Structural Topic Models

Wouter van Atteveldt

CCS Hannover, Feb 2018
Many topic models

- Many extensions to plain LDA
- Variations on plate model structure or assumptions
- Can overcome some of LDA’s limitations
- Often no (good) R support
  - (Except Structural Topic Models)
Correlated Topic Models

Figure 1: Top: Graphical model representation of the correlated topic model. The logistic normal distribution, used to model the latent topic proportions of a document, can represent correlations between topics that are impossible to capture using a single Dirichlet. Bottom: Example densities of the logistic normal on the 2-simplex. From left: diagonal covariance and nonzero-mean, negative correlation between components 1 and 2, positive correlation between components 1 and 2.

Figure 2: A portion of the topic graph learned from 15,744 OCR articles from *Science*. Each node represents a topic, and is labeled with the five most probable words from its distribution; edges are labeled with the correlation between topics.
Dynamic Topic Models

Many topic models

1881 brain movement action right eye hand left muscle nerve sound
1890 movement eye right hand brain sound experiment
1900 movement brain sound nerve active muscle left eye right nervous
1910 movement sound muscle active nerve response fiber reaction brain response
1920 stimulate muscle sound movement response nerve frequency fiber active brain
1930 random sound stimulate muscle active nerve frequency fiber active brain
1940 record nerve stimulate response muscle electrode active frequency electrode potential
1950 response record stimulate condition active potential stimulus nerve subject eye
1960 response cell potential stimulus neuron active nerve eye response abstract
1970 response cell potential stimulus neuron active nerve cell receptor muscle response current
1980 cell channel neuron Ca2+ active brain stimulus neuron active nerve eye synapse signal
1990 neuron active brain cell fig response channel receptor synapse signal
2000 neuron active brain cell response channel receptor synapse signal

"Neuroscience"

1887 Mental Science
1900 Hemicrania in Migraine
1912 A Defence of the "New Phrenology"
1921 The Synchronous Flashing of Fireflies
1932 Myoesthesia and Imageless Thought
1943 Acetylcholine and the Physiology of the Nervous System
1952 Brain Waves and Unit Discharge in Cerebral Cortex
1963 Errorless Discrimination Learning in the Pigeon
1974 Temporal Summation of Light by a Vertebrate Visual Receptor
1983 Hysterectomy in the Force-Calcium Relation in Muscle
1993 GABA-Activated Chloride Channels in Secretery Nerve Endings
Hierarchical Topic Models

Figure 5: A topic hierarchy estimated from 1717 abstracts from NIPS01 through NIPS12. Each node contains the top eight words from its corresponding topic distribution.
Structural Topic Models

http://www.structuraltopicmodel.com/
Roberts et al. (2014). Structural Topic Models for Open Ended Survey Responses. AJPS.
Structural Topic Models

- We often have information about the document
  - Article metadata (year, source)
  - Speaker data (year, party, gender)
  - Respondent data for open questions
- Allow topic model to use that data
  - e.g. topic proportions change over time
  - e.g. words differ between speakers, parties
Structural Topic Models

- We often have information about the document:
  - Article metadata (year, source)
  - Speaker data (year, party, gender)
  - Respondent data for open questions
- Allow topic model to use that data:
  - e.g. topic proportions change over time
  - e.g. words differ between speakers, parties
- Additional data:
  - $D \times P$ topic prevalence covariates
  - $D$ topic content groups
Generative Process

- "Normal" LDA, but
- Topic proportions $\theta$ drawn from logistic normal GLM with covariates
  - Topics can correlate with each other, with covariates
- Topic words $\beta$ multinomial logistic model with covariates
  - $\beta_k \propto \exp(m + \kappa(topic) + \kappa(cov) + \kappa(int))$
  - So word chosen can also depend on
    - covariate (Republicans talk about freedom)
    - interaction (republicans use 'obamacare' to talk about ACA)
Example: newspaper perspectives

(Roberts, Stewart, Airoldi 2017)
Example: newspaper perspectives

(Roberts, Stewart, Airoldi 2017)
Example: Fatwas

Lucas et al 2015, Nielsen 2014
Example: Fatwas
Lucas et al 2015, Nielsen 2014

100 Topics Occuring in "Normal" Fatwas (Jihad Score < 0)

**Favorite Jihadi Topics**
- Shariah
- The Prophet
- Ibn Taymiyya
- Ablutions
- Money
- Prayer
- Permissibility
- Heaven and Hell
- Hajj
- Duty

**Evenly Split Topics**
- Sin
- Sheikh Uthaymeen
- God's Oneness
- Quran
- Knowledge
- Apostasy
- Quran
- Ulama
- Heaven and Earth
- Knowledge

**Favorite Non-Jihadi Topics**
- Hadeeth
- Dating
- Zakat
- Surahs and Verses
- Hadeeth
- Ramadan Fasting
- Hadeeth
- Divorce, Marriage, Sex
- Fatwa Greeting Formula

Bootstrapped 95% Confidence Interval
Money, Pilgrimage, and Marriage
stm: Structural Topic Models in R

Intro

Structural Topic Models

STM

STM: Structural Topic Models in R

Ingest

Process

Estimate

Evaluate

Understand

Visualize

Extensions

stmBrowser

stmCorrViz

TextProcessor

readCorpus

prepDocuments

plotRemoved

stm

searchK

manyTopics

multiSTM

selectModel

permutationTest

labelTopics

findThoughts

estimateEffect

topicCorr

plot.STM

cloud

plotQuote

plot.estimateEffect

plot.topicCorr

Structural Topic Models

Wouter van Atteveldt
Estimating an stm

stm(dfm, K)
stm(dfm, K, prevalence =~ year)
stm(dfm, K, content =~ president)
Inspecting an stm

```
labelTopics(m)
findThoughts(m, texts=texts, n=1, topics=1)
plot(m, type = "summary")
cloud(m, topics=3)
plot(m, type = "perspectives", topics = c(12, 20))
```
Covariate effect

\[
\text{eff} = \text{estimateEffect}(\text{topics} \sim \text{covariates}, \ m, \\
\quad \text{docvars(dfm)})
\]

\text{summary(eff, topics=1)}
\text{plot(eff, "president", type="difference",}
\quad \text{topics=c(1,4,5), model=m)}
\text{plot(eff, "year", type="continuous",}
\quad \text{topics=1, model=m)}
\text{plot(m, type = "perspectives", topics = 11)}
Inspection topic correlations

plot(topicCorr(m3))
Helping with model selection

models = selectModel(..., runs=20)
plotModels(models)

searchK(K=c(10, 20, 30), ...)
stm(K=0, ...)

Structural Topic Models in R
Conclusion

- Topic modeling as dimensionality reduction
- Topics should make sense (have coherence and meaning)
- Preprocessing, hyperparameters are important!
- LDA interpretable process, (often) plausible results
- Many extensions exist
  - Of which \textit{stm} has best R support