Twitter in the Parliament - A Text-based Analysis of German Political Entities

Patrick Schulze, Simon Wiegrebe

Supervisors:
Prof. Dr. Christian Heumann, Prof. Dr. Paul W. Thurner

7. Juli 2020

Covariate-level Topic Analysis

Overview

- Explore estimated topical structure with respect to different dimensions, e.g. membership in political party, time, ...
- Precisely: examine relationship between document-level prevalence covariates $\boldsymbol{x}_d$ and topic proportions $\boldsymbol{\theta}_d$
- Natural idea: regress topic proportions on prevalence covariates
- Problem: $\boldsymbol{\theta}_d$ is *latent* variable and has to be estimated itself!
- In following two approaches to address this problem:
  1. Regression that takes into account uncertainty about $\boldsymbol{\theta}_d$: perform sampling technique known as "method of composition" in social sciences
  2. Direct assessment of STM output via logistic normal distribution with estimated topical prevalence parameters $\hat{\boldsymbol{\Gamma}}$ and $\hat{\boldsymbol{\Sigma}}$

Covariate-level Topic Analysis

Method of Composition

- Let $\boldsymbol{\theta}_{(k)} := (\theta_{1,k}, \ldots, \theta_{D,k})^T \in [0,1]^D$ denote proportion of k-th topic for all D documents
- Method of Composition (repeat m times):
  1. Sample $\boldsymbol{\theta}^*_{(k)}$ from (variational) posterior of $\boldsymbol{\theta}_{(k)}$ estimated by STM
  2. Run regression model with response $\boldsymbol{\theta}^*_{(k)}$ and covariates $\boldsymbol{X}$ to obtain estimate $\hat{\boldsymbol{\xi}}^*$ of regression coefficients $\boldsymbol{\xi}^*$ and covariance of $\hat{\boldsymbol{\xi}}^*$, $\hat{\boldsymbol{V}}^*_{\xi}$
  3. Sample $\tilde{\boldsymbol{\xi}}^*$ from $F(\hat{\boldsymbol{\xi}}^*, \hat{\boldsymbol{V}}^*_{\xi})$, where F is (asymptotic) distribution of $\hat{\boldsymbol{\xi}}^*$
- Idea: samples $\tilde{\boldsymbol{\xi}}^*$ take into account uncertainty in $\boldsymbol{\theta}_{(k)}$
- Visualization of topic-metadata relationship: For observation $\boldsymbol{x}_{\text{pred}}$, plot $\boldsymbol{x}_{\text{pred}}$ vs. predicted response with $\boldsymbol{x}^T_{\text{pred}}\tilde{\boldsymbol{\xi}}^*$ as linear predictor

Covariate-level Topic Analysis

Method of Composition: Problems

Several concerns with method of composition:

1. In STM, regression model in step 2 is OLS; however OLS not appropriate to model (sampled) proportions in open unit interval
2. Mixing of Bayesian and frequentist approach questionable:
   - From Bayesian perspective, $\tilde{\boldsymbol{\xi}}^*$ can only be considered sample from posterior of $\boldsymbol{\xi}$ in certain Bayesian regression models with questionable (uniform) prior assumptions
   - Using $\boldsymbol{x}_{\text{pred}}^T \tilde{\boldsymbol{\xi}}^*$ as linear predictor does *not* yield sample of posterior predictive distribution
3. Separate modeling of topic proportions neglects dependence of different topics among each other

Covariate-level Topic Analysis

Method of Composition: Usage within R Package *stm*

- Problem: OLS regression not suitable for (sampled) proportions, which are restricted to interval (0,1)
- Estimated relationship between proportions and prevalence covariates might even involve negative proportions:

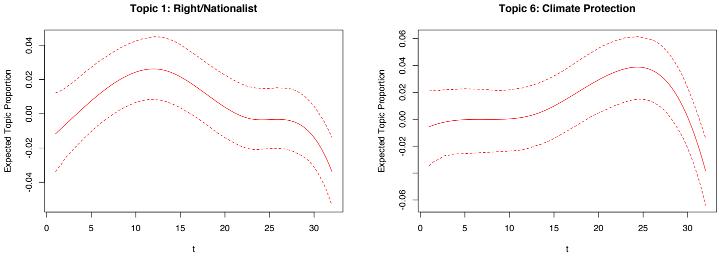

Figure: Emprical mean and 95% credible intervals for topics 1 and 6 over time, estimated using *estimateEffect* from the *stm* package.

Covariate-level Topic Analysis

Method of Composition: Extension of existing approach

- Instead of OLS regression, we can use a beta regression or a quasibinomial GLM (both with logit-link) to adequately model proportions

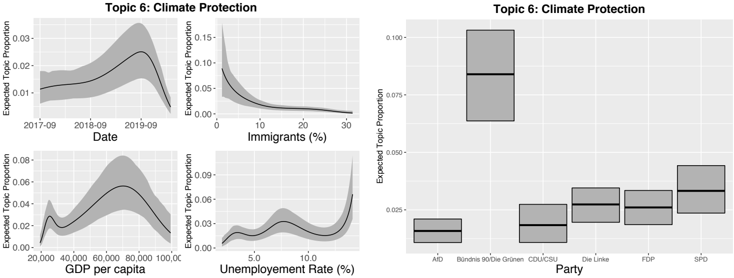

Figure: Empirical mean and 95% credible intervals, obtained using a quasibinomial GLM.

Covariate-level Topic Analysis

Mixing of Bayesian and Frequentist Approach

- Regression within of method of composition is *frequentist* regression
- However, in STM $\tilde{\boldsymbol{\xi}}^*$ considered samples from (marginal, i.e., integrated over latent topic proportions) posterior of regression coefficients; only true by assuming uniform priors for $\boldsymbol{\xi}$
- Caution: uncertainty from previous plots with respect to prediction of mean $\Rightarrow$ does *not* reflect variation of topic proportions in data!
- Better idea: fully Bayesian approach with more realistic priors and sampling from posterior predictive distribution to reflect variation of data

Covariate-level Topic Analysis

Fully Bayesian Approach: Idea

- Idea: *explicitly* perform Bayesian regression in second step of each iteration of method of composition
- Modeling via beta regression (with normal priors centered around zero) in order to model proportions in $(0, 1)$
- Visualization: Sample proportions from posterior predictive distribution at end of each step of method of composition (i.e., conditioning on previously sampled $\boldsymbol{\theta}^*_{(k)}$) with covariate values $\boldsymbol{x}_{\text{pred}}$

Covariate-level Topic Analysis

Fully Bayesian Approach: Results

- Predicted (empirical) mean mostly in line with results from previous analysis
- Uncertainty now w.r.t. variation of topic proportions in data
- Observed variation for topic proportions corresponds well to variation according to predictive posterior

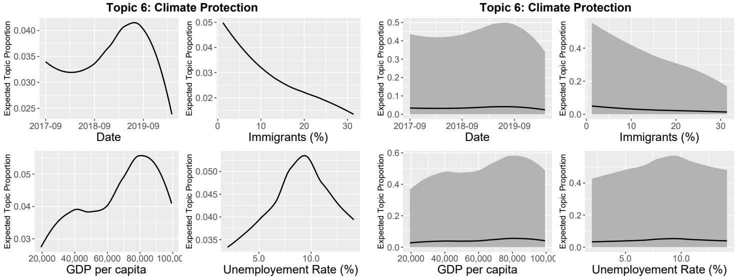

Figure: Smooth effects without credible intervals (left) and smooth effects with 95% credible intervals (right)

Covariate-level Topic Analysis

Multivariate Modeling of Proportions

- Remember, by assumption: $\boldsymbol{\theta}_d \sim \text{LogisticNormal}(\boldsymbol{\Gamma}^T \boldsymbol{x}_d^T, \boldsymbol{\Sigma})$
- Logistic normal distribution assumes high dependence among individual components $\Rightarrow$ not fully taken into account in univariate modeling via, e.g., the beta distribution
- Inference within STM involves finding estimates $\hat{\boldsymbol{\Gamma}}$ and $\hat{\boldsymbol{\Sigma}}$ $\Rightarrow$ Idea: plug estimates into logistic normal distribution
- For given covariate value $\boldsymbol{x}_{\text{pred}}$, obtain topic proportion as $\boldsymbol{\theta}_d^* \sim \text{LogisticNormal}(\hat{\boldsymbol{\Gamma}}^T \boldsymbol{x}_{\text{pred}}^T, \hat{\boldsymbol{\Sigma}})$

Covariate-level Topic Analysis

Multivariate Modeling of Proportions

- Plugging in $\mathbf{\Gamma}$ and $\hat{\mathbf{\Sigma}}$ is "naïve" method: ideally sample prevalence parameters from their posterior $\Rightarrow$ would yield higher variation
- However, not easily possible $\Rightarrow$ should be addressed in future implementations

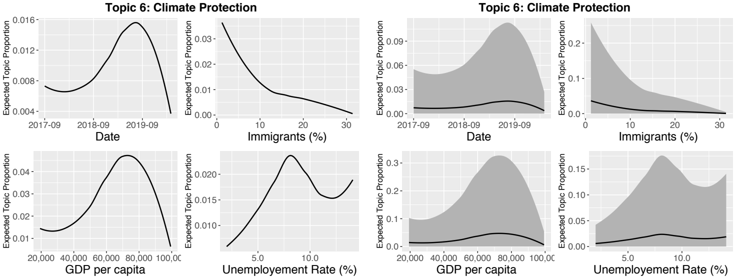

Figure: Smooth effects without credible intervals (left) and smooth effects with credible intervals (right)

Bibliography