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
  - In standard regression analysis, dependent variable is realization of random variable
  - In STM, however, we have access to posterior of topic proportions $\boldsymbol{\theta}_d$
  - If we "naïvely" use mean/mode of this posterior as dependent variable of regression, much information is lost
  - Solution: perform sampling technique known as "method of composition" in social sciences
- Alternatively: direct assessment of logistic normal distribution with estimated topical prevalence parameters $\hat{\boldsymbol{\Gamma}}$ and $\hat{\boldsymbol{\Sigma}}$

Covariate-level Topic Analysis

Method of Composition: Usage within R Package *stm*

- 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 estimates of regression coefficients $\hat{\boldsymbol{\xi}}^*$ 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 $\boldsymbol{x}_{\text{pred}}^T \tilde{\boldsymbol{\xi}}^*$ as linear predictor

Problems

1. In STM, regression model in step 2 is OLS; however OLS not appropriate to model proportions
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.
3. Using $\boldsymbol{x}_{\text{pred}}^T\tilde{\boldsymbol{\xi}}^*$ as linear predictor does *not* yield sample of posterior predictive distribution
4. Separate modeling of topic proportions neglects dependence among variables

Covariate-level Topic Analysis

Method of Composition: Usage within R Package *stm*

- Notation:
  - $\boldsymbol{\theta}_{(k)} := (\theta_{1,k}, \dots, \theta_{D,k})^T \in [0,1]^D$: proportion of k-th topic for all D documents
  - $q(\boldsymbol{\theta}_{(k)}|\boldsymbol{X}, \boldsymbol{W})$: approximate variational posterior of $\boldsymbol{\theta}_{(k)}$
  - $q(\hat{\boldsymbol{\xi}}|\boldsymbol{X}, \boldsymbol{\theta}_{(k)})$: (normal) distribution of estimated regression coefficients $\hat{\boldsymbol{\xi}}$ from OLS regression $\boldsymbol{\theta}_{(k)} = \boldsymbol{X}\boldsymbol{\xi} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})$
- Method of composition:
  1) Draw $\boldsymbol{\theta}^*_{(k)} \sim q(\boldsymbol{\theta}_{(k)}|\boldsymbol{X}, \boldsymbol{W})$.
  2) Draw $\hat{\boldsymbol{\xi}}^* \sim q(\hat{\boldsymbol{\xi}}|\boldsymbol{X}, \boldsymbol{\theta}^*_{(k)})$.
- It then holds that $\hat{\boldsymbol{\xi}}^*_1, \dots, \hat{\boldsymbol{\xi}}^*_m$ is an i.i.d. sample from the marginal posterior of regression coefficients

$$q(\boldsymbol{\xi}|\boldsymbol{X}, \boldsymbol{W}) = \int_{\boldsymbol{\theta}_{(k)}} q(\boldsymbol{\xi}|\boldsymbol{X}, \boldsymbol{\theta}_{(k)}) q(\boldsymbol{\theta}_{(k)}|\boldsymbol{X}, \boldsymbol{W}) \mathrm{d}\boldsymbol{\theta}_{(k)}$$

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)

$\Rightarrow$ Estimated relationship between proportions and prevalence covariates might involve negative estimated proportions

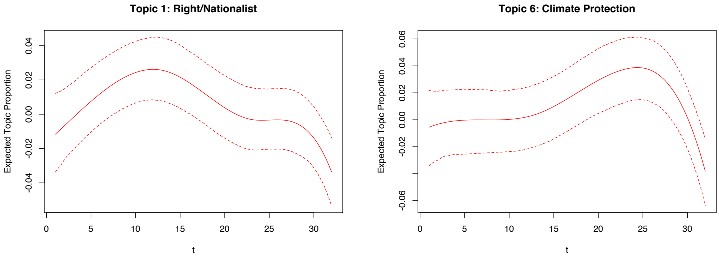

Figure: Emprical mean and 95% credible intervals for topics 1 and 6 over time, estimated using *estimateEffect* from the *stm* package.

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
- In this case, regression coefficients are *asymptotically* normally distributed

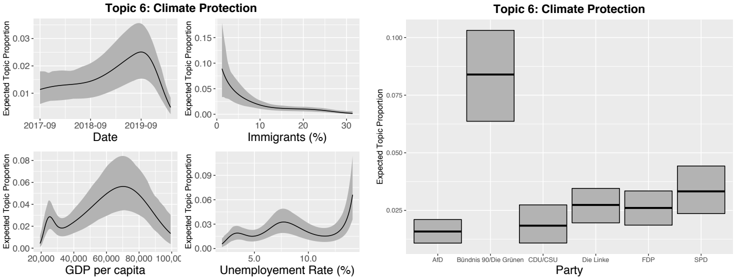

Figure: Empirical mean and 95% credible intervals, obtained using a quasibinomial GLM.

Covariate-level Topic Analysis

Problem: Univariate 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
- However, regression within method of composition uses *univariate* k-th topic proportion as dependant variable
- Problem with this approach: dependence among components neglected $\Rightarrow$ especially uncertainty estimates are unrealistic

Covariate-level Topic Analysis

Multivariate Modeling via Logistic Normal Distribution

- Inference within STM involves finding estimates $\hat{\mathbf{\Gamma}}$ and $\hat{\mathbf{\Sigma}}$
- Idea: plug estimates into logistic normal distribution $\Rightarrow$ for a given covariate value $\boldsymbol{x}_d^*$, "predict" topic proportion as $\boldsymbol{\theta}_d^* \sim \text{LogisticNormal}(\hat{\mathbf{\Gamma}}^T(\boldsymbol{x}_d^*)^T, \hat{\mathbf{\Sigma}})$
- Ideally, we would apply fully Bayesian approach and sample from (variational) posterior of $\mathbf{\Gamma}$ (and update $\mathbf{\Sigma}$, which is obtained via MLE) $\Rightarrow$ "Predictive Posterior" of topic proportions
- However, output obtained using R package *stm* does not allow for simple implementation of such a procedure (i.e., sampling from variational posterior of $\mathbf{\Gamma}$ and updating $\mathbf{\Sigma}$); yet, possible in theory!

Covariate-level Topic Analysis

Multivariate Modeling via Logistic Normal Distribution

- Still, our results suggest a high discrepancy between:
  - Distribution of topic proportions assumed in generative process of STM
  - Impression we gain of this distribution via separate modeling of topics.
- Fully Bayesian approach would most likely yield even higher uncertainty

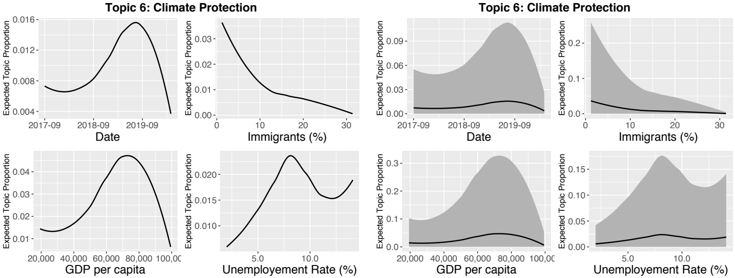

Figure: Smooth effects without credible intervals (left) and smooth effects with credible intervals (right)

Bibliography