2_theoretical_framework

Patrick Schulze, Simon Wiegrebe

June 2020

Contents

1 Introduction

TBD

2 Theoretical Framework

2.1 Topic Modelling - Overview

Topic models seek to discover latent thematic clusters, called topics, within a collection of discrete data, usually text; therefore, topic modelling can be regarded as dimensionality reduction technique. 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 with rows being the documents in the corpus and the columns containing 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, Ng, and Jordan (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, Ng, and Jordan (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 modelling. 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 $\{1, ..., V\}$. 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* $d \in \{1, ..., D\}$ is a sequence of words of length N_d. For a given document d, we denote its words by $d = (w_{d,1}, ..., w_{d,N_d})$. 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, $d \in \{1, ..., D\}$ 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*) $k \in \{1, ..., K\}$.
- A *topic-word distribution* β 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.XXX), 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 β_k and the matrix whose k-th column is topic β_k by $B := \beta_{1:K} = [\beta_1|...|\beta_K]$. Each vector β_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 β_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 $k \in \{1, ..., K\}$ (i.e., to the corresponding word distribution β_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 θ_d, are the proportions of the document's terms assigned to each of the topics $k \in \{1, ..., K\}$. Topic proportions vary across documents. Since for each document d the proportions of all K topics must add up to one ($\sum_{k=1}^{K} \theta_{d,k} = 1, \forall d \in \{1, ..., D\}$), topic proportions are probabilities.
- The *bag-of-word* assumption is an assumption used in all (probabilistic) text models referenced in this paper, including LSI and pLSI, and states that only words themselves, as well as their counts carry meaning, while word order or grammar do not. Statistically, this is equivalent to assuming that words within a document are *exchangeable* (aldous1985exchangeability).

As mentioned above, LDA is the first generative probabilistic model of an entire text corpus. (Recall that pLSI is only probabilisitic 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 $d \in \{1, ..., D\}$:

1) Draw topic proportions $\theta_d \sim Dir_K(\alpha)$.
2) For each word $n \in \{1, ..., N_d\}$:
   a) Draw a topic assignment $z_{d,n} \sim Multinomial_K(\theta_d)$.
   b) Draw a word $w_{d,n} \sim Multinomial_V(\beta_{d,n})$.

Thus, topic proportions are drawn from a Dirichlet distribution with K-dimensional hyperparameter vector α, all components $\alpha_k > 0$. This means that for each document $d \in \{1, ..., D\}$, the corresponding topic proportions θ_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 (i.e., from the topic itself, according to our above definition). These topic-specific word distributions β_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. (Footnote: Due to the constraint $\sum_{k=1}^{K} \theta_k = 1$, there is clearly some degree of dependence between topic proportions. However, the degree of dependence is minimal, as the Dirichlet distribution is characterized by complete neutrality: the (components) $\theta_1/(1-S_0), \theta_2/(1-S_1), ..., \theta_K(1-S_{K-1})$ are mutually independent, where $S_0 := 0$ and $S_k = \sum_{i=1}^{k} \theta_k, k \in \{1, ..., K\}$. Stated differently, for each component $\theta_k, k \in \{1, ..., K\}$, 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, Mosimann, and others (1980)). As a consequence, the occurrence of one topic within a document is not correlated with the occurrence of another topic (Blei, Lafferty, and others (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 α, 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 actual topics (i.e., the topic-specific word distributions) β_k are estimated identically for all documents by construction. Similarly to the second restriction, this prevents researchers 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 used as building block for more advanced (and usually more specified) generative topic models.

One model that builds on LDA but addresses some of its shortcomings is the *Correlated Topic Model* (CTM) by Blei, Lafferty, and others (2007). Specifically, the CTM addresses the first one of the abovementioned restrictions: the inability to cope with inter-topic correlations. Specifically, the CTM 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 $d \in \{1, ..., D\}$:

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

Steps 1 and 2 constitute the sampling from a logistic normal distribution: a K-dimensional vector η_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 as well as the parameters of the normal distribution in step 1, $\mu \in \mathbb{R}^{K-1}$ and $\Sigma \in \mathbb{R}^{(K-1)\times(K-1)}$, are hyperparameters which must be determined ex ante. As for LDA, the topic-specific word distributions β_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, as these rely on conjugacy and the implied closed-form solutions. However, using 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.

Addressing the second restriction: DMR (add to bibliography, search OPAC)

Third restriction: SAGE Eisenstein, Ahmed, and Xing (2011)

Based on the foundational LDA as well as its extensions/modifications, 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.

2.2 The Structural Topic Model

2.2.1 Overview

- explicitly write out the algorithm (step-wise)
- review notation (in particular, β_k)

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. To be precise, topic proportions $(\theta_1, ..., \theta_D)^T$ 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) 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 Σ. 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. (TBD: connection to DMR!!!)

iii) Within the STM, document-level covariate information can also be used to fine-tune the topic-word distributions β_k. In particular, the STM allows for specifying a single categorical document-level *topical content variable* Y with A levels: $Y_d \in \{1, ..., A\}, \forall d \in \{1, ..., D\}$ (Roberts, Stewart, and Tingley (2019)). Consequently, each topic $k \in \{1, ..., K\}$ is now associated with a total of A topic-word distributions $\beta_{k,a}, a \in \{1, ..., A\}$ instead of a single one, β_k. For a given document d, this means the K topic-word distributions β_k are determined according to the level a assumed by Y_d and are identical across all documents with $Y_d = a$ (Roberts, Stewart, and Airoldi (2016)). This way, for a given document d, document-level metadata can not only impact the weighting of topics θ_d, but also the weighting over words for each topic β_k. (TBD: connection to SAGE!!!)

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.

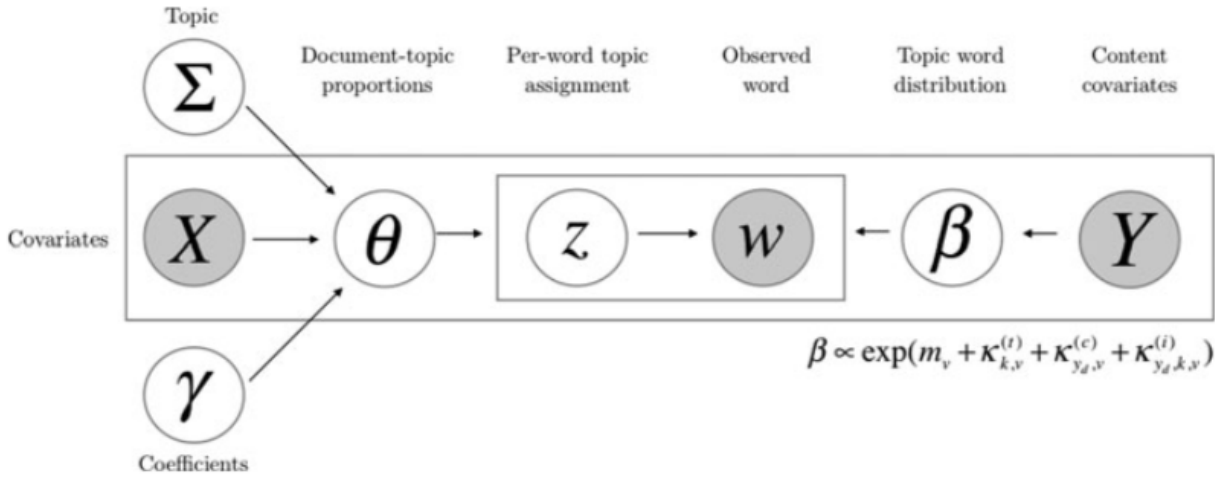
Each of the N_d words in document d is subsequently assigned to one of the K topics, depending on the topic proportions θ_d; this per-word topic assignment is captured by the latent variable $z_{d,n} \sim \text{Multinomial}_K(\theta_d)$, where $n \in \{1, \dots, N_d\}$ denotes the word index.

As stated, the distribution over words that characterizes a topic can vary across documents according to the value of a so-called topical content variable Y, giving rise to $\mathbf{Y} \in \mathbb{R}^{D\times A}$, where A denotes the number of levels of the topical content variable.

Then, a word $w_{d,n}$ is the result of the assigned topic, expressed by $z_{d,n}$, (the level of) content variable 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.

It is important to recall that $z_{d,n}$ singles out the topic k for word $w_{d,n}$, so that $\beta_{d,n}$ is already topic-specific, even though it does not explicitly have a k-subscript; for details on the exact specification of $\beta_{d,n}$ see Roberts, Stewart, and Airoldi (2016), p. 991. 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 variable, where the topic assignment itself is a function of the topical prevalence covariates.



Graphical model representation of the STM (from Roberts, Stewart, and Airoldi (2016), p. 990)

2.2.2 Scope

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 subsequently 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.

2.2.3 Posterior Distribution

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 $\beta_{d,k,\nu} \propto \exp(m_\nu + \kappa^{(t)}_{k,\nu} + \kappa^{(c)}_{y_d,\nu} + \kappa^{(i)}_{y_d,k,\nu})$, $\nu \in \{1, \dots, V\}$, and $\theta_d := \text{softmax}(\eta_d)$.

2.3 Estimation

- variational inference generally (mean-field variational Bayes)
- EM algorithm used in LDA
- no generally applicable algorithm for CTM in CTM itself
- non-conjugacy: blei and wang 2013: laplace approximation
- an approximate variational EM algorithm using a Laplace approximation

Atchison, J, and Sheng M Shen. 1980. "Logistic-Normal Distributions: Some Properties and Uses." *Biometrika* 67 (2): 261–72.

Blei, David M, John D Lafferty, and others. 2007. "A Correlated Topic Model of Science." *The Annals of Applied Statistics* 1 (1): 17–35.

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

Deerwester, Scott, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. "Indexing by Latent Semantic Analysis." *Journal of the American Society for Information Science* 41 (6): 391–407.

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*.

Eisenstein, Jacob, Amr Ahmed, and Eric P Xing. 2011. "Sparse Additive Generative Models of Text."

Hofmann, Thomas. 1999. "Probabilistic Latent Semantic Indexing." In *Proceedings of the 22nd Annual International Acm Sigir Conference on Research and Development in Information Retrieval*, 50–57.

James, Ian R, James E Mosimann, and others. 1980. "A New Characterization of the Dirichlet Distribution Through Neutrality." *The Annals of Statistics* 8 (1): 183–89.

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.

Roberts, Margaret E., Brandon M. Stewart, and Dustin Tingley. 2019. "stm: An R Package for Structural Topic Models." *Journal of Statistical Software* 91 (2): 1–40. https://doi.org/10.18637/jss.v091.i02.

Roberts, Margaret E, Brandon M Stewart, Dustin Tingley, Edoardo M Airoldi, and others. 2013. "The Structural Topic Model and Applied Social Science." In *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation*, 1–20. Harrahs; Harveys, Lake Tahoe.

Salton, Gerard, and M. J. McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill.

Wang, Chong, and David M Blei. 2013. "Variational Inference in Nonconjugate Models." *Journal of Machine Learning Research* 14 (Apr): 1005–31.