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

Topic Modeling: Motivation and Theory

Motivation

- bla

Topic Modeling: Motivation and Theory

Structural Topic Model (STM)

- Topic model that incorporates document-level metadata:
  - *Topical prevalence* covariates $\boldsymbol{X} = [\boldsymbol{x_1}|\ldots|\boldsymbol{x_D}]^T \in \mathbb{R}^{D\times P}$
  - Categorical *topical content* variable $\boldsymbol{Y} \in \mathbb{R}^D$ with A levels, i.e., $Y_d \in \{1,\ldots,A\}$, for all $d \in \{1,\ldots,D\}$
- Generative process for each document $d \in \{1,\ldots,D\}$:
  1) Draw $\boldsymbol{\eta_d} \sim \mathcal{N}_{K-1}(\boldsymbol{\Gamma}^T \boldsymbol{x_d}^T, \boldsymbol{\Sigma})$, with $\eta_{d,K} = 0$ for model identifiability.
  2) Normalize $\boldsymbol{\eta_d}$, for all $k \in \{1,\ldots,K\} : \theta_{d,k} = \frac{exp(\eta_{d,k})}{\sum_{j=1}^{K} exp(\eta_{d,j})}$.
  3) For each word $n \in \{1,\ldots,N_d\}$:
     a) Draw topic assignment $\boldsymbol{z_{d,n}} \sim \text{Multinomial}_K(\boldsymbol{\theta}_d)$.
     b) If no topical content variable specified: $w_{d,n} \sim \text{Multinomial}_V(\boldsymbol{\beta}_{d,n})$.
     c) Otherwise, determine document-specific word distributions $B_a := [\boldsymbol{\beta}_1^a|\ldots|\boldsymbol{\beta}_K^a]$ based on $Y_d = a$, for all topics $k \in \{1,\ldots,K\}$; select $\boldsymbol{\beta}_{d,n} := B_a \boldsymbol{z_{d,n}}$; and draw word $w_{d,n} \sim \text{Multinomial}_V(\boldsymbol{\beta}_{d,n})$.

Topic Modeling: Motivation and Theory

Graphical Model of the STM

- Again, we can visualize the generative process using the representation as a graphical model:

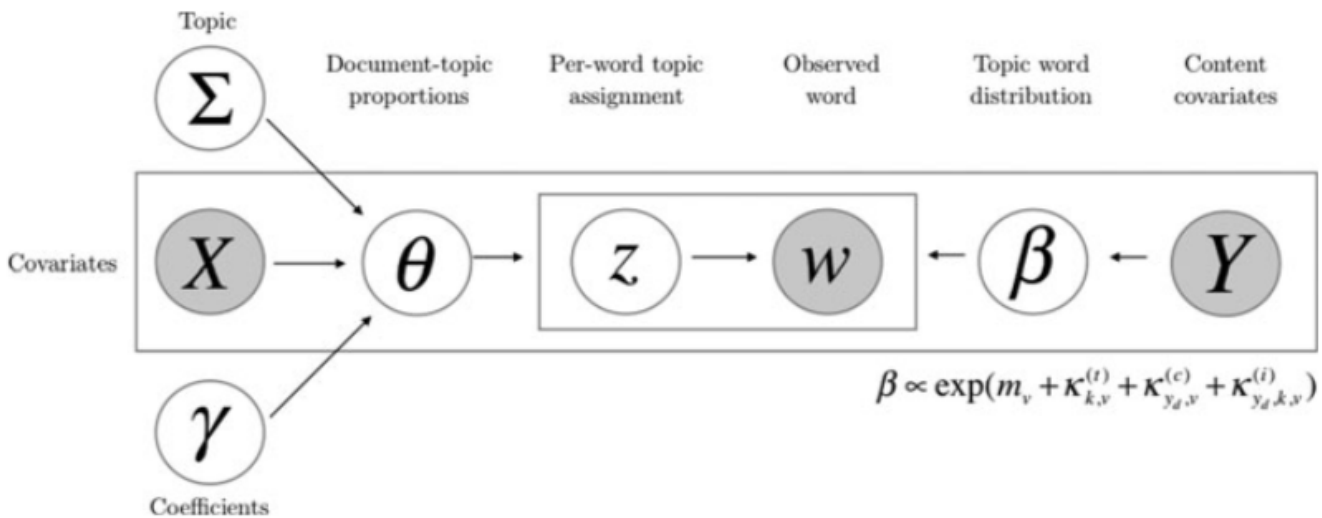

Figure: Graphical Model of the STM

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*

- Notation:
  - $\boldsymbol{\theta}_{(k)} := (\theta_{1,k}, \ldots, \theta_{D,k})^T \in [0,1]^D$: 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 $\hat{\boldsymbol{\xi}}^*_1, \ldots, \hat{\boldsymbol{\xi}}^*_m$ 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

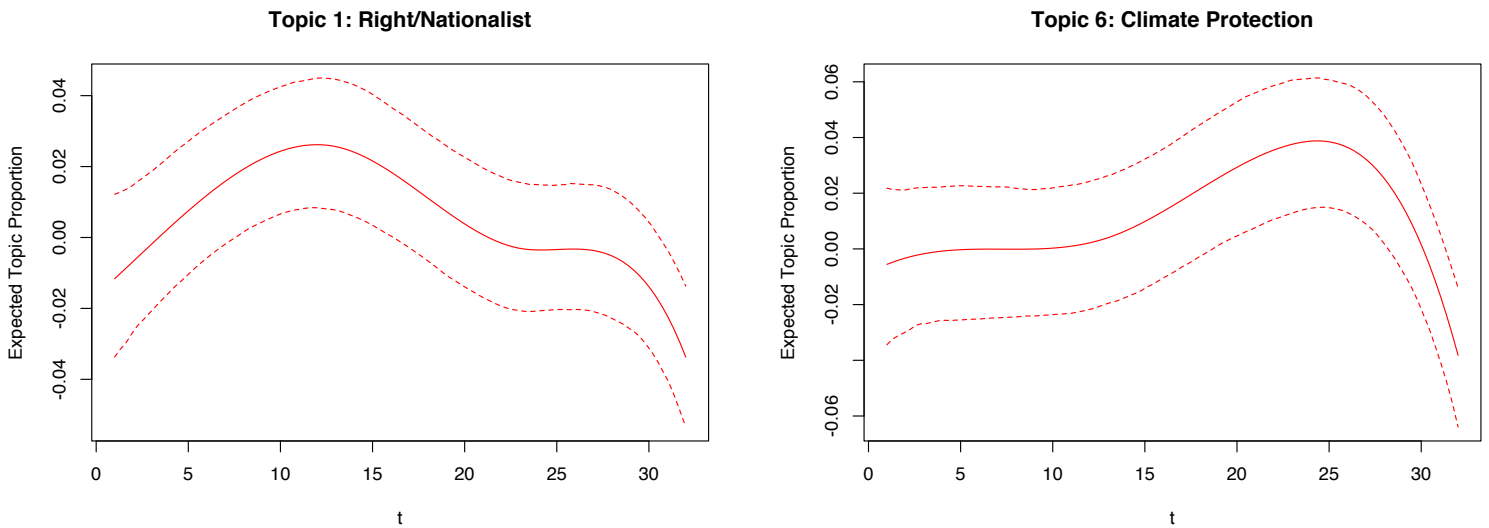

Figure: Emprical mean and 95% credible intervals for topics 1 and 6 over time, estimated using *estimateEffect* from the *stm* package.

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

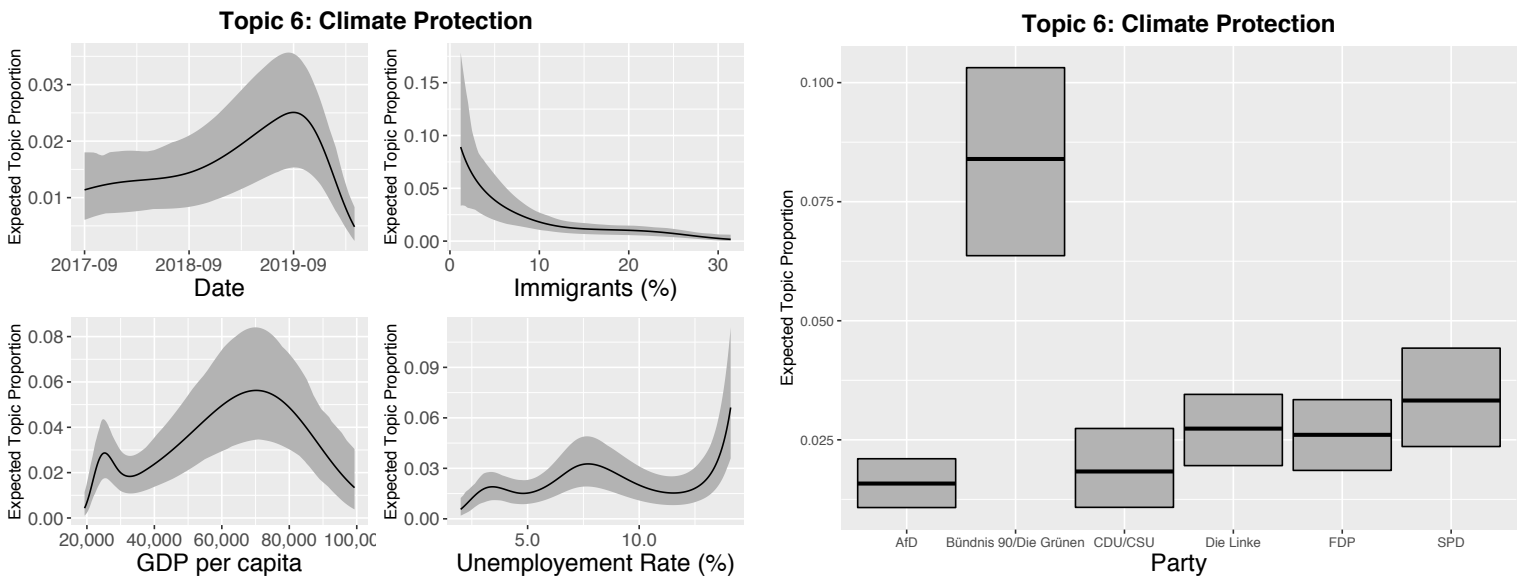

Figure: Empirical mean and 95% credible intervals, obtained using a quasibinomial GLM.

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

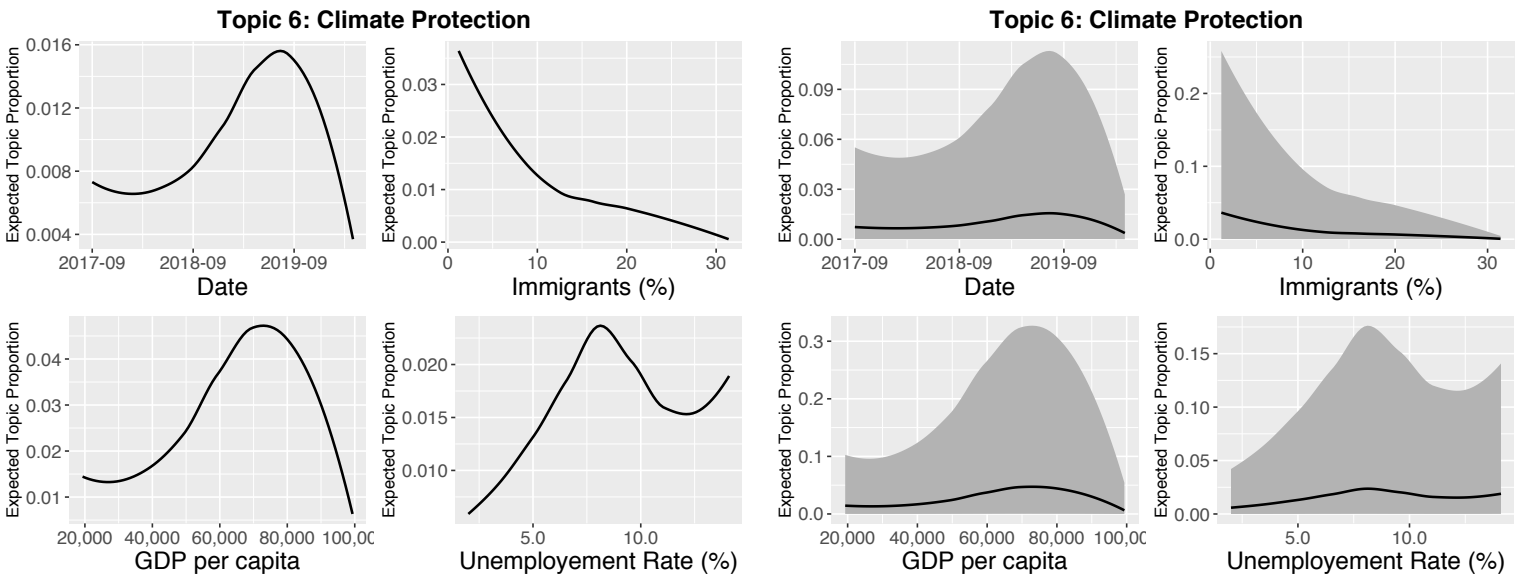

Figure: Smooth effects without credible intervals (left) and smooth effects with credible intervals (right)

Causal Inference

Correlation vs. Causality

- In previous section: assessment of relationship between metadata and topic proportions
- As stated, framework should be used to *explore* topics with respect to different dimensions
- In particular, *causal* interpretation of results is generally not justified ("correlation vs. causality")
- When making causal inference, we have to consider that topic proportions are *latent* variables
- Possible solution: conduct a train-test split

Causal Inference

Identification Problem and Overfitting

- Assume there are two groups, a treatment group and a control group
- Aside from treatment, individuals from both groups are similar
- Objective: quantify treatment effect, in our case effect of treatment on prevalence of specific topic.
- Necessary assumption: response of an individual depends only on treatment of this individual
- *Identification problem*: estimating topic model to discover latent topic proportions can introduce additional dependency among individuals $\Rightarrow$ response of each individual is *not* only determined by treatment of that individual!
- *Overfitting*: fitted topic model might mistake noise for patterns in some way $\Rightarrow$ response again not solely determined by treatment of an individual, but additionally by specific characteristics of other individuals.

Causal Inference

Train-test split

- Idea: split data $\mathcal{D}$ into training set $\mathcal{D}_{\text{train}}$ and test set $\mathcal{D}_{\text{test}}$.
- Training set $\mathcal{D}_{\text{train}}$ used to determine a model that infers latent topic proportions from a given text
- Test set $\mathcal{D}_{\text{test}}$ used in order to assess relation between *predicted* test set topic proportions and test set prevalence covariates.
- Solves identification problem: model used for prediction is determined by training set observations $\Rightarrow$ treatment of test set observations not dependent on other individuals' treatment from test set.
- Overfitting also solved: noise from training set is very unlikely to be replicated on test set

Causal Inference

Implementation within the STM

- Input documents, i.e., words and metadata from the training set $\mathcal{D}_{\text{train}}$, and obtain estimates $(\hat{\boldsymbol{\beta}}_{\text{train}}, \hat{\boldsymbol{\Gamma}}_{\text{train}}, \hat{\boldsymbol{\Sigma}}_{\text{train}})$ using the STM
- Then, estimate (variational) posterior of test set topic proportions, conditional on the model parameters $(\hat{\boldsymbol{\beta}}_{\text{train}}, \hat{\boldsymbol{\Gamma}}_{\text{train}}, \hat{\boldsymbol{\Sigma}}_{\text{train}})$ from training set $\mathcal{D}_{\text{train}}$ as well as words $\boldsymbol{W}_{\text{test}}$ from test set $\mathcal{D}_{\text{test}}$
- Estimation of (variational) posterior conditional on data and training set parameters occurs via E-step of (variational) EM algorithm
- Benefit of using the STM: covariate information from training set directly used to predict topic proportions on test set
- Important: Covariate information from test set must not be used! Otherwise, for two documents from test set with exact same words, different topic proportions are predicted if prevalence covariates differ. However, in such a case causal effect should to be zero.

Causal Inference

Results

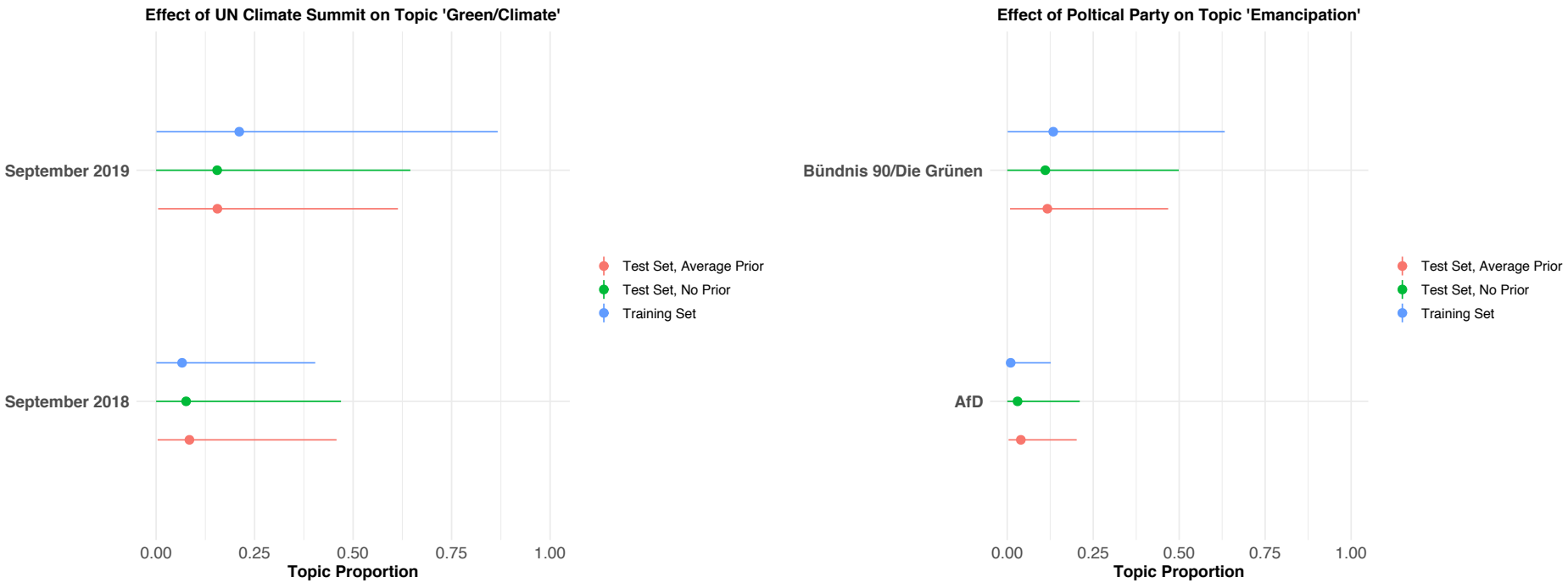

Figure: Maximum-a-posteriori (MAP) estimates of topic proportions on training and test data. Points display the mean, lines 2.5% and 97.5% credible intervals.

Causal Inference

Results

- UN Climate Action Summit 2019 was held on September 23, 2019
- As observed, topic associated with climate issues was discussed to much larger extent during that time than the year before
- While MAP estimates for different prior specifications on test set are rather similar, estimated effect for training data is much larger
- For effect of political party on topic labelled as 'Emancipation', we find similar results: average difference of estimated topic proportions between both parties is larger for the training data
- Further, note that credible intervals on the training data differ compared to credible intervals on the test data in both cases.

- To estimate the treatment effect, we determine the average difference of predicted topic proportions between both groups:

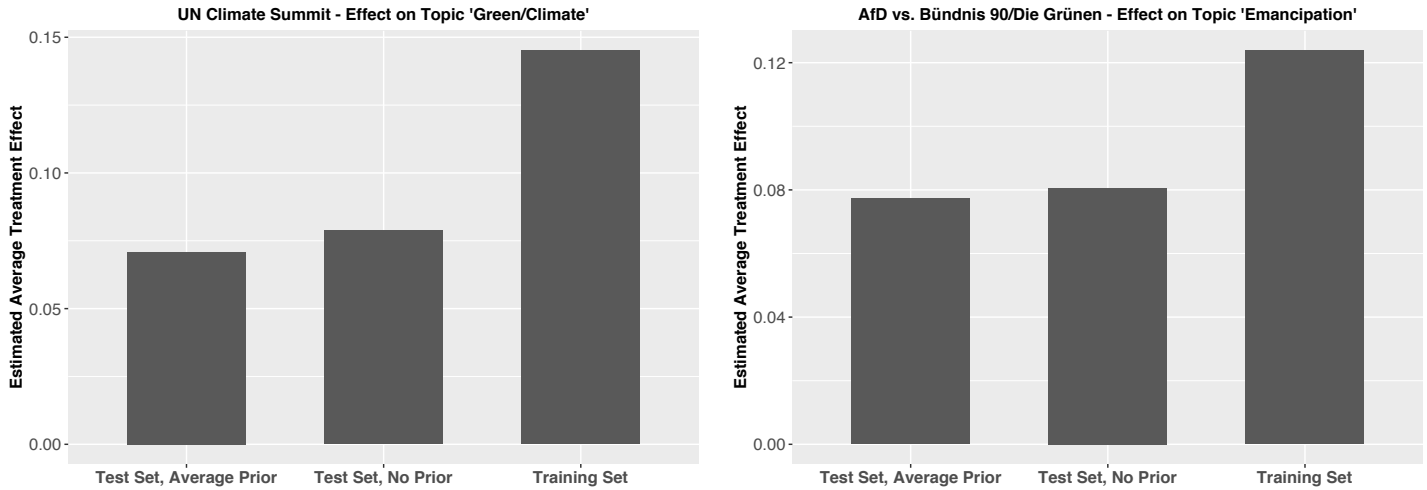

Figure: Estimated Average Treatment Effects (ATE) using training and test data.

- In both cases treatment effect is larger if "naïvely" estimated solely on training data!

Bibliography