1 Introduction

- Reserach Question
  - Analyzing news reports about german elections, where we allow prevalence of topics to evolve over time (before and after the elections) and vary across newswire services.
  - Quantify the effect of news wire source on both topical prevalence (topic distribution) and topical content (word-topic distribution).
- Method
  - Topical content of documents, in which mixing weights are parameterized by observed covariates
  - Topic distribution word-topic distribution are specified as a simple generalized linear model on an arbitrary number of document-level covariates (news source, time of release).

Text as a data source is becoming increasingly popular in empirical economics, as the information encoded in text is a rich complement to conventional data (e.g. price, wage, votes, etc.) (**grimmer˙bayesian˙2010**; **grimmer˙text˙2013**; **quinn˙how˙2010**; **gentzkow˙text˙2017**; **roberts˙model˙2016**). On interesting question is, how document metadata such as author or date influence its content. Depending on the characteristics of a corpus (a collection of documents), documents contain more than one topic, as well as terms most likely appear in multiple topics, thus the use of mixed-membership models (**airoldi˙handbook˙2014**) - or topic model as they are often referred in the literature of text analysis (**blei˙probabilistic˙2012**) - are a popular approach.

Within these models the Latent Dirichlet Allocation (LDA hencefirth) is a widely used topic modelling technique, where each document is viewed as a mixture of topics (represented by the document-topic distribution) and each topic is a mixture of unique terms (represented by the topic-term distribution). To "learn" the topic prevalence and the topic-term distribution, collapsed Gibbs sampling[1] can be used. LDA makes a statistical assumption that all texts in the modeled corpus are generated by the same underlying process (**blei˙latent˙2003**). Thus, it is not ideally suited to examining differences in topical content that are affected by covariates such as author identity or time of writing.

[1] See Section 4.2 for a non-formal description of the Gibbs sampler

roberts˙model˙2016 develop a structural topic model (STM) that allows to incorporate external variables that effect both topical content and topical prevalence. We use this approach to analyze online news about the german federal elections, where we allow prevalence of topics to change over time (before and after the elections) and vary across newswire services (private or public) to quantify the effect of news wire source on both topical prevalence and topical content.

2 Related Literature

In **tetlock˙giving˙2007**, c_i is a bag-of-words representation and the outcome of interest v_i is the latent "sentiment" of Wall Street Journal columns, defined along a number of dimensions such as "positive," "optimistic," and so on. The author defines the function $f(\cdot)$ using a dictionary called the General Inquirer, which provides lists of words associated with each of these sentiment categories.2 The elements of $f(c_i)$ are defined to be the sum of the counts of words in each category.

Topic modeling is a statistical and computational technique for discerning information about the contents of a large corpus of documents without reading or annotating the original texts. A topic model uncovers patterns of word co-occurrence across the corpus, yielding a set of word clusters, together with associated probabilities of occurrence, which constitute the topics.

See (**taddy˙estimation˙2012**) for a review of topic estimation techniques)

Since its introduction into text analysis, topic modeling has become hugely popular.8 (See **blei˙probabilistic˙2012** for an overview.) The model has been especially useful in political science (e.g., (**grimmer˙bayesian˙2010**)), where researchers have been successful in attaching political issues and beliefs to the estimated latent topics.

Topic modeling is alternatively labeled as "latent Dirichlet allocation," (LDA) which refers to the Bayesian model in **blei˙latent˙2003** that treats each $\boldsymbol{v}_i$ and $\boldsymbol{\theta}_l$ as generated from a Dirichlet - distributed prior. The same model was independently introduced in genetics by **pritchard˙inference˙2000** for factorizing gene expression as a function of latent populations; it has been similarly successful in that field.

The basic topic model has been generalized and extended in variety of ways. A prominent example is the dynamic topic model of **blei˙dynamic˙2006**, which considers documents that are indexed by date (e.g., publication date for academic articles) and allows the topics, say $\boldsymbol{\Theta}_t$, to evolve smoothly in time.

A typical application of topic modeling in the social sciences first estimates LDA, then uses estimates of θ_d as the dependent variable in an regression on covariates to test whether different types of documents have different content.

This is contradictory because documents are assumed to be generated by a statistical process that we subsequently reject.

The structural topic model (STM) of Roberts et. al. (2016) explicitly introduces covariates into a topic model, and allows one to estimate the impact of document-level covariates on topic content and prevalence as part of the topic model itself.

3 Statistical Analysis of Text Data

Consider a collection of documents by $d \in \{1...D\}$, each containing $n \in \{1...N_d\}$ words. Primary observations consist of words $w_{d,n,}$ that are instances of unique terms from a vocabulary of terms, indexed by $v \in \{1...V\}$.

To use text as data and reduce the dimensionality, a common strategy is to (a) pre-process the text by imposing some preliminary restrictions (stop-word removal, tokenization) based on the nature of the data (twitter text, newspaper articles, speeches, etc.) and (b) to represent a document d as a vector of word counts, $\boldsymbol{n}_d \in \boldsymbol{N}^V$. This representation is often referred to as the bag of words model, since the order in which words are used within a document is completely disregarded. Nowadays, the bag of words model is a common representation for most of statistic literature about text data analysis (**blei˙latent˙2003**; **erosheva˙mixed-membership˙2004**; **griffiths˙finding˙2004**; **genkin˙large-scale˙2007**).

Term-Document matrices represent frequency distribution of unique terms in the documents. Any one document will contain only a subset of all unique terms, and the rows corresponding to unused terms will all be zero.The key task then becomes how to extract low-dimensional information from documents that are high-dimensional by nature. This is analogous to a situation in which a researcher has a database with thousands of covariates and is attempting to choose which subset of them, or which summary statistics, should be included in regression analysis.

1. Represent raw text D as a numerical array $\boldsymbol{C}$.

2. Map $\boldsymbol{C}$ to predict values $\hat{\boldsymbol{V}}$ of unknown outcomes $\boldsymbol{V}$.

   E.g. the variable of interest $\boldsymbol{V}$ is an indicator whether the email is spam. The prediction $\hat{\boldsymbol{V}}$ determines whether or not to send the email

to a spam filter. Sometimes the attribute of interest is latent, such as the topics of a newspaper article.

3. Use $\hat{\boldsymbol{V}}$ in subsequent descriptive or causal analysis.

Regarding 2, the methods to connect counts $\boldsymbol{c}_i$ to attributes $\boldsymbol{v}_i$ can be roughly divided into four categories (**gentzkow˙text˙2017**):

1. Dictionary-based methods: No statistical inference. Simply specify $\hat{\boldsymbol{v}}_{\boldsymbol{i}} = f(\boldsymbol{c}_i)$ for some unknown function $f(\cdot)$. Sometimes based on a specific dictionary of terms (**tetlock˙giving˙2007**, **baker˙measuring˙2015**).

2. Text regression methods: Directly estimate the conditional outcome distribution $p(\boldsymbol{v}_i|\boldsymbol{c}_i)$. Intuition: If we want to predict $\boldsymbol{v}_{\boldsymbol{i}}$ from $\boldsymbol{c}_i$, we would regress the observed values of the former ($\boldsymbol{V}^{train}$) on the corresponding latter ($\boldsymbol{C}^{train}$). High dimensionality of $\boldsymbol{c}_i$ ($p > n^{train}$) requires use of appropriate regression techniques to avoid overfitting (e.g. L_1 regularized linear or logistic regression)

3. Generative model of $p(\boldsymbol{c}_i|\boldsymbol{v}_i)$. Intuition: In many cases the underlying causal relationship runs from outcomes to language rather than the other way around. E.g. Google searches about flu do not cause flu cases to occur, rather, people with flu are more likely to produce such searches.

   (a) Observed attributes (supervised methods): Supervised machine learning starts with a researcher classifying observations to 'train' an algorithm under human 'supervision' – to 'learn' the correlation between the researcher's ascribed classes and words characteristic of documents in those classes (Grimmer and Stewart (2013)). Fitting the model based on the observed training data $\boldsymbol{V}^{train}$, say $f_{\boldsymbol{\theta}}(\boldsymbol{c}_i; \boldsymbol{v}_i)$ for a vector of parameters $\boldsymbol{\theta}$, to this training set. The fitted model $f_{\hat{\boldsymbol{\theta}}}$ can be inverted in order to infer $\boldsymbol{v}_i$ for documents in the test set.

   (b) Latent attributes (unsupervised methods): The function relating $\boldsymbol{c}_i$ to $\boldsymbol{v}_i$ is unknown, as we cannot observe the true value of v_i. Principal component analysis (PCA), latent Dirichlet allocation (LDA, topic modeling, structural-topic modeling). Unsupervised machine learning involves taking unclassified observations and uncovering hidden patterns that structure them in some meaningful way. The outputs of algorithms for unsupervised machine learning can be used as inputs into econometric models for predicting

some variable of interest, but this is a different approach from intentionally choosing the dimensions of content based on their predictive ability.

4. Deep learning techniques: neural networks, distributed language models.

The goal of this paper is to find the latent topics within newspaper articles and how different types of media outlets (as well as the date?) influence the topic prevalence as well as the language to describe a topic (the word-topic distribution). We implement generative model (topic model).

4 Generative Process

In this unsupervised method, the words in a document are viewed as the realization of some stochastic process. The generative process is defined through a probability model for $p(\boldsymbol{c}_i|\boldsymbol{v}_i)$.

Each observation $\boldsymbol{c}_i$ is a conditionally independent draw from the vocabulary of possible tokens accodring to some document-specific token probability vector $\boldsymbol{q}_i = [q_{i1}...q_{ip}]'$. Conditioning on document length, $m_i = \sum_j c_{ij}$, this implies a multinomial distribution for the counts

$$\boldsymbol{c}_i \sim \boldsymbol{MN}(\boldsymbol{q}_i, m_i). \tag{1}$$

Under the basic model in 1, a connection between text and attributes is defined through the link function $\boldsymbol{q}_i = q(\boldsymbol{v}_i)$.

$$\boldsymbol{E}\left[\frac{\boldsymbol{c}_i}{m_i}\right] = \boldsymbol{q}_i = v_{i1}\boldsymbol{\theta}_1 + v_{i2}\boldsymbol{\theta}_2 + v_{ik}\boldsymbol{\theta}_k = \boldsymbol{\Theta}\boldsymbol{v}_i \tag{2}$$

where attributes v_{il} are referred as topic wights, restricting $v_{il} \geq 0$ and $\sum_{l=1}^{k} v_{il} = 1$, and each topic $\boldsymbol{\theta}_l$ is a probability vector over possible tokens: $\theta_{lj} \geq 0$ and $\sum_{j=1}^{p} \theta_{il} = 1$. Each $\boldsymbol{v}_i$ and $\boldsymbol{\theta}_l$ is generated from a Dirichlet-distributed prior.

Latent Dirichlet allocation (LDA) is a particularly popular method for fitting a topic model. It treats each document as a mixture of topics, and each topic as a mixture of words. This allows documents to "overlap" each other in terms of content, rather than being separated into discrete groups, in a way that mirrors typical use of natural language. Formally speaking, each document has its own probability distribution over topics. Then, for each word in each document, a topic assignment is made and then, conditional on the assignment, a word from the corresponding topic.

Estimation of topic models make use of some alternating inference for $\boldsymbol{V}|\boldsymbol{\Theta}$ and $\boldsymbol{\Theta}|\boldsymbol{V}$.

1. Expectation-maximization algorithm (EM) Either maximize the likelihood implied by 1 and 2 or, after incorporating the usual Dirichlet priors on $\boldsymbol{v}_i$ and $\boldsymbol{\theta}_l$ (**taddy˙estimation˙2012**)

2. Target full posterior distribution $p(\boldsymbol{\Theta}, \boldsymbol{V}|\boldsymbol{c}_i)$

Choice of number of topics k is often fairly arbitrary. In practice it is very common to simply start with a number of topics on the order of ten, and then adjust the number of topics in whatever direction seems to improve interpretability. Whether this ad hoc procedure is problematic depends on the application. In many applications of topic models to date the goal is to provide an intuitive description of text rather than inference on some underlying "true" parameters; in these cases, the ad hoc selection of the number of topics may be reasonable.

Data-driven approaches:

1. **taddy˙estimation˙2012** describes a model selection process for k that is based upon Bayes factors.

2. **airoldi¨reconceptualizing˙2010** provide a cross-validation (CV) scheme

3. **teh˙hierarchical˙2006** use Bayesian nonparametric techniques that view k as an unknown model parameter.

4.1 Structural Topic Model

The process for generating individual words is the same as for plain LDA conditional on the β_k and π_d terms.

However both objects can depend on potentially different sets of document-level covariates. Each document has:

1. Topic Prevalence. Attributes r_d that affect the likelihood of discussing topic k. how much of a document is associated with a topic

2. Topic Content. Attributes r_d that affect the likelihood of discussing term v overall, and of discussing it within topic k. the words used within a topic

The generation of the k and d terms is via multinomial logistic regression, which breaks local conjugacy.

The standard topic modeling technique, Latent Dirichlet Allocation (LDA), may have limited utility in the realm of social media. LDA makes a statistical assumption that all texts in the modeled corpus are generated by the same underlying process (Blei). Thus, it is not ideally suited to examining differences in topical content that are affected by external variables such as author identity or time of writing.

Structural topic modeling (STM) is a recently introduced variant of LDA that is designed to address precisely this limitation . STM can represent the effect of external variables on both topical content and topical prevalence. The external variables can consist of any metadata that distinguishes one text from another, including variables relating to author identity (gender, age, political affiliation, etc.), textual genre (for example, news stories versus academic articles), and time of production.

stmVigenette: "The goal of the Structural Topic Model is to allow researchers to discover topics and estimate their relationship to document metadata. Outputs of the model can be used to conduct hypothesis testing about these relationships."

4.1.1 Estimation of the STM

In STM, metadata can be entered in the topic model in two ways: topical prevalence and topical content. Metadata covariates for topical prevalence allow the observed metadata to affect the frequency with which a topic is discussed. Covariates in topical content allow the observed metadata to affect the word rate use within a given topicthat is, how a particular topic is discussed.

We use the online magazine-type as a covariate in the topic prevalence portion of the model with the data described above. Each document is modeled as a mixture of multiple topics. Topical prevalence captures how much each topic contributes to a document. Because different documents come from different sources, it is natural then to want to allow this prevalence to vary with metadata that we have about document sources.

We will simply let prevalence be a function of the magazine variable, which is coded as either Spiegel Online or FOCUS Online and the variable day which is an integer measure of days running from 01-01-2017 to 31-07-2017.

4.2 Gibbs Sampler

Strategy for discovering topics **griffiths˙finding˙2004**

- Considering the posterior distribution over the assignments of words to topics, $P(z|w)$. (and not explicitly representing ϕ or θ as parameters to be estimated)
- Examine this posterior distribution to obtain ϕ or θ

After making assumptions about the parameters (number of topics K, prior distributions α and β), the procedure of Gibbs sampling is as follows:

1. Go through each document and randomly assign each word in the document to one of K topics, based on the prior distributions.
2. For each document d, go through each word w and compute:
   (a) The document-topic distribution $\theta = p(t|d)$
   (b) The topic-term distribution $\phi = p(w|t)$
3. Reassign word w a new topic t^*, where we choose topic t^* with probability $p(t^*|d) * p(w|t^*)$

On repeating the last step a large number of times, the algorithm reaches a steady state where topic assignments are pretty good. These posterior distributions θ and ϕ are then used to determine the topic mixtures of each document.

4.3 Validate accuracy

- Manual audits: cross checking some subset of the fitted values against the coding a human would produce by hand
- Inspection of fitted parameters
- Interpretation of fitted topics usually proceeds by ranking the tokens in each topic according to token probability.

Caution against the over-interpretation of unsupervised models: posterior distributions informing parameter estimates are highly multimodal, and multiple topic model runs can lead to multiple different interpretations. Add some supervision (**airoldi˙improving˙2016**, **gentzkow˙text˙2017**)