Lecture 1: Topic Models in Social Science

Miriam Hurtado Bodell

LINKÖPING UNIVERSITY

Recap

LDA is a method for discovering latent semantic structures in documents

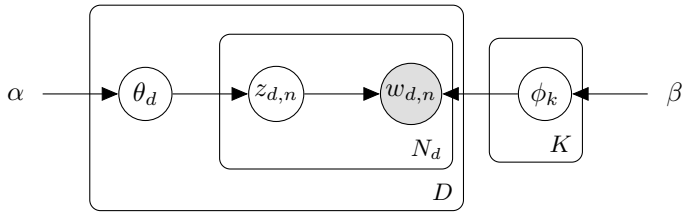


Figure: Generative processes for Latent Dirichlet Allocation

Recap

LDA is a method for discovering latent semantic structures in documents

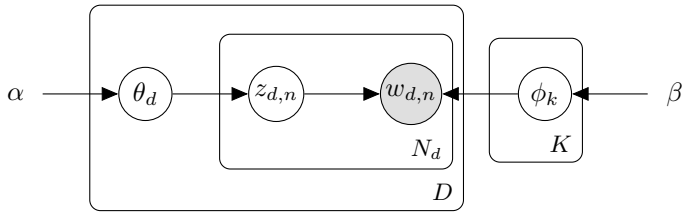


Figure: Generative processes for Latent Dirichlet Allocation

$\rightarrow$ unsupervised method

Topic models in the social sciences

What are topics?

- Technical answer: a distribution over words
- Sociological answer: ?

Topic models in the social sciences

What are topics?

- Technical answer: a distribution over words
- Sociological answer: **theoretically construct**

Examples:

- **Frames** (DiMaggio, Nag & Blei 2013)

"Immigrants are taking our jobs"

"The fact that immigrants are over represented in prisons say something about them."

Topic models in the social sciences

What are topics?

- Technical answer: a distribution over words
- Sociological answer: **theoretically construct**

Examples:

- **Frames** (DiMaggio, Nag & Blei 2013)
- Language domains within academic fields (McFarland et al. 2013)
- Literary themes (Jockers & Mimno 2013)

You have to make a compelling argument that topic = theoretical construct of interest

Topic models in the social sciences

When does it make sense to use topic models?

- When theoretical construct of interest can be conceptualized as a topic/relation of topics
- When documents are a meaningful unit of analysis
- When documents can be viewed as having mixed memberships
- You are open to letting the data help you identifying topics of relevance
- When polysemy is important/present

Topic models in the social sciences

How can we use topic models?

Discovery

- Topics model inductively identifies latent variables: Θ (document-topic distr.), Φ (topic-word distr.), and $z_{n,d}$ (token topic indicators)

  $\rightarrow$ can be used to *explore*/summarize topical structures in corpora

Measurement

- Document classification
- Topic models can be used to *measure* the degree to which a theme is present in corpora

Example **discovery**: DiMaggio, Nag & Blei (2013)

Aim: Explore the frames in five U.S. newspapers articles of public funding of the arts, 1986–1997 [*note: no clear RQ/hypotheses!*]

Method: LDA with $k = 12$ on $\sim$ 6000 newspaper articles

Frame = topic

Results:

- Media frames on public art funding become more focused on controversy/conflict, art become part of "culture wars"
- Differences between news outlets (conservative news focus on polarization)

Example **discovery**: DiMaggio, Nag & Blei (2013)

Topic model take away: Topic models can be used totally explorative if researcher has understanding of context and can make sense of the result

Paper makes case for great overlap between cultural sociological ideas and topic model methodologies

Example **measurement**: Fligstein, Brundage & Schultz (2017)

RQ: Why did the Federal Open Market Committee (FOMC) fail to see the 2007/2008 financial crisis before it happened?

Hypotheses:

1. FOMC primarily used (a) macroeconomic/ (b) financial frame in discussions, because
2. FOMC members with a private banking experience primarily use financial frame (but were in the minority)
3. FOMC exhibited "positive asymmetry" (normalized negative scenarios)
4. Macroeconomic focus limited FOMC's ability to see sources/consequences of the crisis

Example **measurement**: Fligstein, Brundage & Schultz (2017)

Data: Transcripts from 72 FOMC meetings 2000–2008

Method:

- Run LDA with $K = 15$ to measure frames (choose α and β based on interpretability)
- Study distribution of macroeconomic/financial terms across topics
- Manually curated list of words: macroeconomic (e.g. interest, growth, inflation, employment) and finance (e.g. bank, finance, debt, liquidity).

$\rightarrow$ **Primary frame = manual theme words occurs in multiple topics**

Example **measurement**: Fligstein, Brundage & Schultz (2017)

Results:

- Finds that macroeconomics frames are more common across topics
- Members with private banking background are more likely to use financial frame
- Support for hypotheses 3 & 4 via close reading

Topic model take away: *Very simple model* combined with expert knowledge of context/theories can lead to publication in top journal!

Limitations of LDA topic models in the social sciences

Bag-of-words assumption is too simplistic – syntax matters!

Difficult to get reliable and transparent interpretations of topics

- Topics will change between model runs
- Would different people give the topic the same name? (reading tea leaves)
- How do we know that the topics of interest will show up?

Topic models in social sciences

Augmenting social scientists when using topic models, computational grounded theory (Nelson 2017):

1. Inductive computational method (topic model)
2. Deep/close reading (refinement of pattern/hypotheses)
3. Deductive computational method for out-of-sample test

Limitations of LDA topic models in the social sciences

Bag-of-words assumption is too simplistic – syntax matters!

Difficult to get reliable and transparent interpretations of topics

- Topics will change between model runs
- How do we know that the topics of interest will show up?

Number of topics (K) is fixed and needs to be set *a priori* to running the model

Model validation?

Short interlude on model validation in social sciences

Statistical: coherence, likelihood, etc.

Internal: compare with humanly annotated data

External: does topic salience react to external events, predict unseen data

→ best practices are still being developed!

Extensions of LDA topic models in social sciences

Structural topic model (Roberts et al. 2016): study impact of document-level covariates and metadata on topic prevalence & topical content (prior of topic over docs and words over topics document-dependent)

Example RQs:

- How much to party X vs party Y talk about topic T?
- What characterizes topic T for newspaper X and newspaper Y?

Note on caution: implemented using (mean-field) variational inference!

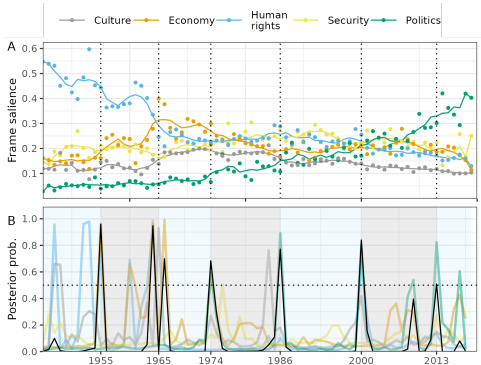
Extensions of LDA topic models in social sciences

Seeded topic models (Watanabe 2023): allow for researcher to decide *a priori* what topics the model should find

Example RQs:

- How much do newspapers discuss immigration?
- How often is immigration and crime discussed together compared to immigration and religion?

Seeded topic models: example from my own research



Summary

- Social scientists use estimated topics from topic models to represent theoretically relevant concepts
  - Perhaps most notably: frames
- Works especially well when important that both documents and words can be "polysemic" (have multiple meanings)
- Been criticized for their lack of transparency and reliability, and difficulty to validate
- *Old, but useful!*

Let's get started!

www.liu.se

LINKÖPING UNIVERSITY