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

Patrick Schulze, Simon Wiegrebe

Project partners: *Prof. Dr. Paul W. Thurner, Sandra Wankmüller* (Geschwister Scholl Institute of Political Science, LMU)

Supervisors: *Prof. Dr. Christian Heumann, Matthias Aßenmacher*

July 16, 2020

Outline

Introduction

Introduction

- Huge amounts of data, especially text, produced by social media
- Field of particular interest in the context of social media and big data: *Politics*
    - e.g., Brexit, 2016 presidential election in the US, Facebook data scandal
- Tools of analysis for such data simultaneously provided by advances in *Natural Language Processing* (NLP)
- *Topic analysis*: analytical tool for discovery and exploration of latent thematic clusters within text

Introduction

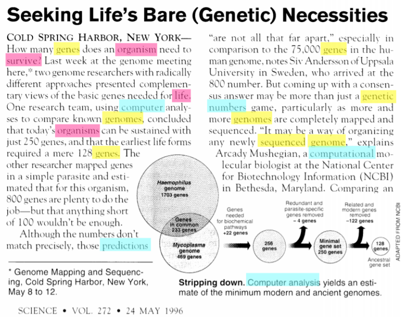
- Key contributions of this project:
  - Construction of dataset containing Twitter posts by members of the German Bundestag and a variety of metadata
  - Application of the *Structural Topic Model* (STM), introduced by (**roberts2016model**), to German MPs' Twitter communication
  - Development of new tools for estimation of relationship between topic proportions and metadata
  - Application of STM-specific train-test split to enable causal inference

Topic Modeling: Motivation and Theory

Topic Modeling: Motivation and Theory

Motivation

- Motivating example: excerpt from a scientific article **blei2012presentation**



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

- Question at hand: how to assign colored words to topics?

Topic Modeling: Motivation and Theory

Notation and Terminology (I)

- *Words* w: instances of a vocabulary of V unique *terms*
- *Documents* $d \in \{1, \ldots, D\}$: sequences of words of length N_d; $w_{d,n}$ denoting n-th word of document d
- *Corpus*: collection (or set) of D documents
- *Topics* $k \in \{1, \ldots, K\}$: latent thematic clusters within a text corpus; (implicit) representation of a corpus
- *Topic-word distributions* β: probability distributions over words; β_k denoting the word distribution corresponding to the k-th topic

Topic Modeling: Motivation and Theory

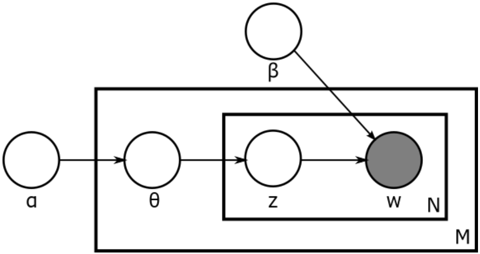
Notation and Terminology (II)

- *Topic assignments* $\boldsymbol{z}_{d,n}$: assignment of $w_{d,n}$ to a specific topic $k \in \{1, \ldots, K\}$; $\boldsymbol{\beta}_{d,n}$ representing the (assigned) word distribution for $w_{d,n}$
- *Topic proportions* $\boldsymbol{\theta}_d$: proportions of document d's words assigned to each of the topics; $\sum_{k=1}^{K} \theta_{d,k} = 1$, for all $d \in \{1, \ldots, D\}$
- *Bag-of-word* assumption: only words themselves meaningful, unlike word order or grammar; equivalent to assuming *exchangeability* **aldous1985exchangeability**

Topic Modeling: Motivation and Theory

Latent Dirichlet Allocation (LDA) (I)

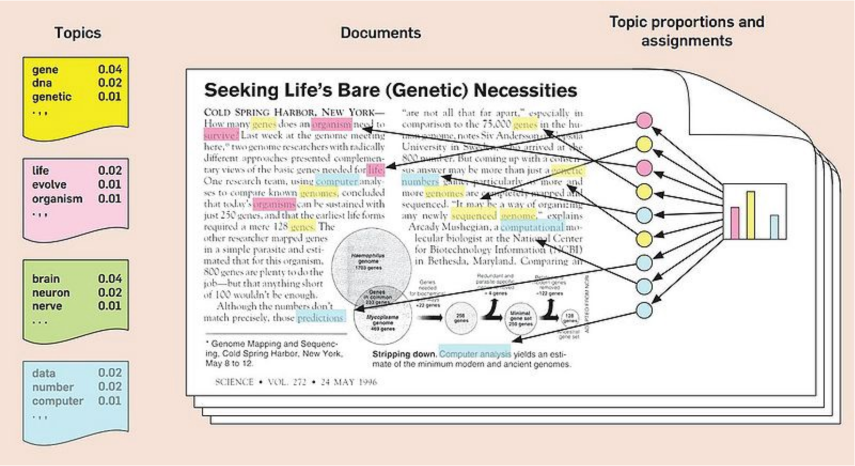
- First topic model with entirely probabilistic generating process: LDA **blei2003latent**
- Generative process for each document $d \in \{1, \ldots, D\}$:
  1. Draw topic proportions $\boldsymbol{\theta}_d \sim \text{Dir}_K(\boldsymbol{\alpha})$.
  2. For each word $n \in \{1, \ldots, N_d\}$:
     a) Draw a topic assignment $\boldsymbol{z}_{d,n} \sim \text{Multinomial}_K(\boldsymbol{\theta}_d)$.
     b) Draw a word $w_{d,n} \sim \text{Multinomial}_V(\boldsymbol{\beta}_{d,n})$.
- Graphical model representation of LDA: **blei2003latent**



Topic Modeling: Motivation and Theory

Latent Dirichlet Allocation (LDA) (II)

- Illustration of topic assignment for the words of a document: **blei2012probabilistic**



Topic Modeling: Motivation and Theory

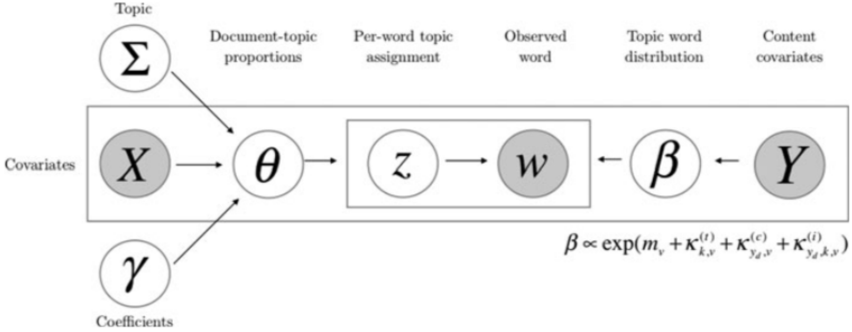
Structural Topic Model (STM)

- 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: $w_{d,n} \sim \text{Multinomial}_V(\boldsymbol{\beta}_{d,n})$. Otherwise, determine document-specific word distributions $\boldsymbol{B_a} := [\beta_1^a| \ldots |\beta_K^a]$ based on $Y_d = a$, for all topics $k \in \{1, \ldots, K\}$; select $\boldsymbol{\beta}_{d,n} := \boldsymbol{B_a z}_{d,n}$; and draw word $w_{d,n} \sim \text{Multinomial}_V(\boldsymbol{\beta}_{d,n})$.

Topic Modeling: Motivation and Theory

Graphical Model of the STM

- Visualization of the generative process again through graphical model **roberts2016model**:



Topic Modeling: Motivation and Theory

Inference and Parameter Estimation

- (Hierarchical) Bayesian model $\Rightarrow$ exact inference impossible due to marginal distributions in the denominator of posterior distribution p
- Variational inference: positing a simple distribution family q for latent variables $\boldsymbol{\theta}$ and $\boldsymbol{z}$
- Mean-field variational inference: positing full factorizability of approximating posterior q, i.e., $q(\boldsymbol{\theta}, \boldsymbol{z}) = q(\boldsymbol{\theta})q(\boldsymbol{z})$
- Then: minimizing Kullback-Leibler divergence between q and p
- STM uses a mean-field variational EM algorithm:
    - E-step: update posterior distributions of latent variables $\boldsymbol{\theta}$ and $\boldsymbol{z}$
    - M-step: update model parameters $\boldsymbol{\Gamma}$, $\boldsymbol{\Sigma}$, and - if present - topical content parameters

Data

Data

Data Collection (I)

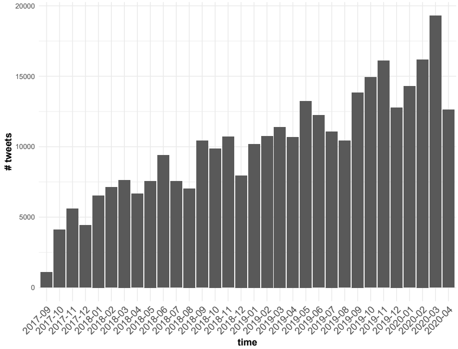
- MP-level data: from `www.bundestag.de/abgeordnete` using Python's *BeautifulSoup* and a *selenium web driver* **van1995python richardson2007beautiful**

- Twitter profiles: from official party homepages
- Socioeconomic data and 2017 German federal election results: from `www.bundeswahlleiter.de`

Data

Data Collection (II)

- Tweets (and further Twitter features): via the official Twitter API using Python's *tweepy* library**roesslein2020tweepy**
- Monthly tweets (after dropping MPs without electoral district) for our period of analysis, September 24, 2017 through April 24, 2020:



- In the following: grouping each MP's tweets on a monthly basis

Data

Data Preprocessing

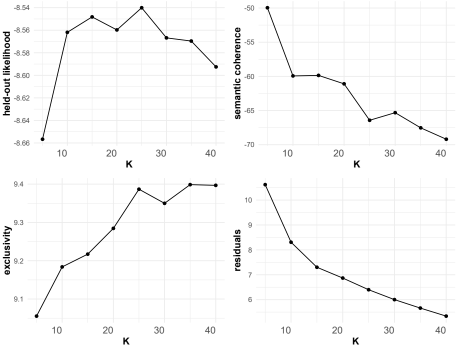
- Preprocessing: in R **R**, using the *quanteda* package **quanteda**
- Transcription of German umlauts (e.g. ä $\rightarrow$ a) and ligature (ß$\rightarrow$ ss)
- Removal of hyphens: relevant for compound words (e.g., *Corona-Krise* vs *Coronakrise*)
- Transformation of text data into document-feature matrix (DFM); conversion to lowercase; removal of stopwords, units (*kg*, *uhr*), interjections (*aaahhh*, *ufff*), etc.
- Word stemming, i.e., cutting off word endings (e.g., *politisch* $\rightarrow$ *polit*) **lucas2015computer**

Model Selection and Global Characteristics

Model Selection and Global Characteristics

Model Selection

- Model evaluation metrics for hyperparameter K (number of topics):



- "Best" trade-off: $K = 15$

Model Selection and Global Characteristics

Labeling (I)

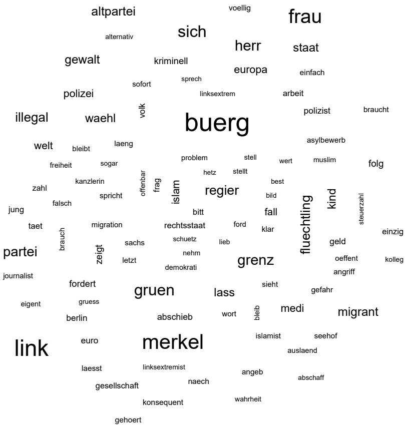
- Three-step procedure for labeling
- First step: top words for different weighting methodologies

Topic 1 Top Words: **Highest Prob:** buerg, link, merkel, frau, sich **FREX:** altpartei, islam, linksextremist, asylbewerb, linksextrem **Lift:** eitan, 22jaehrig, abdelsamad, abgehalftert, afdforder **Score:** altpartei, linksextremist, frauenkongress, islamist, boehring
Topic 3 Top Words: **Highest Prob:** brauch, wichtig, leid, dank, klar **FREX:** emissionshandel, soli, marktwirtschaft, feedback, co2steu **Lift:** aequivalenz, altersvorsorgeprodukt, bildungsqualitaet, co2limit, co2meng **Score:** emissionshandel, co2limit, basisrent, euet, technologieoff
Topic 4 Top Words: **Highest Prob:** sozial, miet, kind, arbeit, brauch **FREX:** mindestlohn, miet, wohnungsbau, mieterinn, loehn **Lift:** auseinanderfaellt, baugipfel, bestandsmiet, billigflieg, binnennachfrag **Score:** miet, mieterinn, mietendeckel, grundsicher, bezahlbar
Topic 6 Top Words: **Highest Prob:** gruen, klimaschutz, brauch, klar, euro **FREX:** fossil, erneuerbar, kohleausstieg, verkehrsminist, verkehrsw **Lift:** abgasbetrug, abgebaggert, abschalteinricht, abschaltet, ammoniak **Score:** erneuerbar, fossil, zdebel, verkehrsminist, klimaschutz

Model Selection and Global Characteristics

Labeling (II)

- Word cloud of **Highest Prob** top words (for topic 1):



- Word size corresponding to word frequency in topic 1

Model Selection and Global Characteristics

Labeling (III)

- Second step: looking at documents (i.e., original tweets) with highest proportion of topic 1

Martin Hess
@Martin_Hess_AfD

Ehem. Verfassungsrichter bestätigt AfD-Forderung: Zurückweisung illegaler Migranten dringend geboten. Gegenwärtige Politik widerspricht dem Verstand und auch der Verfassung. Wir müssen zurück zu Recht & Ordnung, wie die #AfD seit fast 3 Jahren fordert!

Hans-Jürgen Papier hält Zurückweisung von Migranten an deutscher Grenze für ...
Im Asylstreit meldet sich nun Ex-Verfassungsrichter Papier zu Wort. Die Zurückweisung von Migranten an den Grenzen sei zwingend nötig, schreibt er in...
welt.de

9:47 AM · Jun 30, 2018 · Twitter for iPhone

Model Selection and Global Characteristics

Labeling (IV)

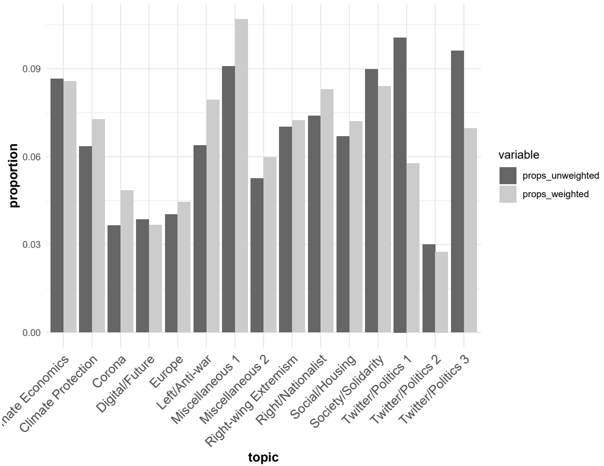
- Third step: assigning labels

Topic 1	Right/Nationalist
Topic 2	Miscellaneous 1
Topic 3	Climate Economics
Topic 4	Social/Housing
Topic 5	Digital/Future
Topic 6	Climate Protection
Topic 7	Europe
Topic 8	Corona
Topic 9	Left/Anti-war
Topic 10	Twitter/Politics 1
Topic 11	Twitter/Politics 2
Topic 12	Miscellaneous 2
Topic 13	Twitter/Politics 3
Topic 14	Right-wing Extremism
Topic 15	Society/Solidarity

Model Selection and Global Characteristics

Global Topic Proportions

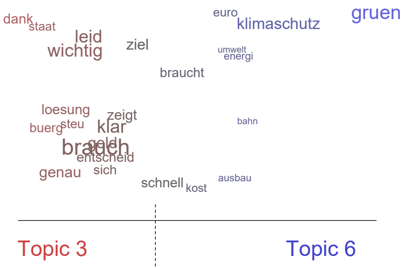
- Illustration of **global** topic proportions:

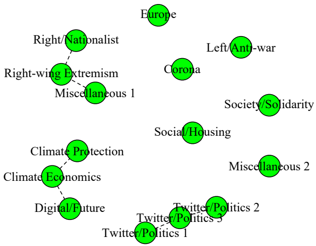


Model Selection and Global Characteristics

Global Topic Correlations

- Vocabulary overlap (left) and topic correlations (right):





Covariate-level Topic Analysis

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 $\boldsymbol{\theta}_{(k)} := (\theta_{1,k}, \dots, \theta_{D,k})^T \in [0,1]^D$ 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 problems 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

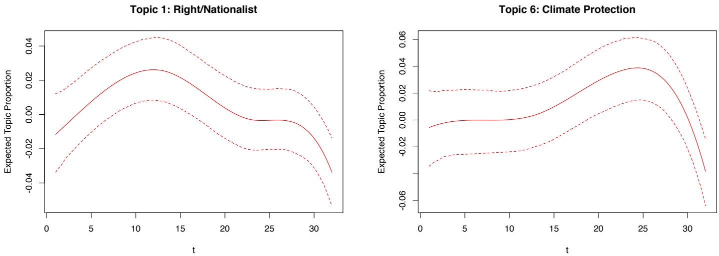
Covariate-level Topic Analysis

Problem 1: OLS Regression

Covariate-level Topic Analysis

Method of Composition: Usage within R Package *stm*

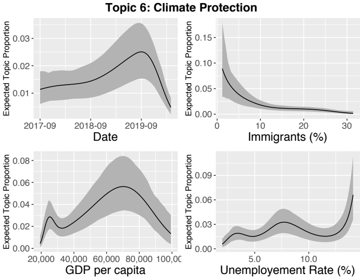
- Problem: OLS regression not suitable for (sampled) proportions, which are restricted to interval (0,1)
- Estimated relationship between proportions and prevalence covariates might even involve negative proportions:

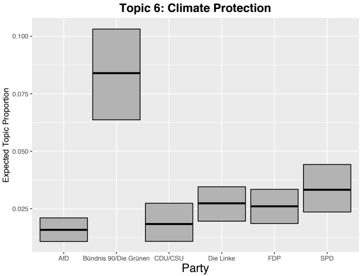


Covariate-level Topic Analysis

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





Covariate-level Topic Analysis

Problem 2: Mixing of Bayesian and Frequentist Approach

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

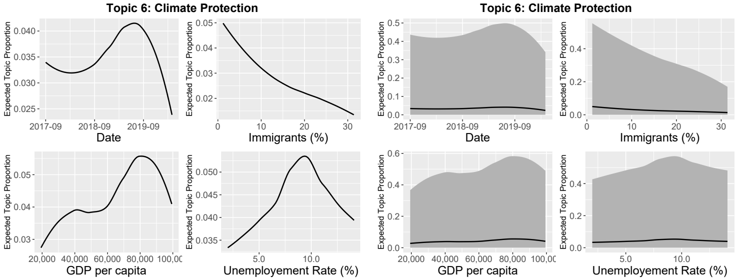
Fully Bayesian Approach: Idea

- 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



Covariate-level Topic Analysis

Problem 3: Univariate Modeling of Topic Proportions

Covariate-level Topic Analysis

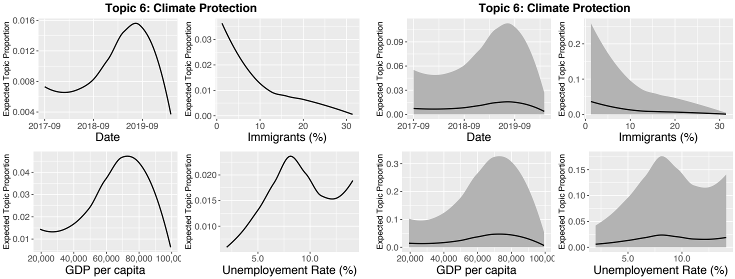
Approach to Multivariate Modeling of Proportions (I)

- 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

Approach to Multivariate Modeling of Proportions (II)

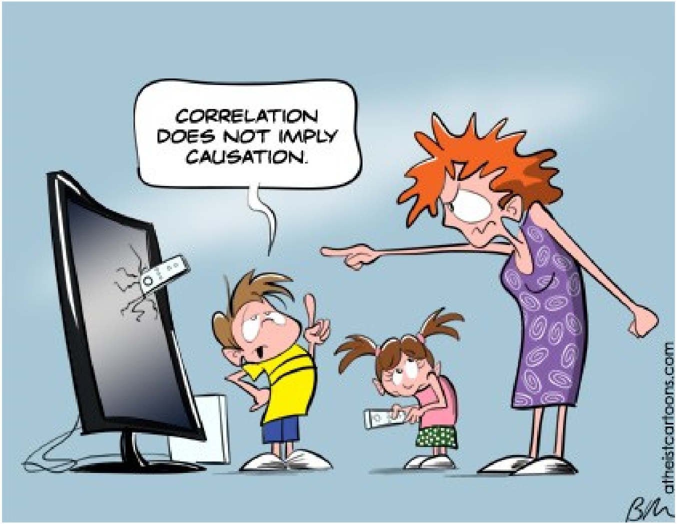
- Plugging in $\mathbf{\Gamma}$ and $\hat{\mathbf{\Sigma}}$ is "naïve" method: ideally sample prevalence parameters from their posterior $\Rightarrow$ would yield higher variation
- However, not easily possible $\Rightarrow$ should be addressed in future implementations



Causal Inference

Causal Inference

Correlation vs. Causality (I)



Causal Inference

Correlation vs. Causality (II)

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

Causal Inference

Identification Problem and Overfitting

- Setup: two groups (treatment and control), individuals otherwise similar
- Objective: quantifying treatment effect, in our case effect of treatment on prevalence of specific topic.
- Necessary assumption: response of an individual depending only on their treatment
- *Identification problem*: estimating topic model to discover latent topic proportions can introduce additional dependency among individuals $\Rightarrow$ response of each individual *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: splitting 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 to assess relation between *predicted* test set topic proportions and test set prevalence covariates
- Identification problem solved: model used for prediction 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 very unlikely to be replicated on test set

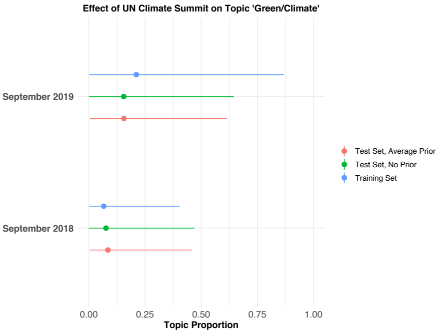
Causal Inference

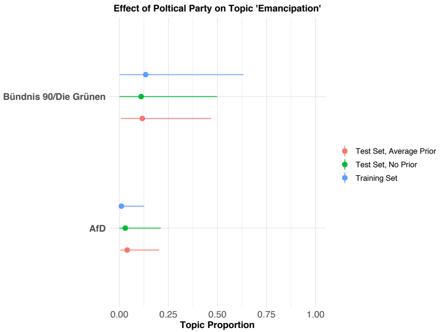
Implementation within the STM

- Inputting documents, i.e., words and metadata from the training set $\mathcal{D}_{\text{train}}$, to obtain estimates $(\hat{\beta}_{\text{train}}, \hat{\mathbf{\Gamma}}_{\text{train}}, \hat{\mathbf{\Sigma}}_{\text{train}})$ using the STM
- Then, estimating (variational) posterior of test set topic proportions, conditional on the model parameters $(\hat{\beta}_{\text{train}}, \hat{\mathbf{\Gamma}}_{\text{train}}, \hat{\mathbf{\Sigma}}_{\text{train}})$ 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 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: predicting different topic proportions for two documents from test set with exact same words if prevalence covariates differ
  - However, causal effect should be zero in such a case!

Causal Inference

Results (I)





Causal Inference

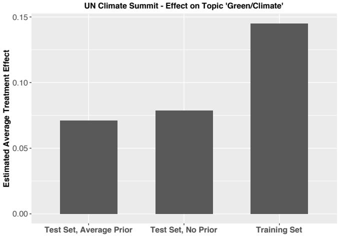
Results (II)

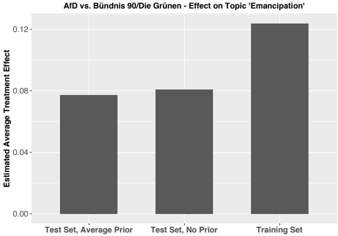
- UN Climate Action Summit 2019 held on September 23, 2019
- As observed, topic associated with climate issues much more prevalent during that time than the year before
- MAP estimates for different prior specifications on test set rather similar, yet estimated effect for training data much larger
- Similar results for effect of political party on topic labeled as 'Emancipation': average difference of estimated topic proportions between both parties larger for the training data
- Additionally: credible intervals on the training data different from those on the test data in both cases

Causal Inference

Results (III)

- Estimation of treatment effect: determining the average difference of predicted topic proportions between both groups





- Treatment effect larger if "naïvely" estimated solely on training data in both cases!

Discussion

Discussion

Summary

- Creation of broad dataset including large-scale unstructured text and variety of metadata *Rightarrow* use in future (politological) analyses
- Exemplification of topic analysis for German parliamentarians' Twitter communication
- Critical discussion of existing tools and development of new approaches regarding estimation of topic-metadata relationships
- Detailed illustration of train-test framework for causal inference within the STM

Discussion

Suggestions for Future Research

- Holistic framework for estimation of topic-metadata relationships *rightarrow* investigation of effect size and especially importance, for instance through fully Bayesian approach using MCMC
- Identification of natural experiments for causal inference
- Research into alternative model designs, beyond STM (and LDA)

Bibliography