American's (and Comparative's) Next Top(ic) Models

Kevin Munger

NYU

December 2, 2015

Topic Models

- These bad boys are getting super popular, and for good reason

Topic Models

- These bad boys are getting super popular, and for good reason
- Explosion in popularity $\rightarrow$ the potential for uninformed applications

Topic Models

- These bad boys are getting super popular, and for good reason
- Explosion in popularity $\rightarrow$ the potential for uninformed applications
- Done well and interpreted correctly, can be a valuable tool

Topic Models

- There is so much text in the world, usually divided into "documents"–papers, speeches, tweets

Topic Models

- There is so much text in the world, usually divided into "documents"–papers, speeches, tweets
- We'd like to summarize the information in these documents, so we use topic models to create topics and assign them to documents

Topic Models

- There is so much text in the world, usually divided into "documents"–papers, speeches, tweets
- We'd like to summarize the information in these documents, so we use topic models to create topics and assign them to documents
- This method is "unsupervised machine learning," and this reduces the role of researcher biases (more on this later)

Topic Models

- There is so much text in the world, usually divided into “documents”–papers, speeches, tweets
- We’d like to summarize the information in these documents, so we use topic models to create topics and assign them to documents
- This method is “unsupervised machine learning,” and this reduces the role of researcher biases (more on this later)
    - This means you don’t choose “topics”

Topic Models

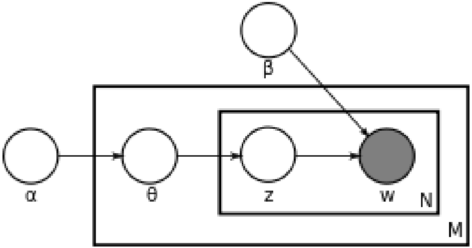
- There is so much text in the world, usually divided into "documents"–papers, speeches, tweets
- We'd like to summarize the information in these documents, so we use topic models to create topics and assign them to documents
- This method is "unsupervised machine learning," and this reduces the role of researcher biases (more on this later)
  - This means you don't choose "topics"
  - In fact, "topics" is sometimes begging the question

Topic Models

- There is so much text in the world, usually divided into "documents"–papers, speeches, tweets
- We'd like to summarize the information in these documents, so we use topic models to create topics and assign them to documents
- This method is "unsupervised machine learning," and this reduces the role of researcher biases (more on this later)
  - This means you don't choose "topics"
  - In fact, "topics" is sometimes begging the question
  - Also means you want a LOT of data

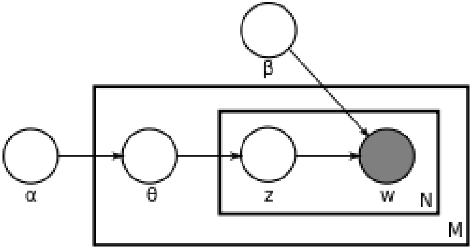
The Godfather–Latent Dirchlet Allocation

- Developed by David Blei, Andrew Ng and *the* Michael I. Jordan (Blei, Ng, and Jordan, 2003)
- Given only a number of topics, concentration parameter α and a collection of documents, produces a distribution of "topics" over those documents
- Does not incorporate covariates about the documents



The Godfather–Latent Dirchlet Allocation

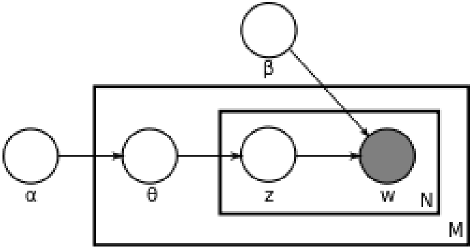
- Developed by David Blei, Andrew Ng and *the* Michael I. Jordan (Blei, Ng, and Jordan, 2003)
- Given only a number of topics, concentration parameter α and a collection of documents, produces a distribution of "topics" over those documents
- Does <u>not</u> incorporate covariates about the documents



- LDA is a <u>generative topic model</u>–the fundamental assumption is that each document is created via draws from some distribution

The Godfather–Latent Dirchlet Allocation

- Developed by David Blei, Andrew Ng and *the* Michael I. Jordan (Blei, Ng, and Jordan, 2003)
- Given only a number of topics, concentration parameter α and a collection of documents, produces a distribution of "topics" over those documents
- Does not incorporate covariates about the documents



- LDA is a generative topic model–the fundamental assumption is that each document is created via draws from some distribution
- With the caveat that word order doesn't matter–"bag of words"

The summary of how LDA works

From Barberà et al (2013)

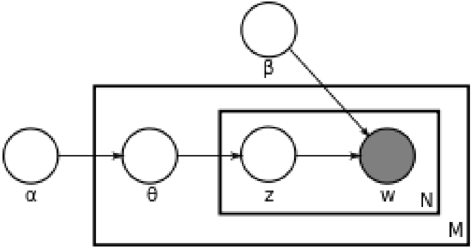
1. The topic distribution for document w is determined by: $\theta \sim \text{Dirichlet}(\alpha)$
2. The word distribution for topic k is determined by: $\beta \sim \text{Dirichlet}(\delta)$
3. For each of the words in document w
   (a) Choose a topic $z_n \sim \text{Multinomial}(\theta)$
   (b) Choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on z_n.

LDA

- Assumptions of LDA:

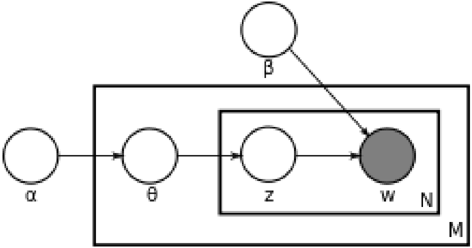
LDA

- Assumptions of LDA:



LDA

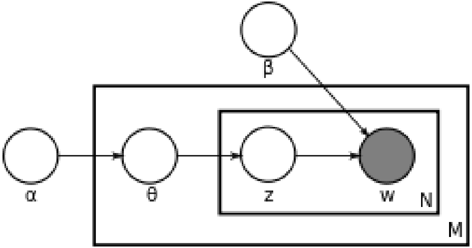
- Assumptions of LDA:



  - The prior distribution of words over topics (β) and topics over documents (α) is a symmetric Dirichlet

LDA

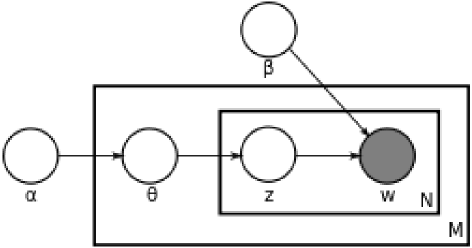
- Assumptions of LDA:



  - The prior distribution of words over topics (β) and topics over documents (α) is a symmetric Dirichlet
  - Topics are uncorrelated

LDA

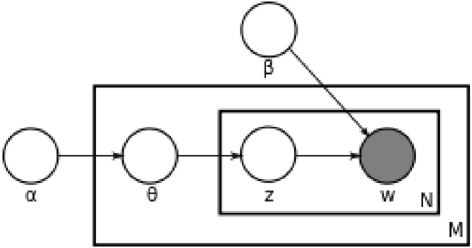
- Assumptions of LDA:



  - The prior distribution of words over topics (β) and topics over documents (α) is a symmetric Dirichlet
  - Topics are uncorrelated
  - Words are all exchangeable (each draw from the generative model is independent)

LDA

- Assumptions of LDA:



  - The prior distribution of words over topics (β) and topics over documents (α) is a symmetric Dirichlet
  - Topics are uncorrelated
  - Words are all exchangeable (each draw from the generative model is independent)
- As Grimmer and Stewart (2013) put it: "all quantitative models of language are wrong–but some are useful"

What does LDA actually do?

- The only observed variables are the words in the documents

What does LDA actually do?

- The only observed variables are the words in the documents
- We're interested in estimating two latent variables: the distribution of words over topics (β) and topics over documents (α)

What does LDA actually do?

- The only observed variables are the words in the documents
- We're interested in estimating two latent variables: the distribution of words over topics (β) and topics over documents (α)
- This was first done with variational inference, most later applications use Gibbs sampling

What does LDA actually do?

- The only observed variables are the words in the documents
- We're interested in estimating two latent variables: the distribution of words over topics (β) and topics over documents (α)
- This was first done with variational inference, most later applications use Gibbs sampling
- The intuition behind each of these methods is to optimize one variable while holding the others constant, iterating across all of the variables many times

From text to topics

- What are the concrete steps of taking documents and generating topics using LDA?

From text to topics

- What are the concrete steps of taking documents and generating topics using LDA?
- I'm going to give an overview before we actually do it

From text to topics

- What are the concrete steps of taking documents and generating topics using LDA?
- I'm going to give an overview before we actually do it
  - Turn each document into a Document-Term Matrix

From text to topics

- What are the concrete steps of taking documents and generating topics using LDA?
- I'm going to give an overview before we actually do it
  - Turn each document into a Document-Term Matrix
  - Remove stop-words

From text to topics

- What are the concrete steps of taking documents and generating topics using LDA?
- I'm going to give an overview before we actually do it
  - Turn each document into a Document-Term Matrix
  - Remove stop-words
  - Remove rare words

From text to topics

- What are the concrete steps of taking documents and generating topics using LDA?
- I'm going to give an overview before we actually do it
  - Turn each document into a Document-Term Matrix
  - Remove stop-words
  - Remove rare words
  - Stemming

From text to topics

- What are the concrete steps of taking documents and generating topics using LDA?
- I'm going to give an overview before we actually do it
  - Turn each document into a Document-Term Matrix
  - Remove stop-words
  - Remove rare words
  - Stemming
  - Choose the number of topics, K

From text to topics

- What are the concrete steps of taking documents and generating topics using LDA?
- I'm going to give an overview before we actually do it
  - Turn each document into a Document-Term Matrix
  - Remove stop-words
  - Remove rare words
  - Stemming
  - Choose the number of topics, K
    - ⋆ This is the hard part, we'll get back to it

From text to topics

- What are the concrete steps of taking documents and generating topics using LDA?
- I'm going to give an overview before we actually do it
  - Turn each document into a Document-Term Matrix
  - Remove stop-words
  - Remove rare words
  - Stemming
  - Choose the number of topics, K
    - This is the hard part, we'll get back to it
  - Choose α

From text to topics

- What are the concrete steps of taking documents and generating topics using LDA?
- I'm going to give an overview before we actually do it
  - Turn each document into a Document-Term Matrix
  - Remove stop-words
  - Remove rare words
  - Stemming
  - Choose the number of topics, K
    - This is the hard part, we'll get back to it
  - Choose α
    - For now, we're going to use a simple rule:

From text to topics

- What are the concrete steps of taking documents and generating topics using LDA?
- I'm going to give an overview before we actually do it
  - Turn each document into a Document-Term Matrix
  - Remove stop-words
  - Remove rare words
  - Stemming
  - Choose the number of topics, K
    - This is the hard part, we'll get back to it
  - Choose α
    - For now, we're going to use a simple rule:
    - $\alpha = \frac{50}{K}$

Let's give it a try

- Go to https://github.com/kmunger/Topic_Modelsand find the R file LDA.R
- We're going to walk through an example

Choosing K

- The current idea is generally to let the data tell you–choose K that maximizes the out-of-sample likelihood of the model

Choosing K

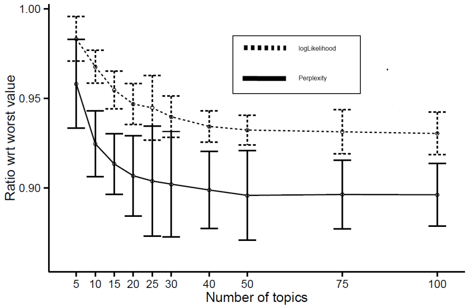
- The current idea is generally to let the data tell you–choose K that maximizes the out-of-sample likelihood of the model
  - Use *n*-fold cross-validation

Choosing K

- The current idea is generally to let the data tell you–choose K that maximizes the out-of-sample likelihood of the model
  - Use *n*-fold cross-validation
  - You can also use perplexity instead of likelihood, they often give similar results

Choosing K

- The current idea is generally to let the data tell you–choose K that maximizes the out-of-sample likelihood of the model
  - Use *n*-fold cross-validation
  - You can also use perplexity instead of likelihood, they often give similar results
  - An example from my work



Topic Stability, K and Multi-Modality

- One common application of LDA is to look at how topic use changes over time

Topic Stability, K and Multi-Modality

- One common application of LDA is to look at how topic use changes over time
- So you've run LDA and named your topics

Topic Stability, K and Multi-Modality

- One common application of LDA is to look at how topic use changes over time
- So you've run LDA and named your topics
- You can use a non-parametric method like loess to establish "significant" changes

Topic Stability, K and Multi-Modality

- One common application of LDA is to look at how topic use changes over time
- So you've run LDA and named your topics
- You can use a non-parametric method like loess to establish "significant" changes
- However, your results might not be robust to changing K, or even to re-running LDA with a different random seed!

Topic Stability, K and Multi-Modality

- The problem is that, because the LDA algorithm follows the gradient of the function, it can only ensure you find a <u>local</u> mode, of which there may be many

Topic Stability, K and Multi-Modality

- The problem is that, because the LDA algorithm follows the gradient of the function, it can only ensure you find a <u>local</u> mode, of which there may be many
- There is no guarantee of finding the <u>global</u> maximum

Topic Stability, K and Multi-Modality

- The problem is that, because the LDA algorithm follows the gradient of the function, it can only ensure you find a local mode, of which there may be many
- There is no guarantee of finding the global maximum
- The "topic" you found in one run might not exist in another!

Topic Stability, K and Multi-Modality

- The problem is that, because the LDA algorithm follows the gradient of the function, it can only ensure you find a local mode, of which there may be many
- There is no guarantee of finding the global maximum
- The "topic" you found in one run might not exist in another!
- Let's look at an example

Topic Stability, K and Multi-Modality

- This problem is discussed in Roberts, Stewart, and Tingley (2014)

Topic Stability, K and Multi-Modality

- This problem is discussed in Roberts, Stewart, and Tingley (2014)
- Even if knew we could find the global optimum of the function, that might not be the most useful choice for our question

Topic Stability, K and Multi-Modality

- This problem is discussed in Roberts, Stewart, and Tingley (2014)
- Even if knew we could find the global optimum of the function, that might not be the most useful choice for our question
- The objection function still matters, but "among locally optimal solutions model fit statistics provide a weak signal of model quality as judged by human analysts"

Topic Stability, K and Multi-Modality

- This problem is discussed in Roberts, Stewart, and Tingley (2014)
- Even if knew we could find the global optimum of the function, that might not be the most useful choice for our question
- The objection function still matters, but "among locally optimal solutions model fit statistics provide a weak signal of model quality as judged by human analysts"
- Two more useful criteria, not currently used often:

Topic Stability, K and Multi-Modality

- This problem is discussed in Roberts, Stewart, and Tingley (2014)
- Even if knew we could find the global optimum of the function, that might not be the most useful choice for our question
- The objection function still matters, but "among <u>locally optimal solutions</u> model fit statistics provide a weak signal of model quality as judged by human analysts"
- Two more useful criteria, not currently used often:
  - Semantic coherence: the tendency of a topic's high probability words to co-occur in the same document

Topic Stability, K and Multi-Modality

- This problem is discussed in Roberts, Stewart, and Tingley (2014)
- Even if knew we could find the global optimum of the function, that might not be the most useful choice for our question
- The objection function still matters, but "among <u>locally optimal solutions</u> model fit statistics provide a weak signal of model quality as judged by human analysts"
- Two more useful criteria, not currently used often:
  - Semantic coherence: the tendency of a topic's high probability words to co-occur in the same document
  - Exclusivity: if a high probability word is specific to a single topic

Topic Stability, K and Multi-Modality

- Does multi-modality matter? How would we know?

Topic Stability, K and Multi-Modality

- Does multi-modality matter? How would we know?
  - Global alignment: in different runs, let each topic find its best match and see the cosine similarity between the topic-word distributions in each match

Topic Stability, K and Multi-Modality

- Does multi-modality matter? How would we know?
  - Global alignment: in different runs, let each topic find its best match and see the cosine similarity between the topic-word distributions in each match
  - Pairwise similarity: how many of the top 10 terms from each topic are shared

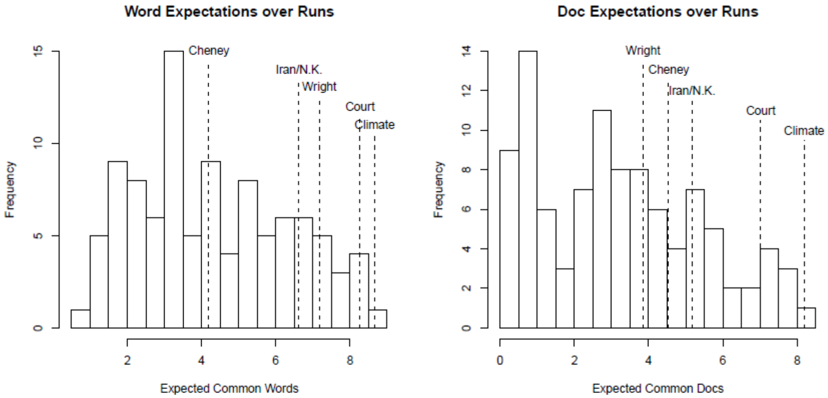
Topic Stability, K and Multi-Modality

- Does multi-modality matter? How would we know?
  - Global alignment: in different runs, let each topic find its best match and see the cosine similarity between the topic-word distributions in each match
  - Pairwise similarity: how many of the top 10 terms from each topic are shared
  - It turns out they're pretty highly correlated

Topic Stability, K and Multi-Modality

- Does multi-modality matter? How would we know?
  - Global alignment: in different runs, let each topic find its best match and see the cosine similarity between the topic-word distributions in each match
  - Pairwise similarity: how many of the top 10 terms from each topic are shared
  - It turns out they're pretty highly correlated
- These are open problems in the field, just to be aware of–but don't use LDA and focus on a single topic without being careful!

Topic Stability: Figure 4 from Roberts, Stewart, and Tingley (2014)

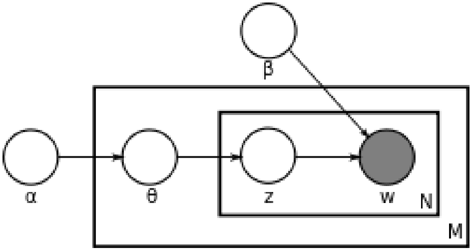


Let's give it a try–AGAIN

- Go to `https://github.com/kmunger/Topic_Models` and find the R file STM.R
- This model takes a while to run, so let's get it started before we talk about it

LDA–changing assumptions

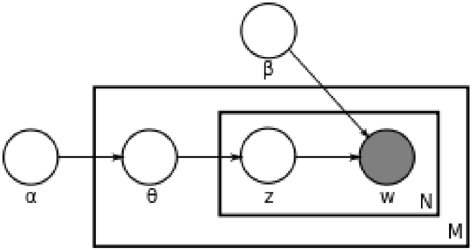
- Assumptions of LDA:



  - The prior distribution of words over topics (β) and topics over documents (α) is a symmetric Dirichlet
  - Topics are uncorrelated
  - Words are all exchangeable (each draw from the generative model is independent)

LDA–changing assumptions

- Assumptions of LDA:



  - The prior distribution of words over topics (β) and topics over documents (α) is a symmetric Dirichlet
  - Topics are ~~uncorrelated~~ potentially correlated
  - Words are all exchangeable (each draw from the generative model is independent)

Correlated Topic Model

- Developed by Blei and Lafferty (2006)

Correlated Topic Model

- Developed by Blei and Lafferty (2006)
- Logically it makes sense that knowing the prevalence of one topic in a document tells us something about the distribution over the other topics

Correlated Topic Model

- Developed by Blei and Lafferty (2006)
- Logically it makes sense that knowing the prevalence of one topic in a document tells us something about the distribution over the other topics
- Blei and Lafferty solve this by letting topics be correlated according to the logistic normal

Correlated Topic Model

- Developed by Blei and Lafferty (2006)
- Logically it makes sense that knowing the prevalence of one topic in a document tells us something about the distribution over the other topics
- Blei and Lafferty solve this by letting topics be correlated according to the logistic normal
  - Logistic normal isn't conjugate to the multinomial distribution, so this requires some additional variational inference

Correlated Topic Model

- Developed by Blei and Lafferty (2006)
- Logically it makes sense that knowing the prevalence of one topic in a document tells us something about the distribution over the other topics
- Blei and Lafferty solve this by letting topics be correlated according to the logistic normal
  - Logistic normal isn't conjugate to the multinomial distribution, so this requires some additional variational inference
- They show that CTM performs better than LDA on the original LDA corpus

Correlated Topic Model

- Developed by Blei and Lafferty (2006)
- Logically it makes sense that knowing the prevalence of one topic in a document tells us something about the distribution over the other topics
- Blei and Lafferty solve this by letting topics be correlated according to the logistic normal
  - Logistic normal isn't conjugate to the multinomial distribution, so this requires some additional variational inference
- They show that CTM performs better than LDA on the original LDA corpus
  - It supports more topics

Correlated Topic Model

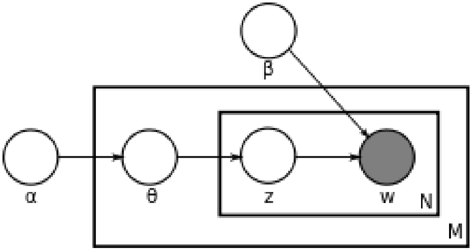
- Developed by Blei and Lafferty (2006)
- Logically it makes sense that knowing the prevalence of one topic in a document tells us something about the distribution over the other topics
- Blei and Lafferty solve this by letting topics be correlated according to the logistic normal
  - Logistic normal isn't conjugate to the multinomial distribution, so this requires some additional variational inference
- They show that CTM performs better than LDA on the original LDA corpus
  - It supports more topics
  - Also better perplexity scores on partially held-out documents

Correlated Topic Model

- Developed by Blei and Lafferty (2006)
- Logically it makes sense that knowing the prevalence of one topic in a document tells us something about the distribution over the other topics
- Blei and Lafferty solve this by letting topics be correlated according to the logistic normal
    - Logistic normal isn't conjugate to the multinomial distribution, so this requires some additional variational inference
- They show that CTM performs better than LDA on the original LDA corpus
    - It supports more topics
    - Also better perplexity scores on partially held-out documents
- They have a package on CRAN, and it's been integrated into later topic models

LDA–changing assumptions

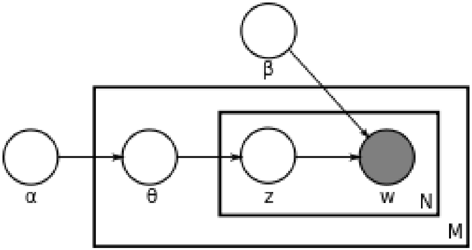
- Assumptions of LDA:



  - The prior distribution of words over topics (β) and topics over documents (α) is a symmetric Dirichlet
  - Topics are uncorrelated
  - Words are all exchangeable (each draw from the generative model is independent)

LDA–changing assumptions

- Assumptions of LDA:



  - The prior distribution of words over topics (β) and topics over documents (α) is a ~~symmetric~~ asymmetric Dirichlet
  - Topics are uncorrelated
  - Words are all exchangeable (each draw from the generative model is independent)

Priors Matter (Wallach, Mimno, and McCallum, 2009)

- The simplifying assumption of fixed symmetric Dirichlet priors is Bad

Priors Matter (Wallach, Mimno, and McCallum, 2009)

- The simplifying assumption of fixed symmetric Dirichlet priors is Bad
- With simulations, they find that relaxing this assumption leads to better, more stable topics

Priors Matter (Wallach, Mimno, and McCallum, 2009)

- The simplifying assumption of fixed symmetric Dirichlet priors is Bad
- With simulations, they find that relaxing this assumption leads to better, more stable topics
  - This is almost all due to making the topic-document prior α asymmetric Dirichlet; doing so for β provides little benefit

Priors Matter (Wallach, Mimno, and McCallum, 2009)

- The simplifying assumption of fixed symmetric Dirichlet priors is Bad
- With simulations, they find that relaxing this assumption leads to better, more stable topics
  - This is almost all due to making the topic-document prior α asymmetric Dirichlet; doing so for β provides little benefit
- Running the model with asymmetric priors is computationally expensive, but can be approximated by optimizing over them rather than integrating them out

Priors Matter (Wallach, Mimno, and McCallum, 2009)

- The simplifying assumption of fixed symmetric Dirichlet priors is Bad
- With simulations, they find that relaxing this assumption leads to better, more stable topics
  - This is almost all due to making the topic-document prior α asymmetric Dirichlet; doing so for β provides little benefit
- Running the model with asymmetric priors is computationally expensive, but can be approximated by optimizing over them rather than integrating them out
- Using this method, LDA is more robust to the number of topics you choose

Priors Matter (Wallach, Mimno, and McCallum, 2009)

- The simplifying assumption of fixed symmetric Dirichlet priors is Bad
- With simulations, they find that relaxing this assumption leads to better, more stable topics
  - This is almost all due to making the topic-document prior α asymmetric Dirichlet; doing so for β provides little benefit
- Running the model with asymmetric priors is computationally expensive, but can be approximated by optimizing over them rather than integrating them out
- Using this method, LDA is more robust to the number of topics you choose
  - Stop words are better allocated to specific topics, you can be more conservative with your list of stop words

Priors Matter (Wallach, Mimno, and McCallum, 2009)

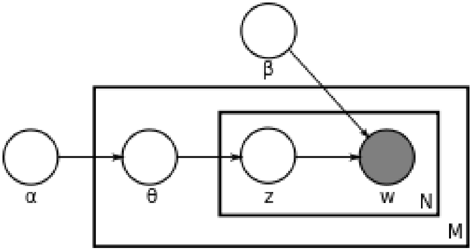
- The simplifying assumption of fixed symmetric Dirichlet priors is Bad
- With simulations, they find that relaxing this assumption leads to better, more stable topics
  - This is almost all due to making the topic-document prior α asymmetric Dirichlet; doing so for β provides little benefit
- Running the model with asymmetric priors is computationally expensive, but can be approximated by optimizing over them rather than integrating them out
- Using this method, LDA is more robust to the number of topics you choose
  - Stop words are better allocated to specific topics, you can be more conservative with your list of stop words
  - Adding more topics "nibbles away" at previous topics, rather than changing them in important ways

Priors Matter (Wallach, Mimno, and McCallum, 2009)

- The simplifying assumption of fixed symmetric Dirichlet priors is Bad
- With simulations, they find that relaxing this assumption leads to better, more stable topics
  - This is almost all due to making the topic-document prior α asymmetric Dirichlet; doing so for β provides little benefit
- Running the model with asymmetric priors is computationally expensive, but can be approximated by optimizing over them rather than integrating them out
- Using this method, LDA is more robust to the number of topics you choose
  - Stop words are better allocated to specific topics, you can be more conservative with your list of stop words
  - Adding more topics "nibbles away" at previous topics, rather than changing them in important ways
  - Thus, too many topics is better than too few

LDA–changing assumptions

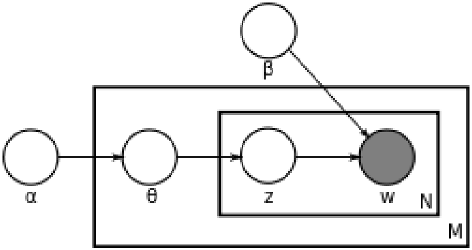
- Assumptions of LDA:



  - The prior distribution of words over topics (β) and topics over documents (α) is a symmetric Dirichlet
  - Topics are uncorrelated
  - Words are all exchangeable (each draw from the generative model is independent)

LDA–changing assumptions

- Assumptions of LDA:



  - The prior distribution of words over topics (β) and topics over documents (α) is a symmetric Dirichlet
  - Topics are uncorrelated
  - Words are all exchangeable (each draw from the generative model is independent) and let's take this seriously

Exchangability is hard–Lemmatization

- Stemming is cruder version of something called lemmatization

Exchangability is hard–Lemmatization

- Stemming is cruder version of something called lemmatization
- Lemmatization is a Natural Language Processing technique that looks at sentences to detect parts of speech and tense

Exchangability is hard–Lemmatization

- Stemming is cruder version of something called lemmatization
- Lemmatization is a Natural Language Processing technique that looks at sentences to detect parts of speech and tense
    - Time flies like an arrow.

Exchangability is hard–Lemmatization

- Stemming is cruder version of something called lemmatization
- Lemmatization is a Natural Language Processing technique that looks at sentences to detect parts of speech and tense
  - Time flies like an arrow.
  - Fruit flies like a banana.

Exchangability is hard–Lemmatization

- Stemming is cruder version of something called lemmatization
- Lemmatization is a Natural Language Processing technique that looks at sentences to detect parts of speech and tense
  - Time flies like an arrow.
  - Fruit flies like a banana.
  - The other words in the sentence change what “flies” means, and even if it’s a verb or a noun

Exchangability is hard–Lemmatization

- Stemming is cruder version of something called lemmatization
- Lemmatization is a Natural Language Processing technique that looks at sentences to detect parts of speech and tense
  - Time flies like an arrow.
  - Fruit flies like a banana.
  - The other words in the sentence change what "flies" means, and even if it's a verb or a noun
- This is akin to sentiment models that don't incorporate negation

Exchangability is hard–Lemmatization

- Stemming is cruder version of something called lemmatization
- Lemmatization is a Natural Language Processing technique that looks at sentences to detect parts of speech and tense
  - Time flies like an arrow.
  - Fruit flies like a banana.
  - The other words in the sentence change what "flies" means, and even if it's a verb or a noun
- This is akin to sentiment models that don't incorporate negation
- The problem is that NLP software is difficult to install and use, and often doesn't work well without hand-processing

Structural Topic Model

- Roberts et al. (2014) have developed the Structural Topic Model
- It allows both the <u>content</u> and the <u>prevalence</u> of topics to vary with covariates

Structural Topic Model

- Roberts et al. (2014) have developed the Structural Topic Model
- It allows both the <u>content</u> and the <u>prevalence</u> of topics to vary with covariates
  - Content can vary with a binary variable (eg Liberal v Conservative)

Structural Topic Model

- Roberts et al. (2014) have developed the Structural Topic Model
- It allows both the <u>content</u> and the <u>prevalence</u> of topics to vary with covariates
  - Content can vary with a binary variable (eg Liberal v Conservative)
  - Prevalence can vary with both categorical and continuous variables (eg time)

Structural Topic Model

- Roberts et al. (2014) have developed the Structural Topic Model
- It allows both the content and the prevalence of topics to vary with covariates
  - Content can vary with a binary variable (eg Liberal v Conservative)
  - Prevalence can vary with both categorical and continuous variables (eg time)
- The R package contains several useful techniques for model selection/results visualization

Structural Topic Model

- Roberts et al. (2014) have developed the Structural Topic Model
- It allows both the content and the prevalence of topics to vary with covariates
  - Content can vary with a binary variable (eg Liberal v Conservative)
  - Prevalence can vary with both categorical and continuous variables (eg time)
- The R package contains several useful techniques for model selection/results visualization
- It *also* has a strong addition to the problem of multimodality–spectral initialization

Spectral Initialization

- Very technical application of the spectral theorem

Spectral Initialization

- Very technical application of the spectral theorem
- Intuition is <u>separability</u>: for each topic, there is at least one "anchor term" assigned only to that topic

Spectral Initialization

- Very technical application of the spectral theorem
- Intuition is separability: for each topic, there is at least one "anchor term" assigned only to that topic
  - All of the other terms of the β matrix are a combination of these anchor terms

Spectral Initialization

- Very technical application of the spectral theorem
- Intuition is separability: for each topic, there is at least one "anchor term" assigned only to that topic
  - All of the other terms of the β matrix are a combination of these anchor terms
  - Result is deterministic, and thus independent of the starting value

Spectral Initialization

- Very technical application of the spectral theorem
- Intuition is separability: for each topic, there is at least one "anchor term" assigned only to that topic
  - All of the other terms of the β matrix are a combination of these anchor terms
  - Result is deterministic, and thus independent of the starting value
- Caveats
  - Works best on large document sets

Spectral Initialization

- Very technical application of the spectral theorem
- Intuition is separability: for each topic, there is at least one "anchor term" assigned only to that topic
  - All of the other terms of the β matrix are a combination of these anchor terms
  - Result is deterministic, and thus independent of the starting value
- Caveats
  - Works best on large document sets
  - Eats up memory–can't be used for vocabularies over 10,000 words

Other Topic Models

- If you have a specific task in mind, it can be best to design your own topic model

Other Topic Models

- If you have a specific task in mind, it can be best to design your own topic model
- Obviously, this is harder than downloading an R package

Other Topic Models

- If you have a specific task in mind, it can be best to design your own topic model
- Obviously, this is harder than downloading an R package
- There are specific applications for which people have designed excellent topic models; don't reinvent the wheel

Other Topic Models–Quinn et al. (2010)

- The goal is to assign political speeches to a single topic

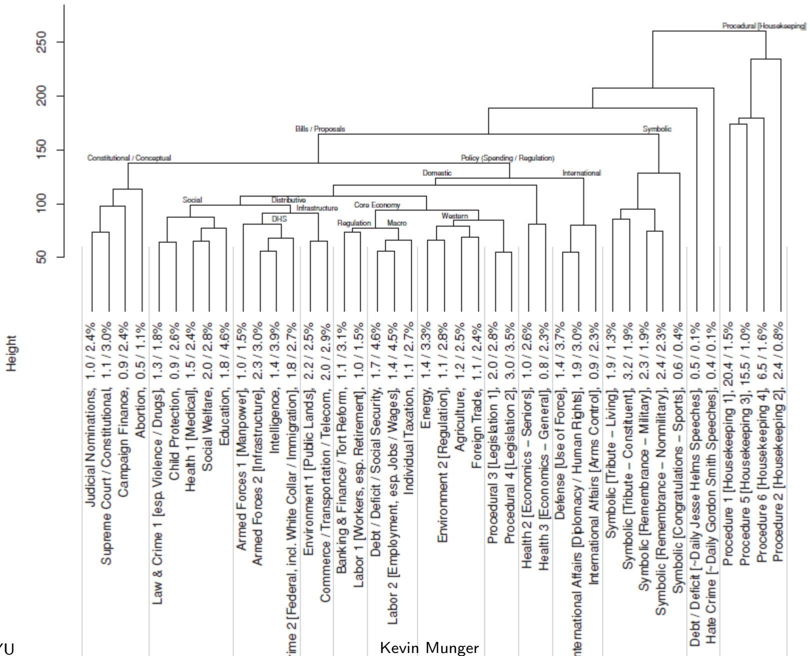
Other Topic Models–Quinn et al. (2010)

- The goal is to assign political speeches to a single topic
- They take the process of topic validation very seriously

Other Topic Models–Quinn et al. (2010)

- The goal is to assign political speeches to a single topic
- They take the process of topic validation very seriously
- They also find that topics are clustered hierarchically

FIGURE 1 Agglomerative Clustering of 42-Topic Model



Other Topic Models–Grimmer (2010)

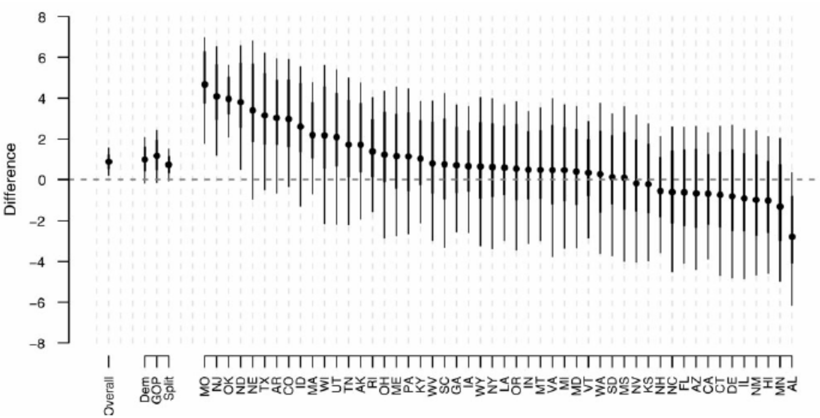
- The goal is to measure "Expressed Agendas in Senate Press Releases"

Other Topic Models–Grimmer (2010)

- The goal is to measure "Expressed Agendas in Senate Press Releases"
- The crucial difference in the model is that documents are clustered at the author level

Other Topic Models–Grimmer (2010)

- The goal is to measure "Expressed Agendas in Senate Press Releases"
- The crucial difference in the model is that documents are clustered at the author level
- Also excellent work in the paper validating the results


Difference
8
6
4
2
0
−2
−4
−6
−8
Overall
Dem
GOP
Split
MO NJ OK ND NE TX AR CO ID MA WI UT TN AK RI OH ME PA KY WV SC GA IA WY NY LA OR IN MT VA MI MD VT WA SD MS NV KS NH NC FL AZ CA CT DE IL NM HI MN AL

Thanks!

km2713@nyu.edu
@kmmunger

Blei, David, and John Lafferty. 2006. "Correlated topic models." Advances in neural information processing systems 18: 147.

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. "Latent dirichlet allocation." the Journal of machine Learning research 3: 993–1022.

Grimmer, Justin. 2010. "A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases." Political Analysis 18 (1): 1–35.

Grimmer, Justin, and Brandon M Stewart. 2013. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." Political Analysis p. mps028.

Kim, In Song, John Londregan, and Marc Ratkovic. 2014. "Voting, Speechmaking, and the Dimensions of Conflict in the US Senate.".

Quinn, Kevin M, Burt L Monroe, Michael Colaresi, Michael H Crespin, and Dragomir R Radev. 2010. "How to analyze political attention with minimal assumptions and costs." American Journal of Political Science 54 (1): 209–228.

Roberts, Margaret E, Brandon M Stewart, and Dustin Tingley. 2014.

"Navigating the local modes of big data: The case of topic models.".

Roberts, Margaret E, Brandon M Stewart, Dustin Tingley, and Maintainer Brandon Stewart. 2014. "Package stm.".

Wallach, Hanna M, David M Mimno, and Andrew McCallum. 2009. Rethinking LDA: Why priors matter. In Advances in neural information processing systems. pp. 1973–1981.