Day 11: Topic Models

ME314: Introduction to Data Science and Machine Learning

Jack Blumenau

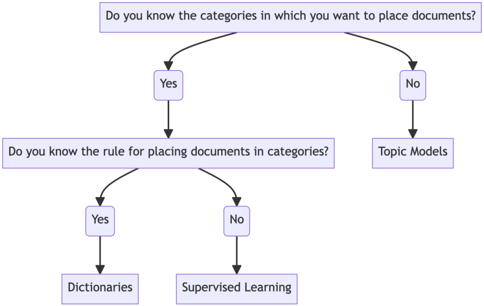
26th July 2023

Day 11 Outline

Topic Models

Topic Models

- Topic models allow us to cluster similar documents in a corpus together.
- Wait. Don't we already have tools for that?
- Yes! Dictionaries and supervised learning.
- So what do topic models add?



Topic Models

- Topic models offer an automated procedure for discovering the main "themes" in an unstructured corpus
- They require no prior information, training set, or labelling of texts before estimation
- They allow us to automatically organise, understand, and summarise large archives of text data.
- Latent Dirichlet Allocation (LDA) is the most common approach (Blei et al., 2003), and one that underpins more complex models
- Topic models are an example of *mixture* models:
    - Documents can contain multiple topics
    - Words can belong to multiple topics

Topic Models as Language Models

- Yesterday, we introduced the idea of a *probabilistic language model*
  - These models describe a story about how documents are generated using probability
- A language model is represented by a probability distribution over words in a vocabulary
- The Naive Bayes text classification model is *one* example of a generative language model where
  - We estimate separate probability distributions for each category of interest
  - Each document is assigned to a single category
- Topic models are also language models
  - We estimate separate probability distributions for each topic
  - Each document is described as belonging to *multiple* topics

What is a “topic”?

A “topic” is a probability distribution over a fixed word vocabulary.

- Consider a vocabulary: gene, dna, genetic, data, number, computer
- When speaking about **genetics**, you will:
  - frequently use the words “gene”, “dna” & “genetic”
  - infrequently use the words “data”, “number” & “computer”
- When speaking about **computation**, you will:
  - frequently use the words “data”, “number” & “computation”
  - infrequently use the words “gene”, “dna” & “genetic”

Topic	gene	dna	genetic	data	number	computer
Genetics	0.4	0.25	0.3	0.02	0.02	0.01
Computation	0.02	0.01	0.02	0.3	0.4	0.25

Note that no word has probability of exactly 0 under either topic.

Documents as mixtures of topics

- In a topic model, each document is described as being composed of a **mixture** of corpus-wide topics
- For each document, we find the topic proportions that maximize the probability that we would observe the words in that particular document

Imagine we have two documents with the following word counts

Table 2: Document word counts

Doc	gene	dna	genetic	data	number	computer
A	2	3	1	3	2	1
B	2	4	2	1	2	1

Table 3: Topic distributions

Topic	gene	dna	genetic	data	numb
Genetics	0.4	0.25	0.3	0.02	0.02
Computation	0.02	0.01	0.02	0.3	0.4

Implication: Our documents may be better described in terms of *mixtures* of different topics than by one topic alone.

A topic model simultaneously estimates two sets of probabilities

1. The probability of observing each word for each topic
2. The probability of observing each topic in each document

These quantities can then be used to organise documents by topic, assess how topics vary across documents, etc.

A motivating example

- Data: UK House of Commons' debates (PMQs)
    - ≈ 30000 parliamentary speeches from 1997 to 2015
    - ≈ 3000 unique words
    - $\approx 2m$ total words
- Sample/feature selection decisions
    - Sample selection: Only PMQs ($\approx 3\%$ of total speeches)
    - Feature selection: Removed frequently occurring & very rare words
    - Feature selection: All words have been "stemmed"
- Results of a 30-topic model

Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA)

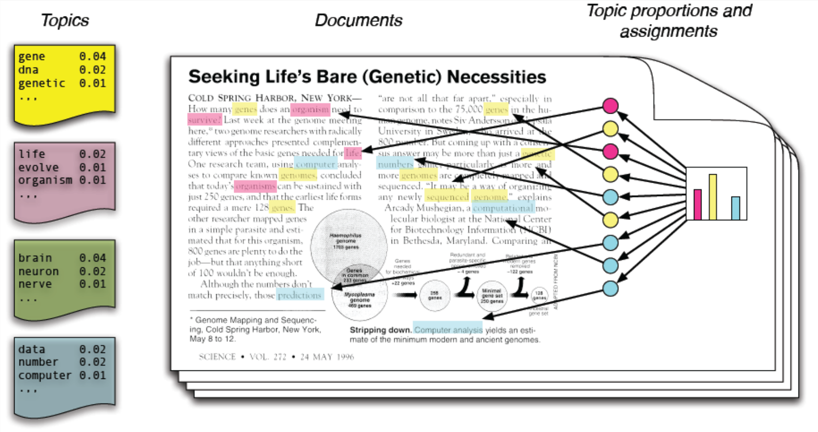
[PDF] **Latent dirichlet allocation**

DM Blei, AY Ng, MI Jordan - Journal of machine Learning research, 2003 - jmlr.org

We describe **latent Dirichlet allocation** (LDA), a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in …

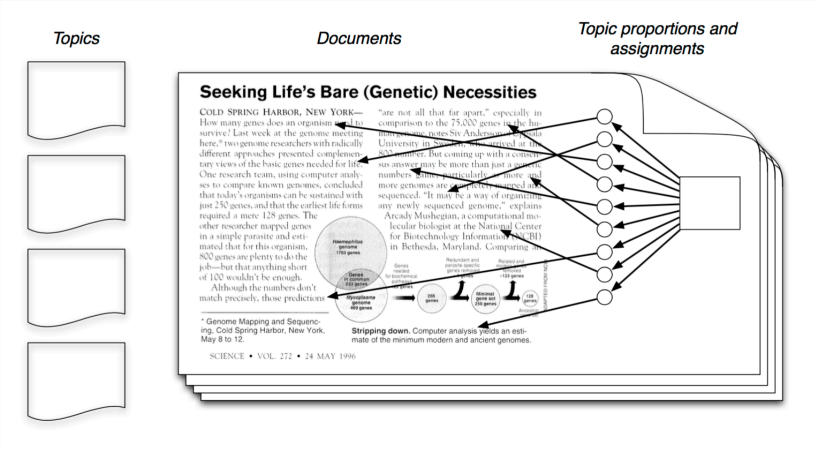
☆ Save Cite Cited by 43350 Related articles All 97 versions Web of Science: 16980

Latent Dirichlet Allocation (LDA)



- The researcher picks a number of topics, K.
- Each *topic* (k) is a distribution over words
- Each *document* (d) is a mixture of corpus-wide topics
- Each *word* (j) is drawn from one of those topics

Latent Dirichlet Allocation (LDA)



- In reality, we only observe the documents
- The other structure are hidden variables

- Our goal is to **infer** the hidden variables
- I.e., compute their distribution conditioned on the documents

Latent Dirichlet Allocation (LDA)

- The LDA model is a Bayesian mixture model for discrete data which describes how the documents in a dataset were created
- The number of topics, K, is selected by the researcher
- Each of the K *topics* is a probability distribution over a fixed vocabulary of J words
- Each of the D *documents* is a probability distribution over the K topics
- Each word in each document is drawn from a multinomial distribution specific to a particular topic
- Inference consists of estimating a posterior distribution over the parameters of the probability model from a combination of what is observed (words in documents) and what is hidden (topic and word parameters)

Probability Distributions Review

- A probability distribution is a function that gives the probabilities of the occurrence of different possible outcomes for a random variable
- Different parameter values change the distribution's shape and describe the probabilities of the different events
  - E.g. In a normal distribution, μ describes the mean and σ^2 describes the variance
- The notation "$\sim$" means to "draw" from the distribution
  - E.g. $x \sim N(0, 1)$ means to draw one value from a standard normal, which might result in $X = 1.123$
- There are two key distributions that we need to know about to understand topic models: the Multinomial and the Dirichlet distributions

Multinomial Distribution

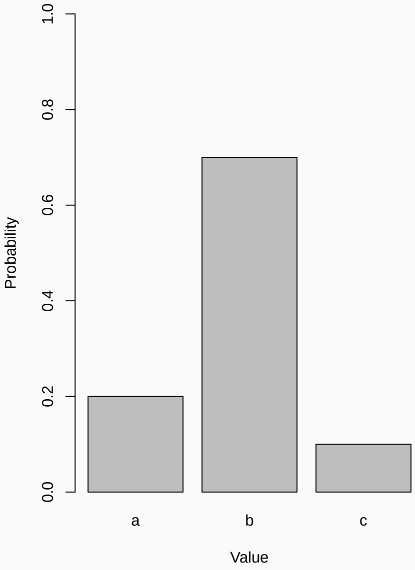
- The multinomial distribution describes the results of a random variable that can take on K possible categories
- The multinomial distribution depicted has probabilities $[0.2, 0.7, 0.1]$
- A draw (of size one) from a multinomial distribution returns one of the categories of the distribution

$c \sim$ Multinom$(1, [0.2, 0.7, 0.1])$ might return $c = 2$

- A larger draw returns several categories of the distribution in proportion to their probabilities

$C \sim$ Multinom$(10, [0.2, 0.7, 0.1])$ might return $c_1 = 2, c_2 = 7, c_3 = 1$

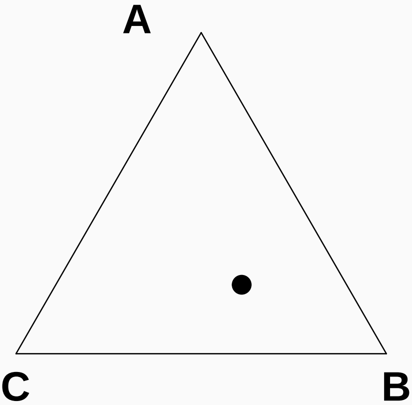
- Naive Bayes uses the multinomial distribution



- The Dirichlet distribution is a distribution over the simplex, i.e., positive vectors that sum to one
- A draw from a dirichlet distribution returns a vector if positive numbers that sum to one

$b \sim$ Dirichlet(α) might return $b = [0.2, 0.7, 0.1]$

- In other words, we can think of draws from a Dirichlet distribution being themselves multinomial distributions
- The parameter α controls the mean shape and sparsity of the multinomials (more on this later).



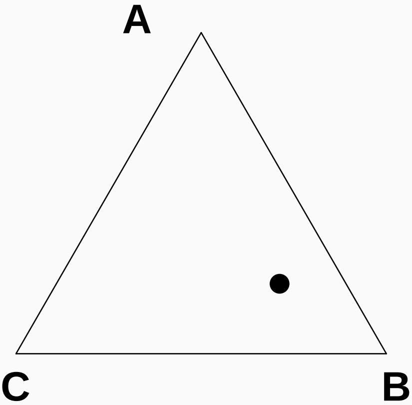

```
## p = [0.21,0.5,0.28]
```

Dirichlet Distribution

- The Dirichlet distribution is a distribution over the simplex, i.e., positive vectors that sum to one
- A draw from a dirichlet distribution returns a vector if positive numbers that sum to one

$b \sim$ Dirichlet(α) might return $b = [0.2, 0.7, 0.1]$

- In other words, we can think of draws from a Dirichlet distribution being themselves multinomial distributions
- The parameter α controls the mean shape and sparsity of the multinomials (more on this later).



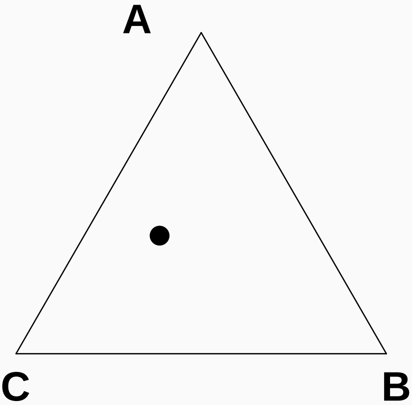

```
## p = [0.22,0.6,0.18]
```

Dirichlet Distribution

- The Dirichlet distribution is a distribution over the simplex, i.e., positive vectors that sum to one
- A draw from a dirichlet distribution returns a vector if positive numbers that sum to one

$b \sim$ Dirichlet(α) might return $b = [0.2, 0.7, 0.1]$

- In other words, we can think of draws from a Dirichlet distribution being themselves multinomial distributions
- The parameter α controls the mean shape and sparsity of the multinomials (more on this later).

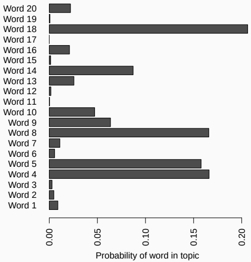


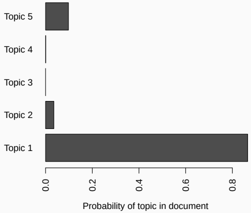

```
## p = [0.37,0.2,0.43]
```

LDA Generative Process

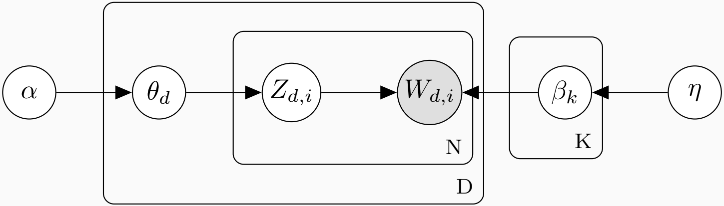
LDA assumes a generative process for documents:

1. For each *topic*, draw a probability distribution over words
   - $\beta_k \sim \text{Dirichlet}(\eta)$
   - $\beta_k \in \{0, 1\}$ and $\sum_{j=1}^{J} \beta_{j,k} = 1$
   - $\rightarrow$ prob. that word j occurs in topic k
2. For each *document*, draw a probability distribution over topics
   - $\theta_d \sim \text{Dirichlet}(\alpha)$
   - $\theta_{d,k} \in \{0, 1\}$ and $\sum_{k=1}^{K} \theta_{d,k} = 1$
   - $\rightarrow$ prob that topic k occurs in document d
3. For each *word* in each document
   - Draw one of K topics from step 2 (θ_d)
     - $z_i \sim \text{Multinomial}(\theta_d)$
     - $\rightarrow$ topic indicator of word i
   - Draw one of J words from step 1 (β_k)
     - $w_i \sim \text{Multinomial}(\beta_{z_i})$
     - $\rightarrow$ actual word of word i





LDA as a graphical model

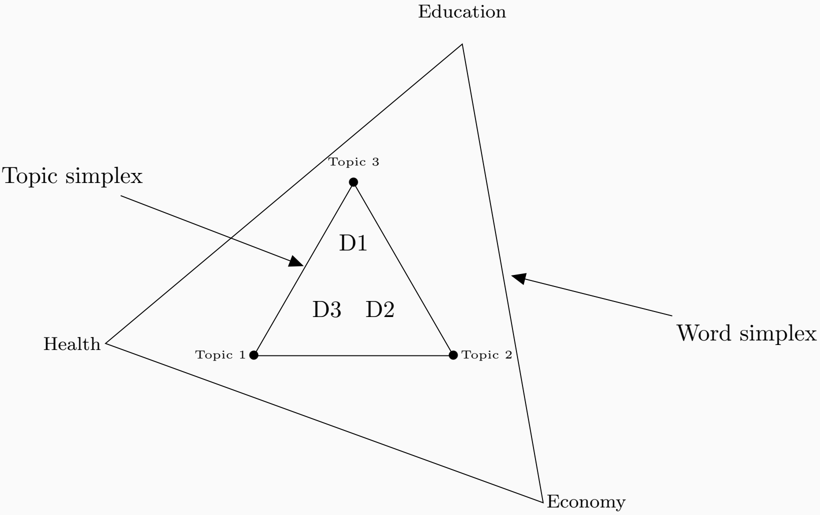


- Nodes are random variables; edges indicate dependence.
- Shaded nodes are observed; unshaded nodes are hidden.
- Plates indicate replicated variables.

The Dirichlet distribution

- The Dirichlet is used twice in LDA:
  - The topics (β_k) are a J dimensional Dirichlet (topics are a probability distribution over words)
  - The topic proportions (θ_d) are a K dimensional Dirichlet (documents are a probability distribution over topics)
- The parameter α (or η) controls the sparsity of the draws from the Dirichlet distribution.
  - When α is larger, the probabilities will be more evenly spread across categories
  - When α is smaller, more probability mass will be allocated to particular categories

Latent Dirichlet allocation (LDA)



Why does LDA "work"?

- LDA trades off two goals.
  1. For each document, allocate its words to as few topics as possible. (α)
  2. For each topic, assign high probability to as few terms as possible. (η)
- These goals are at odds.
  1. Putting a document in a single topic makes (2) hard: All of its words must have probability under that topic.
  2. Putting very few words in each topic makes (1) hard: To cover a document's words, it must assign many topics to it.
- Trading off these goals finds groups of tightly co-occurring words.

LDA Estimation

- Assuming the documents have been generated in such a way makes it possible to back out the shares of topics within documents and the share of words within topics
- Estimation of the LDA model is done in a Bayesian framework
- Our $Dir(\alpha)$ and $Dir(\eta)$ are the prior distributions of the θ_d and β_k
- We combine our data and model using Bayes' rule to update these prior distributions to obtain a posterior distribution for each θ_d and β_k
- The means of these posterior distributions are the outputs of statistical packages and which we use to investigate the θ_d and β_k
- Estimation is performed using either collapsed Gibbs sampling or variational methods
    - See Blei, 2012 for more details
- Fortunately, for us these are easily implemented in R

Imagine we have $D = 1000$ documents, $J = 10,000$ words, and $K = 3$ topics.

The key outputs of the topic model are the β and θ matrices:

$$\theta = \underbrace{\begin{pmatrix} \theta_{1,1} & \theta_{1,2} & \theta_{1,3} \\ \theta_{2,1} & \theta_{2,2} & \theta_{2,3} \\ \dots & \dots & \dots \\ \theta_{D,1} & \theta_{D,2} & \theta_{D,3} \end{pmatrix}}_{D \times K} = \underbrace{\begin{pmatrix} 0.7 & 0.2 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ \dots & \dots & \dots \\ 0.3 & 0.3 & 0.4 \end{pmatrix}}_{1000 \times 3}$$

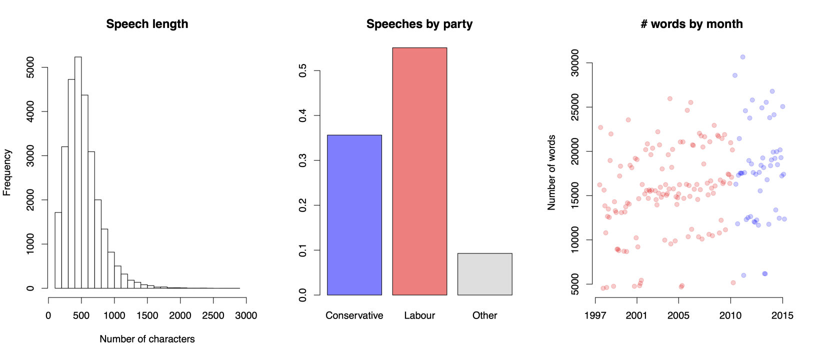
$$\beta = \underbrace{\begin{pmatrix} \beta_{1,1} & \beta_{1,2} & \dots & \beta_{1,J} \\ \beta_{2,1} & \beta_{2,2} & \dots & \beta_{2,J} \\ \beta_{3,1} & \beta_{3,2} & \dots & \beta_{3,J} \end{pmatrix}}_{K \times J} = \underbrace{\begin{pmatrix} 0.04 & 0.0001 & \dots & 0.003 \\ 0.0004 & 0.001 & \dots & 0.00005 \\ 0.002 & 0.0003 & \dots & 0.0008 \end{pmatrix}}_{3 \times 10,000}$$

- Data: UK House of Commons' debates (PMQs)
  - ≈ 30000 parliamentary speeches from 1997 to 2015
  - ≈ 3000 unique words
  - $\approx 2m$ total words

```
## Rows: 27,885
## Columns: 4
## $ name         <chr> "Ian Bruce", "Tony Blair", "Denis MacShane", "Tony Blair"~
## $ party        <chr> "Conservative", "Labour", "Labour", "Labour", "Liberal De~
## $ constituency <chr> "South Dorset", "Sedgefield", "Rotherham", "Sedgefield", ~
## $ body         <chr> "In a written answer, the Treasury has just it made clear~
```

- Estimate a range of topic models ($K \in \{20, 30, ..., 100\}$) using the `topicmodels` package

LDA example



Speech length
Frequency
0
1000
2000
3000
4000
5000
0
500
1000
1500
2000
2500
3000
Number of characters
Speeches by party
0.0
0.1
0.2
0.3
0.4
0.5
Conservative
Labour
Other
words by month
Number of words
5000
10000
15000
20000
25000
30000
1997
2001
2005
2010
2015

Implementation in R

```r
library(quanteda)
library(topicmodels)

## Create corpus
pmq_corpus <- pmq %>%
  corpus(text_field = "body")

pmq_dfm <- pmq_corpus %>%
  tokens(remove_punct = TRUE) %>%
  dfm() %>%
  dfm_remove(stopwords("en")) %>%
  dfm_wordstem() %>%
  dfm_trim(min_termfreq = 5)

## Convert for usage in 'topicmodels' package
pmq_tm_dfm <- pmq_dfm %>%
  convert(to = 'topicmodels')
```

Implementation in R

```r
## Estimate LDA
ldaOut <- LDA(pmq_tm_dfm, k = 60)

save(ldaOut, file = "ldaOut_60.Rdata")
```

LDA example

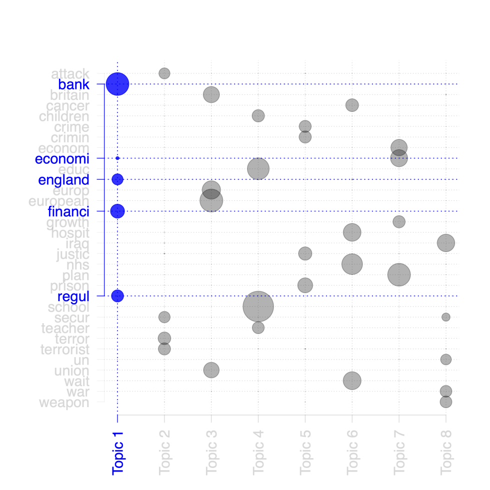
We will make use of the following score to visualise the posterior topics:

$$\text{term-score}_{k,v} = \hat{\beta}_{k,v} \log \left(\frac{\hat{\beta}_{k,v}}{\left(\prod_{j=1}^{K} \hat{\beta}_{j,v}\right)^{\frac{1}{K}}} \right)$$

This formulation is similar to the TFIDF term score, where

- The first term, $\hat{\beta}_{k,v}$, is the probability of term v in topic k and is akin to the term frequency
- The second term is akin to the document frequency (i.e. it down-weights terms that have high probability under all topics)

LDA example


attack
bank
britain
cancer
children
crime
crimin
econom
economi
educ
england
europ
european
financi
growth
hospit
iraq
justic
nhs
plan
prison
regul
school
secur
teacher
terror
terrorist
un
union
wait
war
weapon
Topic 1
Topic 2
Topic 3
Topic 4
Topic 5
Topic 6
Topic 7
Topic 8

LDA example

Topic 1	Topic 2	Topic 3	Topic 4
bank	terror	european	school
financi	terrorist	europ	educ
regul	secur	britain	children
england	attack	union	teacher
crisi	protect	british	pupil
fiscal	agre	referendum	class
market	act	constitut	parent

Topic 5	Topic 6	Topic 7	Topic 8
prison	nhs	plan	iraq
justic	wait	economi	weapon
crimin	hospit	econom	war
crime	cancer	growth	un
releas	patient	grow	resolut
court	list	longterm	iraqi
sentenc	health	deliv	saddam

Top Document by Topic

tax.pay.cut.incom.wage.minimum.rate (27%)

There are now 1.3 million workers in Britain who have benefited thanks to the minimum wage. The lowest paid have had their incomes increased by 1,500 a year as a result of the minimum wage. We are proud that this party introduced the minimum wage. We remember being told by some that introducing a minimum wage would cost 1 million jobs. In fact, we have managed to introduce the minimum wage and gain 1 million jobs.

Advantages and Disadvantages of LDA

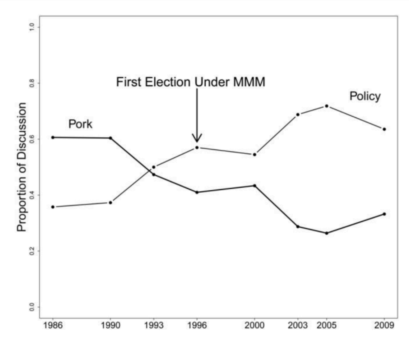
Advantages

- Automatically finds substantively interesting collections of words
- Automatically labels documents in “meaningful” ways
- Easily scaled to large corpora (millions of documents)
- Requires very little prior work (no manual labelling of texts/dictionary construction etc)

Disadvantages

- Generated topics may not reflect substantive interest of researcher
- Many estimated topics may be redundant for research question
- Requires extensive post-hoc interpretation of topics
- Sensitivity to number of topics selected (what is the best choice for K?)

LDA Example (Catalinac, 2014)



- **Research question:** Do different electoral systems create incentives for politicians to focus on different aspects of policy?
- **Theory**: PR electoral reform in 1994 in Japan should increase the amount of attention that politicians devote to "policy" rather than "pork".
- **Conclusion**: "Applying probabilistic topic modeling... shows that candidates for office change tried-and-true electoral strategies when confronted with an electoral reform."

LDA Example (Catalinac, 2014)

Questions:

- LDA on 8000 manifestos
  - Are entire manifestos the appropriate unit of analysis? Would sections, or paragraphs, be more appropriate?
- $K = 69$
  - "We fit the model with 69 topics because this was one of the lowest specifications that produced topics that were fine-grained enough to resemble our quantities of interest."
  - Are 69 topics the appropriate number?
- Is this a good case for topic models? We know the categories of interest ex ante
  - Why not use a dictionary approach here? Or supervised learning?

We will discuss strategies for addressing some of these after the break.

Break

LDA Extensions

Extending LDA

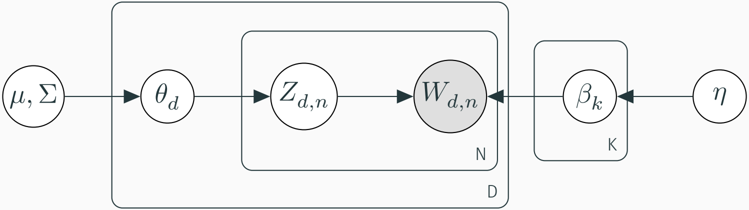
- LDA can be **embedded in more complicated models**, embodying further intuitions about the structure of the texts.
    - E.g., it can be used in models that account for syntax, authorship, word sense, dynamics, correlation, hierarchies, and other structure.
- The **data generating distribution** can be changed. We can apply mixed-membership assumptions to many kinds of data.
    - E.g., we can build models of images, social networks, music, purchase histories, computer code, genetic data, and other types.
- The **posterior** can be used in creative ways.
    - E.g., we can use inferences in information retrieval, recommendation, similarity, visualization, summarization, and other applications.

LDA Extensions

1. **Correlated Topic Model (CTM)**
    - LDA assumes that topics are uncorrelated across the corpus
    - The correlated topic model allows topics to be correlated
    - Closer approximation to true document structure, but estimation is slower
2. **Dynamic Topic Model (DTM)**
    - LDA assumes that topics are fixed across documents
    - In some settings, we have documents from many different time periods
    - The assumption that topics are fixed may not be sensible
    - The dynamic topic model allows topical content to vary smoothly over time
3. **Structural Topic Model (STM)**
    - Social scientists are typically interested in how topics vary with covariates
    - The structural topic model incorporates covariates into the LDA model
    - When estimated without covariates, the STM is the same as the CTM

Correlated Topic Model

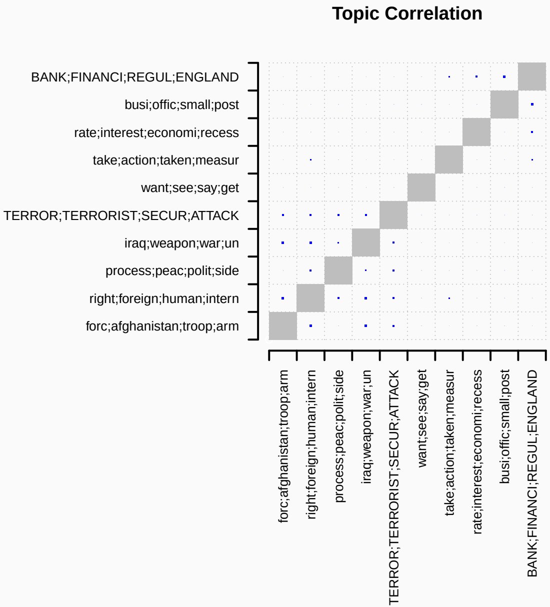
- The Dirichlet is a distribution on the simplex (positive vectors that sum to 1).
- It assumes that components are nearly independent.
- In real data, an article about fossil fuels is more likely to also be about geology than about genetics.
- The logistic normal is a distribution on the simplex that can model dependence between components.
- Amend the model so that the logit transformation of the topic-proportion parameters are drawn from a multivariate normal distribution



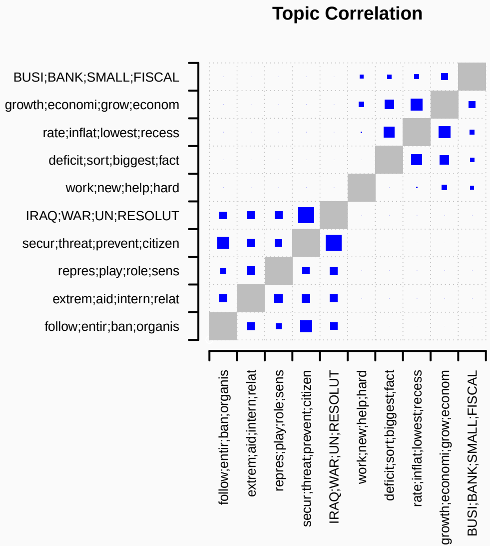
where the first node is logistic normal prior.

- Draw topic proportions from a logistic normal.
- This allows topic occurrences to exhibit correlation.
- Provides a “map’’ of topics and how they are related
- Provides a better fit to text data, but computation is more complex

LDA topic correlation



CTM topic correlation



CTM pros and cons

Advantages:

1. Probably a more reasonable approximation of the “true” data generating process of documents
2. Possible that correlations between topics might be a quantity of interest
3. CTM tends to have better statistical fit to data than LDA

Disadvantages:

1. CTM is somewhat more computationally demanding than LDA
2. CTM tends to have lower topic interpretability than LDA

Dynamic Topic Model

- LDA assumes that the order of documents does not matter.
- Not appropriate for sequential corpora (e.g., that span hundreds of years)
- We may want to track how language changes over time.
  - How has the language used to describe neuroscience developed from "The Brain of Professor Laborde" (1903) to "Reshaping the Cortical Motor Map by Unmasking Latent Intracortical Connections" (1991)
  - How has the language used to describe love developed from "Pride and Prejudice" (1813) to "Eat, Pray, Love" (2006)
- Dynamic topic models let the topics drift in a sequence.

Dynamic Topic Model

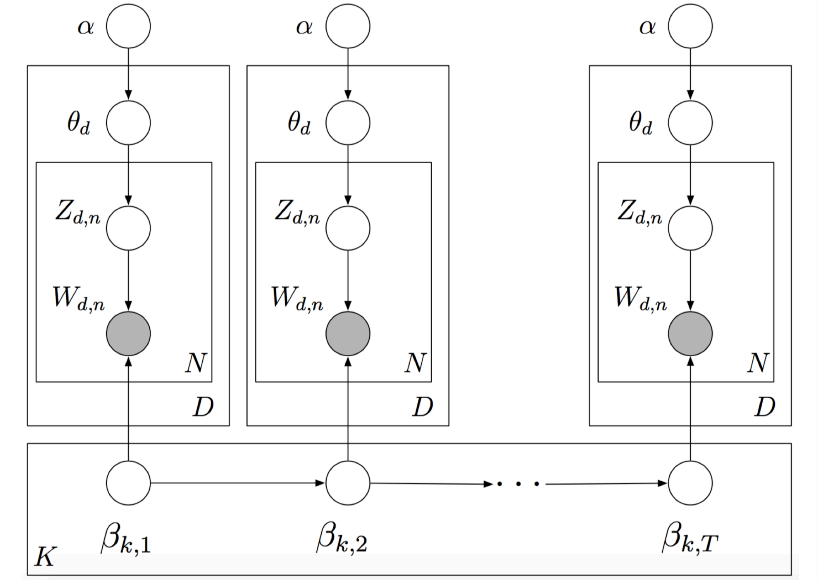
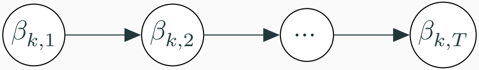


Plate (K) allows topics to "drift" through time.

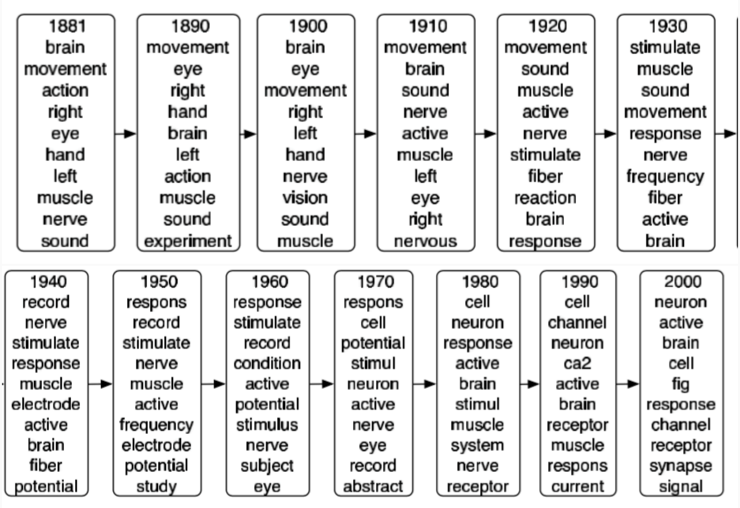
Dynamic Topic Models



- Use a logistic normal distribution to model topics evolving over time.
    - The kth topic at time 2 has evolved smoothly from the kth topic at time 1
- As for CTMs, this makes computation more complex. But it lets us make inferences about sequences of documents.

Dynamic Topic Model Example (Mimno and Lafferty, 2006)

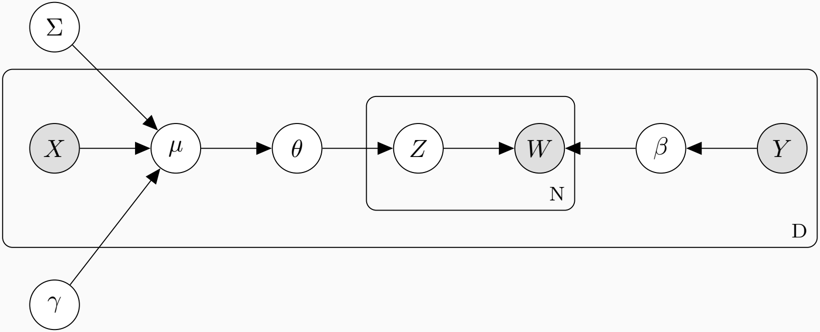
“Neuroscience” topic based on DTM of 30,000 articles from *Science*



Structural Topic Model

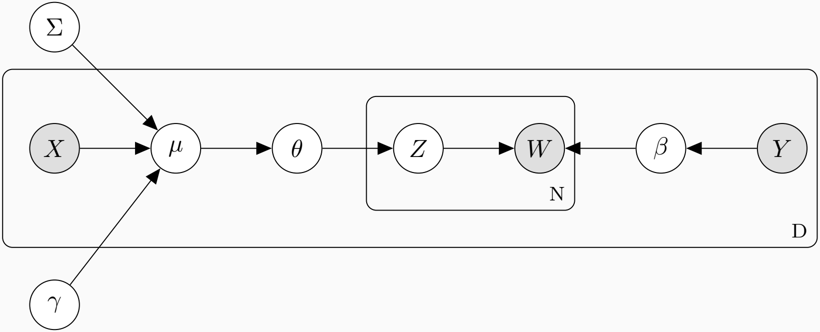
- Typically, when estimating topic models we are interested in how some covariate is associated with the prevalence of topic usage (Gender, date, political party, etc)
- The Structural Topic Model (STM) allows for the inclusion of arbitrary covariates of interest into the generative model
- **Topic prevalence** is allowed to vary according to the covariates X
    - Each document has its own prior distribution over topics, which is defined by its covariates, rather than sharing a global mean
- **Topical content** can also vary according to the covariates Y
    - Word use *within* a topic can differ for different groups of speakers/writers

Structural Topic Model



Topic prevalence model:

- Draw topic proportions (θ) from a logistic normal generalised linear model based on covariates X
- This allows the expected document-topic proportions to vary by covariates, rather than from a single shared prior
- γ coefficients can be interpreted as in regression: the expected change in θ_k for a unit change in X



Topical content model:

- The β coefficients, the word probabilities for a given topic, are allowed to vary according to the covariates Y
- Differences in β capture how documents with different covariates use words differently *within a given topic*

Structural Topic Model Application

- In the legislative domain, we might be interested in the degree to which MPs from different parties represent distinct interests in their parliamentary questions
- We can use the STM to analyse how topic prevalence varies by party
- Specify a linear model with:
  - the topic proportions of speech d, by legislator i as the outcome
  - the party of legislator i as the predictor

$$\theta_{dk} = \alpha + \gamma_{1k} * \text{labour}_{d(i)}$$

- The γ_k coefficients give the estimated difference in topic proportions for Labour and Conservative legislators for each topic

Structural Topic Model Application

```r
library(stm)

## Estimate STM
stmOut <- stm(
            documents = pmq_dfm,
            prevalence = ~party.reduced,
            K = 30,
            seed = 123
            )

save(stmOut, file = "stmOut.Rdata")
```

Structural Topic Model Application

```r
labelTopics(stmOut)
```

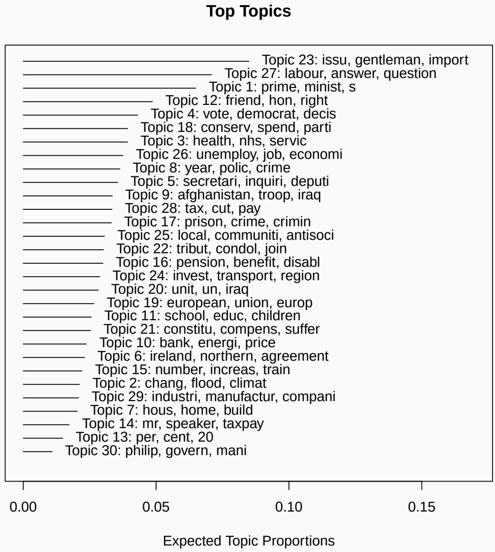
```
## Topic 1 Top Words:
##       Highest Prob: minist, prime, govern, s, tell, confirm, ask
##       FREX: prime, minist, confirm, failur, paymast, lack, fail
##     Lift: protectionist, harrison, roadshow, booki, arrog, googl, pembrokeshir
##       Score: prime, minist, s, confirm, protectionist, govern, tell
## Topic 2 Top Words:
##       Highest Prob: chang, review, target, made, fund, depart, need
##       FREX: climat, flood, review, chang, environ, emiss, carbon
##       Lift: consequenti, parrett, 2050, dredg, climat, greenhous, barnett
##       Score: chang, flood, climat, review, target, environ, emiss
## Topic 3 Top Words:
##       Highest Prob: servic, health, nhs, care, hospit, nation, wait
##       FREX: cancer, patient, nhs, health, hospit, gp, doctor
##     Lift: herceptin, horton, scotsman, wellb, clinician, healthcar, polyclin
##       Score: health, nhs, servic, hospit, cancer, patient, nurs
## Topic 4 Top Words:
##       Highest Prob: decis, vote, made, parti, elect, propos, debat
##       FREX: vote, liber, debat, scottish, decis, recommend, scotland
##       Lift: calman, gould, wakeham, imc, in-built, ipsa, jenkin
##       Score: vote, democrat, decis, parti, debat, liber, elect
## Topic 5 Top Words:
```

Structural Topic Model Application

- `Highest Prob` is the raw β coefficients
- `Score` is the term-score measure we defined above
- `FREX` is a measure which combines word-topic frequency with word-topic exclusivity
- `Lift` is a normalised version of the word-probabilities

Structural Topic Model Application

```
plot(stmOut, labeltype = "score")
```



```r
cloud(stmOut, topic = 3)
```

Structural Topic Model Application

```
findThoughts(model = stmOut,
             texts = texts(pmq_corpus),
             topic = 3)
```

I suspect that many Members from all parties in this House will agree that mental health services have for too long been treated as a poor cousin a Cinderella service in the NHS and have been systematically underfunded for a long time. That is why I am delighted to say that the coalition Government have announced that we will be introducing new access and waiting time standards for mental health conditions such as have been in existence for physical health conditions for a long time. Over time, as reflected in the new NHS mandate, we must ensure that mental health is treated with equality of resources and esteem compared with any other part of the NHS.

```r
dim(stmOut$theta)
```

```
## [1] 27885    30
```

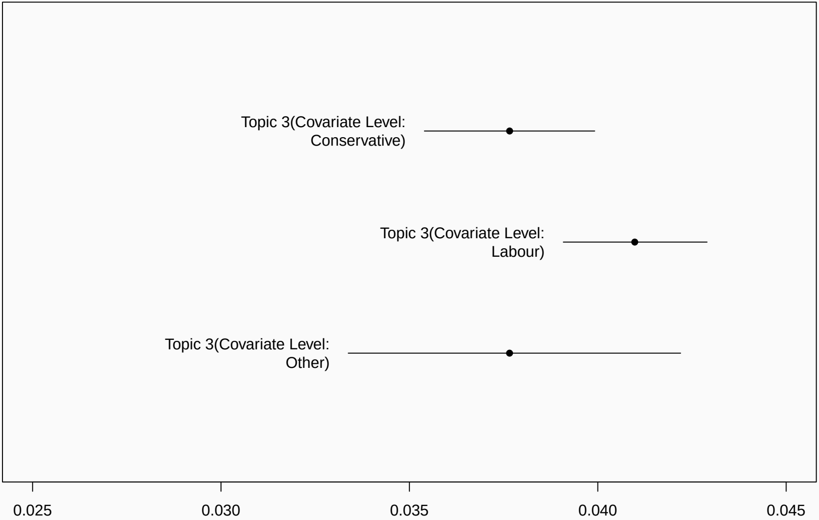
Structural Topic Model Application

Do MPs from different parties speak about healthcare at different rates?

```
stm_effects <- estimateEffect(formula = c(3) ~ party.reduced,
                              stmobj = stmOut,
                              metadata = docvars(pmq_dfm))

plot.estimateEffect(stm_effects,
                    covariate = "party.reduced",
                    method = "pointestimate",
                    xlim = c(0.025, 0.045))
```

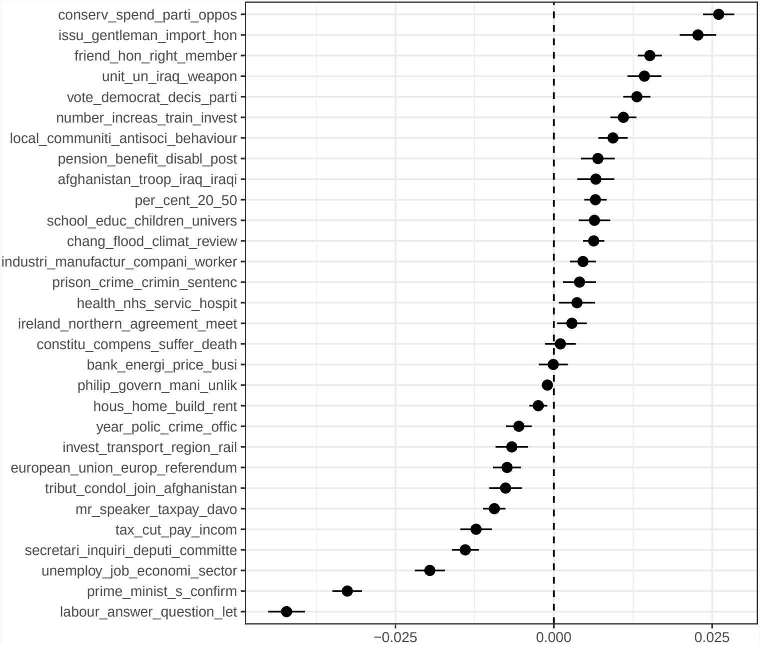
Structural Topic Model Application



On which topics do Conservative and Labour MPs differ the most?

```
stm_effects <- estimateEffect(formula = c(1:30) ~ party.reduced,
                              stmobj = stmOut,
                              metadata = docvars(pmq_dfm))
```

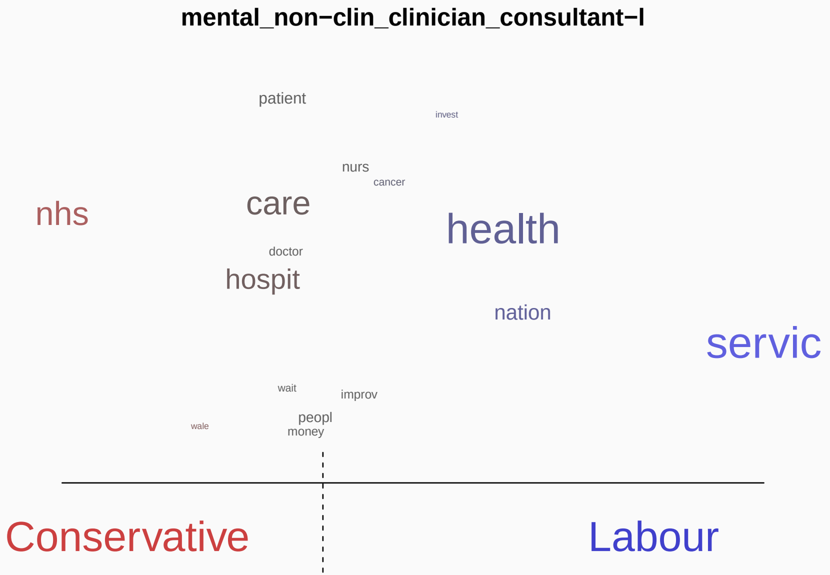
Structural Topic Model Application



```r
library(stm)

## Estimate STM
stmOut2 <- stm(
            documents = pmq_dfm,
            content = ~party.reduced,
            K = 30,
            seed = 123
            )

save(stmOut2, file = "stmOut2.Rdata")
```


mental_non−clin_clinician_consultant−l
patient
invest
nurs
cancer
nhs
care
health
doctor
hospit
nation
servic
wait
improv
wale
peopl
money
Conservative
Labour

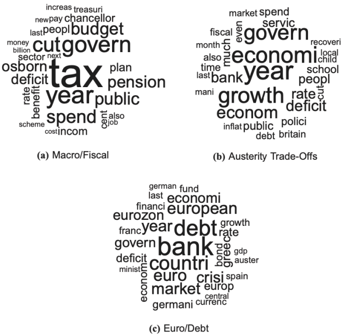
Do liberal and conservative newspapers report on the economy in different ways?

Lucy Barnes and Tim Hicks (UCL) study the determinants of voters' attitudes toward government deficits. They argue that individual attitudes are largely a function of media framing. They examine whether and how the Guardian (a left-leaning) and the Telegraph (a right-leaning) report on the economy.

Data and approach:

- $\approx 10,000$ newspaper articles
    - All articles using the word "deficit" from 2010-2015
- STM model
- $K = 6$
    - "We experimented with topic counts up to 20. Six was the value at which the topics' content could be interpreted as substantively meaningful and distinct."
- Newspaper covariates for both prevalence and content

FIGURE 2 Word Clouds Indicating the Prevalence of Particular Words within Three Fiscal Policy Topics

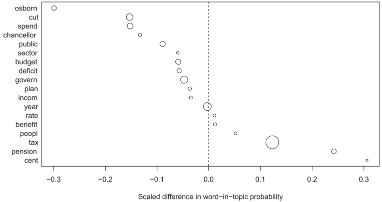


STM Application

Data and approach:

- $\approx 10,000$ newspaper articles
  - All articles using the word "deficit" from 2010-2015
- STM model
- $K = 6$
  - "We experimented with topic counts up to 20. Six was the value at which the topics' content could be interpreted as substantively meaningful and distinct."
- Newspaper covariates for both prevalence and content

FIGURE 3 Relative Frequencies of Most Common Words within Respective Topics



(a) Macro/Fiscal

Validating Topic Models

Validating Topic Models

- LDA, and topic models more generally, require the researcher to make several implementation decisions
- In particular, we must select a value for K, the number of topics
- How can we select between different values of K? How can we tell how well a given topic model is performing?

Held-out likelihood

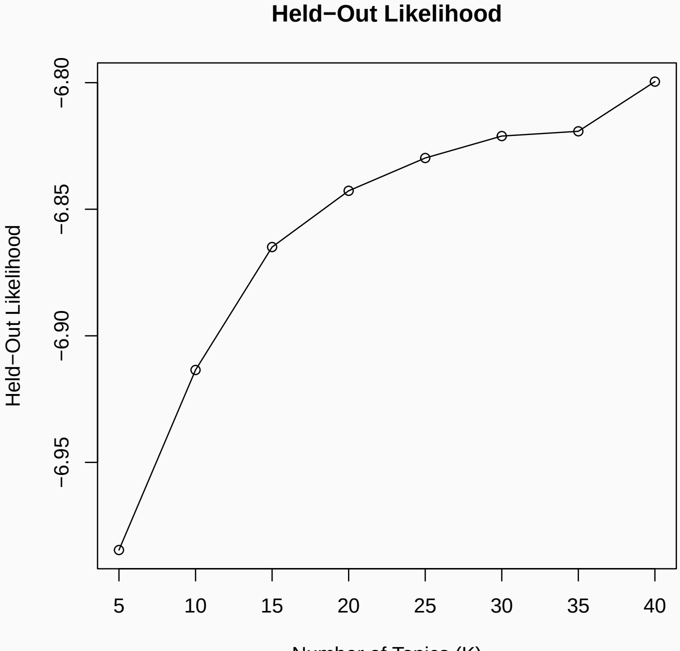
- Ask which words the model believes will be in a given document and compare this to the document's actual word composition
- E.g. Splitting texts in half, train a topic model on one half, calculate the held-out likelihood for the other half

Held-out likelihood

We can apply many of this metric (as well as others) across a range of topic models using the `searchK` function in the `stm` package.

```
search_stm_out <- searchK(documents = pmq_dfm,
                          K = c(5,10,15,20,25,30,35,40),
                          N = 2000)
```

Held-out likelihood



Held-out likelihood

Problems:

- Prediction is not always important in exploratory or descriptive tasks. We may want models that capture other aspects of the data.
- More importantly, there tends to be a negative correlation between held-out likelihood and human judgements of topic coherence!

  "Topic models which perform better on held-out likelihood may infer less semantically meaningful topics." (Chang et al. 2009.)

Semantic validity (Chang et al. 2009)

Word intrusion: Test if topics have semantic coherence by asking humans identify a spurious word inserted into a topic.

Topic	w_1	w_2	w_3	w_4	w_5	w_6
1	bank	financ	**terror**	england	fiscal	market
2	europe	union	eu	referendum	vote	**school**
3	**act**	deliv	nhs	prison	mr	right

Assumption: When humans find it easy to locate the “intruding” *word*, the topics are more coherent.

Semantic validity (Chang et al. 2009)

Topic intrusion: Test if the association between topics and documents makes sense by asking humans to identify a topic that was not associated with a document.

> *Reforms to the banking system are an essential part of dealing with the crisis, and delivering lasting and sustainable growth to the economy. Without these changes, we will be weaker, we will be less well respected abroad, and we will be poorer.*

Topic	w_1	w_2	w_3	w_4	w_5	w_6
1	bank	financ	regul	england	fiscal	market
2	plan	econom	growth	longterm	deliv	sector
3	**school**	**educ**	**children**	**teacher**	**pupil**	**class**

Assumption: When humans find it easy to locate the "intruding" *topic*, the mappings are more sensible.

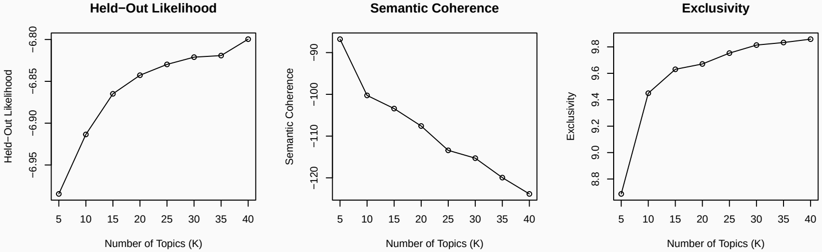
Validating Topic Models – Alternative Quantitative Metrics

Implementing the types of task described above are costly and require a lot of human effort. Some quantitative metrics, however, do a reasonable job at predicting which models perform well on these tasks.

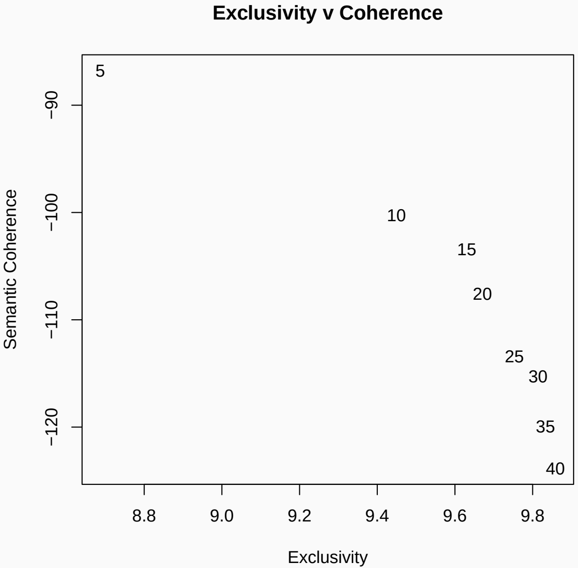
- Semantic coherence
  - Do the most common words from a topic also co-occur together frequently in the same documents?
- Exclusivity
  - Do words with high probability in one topic have low probabilities in others?

One approach to model selection is to find the K that makes these quantities as large as possible.

Quantitative Evaluation of STM



Quantitative Evaluation of STM



Validating Topic Models – Substantive approaches

We can additionally validate a model by comparing the estimated topics to substantive external criteria.

- *Semantic validity*
  - Does a topic identify a coherent groups of texts that are internally homogenous but distinctive from other topics?
- *Predictive validity*
  - How well does variation in topic usage correspond to known events?
- *Construct validity*
  - How well does our measure correlate with other measures?

Implication: All these approaches require *careful human reading of texts and topics*, and comparison with sensible metadata.

Conclusion

- Topic models offer an approach to automatically inferring the substantive themes that exist in a corpus of texts
- A topic is described as a probability distribution over words in the vocabulary
- Documents are described as a mixture of corpus wide topics
- Topic models require very little up-front effort, but require extensive interpretation and validation