

Quantitative text analysis: Topic Models

Blake Miller

MY 459: Quantitative Text Analysis

March 5, 2023

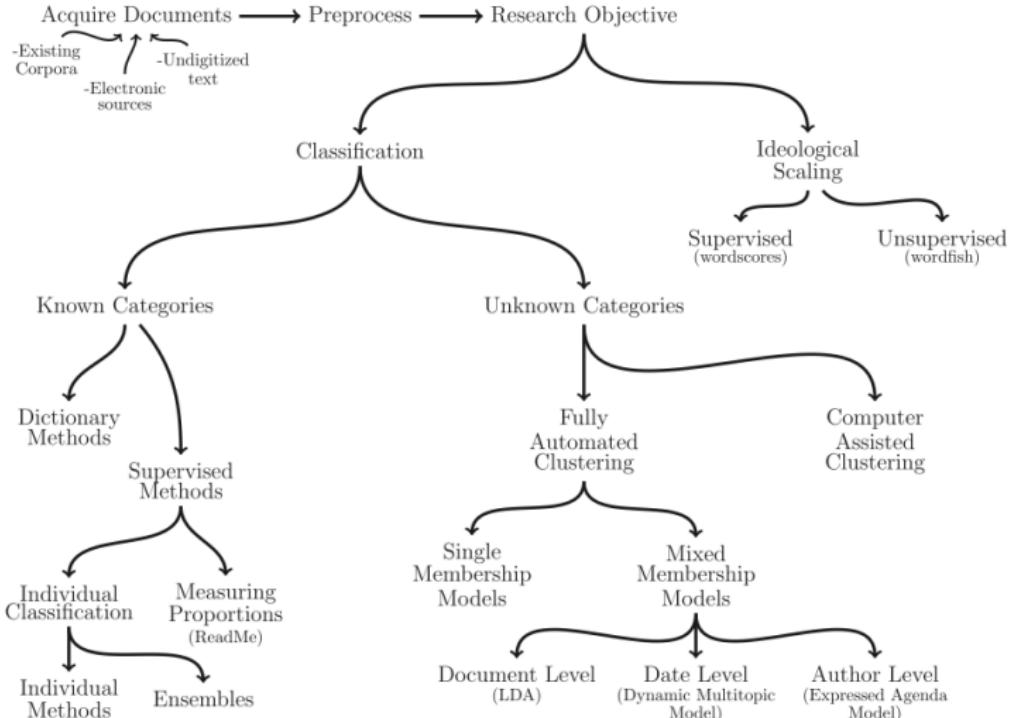
Course website: lse-my459.github.io

1. Overview and Fundamentals
2. Descriptive Statistical Methods for Text Analysis
3. Automated Dictionary Methods
4. Machine Learning for Texts
5. Supervised Scaling Models for Texts
6. *Reading Week*
7. Unsupervised Models for Scaling Texts
8. Similarity and Clustering Methods
9. Topic models
10. Word embeddings/guided coding
11. Current topics in QTA/NLP (if no industrial action)

Discussion of Final Project

- ▶ A reminder of the instructions available on the course website.
- ▶ A reminder of where to submit the final project (Moodle) and the format expected (PDF or Word document, code not required, but can be provided separately from the writeup if you want).
- ▶ Get into groups of three
 - ▶ Share your ideas for what you might want to write your final project on (it's ok if you don't have a good idea on this yet!)
 - ▶ What is your research question? (Again, it's ok if this is not fully formed yet!)
 - ▶ Discuss some ideas for methods from class that might help you answer such a question.
 - ▶ Discuss some questions you may have for me about the project that I can answer for you today.
- ▶ Some common questions:
 - ▶ Can this project be related to my capstone project?
 - ▶ How many different methods from class do I need to use?
- ▶ Other questions on the assignment from the class

Overview of text as data methods



Outline

- ▶ Overview of topic models
- ▶ Latent Dirichlet Allocation (LDA)
- ▶ Model Selection and Validation
- ▶ Examples
- ▶ Extensions of LDA
- ▶ Implementations in R
- ▶ Guided coding

Topic Models

- ▶ Topic models are algorithms for discovering the main “themes” in an unstructured corpus
- ▶ Can be used to organize the collection according to the discovered themes
- ▶ Requires no prior information, training set, or human annotation – only a decision on K (number of topics)
- ▶ Most common: Latent Dirichlet Allocation (LDA) – Bayesian mixture model for discrete data where topics are assumed to be uncorrelated
- ▶ LDA provides a generative model that describes how the documents in a dataset were created
 - ▶ Each of the K *topics* is a distribution over a fixed vocabulary
 - ▶ Each document is a collection of words, generated according to a multinomial distribution, one for each of K topics

Latent Dirichlet Allocation

Figure 1. The intuitions behind latent Dirichlet allocation. We assume that some number of "topics," which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic. The topics and topic assignments in this figure are illustrative—they are not fit from real data. See Figure 2 for topics fit from data.

Topics

| | |
|---------|------|
| gene | 0.04 |
| dna | 0.02 |
| genetic | 0.01 |
| ... | |

| | |
|----------|------|
| life | 0.02 |
| evolve | 0.01 |
| organism | 0.01 |
| ... | |

| | |
|--------|------|
| brain | 0.04 |
| neuron | 0.02 |
| nerve | 0.01 |
| ... | |

| | |
|----------|------|
| data | 0.02 |
| number | 0.02 |
| computer | 0.01 |
| ... | |

Documents

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,¹² two genomic 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

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

"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson, a geneticist at Sweden's Umeå 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 all newly sequenced genomes," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing all

Haemophilus genome
1700 genes

Drosophila genome
13,600 genes

Mycobacterium genome
469 genes

Genes essential for bacterial life
220 genes

Genes essential for Drosophila life
120 genes

Genes essential for Haemophilus life
250 genes

Genes essential for Mycobacterium life
122 genes

Redundant and parasite specific genes
4 genes

Minimal gene set
250 genes

Essential gene set
122 genes

Adapted from NCBI

Topic proportions and assignments

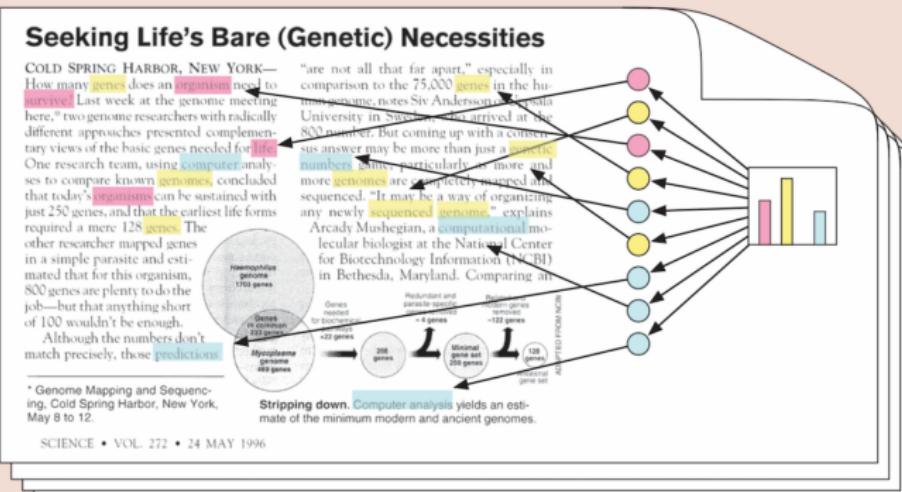


Illustration of the LDA generative process

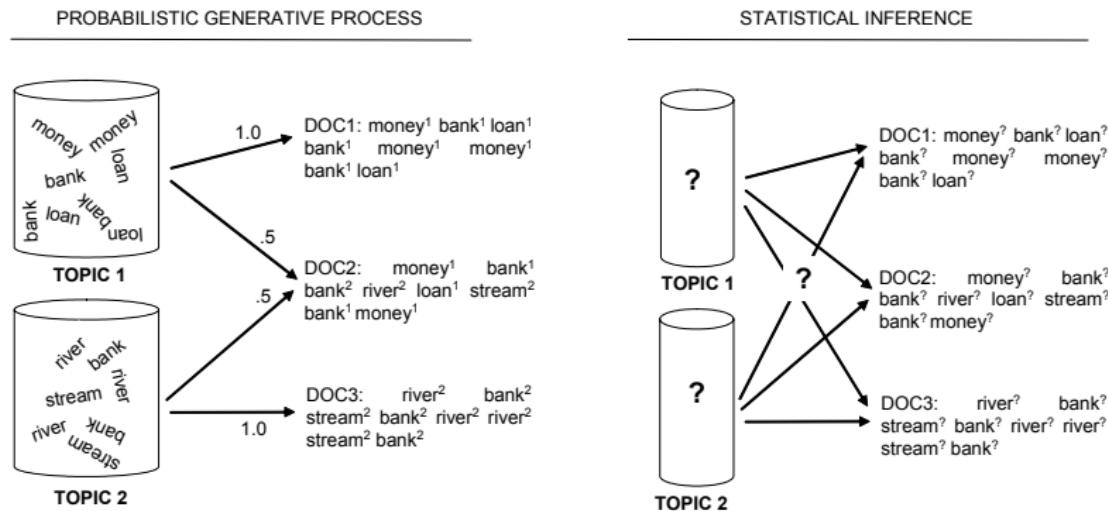


Figure 2. Illustration of the generative process and the problem of statistical inference underlying topic models

(from Steyvers and Griffiths 2007)

Topics example

Topic 247

| word | prob. |
|-----------|-------|
| DRUGS | .069 |
| DRUG | .060 |
| MEDICINE | .027 |
| EFFECTS | .026 |
| BODY | .023 |
| MEDICINES | .019 |
| PAIN | .016 |
| PERSON | .016 |
| MARIJUANA | .014 |
| LABEL | .012 |
| ALCOHOL | .012 |
| DANGEROUS | .011 |
| ABUSE | .009 |
| EFFECT | .009 |
| KNOWN | .008 |
| PILLS | .008 |

Topic 5

| word | prob. |
|---------|-------|
| RED | .202 |
| BLUE | .099 |
| GREEN | .096 |
| YELLOW | .073 |
| WHITE | .048 |
| COLOR | .048 |
| BRIGHT | .030 |
| COLORS | .029 |
| ORANGE | .027 |
| BROWN | .027 |
| PINK | .017 |
| LOOK | .017 |
| BLACK | .016 |
| PURPLE | .015 |
| CROSS | .011 |
| COLORED | .009 |

Topic 43

| word | prob. |
|------------|-------|
| MIND | .081 |
| THOUGHT | .066 |
| REMEMBER | .064 |
| MEMORY | .037 |
| THINKING | .030 |
| PROFESSOR | .028 |
| FELT | .025 |
| REMEMBERED | .022 |
| THOUGHTS | .020 |
| FORGOTTEN | .020 |
| MOMENT | .020 |
| THINK | .019 |
| THING | .016 |
| WONDER | .014 |
| FORGET | .012 |
| RECALL | .012 |

Topic 56

| word | prob. |
|-----------|-------|
| DOCTOR | .074 |
| DR. | .063 |
| PATIENT | .061 |
| HOSPITAL | .049 |
| CARE | .046 |
| MEDICAL | .042 |
| NURSE | .031 |
| PATIENTS | .029 |
| DOCTORS | .028 |
| HEALTH | .025 |
| MEDICINE | .017 |
| NURSING | .017 |
| DENTAL | .015 |
| NURSES | .013 |
| PHYSICIAN | .012 |
| HOSPITALS | .011 |

Figure 1. An illustration of four (out of 300) topics extracted from the TASA corpus.

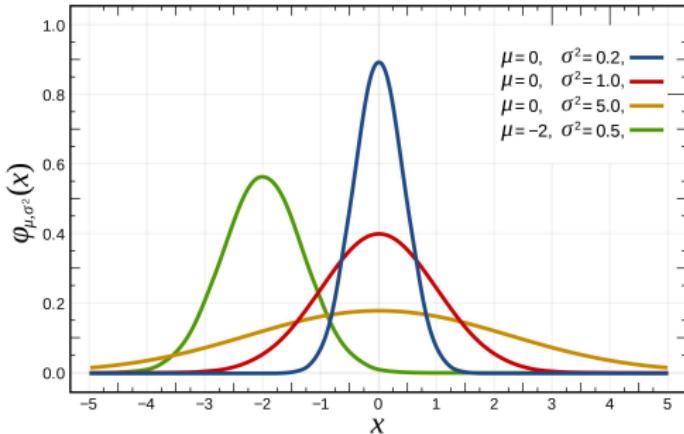
(from Steyvers and Griffiths 2007)

Often K is quite large!

Outline

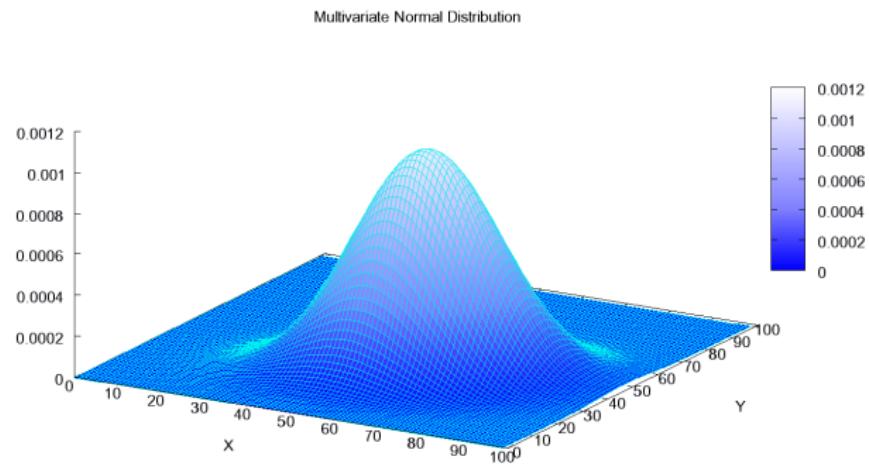
- ▶ Overview of topic models
- ▶ Latent Dirichlet Allocation (LDA)
- ▶ Model Selection and Validation
- ▶ Examples
- ▶ Extensions of LDA
- ▶ Implementations in R
- ▶ Guided coding

Review: Univariate probability density function



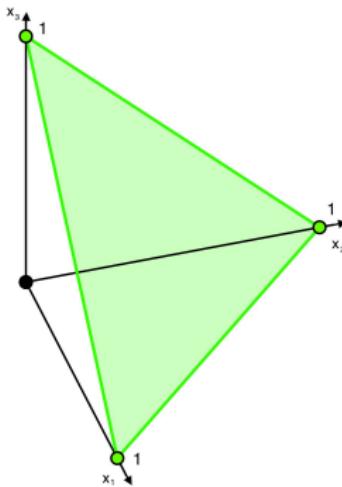
- ▶ Different parameter values (e.g. μ and σ for the normal distribution) change the distributions' shape
- ▶ The notation " $x \sim N(0, 1)$ " denotes to sample or draw " x " from a standard normal distribution. This draw could e.g. return $x = -1.124$

Review: Multivariate probability density function



A draw $a \sim N(\mu, \Sigma)$ from this multivariate normal distribution could e.g. return $a = (-0.12, 1.2)$

Review: Graphical intuition of a standard simplex

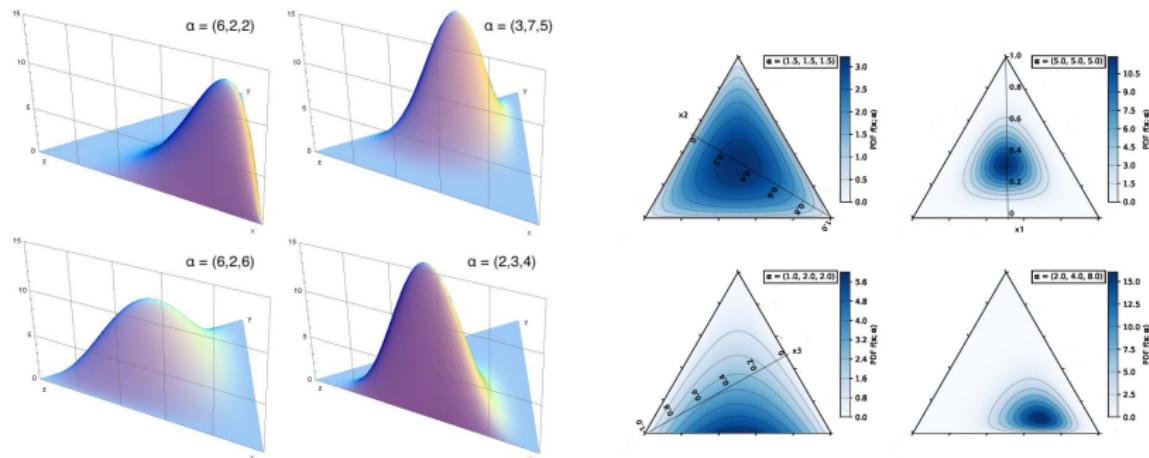


A point on the triangle is x_1, x_2, x_3 , with $x_1, x_2, x_3 \in [0, 1]$ and $x_1 + x_2 + x_3 = 1$. For example, $(0.05, 0.8, 0.15)$

Generalises to higher dimensions with $x_1, \dots, x_K \in [0, 1]$ and $\sum x_k = 1$

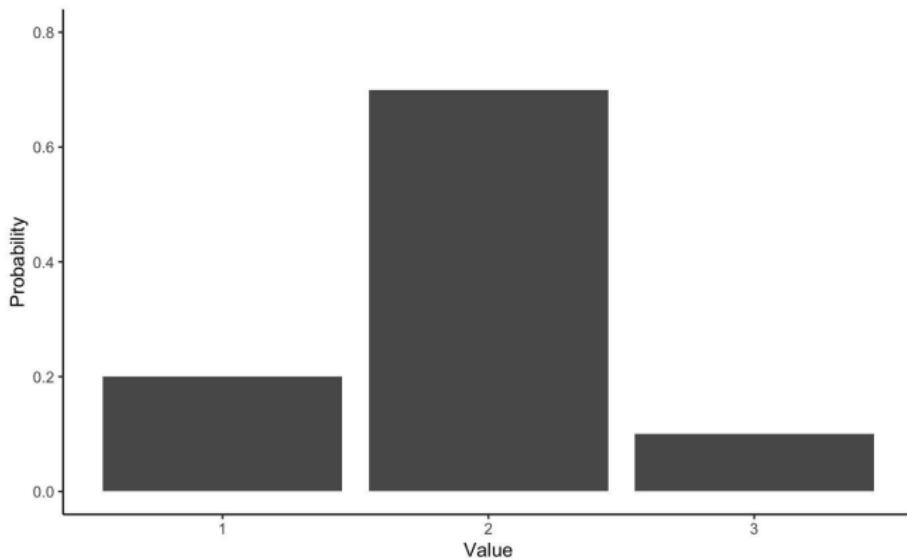
Key distribution 1: The Dirichlet distribution

Dirichlet distribution: Probability distribution over a simplex



- ▶ A draw $b \sim Dir(\alpha)$ from this distribution could e.g. return $b = (0.2, 0.7, 0.1)$
- ▶ Hence, we can think of the draw from a Dirichlet distribution being itself a multinomial distribution (next slide)

Key distribution 2: Multinomial distribution



The multinomial distribution depicted has probabilities $[0.2, 0.7, 0.1]$. A draw $c \sim \text{Multinomial}([0.2, 0.7, 0.1])$ could e.g. return $c = 2$

Latent Dirichlet Allocation

- ▶ Document = random mixture over latent topics
- ▶ Topic = distribution over words

Probabilistic model with 3 steps:

1. Choose $\theta_i \sim \text{Dirichlet}(\alpha)$
2. Choose $\beta_k \sim \text{Dirichlet}(\delta)$
3. For each word in document i :
 - ▶ Choose a topic $z_m \sim \text{Multinomial}(\theta_i)$
 - ▶ Choose a word $w_{im} \sim \text{Multinomial}(\beta_{i,k=z_m})$

where:

α =parameter of Dirichlet prior on distribution of topics over docs.

θ_i =topic distribution for document i

δ =parameter of Dirichlet prior on distribution of words over topics

β_k =word distribution for topic k

LDA: Generative model

- ▶ Now that we have reviewed some key concepts, we can study how the LDA assumes our documents have been generated
- ▶ We consider a corpus of D documents, each with N_d words
- ▶ Idea: Assume a statistical model that generated the observed corpus, then estimate the model and recover latent (unobserved) quantities, i.e. topics, under the assumptions made
- ▶ Each document is assumed to contain weights of topics, and each topic is assumed to contain weights of words
- ▶ For simplicity, let us first imagine a stylised corpus with only $D = 5$ documents, $N = 8$ words per document (and also 8 unique words in the corpus in total), and $K = 3$ topics

LDA: Generative model, stylised example

- ▶ For each of the 5 documents, draw its topic shares from a Dirichlet distribution $\theta_d \sim Dir(\alpha)$. For document 1 this could e.g. yield $\theta_1 = (0.1, 0.8, 0.1)$, i.e. document 1 consists predominantly of topic 2
- ▶ For each of the 3 topics, draw its word shares from a second Dirichlet distribution $\beta_k \sim Dir(\delta)$. For topic 2 this could e.g. yield $\beta_2 = (0, 0, 0, 0.1, 0, 0, 0.4, 0.5)$, i.e. topic 2 consists predominantly of words 7 and 8
- ▶ Now it is possible to fill each document with words. For each document-word position d, n
 - ▶ Choose the topic $z_{d,n}$ of the document-word position by taking a draw from document d 's topic distribution: $z_{d,n} \sim \text{Multinomial}(\theta_d)$. For example, for the first word in document 1 we might draw topic $z_{d,n} = 2$
 - ▶ Choose the word $w_{d,n}$ by taking a draw from the corresponding topic distribution: $w_{d,n} \sim \text{Multinomial}(\beta_{z_{d,n}})$. Since we are drawing from topic 2, we might draw word 8. This fills the word at position d, n

LDA: Generative model, general case

- ▶ For each document $d \in \{1, \dots, D\}$, draw its topic shares from a Dirichlet distribution $\theta_d \sim Dir(\alpha)$
- ▶ For each topic $k \in \{1, \dots, K\}$, draw its word shares from a second Dirichlet distribution $\beta_k \sim Dir(\delta)$
- ▶ Fill each document with words. For each document-word position d, n in $d \in \{1, \dots, D\}$ and $n \in \{1, \dots, N_d\}$
 - ▶ Choose the topic $z_{d,n}$ of the document-word position by taking a draw from document d 's topic distribution: $z_{d,n} \sim \text{Multinomial}(\theta_d)$
 - ▶ Choose the word $w_{d,n}$ by taking a draw from the corresponding topic distribution: $w_{d,n} \sim \text{Multinomial}(\beta_{z_{d,n}})$

Latent Dirichlet Allocation

Key parameters:

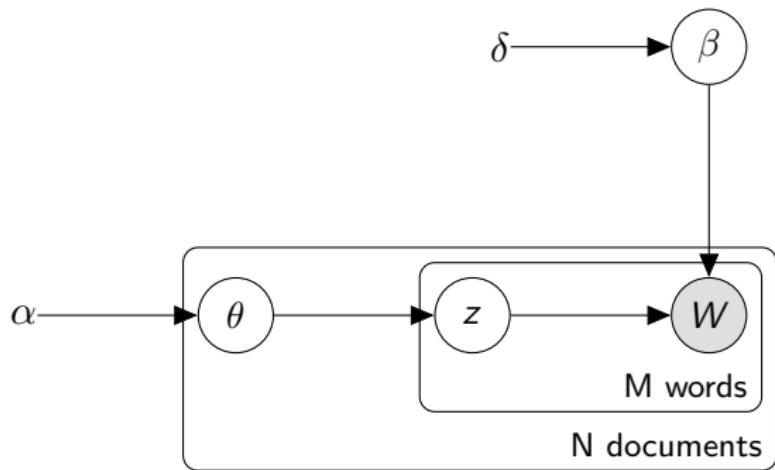
1. θ = matrix of dimensions N documents by K topics where θ_{ik} corresponds to the probability that document i belongs to topic k ; i.e. assuming $K = 5$:

| | T1 | T2 | T3 | T4 | T5 |
|--------------|------|------|------|------|------|
| Document 1 | 0.15 | 0.15 | 0.05 | 0.10 | 0.55 |
| Document 2 | 0.80 | 0.02 | 0.02 | 0.10 | 0.06 |
| ... | | | | | |
| Document N | 0.01 | 0.01 | 0.96 | 0.01 | 0.01 |

2. β = matrix of dimensions K topics by M words where β_{km} corresponds to the probability that word m belongs to topic k ; i.e. assuming $M = 6$:

| | W1 | W2 | W3 | W4 | W5 | W6 |
|-----------|------|------|------|------|------|------|
| Topic 1 | 0.40 | 0.05 | 0.05 | 0.10 | 0.10 | 0.30 |
| Topic 2 | 0.10 | 0.10 | 0.10 | 0.50 | 0.10 | 0.10 |
| ... | | | | | | |
| Topic k | 0.05 | 0.60 | 0.10 | 0.05 | 0.10 | 0.10 |

Plate notation



$\beta : M \times K$ matrix where β_{im} indicates prob(topic= k) for word m

z : The topic assignment for a particular word m

$\theta : N \times K$ matrix where θ_{ik} indicates prob(topic= k) for document i

Algorithms to obtain posterior distributions

- ▶ Not the focus of this lecture
- ▶ Common algorithm to obtain posterior distributions for the θ_d and β_k are from two main groups
 1. Sampling algorithms such as *Gibbs sampling* approximate the posterior distribution with an empirical distribution by drawing a sequence of samples in a way that ensure its limit distribution is the true posterior distribution
 2. Variational methods approximate the posterior distribution with a parametrised family of distributions. The error of this approximation is minimised and obtaining posteriors thus becomes an optimisation problem
- ▶ See “Probabilistic topic models” by Blei (2012) for links to further resources

Outline

- ▶ Overview of topic models
- ▶ Latent Dirichlet Allocation (LDA)
- ▶ **Model Selection and Validation**
- ▶ Examples
- ▶ Extensions of LDA
- ▶ Implementations in R
- ▶ Guided coding

Validation

From Quinn et al, AJPS, 2010:

1. Semantic validity

- ▶ Do the topics identify coherent groups of tweets that are internally homogenous, and are related to each other in a meaningful way?

2. Convergent/discriminant construct validity

- ▶ Do the topics match existing measures where they should match?
- ▶ Do they depart from existing measures where they should depart?

3. Predictive validity

- ▶ Does variation in topic usage correspond with expected events?

4. Hypothesis validity

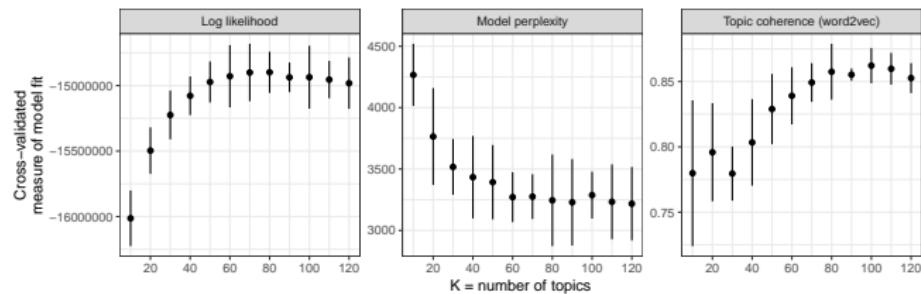
- ▶ Can topic variation be used effectively to test substantive hypotheses?

Selecting K and beyond

- ▶ In particular selecting the amount of topics K , but also parameters, covariates, and potentially initialisations is a very challenging exercise without a single solution
- ▶ Researchers typically consult a combination of quantitative metrics and human judgement

Choosing the number of topics

- ▶ Choosing K is “one of the most difficult questions in unsupervised learning” (Grimmer and Stewart, 2013, p.19)
- ▶ One approach is to decide based on cross-validated model fit



- ▶ **BUT:** “there is often a negative relationship between the best-fitting model and the substantive information provided”.
- ▶ GS propose to choose K based on “substantive fit.”

Model evaluation using “perplexity”

- ▶ Makes use of “held-out” data as we would in a supervised learning problem
- ▶ Perplexity can tell us how well our topic model can predict words on these held-out documents.
- ▶ **perplexity**: can be computed as (using VEM):

$$\text{perplexity}(w) = \exp \left\{ -\frac{\sum_{d=1}^M \log p(w_d)}{\sum_{d=1}^M N_d} \right\}$$

- ▶ A lower perplexity score indicates better performance

Quantitative metrics

- ▶ Held-out likelihood or perplexity: For some held-out documents, how likely would the model have generated/predicted these documents
- ▶ Semantic coherence: For example, how likely do the most common words from a topic also co-occur in the same document?
- ▶ Exclusivity: Do words with high probability in one topic have low probabilities in others?
- ▶ Many automated metrics exist, see e.g. Grimmer and Stewart (2013), Mimno et al. (2011), Taddy (2012)

On held-out likelihood metrics

- ▶ Held-out likelihood and perplexity allow to judge the predictive ability of the model
- ▶ Yet, choosing K such that it achieves the highest held-out likelihood has serious limitations
- ▶ Rather than creating an auto-complete prediction model or something similar, the purpose of topic modelling most often is to obtain coherent topics that tell a story
- ▶ In their paper “Reading tea leaves” (2009) Chang et al. contrast likelihood based metrics with human judgements about topic coherence
- ▶ Also see this short video by one of the authors

Reading tea leaves

Word Intrusion

| | | | | | | |
|--------|-------------|-----------|----------|------------|------------|----------|
| 1 / 10 | floppy | alphabet | computer | processor | memory | disk |
| 2 / 10 | molecule | education | study | university | school | student |
| 3 / 10 | linguistics | actor | film | comedy | director | movie |
| 4 / 10 | islands | island | bird | coast | portuguese | mainland |

Topic Intrusion

| DOUGLAS HOFSTADTER | | | | | | | |
|---|--------|------------|-----------|-----------|------------|-----------|-------------|
| Douglas Richard Hofstadter (born February 15, 1945 in New York, New York) is an American academic whose research focuses on consciousness, thinking and creativity. He is best known for ", first published in | | | | | | | |
| student | school | study | education | research | university | science | learn |
| human | life | scientific | science | scientist | experiment | work | idea |
| play | role | good | actor | star | career | show | performance |
| write | work | book | publish | life | friend | influence | father |

Figure 2: Screenshots of our two human tasks. In the word intrusion task (left), subjects are presented with a set of words and asked to select the word which does not belong with the others. In the *topic intrusion* task (right), users are given a document's title and the first few sentences of the document. The users must select which of the four groups of words does not belong.

Figure: From: “Reading Tea Leaves: How Humans Interpret Topic Models” by Chang et al., 2009

- ▶ Likelihood based metrics were actually negatively correlated with human metrics about topic coherence

Quantitative metrics

- ▶ For a discussion of model selection for (structural) topic models (e.g. different choices of K , initialisations, and covariates) and evaluation, see e.g. Section “3.4. Evaluate: Model selection and search” in the package vignette or the Section *Model Specification and Selection* in “Structural Topic Models for Open-Ended Survey Responses” by Roberts et al. (2014)
- ▶ In Roberts et al. (2014), the authors e.g. argue that “a semantically interpretable topic has two qualities: (1) it is cohesive in the sense that high-probability words for the topic tend to co-occur within documents, and (2) it is exclusive in the sense that the top words for that topic are unlikely to appear with in top words of other topics.”
- ▶ Semantic coherence and exclusivity, with many other quantitative metrics, are outputs of the function ‘searchK’ in the ‘stm’ package

Takeways

- ▶ No quantitative metric can replace human judgement when selecting K or other model parameters, and evaluating the fit of a particular topic model more generally
- ▶ “The most effective method for assessing model fit is to carefully read documents that are closely associated with particular topics to verify that the semantic concept covered by the topic is reflected in the text.” from “A Model of Text for Experimentation in the Social Sciences” by Roberts et al. (2016)
- ▶ The ‘plotQuote’ function in ‘stm’ allows to plot documents highly associated with certain topics
- ▶ Key consideration: What is the goal of the current model?
- ▶ To generate coherent topics which describe themes? Combine careful human reading with quantitative metrics
- ▶ To use topic models e.g. for document embeddings to find similar documents? As input in a predictive model? Try to select K and other parameters such that they maximise whatever the objective function

Outline

- ▶ Overview of topic models
- ▶ Latent Dirichlet Allocation (LDA)
- ▶ Model Selection and Validation
- ▶ Examples
- ▶ Extensions of LDA
- ▶ Implementations in R
- ▶ Guided coding

Example: open-ended survey responses

Bauer, Barberá *et al*, *Political Behavior*, 2016.

- ▶ Data: General Social Survey (2008) in Germany
- ▶ Responses to questions: *Would you please tell me what you associate with the term “left”? and would you please tell me what you associate with the term “right”?*
- ▶ Open-ended questions minimize priming and potential interviewer effects
- ▶ Sparse Additive Generative model instead of LDA (more coherent topics for short text)
- ▶ $K = 4$ topics for each question

Example: open-ended survey responses

Table 1: Top scoring words associated with each topic, and English translations)

| |
|--|
| Left topic 1: Parties (proportion = .26, average lr-scale value = 5.38) linke, spd, partei, linken, pds, politik, kommunisten, parteien, grünen, punks <i>the left, spd, party, the left, pds, politics, communists, parties, greens, punks</i> |
| Left topic 2: Ideologies (proportion = .26, average lr-scale value = 5.36) kommunismus, links, sozialismus, lafontaine, rechts, aber, gysi, linkspartei, richtung, gleichmacherei <i>communism, left, socialism, lafontaine, right, but, gysi, left party, direction, levelling</i> |
| Left topic 3: Values (proportion = .24, average lr-scale value = 4.06) soziale, gerechtigkeit, demokratie, soziales, bürger, gleichheit, gleiche, freiheit, rechte, gleichberechtigung <i>social, justice, democracy, social, citizen, equality, equal, freedom, rights, equal rights</i> |
| Left topic 4: Policies (proportion = .24, average lr-scale value = 4.89) sozial, menschen, leute, ddr, verbinde, kleinen, einstellung, umverteilung, sozialen, vertreten <i>social, humans, people, ddr, associate, the little, attitude, redistribution, social, represent</i> |
| Right topic 1: Ideologies (proportion = .27, average lr-scale value = 5.00) konservativ, nationalsozialismus, rechtsradikal, radikal, ordnung, politik, nazi, recht, menschen, konservative <i>conservative, national socialism, right-wing radicalism, radical, order, politics, nazi, right, people, conservatives</i> |
| Right topic 2: Parties (proportion = .25, average lr-scale value = 5.26) npd, rechts, cdu, csu, rechten, parteien, leute, aber, verbinde, rechtsradikalen <i>npd, right, cdu, csu, the right, parties, people, but, associate, right-wing radicalists</i> |
| Right topic 3: Xenophobia (proportion = .25, average lr-scale value = 4.55) ausländerfeindlichkeit, gewalt, ausländer, demokratie, nationalismus, rechtsradikalismus, diktatur, national, intoleranz, faschismus <i>xenophobia, violence, foreigners, democracy, nationalism, right-wing radicalism, dictatorship, national, intolerance, fascism</i> |
| Right topic 4: Right-wing extremists (proportion = .23, average lr-scale value = 4.90) nazis, neonazis, rechtsradikale, rechte, radikale, radikalismus, partei, ausländerfeindlich, reich, nationale <i>nazis, neonazis, right-wing radicalists, rightists, radicals, radicalism, party, xenophobia, rich, national</i> |

Note: "proportion" indicates the average estimated probability that any given response is assigned to a topic. "average lr-scale value" is the mean position on the left-right scale (from 0 to 10) of individuals whose highest probability belongs to that particular topic.

Example: open-ended survey responses

Fig. 6: Left-right scale means for different subsamples of associations with left (dashed = sample mean, bars = 95% Cis)

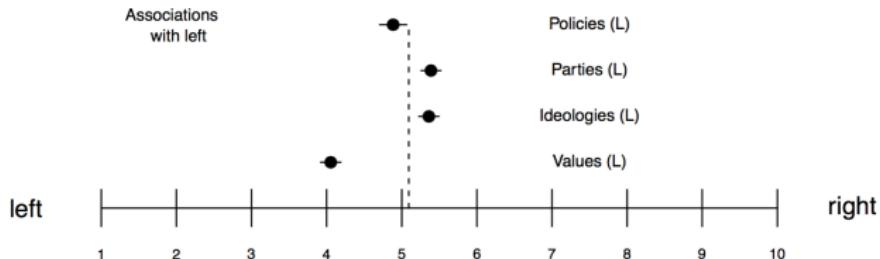
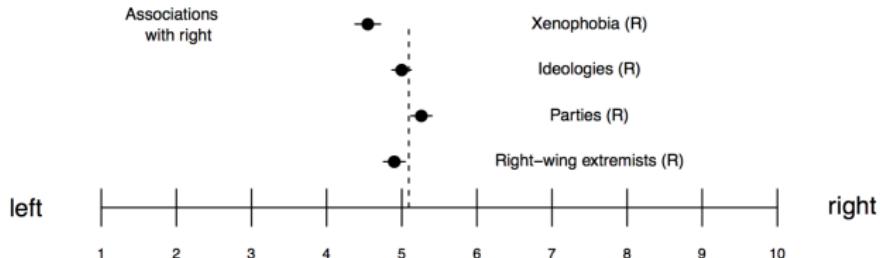
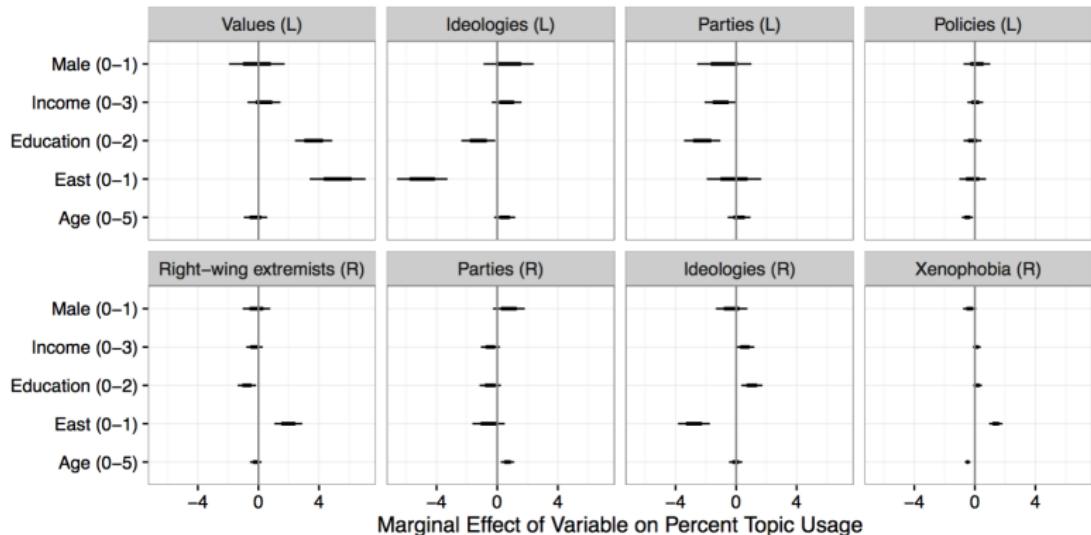


Fig. 7: Left-right scale means for different subsamples of associations with right (dashed = sample mean, bars = 95% Cis)



Example: open-ended survey responses

Fig. 9: Systematic relationship between associations with “left” and “right” and characteristics of respondents



Note: Each line indicates a 95% confidence interval (and 66% confidence interval in darker color) for the coefficient of eight different regressions of topic usage (in a scale from 0 to 100) at the respondent level on seven individual-level characteristics. The line on the bottom right corner (second row, second plot), for example, shows that individual a one-category change in age is associated with around one percentage point increase in the probability that the individual associated “right” with political parties.

Bauer, Barberá *et al*, *Political Behavior*, 2016.

Example: topics in US legislators' tweets

- ▶ A carefully documented project with very good average topic coherence is e.g. "Leaders or Followers? Measuring Political Responsiveness in the U.S. Congress Using Social Media Data" by Barberá et al. (2014)
- ▶ Data: 651,116 tweets sent by US legislators from January 2013 to December 2014
- ▶ 2,920 documents = $730 \text{ days} \times 2 \text{ chambers} \times 2 \text{ parties}$
- ▶ Why aggregating? Applications that aggregate by author or day outperform tweet-level analyses (Hong and Davidson, 2010)
- ▶ Sidenote: For short texts, such as also e.g. survey responses, topic models such as sparse additive generative models (SAGE) might create more coherent topics than e.g. LDA (see e.g. Bauer et al, Political Behavior, 2017)
- ▶ $K = 100$ topics
- ▶ Validation: <http://j.mp/lda-congress-demo>

Outline

- ▶ Overview of topic models
- ▶ Latent Dirichlet Allocation (LDA)
- ▶ Model Selection and Validation
- ▶ Examples
- ▶ Extensions of LDA
- ▶ Implementations in R
- ▶ Guided coding

Extensions of LDA

1. Structural topic model (Roberts et al, 2014, AJPS)
2. Dynamic topic model (Blei and Lafferty, 2006, ICML; Quinn et al, 2010, AJPS)
3. Hierarchical topic model (Griffiths and Tenenbaum, 2004, NIPS; Grimmer, 2010, PA)

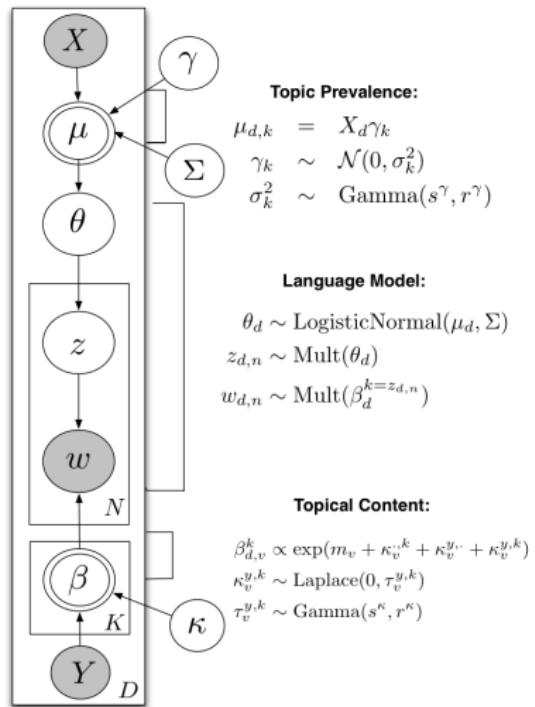
Why?

- ▶ Substantive reasons: incorporate specific elements of DGP into estimation
- ▶ Statistical reasons: structure can lead to better topics.

Structural topic model (STM)

- ▶ Basic idea: STM = LDA + Contextual Information
- ▶ STM provides two ways to include contextual information
 - ▶ Topic prevalence can vary by metadata (e.g. Democrats talk more about education than Republicans)
 - ▶ Topic content can vary by metadata (e.g. Democrats are less likely to use the word “life” when talking about abortion than Republicans)
- ▶ Including context improves the model:
 - ▶ more accurate estimation
 - ▶ better qualitative interpretability

Structural topic model (STM)

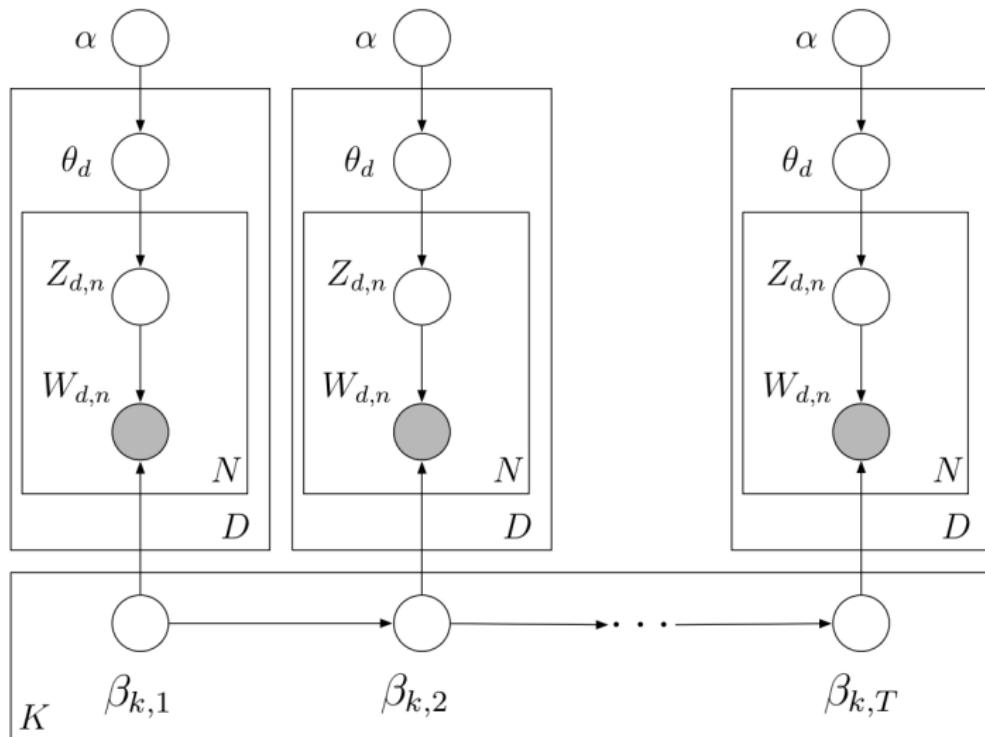


- ▶ **Prevalence:** Prior on the mixture over topics is now document-specific, and can be a function of covariates (documents with similar covariates will tend to be about the same topics)
- ▶ **Content:** distribution over words is now document-specific and can be a function of covariates (documents with similar covariates will tend to use similar words to refer to the same topic)

Structural topic model (STM)

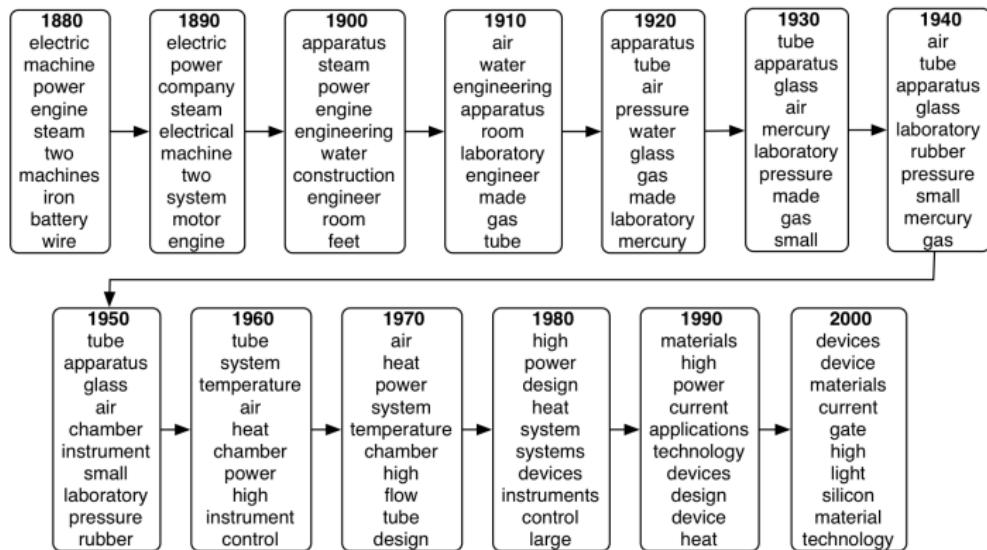
- ▶ User specifies the number of topics: K
- ▶ Observed data for standard topic models
 - ▶ Each document ($d \in 1 \dots D$) is a collection of N_d tokens
 - ▶ Each token is a particular word from a dictionary of V entries
 - ▶ Data summarized in a single matrix $D \times V$ matrix \mathbf{W}
- ▶ Additional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

Dynamic topic model



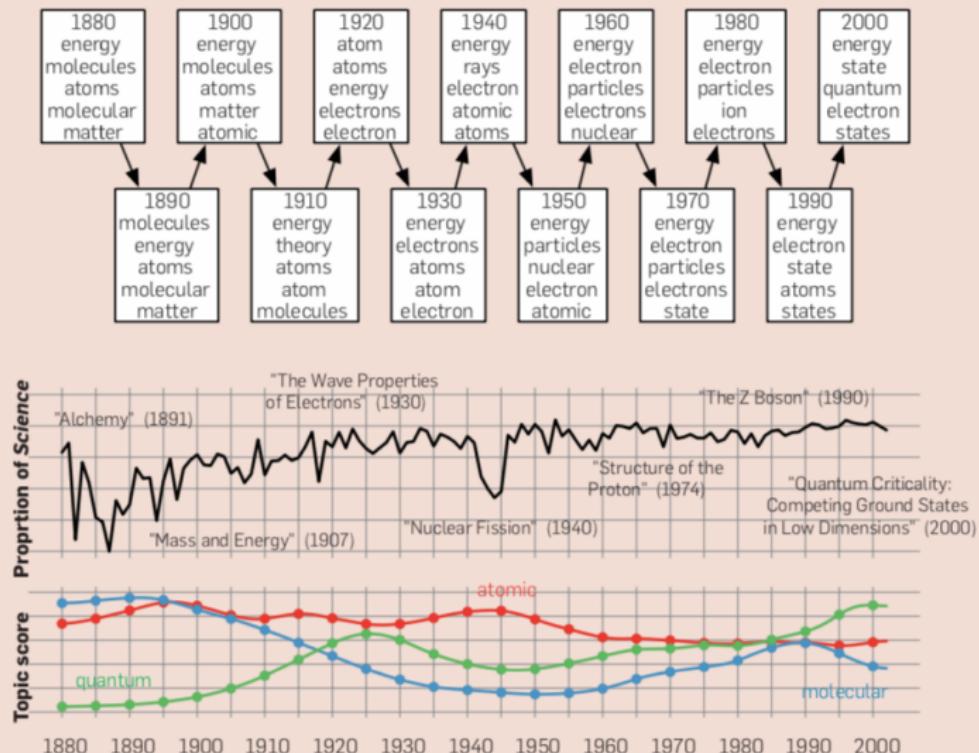
Source: Blei, "Modeling Science"

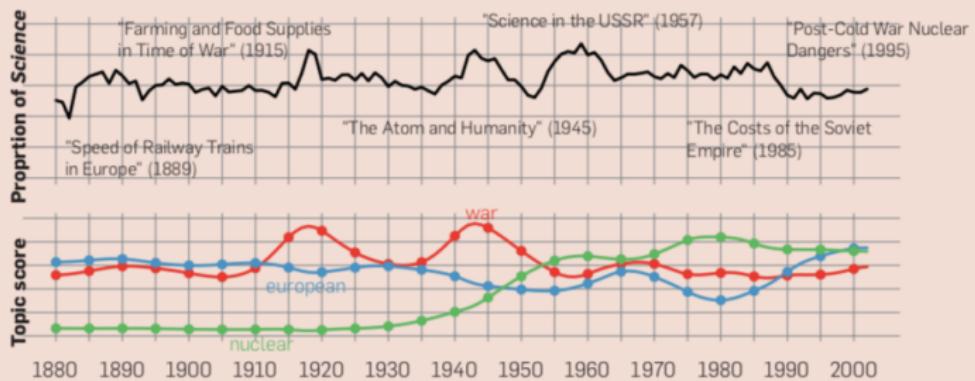
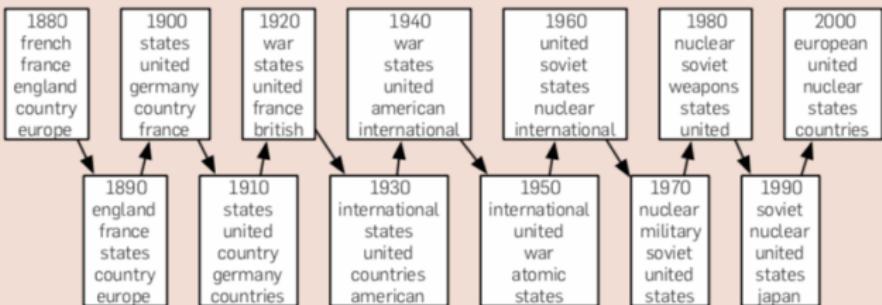
Dynamic topic model



Source: Blei, "Modeling Science"

Figure 5. Two topics from a dynamic topic model. This model was fit to *Science* from 1880 to 2002. We have illustrated the top words at each decade.





Outline

- ▶ Overview of topic models
- ▶ Latent Dirichlet Allocation (LDA)
- ▶ Model Selection and Validation
- ▶ Examples
- ▶ Extensions of LDA
- ▶ **Implementations in R**

Implementations in R

- ▶ For an implementation of the LDA and CTM, see e.g. the package 'topicmodels'
- ▶ We will focus on the 'stm' package which offers a range of helpful functionalities
- ▶ Without added covariates, the 'stm' function also estimates a standard CTM, with covariates a structural topic model
- ▶ Further helpful package to visualise topic models are 'LDAvis' or, the stm-specific, 'stminsights'

Functionalities 'stm' package

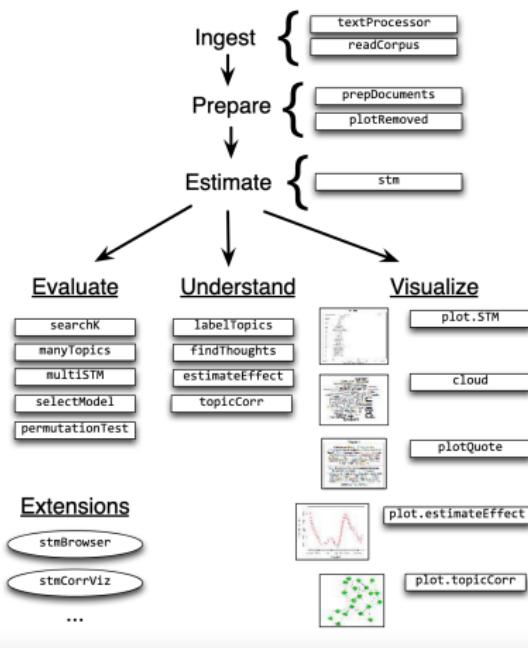


Figure: From the vignette “stm: RPackage for Structural Topic Models” by Roberts, Stewart, and Tingley

Next week

- ▶ Topic models guided coding
 - ▶ Latent dirichlet example
 - ▶ Using STM for open-ended survey responses
- ▶ Overview of word embeddings
- ▶ Different algorithms: word2vec and GloVe
- ▶ Social Science Applications
- ▶ Algorithmic Bias and Fairness
- ▶ Word embedding guided coding
 - ▶ Dictionary term expansion
 - ▶ Diachronic embeddings (how do word meanings change over time)