Topic models

William Lowe

Hertie School

6th October 2020

Plan

- → Geometry
- → Topic models, the very idea
- → Latent Dirichlet Allocation
- → Dirichlet distributions for beginners
- → Training and output
- → Interpretation
- → Explaining topic prevalence
- → The Structural Topic Model

Word geometry

Consider the i-th row of a document term matrix as a vector of proportions $P_i = [P_{i1} \ldots P_{iV}]$

- → (Make the proportions from counts W_{iv} by dividing every count by the row sum)

This is the document's vocabulary profile

For fixed vocabulary, all possible documents live on a *simplex*

- → The 'corners' are where the document contains 100% tokens of one vocabulary word
- → There are V of them, i.e. a lot…

Topic geometry

Now consider the message $\theta_i = [\theta_{i1} \ldots \theta_{iK}]$ that we think the i-th document expresses.

This is a vector of K *topic proportions* for document i

→ From applying a dictionary
→ From manual coding
→ …

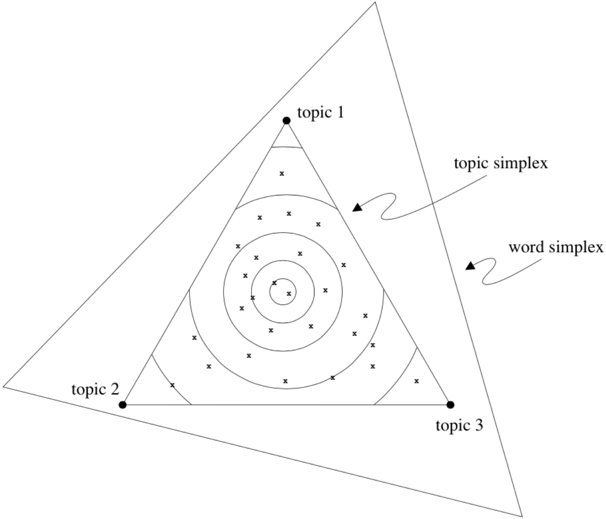
This is the document's topic profile

For fixed vocabulary, all possible documents also live on a *simplex*, but it

→ only has K corners
→ is embedded in the space of document profiles

Huge dimensionality reduction…

Topic geometry



Topic geometry

The ‘corners’ of the topic simplex represent the balance of words this topic generates

→ i.e. a dictionary category

but with two differences to the ones we’ve been looking at. Each one

→ contains an ‘entry’ for every possible word, not just some of them
→ expressed generatively, so we could ‘write’ on that topic
(by treating the entry as a set of multinomial parameters)

Topic models

Topic models answer the question: If we only knew K

- → What would a good set of topics be?
- → What would be a good set of topic proportions be

Basically

- → $B = [\beta_1, \ldots, \beta_K]$ a matrix of word probabilities (V words by K topics)
- → $\Theta = [\theta_1, \ldots, \theta_D]^T$ a matrix of topic proportions (D documents by K topics)

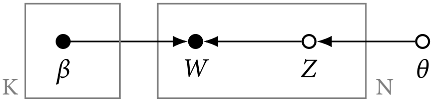
We'll look at the oldest and simplest way to answer these questions

- → Latent Dirichlet Allocation (a.k.a LDA, Blei et al., 2003)

Dictionaries vs topic models

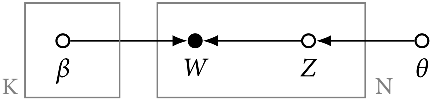
Dictionary-based content analysis

We know the dictionary



Topic model

We *don't* know the dictionary

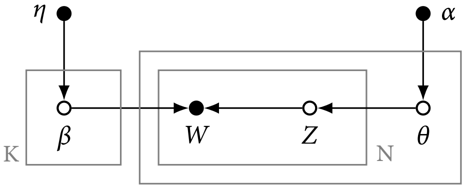


Latent Dirichlet Allocation

We *don't* know the dictionary, but we do have *prior expectations*

→ about the βs

→ about the θs



$$\beta_k \sim \text{Dirichlet}(\eta) \qquad \theta_d \sim \text{Dirichlet}(\alpha)$$

$$W_i \sim \text{Multinomial}(\beta_{Z_i=k}, 1) \qquad Z_i \sim \text{Multinomial}(\theta_d, N)$$

Dirichlet distributions

Two sources of information to figure out β

- → Prior: how 'sparsely' populated is each topic with words?
- → Likelihood: what words W to we see?

Two sources of information to figure out θ

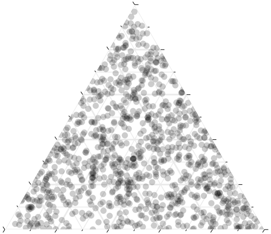
- → Prior: how 'sparsely' are documents populated with topics
- → Likelihood: what topics Z do we see?

We use a Dirichlet distribution for both priors:

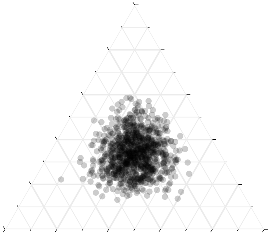
- → Dirichlet distributions generate probabilities i.e. profiles on a simplex

Dirichlet parameters

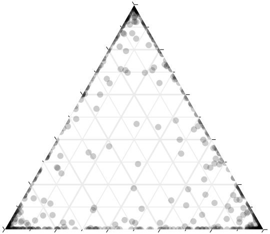
$\alpha = [1, 1, 1]$



$\alpha = [10, 10, 10]$



$\alpha = [0.1, 0.1, 0.1]$



Why Dirichlet?

Because Multinomial!

Why Dirichlet?

Because Multinomial! Nice things about the Dirichlet distribution

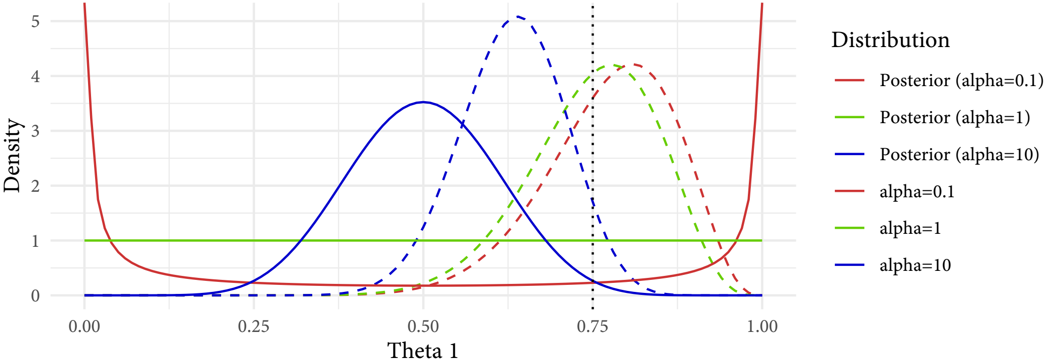
→ It's *conjugate* to the Multinomial

Let's (briefly) consider a special case: Beta (Dirichlet with 2 probabilities) and Binomial (Multinomial with 2 outcomes)

→ *two* possible topics (1 and 0) in proportions $[\theta_1, (1-\theta_1)]$ and a single document with 20 words whose Zs are known: 15 Z=1s and 5 Z=0s

$$\begin{aligned}
P(\theta_1) &\propto \theta_1^{\alpha_1-1}(1-\theta_1)^{\alpha_2-1} && \text{Beta} \\
P(Z_1 \ldots Z_{20} \mid \theta_1) &\propto \prod_i^{20} \theta_1^{Z_i}(1-\theta_1)^{Z_i} && \\
&\propto \theta_1^{15}(1-\theta_1)^{5} && \text{Binomial} \\
P(\theta_1 \mid Z_1 \ldots Z_{20}) &\propto \theta_1^{(15+\alpha_1-1)}(1-\theta_1)^{(5+\alpha_2-1)} && \text{Beta again!}
\end{aligned}$$

Why Dirichlet?



Prior distributions are solid lines and the resulting posterior distributions over θ for 15 Z=1 and 5 Z=0 are dashed.

Topic model training

Topic models can be quite time consuming to estimate.

→ Lots of coupled unknowns all at once

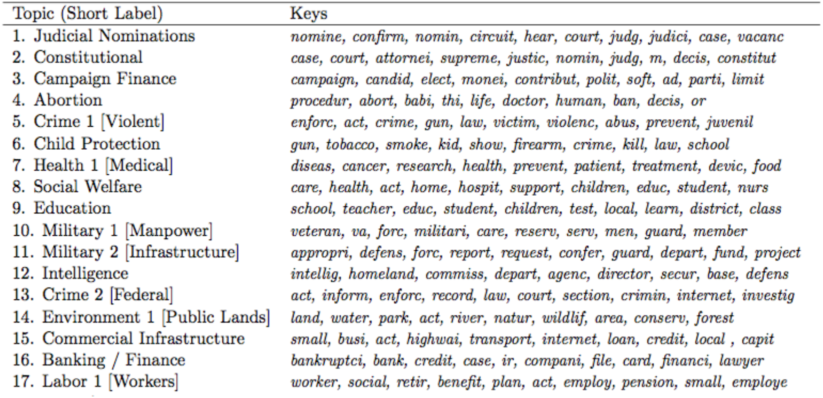
Intuition:

→ Any set of parameters make the observed word counts more or less probable
→ If we knew the Z's then estimating β and θ would be straightforward
→ If we new β and θ then estimating Z would be straightforward
→ So alternate between these steps

This simple approach is called *Gibbs sampling*

A more complete machine learning course will tell you all about; we won't linger…

Topic models: β

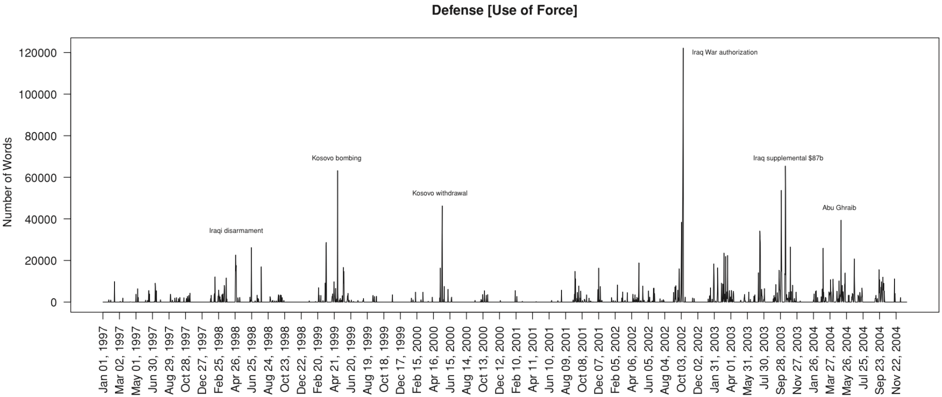


Topic (Short Label)	Keys
1. Judicial Nominations	*nomine, confirm, nomin, circuit, hear, court, judg, judici, case, vacanc*
2. Constitutional	*case, court, attornei, supreme, justic, nomin, judg, m, decis, constitut*
3. Campaign Finance	*campaign, candid, elect, monei, contribut, polit, soft, ad, parti, limit*
4. Abortion	*procedur, abort, babi, thi, life, doctor, human, ban, decis, or*
5. Crime 1 [Violent]	*enforc, act, crime, gun, law, victim, violenc, abus, prevent, juvenil*
6. Child Protection	*gun, tobacco, smoke, kid, show, firearm, crime, kill, law, school*
7. Health 1 [Medical]	*diseas, cancer, research, health, prevent, patient, treatment, devic, food*
8. Social Welfare	*care, health, act, home, hospit, support, children, educ, student, nurs*
9. Education	*school, teacher, educ, student, children, test, local, learn, district, class*
10. Military 1 [Manpower]	*veteran, va, forc, militari, care, reserv, serv, men, guard, member*
11. Military 2 [Infrastructure]	*appropri, defens, forc, report, request, confer, guard, depart, fund, project*
12. Intelligence	*intellig, homeland, commiss, depart, agenc, director, secur, base, defens*
13. Crime 2 [Federal]	*act, inform, enforc, record, law, court, section, crimin, internet, investig*
14. Environment 1 [Public Lands]	*land, water, park, act, river, natur, wildlif, area, conserv, forest*
15. Commercial Infrastructure	*small, busi, act, highwai, transport, internet, loan, credit, local , capit*
16. Banking / Finance	*bankruptci, bank, credit, case, ir, compani, file, card, financi, lawyer*
17. Labor 1 [Workers]	*worker, social, retir, benefit, plan, act, employ, pension, small, employe*

From Quinn et al. (2007)

Note: only the top most probable words are shown and topic labels are manually assigned.

Topic models: θ_k



From Quinn et al. (2007)

Interpreting topics

Ideally we'd like to be able to say: "make this one about defense"

Unfortunately, that level of high level control is an unsolved problem

We can only *after the fact* assign our own labels the topics, and hope some are topics that we want.

Are they good, these topics?

→ often better the *statistical* properties of the model the less interpretable it tends to be (Chang et al., 2009)

→ Clearly we're missing something with the model structure…

Explaining topic prevalence

Often we want to both *measure* but also *explain* the prevalence of topic mentions

Example: What are the effects of a Japanese house electoral reform on candidate platforms? (Catalinac, 2018)

- → Fit a topic model to LDP platforms
- → Extract two topics that look like ‘pork’ and ‘policy’
- → Average these per year and plot
- → Compare relative prevalence to electoral change timeline

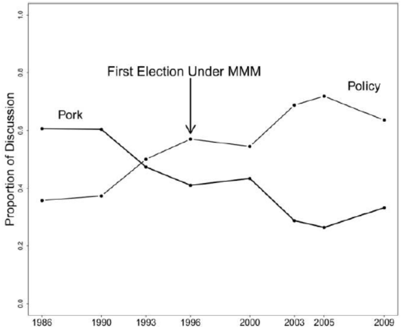


Figure 1. LDP candidates switched to more policy and less pork in the 1993 election and continued with this strategy under MMM. This figure plots the mean proportions of discussion devoted to pork and policy, respectively, in the 2,355 manifestos produced by LDP candidates in these eight elections.

Structural topic model

If we like some of the topics, we might want to know how they vary with external information, e.g.

→ How does rate of topic 3, say 'defence', change with the party of the speaker?

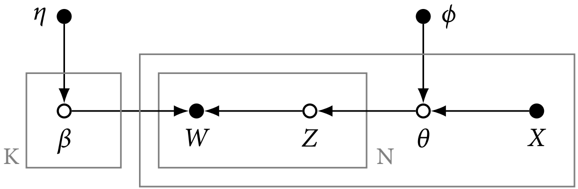
Structural topic model

If we like some of the topics, we might want to know how they vary with external information, e.g.

→ How does rate of topic 3, say 'defence', change with the party of the speaker?

This is a regression model (Roberts et al., 2014) with

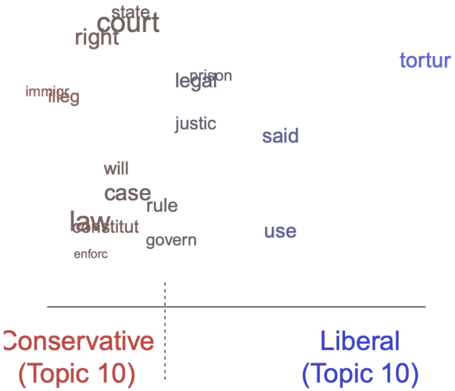
→ speaker party indicator, convariates etc. as X (observed)
→ proportion of the speech assigned to topic 3 as θ_3 (inferred, not observed)
→ The words W (observed)



Structural topic model

Having a topic model allows us to get contrast vocabulary *within topic* too.

Here’s contrasting usage when talking about Guatanamo Bay in a Bush era data set



Even more topic models

There's a small industry developing new types of topic model

→ A brief search will acquaint you with more than enough to play with

Check if they have stable code!

We'll look closer into topic model evaluation and related matters next week!

References

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). ‘Latent dirichlet allocation’. *Journal of Machine Learning Research*, *3*, 993–1022.

Catalinac, A. (2018). ‘Positioning under alternative electoral systems: Evidence from japanese candidate election manifestos’. *American Political Science Review*, *112*(1), 31–48.

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C. & Blei, D. M. (2009). ‘Reading tea leaves: How humans interpret topic models’. *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems*, 288–296.

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H. & Radev, D. R. (2007). ‘How to analyze political attention with minimal assumptions and costs’. *American Journal of Political Science*, *54*(1), 209–228.

References

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B. & Rand, D. G. (2014). ‘Structural topic models for open-ended survey responses’. *American Journal of Political Science*, *58*(4), 1064–1082.