Ludwig-Maximilians-Universität München

Statistical Consulting Project

E-Complaints of Citizens and Responsiveness Strategies of Political Candidates and Parties: Investigating Large-Scale Unstructured Text

Comparison of homepages of national and local political entities

Simon Wiegrebe, Patrick Schulze

supervised by
Prof. Dr. Christian Heumann and Prof. Dr. Paul W. Thurner

June 27, 2020

Contents

1 Covariate-level Topic Analysis

We now proceed to analyze the relationship between metadata information (i.e., document-level covariates) and topic proportions. We specify topical prevalence as

$$\mu_{d,k} = x_d^T \gamma_k = \text{party}_{d,k} + \text{state}_{d,k} + f_k(\text{t}_d) + g_k(\text{struct}_d), \tag{1.1}$$

for all documents $d = 1, \dots, D$, and for all topics $k = 1, \dots, K$, where

$$g_k(\text{struct}_d) = g_k^{(1)}(\text{GDP}_d) + g_k^{(2)}(\text{unemployment}_d) + g_k^{(3)}(\text{immigrants}_d) + g_k^{(4)}(\text{votes}_d).$$

That is, the political party and federal state of the respective parliamentarian associated with a document are specified as simple categorial dummy effects, while date and electoral-district structural covariates (GDP per capita, unemployment rate, percentage of immigrants, and the 2017 vote share) are modeled as additive smooth functions.

Note that approximate inference implies replacing $\mu_{d,k}$ with $\lambda_{d,k}$, i.e., with the mean of the approximate Gaussian posterior $q(\eta_{d,k})$. The estimates of $\Gamma = [\gamma_1 | \dots | \gamma_K]$ are updated in a Bayesian linear regression during each iteration of the EM algorithm in the M-step; for details see Roberts et al. (2013), p. 993.

While topical prevalence has an effect on the estimated topic proportions, the exact specification of topical prevalence is not a decisive factor. Both estimated topic proportions as well as heldout likelihood are in general only marginally affected by the concrete choice of the functional form. However, completely removing topical prevalence, in which case the model reduces to a CTM, does result in different topic proportions, as we show in section XXX. Since evaluation metrics such as heldout likelihood are mostly unaffected by the exact choice of topical prevalence and because the computational cost of fitting an stm is rather high, automatic model selection methods w.r.t. topical prevalence are not available. A reasonable specification of topical prevalence therefore relies on the domain knowledge of the researcher.

There exist different approaches to study the relationship between topic proportions and prevalence covariates. One possibility is to directly assess the estimates $\hat{\Gamma}$ and $\hat{\Sigma}$, which are generated by the stm. Since the document-level topic proportions θ_d follow a logistic normal distribution (with mean μ_d and covariance matrix Σ), interpretation of the results can be difficult, since the logistic normal distribution is not very accessible. Nonetheless, we can visualize the relationship between a topic and a prevalence covariate, fixing other covariates at their median (for categorial variables the majority vote is used).

Alternatively, the estimated topic proportions can be used as the dependent variable of a new regression on prevalence covariates. However, in contrast to a standard regression setting, in this case the dependent variable has been estimated itself, before the regression is performed. Instead of simply using the maximum-a-posteriori (MAP) estimates of θ_d as the dependent variable, having access to the posterior distribution of the topic proportions, we can take account for the uncertainty of the dependent variable. This can be achieved by employing a sampling procedure known as the method of composition in the social sciences; see Tanner (2012), p.52. This procedure is implemented in the *stm* package through its function *estimateEffect*.

In the following, we will first introduce the method of composition. We will discuss its implementation in the *stm* package and provide alternative regression approaches based on the method of composition. Subsequently, we will evaluate the relationship between prevalence covariates and topic proportions by directly assessing the estimates $\hat{\Gamma}$ and $\hat{\Sigma}$, as outlined above, and compare the results of both approaches.

1.1 Method of Composition

Let $\theta_{(k)} := (\theta_{1,k}, \ldots, \theta_{D,k})^T \in [0,1]^D$ denote the proportions of the k-th topic for all D documents. As stated, we want to perform a regression of these topic proportions $\theta_{(k)}$ on a subset $\tilde{X} \in \mathbb{R}^{D \times \tilde{P}}$ of prevalence covariates X. The true topic proportions are unknown, but the stm produces an estimate of the approximate posterior of $\theta_{(k)}$. A naïve approach would be to regress the estimated mode of the approximate posterior distribution on $\tilde{X}$. However, this approach neglects much of the information contained in the distribution.

Instead, repeatedly sampling $\theta^*_{(k)}$ from the approximate posterior distribution, performing a regression for each sampled $\theta^*_{(k)}$ on $\tilde{X}$, and then sampling from the estimated distribution of icients, provides an i.i.d. sample from the marginal posterior distribution of regression coefficients.

Sampling $\theta^*_{(k)}$ is achieved by first sampling the unnormalized topic proportions η^* from the approximate posterior $q(\eta)$, applying the softmax $\theta^* = \text{softmax}(\eta^*)$ (element-wise, i.e., for each of the K-dimensional vectors of topic proportions), and lastly selecting the k-th column of θ^*. Precisely, $q(\eta) = \prod_d q(\eta_d)$ is a normal distribution, which emerges from the laplace approximation within the variational inference scheme; for details see Roberts et al. (2016), pp. 992-993. For clarity, we denote the approximate posterior of topic proportions as $q(\theta_{(k)}|X, W)$, in order to emphasize that the parameters of this distribution are learned from the observed data, i.e. prevalence covariates and words (note that we have no content variables included). Furthermore, let ξ denote the regression coefficients from a regression of $\theta_{(k)}$ on $\tilde{X}$, and let $q(\xi|\tilde{X}, \theta_{(k)})$ be the approximate posterior distribution of these coefficients, i.e. given design matrix $\tilde{X}$ and response $\theta_{(k)}$.

The method of composition can now be described by repeating the following process m times:

1. Draw $\theta^*_{(k)} \sim q(\theta_{(k)}|X, W)$.

2. Draw $\xi^* \sim q(\xi|\tilde{X}, \theta_{(k)})$.

It then holds that $\xi_1^*, \dots, \xi_m^*$ is an i.i.d. sample from the marginal posterior

$$q(\xi|X, W) := \int_{\theta_{(k)}} q(\xi|\tilde{X}, \theta_{(k)}) q(\theta_{(k)}|X, W) \mathrm{d}\theta_{(k)} = \int_{\theta_{(k)}} q(\xi, \theta_{(k)}|X, W) \mathrm{d}\theta_{(k)},$$

where $q(\xi, \theta_{(k)}|X, W) := q(\xi|\tilde{X}, \theta_{(k)}) q(\theta_{(k)}|X, W)$. Thus, by integrating over $\theta_{(k)}$, this approach allows incorporating information contained in the posterior distribution of $\theta_{(k)}$ when determining ξ.

1.1.1 Implementation in the *stm* package

The R package *stm* implements a simple OLS regression through its *estimateEffect* function. However, this approach ignores that the sampled topic proportions are restricted to $(0, 1)$. As expected, using this framework we frequently observe predicted proportions outside of $(0, 1)$. Moreover, credible intervals are non-informative, due to violated model assumptions.

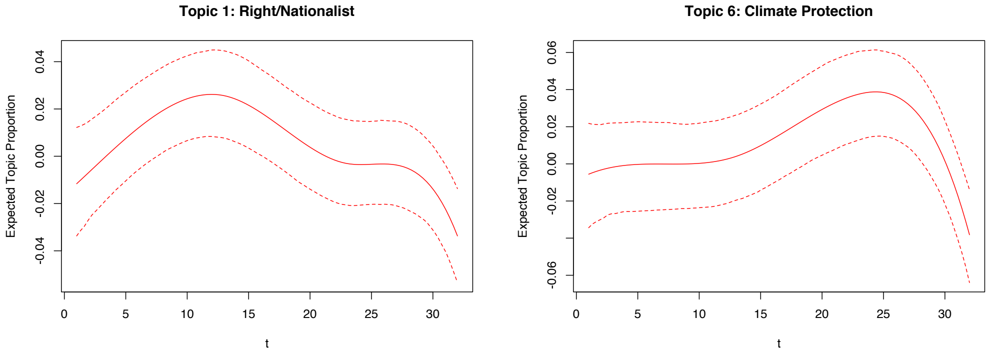

Figure 1: Estimated prevalence of topics 1 and 6 over time, generated using *estimateEffect* from the *stm* package

1.1.2 Alternative implementation

We can attempt to improve the approach employed within the *stm* package by replacing the OLS regression with a regression model that assumes a dependent variable in the interval $(0, 1)$. However, note that since topic proportions are modeled separately, regardless of the specific model implied, distributional assumptions about $\theta_{(k)}$ will be violated. This is due to the fact that the the distribution of a subvector - and thus particularly of a single component - of θ_d is not of a simple form, when θ_d follows a logistic normal distribution, see e.g. Atchison and Shen (1980).

As shown by Atchison and Shen (1980), a distribution that can be used to approximate a logistic normal distribution is the Dirichlet distribution. However, note that the Dirichlet distribution assumes less interdependence among components than implied by the logistic normal distribution. In case of the Dirichlet distribution the univariate marginal distributions are beta. One possibility is thus to perform a separate beta regression for each topic proportion on $\tilde{X}$.

As an alternative approximation we can employ a quasibinomial generalized linear model (GLM). Topic proportions can be rescaled and discretized and topics comprehended as classes, such that each rescaled topic proportion can be interpreted as the "number of successes" for the respective class. To match the underlying logistic normal distribution more closely, the quasi-likelihood furthermore allows for a flexible variance specification.

Note that $q(\xi|\theta_{(k)}, \tilde{X})$ is asymptotically normal for both the beta regression, see Ferrari and Cribari-Neto (2004), p. 17, and the quasibinomial GLM, see e.g. Fahrmeir et al. (2007), p. 285. Furthermore, in both cases we use a logit-link.

1.1.3 Visualization

We now apply the method of composition, based on either a beta regression or a quasibinomial GLM, in order to visualize covariate effects. Here we only visualize the results obtained by the quasibinomial GLM; the results of the beta regression, which show similar trends, are found in the appendix. Setting the number of simulations to 100, we sample $\xi_1^*, \dots, \xi_{100}^*$ from the marginal posterior distribution $q(\xi|X, W)$. As mentioned, when visualizing the impact of a particular covariate, all other covariates are held at their median (or majority vote, if categorial), in line with the methodology employed in the *stm* package. Let $\tilde{X}^*$ denote the subset of X where, apart from the variable of interest, each selected column consists of the median of the respective column of X. In order to plot the predicted effects, we then input $\tilde{X}^*\xi^*$ into the sigmoid function, which is the response function corresponding to a regression with logit-link, and calculate the predicted proportions.

We exemplarily illustrate the relationship between covariates and topic proportions for topic 4 ("Social/Housing") and topic 6 ("Climate Protection"). The linear predictor of our regressions takes the same form as in (1.1), i.e., we do not use a subset $\tilde{X}$, but the full set of prevalence covariates X, in order to estimate the effects, although we do not display each covariate included. For smooth effects, it is important to recall that their borders are inherently unstable, which is why one should refrain from (over-)interpreting them. For both continuous and categorical variables, black lines indicate the mean, and the shaded area represents 95% credible intervals.

For topic 4, "Social/Housing", we observe that most continuous variables have a small effect in absolute terms: the absolute variation in topic proportion across the covariate domains merely amounts to 4%, compared to 8% for topic 6. For most covariates the trend is rather ambiguous. Somewhat surprisingly, a very high unemployment rate is negatively

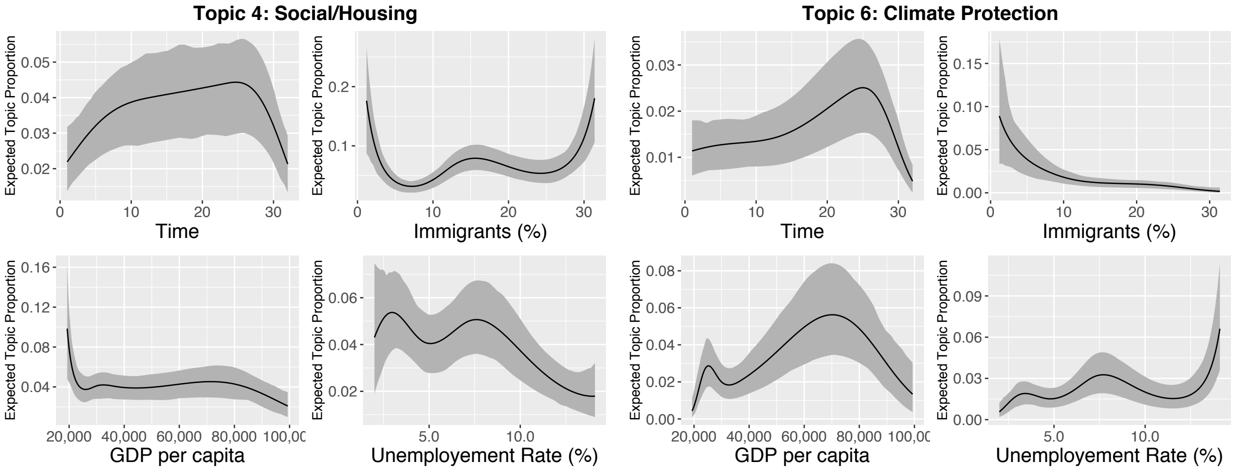

Figure 2: Mean and 95% credible intervals for smooth effects, obtained using a quasibinomial GLM.

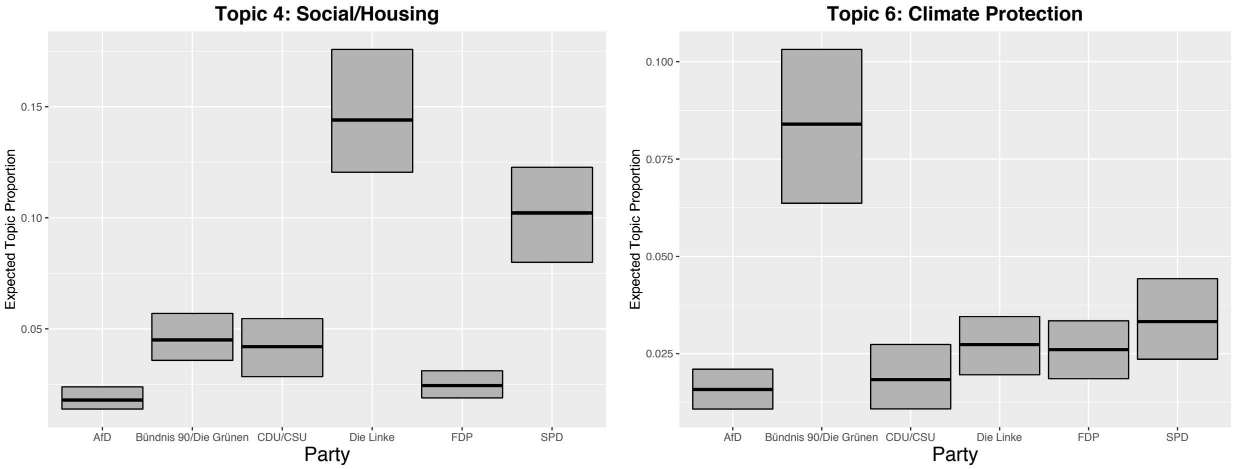

Figure 3: Mean and 95% credible intervals for different political parties, obtained using a quasibinomial GLM.

linked to topic 4.

The effect of the political party on the relevance assigned to the topic "Social/Housing" is very much in line with a priori expectations: the left party and social democrats have the highest topical prevalence (15% and 10%, respectively), and the nationalist party the lowest (2%).

For the smooth effects of topic 6, we observe its prevalence peaks in September 2019, corresponding to month t=25, decreasing afterwards. The absolute changes in topic proportions over time are rather small (around 3%). The percentage of immigrants within an electoral district shows a negative relation to topic 6. Furthermore, topic 6 tends to be discussed more frequently in mid-income electoral districts than in high- and low-income districts. Finally, the link to the unemployment rate is somewhat ambiguous, although generally rather positive.

Regarding the relationship between the political party and the prevalence of topic "Climate Protection", as to be expected, we find high topical prevalence for the green party. Similar to the smooth effects, total variation in topic proportions across parties amounts to approximately 8%.

Finally, the graph below shows a summary comparison of topical prevalence across all parties, for topics "Right/Nationalist", "Climate Protection" and "Social/Housing". The results are generally consistent with expectations. The proportions of topics "Climate Protection" and "Social/Housing" vary between 2% and 9% and between 2% and 15%, respectively. For topic 1, "Right/Nationalist", note how topical prevalence for the AfD party amounts to more than 40%, implying that more than 40% of the total content tweeted by AfD party members is about right-wing/nationalist issues, particularly immigration; for all other parties, topic 1 is rather marginal below 3%.

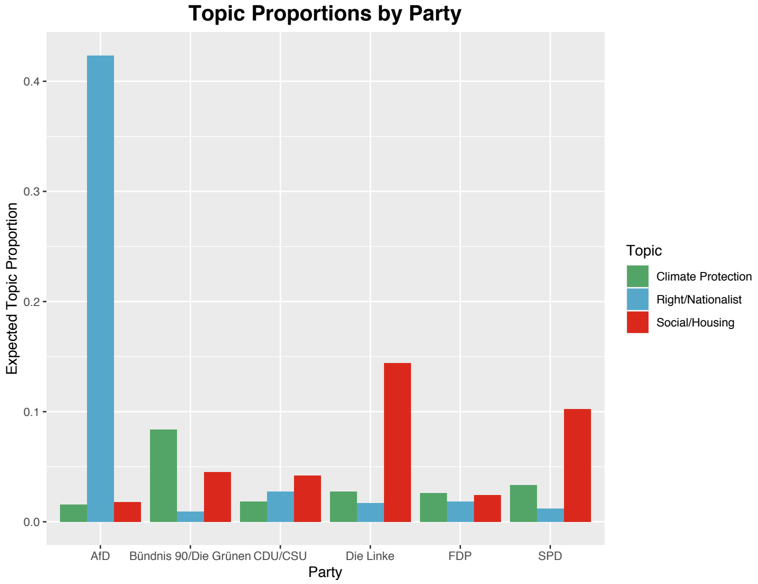

Figure 4: Topical prevalence by political party for topics 1, 4, and 6.

1.2 Direct assessment using $\hat{\Gamma}$ and $\hat{\Sigma}$

The *stm* being an extension to the correlated topic model (CTM), it is assumed that the topic proportions follow a logistic normal distribution, such that $\theta_d \sim \text{LogisticNormal}_{K-1}(\Gamma^T x_d^T, \Sigma)$. Within the CTM, the Dirichlet distribution of the LDA has been replaced with a logistic normal distribution, in order to allow for a joint dependence among topics. Therefore, as mentioned above, separately modeling topic proportions is a simplification; in particular credible intervals should be treated with caution.

In order to examine the relation of prevalence covariates and topic proportions considering the joint dependence among the latter, we can attempt to directly use the output produced by the *stm*: inference of the *stm* involves finding the maximum-a-posteriori (MAP) estimate $\hat{\Gamma}$ and the maximum likelihood estimate $\hat{\Sigma}$.

If we are interested how a specific prevalence variable is related to topic proportions, similar to previous analyses, we can attempt to predict topic proportions based on a new design matrix X^*, where each column apart from the variable of interest corresponds to the median of the respective column of X. Ideally, in order to directly predict topic proportions, we would first draw a sample Γ^* from the posterior distribution of Γ, and subsequently sample the topic proportions θ_d^* from a logistic normal with mean parameters $((\Gamma^*)^T(x_d^*)^T, \hat{\Sigma})$, where $\hat{\Sigma}$ is the maximum likelihood estimation of Σ. The resulting topic proportions would then correspond to a sample of the posterior predictive distribution of topic proportions. Unfortunately, the output of the stm does not allow for the possibility to draw a sample from the posterior distribution of Γ, but only provides its MAP estimate $\hat{\Gamma}$.

Nevertheless, in order to get an impression how the assumed generative process of topic proportions in the stm behaves, we can plug in the estimates $\hat{\Gamma}$ and $\hat{\Sigma}$ into the logistic normal distribution and visualize sampled values from this distribution. Given a new observation x_d^*, we can sample θ_d^* from $\text{LogisticNormal}_{K-1}(\hat{\Gamma}^T(x_d^*)^T, \hat{\Sigma})$ by

1. Drawing $\eta_d^* \sim \mathcal{N}_{K-1}(\hat{\Gamma}^T(x_d^*)^T, \hat{\Sigma})$ and setting $\eta_{d,K}^* = 0$.

2. Mapping to the simplex, i.e., for all $k = 1, \ldots, K$: $\theta_{d,k}^* = \frac{\exp\left(\eta_{d,k}^*\right)}{\exp\left(\sum_{i=1}^K \eta_{d,i}^*\right)}$.

3. Setting $\theta_d^* := (\theta_{d,1}^*, \ldots \theta_{d,K}^*)^T$.

We have repeated the above steps 1000 times for each input value of a selected variable, while fixing other variables at their median, and obtained the empirical mean as well as 95% credible intervals. Plotting the results, we observe that while the mean shows a similar trend to our previous analyses, the obtained credible intervals are much broader.

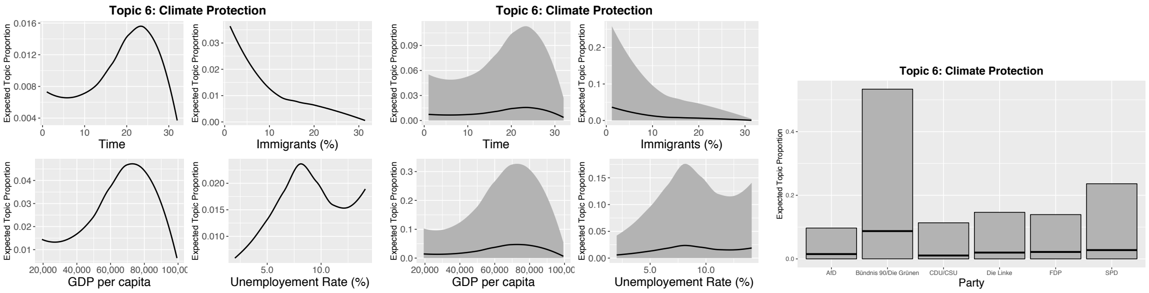

Figure 5: Smooth effects without credible intervals (left), smooth effects with credible intervals (mid), and effect of the political party (right).

The large fluctuations for a specific topic proportion can be ascribed to the fact that the unnormalized topic proportions are drawn from a $K-1$-dimensional *multivariate* normal distribution, before the softmax is applied. Therefore, a single normalized proportion depends heavily on the sampled unnormalized proportions of the remaining topics. While the variance of a topic-specific unnormalized proportion is independent of the remaining unnormalized proportions and c.p. constant for an increasing number of topics, the application of the softmax function induces a large increase in the variance of a topic-specific normalized proportion.

We suspect that the magnitude of credible intervals in figure 5 provides a more realistic picture than in case of a separate modeling of topic proportions, since the usage of the logistic normal distribution of topic proportions is an implicit assumption made within the stm that there is a dependence among topics, as argued above. This ultimately produces a large variance of the univariate marginal distributions of topic proportions, as can be observed. While ideally we should sample Γ from its posterior distribution instead of plugging in its MAP estimate, our results nevertheless suggest that there is a discrepancy between the assumed distribution of topic proportions in the generative process of the stm, and the impression we gain of the distribution of topic proportions from a separate modeling of topics within the method of composition.

2 Train-test Split

In our analyses from section 4.4, we first estimated the latent topic proportions using the stm, and then assessed the relation between these document-level topic proportions and prevalence covariates. In particular, the documents that were used to obtain the topic proportions were the same that were subsequently used to quantify relationships between covariates and topic proportions. As Egami et al. (2018) argue, this double usage of data is a form of overfitting and hence inferences about covariate effects are biased. Additionally, since in the stm prevalence covariates affect estimated topic proportions, there is not only a mere double usage of data (i.e., in the sense that the same documents are used twice), but also a direct double usage of prevalence covariates, as the estimated latent topic proportions are regressed on the former.

Both problems can be addressed using the framework proposed by Egami et al. (2018). The general idea is to split the data $\mathcal{D}$ into a training set $\mathcal{D}_{\text{train}}$ and a test set $\mathcal{D}_{\text{test}}$, and utilize the training set in order to determine a model to infer latent topic proportions from any text assumed to be generated by the same underlying process as the training set. Subsequently, this estimated model is applied on the test set, in order to assess the relation between test set topic proportions and test set prevalence covariates. In the following, we will explain the exact procedure for the stm (note that Egami et al. (2018) focus, for the most part, on the general framework, while the exact application within the stm is not discussed in-depth) and evaluate the results when applied to our data.

2.1 Model Estimation on the Training Set

On the training set, we estimate components of the stm similarly to the estimation on the full data set. That is, we input documents, i.e., words and metadata from the training set, and obtain estimates $(\hat{\beta}_{\text{train}}, \hat{\Gamma}_{\text{train}}, \hat{\Sigma}_{\text{train}})$, where $\hat{\beta}_{\text{train}}$ is associated with the topic-word distribution, and $\hat{\Gamma}_{\text{train}}$ as well as $\hat{\Sigma}_{\text{train}}$ are the topical prevalence parameters.

2.2 Prediction of Topic Proportions on the Test Set

Prediction of the topic proportions on the test is not straightforward, since the topic proportions are latent and the stm is not built for the purpose of predicting these latent variables on a set of new, unseen data. The fundamental idea is to estimate the variational posterior of the latent variables, that is, the topic proportions θ_d, where $d \in \mathcal{D}_{\text{test}}$ (note that z_d is integrated out in the stm), conditioned on the model parameters $(\hat{\beta}_{\text{train}}, \hat{\Gamma}_{\text{train}}, \hat{\Sigma}_{\text{train}})$ from the training set, as well as the words W_{test} from the test set. This functionality is implemented in the *stm* package through the function *fitNewDocuments*, which per default outputs the MAP estimates of topic proportions θ_d, for all $d \in \mathcal{D}_{\text{test}}$. Note that estimating the variational posterior of the latent variables, conditioned on the parameters and the words, is precisely what occurs during each E-step of the EM Algorithm. Thus, the implementation of *fitNewDocuments* simply consists of one E-step with inputs $(\hat{\beta}_{\text{train}}, \hat{\Gamma}_{\text{train}}, \hat{\Sigma}_{\text{train}}, W_{\text{test}})$. It is, however, not obvious how to exactly input $\hat{\Gamma}_{\text{train}}$ and $\hat{\Sigma}_{\text{train}}$ into the E-step. Depending on the characteristics of the specific analysis conducted by the researcher, Egami et al. (2018) propose three different alternatives:

1. **Covariate-specific prior**: Before applying the E-step, $\hat{\Gamma}_{\text{train}}$ is used to obtain $\hat{\mu}_d := (\hat{\Gamma}_{\text{train}})^T (x_d)^T$, for each document $d \in \mathcal{D}_{\text{test}}$ in the test set. Each document is then updated performing the E-step with inputs $(\mu_d, \Sigma) = (\hat{\mu}_d, \hat{\Sigma}_{\text{train}})$ together with the respective document specific words as well as $\hat{\beta}_{\text{train}}$ (for the exact update machanism see pp. 992-993, Roberts et al. (2013)). The problem with this approach is, however, that for two documents from the test set containaing the exact same words, different topic proportions are predicted if the prevalence covariates differ. However, in such a case we would want the causal effect of the covariates on the topic proportions to be zero.

2. **Average prior**: The average prior circumvents the above described problem of the covariate-specific prior by simply using - for each document in the test set - the average $\overline{\mu}_{\text{train}} := \frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{d \in \mathcal{D}_{\text{train}}} (\hat{\Gamma}_{\text{train}})^T (x_d)^T$ of all document-specific means from the training set. The covariance $\hat{\Sigma}_{\text{train}}$ is recalculated based on the new average $\overline{\mu}_{\text{train}}$ according to formula (11) on p. 993, Roberts et al. (2013). In this scenario, prevalence covariates from the test have no influence at all on the predicted topic proportions.

3. **No prior**: If no prior is used, then for each document $d \in \mathcal{D}_{\text{test}}$ in the test set the E-step is performed using $\mu_d = 0$ and replacing $\hat{\Sigma}_{\text{train}}$ with a diagonal covariance matrix with very large diagonals.

The covariate-specific prior cannot be used in our case due to the above described problem, that different topic proportions are predicted for identically worded test set documents, if their prevalence covariates differ. The option "no prior" can be useful if the metadata on the test set is believed to be linked differently to topics than is the case on the training set. In most cases the second option, "average prior", should provide the best trade-off, since in this case metadata from the training set is directly used to predict topic proportions, but the problem of the covariate-specific prior is solved. Note that hence in this case there is no double usage of covariates.

2.3 Results

We now depict the results obtained conducting a train-test split, where we split the data into two equally sized sets, and choose the option "average prior". Note that the test data cannot consist fo words which have not been seen in the training data. Therefore, all previously unseen words are removed from the test data. After removing the words, the test data contains 80.6% of the original words.

In contrast to section 4.4., the focus in this section will be on quantifying causal effects between covariates and the amount a topic is discussed, since the train-test framework is most appropriate in order to conduct such analyses. As mentioned, the function *fitNewDocuments* outputs the MAP estimates of the variational posterior of topic proportions for the test set. In Figure 6 we these MAP estimates of topic proportions, together with the topic proportions obtained for the training data.

The UN Climate Action Summit 2019 was held on 23 September 2019. As can be observed, the topic associated with climate issues was discussed to a much larger extent during this time than a year earlier. While the map estimates for the different prior specifications on the test set are rather similar, the estimated effect is much larger training for the training data. If we compare the estimated topic proportions for a topic we labelled as 'Emancipation' for the two opposing parties 'AfD' and 'Bündnis 90/Die Grünen', we find similar results: the average difference of estimated topic proportions between both parties is larger for the training data. Also, note that the variation is higher on the training data compared to the test data in both cases.

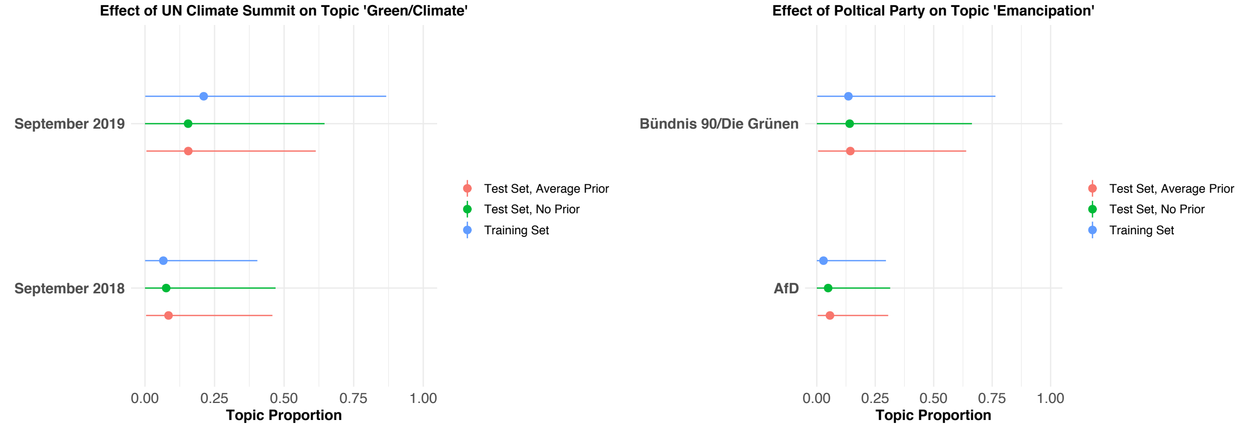

Figure 6: Maximum-a-posteriori (MAP) estimates of topic proportions on training and test data.Points display the mean, the lines 2.5% and 97.5% credible intervals

.

If we want to estimate the treatment effect, we can estimate the average difference of MAP estimates between both groups. Following Egami et al. (2018) we obtain an estimate of the Average Treatment Effect (ATE) on a data set $\mathcal{D}$ as

$$\widehat{\text{ATE}} = \frac{1}{|\mathcal{D}_{\text{treatment}}|} \sum_{i \in \mathcal{D}_{\text{treatment}}} \hat{\theta}_i - \frac{1}{|\mathcal{D}_{\text{control}}|} \sum_{i \in \mathcal{D}_{\text{control}}} \hat{\theta}_i, \tag{2.1}$$

where $\hat{\theta}_i$ are the MAP estimates for the i-th document. Egami et al. (2018) show that, if additional conditions hold, the estimated $\widehat{\text{ATE}}$ on previously unseen test data $\mathcal{D}_{\text{test}}$ is an unbiased estimate of the ATE.

In Figure 7 we visualize the ATE estimated on training and on test data with different prior specifications (note that this is simply the difference of the means depicted in Figure 6). The results correspond to our classical idea of overfitting: since the characteristics of each parliamentarian associated with a document have been used to estimate the topic proportions in the first place, when evaluating these topic proportions on the same data, the fit is optimistically biased.

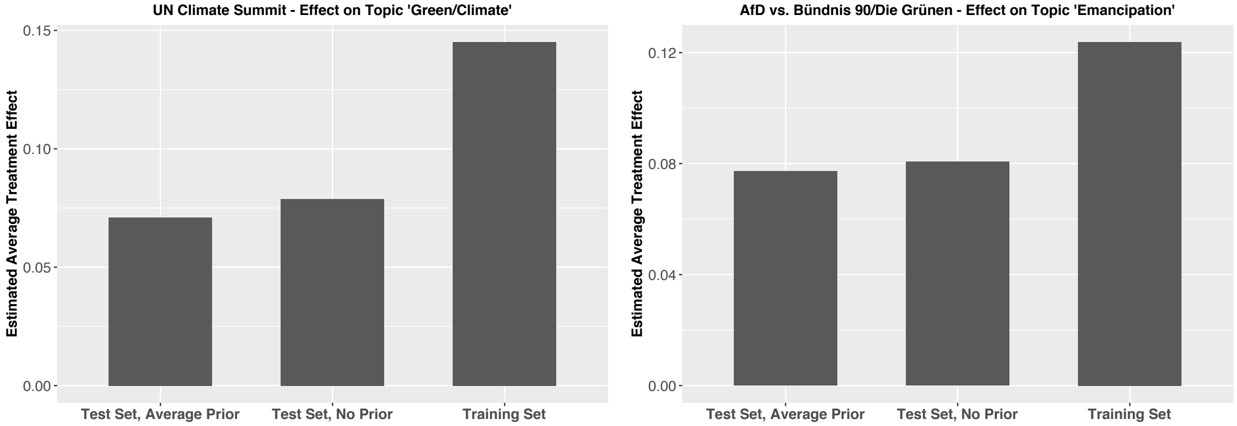

Figure 7: Estimated Average Treatment Effects (ATE) using training and test data

3 Appendix 1

In line with Wang and Blei (2013), consider a generic topic model with latent variables θ and z as well as observed data x:

$$p(\theta, z, x) = p(x|z)p(z|\theta)p(\theta).$$

The exact posterior distribution

$$p(\theta, z|x) = \frac{p(\theta, z, x)}{\int p(\theta, z, x)dzd\theta}$$

is usually intractable due to the high-dimensional integral, which is why the distribution needs to be approximated.

As stated in section 2.3, in variational inference a simple distribution family $q(\theta, z)$ is posited and subsequently, we determine the member of this family - that is, the variational parameter(s) - that minimizes the KL divergence. Note that, for computational purposes, we compute KL divergence of the true posterior p from the approximating posterior q, $KL(q||p)$, whereas intuitively one would seek to minimize $KL(p||q)$.

The most popular variational inference technique is mean-field variational inference (also: mean-field variational Bayes), where we posit full factorizability of $q(\theta, z)$: $q(\theta, z) = q(\theta)q(z)$. That is, θ and z are assumed to be independent with their own distributions and variational parameters ϕ (which we suppress for improved readability). Since θ and z are actually dependent, this approximate distribution family $q(\theta, z)$ does not contain the true posterior $p(\theta, z|x)$.

Let us now write out the KL divergence of p from q:

$$\begin{aligned} KL(q||p) &= \mathbb{E}_q[log\frac{q(\theta, z)}{p(\theta, z|x)}] \\ &= \mathbb{E}_q[log(q(\theta, z))] - \mathbb{E}_q[log(p(\theta, z|x))] \\ &= \mathbb{E}_q[log(q(\theta, z))] - \mathbb{E}_q[log(p(\theta, z, x))] + log(p(x)) \end{aligned}$$

Since $KL(q||p) \geq 0$ (which can be easily shown using Jensen's inequality), it follows that:

$$log(p(x)) \geq \mathbb{E}_q[log(p(\theta, z, x))] - \mathbb{E}_q[log(q(\theta, z))].$$

The left-hand side of the above inequality is the marginal log likelihood of observed data x and is also called evidence (of the observed data). Note that the evidence is not computable - otherwise we would not need to resort to variational inference in the first place. The right-hand side thus presents a lower bound on the evidence and we define the *Evidence Lower*

BOund (ELBO) as:

$$ELBO := \mathbb{E}_q[log(p(\theta, z, x))] - \mathbb{E}_q[log(q(\theta, z))],$$

where the second component of the ELBO, $\mathbb{E}_q[log(q(\theta, z))$, is the entropy of the approximate distribution q. Equivalently, we could say that the evidence constitutes an upper bound for the ELBO. This means that we actively maximize the ELBO (which is therefore also called *variational objective*), which in turn is equivalent to minimizing the KL divergence of the true posterior $p(\theta, z|x)$ from the approximate distribution $q(\theta, z)$. Therefore, the approximation $q(\theta, z)$ - or, more precisely, the variational parameters ϕ of $q(\theta)$ and $q(z)$ - that maximizes the ELBO simultaneously minimizes KL divergence (Blei et al., 2003; Wang and Blei, 2013). Wang and Blei (2013) show that for the chosen factorization of the joint distribution $p(\theta, z, x)$, and using the optimality conditions as derived in Bishop (2006), we obtain the following solutions when setting $\frac{\partial ELBO}{\partial q} \overset{!}{=} 0$:

$$q^*(\theta) \propto exp\{\mathbb{E}_{q(z)}[log(p(z|\theta))p(\theta)]\},$$
$$q^*(z) \propto exp\{\mathbb{E}_{q(\theta)}[log(p(x|z))p(z|\theta)]\}.$$

The coordinate ascent algorithm iteratively updates one of these two expressions while holding the other one constant, but requires closed-form updates to do so. This requirement is fulfilled as long as all model nodes are conditionally conjugate, i.e., as long as for each node in the model "its conditional distribution given its Markov blanket (i.e., the set of random variables that it is dependent on in the posterior) is in the same family as its conditional distribution given its parents (i.e., its factor in the joint distribution)" (Wang and Blei (2013), p. 1008). The authors consequently define a class of models where some nodes are not conditionally conjugate, the so-called *nonconjugate models*; for this class, using Laplace approximations, the variational family is shown to be $q(\theta, z) = q(\theta|\mu, \Sigma)q(z|\phi)$; that is, $q(\theta)$ is now Gaussian with variational parameters μ and Σ.

The STM in particular constitutes a nonconjugate model, since $p(\theta)$ is logistic normal and thus not conjugate with respect to the multinomial distribution $p(z|\theta)$. Consequently, no closed-form update is available for $q(\eta)$. Using mean-field variational inference, the approximate posterior family is $\prod_{d=1}^{D} q(\eta_d)q(z_d)$, where $q(\eta_d)$ is Gaussian and $q(z)$ is binomial (Roberts et al., 2016). Given the posterior, inference now consists in finding the particular member of the posterior distribution family that maximizes the approximate ELBO. (Due to the subsequent Laplace approximation, ELBO does not constitute a true lower bound on the evidence and the updates do not maximize ELBO directly, which is why Roberts et al. (2013) use the term *approximate* ELBO. See Wang and Blei (2013) for further discussion.) Applying Laplace variational inference, we approximate $q(\eta_d)$ using a (quadratic) Taylor expression around the maximum-a-posteriori (MAP) estimate $\hat{\eta}_d$, which yields a Gaussian variational posterior $q(\eta_d)$, centered around $\hat{\eta}_d$, and allows for a closed-form solution of $q(z_d)$. Iteratively updating $q(\eta_d)$ and $q(z_d)$ thus constitutes the E-step of the EM algorithm.

The M-step consists in maximizing the approximate ELBO with respect to model parameters. Prevalence parameters Γ and Σ are updated through linear regression and maximum likelihood estimation (MLE), respectively. The updates for topic-word distributions β_k (or $\beta_{k,a}$ if a content covariate is specified) are obtained through multinomial logistic regression. Further details are provided in Roberts et al. (2013) and in the appendix of Roberts et al. (2013). Moreover, the appendix of Blei et al. (2003) provides a detailed description of variatonal inference and empirical parameter estimation for the (conditionally conjugate) LDA model.

4 Appendix

4.1 Plots of section 4.4

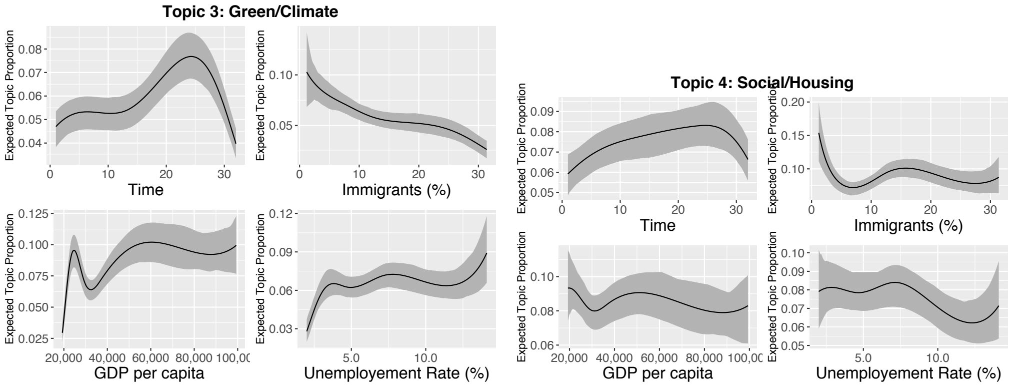

Figure 8: bla

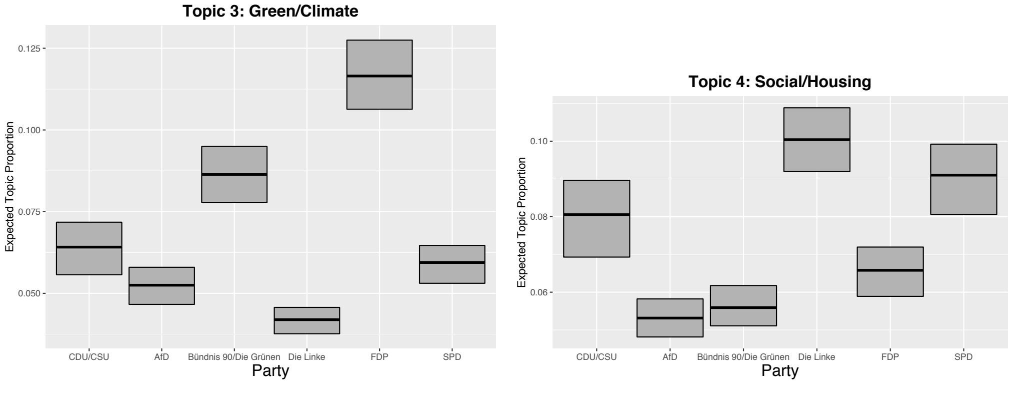

Figure 9: blabla

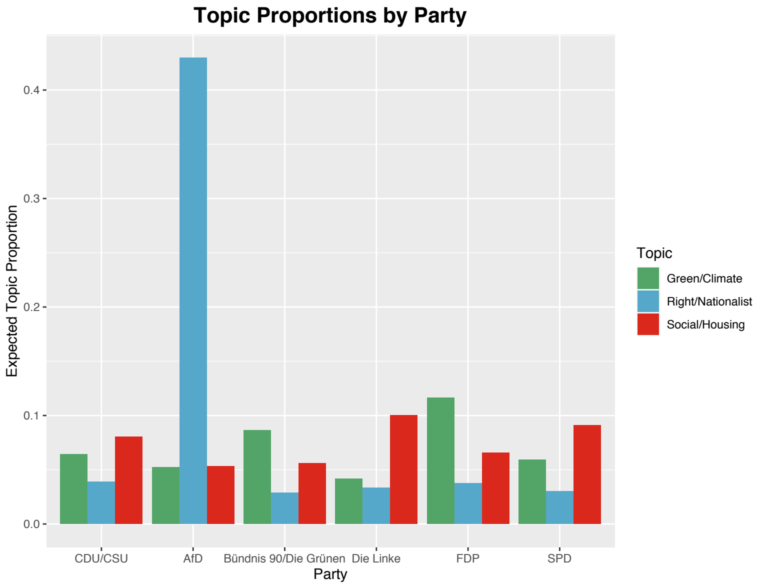

Figure 10: Topical prevalence by political party for topics 1, 2, and 3.

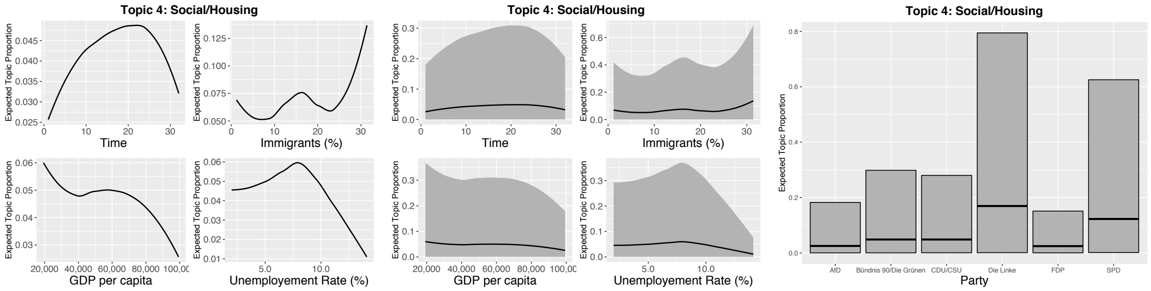

Figure 11: bla

References

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

Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Naoki Egami, Christian J Fong, Justin Grimmer, Margaret E Roberts, and Brandon M Stewart. How to make causal inferences using texts. *arXiv preprint arXiv:1802.02163*, 2018.

Ludwig Fahrmeir, Thomas Kneib, Stefan Lang, and Brian Marx. *Regression*. Springer, 2007.

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

Margaret E Roberts, Brandon M Stewart, Dustin Tingley, Edoardo M Airoldi, et al. The structural topic model and applied social science. In *Advances in neural information processing systems workshop on topic models: computation, application, and evaluation*, pages 1–20. Harrahs and Harveys, Lake Tahoe, 2013.

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

Martin A Tanner. *Tools for statistical inference*. Springer, 2012.

Chong Wang and David M Blei. Variational inference in nonconjugate models. *Journal of Machine Learning Research*, 14(Apr):1005–1031, 2013.