)
  - Draw observed word $w_{d,n}$ from Discrete($\beta_{k=z,}$)

Technical Details: The Structural Topic Model

- Low rank approximation to expected counts: $\underset{D\times V}{\tilde{W}} \approx \underset{D\times K}{\theta}\ \underset{K\times V}{\beta}$
- θ, $D \times K$ document-topic matrix $\Leftarrow$ logistic normal glm with covariates
- β, $K \times V$ topic-word matrix
- Each token has a topic drawn from the document mixture
  - Draw token topic $z_{d,n}$ from Discrete(θ_d)
  - Draw observed word $w_{d,n}$ from Discrete($\beta_{k=z,}$)

Technical Details: The Structural Topic Model

- Low rank approximation to expected counts: $\underset{D\times V}{\tilde{W}} \approx \underset{D\times K}{\theta}\underset{K\times V}{\beta}$
- θ, $D \times K$ document-topic matrix $\Leftarrow$ logistic normal glm with covariates
  - Covariate-specific prior with global topic covariance
  - $\theta_{d,\cdot} \sim \text{LogisticNormal}(X_d\gamma, \Sigma)$
- β, $K \times V$ topic-word matrix

- Each token has a topic drawn from the document mixture
  - Draw token topic $z_{d,n}$ from Discrete(θ_d)
  - Draw observed word $w_{d,n}$ from Discrete$(\beta_{k=z,})$

Technical Details: The Structural Topic Model

- Low rank approximation to expected counts: $\underset{D\times V}{\tilde{W}} \approx \underset{D\times K}{\theta}\underset{K\times V}{\beta}$
- θ, $D \times K$ document-topic matrix $\Leftarrow$ logistic normal glm with covariates
  - Covariate-specific prior with global topic covariance
  - $\theta_{d,\cdot} \sim \text{LogisticNormal}(X_d\gamma, \Sigma)$
- β, $K \times V$ topic-word matrix $\Leftarrow$ multinomial logit with covariates

- Each token has a topic drawn from the document mixture
  - Draw token topic $z_{d,n}$ from Discrete(θ_d)
  - Draw observed word $w_{d,n}$ from Discrete($\beta_{k=z,}$)

Technical Details: The Structural Topic Model

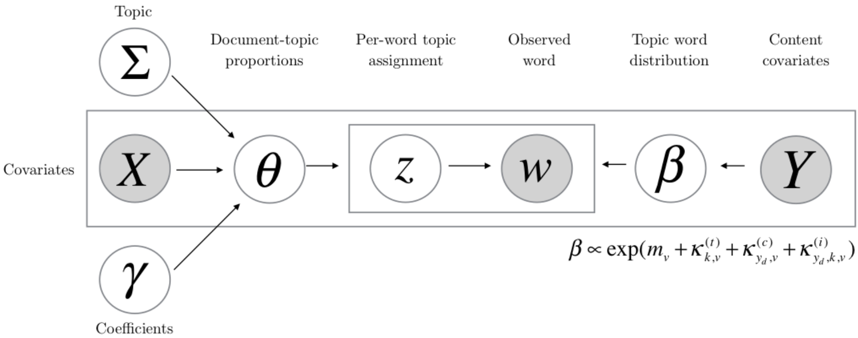
- Low rank approximation to expected counts: $\underset{D \times V}{\tilde{W}} \approx \underset{D \times K}{\theta} \underset{K \times V}{\beta}$
- θ, $D \times K$ document-topic matrix $\Leftarrow$ logistic normal glm with covariates
  - Covariate-specific prior with global topic covariance
  - $\theta_{d,\cdot} \sim \text{LogisticNormal}(X_d \gamma, \Sigma)$
- β, $K \times V$ topic-word matrix $\Leftarrow$ multinomial logit with covariates
  - Each topic is now a sparse, covariate-specific deviation from a baseline distribution.
  - $\vec{\beta}_{k,\cdot} \propto \exp\left(m + \kappa^{(\text{topic})} + \kappa^{(\text{cov})} + \kappa^{(\text{int})}\right)$
  - Thee parts: topic, covariate, topic-covariate interaction

- Each token has a topic drawn from the document mixture
  - Draw token topic $z_{d,n}$ from Discrete(θ_d)
  - Draw observed word $w_{d,n}$ from Discrete$(\beta_{k=z,})$

Technical Details: The Structural Topic Model

- Low rank approximation to expected counts: $\underset{D \times V}{\tilde{W}} \approx \underset{D \times K}{\theta} \underset{K \times V}{\beta}$
- θ, $D \times K$ document-topic matrix $\Leftarrow$ logistic normal glm with covariates
  - Covariate-specific prior with global topic covariance
  - $\theta_{d,\cdot} \sim \text{LogisticNormal}(X_d\gamma, \Sigma)$
- β, $K \times V$ topic-word matrix $\Leftarrow$ multinomial logit with covariates
  - Each topic is now a sparse, covariate-specific deviation from a baseline distribution.
  - $\vec{\beta}_{k,\cdot} \propto \exp\left(m + \kappa^{(\text{topic})} + \kappa^{(\text{cov})} + \kappa^{(\text{int})}\right)$
  - Thee parts: topic, covariate, topic-covariate interaction
  - β may instead by point-estimated
- Each token has a topic drawn from the document mixture
  - Draw token topic $z_{d,n}$ from Discrete(θ_d)
  - Draw observed word $w_{d,n}$ from Discrete($\beta_{k=z,}$)

Structural Topic Model



Estimation and Implementation of the STM

Estimation and Implementation of the STM

- Define a probabilistic model and estimate parameters

Estimation and Implementation of the STM

- Define a probabilistic model and estimate parameters
    - bayesian estimation using variational inference

Estimation and Implementation of the STM

- Define a probabilistic model and estimate parameters
  - bayesian estimation using variational inference (initialization from spectral method of moments estimator)

Estimation and Implementation of the STM

- Define a probabilistic model and estimate parameters
  - bayesian estimation using variational inference (initialization from spectral method of moments estimator)
  - essentially word co-occurences used to discover topics

Estimation and Implementation of the STM

- Define a probabilistic model and estimate parameters
  - bayesian estimation using variational inference (initialization from spectral method of moments estimator)
  - essentially word co-occurences used to discover topics
- General to many kinds of corpus structure using covariates

Estimation and Implementation of the STM

- Define a probabilistic model and estimate parameters
  - bayesian estimation using variational inference (initialization from spectral method of moments estimator)
  - essentially word co-occurences used to discover topics
- General to many kinds of corpus structure using covariates
- `stm` Package in `R`

Estimation and Implementation of the STM

- Define a probabilistic model and estimate parameters
  - bayesian estimation using variational inference (initialization from spectral method of moments estimator)
  - essentially word co-occurences used to discover topics
- General to many kinds of corpus structure using covariates
- `stm` Package in `R`
  - complete workflow: raw texts → figures

Estimation and Implementation of the STM

- Define a probabilistic model and estimate parameters
  - bayesian estimation using variational inference (initialization from spectral method of moments estimator)
  - essentially word co-occurences used to discover topics
- General to many kinds of corpus structure using covariates
- `stm` Package in `R`
  - complete workflow: raw texts $\rightarrow$ figures
  - simple regression style syntax using formulas

```
mod.out <- stm(documents,vocab, K=10,
               prevalence= ~paper + s(time),
               data=metadata, init.type="Spectral")
```

Estimation and Implementation of the STM

- Define a probabilistic model and estimate parameters
  - bayesian estimation using variational inference (initialization from spectral method of moments estimator)
  - essentially word co-occurences used to discover topics
- General to many kinds of corpus structure using covariates
- `stm` Package in R
  - complete workflow: raw texts → figures
  - simple regression style syntax using formulas

```
mod.out <- stm(documents,vocab, K=10,
               prevalence= ~paper + s(time),
               data=metadata, init.type="Spectral")
```

  - many functions for summarization, visualization and checking

Estimation and Implementation of the STM

- Define a probabilistic model and estimate parameters
  - bayesian estimation using variational inference (initialization from spectral method of moments estimator)
  - essentially word co-occurences used to discover topics
- General to many kinds of corpus structure using covariates
- `stm` Package in `R`
  - complete workflow: raw texts $\rightarrow$ figures
  - simple regression style syntax using formulas

```
mod.out <- stm(documents,vocab, K=10,
               prevalence= ~paper + s(time),
               data=metadata, init.type="Spectral")
```

  - many functions for summarization, visualization and checking
- Complete vignette online with examples

Estimation and Implementation of the STM

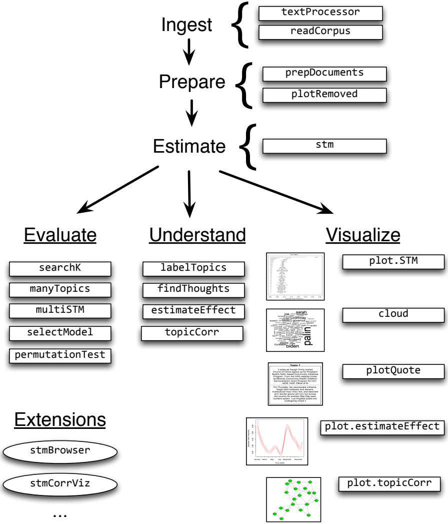
- Define a probabilistic model and estimate parameters
  - bayesian estimation using variational inference (initialization from spectral method of moments estimator)
  - essentially word co-occurences used to discover topics
- General to many kinds of corpus structure using covariates
- `stm` Package in R
  - complete workflow: raw texts $\rightarrow$ figures
  - simple regression style syntax using formulas

```
mod.out <- stm(documents,vocab, K=10,
               prevalence= ~paper + s(time),
               data=metadata, init.type="Spectral")
```

  - many functions for summarization, visualization and checking
- Complete vignette online with examples

You can do this with your data!

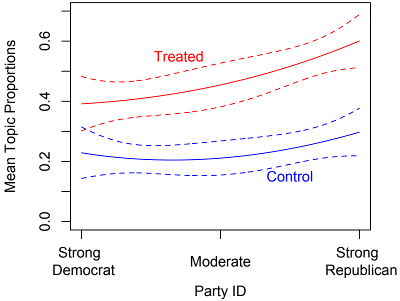
stm is Full of Functions to Help You!



Applications

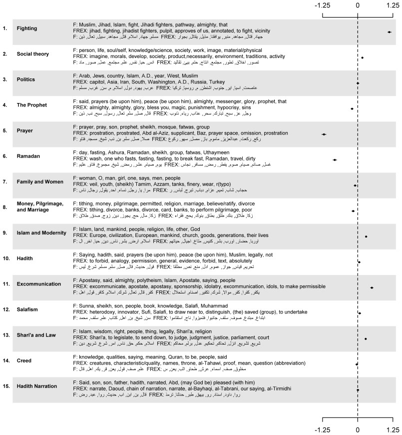
Applications

- Open-Ended Survey Response (Roberts et al 2014)



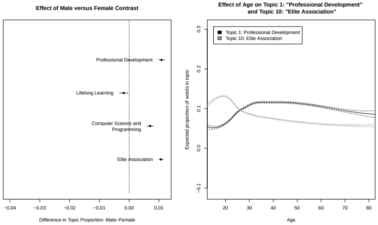
Applications

- Open-Ended Survey Response (Roberts et al 2014)
- Fatwas of Jihadi Clerics (Lucas et al 2015)



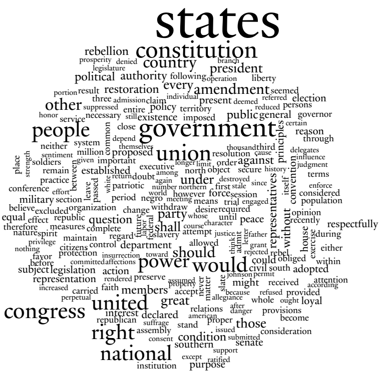
Applications

- Open-Ended Survey Response (Roberts et al 2014)
- Fatwas of Jihadi Clerics (Lucas et al 2015)
- Student Text in MOOCs (Reich et al 2015)



Applications

- Open-Ended Survey Response (Roberts et al 2014)
- Fatwas of Jihadi Clerics (Lucas et al 2015)
- Student Text in MOOCs (Reich et al 2015)
- Constitutional Moments (Stewart and Young)



Applications

- Open-Ended Survey Response (Roberts et al 2014)
- Fatwas of Jihadi Clerics (Lucas et al 2015)
- Student Text in MOOCs (Reich et al 2015)
- Constitutional Moments (Stewart and Young)
- Chinese/Western Newswires (Roberts et al 2016)

Applications

- Open-Ended Survey Response (Roberts et al 2014)
- Fatwas of Jihadi Clerics (Lucas et al 2015)
- Student Text in MOOCs (Reich et al 2015)
- Constitutional Moments (Stewart and Young)
- Chinese/Western Newswires (Roberts et al 2016)
- How Teams Make Forecasts (Horowitz et al 2017)

Applications

- Open-Ended Survey Response (Roberts et al 2014)
- Fatwas of Jihadi Clerics (Lucas et al 2015)
- Student Text in MOOCs (Reich et al 2015)
- Constitutional Moments (Stewart and Young)
- Chinese/Western Newswires (Roberts et al 2016)
- How Teams Make Forecasts (Horowitz et al 2017)
- Self-Censorship of Bloggers (Roberts et al 2017)

Applications

- Open-Ended Survey Response (Roberts et al 2014)
- Fatwas of Jihadi Clerics (Lucas et al 2015)
- Student Text in MOOCs (Reich et al 2015)
- Constitutional Moments (Stewart and Young)
- Chinese/Western Newswires (Roberts et al 2016)
- How Teams Make Forecasts (Horowitz et al 2017)
- Self-Censorship of Bloggers (Roberts et al 2017)
- Snowden Leak Response (Lucas et al 2015)

Applications

- Open-Ended Survey Response (Roberts et al 2014)
- Fatwas of Jihadi Clerics (Lucas et al 2015)
- Student Text in MOOCs (Reich et al 2015)
- Constitutional Moments (Stewart and Young)
- Chinese/Western Newswires (Roberts et al 2016)
- How Teams Make Forecasts (Horowitz et al 2017)
- Self-Censorship of Bloggers (Roberts et al 2017)
- Snowden Leak Response (Lucas et al 2015)

Many applications outside our group:

Applications

- Open-Ended Survey Response (Roberts et al 2014)
- Fatwas of Jihadi Clerics (Lucas et al 2015)
- Student Text in MOOCs (Reich et al 2015)
- Constitutional Moments (Stewart and Young)
- Chinese/Western Newswires (Roberts et al 2016)
- How Teams Make Forecasts (Horowitz et al 2017)
- Self-Censorship of Bloggers (Roberts et al 2017)
- Snowden Leak Response (Lucas et al 2015)

Many applications outside our group:

- Over 25 published articles in over 20 journals

Applications

- Open-Ended Survey Response (Roberts et al 2014)
- Fatwas of Jihadi Clerics (Lucas et al 2015)
- Student Text in MOOCs (Reich et al 2015)
- Constitutional Moments (Stewart and Young)
- Chinese/Western Newswires (Roberts et al 2016)
- How Teams Make Forecasts (Horowitz et al 2017)
- Self-Censorship of Bloggers (Roberts et al 2017)
- Snowden Leak Response (Lucas et al 2015)

Many applications outside our group:

- Over 25 published articles in over 20 journals
- Used in education, politics, climate science, sociology, law, statistics, public policy, transportation systems

Applications

- Open-Ended Survey Response (Roberts et al 2014)
- Fatwas of Jihadi Clerics (Lucas et al 2015)
- Student Text in MOOCs (Reich et al 2015)
- Constitutional Moments (Stewart and Young)
- Chinese/Western Newswires (Roberts et al 2016)
- How Teams Make Forecasts (Horowitz et al 2017)
- Self-Censorship of Bloggers (Roberts et al 2017)
- Snowden Leak Response (Lucas et al 2015)

Many applications outside our group:

- Over 25 published articles in over 20 journals
- Used in education, politics, climate science, sociology, law, statistics, public policy, transportation systems
- More on the way. . .

Conclusion

Conclusion

- Emerging opportunities for text analysis in the social sciences

Conclusion

- Emerging opportunities for text analysis in the social sciences
  - new models for text analysis using document context (`STM`)

Conclusion

- Emerging opportunities for text analysis in the social sciences
  - new models for text analysis using document context (`STM`)
  - open-source R package `stm`

Conclusion

- Emerging opportunities for text analysis in the social sciences
  - new models for text analysis using document context (STM)
  - open-source R package `stm`
    (auxiliary packages `stmCorrViz` and `stmBrowser`)

Conclusion

- Emerging opportunities for text analysis in the social sciences
  - new models for text analysis using document context (STM)
  - open-source R package `stm`
    (auxiliary packages `stmCorrViz` and `stmBrowser`)
  - other great packages for LDA in R (`mallet`, `topicmodels`, `lda`)

Conclusion

- Emerging opportunities for text analysis in the social sciences
  - new models for text analysis using document context (STM)
  - open-source R package stm
    (auxiliary packages stmCorrViz and stmBrowser)
  - other great packages for LDA in R (mallet, topicmodels, lda)
- Talk has necessarily skipped over many, many important details- be sure to read more!

Go try out the software today!

Suggested Reading

- Blei (2012) "Probabilistic Topic Models" *Transactions of the ACM*.
- Wallach, Mimno and McCallum (2009) "Rethinking LDA: Why Priors Matter" *NIPS*
- Grimmer and Stewart (2013) "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts" *Political Analysis*
- Roberts et al. (2014) "Structural topic models for open-ended survey responses" *American Journal of Political Science*
- Boyd-Graber, Mimno and Newman (2016) "Care and Feeding of Topic Models: Problems, Diagnostics and Improvements" in *Handbook of Mixed Membership Models*

For more information

BrandonStewart.org

structuraltopicmodel.com