Theoretical Considerations

How the STM works

The Structural Topic Model (STM) is a topic model which extends classical topic models such as Latent Dirichlet Allocation (LDA) by incorporating information of covariates. Topic models can be used to infer topics from a large text corpus grouped into documents. In topic modelling it is assumed that this corpus is generated from a small number of distributions over words, the topics. The proportions of these topics are document-specific. In contrast to simpler topic models such as LDA, the STM relates topic proportions to document-level covariates. Furthermore, each distribution over words, i.e. each topic, can vary for different documents dependent on the covariate values of this document.

The detailed mechanism underlying the STM can be illustrated using its graphical model representation (see Figure 1). As outlined above, for each document indexed by $d \in \{1, \dots, D\}$ there exists a $K-1$-dimensional vector θ_d of topic proportions. Topic proportions are assumed to depend on P document-specific so-called topical prevalence covariates $X \in \mathbb{R}^{D \times P}$, by following a logistic normal distribution with mean $X_d\Gamma$, where $\Gamma = [\gamma_1|\dots|\gamma_k]$, and covariance Σ. Each of the N_d words in document d is subsequently assigned to one of the K topics dependent on the topic proportions θ_d; this per-word topic assignment is captured by the latent variable $z_{d,n} \sim \text{Multinomial}_K(\theta_d)$, where $n \in \{1, \dots, N_d\}$ denotes the word index. As stated, the distribution over words that characterizes a topic can vary for each document dependent on document-specific covariates $Y \in \mathbb{R}^{D \times A}$, the so-called topical content covariates. A word $w_{d,n}$ is then the result of the assigned topic, expressed by $z_{d,n}$, the content covariates Y_d, and their interactions. More precisely, this last step is intuitively best understood as a multinomial logistic regression of the words on the latter variables. A word $w_{d,n}$ then ultimately follows a multinomial distribution with probabilities $\beta_{d,n} := \beta(z_{d,n}, Y_d) \in \mathbb{R}^V$, i.e. $w_{d,n} \sim \text{Multinomial}_V(\beta_{d,n})$, where V denotes the total number of distinct words in the corpus; for details on the exact specification of $\beta_{d,n}$ see p. 991, Roberts, Stewart, and Airoldi (2016). Thus the occurence of a word (which is equivalent to being drawn from a corresponding multinomial distribution) depends on the topic assignment as well as on the topical content covariates, where the topic assignment itself is a function of the topical prevalence covariates.

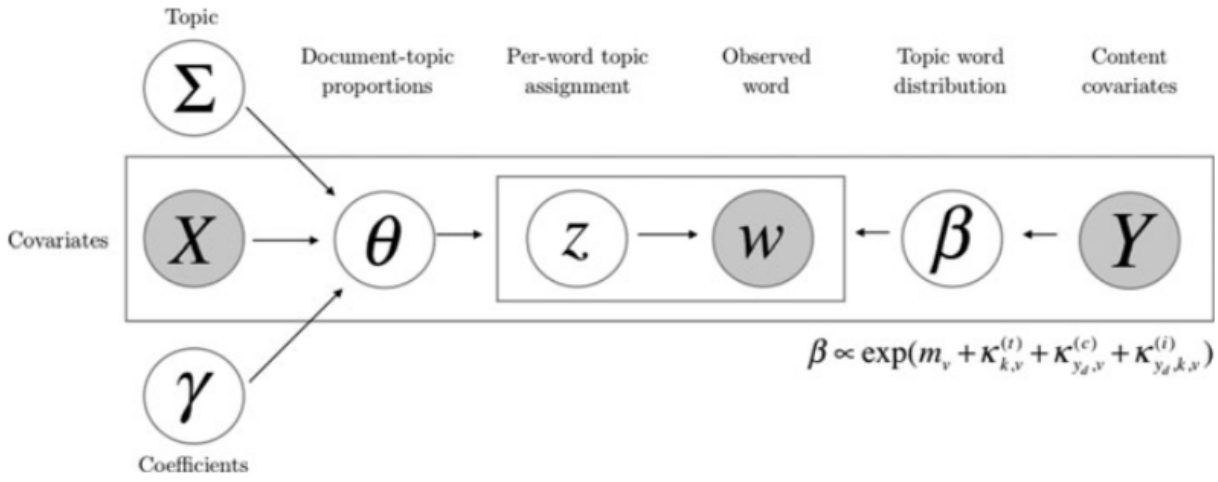

Figure 1: Graphical model representation of the STM

Scope of the STM

Topic models are unsupervised learning methods, since the true topics from which the text was generated are not known. Thus, traditionally topic models have been used as an exploratory tool providing a concise summary of topics, where it is hoped that the posterior induces a good decomposition of the corpus. Topic models have also been used for tasks such as collaborative filtering and classification (see e.g. Blei, Ng, and Jordan (2003)). In particular, they can be used as a dimensionality reducing method in semi-supervised learning methods. Such a process can in general be described as a two-stage approach, where in the first stage topic proportions and content are learned, and in the second stage a supervised method such as regression takes this learned representation as input.

The fundamental idea of STMs is to combine these two steps: Topics and their relation to covariates are jointly estimated. For instance, the estimated effect of topical prevalence covariates X_d on topic proportions is reflected in the estimate of Γ. However, since the topic proportions θ_d are random variables, it is a better approach to incorporate the uncertainty of θ_d, accesible through the estimated approximation of the posterior $p(\theta_d|\Gamma, \Sigma, X)$, when determinig the effect of covariates on topic proportions. This is achieved by what is called the "method of composition" in social sciences: By sampling from the approximate posterior for θ and subsequently regressing these topic proportions on X it is possible to integrate out the topic proportions (since these are latent variables!) and obtain an i.i.d. sample from the marginal posterior of the regression coefficients for the topical prevalence covariates.

A problem we see with this approach is, however, that the same covariates and in general the same data used to infer the topical structure are subseqeuntly used to determine effetcs of the former on the latter (or vice versa). This problem has recently also been adressed by Egami et al. (2018). In practice, in case of the regression coefficients for the topical prevalence covariates (obtained using the method of composition as outlined above), due to the regularizing priors for Γ we have found that the prevalence covariates have almost no influence on the estimated topic proportions. Thus the regression coefficients (with the topic proportions as the dependent variable) should not be largely affected by this problem. However, the question then appears why the covariate variables have been used to obtain the topical structure in the first place. In an empirical evaluation Roberts, Stewart, and Airoldi (2016) showed that the STM consistently outperformed other topic models such as LDA, when comparing the respective heldout likelihoods in different settings. This indicates that the STM performs better at predicting the topical structure by incorporating covariates, regardless of the concrete specification of these covariates.

Nevertheless, it should in each case be investigated whether the relationship of variables implied by the STM is valid. For instance, we have split our data into training and test sets and found that the topical structure predicted on the test set differs starkly from the structure on the training set. This could of course be caused by a misspecification of the topical prevalence and content variables. However, since the topical prevalence covariates have almost no influence on the estimated topic proportions on the training set due to the regularizing priors (and e.g. likewise on the heldout likelihood that can be used for validation), it is practically impossible to validate a good prevalence specification.

Posterior

The posterior given on p. 992, Roberts, Stewart, and Airoldi (2016), can be derived as follows:

$$
\begin{aligned}
p(\eta, z, \kappa, \Gamma, \Sigma | w, X, Y) &\propto \underbrace{p(w|\eta, z, \kappa, \Gamma, \Sigma, X, Y)}_{=p(w|z,\kappa,Y)} p(\eta, z, \kappa, \Gamma, \Sigma | X, Y) \\
&\propto p(w|z, \kappa, Y) p(z|\eta) p(\eta|\Gamma, \Sigma, X) \prod p(\kappa) \prod p(\Gamma) p(\Sigma) \\
&\propto \Big\{ \prod_{d=1}^{D} p(\eta_d|\Gamma, \Sigma, X_d) \Big(\prod_{n=1}^{N} p(w_n|\beta_{d,n}) p(z_{d,n}|\theta_d) \Big) \Big\} \prod p(\kappa) \prod p(\Gamma) p(\Sigma) \\
&\propto \Big\{ \prod_{d=1}^{D} \text{Normal}(\eta_d | X_d \Gamma, \Sigma) \Big(\prod_{n=1}^{N} \text{Multinomial}(z_{n,d}|\theta_d) \\
&\quad \times \text{Multinomial}(w_n|\beta_{d,n}) \Big) \Big\} \times \prod p(\kappa) \prod p(\Gamma) p(\Sigma),
\end{aligned}
$$

where $\beta_{d,n} := \beta(z_{d,n}, Y_d) \in \mathbb{R}^V$ with entries $\beta_{d,k,\nu} \propto \exp(m_\nu + \kappa^{(t)}_{k,\nu} + \kappa^{(c)}_{y_d,\nu} + \kappa^{(i)}_{y_d,k,\nu})$, $\nu \in \{1, \ldots, V\}$, and $\theta_d := \text{softmax}(\eta_d)$.

Literature

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3 (Jan): 993–1022.

Egami, Naoki, Christian J Fong, Justin Grimmer, Margaret E Roberts, and Brandon M Stewart. 2018. "How to Make Causal Inferences Using Texts." *arXiv Preprint arXiv:1802.02163*.

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). Taylor & Francis: 988–1003.