A Bayesian approach to modeling topic-metadata relationships

Patrick Schulze[1†], Simon Wiegrebe[1,3*†], Paul W. Thurner[2], Christian Heumann[1] and Matthias Aßenmacher[1]

[1]Department of Statistics, LMU, Munich, Germany.
[2]Geschwister Scholl Institute of Political Science (GSI), LMU, Munich, Germany.
[3]Department of Genetic Epidemiology, University of Regensburg, Regensburg, Germany.

*Corresponding author(s). E-mail(s): simon.wiegrebe@stat.uni-muenchen.de;
Contributing authors: pa.schulze@campus.lmu.de; paul.thurner@gsi.uni-muenchen.de; chris@stat.uni-muenchen.de; matthias@stat.uni-muenchen.de;
†These authors contributed equally to this work.

Abstract

The objective of advanced topic modeling is not only to explore latent topical structures, but also to estimate relationships between the discovered topics and theoretically relevant metadata. Methods used to estimate such relationships must take into account that the topical structure is not directly observed, but instead being estimated itself in an unsupervised fashion, usually by common topic models. A frequently used procedure to achieve this is the *method of composition*, a Monte Carlo sampling technique performing multiple repeated linear regressions of sampled topic proportions on metadata covariates. In this paper, we propose two modifications of this approach: First, we substantially refine the existing implementation of the method of composition from the R package `stm` by replacing linear regression with the more appropriate Beta regression. Second, we provide a fundamental enhancement of the entire estimation framework by substituting the current blending of frequentist and Bayesian methods with

a fully Bayesian approach. This allows for a more appropriate quantification of uncertainty. We illustrate our improved methodology by investigating relationships between Twitter posts by German parliamentarians and different metadata covariates related to their electoral districts, using the Structural Topic Model to estimate topic proportions.

Keywords: Natural Language Processing, Topic Modeling, Topic-Metadata Relationships, Bayesian Statistics, Beta Regression, Twitter data

1 Introduction

The rise of social media has led to an unprecedented increase in the supply of publicly available unstructured text data. Researchers often wish to examine relationships between observable metadata (e.g., characteristics of a document's author) and in-text patterns (Farrell, 2016; Kim, 2017). Probabilistic topic models identify such in-text patterns by producing a posterior distribution over different topics. Yet estimating relationships with observed metadata is not trivial as the target variable is latent and itself being estimated from the text data. In this work we focus on exploring and estimating relationships between metadata and topics learned by the Structural Topic Model (STM; Roberts, Stewart, & Airoldi, 2016). We selected this model due to its high relevance in the social sciences - see Appendix A.[1] The R package `stm` (Roberts, Stewart, & Tingley, 2019) implements the STM itself and additionally provides a framework for estimating topic-metadata relationships via the *method of composition*, a combination of Monte Carlo sampling and frequentist linear regression. Even though this estimation technique is prone to producing predictions incompatible with standard definitions of probability, it is frequently applied in the literature (cf. Appendix A). This leads to implausibilities of two different forms: authors sometimes report negative expected topic proportions (e.g. Farrell, 2016; Moschella & Pinto, 2019, see also our Fig. 1); whereas in other cases "only" the confidence bands partly include negative values (e.g. Bohr & Dunlap, 2018; Chandelier, Steuckardt, Mathevet, Diwersy, & Gimenez, 2018; Cho et al., 2017; Heberling, Prather, & Tonsor, 2019). In both cases, it is ignored that sampled topic proportions are confined to $(0, 1)$ by definition, which severely harms the interpretability of results.

In this paper, we suggest two key modifications to the `stm` implementation in R (Roberts et al., 2019): First, our proposed Beta regression approach is a natural correction of the linear regression approach, accounting for topic proportions being restricted to the interval $(0, 1)$. Second, we develop a *Bayesian*

[1] However, it is crucial to understand that the choice of the topic model is only relevant for the estimation of topic proportions and does not affect the methodology for subsequent estimation of topic-metadata relationships. Therefore, the contributions presented in this work are equally valid and applicable when other topic models - such as the Latent Dirichlet Allocation (LDA; Blei, Ng, & Jordan, 2003) or the Correlated Topic Model (CTM; Blei & Lafferty, 2007) - are used for the initial estimation of topic proportions.

design within the *method of composition* to allow for a more coherent estimation and interpretation of topic-metadata relationships; in particular, we obtain a posterior predictive distribution of topic proportions at different values of metadata covariates.

We demonstrate the added value of our corrections by analyzing Twitter posts of German politicians, gathered from September 2017 through April 2020. Politics has been particularly impacted by the increasing usage of social media as evidenced by the Brexit vote and US presidential elections, with Twitter being extensively used for direct communication by politicians. We investigate relationships between latent topics in the tweets of German members of parliament (MPs) and corresponding metadata, such as tweet date or unemployment rate in the respective MP's electoral district. In doing so, we attempt to link the topics discussed to specific events as well as to socioeconomic characteristics of the MP's electoral districts.

2 Background

Topic models seek to discover latent thematic clusters, called topics, within a collection of discrete data, usually text documents. In addition to identifying such clusters, topic models estimate the proportions of the discovered topics within each document. Many topic models build upon the well-known LDA, which is a generative probabilistic three-level hierarchical Bayesian mixture model that assumes a Dirichlet distribution for topic proportions. The Correlated Topic Model (CTM; Blei & Lafferty, 2007), for instance, builds on the LDA but replaces the Dirichlet distribution with a logistic normal distribution in order to capture inter-topic correlations. The STM adopts this approach, but additionally incorporates document-level metadata into the estimation of topics:[2]

- For each document, indexed by $d \in \{1, \dots, D\}$, and each topic, indexed by $k \in \{1, \dots, K\}$, a topic proportion $\theta_{d,k}$ is drawn from a logistic normal distribution.[3]
- The parameters of the logistic normal distribution depend on document-level metadata covariates $\mathbf{x}_d$.

For parameter estimation, the STM employs a variational expectation maximization (EM) algorithm, where in the E-step the variational posteriors are updated using a Laplace approximation (Roberts et al., 2016; Wang & Blei, 2013). In the M-step, the approximated Kullback-Leibler (KL) divergence is minimized with respect to the model parameters.

[2] Within the STM, document-level covariates can also be used to fine-tune topic-word distributions (Roberts et al., 2016), but we do not further discuss this here.

[3] The `stm` package provides several metrics to choose the hyperparameter K, as will be discussed in Section 5.2.

3 Modeling Topic-Metadata Relationships in the STM

The STM produces an approximate posterior distribution of topic proportions. A point estimate can be obtained for example as the mode of this distribution. Topic proportions are often used in subsequent analysis, e.g., for determining their relationship with metadata. We argue that the usual practice of simply regressing point estimates of topic proportions on document-level covariates is not adequate for estimating topic-metadata relationships. This approach ignores that topic proportions are themselves estimates, neglecting much of the information contained in their posterior distribution. In this section, we propose a method to adequately explore the relationship between topic proportions and metadata covariates.

One way to account for the uncertainty in topic proportions is the "method of composition" (p.52; Tanner, 2012), which is a simple Monte Carlo sampling technique. Let y be a random variable with unknown distribution $p(y)$ from which we would like to sample and let z be another random variable with known distribution $p(z)$. If $p(y|z)$ is known, we can sample from

$$p(y) = \int p(y|z)p(z)dz, \tag{1}$$

using the following procedure:

1. Draw $z^* \sim p(z)$.
2. Draw $y^* \sim p(y|z^*)$.

Discarding z^*, the resulting y^* are samples from $p(y)$.[4]

In Roberts et al. (2016), the authors employ a variant of the method of composition established by Treier and Jackman (2008), which uses linear regression to obtain the conditional distribution $p(y|z)$. To demonstrate this variant, let $\boldsymbol{\theta}_{\bullet k} = (\theta_{1,k}, \ldots, \theta_{D,k})^T \in (0,1)^D$ denote the proportions of topic k and let $\mathbf{X} := [\mathbf{x}_1|\ldots|\mathbf{x}_D]^T$ be the covariates for all D documents. Let further $q(\boldsymbol{\theta}_{\bullet k})$ be the approximate posterior distribution of topic proportions given observed documents and metadata, as produced by the STM. The idea now is to repeatedly draw samples $\boldsymbol{\theta}^*_{\bullet k}$ from $q(\boldsymbol{\theta}_{\bullet k})$ and subsequently perform a regression of each sample $\boldsymbol{\theta}^*_{\bullet k}$ on covariates $\mathbf{X}$ to obtain coefficient estimates $\hat{\boldsymbol{\xi}}$. Treier and Jackman (2008) consider the asymptotic distribution of $\hat{\boldsymbol{\xi}}$ as posterior density for $\boldsymbol{\xi}$, i.e., as $p(\boldsymbol{\xi}|\boldsymbol{\theta}^*_{\bullet k}, \mathbf{X})$.

That is, the method of composition draws samples from the asymptotic distribution of the Maximum Likelihood Estimate (MLE) for the regression parameters. This use of the asymptotic distribution of the MLE can be motivated by the idea that the prior distribution is dominated by the likelihood for larger samples. Therefore, the posterior can be shown to be approximately

[4] Note that this method is an exact sampling method.

normal with mean vector equal to the MLE and variance equal to the inverse observed information matrix (see, e.g., Walker, 1969).

Using samples $\boldsymbol{\xi}^*$ from this distribution $p(\boldsymbol{\xi}|\boldsymbol{\theta}^*_{\bullet k}, \mathbf{X})$, we can "predict" topic proportions $\theta^*_{pred,k} = g(\mathbf{x}^T_{pred}\boldsymbol{\xi}^*)$ at new covariate values $\mathbf{x}_{pred}$ (g is the regression response function, e.g., identity function for linear regression). Algorithm 1 summarizes the method. Note that sampling from the posterior of topic proportions in the first step of Algorithm 1 accounts for the uncertainty in $\boldsymbol{\theta}_{\bullet k}$, while the uncertainty of the regression estimation itself is addressed by sampling from the (asymptotic) distribution of the regression coefficient estimator.

Algorithm 1: Method of composition with frequentist regression

1 **repeat procedure** m **times**:
Draw $\boldsymbol{\theta}^*_{\bullet k} \sim q(\boldsymbol{\theta}_{\bullet k})$, where q is the approximate posterior of $\boldsymbol{\theta}_{\bullet k}$.
Regress $\boldsymbol{\theta}^*_{\bullet k}$ on $\mathbf{X}$; store estimated regression coefficients $\hat{\boldsymbol{\xi}}$ and corresponding covariance matrix.
Draw $\boldsymbol{\xi}^*$ from the (asymptotic) distribution of $\hat{\boldsymbol{\xi}}$.
Predict topic proportions $\theta^*_{pred,k} = g(\mathbf{x}^T_{pred}\boldsymbol{\xi}^*)$ at new covariate values $\mathbf{x}_{pred}$.
end procedure

To visualize topic-metadata relationships, Roberts et al. (2016) generate multiple "predictions" $\theta^*_{pred,k}$ and calculate empirical quantities such as the mean and quantiles. Calculating mean and credible intervals in such a Bayesian fashion implicitly assumes a (posterior predictive) distribution for $\theta^*_{pred,k}$. This distribution, however, directly depends on the regression - which is frequentist as implemented in the `stm` package. We address this point in detail in Section 4.2.

4 Methodological Improvements

While we agree with performing Monte Carlo sampling of topic proportions in order to integrate over latent variables, we aim to address two inconsistencies:

1. **Inadequate modeling of proportions**: The method of composition is implemented in the R package `stm` via the `estimateEffect` function, which employs a linear regression in the second step of Algorithm 1 (implying $g = id$ in the last step). This implementation ignores that topic proportions are naturally restricted to the interval $(0, 1)$. As a consequence, when using the `estimateEffect` function, we frequently observed predicted topic proportions outside of $(0, 1)$, as is exemplarily shown for one specific topic-covariate combination in Figure 1.
2. **Mixing Bayesian and frequentist methods**: The method of composition used by Treier and Jackman (2008) and Roberts et al. (2016) mixes Bayesian and frequentist methods. As described in Section 3, a frequentist regression is used inside the method of composition, yet estimates

are obtained in a Bayesian manner via calculation of empirical mean and quantiles. Recall that according to Treier and Jackman (2008), $\boldsymbol{\xi}^*$ can be considered a sample from the posterior of regression coefficients. However, the coefficients resulting from a frequentist regression do not have any distribution because the frequentist framework assumes them to be fixed parameters. As a consequence, one cannot sample from the distribution of regression coefficients, which is why Treier and Jackman (2008) sample $\boldsymbol{\xi}^*$ from the distribution of coefficient *estimators*. This distribution, however, only exists by making frequentist assumptions.

In Sections 4.1 and 4.2 below we further discuss these problems and present corrections and alternatives, all of which are implemented in the R package `stmprevalence`. [5]

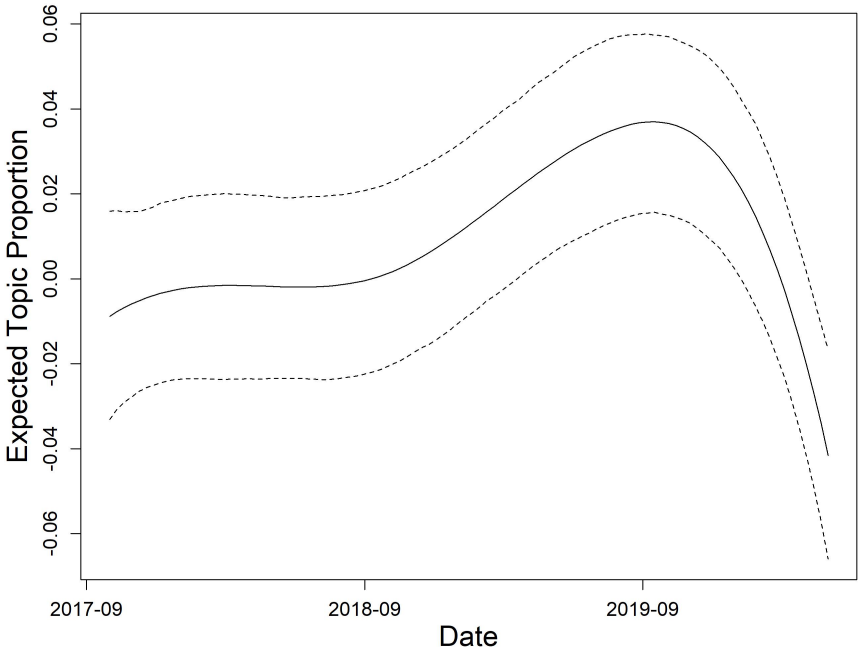

Fig. 1: Mean prediction and 95% confidence intervals for the topic proportion of topic "Climate Protection" over time, generated using `estimateEffect` from the R package `stm`.

[5] Available at https://github.com/PMSchulze/stmprevalence.

4.1 Frequentist Beta Regression

As noted above, the linear regression approach is often used carelessly in the literature, neglecting that topic proportions are non-negative by definition. Farrell (2016) and Moschella and Pinto (2019), for instance, produce figures containing negative expected topic proportions, while Cho et al. (2017), Chandelier et al. (2018), Bohr and Dunlap (2018), and Heberling et al. (2019) display confidence bands partly covering negative values.

Therefore, we correct the approach employed within the `stm` package by replacing the linear regression with a regression model that assumes a dependent variable in the interval $(0, 1)$. As shown by Atchison and Shen (1980), the Dirichlet distribution is well suited to approximate a logistic normal distribution, though inducing less interdependence among the different topics. When employing a Dirichlet distribution, the univariate marginal distributions are Beta distributions. We thus perform a separate Beta regression for each topic proportion on $\mathbf{X}$, using a logit-link.[6] This approach now again corresponds to Algorithm 1, but with g being the logistic sigmoid function in this case.[7]

4.2 Bayesian Beta Regression

Treier and Jackman (2008) and the authors of the STM consider $\boldsymbol{\xi}^*$ to be samples from the posterior of regression coefficients. While it is possible to view frequentist regression from a Bayesian perspective, it implies assuming a uniform prior distribution for regression coefficients $\boldsymbol{\xi}$ - which is rather implausible. More generally, the mixing of Bayesian and frequentist frameworks within the method of composition lacks a theoretical foundation, especially when employing an *asymptotic* distribution of regression coefficient estimators. This applies to the model of Treier and Jackman (2008) as well as to the Beta regression presented in Section 4.1. Furthermore, note that when using a frequentist regression, the estimated uncertainty is with respect to the prediction of the mean of topic proportions. However, when exploring topic-metadata relationships it might be preferable to examine the variation of individual topic proportions among documents at different values of metadata covariates.

Algorithm 2: Method of composition with Bayesian Beta regression

1 **repeat procedure** m **times**:
Draw $\boldsymbol{\theta}^*_{\bullet k} \sim q(\boldsymbol{\theta}_{\bullet k})$, where q is the approximate posterior of $\boldsymbol{\theta}_{\bullet k}$.
Perform a Bayesian Beta regression of $\boldsymbol{\theta}^*_{\bullet k}$ on $\mathbf{X}$ using normal priors centered around zero.
Draw $\theta^*_{pred,k} \sim p(\boldsymbol{\theta}_{pred,k}|\boldsymbol{\theta}^*_{\bullet k}, \mathbf{X}, \mathbf{x}_{pred})$, i.e., conditional on sample $\boldsymbol{\theta}^*_{\bullet k}$.
end procedure

[6] Note that the distribution of regression coefficient estimators is asymptotically normal for Beta regression (p.17; Ferrari & Cribari-Neto, 2004).

[7] While runtime for estimating Beta regressions is considerably longer in relative terms, it is still short in absolute terms, which is why runtime concerns can be disregarded for the practical use of our approach.

Therefore, we propose to replace the frequentist regression in Algorithm 1 by a Bayesian Beta regression with normal priors centered around zero. This enables modeling topic-metadata relationships in a fully Bayesian manner while preserving the methodological improvements from Section 4.1. Algorithm 2 summarizes this approach. By drawing $\theta^*_{pred,k}$ at covariate values $\mathbf{x}_{pred}$, we obtain samples from the posterior predictive distribution

$$p(\boldsymbol{\theta}_{pred,k}|\boldsymbol{\theta}^*_{\bullet k}, \mathbf{X}, \mathbf{x}_{pred}) = \tag{2}$$

$$\int p(\boldsymbol{\theta}_{pred,k}|\mathbf{x}_{pred}, \boldsymbol{\xi}) p(\boldsymbol{\xi}|\boldsymbol{\theta}^*_{\bullet k}, \mathbf{X}) d\boldsymbol{\xi}, \tag{3}$$

where $p(\boldsymbol{\xi}|\boldsymbol{\theta}^*_{\bullet k}, \mathbf{X})$ denotes the posterior distribution of regression coefficients. This allows displaying the (predicted) variation of topic proportions at different covariate levels. As before, quantities of interest, such as the mean and quantiles, are obtained by averaging across samples; now, however, these samples are generated within a fully Bayesian framework.

5 Application[8]

In this section, we first apply the STM to German parliamentarians' Twitter data and subsequently demonstrate both the original (`stm`) and our new method (`stmprevalence`) to explore topic-metadata relationships. Here, we chose to apply the STM in particular for illustrative purposes, because of its flexibility and its relevance in the social sciences. We would like to emphasize again, however, that our methods work with any other topic model, such as LDA or CTM, as long as it produces an (approximate) posterior distribution of topic proportions. This is because our methods focus on the step subsequent to the estimation of a topic model, i.e., on the exploration of relationships between previously estimated topic proportions and metadata covariates.

5.1 Data[9]

For all German MPs during the 19th election period (starting on September 24, 2017), we gathered personal information such as name, party affiliation, and electoral district from the official parliament website as well as Twitter profiles from the official party websites, using *BeautifulSoup* (Richardson, 2007). Next, after excluding MPs without a public Twitter profile, we used *tweepy* (Roesslein, 2020) to scrape all tweets by German MPs from September 24, 2017 through April 24, 2020. We also gathered socioeconomic data, such as GDP per capita and unemployment rate, as well as 2017 election results on an electoral-district level. Text preprocessing, such as transcription of German umlauts, removal of stopwords, and word-stemming, was performed with *quanteda* (Benoit et al., 2018).[10]

[8] Source code available at https://github.com/PMSchulze/topic-metadata-stm.
[9] Raw data: https://figshare.com/s/7a728fcb6d67a67fc3d6.
[10] An in-depth discussion of topic model preprocessing and its application to Twitter data can be found in Lucas et al. (2015).

We define a document as the concatenation of an individual MP's tweets during a single calendar month in order to achieve sufficient document length. Our final data set includes 10,998 monthly MP-level documents, each one associated with 90 covariates.

5.2 Model Fitting and Global-level Analysis

Before fitting the STM, we need to decide on the number of topics, K. To do so, we use the following four model evaluation metrics: *held-out likelihood*, *semantic coherence*, *exclusivity*, and *residuals*. The held-out likelihood approach is based on document completion. The higher the held-out likelihood, the more predictive power the model has on average (Wallach, Murray, Salakhutdinov, & Mimno, 2009). Semantic coherence means that words characterizing a specific topic also appear together in the same documents (Mimno, Wallach, Talley, Leenders, & McCallum, 2011). Exclusivity, on the other hand, indicates to which degree words characterizing a given topic *only* occur in that topic. Finally, the residuals metric, which is based on residual dispersion, indicates a (potentially) insufficiently small value of K whenever the residual dispersion is larger than one (Taddy, 2012).

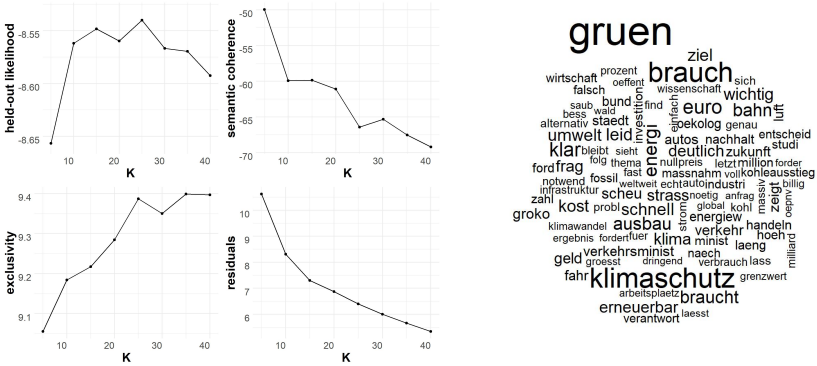

Fig. 2: Left: Model evaluation metrics for hyperparameter K (number of topics). Right: Word cloud for the topic labeled as "Climate Protection".

The left part of Figure 2 shows these four metrics for a grid of K between five and 40 with step size five. Both $K = 15$ and $K = 20$ seem to be good choices. Given the better interpretability for models with fewer topics, we choose $K = 15$.

After fitting the model, we label all topics manually with human interpretable labels; to do so, we use word clouds and top words (see Figure 2 (right panel) and Appendix B). Throughout this work, we consider the topics "Climate Protection", "Right/Nationalist", "Social/Housing", and "Europe" for illustration, in particular the first one. To obtain an overview of the model

output, different global-level analyses are conducted, such as inspecting global topic proportions $\bar{\theta}_k = \frac{1}{D}\sum_{d=1}^{D}\theta_{d,k}$ or creating a network graph.

5.3 Topic-Metadata Relationships

Moving from global- to document-level, we now visualize relationships between document-level topic proportions $\theta_{d,k}$ and covariates $\mathbf{x}_d$. In particular, we examine the extent to which German MPs discussed the abovementioned topics over time and in relation to several socioeconomic variables regarding their respective electoral districts. These relationships were estimated by regressing the previously estimated topic proportions on metadata covariates, using either the linear regression-based method of composition (see Fig. 1) or our Beta regression-based methods (see Fig. 3 and 4).[11]

For all regressions, we choose the same linear predictor, containing the date of the Twitter posts, the MP-level categorical covariates *political party affiliation* and *federal state*, as well as the electoral district-level continuous socioeconomic covariates *immigration share*, *GDP per capita*, and *unemployment rate*; the effects of the latter three, due to being continuous, are estimated as smooth functions using B-splines.

To demonstrate the shortcomings of the approach implemented in the `stm` package, we first apply the `estimateEffect` function to produce "naïve" estimates for the relationship between estimated topic proportions and document-level covariates. Figure 1 shows the estimated proportion of the topic "Climate Protection" over time, peaking during the UN Climate Action Summit 2019 held in September 2019. Importantly, notice that `estimateEffect` produces predicted topic proportions outside of $(0, 1)$. This is due to using a linear regression, which places no restrictions on the range of the dependent variable.

Next, we evaluate the results when replacing the linear regression by a Beta regression, which restricts the dependent variable to the $(0, 1)$-interval.

Figure 3 consists of four panels, one for each topic, each panel being made up of four (sub)plots. The top left plot in the top left panel corresponds to the time trend of the climate protection topic. It shows that the overall trend over time is similar to the one in Figure 1, yet the range is shifted upwards and no negative values are estimated. The three remaining plots of the top left panel depict the relationship of the climate protection topic with the socioeconomic covariates immigration, GDP per capita, and unemployment as measured at the electoral district-level. First, note that only non-negative values are obtained - as desired. Regarding GDP per capita, we notice an increase in the relevance of the climate protection topic until around EUR 70k, yet for very high income electoral districts this trend is reversed. The unemployment rate shows an ambiguous relationship, with rather large fluctuations. Finally, the higher the share of immigrants in an electoral district, the less frequently the district's MPs tend to discuss climate-related subjects on average.

[11] Again, note that the topic proportions could alternatively have been estimated via, e.g., LDA or CTM. Our methods concern the subsequent step, i.e., estimating topic-metadata relationships, and are unrelated to the topic model choice.

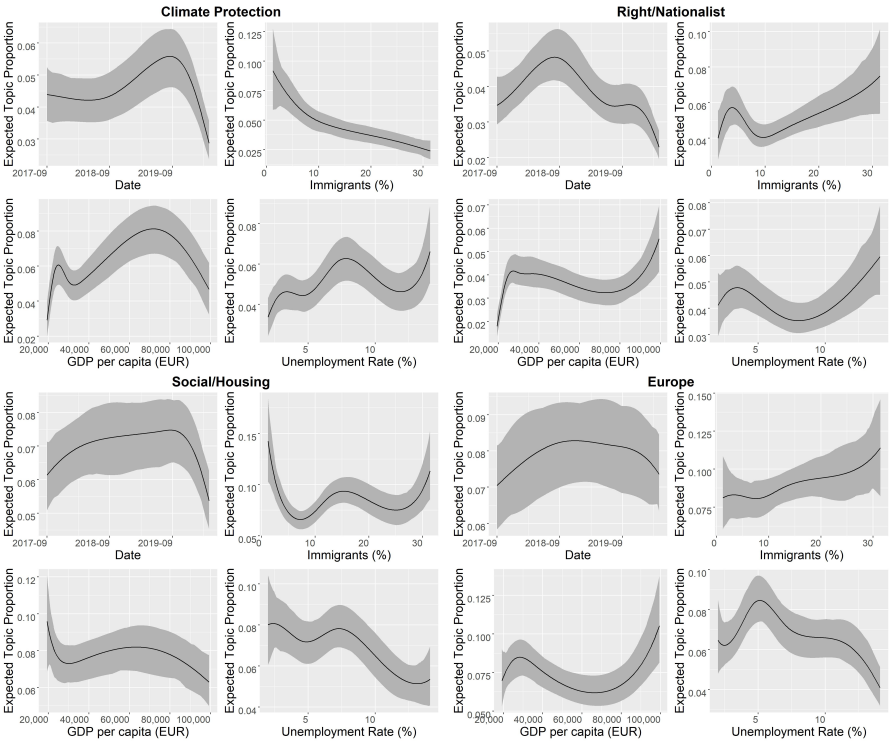

Fig. 3: Mean prediction and 95% confidence intervals for the topic proportion of topics "Climate Protection", "Right/Nationalist", "Social/Housing", and "Europe" for different document-level covariates, obtained using a frequentist Beta regression from the R package `stmprevalence`.

However, one might suspect that this negative relationship between climate protection relevance and immigration is the consequence of spurious correlation: one immigration-related topic might simply be suppressing all other topics.[12] To investigate this, and also in order to evaluate our approach more broadly, we consider three further topics, "Right/Nationalist", "Social/-Housing", and "Europe". Actually, the frequency of the "Right/Nationalist" topic increases as electoral district-level immigrant share increases, yet a similar association can also be found for the Europe-related topic; for the topic regarding social issues and housing, no clear trend is recognizable. This leads us to conclude that the negative association between the relevance of the climate protection topic and the immigration share is not only an effect of the mechanics of compositional data such as topic proportions.

Regarding time, the social and European topics do not show any temporal trend, whereas the nationalist topic clearly peaks around September 2018.

[12] Recall that topic proportions must sum to 1, so an increase in the proportion of one topic mechanically decreases the relevance of all other topics.

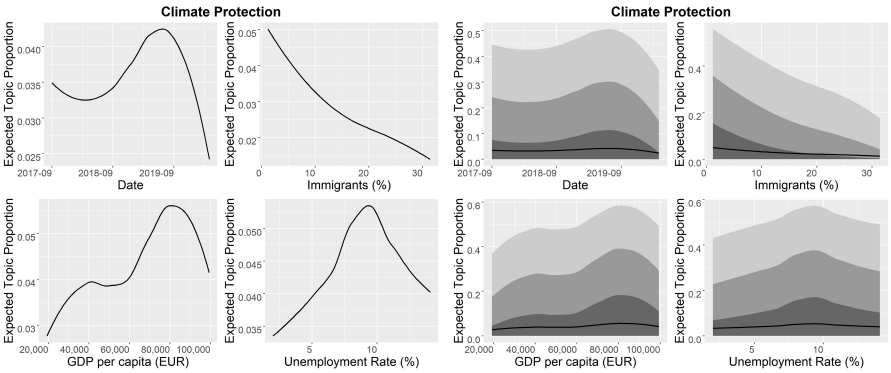

Fig. 4: Left: Mean prediction for the topic proportion of topic "Climate Protection" for different document-level covariates, obtained using a Bayesian Beta regression from the R package `stmprevalence`. Right: 95% (light grey), 90% (grey), and 85% (dark grey) quantiles of the posterior predictive distribution for the topic proportion of topic "Climate Protection".

As for GDP per capita and unemployment rate, only few more or less clear trends can be recognized, such as the decrease in the relevance of the European as well as the social topic with increasing unemployment rate. However, while some interesting and reasonable patterns emerge, we do caution against (quantitative) over-interpretation of the observed patterns.

Finally, we display the results from the fully Bayesian approach discussed in Section 4.2, though here we only focus on the climate protection topic for the sake of brevity. As can be seen in the left plot of Figure 4, the predicted progressions of mean topic proportions at different covariate values are mostly similar to those obtained with the frequentist Beta regression, yet the range is compressed and shifted downwards. In addition to the empirical mean, the right plot of Figure 4 depicts different empirical quantiles of the posterior predictive distribution of topic proportions. Here we can see that topic proportions at different covariate values vary starkly for different MPs. More generally, we find that a fully Bayesian approach enables a much more comprehensive analysis of topic-metadata relationships because it allows for displaying the variation of individual topic proportions observed in the data.

6 Conclusion

Nowadays, large-scale unstructured text from a wide variety of fields is publicly available on social media and various other forms of online appearances. Topic modeling plays an important role in the extraction of specific information from such data. At the same time, researchers - in particular from the social sciences - increasingly move beyond purely exploratory topic analyses, wishing to associate identified topics with metadata. In order to investigate topic-metadata relationships while accounting for the probabilistic nature of topic proportions,

the R package `stm` implements repeated linear regressions of sampled topic proportions on metadata covariates using the method of composition.

In this paper, we identify two main inconsistencies of this original implementation: the inadequate modeling of proportions via linear regression, allowing topic proportions to take on values outside of $(0, 1)$; and the mixing of frequentist regression with Bayesian computations of empirical quantities. We propose improvements to both shortcomings: the more appropriate Beta regression to account for the distributional nature of topic proportions; and a fully Bayesian approach to replace the current mixture of frequentist and Bayesian methods within the method of composition.

We illustrate our proposed improvements by first applying the STM to a data set containing Twitter posts by German MPs and subsequently employing our methods to estimate relationships between estimated topic proportions and MP-level metadata covariates. It is important to note that our methods merely concern the second-step estimation of topic-metadata relationships and are thus equally applicable to other topic models and beyond.

Limitations and Outlook

There are some limitations to our approach, which in turn give rise to future research. Regarding the application case presented in this paper, the relationship with Twitter-related metadata such as retweets or likes would be interesting - especially because such metadata would be actively influenced by the topics of the tweets, whereas the socioeconomic covariates used here are of a more explanatory nature. Unfortunately, Twitter-related metadata are not contained in the data set. Another use case-related aspect is the document length. Longer documents are beneficial for topic models such as the STM in general, yet in our specific case hamper the content-related interpretability of the resulting "tweet documents". We experimented extensively with different document lengths, including days and weeks, but finally came to the conclusion that aggregating tweets at a monthly interval constitutes the best compromise between content-related interpretability and sufficient text length.

Both frequentist and Bayesian Beta regression are well established approaches in the statistical literature, necessarily implying a lower degree of methodological novelty of our approach. However, the correct modeling and illustration of topic-metadata relationships and the corresponding uncertainty is of paramount importance: because of the enormous popularity of topic models such as the STM and the fact that conclusions drawn from a misspecified model can be (substantially) misleading (cf. Appendix A).

Several possibilities exist to build upon our exploratory methods. For instance, our approach could be used in combination with MCMC-based methods in order to make inference in a Bayesian setting. If the goal is to make causal inference beyond exploratory purposes, one must take into account that the estimation of topic proportions induces additional dependence across documents. Developing methods to identify underlying causal mechanisms is the

subject of current research (e.g., Egami, Fong, Grimmer, Roberts, & Stewart, 2018).

Acknowledgments. We would like to thank the editor and the two anonymous referees for their comments, which helped improve this paper considerably. All other acknowledgments regarding funding, etc. are not shown at the moment due to the anonymization guidelines and will be added upon publication.

Declarations

There were no conflicts of interest or competing interests.

References

Atchison, J., & Shen, S.M. (1980). Logistic-normal distributions: Some properties and uses. *Biometrika*, *67*(2), 261–272.

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., Matsuo, A. (2018). quanteda: An r package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), 774. Retrieved from https://quanteda.io

10.21105/joss.00774

Blei, D.M., & Lafferty, J.D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, *1*(1), 17–35.

Blei, D.M., Ng, A.Y., Jordan, M.I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.

Bohr, J., & Dunlap, R.E. (2018). Key topics in environmental sociology, 1990–2014: results from a computational text analysis. *Environmental Sociology*, *4*(2), 181–195.

Chandelier, M., Steuckardt, A., Mathevet, R., Diwersy, S., Gimenez, O. (2018). Content analysis of newspaper coverage of wolf recolonization in france using structural topic modeling. *Biological conservation*, *220*, 254–261.

Cho, I., Wesslen, R., Karduni, A., Santhanam, S., Shaikh, S., Dou, W. (2017). The anchoring effect in decision-making with visual analytics. *2017 ieee conference on visual analytics science and technology (vast)* (pp. 116–126).

Egami, N., Fong, C.J., Grimmer, J., Roberts, M.E., Stewart, B.M. (2018). How to make causal inferences using texts. *arXiv preprint arXiv:1802.02163*.

Farrell, J. (2016). Corporate funding and ideological polarization about climate change. *Proceedings of the National Academy of Sciences*, *113*(1), 92–97.

Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of applied statistics*, *31*(7), 799–815.

Heberling, J.M., Prather, L.A., Tonsor, S.J. (2019). The changing uses of herbarium data in an era of global change: An overview using automated content analysis. *BioScience*, *69*(10), 812–822.

Kim, I.S. (2017). Political cleavages within industry: Firm-level lobbying for trade liberalization. *The American Political Science Review*, *111*(1), 1.

Lucas, C., Nielsen, R.A., Roberts, M.E., Stewart, B.M., Storer, A., Tingley, D. (2015). Computer-assisted text analysis for comparative politics. *Political Analysis*, *23*(2), 254–277.

Mimno, D., Wallach, H.M., Talley, E., Leenders, M., McCallum, A. (2011). Optimizing semantic coherence in topic models. *Proceedings of the conference on empirical methods in natural language processing* (pp. 262–272).

Moschella, M., & Pinto, L. (2019). Central banks' communication as reputation management: How the fed talks under uncertainty. *Public Administration*, *97*(3), 513–529.

Richardson, L. (2007). Beautiful soup documentation. *April*.

Roberts, M.E., Stewart, B.M., Airoldi, E.M. (2016). A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, *111*(515), 988–1003.

Roberts, M.E., Stewart, B.M., Tingley, D. (2019). stm: An R package for structural topic models. *Journal of Statistical Software*, *91*(2), 1–40.

10.18637/jss.v091.i02

Roesslein, J. (2020). Tweepy: Twitter for python! *URL: https://github.com/tweepy/tweepy*.

Taddy, M. (2012). On estimation and selection for topic models. *Artificial intelligence and statistics* (pp. 1184–1193).

Tanner, M.A. (2012). *Tools for statistical inference*. Springer.

Treier, S., & Jackman, S. (2008). Democracy as a latent variable. *American Journal of Political Science*, *52*(1), 201–217.

Walker, A.M. (1969). On the asymptotic behaviour of posterior distributions. *Journal of the Royal Statistical Society: Series B (Methodological)*, *31*(1), 80–88.

Wallach, H.M., Murray, I., Salakhutdinov, R., Mimno, D. (2009). Evaluation methods for topic models. *Proceedings of the 26th annual international conference on machine learning* (pp. 1105–1112).

Wang, C., & Blei, D.M. (2013). Variational inference in nonconjugate models. *Journal of Machine Learning Research*, *14*(Apr), 1005–1031.

A Exemplary figures with implausible predictions

To demonstrate the importance of our proposed corrections of the STM, we collected figures from a selection of research papers where using the original implementation led to implausible estimates. Due to copyright issues, however, we do not show them here but instead merely reference them, along with a short description of how the uncorrected method of composition produces implausible results in the respective cases.

- Cho et al. (2017), p.10, Fig. 10 (actually p.125): negative confidence bands for covariate effects https://ieeexplore.ieee.org/document/8585665
- Bohr and Dunlap (2018), p.9, Fig. 9: negative confidence bands *and* negative covariate effects https://doi.org/10.1080/23251042.2017.1393863
- Moschella and Pinto (2019), p.11, Fig. 2 (actually p.523): negative confidence bands *and* negative covariate effects https://doi.org/10.1111/padm.12543
- Chandelier et al. (2018), p.6, Fig. 2 (actually p.259) : negative confidence bands for covariate effects https://doi.org/10.1016/j.biocon.2018.01.029
- Heberling et al. (2019), p.8, Fig. 5 (actually p.819) : negative covariate effects, https://doi.org/10.1093/biosci/biz094

Finally, another example of confidence bands of topic proportions becoming negative when using the `estimateEffect` function is Figure 7 (p. 20) of the vignette of the `stm` package. In the README file of our `stmprevalence` package, we reproduce this figure and furthermore show how the uncertainty estimation is corrected when using our approaches.

B Word clouds and top words for selected topics

The top words for the four topics "Climate Protection", "Right/Nationalist", "Social/Housing", and "Europe", which are used for illustration in Figure 3, are shown in Table 1 below.

Topic	Word 1	Word 2	Word 3	Word 4	Word 5
Climate Protection	gruen	klimaschutz	brauch	klar	euro
Right/Nationalist	buerg	link	merkel	frau	sich
Social/Housing	sozial	miet	kind	arbeit	brauch
Europe	europaeisch	wichtig	europa	international	thank

Table 1: Top five words (in terms of absolute frequency across all text documents) within the topics "Climate Protection", "Right/Nationalist", "Social/Housing", and "Europe".

The word cloud for the "Climate Protection" topic has already been shown in Figure 2 (right panel). Figure 5 below shows the word clouds for the topics "Right/Nationalist", "Social/Housing", and "Europe", respectively.

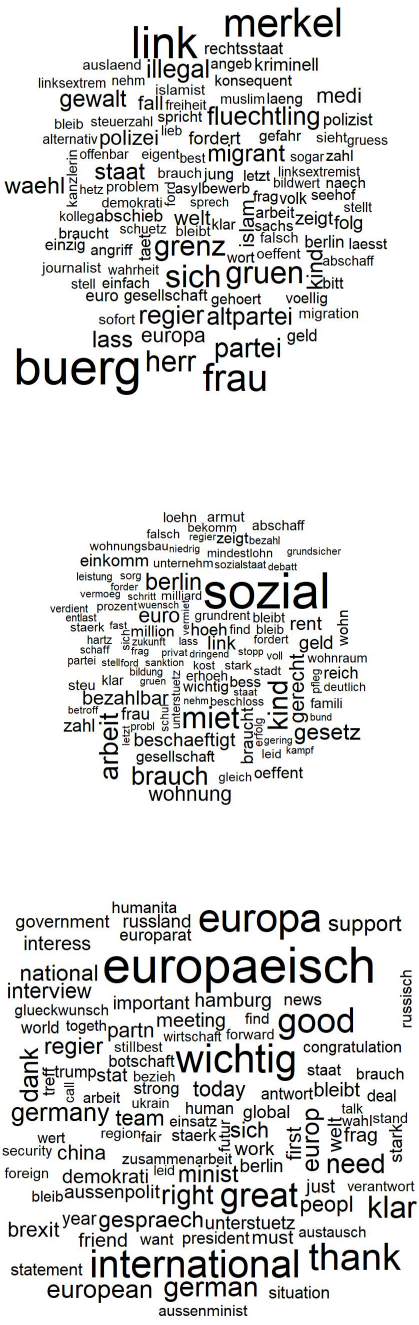

Fig. 5: Word clouds for the topics "Right/Nationalist" (top), "Social/Housing" (center), and "Europe" (bottom).