4_5

Patrick Schulze, Simon Wiegrebe

June 2020

Contents

1 Results

1.1 Hyperparameter Search and Model Fitting

1.2 Labelling

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

As briefly mentioned in section 2 already, a point of concern when using the STM is the double usage of (prevalence) 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.

[Insert some of the initial part of section 4.7 on train-test-split]

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. 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 figure below, we visualize the CTM topic proportion of three topic across parties. Comparing the results to the corresponding ones of the STM (Figure 4.XXX), we see a similar pattern for all but the green party and a general shift in scale across all parties when considering the "green/climate" topic. For the "social/housing" topic the average proportions by party show a higher degree of resemblance across models. Finally, for the right/national topic, there is a substantial difference between the two models in terms of the topic proportion of the AfD party. 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).

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

For continuous covariates (see appendix), the comparison of STM and CTM yields very similar results to those of the per-party comparison: for the "green/climate" topic, the trends are not that similar anymore and the scale differs (as above, with higher overall topic proportions according to the CTM). For the "social/housing" topic, on the other hand, the relationship between the covariates and topic proportion does not differ very much, neither in trend nor in scale.

All in all, the comparison of

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.