4_5

Patrick Schulze, Simon Wiegrebe

June 2020

Contents

1 Results

1.1 Hyperparameter Search and Model Fitting

1.2 Labeling

1.3 Global-level Topic Analysis

1.4 Covariate-level Topic Analysis

1.5 Content Model

1.6 2-Step Procedure: CTM

1.7 2-Step Approach: CTM

In sections 4.4 and 4.5, we analyzed the relationship between topic proportions and metadata, visualizing the effect of prevalence covariates and deciding against the further inclusion of a topical content variable. As briefly mentioned in section 2 already, a point of concern when using the STM is the double usage of covariates: they are used in the estimation of the topic itself (and thus, in the estimation of the latent topic proportions) and subsequently they are again used in metadata inference. From classical statistical modelling, we are used to interpret such relationships, oftentimes ascribing a causal interpretation to the corresponding coefficients; in our case, this would go along the lines of stating, for instance, that "a higher percentage of immigrants within an electoral district makes politicians prioritize issues other than climate protection", refering to Figure XXX.

Topic models, however, present a crucial difference as compared to classical statistical models: the target variable - θ - is latent and thus itself being estimated. For explorative or descriptive purposes, this does not pose a problem, because there is only a single step: discovering topics in the text documents. Yet whenever in a second step, after estimating the model, we wish to conduct (causal) inference, we face an overfitting problem, since the *same* documents and covariates are used in both steps. In this section, we focus on the double usage of (prevalence) covariates, while section 4.7 deals with double usage of documents, i.e., words.

To avoid overfitting due to double usage of covariates, we fit an STM without including any covariates in the model estimation, thus reducing the model to a simple CTM. In a second, isolated step, we estimate the relationship between topic proportions and covariates. That is, we forgo the potential (small) gains of joint estimation of the STM in favor of a clear-cut two-step procedure which avoids overfitting.

As a first step, we fit the CTM analogously to the original STM (which includes topical prevalence variables), the only difference being that no document-level metadata is used in the estimation of the CTM. In line with the performance results in Roberts, Stewart, and Airoldi (2016), we observe a slightly higher held-out likelihood for the STM (-8.5478) than for the CTM (-8.5492) when holding out a random 50% of the words from a randomly chosen 10% of the documents. Moreover, we notice that the topics themselves (in terms of their top words) are almost identical to those of the STM, which is why we use the same topic labeling as in section 4.4. As for differences in topic proportions between the two models on a document level, we consider the average topic proportion deviation per document, $\frac{1}{K}\sum_{k=1}^{K}|\theta_{d,k}(STM) - \theta_{d,k}(CTM)|$. The resulting average difference between topic proportions per topic, averaged across all documents, amounts to 1.61%; that is, for an average document, the absolute difference in the proportion of each topic is less than 2%, which is rather moderate. These differences in topic proportions between STM and CTM further cancel each other out across documents: when comparing *global* topic proportions (i.e., topic proportions simply averaged across all documents), the results are very similar, with the average difference per topic only amounting to 0.23%. Altogether, topic proportions seem to be affected by the topical prevalence covariates to a small degree on an individual document level, and this effect almost disappeares entirely if we consider corpus-wide topic proportions.

[Update the below according to section 4.4 updates in graphs]

In the second step, we consider the relationship between topic proportions and covariates for the CTM and compare the resulting relationships with those of the originally fitted STM (which contains prevalence covariates). For comparability, we use the same methodology as in section 4.4: applying the method of composition with a quasibinomial regression of individual topic proportions on covariates. The only difference is that prevalence covariates were not included in the model used to generate topic proportions. Consequently, sampling all (unnormalized) topic proportions jointly via the logistic normal distribution (as in Figure 4.XXX) is not applicable here, as no Γ-vector is being estimated at all.

In the figures below, we visualize the CTM topic proportions of topics 4 and 6 in relation with continuous covariate values and across parties and compare the results to those of the STM (Figure XXX). As for the relationship between continuous covariates and topic proportions, the results for STM and CTM are very similar: for both topic 4 and topic 6, the trends across the respective covariate range are almost identical for the two models, while the scale differs slightly (with scale differences hardly exceeding 2%). Turning to the categorical variables, in particular party, the conclusion is very similar: minor scale differences, very similar patterns. TBD!!! We observe the same similarities and differences between the two models if we use beta regression instead of quasibinomial regression within the method of composition, corroborating our results (see appendix XXX).

[Graph quasi_t134_cont_ctm from ../plots/4_5]

All in all, the relationships between topical prevalence covariables and topic proportions are very similar to the STM when instead using a clean 2-step estimation procedure where no covariate information was used in the model estimation. This indicates that the problem of double usage of covariate information in the STM, potentially leading to overfitting, is not overly severe.

To be addressed: * metadata for test data is entirely meaningless, does not affect topic proportions at all (given words!!!) * manipulating covariate values neither * train/test: once words are given, covariates do not have any further impact: change party for MD (or exclude newData), show: no effect on predicted topic proportions (-> for future research: predict topic proportions based on document covariates only) * train/test: change formulation, since covariates do not really "generate" topic proportions; don't mention causal inference * train/test: main point of section: validate model: do topics (and their proportions) make sense? * additional paragraph for two-step procedure (+ 1 top graph)

Roberts, Margaret E., Brandon M. Stewart, and Edoardo M. Airoldi. 2016. "A Model of Text for Experimentation in the Social Sciences." *Journal of the American Statistical Association* 111 (515): 988–1003.