Topic Modeling (LDA)

1/28/26

Recap

- Last class:
  - Metrics to measure differences in word usage across subsets of corpora
    - Log Odds with Dirichlet Prior (Fightin' Words)
    - PMI Scores
- Today
  - Topic modeling (LDA)
  - Inference method 1: Gibbs sampling
  - Practical considerations

LDA Introduction

Odds ratio in Congressional data

| Top Republican Words	Score	Top Democrat Words	Score
spending	-66.26	republican	56.63
obamacare	-59.90	wealthiest	40.78
government	-47.92	rhode	39.43
going	-45.33	women	38.16
that	-44.58	pollution	33.66
trillion	-43.43	republicans	32.86
taxes	-42.39	gun	32.45
you	-40.85	investments	32.22
administration	-39.07	families	31.93
debt	-38.92	violence	30.88

Probably all about budget and government spending

"Gun violence" is probably one topic

Topic Modeling: Motivation

- Sometimes we care about specific words (more on this later)
- Often we want to group words into broader *topics*
  - But we don't know these topics in advance, we need to discover them from the data

Latent Dirichlet Allocation
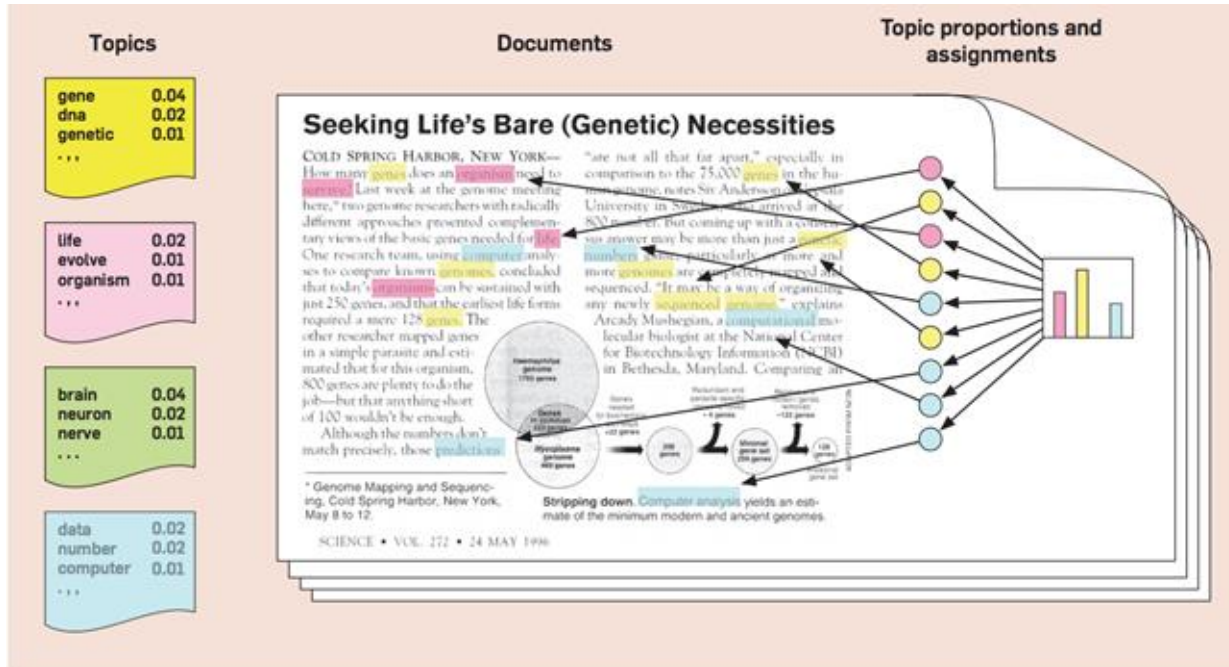
- Assume each document contains a mixture of "topics"
- Each topic uses mixtures of vocabulary words
- **Goal: recover topic and vocabulary distributions**

Definitions

	Topic 1	Topic 2	...	Topic 30
administration	0.01	**0.12**	...	0.02
advertising	0.02	0.001	...	**0.25**
debt	0.1	0.001	...	0.01
...	...	...	...	...
government	0.01	**0.15**	...	0.01
...	...	...	...	...
spending	**0.12**	0.01	...	0.03
taxes	**0.15**	0.02	...	**0.35**
trillion	**0.19**	0.003	...	0.02

	Doc 1	Doc 2	...	Doc N
Topic 1	0.10	**0.60**	...	
Topic 3	0.02	0.05	...	
Topic 4	**0.30**	0.1	...	
...	...	...	...	...
Topic 15	**0.20**	0.01	...	**0.40**
...	...	...	...	...
Topic 28	0.01	0.03	...	**0.20**
Topic 29	**0.25**	0.15	...	
Topic 30	0.03	0.01	...	

Each "topic" is defined by **β**, a multinomial distribution over the entire vocabulary

Each document has associated **θ**, a multinomial distribution over topics

LDA Generative Story

Basic idea:

- Assume a story for generating our data (sampling from distributions)
- Estimate the parameters of the distribution
- [There are other approaches to topic modeling, this is specifically LDA]

Fightin' Words Generative Story

Generative story for log-odds with a Dirichlet Prior:

1. Draw $\boldsymbol{\pi}^{(i)} \sim$ Dirichlet($\boldsymbol{\alpha}$)

2. For $n^{(i)}$ steps:
    1. Draw $w \sim$ Multinomial($\boldsymbol{\pi}^{(i)}$)

Plate Notation: Log-odds with Dirichlet prior

- Shaded circle: value we observe
- Rectangles: values that are repeated (with number in corner reflecting # of repetitions)

α $\boldsymbol{\pi}^{(i)}$ w $n^{(i)}$ 2
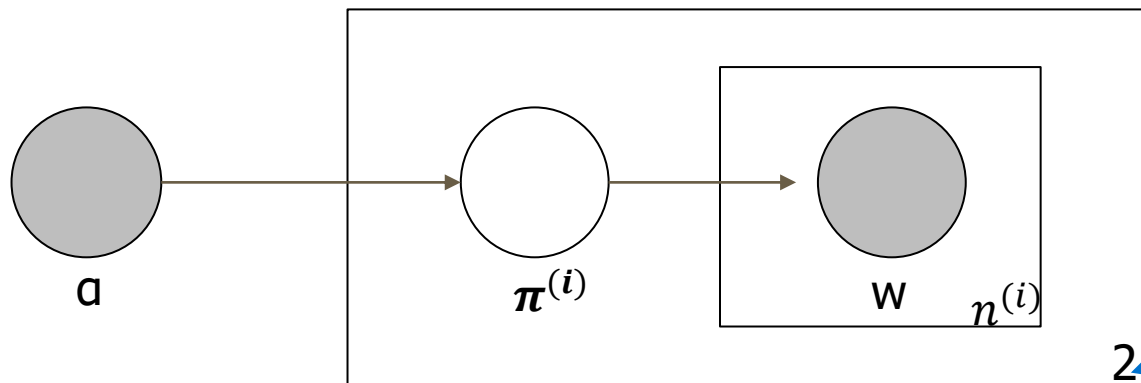
We drew $\boldsymbol{\pi}$ for Democrats and $\boldsymbol{\pi}$ for Republicans

LDA Generative Story

- For each topic k:
  - Draw $\beta_k \sim Dir(\eta)$
- For each document d:
  - Draw $\theta_d \sim Dir(\alpha)$
  - For each word in d:
    - Draw topic assignment $z \sim$ Multinomial(θ_d)
    - Draw $w \sim$ Multinomial(β_z)

We use the data to estimate these two sets of parameters:

- β, a distribution over vocabulary (1 for each topic)

- θ, a distribution over topics (1 for each document)

LDA Generative Story

- For each topic k:
  - Draw $\beta_k \sim Dir(\eta)$
- For each document d:
  - Draw $\theta_d \sim Dir(\alpha)$ → As long as θ_d is sparse, each document should be most affiliated with a few topics
  - For each word in d:
    - Draw topic assignment $z \sim Multinomial(\theta_d)$
    - Draw $w \sim Multinomial(\beta_z)$

The document's topic influences what words are in it:
- words that co-occur in the same document should end up affiliated with the same topic
- Documents with similar words will end up with similar topics
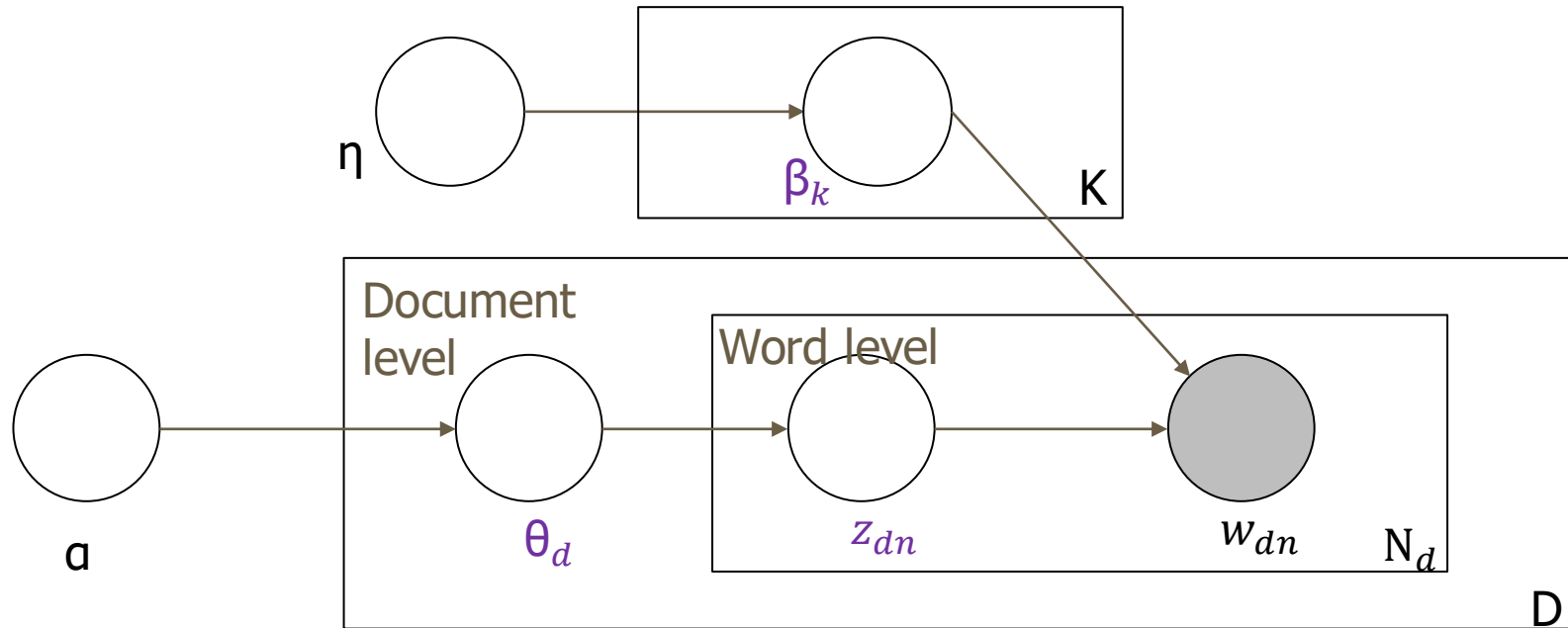
We use the data to estimate these two sets of parameters:

- β, a distribution over vocabulary (1 for each topic)

- θ, a distribution over topics (1 for each document)

η

β_k

K

Document level

Word level

α

θ_d

z_{dn}

w_{dn}

N_d

D
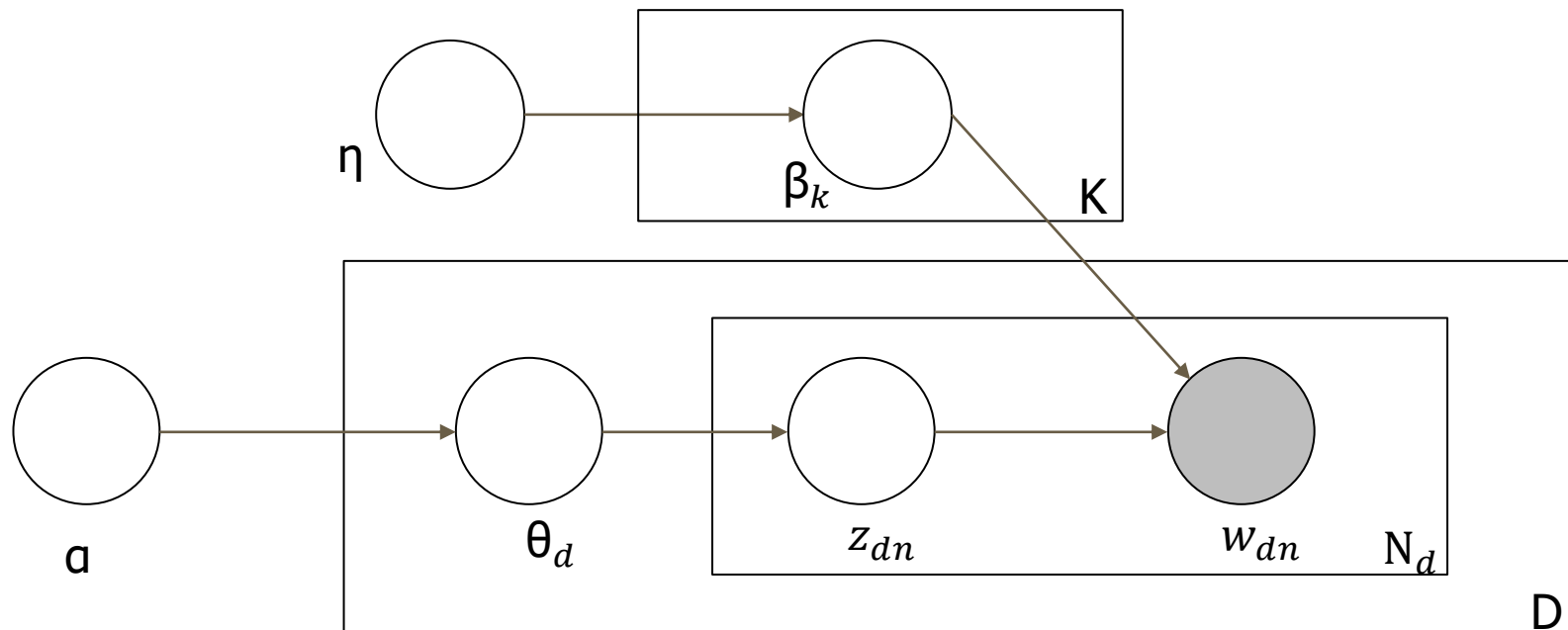
Variables we observe: D = number of documents; N = number of words per document, w words in document
Variables we want to estimate: θ, β, z are latent variables
Variables we choose: α, η are hyperparameters. K = number of topics

General Estimators [Heinrich, 2005]

Goal: estimate θ, β

$$p(\theta, \beta, z \,|w) = \frac{p(w\,|\theta, \beta, z)p(\theta, \beta, z)}{p(w)}$$

- MLE approach
  - Maximize likelihood: $p(w\,|\theta, \beta, z)$
  - MAP approach
  - Maximize posterior: $p(\theta, \beta, z\,|w)$ OR $p(w\,|\theta, \beta, z)p(\theta, \beta, z)$
- Bayesian approach
  - Approximate posterior: $p(\theta, \beta, z\,|w)$
  - Take expectation of posterior to get point estimates

LDA: Bayesian Inference

- Goal: estimate θ, β
- Bayesian approach: we estimate full posterior distribution

$$p(\theta, \phi, z \mid w) = \frac{p(w \mid \theta, \beta, z) p(\theta, \beta, z)}{p(w)}$$

$p(w)$ is the probability of your data set occurring under *any* parameters -- this is intractable!

Solutions: Gibbs Sampling, Variational Inference

Gibbs Sampling

Gibbs Sampling

- A Markov chain Monte Carlo (MCMC) algorithm
  - Algorithms for drawing samples from a probability distribution

- We draw samples by constructing a Markov Chain: the probably of the next sample is calculated from the previous sample

- We construct the chain so that if we draw enough samples (the "burn-in" period), we eventually start drawing samples from our real target distribution

- Once we have samples from the target distribution, we can use them to estimate the parameters we care about

Griffiths, Thomas L., and Mark Steyvers. "Finding scientific topics." Proceedings of the National academy of Sciences 101.suppl_1 (2004): 5228-5235.

Gibbs Sampling

- Assume we know topic assignments z for all words in the corpus

Vastly available digitized text data has created new opportunities for understanding social phenomena. Relatedly, social issues like toxicity, discrimination, and propaganda frequently manifest in text, making text analyses critical for understanding and mitigating them. In this

- We know how many times each word has been assigned to each topic

	Topic 1	Topic 2	Topic 3	...
Social	5	0	2	
analyses	10	3	2	
Discrimination	1	10	2	
...	...	...	...	

- We know how many times each topic has been assigned to each document

	Topic 1	Topic 2	Topic 3	...
Doc 1	11	7	30	
Doc 2	2	22	1	
Doc 3	16	15	17	
...	...	...	...	
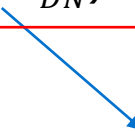
Gibbs Sampling

Vastly available digitized text data has created new opportunities for understanding social phenomena. Relatedly, social issues like toxicity, discrimination, and propaganda frequently manifest in text, making text analyses critical for understanding and mitigating them. In this

- One word at a time, remove the topic assignment and resample it

Remember the high-level: if we do this enough times, we start getting "good" topic assignments that we can use to estimate the parameters we care about

Griffiths, Thomas L., and Mark Steyvers. "Finding scientific topics." Proceedings of the National academy of Sciences 101.suppl_1 (2004): 5228-5235.
Why does this work? http://www.stat.columbia.edu/~gelman/book/BDA3.pdf Chap 11

Gibbs Sampling

▪ Initialize z (e.g. randomly)

▪ for t = 1 to T do: —————————————→ For each iteration
    ○ for d = 1 to D; for n = 1 to N_d do: —————→ For each word in the corpus
        • $z_{dn}^{(t+1)} \sim$ $P(Z_{dn} \mid z_{11}^{(t+1)} \ldots , z_{dn-1}^{(t+1)}, z_{dn+1}^{(t)}, \ldots , z_{DN}^{(t)})$
    ○ end for

▪ end for

Sample a new topic assignment

Gibbs Sampling

$$P(z_{dn} = k | z_{d,-n}, w, \alpha, \eta, \beta, \theta)$$

- We integrate out θ, β (we can do this because of conjugacy)

Number of times document d uses topic k

From prior

Number of times topic k uses word w_{dn}

From prior
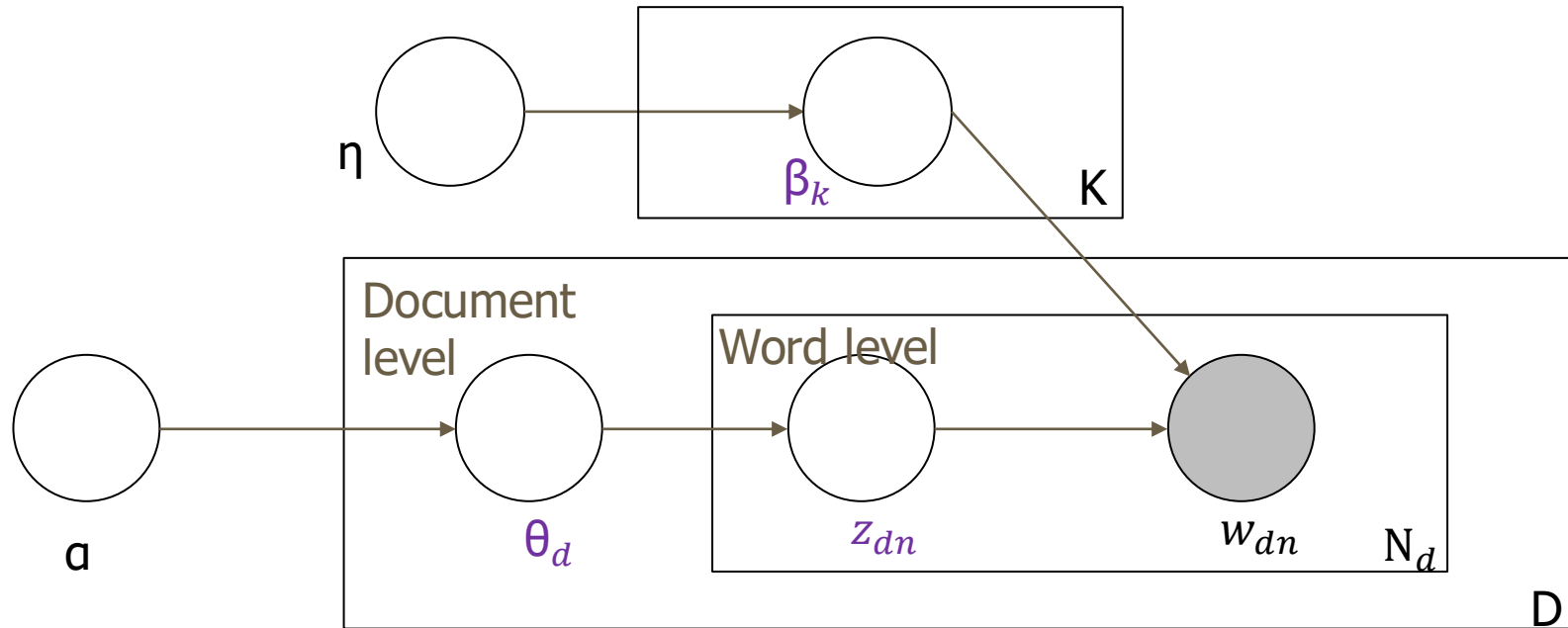
$$P(z_{dn} = k | z_{d,-n}, w, \alpha, \eta) = \frac{c_{dk} + \alpha_k}{\sum_i^K c_{di} + \alpha_i} \frac{v_{kw_{dn}} + \eta_{w_{dn}}}{\sum_i v_{ki} + \eta_i}$$

Variables we observe: D = number of documents; N = number of words per document, w words in document
Variables we want to estimate: θ, β, z are latent variables
Variables we choose: α, η are hyperparameters. K = number of topics

Gibbs Sampling

$$P\big(z_{dn} = k | z_{d,-n}, w, \alpha, \eta\big) = \frac{c_{dk} + \alpha_k}{\sum_i^K c_{di} + \alpha_i} \frac{v_{kw_{dn}} + \eta_{w_{dn}}}{\sum_i v_{ki} + \eta_i}$$

Prevalence of topic in document

Prevalence of word in topic

What make a topic k more likely to be assigned to z_{dn}? What properties does that mean we would expect to see in our final topic estimates?

What happens if α is very high?

Gibbs Sampling

- Initialize z (e.g. randomly)
- for t = 1 to T do:
  - for d = 1 to D; for n = 1 to N_d do:
    - $z_{dn}^{(t+1)} \sim \mathsf{P}(Z_{dn} \mid z_{11}^{(t+1)} \ldots, z_{dn-1}^{(t+1)}, z_{dn+1}^{(t)}, \ldots, z_{DN}^{(t)})$
  - end for
- end for

- We can similarly use counts of topic assignments across multiple samples to estimate (β, θ)

Recap

- Goal: estimate θ, β
- Bayesian approach: we estimate full posterior distribution

$$p(\theta, \beta, z \,|\, w) = \frac{p(w \,|\, \theta, \beta, z)\, p(\theta, \beta, z)}{p(w)}$$

Gibbs Sampling:

- We generate samples from the posterior distribution
- We estimate θ, β from those samples
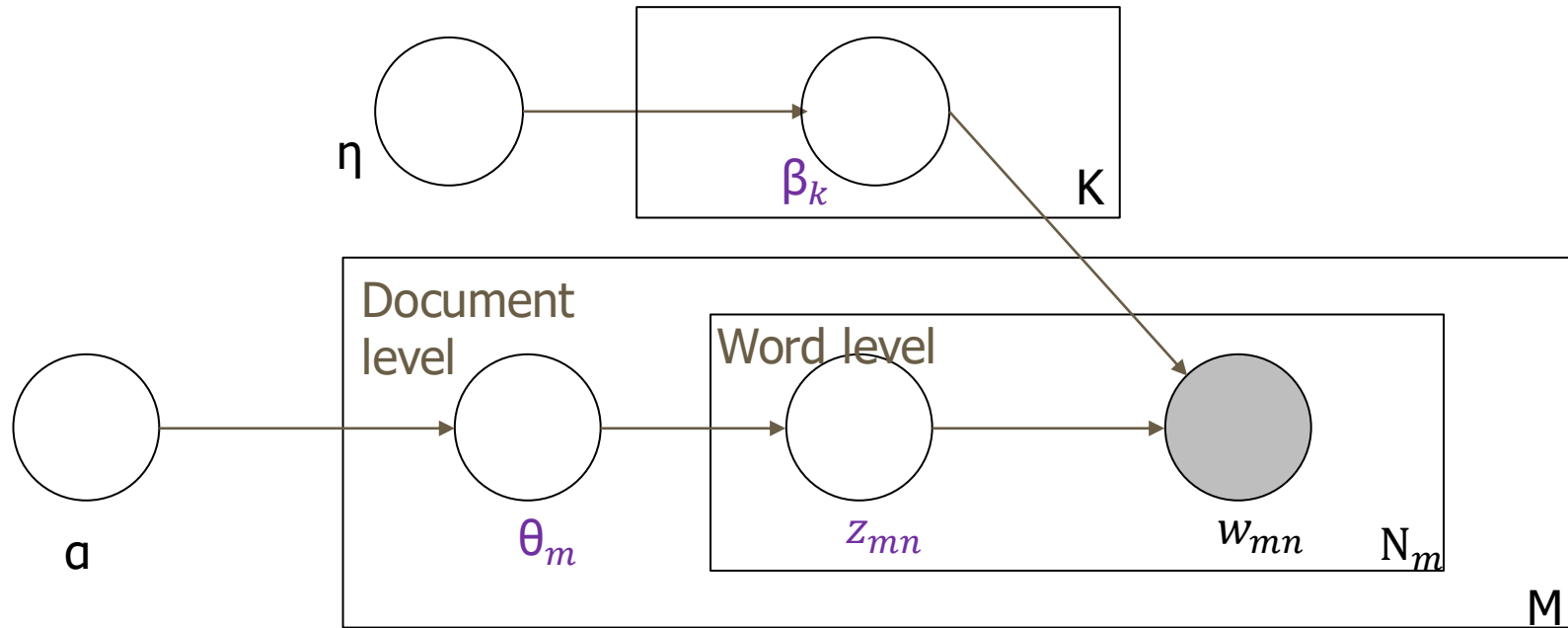
LDA In Practice

Variables we observe: M = number of documents; N = number of words per document, w words in document

Variables we want to estimate: θ, β, z are latent variables

Variables we choose: α, η are hyperparameters. K = number of topics

Choosing α, η and K

- In practice, typically choose *symmetric* Dirichlet priors, e.g. α, η = [1, 1, 1, 1,…], [0.1, 0.1, 0.1, 0.1,…] but some research has explored alternatives

- In practice, try a few K values and judge if topics look reasonable, but there are approaches that estimate the best value

Wallach, Hanna, David Mimno, and Andrew McCallum. "Rethinking LDA: Why priors matter." Advances in neural information processing systems 22 (2009).

Sample Topics from NYT Corpus

#5	#6	#7	#8	#9	#10
10	0	he	court	had	sunday
30	tax	his	law	quarter	saturday
11	year	mr	case	points	friday
12	reports	said	federal	first	van
15	million	him	judge	second	weekend
13	credit	who	mr	year	gallery
14	taxes	had	lawyer	were	iowa
20	income	has	commission	last	duke
sept	included	when	legal	third	fair
16	500	not	lawyers	won	show

How do we describe a topic?

- Most probable words for each topic

- Words common in this topic *relative* to other topics
    - We could use PMI scores!

- Examining documents that contain high proportion of topic

Jey Han Lau, David Newman, Sarvnaz Karimi, and Timothy Baldwin. 2010. Best topic word selection for topic labelling. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters (COLING '10). Association for Computational Linguistics, USA, 605–613.

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

LDA: Evaluation

- Held-out likelihood
    - Hold out some subset of your corpus
    - Compute the likelihood of the held-out data under the parameters you estimated
    - Says NOTHING about coherence of topics

Intruder Detection Tasks

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

6 / 10

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

Show entire excerpt

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

Key idea: If topics are coherent, annotators should easily be able to identify the intruder

Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. Reading tea leaves: how humans interpret topic models. In Proceedings of the 23rd International Conference on Neural Information Processing Systems (NIPS'09). Curran Associates Inc., Red Hook, NY, USA, 288–296.

LDA: Evaluation

- Can we automate these judgements?

- Word intrusion detection:
  - Compute PMI scores between each pair of words in the set of real and intruder words for each topic; train an SVM model to learn intruder words

- [Similar heuristics for topic intrusion detection]

- But follow-up work has suggested these kinds metrics don't always correlate with human judgement

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 530−539
Is Automated Topic Model Evaluation Broken? The Incoherence of Coherence by Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, Philip Resnik (NeurIPS, 2021)

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Practical advice for getting coherent topics

- Evaluate the topics by hand
- Hyperparameter selection (α, η, K):
    o Test different numbers of topics
    o Tune the parameter controlling the topic distributions
- Pay attention to your data:
    o Better for long documents
    o Keep stopwords, don't stem
    o Remove (high numbers of) duplicates
- Just use LDA!

https://maria-antoniak.github.io/2022/07/27/topic-modeling-for-the-people.html
Pulling out the stops: Rethinking stopword removal for topic models by Alexandra Schofield, Måns Magnusson, and David Mimno (EACL, 2017)

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

LDA: Advantages and Drawbacks

- When to use it
  - Initial investigation into unknown corpus
  - Concise description of corpus (dimensionality reduction)
  - [Features in downstream task]
- Limitations
  - Can't apply to specific questions (completely unsupervised)
  - Simplified word representations
    - BOW model
    - Can't take advantage of similar words
  - Strict assumptions
    - E.g. Independence assumptions

Today's takeaways

- Motivation behind topic modeling
- **LDA formulation**
- LDA inference
  - Gibbs sampling (overview of process, e.g. slide 24)
- Evaluation and practical considerations

- Next class:
  - Variational inference
  - LDA extensions
  - More in-depth use cases

Logistics

- HW 1 to be released by the end of this week
- We haven't yet covered all of it, but we have covered part 1

References

1. Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Journal of machine Learning research* 3.Jan (2003): 993-1022.

Optional sources for more depth:

- Gibbs Sampling:
  - Jordan Boyd-Graber's Introduction: https://www.youtube.com/watch?v=u7l5hhmdc0M
  - https://api.drum.lib.umd.edu/server/api/core/bitstreams/a36ce44d-0732-427d-8a81-a18c9b0b4dfa/content
  - UMD Technical Report: https://drum.lib.umd.edu/items/d5aa258e-d2ac-4529-8831-ec0e08a5f2cc

- Variational Inference
  - Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe. "Variational inference: A review for statisticians." Journal of the American statistical Association 112.518 (2017): 859-877.
  - Xanda Schofield and Jordan Boyd-Graber
    - https://www.youtube.com/watch?v=-tKmyHoVZ-g
    - https://www.youtube.com/watch?v=smfWKhDcaoA
  - David Blei Lecture Notes https://www.cs.princeton.edu/courses/archive/fall11/cos597C/lectures/variational-inference-i.pdf