Covariate-level Topic Analysis

Patrick Schulze, Simon Wiegrebe

Contents

0.1 Covariate-level Topic Analysis

After this analysis of topics at a global level, in particular of their labeling and proportions, we now proceed to analyze metadata information (i.e., document-level covariates) and its impact on topic proportions. As mentioned before, the covariates included are party, state (both categorical), date (smooth effect), percentage of immigrants, GDP per capita, unemployment rate, and the 2017 vote share (the last four as smooth effects, on an electoral-district level). Since the target variable $\theta_{(k)}$ is not observable and hence estimated itself within the estimation of the STM, we recur to the method of composition to account for the uncertainty contained within $\theta_{(k)}$.

0.1.1 Method of Composition

Let $\theta_{(k)} \in [0,1]^D$ denote the proportions of the k-th topic for all D documents. Suppose that 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 $\theta_{(k)}$, $q(\theta_{(k)}|\Gamma, \Sigma, X)$, where $\Gamma := \Gamma(w, X, Y)$ and $\Sigma := \Sigma(w, X, Y)$. 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, sampling $\theta^*_{(k)}$ from the posterior distribution, performing a regression for each sampled $\theta^*_{(k)}$ on $\tilde{X}$, and then sampling from the estimated distributions of regression coefficients, provides an i.i.d. sample from the marginal posterior distribution of regression coefficients. This procedure is known as the method of composition in the social sciences (Tanner 2012, 52).

Formally, let ξ denote the regression coefficients from a regression of $\theta_{(k)}$ on $\tilde{X}$, and let $q(\xi|\theta_{(k)}, \tilde{X})$ be the approximate posterior distribution of these coefficients, i.e. given design matrix $\tilde{X}$ and response $\theta_{(k)}$.

The R package *stm* implements a simple OLS regression through its *estimateEffect* function. Using this framework we frequently observe predicted proportions outside of $(0, 1)$, given that the restricted domain of $\theta_{(k)}$ is not taken into account. Moreover, credible intervals are non-informative, due to violated model assumptions. Therefore, to adequately model the topic proportions we perform a beta regression (with logit-link), since the sampled proportions are restricted to the interval $(0, 1)$. More information on why beta regression is useful in such a scenario can be found in Ferrari and Cribari-Neto (2004). In case of a beta regression, $q(\xi|\theta_{(k)}, \tilde{X})$ is a normal distribution (see e.g. Ferrari and Cribari-Neto (2004), p. 17).

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

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

Then, $\xi_1^*, \ldots, \xi_m^*$ is an i.i.d. sample from the marginal posterior

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

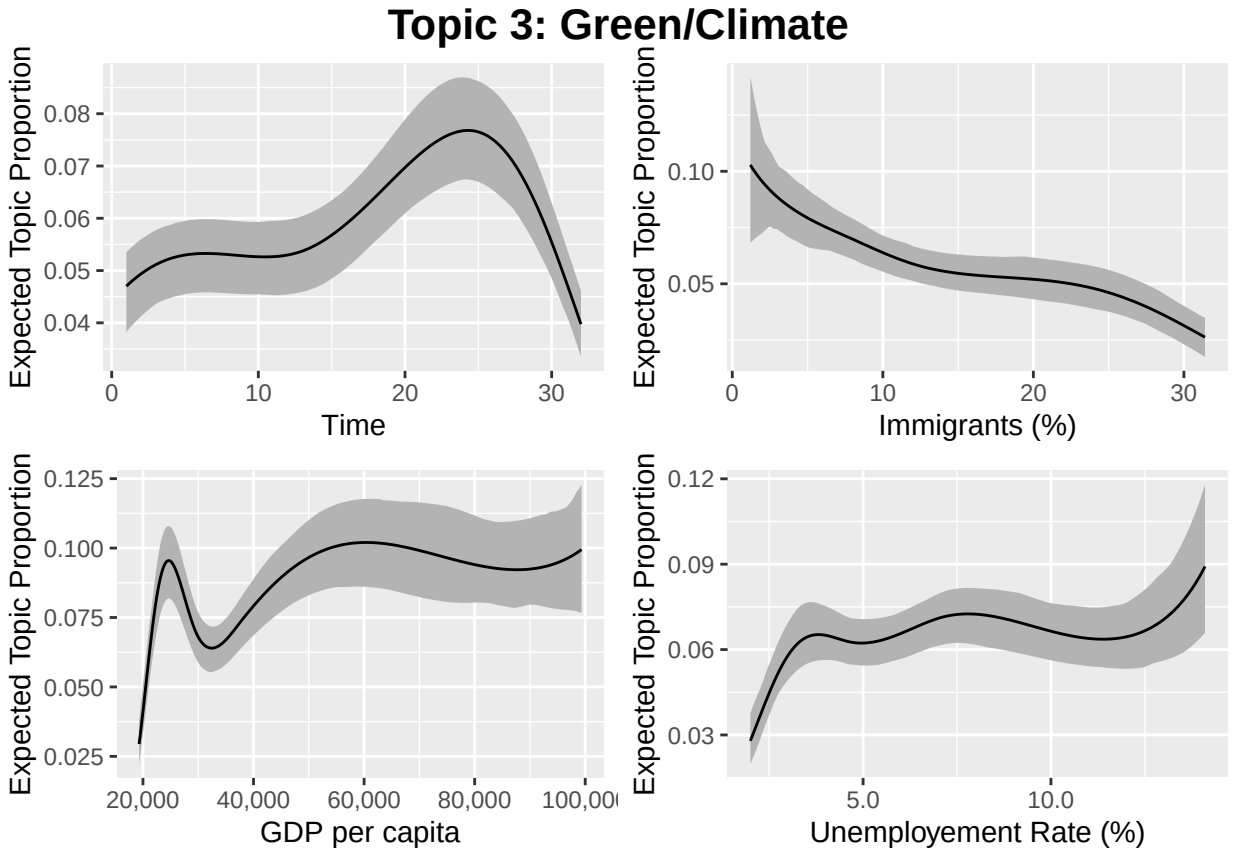
where $q(\xi, \theta_{(k)}|\Gamma, \Sigma, X) := q(\xi|\theta_{(k)}, \tilde{X}) q(\theta_{(k)}|\Gamma, \Sigma, X)$. Thus, by integrating over $\theta_{(k)}$, this approach allows incorporating uncertainty about $\theta_{(k)}$ when determining ξ.

0.1.2 Visualization

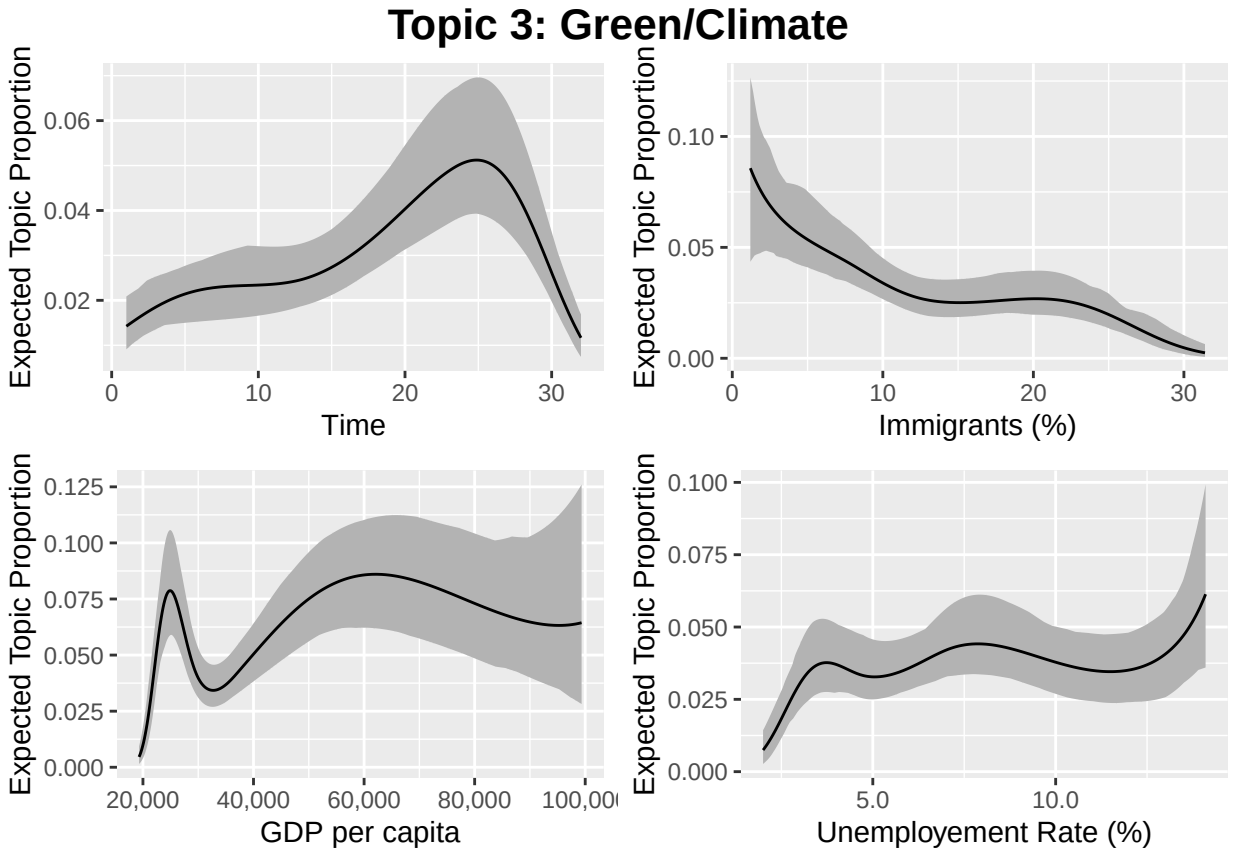
We can now apply the method of composition, based on a beta regression, in order to quantify covariate effects. Setting the number of simulations to 100, we thus sample $\xi_1^*, \ldots, \xi_{100}^*$ from the approximate posterior distribution $q(\xi|\Gamma, \Sigma, X)$. In order to plot the predicted effects, we input $\tilde{X}\xi^*$ into the sigmoid function, which is the response function corresponding to a beta regression with logit-link, and calculate the predicted proportions. When visualizing the impact of a particular covariate, all other covariates are held at their median, in line with the methodology employed in the *stm* package.

We discuss the impact of covariates on topic proportions for topics 3 (green/climate) and 4 (social/housing), sub-dividing the analysis into smooth effects (time, immigration, GDP, and unemployment) and categorical variables (party and state). 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*, the shaded area represents 95% credible intervals.

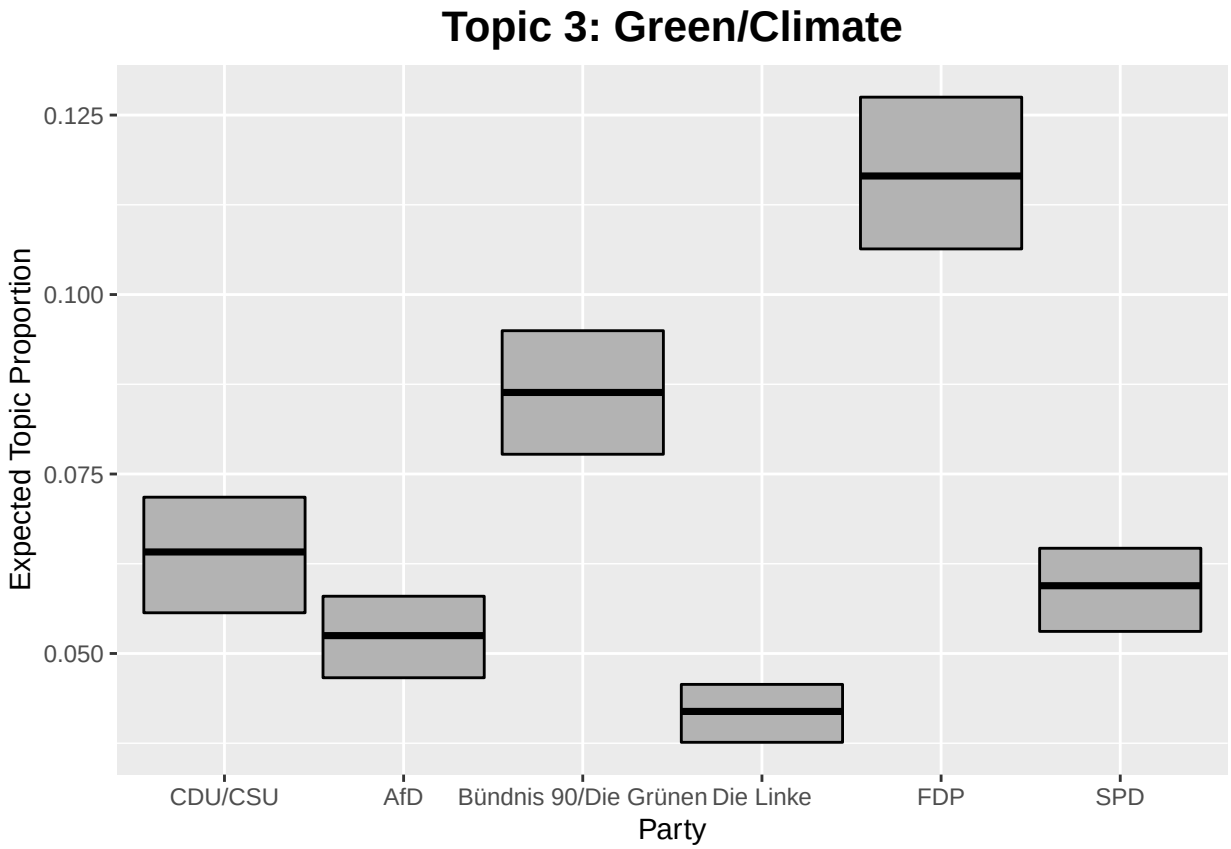
By looking at the smooth effects for topic 3 below, we find that its proportion increases over time until the 25th month, corresponding to September 2019, decreasing sharply afterwards. However this sharp decline is to be taken with caution due to the instability of splines at the borders of the covariate domain. Note that the absolute changes in topic proportions over time for the green/climate topic are rather small (around 4%). The effect of immigrants (as percentage of the total population) is negative across the entire domain, and rather steadily so. The impact of GDP per capita on topic 3 is unclear/constant, while unemployment rate show an overall positive effect.



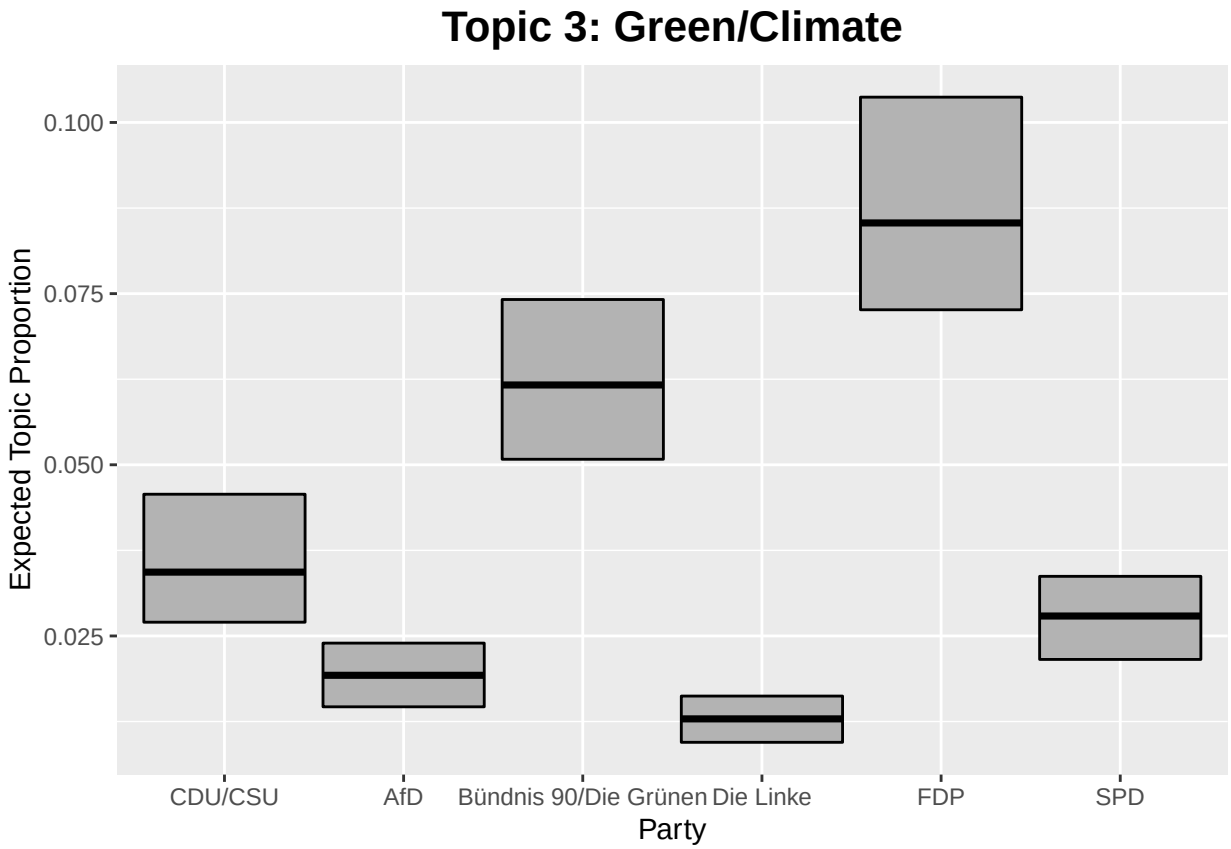
We can compare these results with the predictions using a quasibinomial GLM instead of a beta regression:



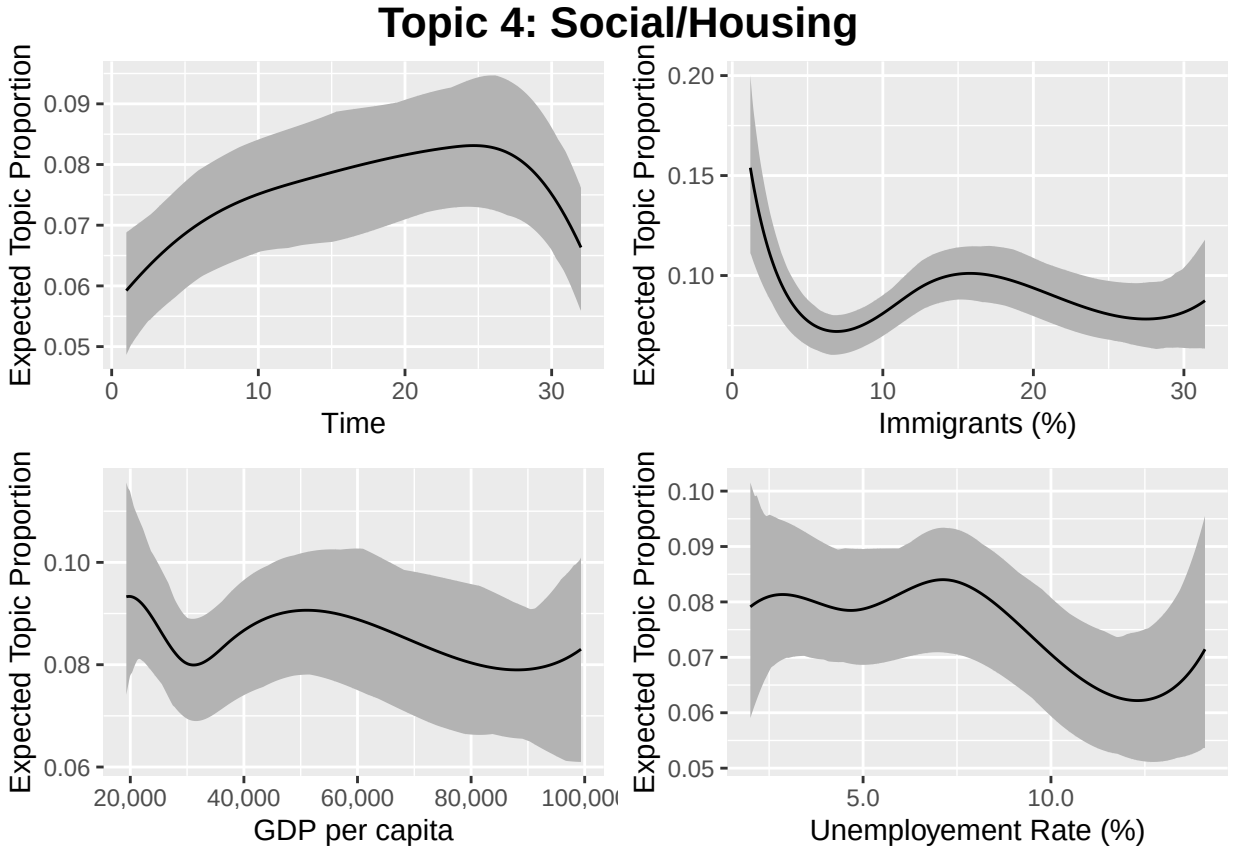
Regarding the effect of categorical variables on topic green/climate, we consider the political party, arguably the most decisive covariate. As was to be expected, we find high topic prevalence for the green party, yet the liberal party is, somewhat surprisingly, the party with the highest prevalence. Similar to the smooth effects, total variation in topic proportions across parties amounts to approximately 8%, as can be seen in the graph below.



Again, we can compare this to the quasbinomial GLM:

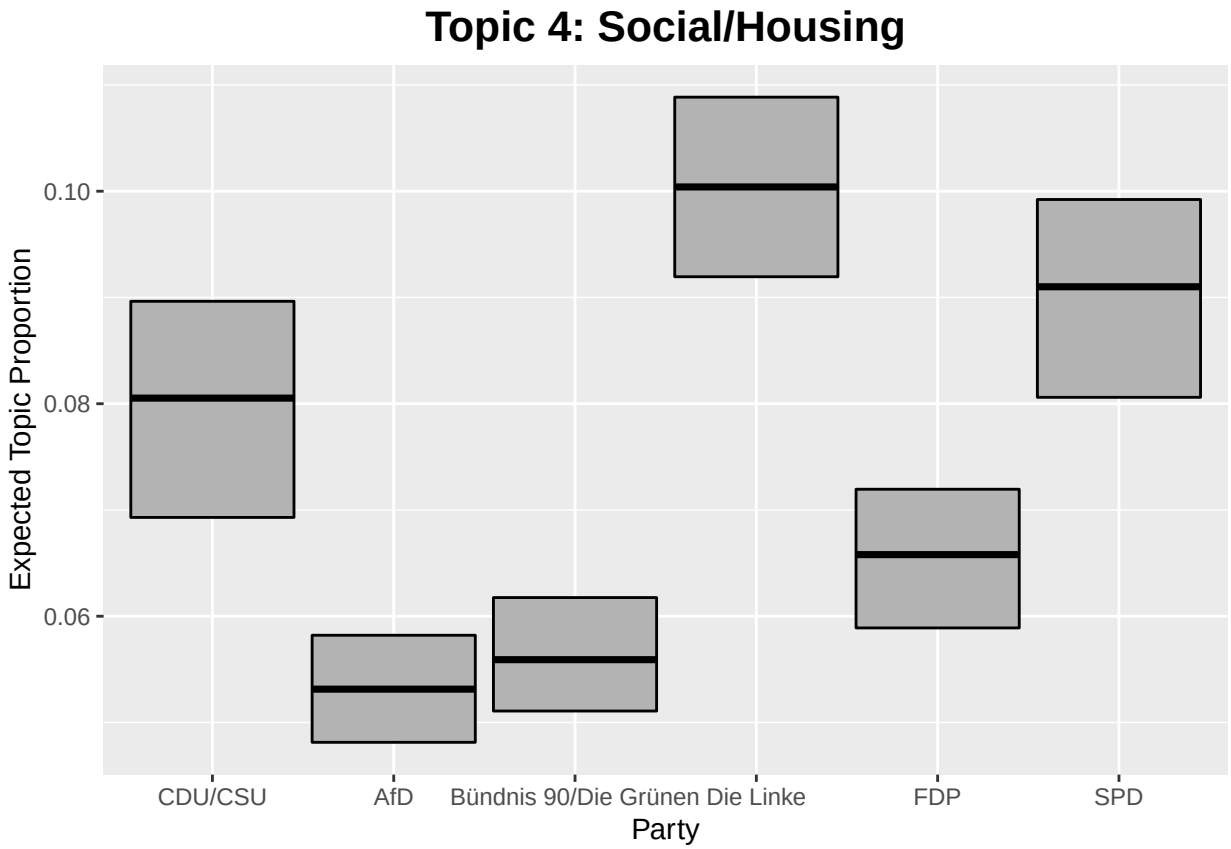


As for topic 4, social/housing, we observe that most (quasi-)continuous variables have a small effect in absolute terms: the absolute variation in topic proportion across the covariate domains merely amounts to 4%, compared to around 8% for the green/climate topic. The time effect is similar to the one for topic 3, particularly the decreasing topic prevalence since September 2019. For the other variables, no clear effect is discernible.

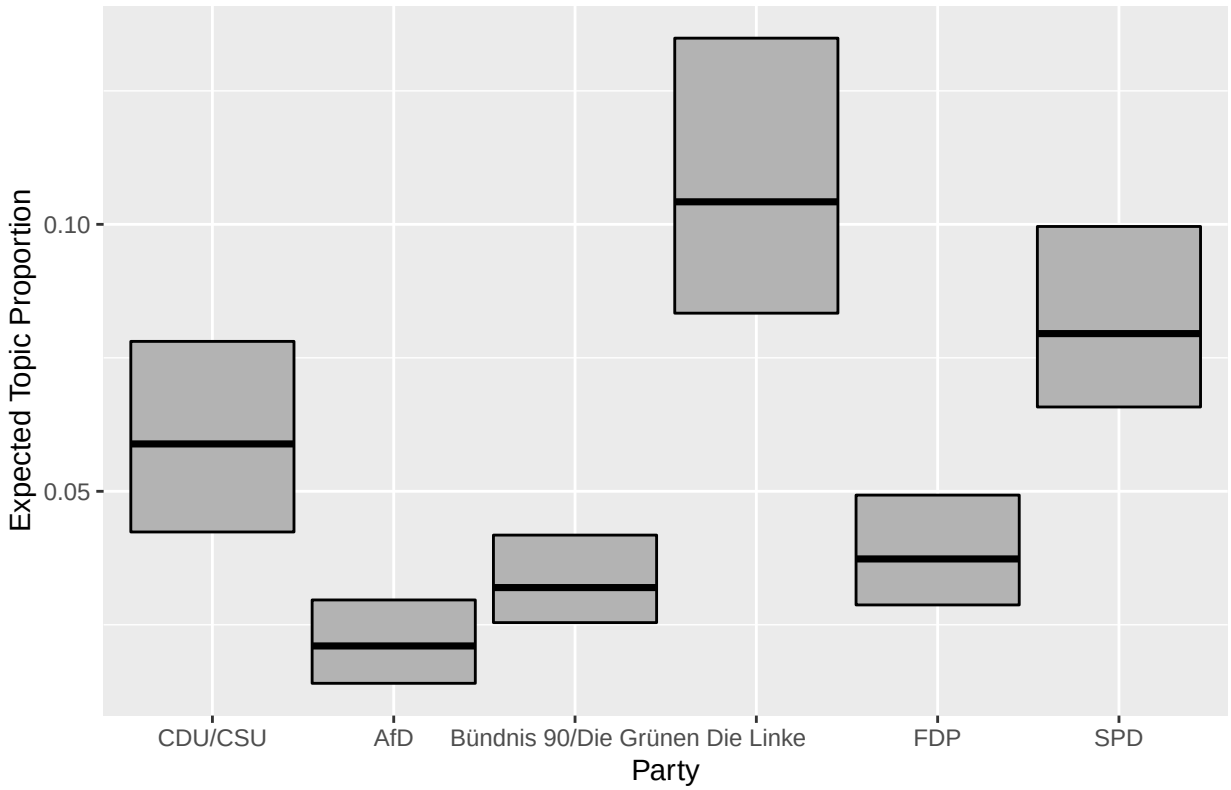


Again, we can compare this to the quasbinomial GLM:

The effect of political party on the relevance assigned to the social/housing topic is very much in line with a priori expectations: the left party and social democrats have the highest topic prevalence, at around 10%, the nationalist party the lowest one at 5%. The overall effect of covariate party is thus similar for topics green/climate and social/housing.



Again, we can compare this to the quasbinomial GLM:



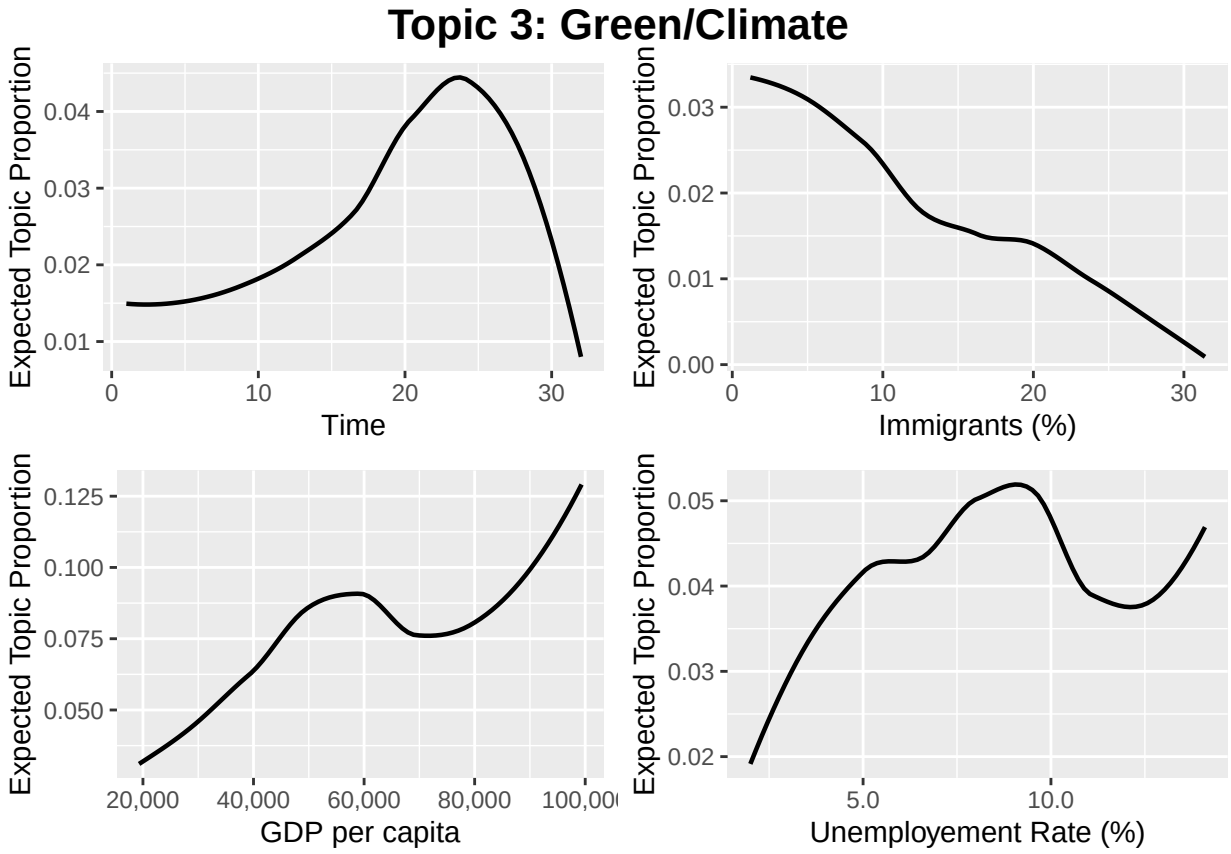
Finally, the graph below shows a summary comparison of topic prevalence across all parties, for topics right/nationalist, green/climate, and social/housing. The results are generally consistent with expectations. The proportions of topics green/climateand social/housing vary between 4% and 12% and between 5% and 10%, respectively. For topic 1, right/nationalist, note how topic 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, particulary immigration; for all other parties, topic 1 is rather marginal at 3-4%.

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

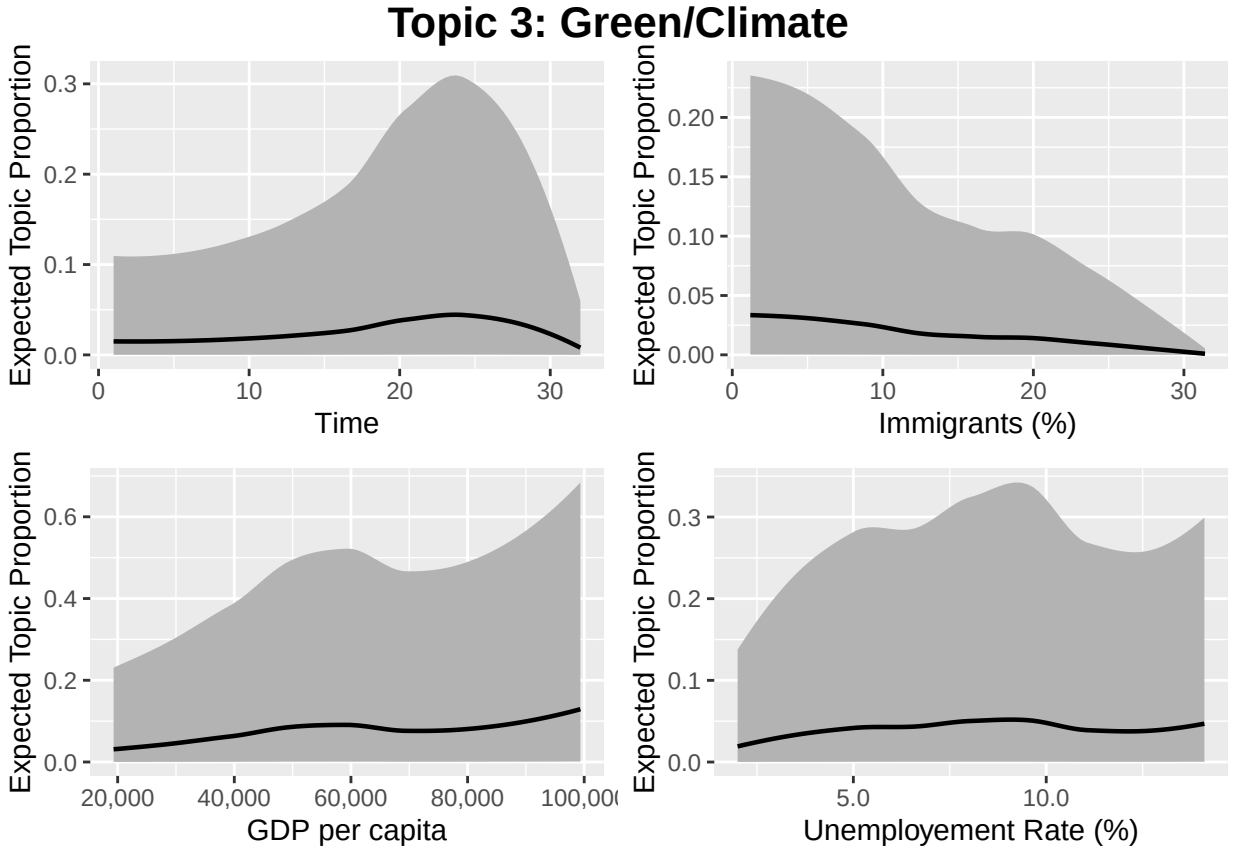
Within the *stm* it is assumed that the topic proportions follow a logistic normal distribution, such that $\theta_d \sim \text{LogisticNormal}_{K-1}(\Gamma^T x_d^T, \Sigma)$. Inference of the *stm* involves finding MAP estimates $\hat{\Gamma}$ and $\hat{\Sigma}$ and hence we can attempt to directly quantify the effect of the prevalence covariates on θ_d. For a given x_d^* we can sample θ_d^* from $\text{LogisticNormal}_{K-1}(\hat{\Gamma}^T (x_d^*)^T, \hat{\Sigma})$ by performing the following steps:

1. Draw $\eta_d^* \sim \mathcal{N}_{K-1}(\hat{\Gamma}^T (x_d^*)^T, \hat{\Sigma})$.
2. For all $k = 1, \ldots, K$: $\theta_{d,k}^* = \exp(\eta_{d,k}^*) / \exp(\sum_{i=1}^K \eta_{d,i}^*)$.
3. $\theta_d^* = (\theta_{d,1}^*, \ldots \theta_{d,K}^*)^T$.

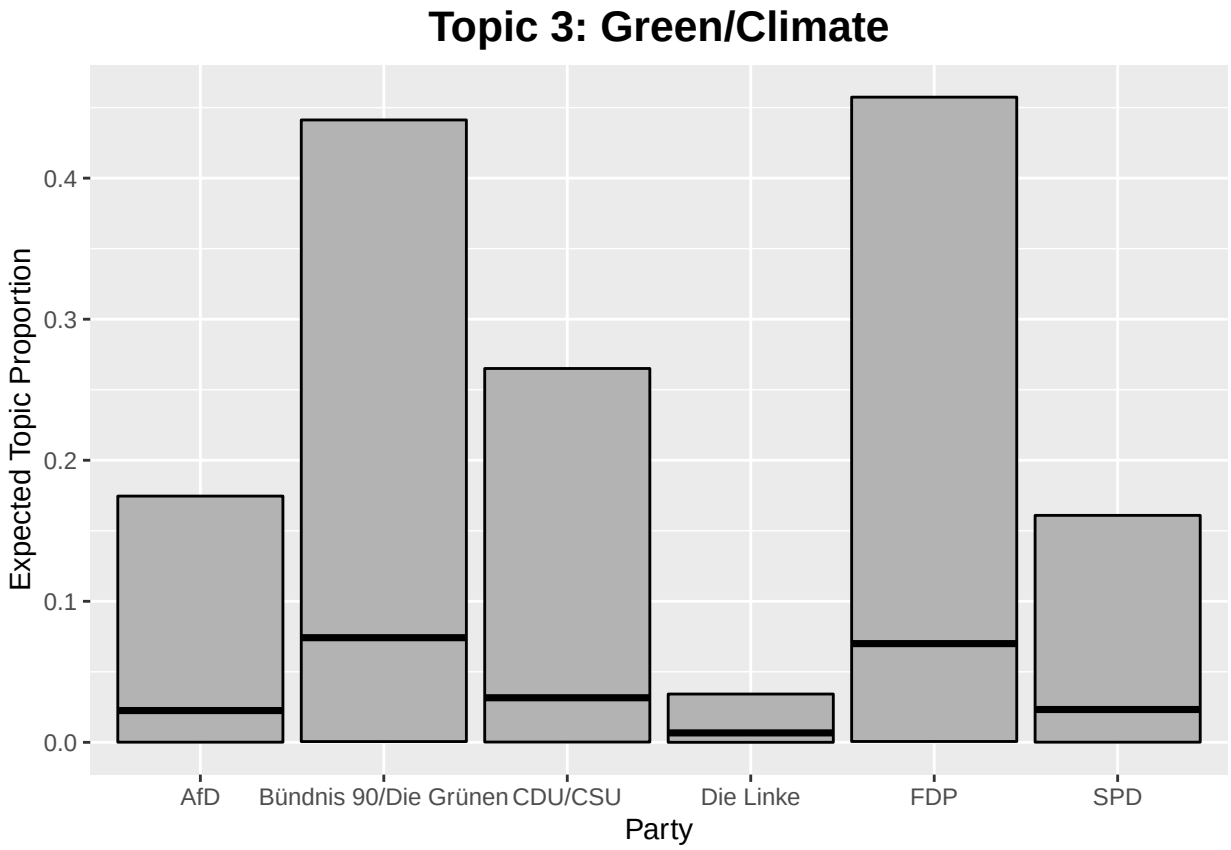
Here η_d^* denote the unnormalized topic proportions and $\eta_{d,K}^*$ is fixed to zero.



Plotting the credible intervals we observe that the spectrum of expected topic proportions is very broad:

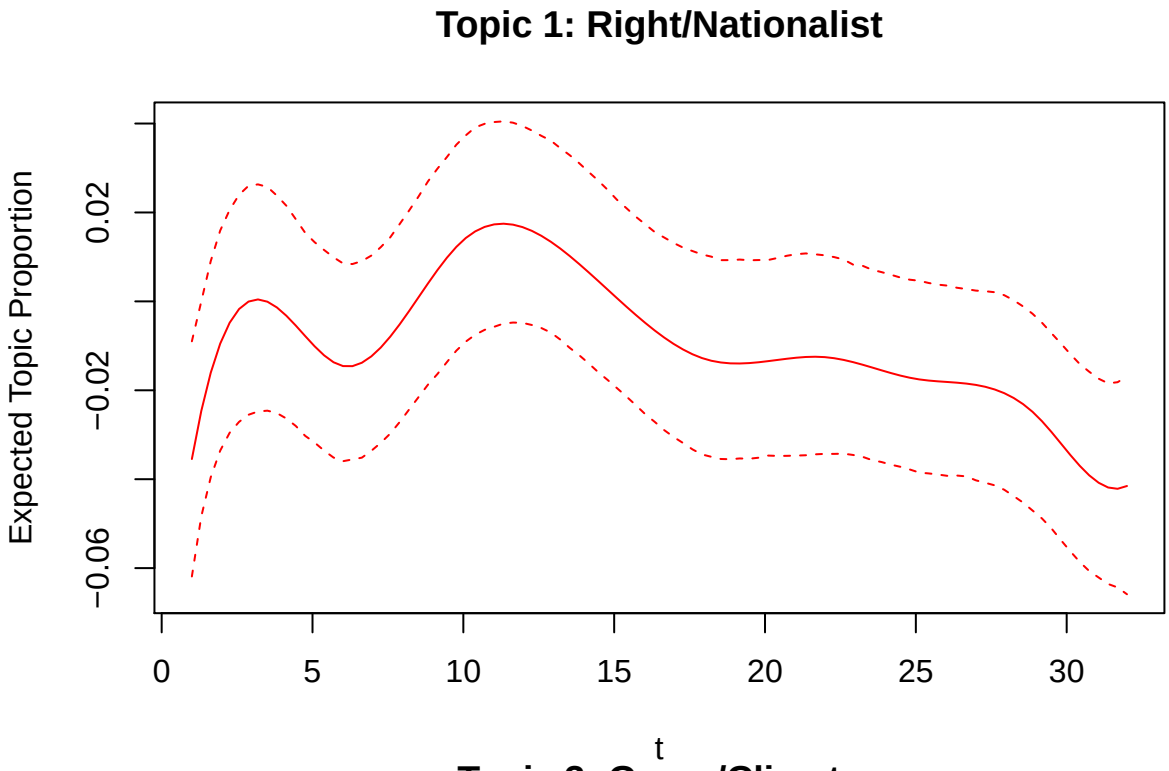


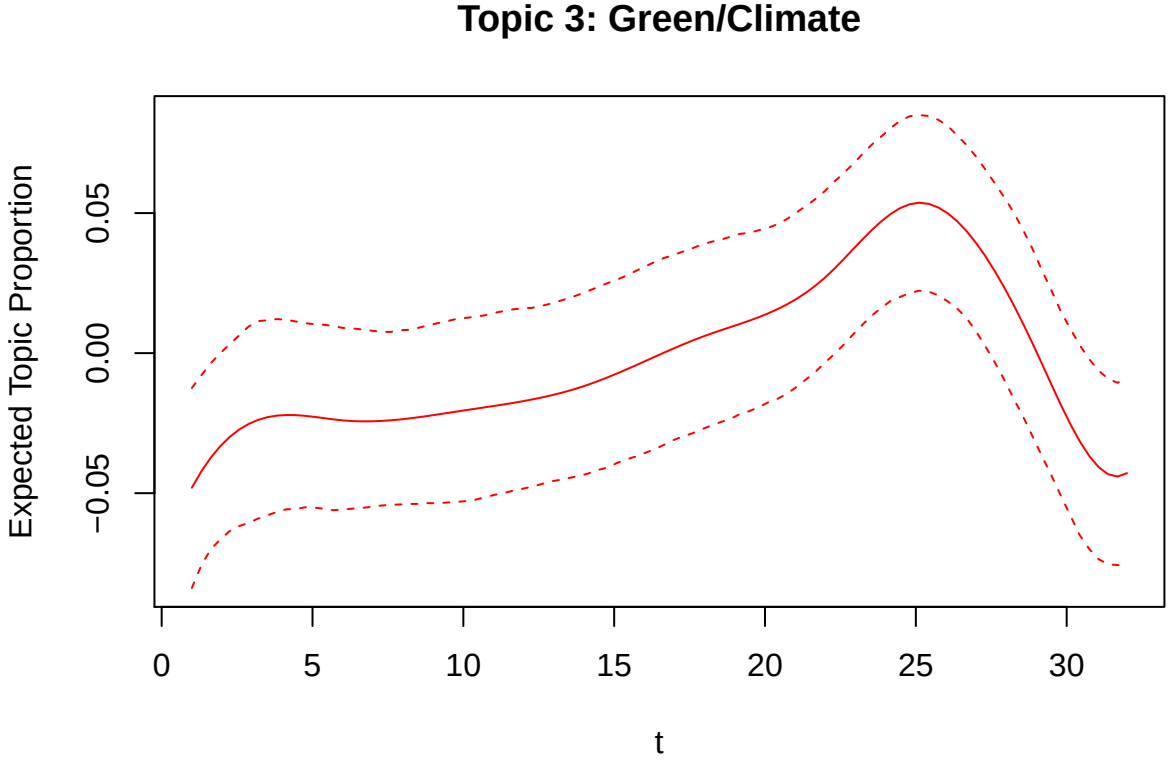
The large fluctuations for a specific topic proportion can be ascribed to the fact that the unnormalized topic proportions are drawn from a $K-1$-dimensional *multivariate* normal distribution, before the softmax is applied: The normalization of a single proportion depends heavily on the sampled unnormalized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



0.1.4 Problems with *estimateEffect*

As expected, we observe predictd topic proportions outside of $[0, 1]$:





Ferrari, Silvia, and Francisco Cribari-Neto. 2004. "Beta Regression for Modelling Rates and Proportions." *Journal of Applied Statistics* 31 (7). Taylor & Francis: 799–815.

Tanner, Martin A. 2012. *Tools for Statistical Inference*. Springer.