

# 1 Theoretical Framework

## 1.1 Topic Modeling - Overview

Topic models seek to discover latent thematic clusters, called topics, within a collection of discrete data, usually text; therefore, topic modeling can be regarded as dimensionality reduction technique. Furthermore, since both the number and content of topics is unknown beforehand (and can never be truly verified), topic modeling is an instance of unsupervised learning. Information retrieval (IR) research generally proposes the reduction of text documents to vectors of real numbers, each number representing (modified) counts of "words" or "terms". An instance of this proposed methodology is the *tf-idf* scheme by Salton and McGill (1983), which for a collection of documents returns a term-by-document matrix where each row corresponds to a document in the corpus and the columns contain the respective *tf-idf* term count. Since only words in a vocabulary of fixed length  $V$  are considered, documents of unrestricted length are being reduced to vectors of a fixed length  $V$ . To further reduce dimensionality, the *latent semantic indexing* (LSI) by Deerwester et al. (1990) applied singular value decomposition (SVD) to the *tf-idf* document-term matrix. However, as Blei et al. (2003) put it, the idea should be to develop a generative probabilistic model of text, in order to estimate to which extent the LSI methodology can align data with the generative text model; yet, given such a model, "it is not clear why one should adopt the LSI methodology - one can attempt to proceed more directly, fitting the model to data using maximum likelihood or Bayesian methods" (p. 994). Picking up this shortcoming of LSI, Hofmann (1999) introduced the *probabilistic LSI* (pLSI) model. This generative data model allows for individual words to be sampled from a mixture model: they are drawn from a multinomial distribution, with latent random variables determining the mixture proportions, which in turn can be viewed as topics. However, the pLSI can only be regarded as partly probabilistic text model, since the mixing components themselves are fixed on a document level, thus lacking a probabilistic generating process.

In their *Latent Dirichlet Allocation* (LDA) model, Blei et al. (2003) included the generation of topic proportions into the generative probabilistic model, the resulting 3-level hierarchical Bayesian mixture model marking the starting point of modern topic modeling. In order to present the main idea of LDA, we first introduce some notation and terminology that we will use throughout the remainder of this paper.

- A *word* (also called *term*) is the smallest unit of discrete text data. Words are elements of a vocabulary of length  $V$  and can thus be indexed by  $\{1, \dots, V\}$ . Mathematically, the  $v$ -th word in the vocabulary can be represented as a vector of length  $V$ , whose  $v$ -th component equals one, with all other components equalling zero. We will sometimes refer to the  $v$ -th word of the vocabulary simply as  $v$ . Apart from document-level

covariates, words are the only random variables within the topic model that we actually observe, the rest being latent.

- A *document*  $d \in \{1, \dots, D\}$  is a sequence of words of length  $N_d$ . For a given document  $d$ , we denote its words by  $d = (w_{d,1}, \dots, w_{d,N_d})$ . Consequently, the  $n$ -th word of document  $d$  is denoted by  $w_{d,n}$ .
- A *corpus* is a collection (or set) of  $D$  documents. Therefore,  $d \in \{1, \dots, D\}$  means that our corpus contains  $D$  documents.
- A *topic* is a latent thematic cluster within a text corpus. The idea is that any collection of documents is made up of  $K$  such topics, where the number of topics  $K$  is an (unknown) hyperparameter which needs to be determined ex ante (see Section 4.1 for hyperparameter determination in our specific use case). We will refer to topics simply by the actual *topic index* (or *topic number*)  $k \in \{1, \dots, K\}$ .
- A *topic-word distribution*  $\beta$  is a probability distribution over words, i.e., over the vocabulary. This is what actually characterizes a topic. For a model containing  $K$  topics (and no topical content variable, see section 2.2 below), topic-word distributions do not vary across documents and uniquely characterize a topic: we denote the word distribution corresponding to the  $k$ -th topic by  $\beta_k$  and the matrix whose  $k$ -th column is topic  $\beta_k$  by  $B := \beta_{1:K} = [\beta_1 | \dots | \beta_K]$ . Each vector  $\beta_k$  thus has length  $V$ , while  $B$  is a  $V \times K$ -matrix. Therefore,  $k$  refers to the latent thematic cluster with topic index  $k$  in general, and  $\beta_k$  refers to the underlying word distribution in particular.
- A *topic assignment*  $z_{d,n}$  is the assignment of the  $n$ -th word of document  $d$  to a specific topic  $k \in \{1, \dots, K\}$  (i.e., to the corresponding word distribution  $\beta_k$ ). Therefore,  $z_{d,n}$  is simply a vector of length  $K$  whose  $k$ -th entry equals one and all other entries equal zero. This way, we can represent the word distribution corresponding to the  $n$ -th word in document  $d$  as  $\beta_{d,n} := Bz_{d,n}$  (again, for a model without topical content variable).
- For a given document  $d$ , the corresponding *topic proportions*, denoted by  $\theta_d$ , are the proportions of the document's terms assigned to each of the topics  $k \in \{1, \dots, K\}$ . Topic proportions vary across documents. Since for each document  $d$  the proportions of all  $K$  topics must add up to one ( $\sum_{k=1}^K \theta_{d,k} = 1, \forall d \in \{1, \dots, D\}$ ), topic proportions represent probabilities.
- The *bag-of-words* assumption is an assumption used in all (probabilistic) text models referenced in this paper, including LSI and pLSI, and states that only words themselves (and their counts) carry meaning, while word order or grammar do not. Statistically, this is equivalent to assuming that words within a document are *exchangeable* (Aldous, 1985).

As mentioned above, LDA is the first generative probabilistic model of an entire text corpus. (Recall that pLSI is only probabilistic for a fixed document.) Now, LDA can be

neatly described by the following 2-step procedure, given the hyperparameter (number of topics)  $K$ : For each document  $d \in \{1, \dots, D\}$ :

- 1) Draw topic proportions  $\theta_d \sim \text{Dir}_K(\alpha)$ .
- 2) For each word  $n \in \{1, \dots, N_d\}$ :
  - a) Draw a topic assignment  $z_{d,n} \sim \text{Multinomial}_K(\theta_d)$ .
  - b) Draw a word  $w_{d,n} \sim \text{Multinomial}_V(\beta_{d,n})$ .

Thus, topic proportions are drawn from a Dirichlet distribution with  $K$ -dimensional hyperparameter vector  $\alpha$ , with all components  $\alpha_k > 0$ ; this vector is estimated from the data. This means that for each document  $d \in \{1, \dots, D\}$ , the corresponding topic proportions  $\theta_d$  represent a  $K$ -dimensional vector which can take on values on the  $(K - 1)$ -simplex:  $\theta_{d,k} \geq 0, \sum_{k=1}^K \theta_{d,k} = 1$ . Also note that the Dirichlet distribution is the conjugate prior of the multinomial distribution, which greatly facilitates estimation (see section 2.3 on variational inference below). Put simply, for each document LDA first generates topic proportions, which are then used as weights for topic assignment. Finally, each "word spot" - which now already has a topic assigned to it - is filled with a draw from the topic-specific word distribution. These topic-specific word distributions  $\beta_k$  need to be estimated from data. Note that LDA is a very simple, restrictive model in (at least) three ways:

- (i) By using the Dirichlet distribution to generate topic proportions, potential correlations between topics cannot be captured due to the neutrality of the Dirichlet distribution.<sup>1</sup>. As a consequence, the occurrence of one topic within a document is not correlated with the occurrence of another topic (Blei et al., 2007). This is a restrictive simplification, as topics such as "sports" and "health" are much more likely to co-occur within a document than, say, "sports" and "war".
- (ii) Second, while topic proportions vary stochastically across documents, they do so given a single, global hyperparameter vector  $\alpha$ , essentially implying that topic proportions are generated based merely on word counts (occurrences and co-occurrences). This is another unrealistic and limiting simplification, since researchers usually possess further document-specific information indicative of the topics addressed within the individual documents.
- (iii) Third, the topic-specific word distributions  $\beta_k$  are estimated identically for all documents, by construction. Similarly to the second restriction, this prevents researchers

---

<sup>1</sup>Due to the constraint  $\sum_{k=1}^K \theta_k = 1$ , there is clearly some degree of dependence between topic proportions. However, the dependence is minimal, as the Dirichlet distribution is characterized by complete neutrality: the components  $\theta_1/(1 - S_0), \theta_2/(1 - S_1), \dots, \theta_K/(1 - S_{K-1})$  are mutually independent, where  $S_0 := 0$  and  $S_k = \sum_{i=1}^k \theta_i, k \in \{1, \dots, K\}$ . Stated differently, for each component  $\theta_k, k \in \{1, \dots, K\}$ , it holds that  $\theta_k/(1 - S_{k-1})$  is independent of the vector constructed by weighting all *remaining* components by their total proportion (James et al., 1980)

from using (document-level) information which might potentially influence the weighting of specific words within a topic.

Due to its simplicity and the resulting restrictions, the LDA has been used as building block for more advanced (and usually more specified) generative topic models. One model that builds on LDA, addressing some of its shortcomings, is the *Correlated Topic Model* (CTM) by Blei et al. (2007). Specifically, the CTM addresses the first one of the above-mentioned restrictions: the inability to cope with inter-topic correlations. The model no longer uses a Dirichlet distribution to sample topic proportions; instead, a logistic normal distribution is employed, which can capture correlations between topics due to the incorporated covariance structure between its components (Atchison and Shen, 1980). The resulting generative process for the CTM can be stated as follows:

For each document  $d \in \{1, \dots, D\}$ :

- 1) Draw unnormalized topic proportions  $\eta_d \sim N_{K-1}(\mu, \Sigma)$ , with  $\eta_{d,k} := 0$  for model identifiability.
- 2) Normalize  $\eta_d$  by mapping it to the simplex:  $\theta_{d,k} = \frac{\exp(\eta_{d,k})}{\sum_{j=1}^K \exp(\eta_{d,j})}, \forall k \in \{1, \dots, K\}$ .
- 3) For each word  $n \in \{1, \dots, N_d\}$ :
  - a) Draw a topic assignment  $z_{d,n} \sim \text{Multinomial}_K(\theta_d)$ .
  - b) Draw a word  $w_{d,n} \sim \text{Multinomial}_V(\beta_{d,n})$ .

Steps 1 and 2 constitute the sampling from a logistic normal distribution: a  $K$ -dimensional vector  $\eta_d$  is drawn from a multivariate normal distribution and subsequently transformed to a vector of proportions (or probabilities) by applying the *softmax* function to each of its elements. The number of topics  $K$  is again a hyperparameters which must be determined ex ante. As in LDA, the parameters of the normal distribution in step 1,  $\mu \in \mathbb{R}^{K-1}$  and  $\Sigma \in \mathbb{R}^{(K-1) \times (K-1)}$ , as well as the topic-specific word distributions  $\beta_k$  need to be estimated from the data. As mentioned above, this process now allows for inter-topic correlation. Yet this comes at a cost: unlike the Dirichlet distribution, the logistic normal distribution is no longer conjugate to the multinomial distribution. As explained in more detail in section 2.3 below, this renders standard variational inference algorithms inapplicable, since these rely on conjugacy and the implied closed-form solutions. However, using the Laplace variational inference developed by Wang and Blei (2013), which is a generic method for variational inference when dealing with nonconjugate models, solves the inference problem for the CTM.

As for the inability to integrate covariate information into the determination of topic proportions, Mimno et al. (2011) were the first to model topic proportions as a function of *observable* document-level metadata. Specifically, their *Dirichlet-Multinomial Regression* (DMR) model still samples topic proportions  $\theta_d$  from a Dirichlet distribution (thus, not allowing for inter-topic correlations), yet unlike in LDA, the Dirichlet prior  $\alpha_d$  is no longer

global but topic-specific. This topic prior  $\alpha_d$ , in turn, is log-linear in the document-level features  $\mathbf{x}_d$  and the (topic-specific) priors for the coefficients of these features,  $\lambda_t$ , have a normal prior. With coefficients being updated through numerical optimization as part of the EM algorithm used for training, the DMR model thus actively uses document features to model topic proportions.

Finally, the third restrictiveness of LDA, the inflexibility of the topic-word distributions  $\beta_k$  when document-level metadata is available, is addressed by Eisenstein et al. (2011) in their *Sparse Additive General* model (SAGE). The authors propose to start off with a background word distribution  $m$  containing log frequencies and to model additive deviations from this baseline for each class. The idea behind SAGE can be used to model differences in topic-word distributions according to the category of some document-level covariate.

Based on the foundational LDA as well as its extensions, Roberts et al. (2013) developed the *Structural Topic Model* (STM), which combines the improvements over the original LDA discussed in this section. Due to its flexibility regarding the incorporation of document-level information, we choose the STM for our specific use case, a text-based analysis of German political entities (TBD, depends on final title of paper). Therefore, we discuss the model in greater detail in section 2.2 below.

## 1.2 The Structural Topic Model

### Overview

The STM addresses the three main shortcomings of the LDA, as discussed in the previous section. In this subsection, we explain the corresponding modifications with respect to LDA and present the generative process of the STM.

- (i) To allow for correlation among topics, the STM uses a logistic normal distribution to sample topic proportions. In fact, if no document-level metadata is fed into the STM, it simply reduces to the CTM.
- (ii) The STM allows for the incorporation and use of document-level metadata when determining topic proportions. Similar to the DMR, topic proportions  $(\theta_1, \dots, \theta_D)^T$  are assumed to depend on  $P$  document-level *topical prevalence variables* (such as the author’s name, her political party or her popularity on Twitter), yet now by following a multivariate logistic normal distribution with mean vector  $X_d\Gamma$ , where  $X \in \mathbb{R}^{D \times P}$  and  $\Gamma = [\gamma_1 | \dots | \gamma_k]$ , and covariance matrix  $\Sigma$ . This way, the model accounts for the fact that document-level covariates might influence how much (that is, which percentage of the total number of words) the corresponding documents attribute to the different topics.
- (iii) Within the STM, document-level covariate information can also be used to fine-tune the topic-word distributions  $\beta_k$ , the methodology being similar to the one in the SAGE

model. In particular, the STM allows for specifying a single categorical document-level *topical content variable*  $Y$  with  $A$  levels:  $Y_d \in \{1, \dots, A\}, \forall d \in \{1, \dots, D\}$  (Roberts et al., 2019). Consequently, each topic  $k \in \{1, \dots, K\}$  is now associated with a total of  $A$  topic-word distributions  $\beta_{k,a}, a \in \{1, \dots, A\}$  instead of a single one,  $\beta_k$ . For a given document  $d$ , this means the  $K$  topic-word distributions  $\beta_k$  are determined according to the level  $a$  assumed by  $Y_d$  and are identical across all documents with  $Y_d = a$  (Roberts et al., 2016). This way, for a given document  $d$ , document-level metadata can not only impact the weighting of topics  $\theta_d$ , but also the weighting over words for each topic  $\beta_k$ . Note that for a given topic  $k$ , the word distributions  $\beta_{k,a}$  are similar to each other for all values of  $a$ ; that is, the content variable  $Y$  is really an  $A$ -level refinement of  $\beta_k$  and does *not* affect the number of topics  $K$ .

The generative process of the STM can be stated as follows (Roberts et al., 2016):

For each document  $d \in \{1, \dots, D\}$ :

- 1) Draw unnormalized topic proportions  $\eta_d \sim N_{K-1}(X_d\Gamma, \Sigma)$ , with  $\eta_{d,k}$  set to zero for model identifiability.
- 2) Normalize  $\eta_d$  by mapping it to the simplex:  $\theta_{d,k} = \frac{\exp(\eta_{d,k})}{\sum_{j=1}^K \exp(\eta_{d,j})}, \forall k \in \{1, \dots, K\}$ .
- 3) For each word  $n \in \{1, \dots, N_d\}$ :
  - a) Draw a topic assignment  $z_{d,n} \sim \text{Multinomial}_K(\theta_d)$ .
  - b) If no topical content variable has been specified, simply draw a word  $w_{d,n} \sim \text{Multinomial}_V(\beta_{d,n})$ . Otherwise, first determine the document-specific word distributions  $\beta_{k,a}$  based on the level  $a$  taken on by  $Y_d$ , for all topics  $k \in \{1, \dots, K\}$ :  $B_a := [\beta_{1,a} \dots \beta_{K,a}]$ ; next, analogously define  $\beta_{d,n} := B_a z_{d,n}$ ; finally, draw a word  $w_{d,n} \sim \text{Multinomial}_V(\beta_{d,n})$ .

This means that unnormalized topic proportions are sampled from a normal distribution with mean  $\Gamma = [\gamma_1 \dots \gamma_K]$  and covariance  $\Sigma$ .  $\Gamma$  is the vector of coefficients corresponding to the topical prevalence covariates contained in  $X$ , with prior distributions  $\gamma_k \sim N_p(0, \sigma_k^2 I_p)$ . The unnormalized topic proportions  $\eta_d$  are then "sent through the softmax function" to yield normalized topic proportions  $\theta_d$ , which in turn are used as weights for the subsequent topic assignment  $z_{d,n}$ . Finally, each word is sampled from the corresponding multinomial word probability distribution (over the vocabulary of length  $V$ ), which depends on topic assignment  $z_{d,n}$  and, for models containing a topical content variable, on its level  $a$ . In line with SAGE methodology, the topic-word distributions are modelled as deviations in log-frequency from a baseline vocabulary. (See Roberts et al. (2016), p. 991 for further details.)  $K$  and  $\sigma_k^2$  are hyperparameters. The graphical model representation in Figure 1 below visualizes the generative process described.

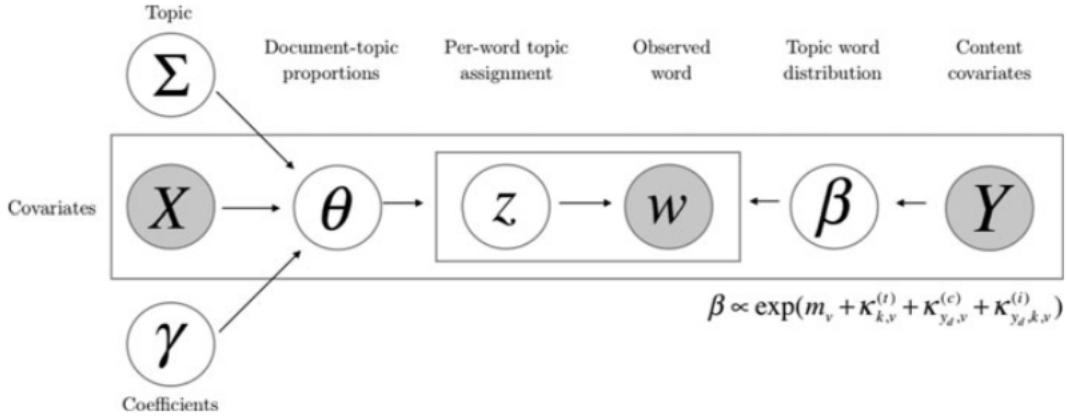


Figure 1: Graphical model representation of the STM (from Roberts et al. (2016), p. 990).

## Scope

Topic models are unsupervised learning methods, since the true topics from which the text was generated are not known. Thus, topic models have been traditionally used as an exploratory tool providing a concise summary of topics, with the posterior ideally inducing a good decomposition of the corpus. Topic models have also been used for tasks such as collaborative filtering and classification (see e.g. (Blei et al., 2003)). In particular, they can be used as dimensionality reduction 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 the STM is to combine these two steps: Topics and their association with covariates are estimated jointly. For instance, the estimated effect of topical prevalence covariates  $X_d$  on topic proportions  $\theta_d$  is reflected in the estimate of  $\Gamma$ . However, since the topic proportions are latent random variables, it is preferable to incorporate the uncertainty of  $\theta_d$ , accessible through the estimated approximation of the posterior  $p(\theta_d|\Gamma, \Sigma, X)$ , when determining 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  $\theta$  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 (see Section 4.4).

A problem we see with this approach, however, is that the same covariates - and in general the same data - used to infer the topical structure are subsequently used to determine effects of the former on the latter (or vice versa). This problem has recently also been addressed by Egami et al. (2018). In our specific case, we find that the prevalence covariates do

not have much impact on the estimated topic proportions due to the regularizing priors for  $\Gamma$  (see Section 4.6). Thus, the regression coefficients (with topic proportions as the dependent variable) should not be largely affected by this problem of double usage. However, this begs the question why document-level covariates are being used to obtain the topical structure in the first place. In an empirical evaluation, Roberts et al. (2016) showed that the STM consistently outperforms 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 their concrete specification.

Nevertheless, in each case it should be investigated whether the relationship of variables implied by the STM is valid. In line with Egami et al. (2018), we address this issue in section 4.7, where we split our data into a training and a test set. Similar topical structures on both datasets (as we find in our case) indicate that misspecification of topical prevalence or content variables is not a concern. 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 likewise on the heldout likelihood that can be used for validation), it is practically impossible to validate a good prevalence specification.

## Posterior Distribution

In this subsection, we briefly derive the posterior distribution of the STM (up to proportionality), as stated on p. 992 of Roberts et al. (2016). Recall that only words  $w$ , prevalence covariates  $X$ , and the content covariate  $Y$  are observable, while all other variables - unnormalized topic proportions  $\eta$ , topic assignments  $z$ , topic-word distribution deviations  $\kappa$ , prevalence coefficients  $\Gamma$ , and unnormalized topic proportion variance  $\Sigma$  - are latent.

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

where  $\beta_{d,n} \in \mathbb{R}^V$  is the topic-word distribution for word  $n$  in document  $d$ , which has been assigned to topic  $k$  through  $z_{d,n}$ . The topic-word distribution vectors  $\beta_{k,a}$  have entries  $\beta_{k,a,v} \propto \exp\left(m_v + \kappa_{k,v}^{(t)} + \kappa_{a,v}^{(c)} + \kappa_{k,a,v}^{(i)}\right)$ ,  $v \in \{1, \dots, V\}$ , where  $\kappa_{k,v}^{(t)}$ ,  $\kappa_{a,v}^{(c)}$ , and  $\kappa_{k,a,v}^{(i)}$  are the



log-transformed rate deviations of word  $v$  for topic  $k$ , for content variable level  $a$ , and for the interaction of  $k$  and  $a$ , respectively.

### 1.3 Inference and Parameter Estimation

In this section, we describe how inference and parameter estimation for topic models, in particular for the STM, are performed. Inference is done using variational inference, and a variational Expectation-Maximization (EM) algorithm is used for empirical parameter estimation. As a detailed discussion of the underlying workings is outside the scope of this paper, we refer the reader to the appendix and the referenced papers.

Since the STM, as well as all models it builds on, are (hierarchical) Bayesian models, the central challenge we face is the exact determination of the posterior distribution. Recall that in the section above, we derived the posterior *up to proportionality*, neglecting the division by marginal distributions. The exact posterior distribution is intractable to compute due to the (high-dimensional) marginal distributions in the denominator, which is why exact inference is infeasible and variational inference is used instead. Generally, for a model with latent variables  $\theta$  and  $z$  and observable data  $x$ , variational inference involves approximating the posterior  $p(\theta, z|x)$  by postulating a simple distribution family for the (joint) distribution of latent model variables  $\theta$  and  $z$  -  $q(\theta, z)$  - and subsequently determining the member of this family which minimizes the "distance" to the true posterior distribution, measured using the Kullback-Leibler (KL) divergence (Wang and Blei, 2013). The approximations of variational inference bring a great amount of flexibility, but come at the cost of some bias, since the approximative distribution family usually does not contain the true posterior.

In the appendix, we show that minimizing KL divergence between true posterior  $p$  and the approximating variational distribution  $q$  is equivalent to maximizing a lower bound on  $\log(p(x))$ , the log-likelihood of the observed data  $x$ . This lower bound is called *ELBO* and is defined as

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

whose second component,  $\mathbb{E}_q[\log(q(\theta, z))]$ , is the entropy of the approximate distribution  $q$ . To be precise, maximizing *ELBO* (or minimizing KL divergence) refers to finding the governing parameter of the approximating distribution  $q$  which maximizes *ELBO*.

The optimality conditions resulting from maximizing *ELBO* lead to the *coordinate ascent algorithm* for variational inference (Wang and Blei, 2013), which converges towards a local optimum (Bishop, 2006). However, this algorithm only works for *conditionally conjugate* models, such as the LDA: all nodes in this model - in particular, the Dirichlet distribution for drawing topic proportions, the multinomial distribution for assigning topics, and the

multinomial for eventually picking words - are conditionally conjugate. The STM, however, as well as the CTM before it, are non-conjugate models due to the logistic normal distribution used to sample topic proportions, which is why algorithm updates are not feasible and the algorithm is not (directly) applicable. As a remedy, Wang and Blei (2013) developed Laplace variational inference, which uses Laplace approximations within coordinate ascent algorithm updates and this way enables its use for the broader class of nonconjugate models, in particular for CTM and STM.

As stated above, the STM uses an Expectation-Maximization (EM) algorithm for empirical parameter estimation. In the E-step, the variational posterior distributions for topic proportions,  $q(\theta_d)$ , and for topic assignment,  $q(z_{d,n})$ , are updated using Laplace variational inference and coordinate ascent. In the M-step, the model parameters - specifically topical prevalence and content coefficients - are updated by maximizing *ELBO* with respect to them (Roberts et al., 2016).

## References

- David J Aldous. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII—1983*, pages 1–198. Springer, 1985.
- 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.
- David M Blei, John D Lafferty, et al. A correlated topic model of science. *The Annals of Applied Statistics*, 1(1):17–35, 2007.
- Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.
- 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.
- Jacob Eisenstein, Amr Ahmed, and Eric P Xing. Sparse additive generative models of text. 2011.

- Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, 1999.
- Ian R James, James E Mosimann, et al. A new characterization of the dirichlet distribution through neutrality. *The Annals of Statistics*, 8(1):183–189, 1980.
- David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the conference on empirical methods in natural language processing*, pages 262–272. Association for Computational Linguistics, 2011.
- 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.
- Margaret E. Roberts, Brandon M. Stewart, and Dustin Tingley. stm: An R package for structural topic models. *Journal of Statistical Software*, 91(2):1–40, 2019. doi: 10.18637/jss.v091.i02.
- Gerard Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.
- Chong Wang and David M Blei. Variational inference in nonconjugate models. *Journal of Machine Learning Research*, 14(Apr):1005–1031, 2013.