Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://amstat.tandfonline.com/loi/uasa20

A Model of Text for Experimentation in the Social Sciences

Margaret E. Roberts, Brandon M. Stewart & Edoardo M. Airoldi

To cite this article: Margaret E. Roberts, Brandon M. Stewart & Edoardo M. Airoldi (2016) A Model of Text for Experimentation in the Social Sciences, Journal of the American Statistical Association, 111:515, 988-1003, DOI: 10.1080/01621459.2016.1141684

To link to this article: https://doi.org/10.1080/01621459.2016.1141684

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2016, VOL. 111, NO. 515, 988–1003, Applications and Case Studies
http://dx.doi.org/10.1080/01621459.2016.1141684

Taylor & Francis
Taylor & Francis Group

A Model of Text for Experimentation in the Social Sciences

Margaret E. Roberts[a], Brandon M. Stewart[b], and Edoardo M. Airoldi[c]

[a]Department of Political Science, University of California San Diego, San Diego, CA, USA; [b]Department of Sociology, Princeton University, Princeton, NJ, USA; [c]Department of Statistics, Harvard University, Cambridge, MA, USA

ABSTRACT

Statistical models of text have become increasingly popular in statistics and computer science as a method of exploring large document collections. Social scientists often want to move beyond exploration, to measurement and experimentation, and make inference about social and political processes that drive discourse and content. In this article, we develop a model of text data that supports this type of substantive research. Our approach is to posit a hierarchical mixed membership model for analyzing topical content of documents, in which mixing weights are parameterized by observed covariates. In this model, topical *prevalence* and topical *content* are specified as a simple generalized linear model on an arbitrary number of document-level covariates, such as news source and time of release, enabling researchers to introduce elements of the experimental design that informed document collection into the model, within a generally applicable framework. We demonstrate the proposed methodology by analyzing a collection of news reports about China, where we allow the prevalence of topics to evolve over time and vary across newswire services. Our methods quantify the effect of news wire source on both the frequency and nature of topic coverage. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received March 2014
Revised October 2015

KEYWORDS
Causal inference; Experimentation; High-dimensional inference; Social sciences; Text analysis; Variational approximation

1. Introduction

Written documents provide a valuable source of data for the measurement of latent linguistic, political, and psychological variables (e.g., Socher et al. 2009; Grimmer 2010; Quinn et al. 2010; Grimmer and Stewart 2013). Social scientists are primarily interested in how document metadata, that is, observable covariates such as author or date, influence the content of the text. With the rapid digitization of texts, larger and larger document collections are becoming available for analysis, for which such metadata information is recorded. A fruitful approach for the analysis of text data is the use of mixtures and mixed membership models (Airoldi et al. 2014a), often referred to as *topic models* in the literature (Blei 2012). While these models can provide insights into the topical structure of a document collection, they cannot easily incorporate the observable metadata information. Here, we develop a framework for modeling text data that can flexibly incorporate a wide range of document-level covariates and metadata, and capture their effect on topical content. We apply our model to learn about how media coverage of China's rise varies over time and by newswire service.

Quantitative approaches to text data analysis have a long history in the social sciences (Mendenhall 1887; Zipf 1932; Yule 1944; Miller, Newman, and Friedman 1958). Today, the most common representation of text data involves representing a document d as a vector of word counts, $\mathbf{w}_d \in \mathbb{Z}_+^V$, where each of the V entries map to a unique term in a vocabulary of interest (with V in the order of thousands to tens of thousands) specified prior to the analysis. This representation is often referred to as the *bag of words* representation, since the order in which words are used within a document is completely disregarded. One milestone in the statistical analysis of text was the analysis of the disputed authorship of "The Federalist" articles (Mosteller and Wallace 1963, 1964, 1984), which featured an in-depth study of the extent to which assumptions used to reduce the complexity of text data representations hold in practice. Because the bag of words representation retains word co-occurrence information, but loses the subtle nuances of grammar and syntax, it is most appropriate for settings where the quantity of interest is a coarse summary such as topical content (Manning, Raghavan, and Schütze 2008; Turney and Pantel 2010). In recent years, there has been a surge of interest in methods for text data analysis in the statistics literature, most of which use the bag of words representation (e.g., Blei, Ng, and Jordan 2003; Erosheva, Fienberg, and Lafferty 2004; Griffiths and Steyvers 2004; Genkin, Lewis, and Madigan 2007; Jeske and Liu 2007; Airoldi et al. 2010; Taddy 2013; Jia et al. 2014). A few studies also test the appropriateness of the assumptions underlying such a representation (e.g., Airoldi and Fienberg 2003; Airoldi et al. 2006).

Perhaps the simplest topic model, to which, arguably, much of the recent interest in statistical text analysis research can be ascribed, is known as the *latent Dirichlet allocation* (LDA henceforth), or also as the *generative aspect model* (Blei, Ng, and Jordan 2001, 2003; Minka and Lafferty 2002). Consider a collection of D documents, indexed by d, each containing N_d words, a vocabulary of interest of V distinct terms, and K

CONTACT Edoardo M. Airoldi *airoldi@fas.harvard.edu* Department of Statistics, Harvard University, Cambridge, MA 02138.
Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/JASA*.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

subpopulations, indexed by k and referred to as *topics*. Each topic is associated with a V-dimensional probability mass function, $\boldsymbol{\beta}_k$, that controls the frequency according to which terms are generated from that topic. The data-generating process for document d assigns terms in the vocabulary to each of the N_d positions; instances of terms that fill these positions are typically referred to as the *words*. Terms in the vocabulary are unique, while distinct words in a document may instantiate multiple occurrences of the same term. The process begins by drawing a K-dimensional Dirichlet vector $\boldsymbol{\theta}_d$ that captures the expected proportion of words in document d that can be attributed to each topic. Then for each position (or, equivalently, for each word) in the document, indexed by n, it proceeds by sampling an indicator $z_{d,n}$ from a $\text{Multinomial}_K(\boldsymbol{\theta}_d, 1)$ whose positive component denotes which topic such position is associated with. The process ends by sampling the actual word indicator $w_{d,n}$ from a $\text{Multinomial}_V(\mathbf{B}\, z_{d,n}, 1)$, where the matrix $\mathbf{B} = [\boldsymbol{\beta}_1 | \cdots | \boldsymbol{\beta}_K]$ encodes the distributions over terms in the vocabulary associated with the K topics.

In practice, social scientists often know more about a document than its word counts. For example, open-ended responses collected as part of a survey experiment include additional information about the respondents, such as gender or political party (Roberts et al. 2014b). From a statistical perspective, it would be desirable to include additional covariates and information about the experimental design into the model to improve estimation of the topics. In addition, the relationships between the observed covariates and latent topics are most frequently the estimand of scientific interest. Here, we allow for such observed covariates to affect two components of the model, the proportion of a document devoted to a topic, which we refer to as *topic prevalence* and the word rates used in discussing a topic, which we refer to as *topical content*.

We leverage generalized linear models (GLMs henceforth) to introduce covariate information into the model. Prior distributions with globally shared mean parameters in the latent Dirichlet allocation model are replaced with means parameterized by a linear function of observed covariates. Specifically, for topic prevalence, the Dirichlet distribution that controls the proportion of words in a document attributable to the different topics is replaced with a logistic Normal distribution with a mean vector parameterized as a function of the covariates (Aitchison and Shen 1980). For topical content, we define the distribution over the terms associated with the different topics as an exponential family model, similar to a multinomial logistic regression, parameterized as a function of the marginal frequency of occurrence deviations for each term, and of deviations from it that are specific to topics, covariates, and their interactions. We shall often refer to the resulting model as the *structural topic model* (STM), because the inclusion of covariates is informative about structure in the document collection and its design. From an inferential perspective, including covariate information allows for partial pooling of parameters along the structure defined by the covariates.

As with other topic models, the exact posterior for the proposed model is intractable, and suffers from identifiability issues in theory (Airoldi et al. 2014a). Inference is further complicated in our setting by the nonconjugacy of the logistic Normal with the multinomial likelihood. We develop a partially collapsed variational expectation-maximization algorithm that uses a Laplace approximation to the nonconjugate portion of the model (Dempster, Laird, and Rubin 1977; Liu 1994; Meng and Van Dyk 1997; Blei and Lafferty 2007; Wang and Blei 2013). This inference strategy provides a computationally efficient approach to model fitting that is sufficiently fast and well behaved to support the analysis of large collections of documents, in practice. We use posterior predictive checks (Gelman, Meng, and Stern 1996) to examine the model and assess model fit, and tools for model selection and interpretation we developed in substantive companion articles (Roberts et al. 2014b; Lucas et al. 2015).

The central contribution of this article is twofold: we introduce a new model of text that can flexibly incorporate various forms of document-level information, and we demonstrate how this model enables an original analysis of the differences among newswire services, in the frequency with which they cover topics and the vocabulary with which they describe topics. In particular, we are interested in characterizing how Chinese sources represent topics differently than foreign sources, or whether they leave out specific topics completely. The model allows us to produce the first quantification of media slant in various Chinese and international newswire services, over a 10 year period of China's rise. In addition, the model allows us to summarize slant more quickly than reading large swaths of text, the method more frequently used by China scholars. The article is organized as follows. We motivate the use of text analysis in the social sciences and provide the essential background for our model. We describe the structural topic model and discuss the proposed estimation strategy. We empirically validate the frequentist coverage of STM in a realistic simulation, and provide a comparative performance analysis with state-of-the-art models on real data. We use STM to study media coverage of China's rise by analyzing variations in topic prevalence and content across five different newswire services over time.

To make the model accessible to social scientists, we developed the R package `stm` (Roberts, Stewart, and Tingley 2014a), which handles model estimation, summary, and visualization (*http://cran.r-project.org/web/packages/stm*).

1.1. Statistical Analysis of Text Data in the Social Sciences

Our development of the proposed model is motivated by a common structure in the application of models for text data within the social sciences. In these settings, the typical application involves estimating latent topics for a corpus of interesting documents and subsequently comparing how topic proportions vary with an external covariate of interest. While informative, these applications raise a practical and theoretical tension. Documents are typically assumed to be exchangeable according to the model used for data analysis, but the exchangeability assumption is then often invalidated by the research findings reported.

This problem has motivated the development of a series of application-specific models designed to capture particular quantities of interest (Ahmed and Xing 2010; Grimmer 2010; Quinn et al. 2010; Gerrish and Blei 2012). Many of the models designed to incorporate various forms of meta-data allow the topic mixing proportions ($\boldsymbol{\theta}_d$) or the observed words (w) to be drawn from document-specific prior distributions rather than globally shared priors $\boldsymbol{\alpha}$, $\boldsymbol{\beta}_k$ in the LDA model. We refer to the distribution over the document-topic proportions as the prior on topical *prevalence* and we refer to the topic-specific distribution over words as the topical *content* prior. For example, the author-topic model allows the prevalence of topics to vary by author (Rosen-Zvi et al. 2004), the geographic topic model allows topical content to vary by region (Eisenstein et al. 2010), and the dynamic topic model allows topic prevalence and topic content to drift over time (Blei and Lafferty 2006).

However, for the vast majority of social scientists, designing a specific model for each application is prohibitively difficult. These users would need a general model that would balance flexibility to accommodate unique research problems with ease of use.

Our approach to this task builds on two prior efforts to incorporate general covariate information into topic models, the Dirichlet-multinomial regression topic model of Mimno and McCallum (2008) and the sparse additive generative model of Eisenstein, Ahmed, and Xing (2011). The model of Mimno and McCallum (2008) replaces the Dirichlet prior on the topic mixing proportions in the LDA model with a Dirichlet-multinomial regression over arbitrary covariates. This allows the prior distribution over document-topic proportions to be specific to a set of observed document features through a linear model. Our model extends this approach by allowing covariance among topics and emphasizing the use of nonlinear functional forms of the features.

While the Dirichlet-multinomial regression model focuses on topical prevalence, the sparse additive generative model allows topical content to vary by observed categorical covariates. In this framework, topics are modeled as sparse log-transformed deviations from a baseline distribution over words. Regularization to the corpus mean ensures that rarely occurring words do not produce the most extreme loadings onto topics (Eisenstein, Ahmed, and Xing 2011). Because the model is linear in the log-probability it becomes simple to combine several effects (e.g., topic, covariate, or topic-covariate interaction) by simply including the deviations additively in the linear predictor. We adopt a similar infrastructure to capture changes in topical content and extend the setting to any covariates.

An alternative is to fit word counts directly as a function of observable covariates and fixed or random effects (Taddy 2013), at the cost of specifying thousands of such effects.

Our solution to the need for a flexible model combines and extends these existing approaches to create the structural topic model (STM henceforth), so-called because we use covariates to structure the corpus beyond exchangeable documents.

2. A Model of Text that Leverages Covariate Information

We introduce the basic structural topic model and notation in Section 2.1. We discuss how covariates inform the model in the Section 2.1.1, and prior specifications in Section 2.1.2.

2.1. Basic Structural Topic Model

Recall that we index the documents by $d \in \{1 \ldots D\}$ and the words (or positions) within the documents by $n \in \{1 \ldots N_d\}$. Primary observations consist of words $w_{d,n}$ that are instances of unique terms from a vocabulary of terms, indexed by $v \in \{1 \ldots V\}$, deemed of interest in the analysis. The model also assumes that the analyst has specified the number of topics K indexed by $k \in \{1 \ldots K\}$. Additional observed information is given by two design matrices, one for topic prevalence and one for topical content, where each row defines a vector of covariates for a given document specified by the analyst. The matrix of topic prevalence covariates is denoted by $\mathbf{X}$, and has dimension $D \times P$. The matrix of topical content covariates is denoted by $\mathbf{Y}$ and has dimension $D \times A$. Rows of these matrices are denoted by $\mathbf{x}_d$ and $\mathbf{y}_d$, respectively. Last, we define m_v to be the marginal log frequency of term v in the vocabulary, easily estimable from total counts (see, e.g., Airoldi, Cohen, and Fienberg 2005).

The proposed model can be conceptually divided into three components: (1) a topic prevalence model, which controls how words are allocated to topics as a function of covariates, (2) a topical content model, which controls the frequency of the terms in each topic as a function of covariates, and (3) a core language (or observation) model, which combines these two sources of variation to produce the actual words in each document. Next, we discuss each component of the model in turn. A graphical illustration of the full data-generating process for the proposed model is provided in Figure 1.

To illustrate the model clearly, we will specify a particular default set of priors. The model, however, as well as the R package `stm`, allow for a number of alternative prior specifications, which we discuss in Section 2.1.2.

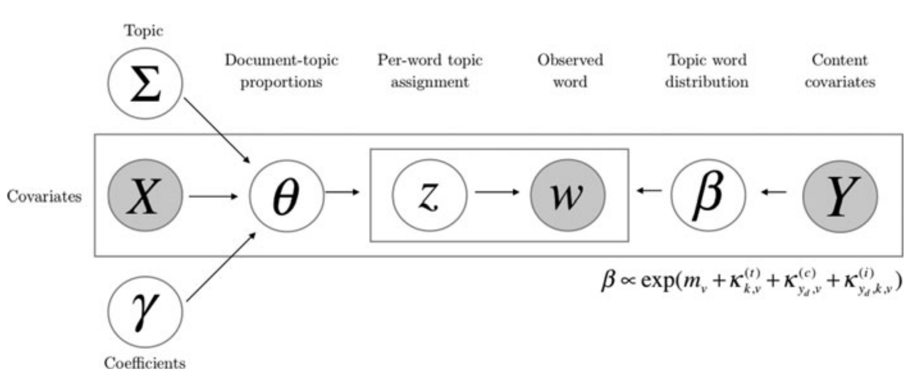

Figure 1. A graphical illustration of the structural topic model.

The data-generating process for document d, given the number of topics K, observed words $\{w_{d,n}\}$, the design matrices for topic prevalence $\mathbf{X}$, and topical content $\mathbf{Y}$, scalar hyper-parameters s, r, ρ, and K-dimensional hyper-parameter vector $\boldsymbol{\sigma}$, is as follows:

$$\boldsymbol{\gamma}_k \sim \text{Normal}_P(0, \sigma_k^2 I_P), \quad \text{for } k = 1 \ldots K-1, \tag{1}$$

$$\boldsymbol{\theta}_d \sim \text{LogisticNormal}_{K-1}(\boldsymbol{\Gamma}'\mathbf{x}_d', \boldsymbol{\Sigma}), \tag{2}$$

$$\mathbf{z}_{d,n} \sim \text{Multinomial}_K(\boldsymbol{\theta}_d), \quad \text{for } n = 1 \ldots N_d, \tag{3}$$

$$\mathbf{w}_{d,n} \sim \text{Multinomial}_V(\mathbf{B}\,\mathbf{z}_{d,n}), \quad \text{for } n = 1 \ldots N_d, \tag{4}$$

$$\beta_{d,k,v} = \frac{\exp\left(m_v + \kappa^{(\mathrm{t})}_{k,v} + \kappa^{(\mathrm{c})}_{y_d,v} + \kappa^{(\mathrm{i})}_{y_d,k,v}\right)}{\sum_v \exp\left(m_v + \kappa^{(\mathrm{t})}_{k,v} + \kappa^{(\mathrm{c})}_{y_d,v} + \kappa^{(\mathrm{i})}_{y_d,k,v}\right)}, \quad \text{for } v = 1 \ldots V \text{ and } k = 1 \ldots K, \tag{5}$$

where $\Gamma = [\boldsymbol{\gamma}_1|\ldots|\boldsymbol{\gamma}_K]$ is a $P \times (K-1)$ matrix of coefficients for the topic prevalence model specified by Equations (1) and (2), and $\{\kappa^{(t)}_{.,.}, \kappa^{(c)}_{.,.}, \kappa^{(i)}_{.,.}\}$ is a collection of coefficients for the topical content model specified by Equation (5) and further discussed below. Equations (3) and (4) denote the core language model.

The core language model allows for correlations in the topic proportions using the logistic normal distribution (Aitchison and Shen 1980; Aitchison 1982). For a model with K topics, we can represent the logistic normal by drawing $\boldsymbol{\eta}_d \sim \text{Normal}_{K-1}(\boldsymbol{\mu}_d, \boldsymbol{\Sigma})$ and mapping to the simplex, by specifying $\theta_{d,k} = \exp(\eta_{d,k})/(\sum_{i=1}^K \exp(\eta_{d,i}))$, where $\eta_{d,K}$ is fixed to zero to render the model identifiable. Given the topic proportion vector, $\boldsymbol{\theta}_d$, for each word within document d, a topic is sampled from a multinomial distribution $\mathbf{z}_{d,n} \sim \text{Multinomial}(\boldsymbol{\theta}_d)$, and conditional on such a topic, a word is chosen from the appropriate distribution over terms $\mathbf{B}\,\mathbf{z}_{d,n}$, also denoted $\boldsymbol{\beta}_{z_{d,n}}$ for simplicity. While in previous research (e.g., Blei and Lafferty 2007) both $\boldsymbol{\mu}$ and $\mathbf{B}$ are global parameters shared by all documents, in the proposed model they are specified as a function of document-level covariates.

2.1.1. Modeling Topic Prevalence and Topic Content With Covariates

The topic prevalence component of the model allows the expected document-topic proportions to vary as a function of the matrix of observed document-level covariates ($\mathbf{X}$), rather than arising from a single prior shared by all documents. We model the mean vector of the logistic normal as a simple linear model such that $\boldsymbol{\mu}_d = \boldsymbol{\Gamma}'\mathbf{x}_d'$, with an additional regularizing prior on the elements of $\boldsymbol{\Gamma}$ to avoid over-fitting. Intuitively, the topic prevalence model takes the form of a multivariate normal linear model with a single shared variance-covariance matrix of parameters. In the absence of covariates, but with a constant intercept, this portion of the model reduces to the model by Blei and Lafferty (2007).

To model the way covariates affect topical content, we draw on a parameterization that has proved useful in the text analysis literature for modeling differential word usage (e.g., Mosteller and Wallace 1984; Airoldi et al. 2006; Eisenstein, Ahmed, and Xing 2011). The idea is to parameterize the (multinomial) distribution of word occurrences in terms of log-transformed rate deviations from the rates of a corpus-wide background distribution $\mathbf{m}$, which can be estimated or fixed to a distribution of interest. The log-transformed rate deviations can then be specified as a function of topics, of observed covariates, and of topic-covariate interactions. In the proposed model, the log-transformed rate deviations are denoted by a collection of parameters $\{\boldsymbol{\kappa}\}$, where the superscript indicates which set they belong to, that is, t for topics, c for covariates, or i for topic-covariate interactions. In detail, $\boldsymbol{\kappa}^{(t)}$ is a K-by-V matrix containing the log-transformed rate deviations for each topic k and term v, over the baseline log-transformed rate for term v. These deviations are shared across all A levels of the content covariate Y_d. The matrix $\boldsymbol{\kappa}^{(c)}$ has dimension $A \times V$, and it contains the log-transformed rate deviation for each level of the covariate Y_d and each term v, over the baseline log-transformed rate for term v. These deviations are shared across all topics. Finally, the array $\boldsymbol{\kappa}^{(i)}$ has dimension $A \times K \times V$, and it collects the covariate-topic interaction effects. For example, for the simple case where there is a single covariate (Y_d) denoting a mutually exclusive and exhaustive group of documents, such as newswire source, the distribution over terms is obtained by adding these log-transformed effects such that the rate $\beta_{d,k,v} \propto \exp(m_v + \kappa^{(\mathrm{t})}_{k,v} + \kappa^{(\mathrm{c})}_{y_d,v} + \kappa^{(\mathrm{i})}_{y_d,k,v})$, where m_v is the marginal log-transformed rate of term v. Typically, m_v is specified as the estimated (marginal) log-transformed rate of occurrence of term v in the document collection under study (see, e.g., Airoldi, Cohen, and Fienberg 2005), but can alternatively be specified as any baseline distribution of interest. The content model is completed by positing sparsity inducing priors for the $\{\kappa\}$ parameters, so that topic and covariate effects represent sparse deviations from the background distribution over terms. We defer discussion of prior specification to Section 2.1.2. Intuitively, the proposed topical content model replaces the multinomial likelihood for the words with a multinomial logistic regression, where the covariates are the word-level topic latent variables $\{z_{d,n}\}$, the user-supplied covariates $\{Y_d\}$, and their interactions. In principle, we need not restrict ourselves to models with single categorical covariates; in practice, computational considerations dictate that the number of levels of topical content covariates be relatively small.

The specification of the topic prevalence model is inspired by generalized additive models (Hastie and Tibshirani 1990). Each covariate is included with B-splines (De Boor et al. 1978), which allows nonlinearity in the effects on the latent topic prevalence, but the covariates themselves remain additive in the specification. The inclusion of a particular covariate allows the model to borrow strength from documents with similar covariate values when estimating the document-topic proportions, analogously to partial pooling in other Bayesian hierarchical models (Gelman and Hill 2007). We also include covariates that affect the rate at which terms are used within a topic through the topical content model. Unlike covariates for topical prevalence, for each observed content covariate combination it is necessary to maintain a dense $K \times V$ matrix, namely, the expected number of occurrences of term v attributable to topic k, within documents having that observed covariate level.

2.1.2. Prior Specifications

The prior specification for the topic prevalence parameters is a zero mean Gaussian distribution with shared variance parameter, that is, $\gamma_{p,k} \sim \text{Normal}(0, \sigma_k^2)$, and $\sigma_k^2 \sim$ Inverse-Gamma(a, b), where p indexes the covariates, k indexes the topics, and a, b are fixed hyperparameters (see the Appendix for more details). There is no prior on the intercept, if included as a covariate. This prior shrinks coefficients toward zero, but does not induce sparsity.

In the topical content specification, we posit a Laplace prior (Friedman, Hastie, and Tibshirani 2010) to induce sparsity on the collection of $\{\boldsymbol{\kappa}\}$ parameters. This is necessary for interpretability. See the Appendix for details of how the hyperparameters are calibrated.

2.2. Estimation and Interpretation

The full posterior of interest, $p(\boldsymbol{\eta}, \mathbf{z}, \boldsymbol{\kappa}, \boldsymbol{\gamma}, \Sigma|\mathbf{w}, \mathbf{X}, \mathbf{Y})$, is proportional to

$$\left(\prod_{d=1}^{D} \text{Normal}(\boldsymbol{\eta}_d|\mathbf{X}_d\boldsymbol{\gamma}, \Sigma) \left(\prod_{n=1}^{N} \text{Multinomial}(z_{n,d}|\boldsymbol{\theta}_d) \times \text{Multinomial}(w_n|\boldsymbol{\beta}_{d,k=z_{d,n}})\right)\right) \times \prod p(\boldsymbol{\kappa}) \prod p(\boldsymbol{\Gamma})$$

with $\theta_{d,k} = \exp(\eta_{d,k})/(\sum_{i=1}^{K}\exp(\eta_{d,i}))$ and $\beta_{d,k,v} \propto \exp(m_v + \kappa^{(\text{t})}_{k,v} + \kappa^{(\text{c})}_{y_d,v} + \kappa^{(\text{i})}_{y_d,k,v})$, and the priors on the prevalence and content coefficients $\boldsymbol{\Gamma}$, $\boldsymbol{\kappa}$ specific to the options chosen by the user. As with most topic models, the posterior distribution for the structural topic model is intractable and so we turn to methods of approximate inference. To allow for ease of use in iterative model fitting, we use a fast variant of nonconjugate variational expectation-maximization (EM).

Traditionally, topic models have been fit using either collapsed Gibbs sampling or mean field variational Bayes (Blei, Ng, and Jordan 2003; Griffiths and Steyvers 2004). Because the logistic normal distribution introduces nonconjugacy, these standard methods are not available. The original work on logistic normal topic models used an approximate variational Bayes procedure by maximizing a novel lower bound on the marginal likelihood (Blei and Lafferty 2007) but the bound can be quite loose (Ahmed and Xing 2007; Knowles and Minka 2011). Later work drew on inference for logistic regression models (Groenewald and Mokgatlhe 2005; Holmes and Held 2006) to develop a Gibbs sampler using auxiliary variable schemes (Mimno, Wallach, and McCallum 2008). Recently, Chen et al. (2013) developed a scalable Gibbs sampling algorithm by leveraging the Polya-Gamma auxiliary variable scheme of Polson, Scott, and Windle (2013).

Instead, we developed an approximate variational EM algorithm using a Laplace approximation to the expectations rendered intractable by the nonconjungacy (Wang and Blei 2013). To speed convergence, empirically, we also integrate out the word-level topic indicator z while estimating the variational parameters for the logistic normal latent variable, and then reintroduce it when maximizing the topic-word distributions, β. Thus inference consists in optimizing the variational posterior for each document's topic proportions in the E-step, and estimating the topical prevalence and content coefficients in the M-step.

2.2.1. Variational Expectation-Maximization

Recall that we can write the logistic normal document-topic proportions in terms of the $K-1$-dimensional Gaussian random variable such that $\boldsymbol{\theta}_d = \frac{\exp(\boldsymbol{\eta}_d)}{\sum_{k=1}^{K}\exp(\eta_{d,k})}$ where $\boldsymbol{\eta}_d \sim \text{Normal}(\mathbf{x}_d\boldsymbol{\Gamma}, \Sigma)$ where $\eta_{d,K}$ is set to 0 for identification. Inference involves finding the approximate posterior $\prod_d q(\boldsymbol{\eta}_d)q(\mathbf{z}_d)$, which maximizes the approximate evidence lower bound (ELBO),

$$\begin{aligned}\text{ELBO} \approx &\sum_{d=1}^{D} E_q[\log p(\boldsymbol{\eta}_d|\boldsymbol{\mu}_d, \boldsymbol{\Sigma})] \\ &+ \sum_{d=1}^{D}\sum_{n=1}^{N} E_q[\log p(z_{n,d}|\boldsymbol{\eta}_d)] \\ &+ \sum_{d=1}^{D}\sum_{n=1}^{N} + E_q[\log p(w_{n,d}|z_{n,d}, \boldsymbol{\beta}_{d,k=z_{d,n}})] - H(q),\end{aligned} \tag{6}$$

where $q(\boldsymbol{\eta}_d)$ is fixed to be Gaussian with mean $\boldsymbol{\lambda}_d$ and covariance $\boldsymbol{\nu}_d$ and $q(\mathbf{z}_d)$ is a variational multinomial with parameter $\boldsymbol{\phi}_d$. $H(q)$ denotes the entropies of the approximating distributions. We qualify the ELBO as approximate to emphasize that it is not a true bound on the marginal likelihood (due to the Laplace approximation) and it is not being directly maximized by the updates (see, e.g., Wang and Blei 2013, for more discussion).

In the E-step, we iterate through each document updating the variational posteriors $q(\boldsymbol{\eta}_d)$, $q(\boldsymbol{\phi}_d)$. In the M-step, we maximize the approximate ELBO with respect to the model parameters $\boldsymbol{\Gamma}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\kappa}$. After detailing the E-step and M-step, we discuss convergence, properties, and initialization before summarizing the complete algorithm.

In practice, one can monitor convergence in terms of relative changes to the approximate ELBO. This boils down to a sum over the document level contributions, and can be dramatically simplified from Equation (6) to the following,

$$\begin{aligned}\mathcal{L}_{\text{ELBO}} = \sum_{d=1}^{D}\Bigg(\Bigg(&\sum_{i=1}^{V} w_{d,v}\log(\boldsymbol{\theta}_d\boldsymbol{\beta}_{d,v})\Bigg) - 0.5\log|\boldsymbol{\Sigma}| \\ &-0.5(\boldsymbol{\lambda}_d - \boldsymbol{\mu}_d)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\lambda}_d - \boldsymbol{\mu}_d) + 0.5\log(|\boldsymbol{\nu}_d|)\Bigg)\end{aligned} \tag{7}$$

Variational E-Step. Because the logistic-normal is not conjugate with the multinomial, $q(\boldsymbol{\eta}_d)$ does not have a closed-form update. We instead adopt the Laplace approximation advocated in Wang and Blei (2013), which involves finding the maximum a posteriori (MAP) estimate $\hat{\boldsymbol{\eta}}_d$ and approximating the posterior with a quadratic Taylor expansion. This results in a Gaussian form for the variational posterior $q(\boldsymbol{\eta}_d) \approx \mathcal{N}(\hat{\boldsymbol{\eta}}_d, -\nabla^2 f(\hat{\boldsymbol{\eta}}_d)^{-1})$, where $\nabla^2 f(\hat{\boldsymbol{\eta}}_d)$ is the Hessian of $f(\boldsymbol{\eta}_d)$ evaluated at the mode. In standard variational approximation, algorithms for the correlated topic model (CTM) inference iterates between the word-level latent variables $q(\mathbf{z}_d)$ and the document-level latent variables $q(\boldsymbol{\eta}_d)$ until local convergence. This process can be slow, and so we integrate out the latent variables z and find the joint optimum using quasi-Newton methods

(Khan and Bouchard 2009). Solving for $\hat{\boldsymbol{\eta}}_d$ for a given document amounts to optimizing the function,

$$f(\hat{\boldsymbol{\eta}}_d) \propto -\frac{1}{2}(\boldsymbol{\eta}_d - \boldsymbol{\mu}_d)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\eta}_d - \boldsymbol{\mu}_d) + \left(\sum_v c_{d,v} \log \sum_k \beta_{k,v} e^{\eta_{d,k}} - W_d \log \sum_k e^{\eta_{d,k}}\right), \quad (8)$$

where $c_{d,v}$ is the count of the vth term in the vocabulary within the dth document and W_d is the total count of words in the document. We optimize the objective with quasi-Newton methods using the gradient

$$\nabla f(\boldsymbol{\eta}_d)_k = \left(\sum_v c_{d,v}\langle \phi_{d,v,k}\rangle\right) - W_d\theta_{d,k} - \left(\Sigma^{-1}(\boldsymbol{\eta}_d - \boldsymbol{\mu}_d)\right)_k, \quad (9)$$

where $\boldsymbol{\theta}_d$ is the simplex mapped version of $\boldsymbol{\eta}_d$ and we define the expected probability of observing a given topic-word as $\langle \phi_{d,v,k}\rangle = (\frac{\exp(\eta_{d,k})\beta_{d,v,k}}{\sum_k \exp(\eta_{d,k})\beta_{d,v,k}})$. This gives us our variational posterior $q(\boldsymbol{\eta}_d) = \mathcal{N}(\boldsymbol{\lambda}_d = \hat{\boldsymbol{\eta}}_d, \boldsymbol{\nu}_d = -\nabla^2 f(\hat{\boldsymbol{\eta}}_d)^{-1})$. We then solve for $q(\mathbf{z}_d)$ in closed form,

$$\phi_{d,n,k} \propto \exp(\lambda_{d,k})\beta_{d,k,w_n}. \quad (10)$$

M-Step. In the M-step, we update the coefficients in the topic prevalence model, topical content model, and the global covariance matrix.

The prior on document-topic proportions maximizes the approximate ELBO with respect to the document-specific mean $\mu_{d,k} = \mathbf{X}_d\boldsymbol{\gamma}_k$ and the topic covariance matrix $\boldsymbol{\Sigma}$. Updates for $\boldsymbol{\gamma_k}$ correspond to linear regression for each topic under the user specified prior with $\boldsymbol{\lambda}_k$ as the outcome variable. By default we give the $\boldsymbol{\gamma}_k$ a Normal$(0, \sigma_k^2)$ where σ_k^2 is either manually selected or given a broad inverse-gamma prior. We also provide an option to estimate $\boldsymbol{\gamma}_k$ using an L_1 penalty.

The matrix $\boldsymbol{\Sigma}$ is then estimated as the convex combination of the maximum likelihood estimation (MLE) and a diagonalized form of the MLE,

$$\hat{\boldsymbol{\Sigma}}_{\text{MLE}} = \frac{1}{D}\sum_d \boldsymbol{\nu}_d + (\boldsymbol{\lambda}_d - \mathbf{X}_d\hat{\boldsymbol{\Gamma}})(\boldsymbol{\lambda}_d - \mathbf{X}_d\hat{\boldsymbol{\Gamma}})^T$$
$$\hat{\boldsymbol{\Sigma}} = w_\Sigma(\text{diag}(\hat{\boldsymbol{\Sigma}}_{\text{MLE}})) + (1 - w_\Sigma)(\hat{\Sigma}_{\text{MLE}}), \quad (11)$$

where the weight $w_\Sigma \in [0, 1]$ is set by the user and we default to zero.

Updates for the topic-word distributions correspond to estimation of the coefficients ($\boldsymbol{\kappa}$) in a multinomial logistic regression model where the observed words are the output, and the design matrix includes the expectations of the word-level topic assignments $E[q(\mathbf{z}_d)] = \boldsymbol{\phi}_d$, topical content covariates Y_d, and their interactions. The intercept $\mathbf{m}$ is fixed to be empirical log probability of the terms in the corpus. (See the Appendix for details.)

Remarks on Inference. Much progress on the analysis of behavior of the inference task in mixed membership models has been accomplished in the past few years. A thread of research in applied statistics has explored the properties of the inference task in mixed membership models, empirically, for a number of model variants (see, e.g., Pritchard et al. 2000; Blei, Ng, and Jordan 2003; Erosheva, Fienberg, and Lafferty 2004; Braun and McAuliffe 2010). While, from a theoretical perspective, mixed membership models similar to the one we consider in this article suffer from multiple symmetric modes in the likelihood defining an equivalence class of solutions (see, e.g., Stephens 2000; Buot and Richards 2006; Airoldi et al. 2014b), a number of successful solutions exist to mitigate the issue in practice, such as using multiple starting points, clever initialization, and procrustes transforms to identify and estimate a canonical element of the equivalence class of solutions (Hurley and Cattell 1962; Wallach et al. 2009b). The takeaway from these articles, which report extensive empirical evaluations of the inference task in mixed membership models, is that inference is expected to have good frequentist properties. More recently, a few articles have been able to analyze theoretical properties of the inference task (Mukherjee and Blei 2009; Tang et al. 2014; Nguyen 2015). These articles essentially show that inference on the mixed membership vectors has good frequentist properties, thus providing a welcome confirmation of the earlier empirical studies, but also conditions under which inference is expected to behave well.

While exactly characterizing the theoretical complexity of the optimization problem is beyond the scope of this article, we note that inference even in simple topic models has been shown to be NP-hard (Arora, Ge, and Moitra 2012). In the next section, we carry out an extensive empirical evaluation, including a frequentist coverage analysis, in scenarios that closely resemble real data, and a comparative performance analysis with state-of-the-art methods, in out-of-sample experiments on real data. These evaluations provide confidence in the results and conclusions we report in the case study. An important component of our strong performance in these setting is the use of an initialization strategy based on the spectral method of moments algorithm of Arora et al. (2013). We describe this approach and compare its performance to a variety of alternatives in Roberts, Stewart, and Tingley (2016).

2.2.2. *Interpretation*

After fitting the model, we are left with the task of summarizing the topics in an interpretable way (Chang et al. 2009). The majority of topic models are summarized by the most frequent terms within a topic, although there are several methods for choosing higher order phrases (Mei, Shen, and Zhai 2007; Blei and Lafferty 2009). Instead, here we use a metric to summarize topics that combines term frequency and exclusivity to that topic into a univariate summary statistic referred to as FREX (Bischof and Airoldi 2012; Airoldi and Bischof in press). This statistics calculates the harmonic mean of the empirical CDF of a term's frequency under a topic with the empirical CDF of exclusivity to that topic. Denoting the $K \times V$ matrix of topic-conditional term probabilities as $\mathbf{B}$, the FREX statistic is defined as

$$\text{FREX}_{k,v} = \left(\frac{\omega}{\text{ECDF}(\beta_{k,v}/\sum_{j=1}^{K}\beta_{j,v})} + \frac{1-\omega}{\text{ECDF}(\beta_{k,v})}\right)^{-1},$$

where ω is a weight, which balances the influence of frequency and exclusivity, which we set to 0.5. The harmonic mean ensures that chosen terms are both frequent and exclusive, rather than simply an extreme on a single dimension. We use a plugin estimator for the FREX statistics using the collection $\{\mathbf{B}\}$ coefficients estimated using variational EM.

3. Empirical Evaluation and Data Analysis

In this section, we demonstrate that our proposed model is useful with a combination of simulation evidence and an example application in political science. From a social science perspective, we are interested in studying how media coverage of China's rise varies between mainstream Western news sources and the Chinese state-owned news agency, Xinhua. We use the STM on a corpus of newswire reports to analyze the differences in both topic prevalence and topical content across five major news agencies.

Before proceeding to our application, we present series of simulation studies. In Section 3.1, we start with a very simple simulation that captures the intuition of why we expect the model to be useful in practice. This section also lays the foundation for our simulation procedures. In Section 3.2, we demonstrate that the model is able to recover parameters of interest in a more complicated simulation setting that closely parallels our real data. In Section 3.3, we further motivate our applied question and present our data. Using the China data we perform a held-out likelihood comparison to three competing models (Section 3.3.1) and check model fit using posterior predictive checks (Section 3.3.2). Finally having validated the model through simulation, held-out experiments and model checking, we present our results in Section 3.3.3.

3.1. Estimating Nonlinear Covariate Effects

In this simulation, we build intuition for why including covariate information into the topic model is useful for recovering trends in topical prevalence. We compare STM with Latent Dirichlet Allocation (LDA) using a very simple data-generating process that generates 100 documents using three topics and a single continuous covariate. We start by drawing the topic word distributions for each topic $\boldsymbol{\beta}_k \sim \text{Dirichlet}_{49}(0.05)$. Collecting the topic word distributions into the 3 by 50 matrix $\mathbf{B}$, each document is simulated by sampling: $N_d \sim \text{Pois}(50)$, $x_d \sim \text{Uniform}(0, 1)$, $\boldsymbol{\theta}_d \sim \text{LogisticNormal}_2(\boldsymbol{\mu} = (0.5, \cos(10x_d)), \boldsymbol{\Sigma} = 0.5\mathbf{I})$, and $w_{d,n} \sim \text{Multinomial}(\mathbf{B}\boldsymbol{\theta}_d)$, where we have omitted the token level latent variable $\mathbf{z}$ to reduce sampling variance.

We simulate from this data generating 50 times. For each simulated dataset, we fit an LDA model using collapsed Gibbs sampling and an STM model. For both cases, we use the correctly specified number of topics. For STM, we specify the model with the covariate x_d for each document using a B-spline with 10 degrees of freedom. Crucially we do not provide it any information about the true functional form. LDA cannot use the covariate information.

Interpreting the simulation results is complicated due to posterior invariance to label switching. For both LDA and STM, we match the estimated topics to the simulated parameters using the Hungarian algorithm to maximize the dot product of the true $\boldsymbol{\theta}$ and the MAP estimate (Papadimitriou and Steiglitz 1998; Hornik 2005).

In Figure 2, we plot the Loess-smoothed (span = 1/3) relationship between the covariate and the MAP estimate for $\boldsymbol{\theta}_d$ of the second topic. Each line corresponds to one run of the model and the true relationship is depicted with a thick black line. For comparison, the third panel shows the case using the true values of $\boldsymbol{\theta}$. While the fits based on the LDA model vary quite widely, the proposed model fits essentially all 50 samples with a recognizable representation of the true functional form. This is in some sense not at all surprising, the proposed model has access to valuable information about the covariate that LDA does not incorporate. The result is a very favorable bias-variance tradeoff in which our prior produces a very mild bias in the estimate of the covariate effects in return for a substantial variance reduction across simulations.

This simulation demonstrates that STM is able to capture a nonlinear covariate effect on topical prevalence. The focus here on the document-topic proportions ($\boldsymbol{\theta}_d$) differs from prior work in computer science, which typically focuses on the recovery of the topic-word distributions ($\boldsymbol{\beta}_k$). Recovery of $\boldsymbol{\beta}_k$ is an easier task in the sense that the parameters are global and our estimates can be expected to improve as the number of documents increases (Arora et al. 2013). By contrast $\boldsymbol{\theta}_d$ is a document level parameter where it makes less sense to speak of the number of words increasing toward infinity. Nevertheless, estimates of covariate relationships based on the document level parameters $\boldsymbol{\theta}_d$ are often the primary focus for applied social scientists and thus we emphasize them here.

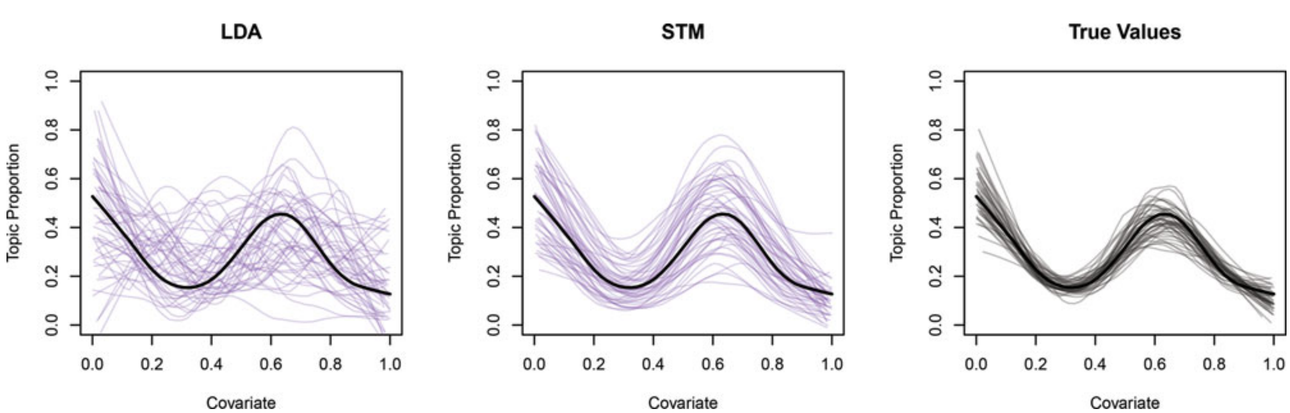

Figure 2. Plot of fitted covariate-topic relationships from 50 simulated datasets using LDA and the proposed structural topic model of text. The third panel shows the estimated relationship using the true values of the topic and thus only reflects sampling variability in the data-generating process.

3.2. Frequentist Coverage Evaluation in a Realistic Setting

In this section, we expand the quantitative evaluation of the proposed model to a more complex and realistic setting. Using the fitted model from the application in Section 3.3.3 as a reference, we simulate synthetic data from the estimated model parameters. The simulated dataset includes 11,980 documents, a vocabulary of $V = 2518$ terms, $K = 100$ topics, and covariates for both topic prevalence and topical content. We set the true values of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ to the MAP estimates of the reference model and simulate new observed words as above. We then fit the model to the synthetic documents using the same settings (and observed covariates) as we did in estimating the reference model. We repeat this process 100 times, and, as above, align the topics to the reference model using the Hungarian algorithm. This is a substantially more rigorous test of the inference procedure. With 100 topics, a content covariate with 5 levels and 2518 vocabulary terms, there are over 1.2 million topic-word probabilities that need to be estimated. The documents themselves are on average 167 words long, and for each one of them over 100 topic proportions need to be estimated.

We evaluate the simulations by examining the frequentist coverage of the credible interval for $\boldsymbol{\theta}$ and the expected error between the MAP estimate and the truth. The most straightforward method for defining credible intervals for $\boldsymbol{\theta}$ is using the Laplace approximation to the unnormalized topic proportions $\boldsymbol{\eta}$. By simulating draws from the variational posterior over $\boldsymbol{\eta}$ and applying the softmax transformation, we can recover the credible intervals for $\boldsymbol{\theta}$. However, this procedure poses a computational challenge as the covariance matrix $\boldsymbol{\nu_d}$, which is of dimension $K - 1 \times K - 1$, cannot easily be stored for each document, and recalculating $\boldsymbol{\nu_d}$ can be computationally unfeasible. Instead, we introduce a simpler global approximation of the covariance matrix $\boldsymbol{\nu_d}$, which leverages the MLE of the global covariance matrix $\boldsymbol{\Sigma}$

$$\tilde{\boldsymbol{\nu}} = \hat{\Sigma} - (\boldsymbol{\lambda}_d - \mathbf{X_d}\hat{\boldsymbol{\Gamma}})(\boldsymbol{\lambda}_d - \mathbf{X_d}\hat{\boldsymbol{\Gamma}})^T = \frac{1}{D}\sum_d \boldsymbol{\nu}_d. \qquad (12)$$

The approximation $\tilde{\boldsymbol{\nu}}$ equals the sample average of the estimated document-specific covariance matrices $\{\boldsymbol{\nu}_d\}$. Under this approximation, it is still necessary to simulate from the multivariate Normal variational posterior, but there are substantial computational gains from avoiding the need to recalculate the covariance matrix for each document. As we show next, this approximation yields credible intervals with good coverage properties. To summarize, for each document we simulate 2500 draws from the variational posterior $\mathcal{N}(\boldsymbol{\lambda}_d, \hat{\boldsymbol{\nu}}_d)$ using the document-specific variational mode λ_d and the global approximation to the covariance matrix $\tilde{\nu}$. We then apply the softmax transformation to these draws and recover the 95% credible interval of $\boldsymbol{\theta}_d$. We calculate coverage along each topic separately.

The left panel of Figure 3 shows boxplots of the coverage rates grouped by size of the true θ with the dashed line indicating the nominal 95% coverage. We can see that for very small values of θ (<0.05) and moderate to large values (>0.15). coverage is extremely close to the nominal 95% level. The observed discrepancies between empirical and nominal coverage are reasonable. There are several sources of variability that contribute to these deviations. First, the variational posterior is conditional on the point estimates of the topic-word distributions $\hat{\beta}$, which are estimated with error. Many of the documents are quite short relative to the total number of topics, thus the accuracy of the Laplace approximation may suffer. Finally, the optimization procedure only finds a local optimum.

Next we consider how well the MAP estimates of θ compare to the true values. The right panel of Figure 3 provides a series of boxplots of the expected L_1 error grouped by the true θ. For very small values of θ, the estimates are extremely accurate, and

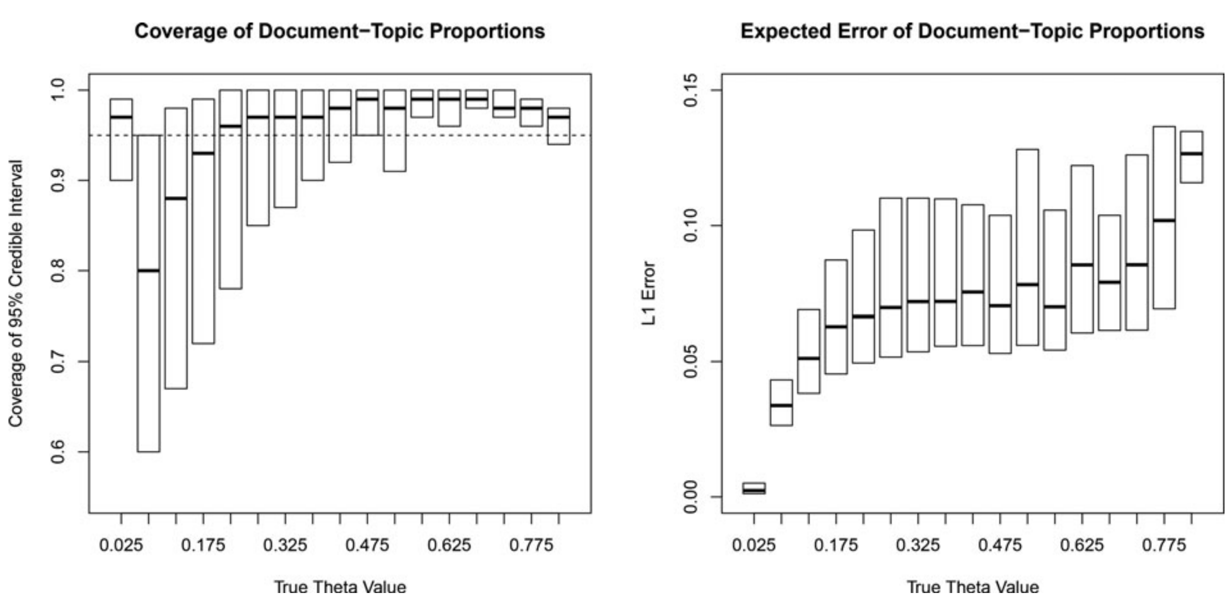

Figure 3. Coverage rates for a 95% credible interval on the document-topic proportions ($\boldsymbol{\theta}_d$) in a simulated $K = 100$ topic model. The left panel shows the distribution of coverage rates on a nominal 95% credible interval grouped by the size of the true $\boldsymbol{\theta}_d$. The right panel shows the distribution of the L_1 errors, $E[|\boldsymbol{\theta}_d - \hat{\boldsymbol{\theta}}_d|]$, where the $\hat{\boldsymbol{\theta}}_d$ is the MAP estimate.

the size of the errors grows little as the true parameter value increases. For very large values of θ, there is a small, but persistent, negative bias that results in underestimation of the large elements of θ.

This simulation represents a challenging case but the model performs well. Additional simulation results can be found in Roberts et al. (2014b) including a permutation style test for topical prevalence covariate effects in which a covariate is randomly permuted and the model is repeatedly reestimated. This can help the analyst determine if there is a risk of overfitting in reported covariate effects. Next, we validate the model using real data.

3.3. Media Coverage of China's Rise

Over the past decade, "rising" China has been a topic of conversations, news sources, speeches, and lengthy books. However, what rising China means for China, the West and the rest of the world is subject to much intense debate (Ikenberry 2008; Ferguson 2010). Tellingly, both Western countries and China accuse each other of slanting their respective medias to obfuscate the true quality of Chinese governance or meaning of China's newfound power (Fang 2001; Johnston and Stockmann 2007). Western "slant" and Chinese censorship and propaganda have been blamed for polarizing views among the American and Chinese public (Roy 1996; Johnston and Stockmann 2007), possibly increasing the probability of future conflict between the two countries.

In Section 3.3, we study both Western and Chinese media slant about China's rise through a collection of newspapers containing the word China over a decade of its development. We give a brief analysis of how different media agencies have characterized China's rise, focusing particularly on key differences in the way the Chinese news agency, Xinhua, represents and covers news topics differently than mainstream Western sources. In doing so, we seek to measure "slant" on a large scale. Proceeding this substantive analysis, in Section 3.3.1 we first show the extent to which our model leads to better prediction out-of-sample than existing models on the data, and the extent to which the proposed model fits the data (using posterior predictive checks).

To explore how different news agencies have treated China's rise differently, we analyze a stratified random sample (Rosenbaum 1987) of 11,980 news reports containing the term "China" dated from 1997–2006 and originating from five different international news sources. For each document in our sample, we observe the day it was written and the newswire service publishing the report. Our data include five news sources: Agence France Presse (AFP), the Associated Press (AP), British Broadcasting Corporation (BBC), Japan Economic Newswire (JEN), and Xinhua (XIN), the state-owned Chinese news agency. We include the month a document was written and the news agency as covariates on topical prevalence. We also include news agency as a covariate affecting topical content to estimate how topics are discussed in different ways by different news agencies. In our case study, we estimated the number of topics to be 100, by evaluating and maximizing topics' coherence using a cross-validation scheme while changing the number of topics (Airoldi et al. 2010).

3.3.1. Comparative Performance Evaluation with State-of-the-Art

To provide a fully automated comparison of our model to existing alternatives, we estimate the heldout likelihood using the document completion approach (Asuncion et al. 2009; Wallach et al. 2009b). To demonstrate that the covariates provide useful predictive information, we compare the proposed structural topic model (STM) to latent Dirichlet allocation (LDA), the Dirichlet multinomial regression topic model (DMR), and the sparse additive generative text model (SAGE). We use a measure of predictive power to evaluate comparative performance among these models: for a subset of the documents we hold back half of the document and evaluate the likelihood of the held out words (Asuncion et al. 2009; Paisley, Wang, and Blei 2012). Higher numbers indicate a more predictive model.

Figure 4 shows the heldout likelihood for a variety of topic values. We show two plots. On the left is the average heldout likelihood for each model on 100 datasets, and their 95% quantiles. At first glance, in this plot, it seems that STM is doing much better or about the same as the other three models. However, looking at the second plot, the paired differences between

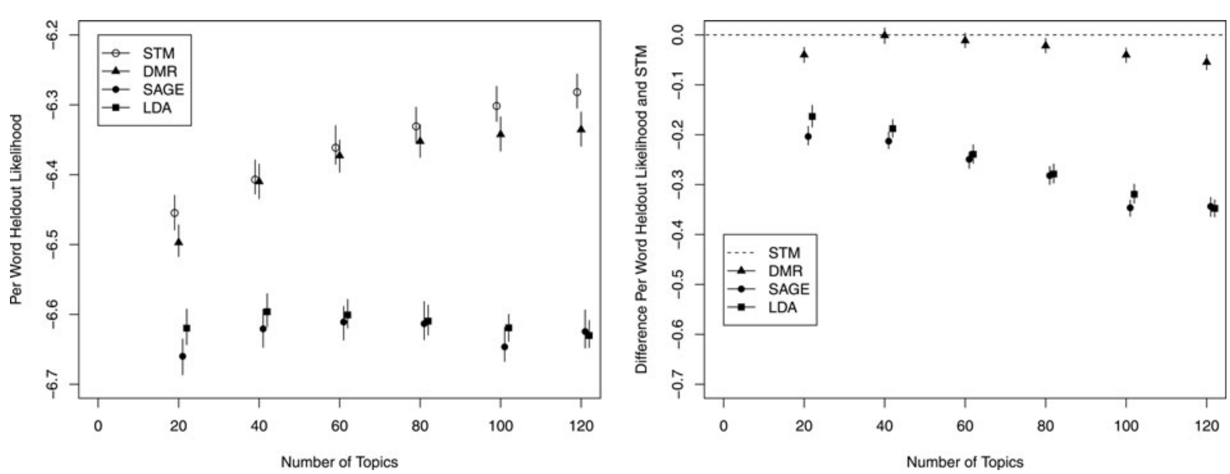

Figure 4. STM versus SAGE, LDA, and DMR heldout likelihood comparison. On the left is the mean heldout likelihood and 95% quantiles. On the right is the mean paired difference between the three comparison models and STM.

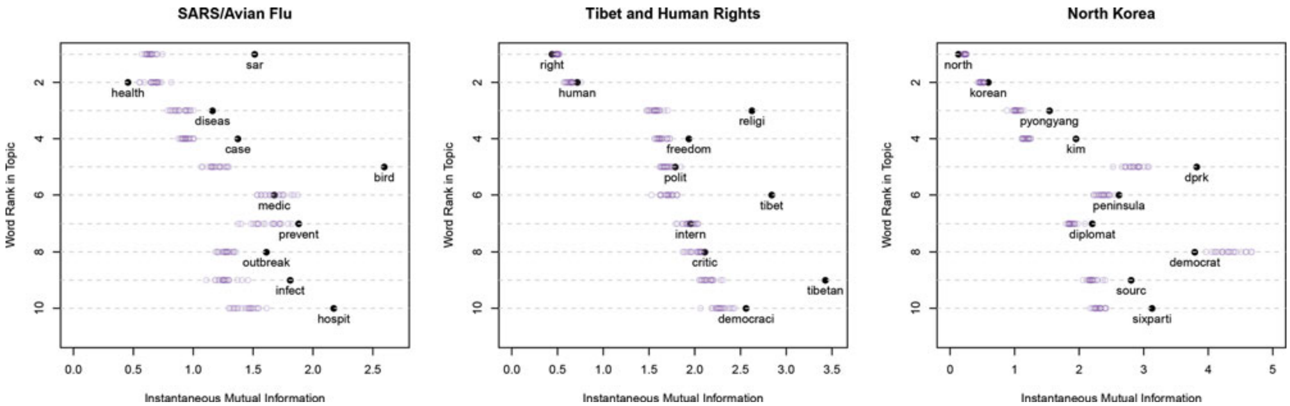

Figure 5. Posterior predictive checks using the methodology outlined in Mimno and Blei (2011). The plot shows the top ten most probable words for each of three topics marginalizing over the covariate-specific word distributions. The x-axis gives the instantaneous mutual information that would be 0 in the true data-generating process. The black closed circle gives the observed value.

the models on each individual dataset, we see that STM consistently outperforms all other models when the models are run on the same dataset. With the exception of the 40 topic run, STM does better than all models in every dataset for every topic number. Focusing on paired comparison suggests that STM is the preferred choice for prediction.

The main takeaway from this table is that STM performs significantly better than competing models, except for the case of 40 topics, when it has comparable predictive ability to Dirichlet multinomial regression model. This suggests that including information on topical prevalence and topical content aids in prediction. Further, STM has more interpretable quantities of interest than its closest competitor because it allows correlations between topics and covariates on topic content. We cover these qualitative advantages in the next section.

3.3.2. Assessing Model Fit

The most effective method for assessing model fit is to carefully read documents that are closely associated with particular topics to verify that the semantic concept covered by the topic is reflected in the text. The parameter $\boldsymbol{\theta}_d$ provides an estimate of each document's association with every topic making it straightforward to effectively direct analyst engagement with the texts (see, e.g., Roberts, Stewart, and Tingley 2014a). An overview of manual validation procedures can be found in Grimmer and Stewart (2013).

When automated tools are required, we can use the framework of posterior predictive checks to assess components of model fit (Gelman, Meng, and Stern 1996). Mimno and Blei (2011) outlined a framework for posterior predictive checks for the latent Dirichlet allocation model using mutual information between document indices and observed words as the realized discrepancy function. Under the data-generating process, knowing the document index would provide us no additional information about the terms it contains after conditioning on the topic. In practice, topical words often have heavy tailed distributions of occurrence, thus we may not expect independence to hold (Doyle and Elkan 2009).

As in Mimno and Blei (2011), we operationalize the check using the instantaneous mutual information between words and document indices conditional on the topic: $\text{IMI}(w, D|k) = H(D|k) - H(D|W = w, k)$, where D is the document index, w is the observed word, k is the topic, and $H()$ is the entropy. When the model assumptions hold, we expect this quantity to be close to zero because the entropy of each word should be the same as the entropy of the topic distribution under the data-generating process. To provide a reference for the observed value, we plot the value for the top 10 words for three different topics along with 20 draws from the simulated posterior predictive distribution (Gelman, Meng, and Stern 1996; Mimno and Blei 2011).

Figure 5 gives an example of these checks for three topics. The posterior predictive checks give us an indication of where the model assumptions do and do not hold. Cases where there is a large gap between the observed value (dark circle) and the reference distribution (open circles) indicate cases where the model assumptions do not hold. Generally, these discrepancies occur for terminology, which is specific to a sub-component of the topic. For example, in the left plot on SARS/avian flu, the two terms with the greatest discrepancies are the word stems for “SARS” and “bird.” The distribution of occurrence for these terms would naturally be heavy-tailed, in the sense that once we have observed the occurrence of the term in a document, the likelihood observing it again would increase. A model that splits SARS and avian flu into separate topics would be unlikely to have this problem. However, for our purposes here combining them into one topic is not a problem.

3.3.3. Substantive Analysis of Differential Newswire Reporting

While it is useful to demonstrate that STM shows predictive gains, our primary motivation in developing the STM is to create a tool that can help us answer social science questions. Specifically, we want to study how the various news sources cover topics related to the last 10 years of China's development and the vocabulary with which these newswires describes the same events. We are interested in how Chinese and Western sources represent prominent international events during this time period differently, that is, describe the same event with different vocabulary, and the differences between how much Chinese and Western sources discuss a particular topic. Accusations of “slant” have been largely anecdotal, and the STM provides us with a unique opportunity to measure characterizations of news about China on a large scale.

AFP	parti communist leader polit deng rule leadership power mao call former democrat central reform death chang zhao china xiaop late democraci public zedong show tiananmen
AP	parti communist polit leader power reform mao rule leadership revolut deng corrupt chang china decad social central former though stand keep one said zhao open
BBC	parti ccp communist central polit build leadership work democrat secretari socialist social general must advanc cultur modern reform china leader mass three think serv studi
JEN	parti democrat polit communist rule leader lawmak former liber leadership secretari opposit post public power late reform chang china coalit formal main central serv select
XIN	parti cpc central communist secretari work social polit general deng socialist leadership reform xiaop build call democrat theori china open modern forward movement characterist mao

Figure 6. China trajectory topic. Each group of words are the highest probability words for the news source.

For the purposes of the following analyses, we labeled each topic individually by looking at the most frequent words and at the most representative articles. We start with a general topic related to Chinese governance, which includes Chinese government strategy, leadership transitions, and future policy. We might call this a "China trajectory" topic. Figure 6 shows the highest probability of words in this topic for each of the news sources. The news sources have vastly different accounts of China's trajectory. AFP and AP talk about China's rule with words like "Tiananmen," referring to the 1989 Tiananmen student movement, and "Zhao," referring to the reformer Zhao Ziyang who fell out of power during that incident due to his support of the students. Even though Tiananmen occurred 10 years before our sample starts, these Western news sources discuss it as central to China's current trajectory.

Xinhua, on the other hand, has a more positive view of China's direction, with words like "build" and "forward," omitting words like "corrupt" or mentions of the Tiananmen crackdown. Interestingly, the BBC and JEN also have a forward-looking view on China's trajectory, discussing "reform,""advancing," and references to the formation of laws in China. The analysis provides clear evidence of varying perspectives in both Western and Chinese sources on China's political future, and surprisingly shows significant variation within Western sources.

Second, we turn to a very controversial event within China during our time period, the crackdown on Falungong. Falungong is a spiritual group that became very popular in China during the 1990s. Due to the scale and organization of the group, the Chinese government outlawed Falungong beginning in 1999, arresting followers, and dismantling the organization.

This topic appears within all of our news sources, since the crackdown occurred within the time period we are studying. Figure 7 (left panel) shows the different ways in which the news sources portray the Falungong incident. Again, we see that the AP and AFP have the most "Western" view of the incident, using words like "dissident," "crackdown," and "activist." The BBC, on the other hand, takes a much milder language to talk about the incident, with words such as "illegal," or "according." JEN talks a lot about asylum for those fleeing China, with words such as "asylum," "refugee," and "immigration." Xinhua, on the other hand, talks about the topic using exclusively language about crime, for example, "crime," "smuggle," "suspect," and "terrorist." Again, we see not only the difference between Western and Chinese sources, but interestingly large variation in language within Western sources.

Since we included news source as a covariate in estimating topical prevalence part within the model, we can estimate the differences in frequency, or how much each of the news sources discussed the Falungong topic. As shown in Figure 7 (right panel), we see unsurprisingly that Xinhua talks significantly less about the topic than Western news sources. This would be unsurprising to China scholars, but reassuringly agrees with expectations. Interestingly, the Western news sources we would identify to have the most charged language, AFP and AP, also talk about the topic more. Slant has a fundamental relationship with topical prevalence, where those with a positive slant on China talk about negative topics less, and those with negative slant on China talk about negative topics more.

In general, our model picks up both short-lived events like the Olympics and invasion of Iraq, and long-term topical trends, such as discussion about North Korea and nuclear weapons over

AFP	polic arrest protest releas dissid detain author follow squar crackdown demonstr movement member tiananmen prison activist ban inform democraci church told china sinc tri wang
AP	polic protest gong arrest falun detain author releas follow prison inform offic wang crackdown activist movement sentenc center squar china demonstr zhang tiananmen dissid investig
BBC	polic offic author arrest investig public gong falun incid member charg illeg inform citizen china accord detain man consul day order document want regist protest
JEN	polic protest investig offic arrest author refuge consul famili asylum five enter letter suspect releas inform alleg china detain immigr yearold demonstr incid sent document
XIN	polic illeg smuggl public gong investig arrest offic crimin cult falun immigr suspect case custom organ author depart order china properti told accord sentenc terrorist

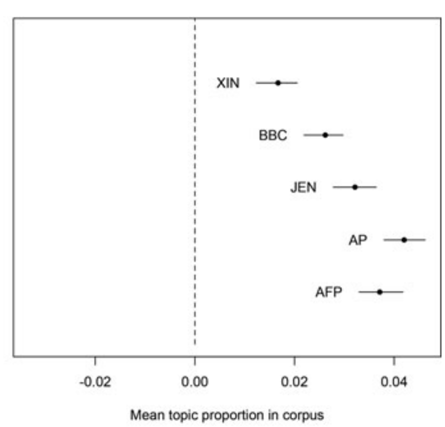

Figure 7. Falungong topic. Each group of words are the highest probability words for the news source (left panel). Mean prevalence of Falungong topic within each news source corpus (right panel).

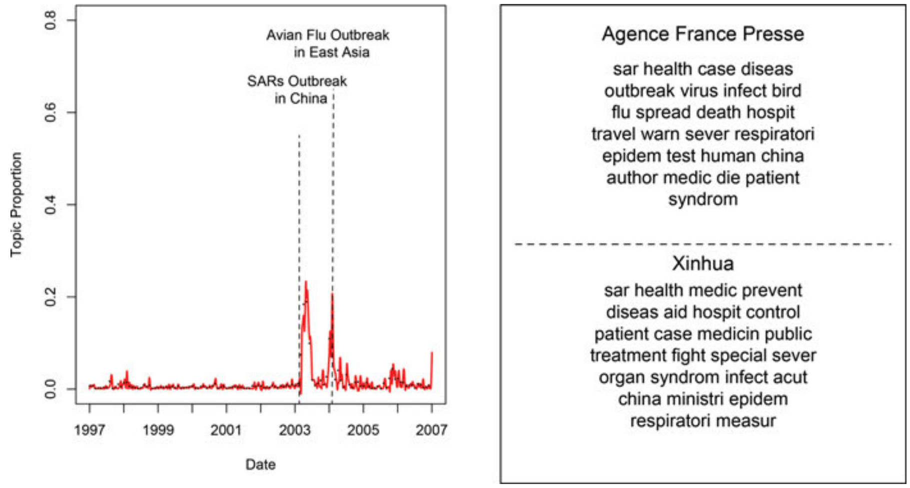

Figure 8. SARS and avian flu. Each dot represents the average topic proportion in a document in that month and the line is a smoothed average across time (left panel). Comparisons between news sources (right panel).

time and discussion of the environment, both increasing over time.

As an illustration, we turn to the differing news coverage of SARS during the outbreak of the disease during 2003 in China. First, in Figure 8 (left panel) we show that by smoothing over time, our model is able to capture the SARS and subsequent avian flu events, described above. The topic model shows how the news both quickly picked up outbreaks of SARS and avian flu and quickly stopped talking about them when the epidemics were resolved. The Chinese government received a lot of

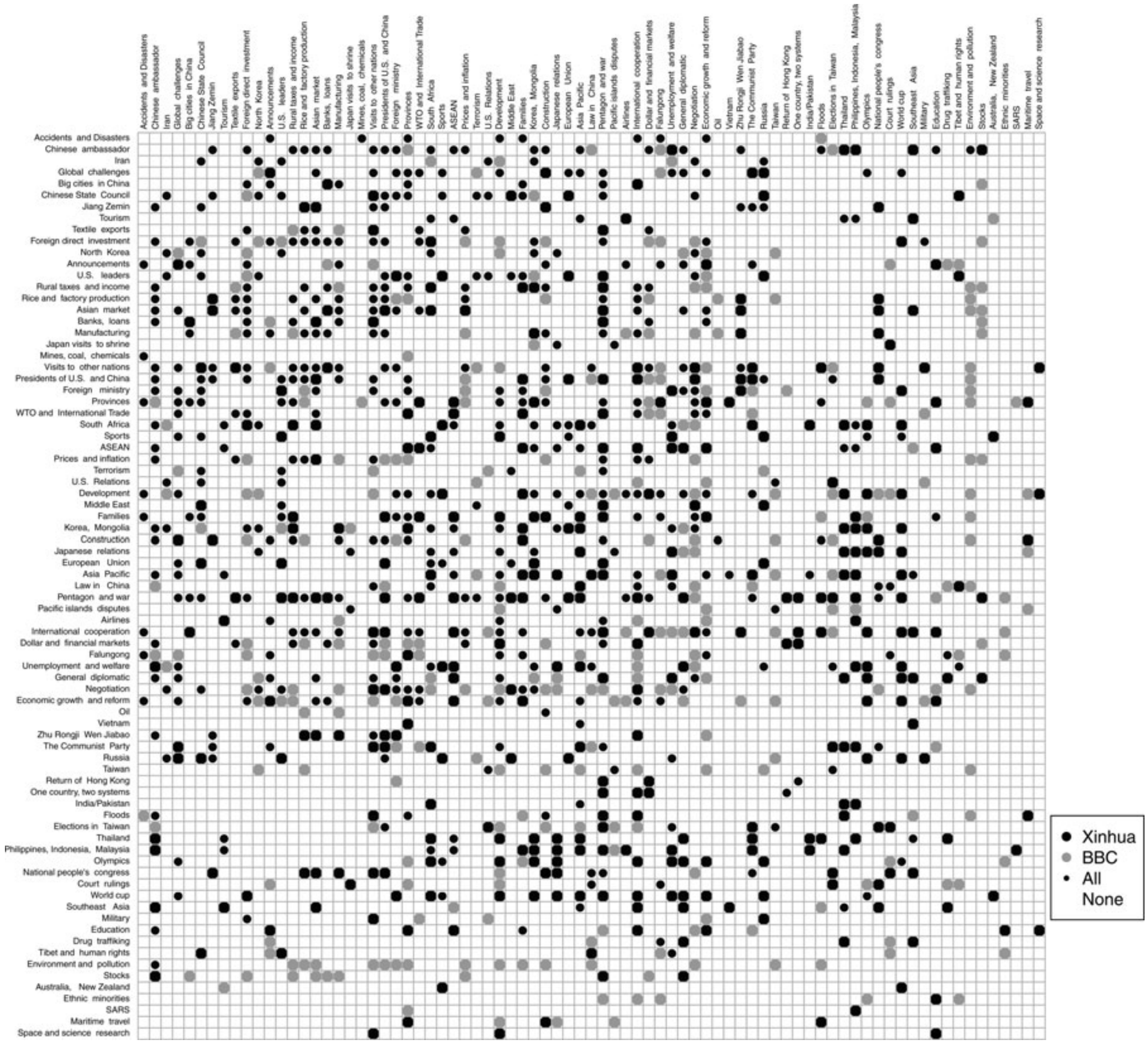

Figure 9. Correlation between topics, Xinhua versus BBC.

international criticism for its news coverage of SARS, mostly because it reported on the disease much later than it knew that the epidemic was occurring. As shown in Figure 8 (right panel), our model picks up small differences in news coverage between Chinese and Western sources once news coverage began happening, although not substantial. In particular, while Western news sources seemed to talk a lot about death, Chinese news sources mainly focused on policy-related words, such as "control," "fight," and "aid," and avoided mentions of death by the disease.

Finally, because the model allows for the inclusion of correlated topics, we can also visualize the relationship between China-related topics in the 1997–2006 period. In particular, we can see how topics are correlated differently for different news wires, indicating how topics are connected and framed differently in each newswire. In Figure 9, we find all edges between topics where they exhibit a positive correlation above 0.1. Pairs of topics where an edge exists in both Xinhua and BBC, we denote with a light blue dot. Pairs of topics where Xinhua, but not BBC have an edge between them, we denote with a red square, and those where BBC, but not Xinhua have an edge between them we denote with a blue square.

We then sort the matrix by topics that are similarly correlated with other topics in Xinhua and BBC to those that are not similarly correlated. Topics such as accidents and disasters, tourism, factory production, and manufacturing are correlated with similar topics in both BBC and Xinhua. However, topics related to democracy, human rights, and the environment have different topic correlations between the two corpuses. For example, in BBC, the environment and pollution is correlated with factory production, construction, development, and families. In Xinhua, on the other hand, environment and pollution is only correlated with the topic associated with the Chinese ambassador, meaning it is mainly talked about in articles related to international relations, rather than internal economic development.

In conclusion, the STM allows us to measure how of the various newswire services differentially treat China's rise over a 10 year period. We see much variation in how the different newswires discuss the rise of China. Unsurprisingly, Xinhua news services talk about negative topics less than Western sources, and focus on the positive aspects of topics related to China's rise. Interestingly, however, we see high variation within Western news sources, with AFP and AP taking a much more negative slant on China's rise than the BBC. We believe we are the first to quantify media slant in news sources all over the world on China's rise, adding to the discussion of how China is perceived across many different countries. Conveniently, the STM allows us to summarize the newspapers' perspectives more quickly than would reading large swaths of text, the method that is currently most frequently used by China scholars.

4. Concluding Remarks

In this article, we have outlined a new mixed membership model for the analysis of documents with meta-information. We also have outlined some of the features of the proposed models, which are important for analyzing experiments and for carrying out other causal analyses when the outcome comes in the form of text data. We then demonstrated the proposed methods to address questions about the variation in news coverage of China's rise. In related work, we have applied these methods to study open-ended survey responses (Roberts et al. 2014b), comparative politics literature (Lucas et al. 2015), and student-generated text in massive open online courses (Reich et al. 2015).

We conclude by highlighting some areas of work that would be fruitful for expanding the role of this type of analysis, especially in the social sciences.

A productive line of inquiry has focused on the interpretation of topic models (Chang et al. 2009; Mimno et al. 2011; Airoldi and Bischof in press). These methods are aided by techniques for dealing with the practical threats to interpretation such as excessive stop-words and categories with overlapping keywords (Wallach, Mimno, and McCallum 2009a; Zou and Adams 2012). In addition to fully automated approaches, work on interactive topic modeling and user-specified constraints is particularly appropriate to social scientists who may have a deep knowledge of their particular document sets (Ramage et al. 2009; Andrzejewski et al. 2011; Hu, Boyd-Graber, and Satinoff 2011). One advantage of our approach is that the meta-information is incorporated by means of generalized linear models, which are already familiar to social scientists.

A second area we want to emphasize is the recent work on general methods for evaluation and model checking (Wallach et al. 2009b; Mimno and Blei 2011; Airoldi and Bischof in press). As noted in both the computer science literature (Blei 2012) and the political science literature (Grimmer and Stewart 2013), validation of the model becomes even more important when using unsupervised methods for *inference* or *measurement* than it is when used for prediction or exploration. While model-based fit statistics are an important part of the process, we also believe that recent work in the automated visualization of topic models (Chaney and Blei 2012; Chuang et al. 2012b; Chuang, Manning, and Heer 2012a) is of equal or greater importance for helping users to substantively engage with the underling texts. And user engagement is important to ultimately deliver interesting substantive conclusions (Grimmer and King 2011).

Alternative inference strategies for the proposed model, and for topic models generally, are an area of current research. With regard to our model, an alternative inference approach would be to develop an Markov chain Monte Carlo (MCMC) sampler based on the polya-gamma data augmentation scheme (Chen et al. 2013; Polson, Scott, and Windle 2013). This has the advantage of retaining asymptotic guarantees on recovering the true posterior. However, while MCMC get to the right answer in theory, in the limit, in practice they also get stuck in local modes, and they often converge slower than variational approaches. A second approach would be to explore techniques based on stochastic approximations (Toulis and Airoldi 2014, 2015). This has the advantage of providing a solution, which scales well to larger collections of documents, while retaining the asymptotic properties of MCMC. Elsewhere, we have also developed a strategy to appropriately include measurement uncertainty from the variational posterior in regressions where the latent topic is used as the outcome variable (Roberts et al. 2014b).

Software availability. The R package `stm` implements the methods described here, in addition to a suite of visualization and post-estimation tools (*http://cran.r-project.org/web/packages/stm*).

Supplementary Materials

The supplementary materials contain the article's appendix. Additionally, replication materials are available on Dataverse and can be found at *http://dx.doi.org/10.7910/DVN/SIGIAU*.

Acknowledgments

The authors thank Ryan Adams, Ken Benoit, David Blei, Patrick Brandt, Amy Catalinac, Sean Gerrish, Adam Glynn, Justin Grimmer, Gary King, Christine Kuang, Chris Lucas, Brendan O'Connor, Arthur Spirling, Alex Storer, Hanna Wallach, Daniel Young, and in particular Dustin Tingley, for useful discussions, and the editor and two anonymous reviewers for their valuable input. The first two authors contributed equally to this work.

Funding

This research supported, in part, by The Eunice Kennedy Shriver National Institute of Child Health & Human Development under grant P2-CHD047879 to the Office of Population Research at Princeton University, by the National Science Foundation under grants CAREER IIS-1149662, and IIS-1409177, and by the Office of Naval Research under grant YIP N00014-14-1-0485 to Harvard University. This research was largely performed when Brandon M. Stewart was a National Science Foundation Graduate Research Fellow at Harvard University. Edoardo M. Airoldi is an Alfred P. Sloan Research Fellow. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH, of the NSF, nor of the ONR.

References

Ahmed, A., and Xing, E. (2007), "On Tight Approximate Inference of the Logistic-Normal Topic Admixture Model," in *Proceedings of AISTATS*, pp. 19–26. [992]

——— (2010), "Staying Informed: Supervised and Semi-Supervised Multi-View Topical Analysis of Ideological Perspective," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, pp. 1140–1150. [989]

Airoldi, E., Erosheva, E., Fienberg, S., Joutard, C., Love, T., and Shringarpure, S. (2010), "Reconceptualizing the Classification of PNAS Articles," *Proceedings of the National Academy of Sciences*, 107, 20899–20904. [988,996]

Airoldi, E. M., Anderson, A. G., Fienberg, S. E., and Skinner, K. K. (2006), "Who Wrote Ronald Reagan's Radio Addresses?" *Bayesian Analysis*, 1, 289–320. [988,991]

Airoldi, E. M., and Bischof, J. M. (in press), "A Regularization Scheme on Word Occurrence Rates That Improves Estimation and Interpretation of Topical Content" (with discussion), *Journal of American Statistical Association*. [993,1000]

Airoldi, E. M., Blei, D. M., Erosheva, E. A., and Fienberg, S. E. (eds.) (2014a), *Handbook of Mixed Membership Models and Their Applications (Handbooks of Modern Statistical Methods)*, Boca Raton, FL: Chapman & Hall/CRC. [988,989]

——— (2014b), "Introduction to Mixed Membership Models and Methods," in *Handbook of Mixed Membership Models and Their Applications (Handbooks of Modern Statistical Methods)*, Boca Raton, FL: Chapman & Hall/CRC, pp. 3–14. [993]

Airoldi, E. M., Cohen, W. W., and Fienberg, S. E. (2005), "Bayesian Models for Frequent Terms in Text," in *Proceedings of the Classification Society of North America and INTERFACE Annual Meetings*. [990,991]

Airoldi, E. M., and Fienberg, S. E. (2003), "Who Wrote Ronald Reagan's Radio Addresses?" Technical Report no. CMU-STAT-03-789, Carnegie Mellon University. [988]

Aitchison, J. (1982), "The Statistical Analysis of Compositional Data," *Journal of the Royal Statistical Society*, Series B, 44, 139–177. [991]

Aitchison, J., and Shen, S. M. (1980), "Logistic-Normal Distributions: Some Properties and Uses," *Biometrika*, 67, 261–272. [989,991]

Andrzejewski, D., Zhu, X., Craven, M., and Recht, B. (2011), "A Framework for Incorporating General Domain Knowledge Into Latent Dirichlet Allocation using First-Order Logic," in *IJCAI*, pp. 25–32. [1000]

Arora, S., Ge, R., Halpern, Y., Mimno, D., Moitra, A., Sontag, D., Wu, Y., and Zhu, M. (2013), "A Practical Algorithm for Topic Modeling With Provable Guarantees," in *Proceedings of The 30th International Conference on Machine Learning*, pp. 280–288. [993,994]

Arora, S., Ge, R., and Moitra, A. (2012), "Learning Topic Models–Going Beyond Svd," in *IEEE 53rd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 1–10. [993]

Asuncion, A., Welling, M., Smyth, P., and Teh, Y.-W. (2009), "On Smoothing and Inference for Topic Models," in *UAI*, pp. 27–34. [996]

Bischof, J. M., and Airoldi, E. M. (2012), "Summarizing Topical Content With Word Frequency and Exclusivity," in *International Conference on Machine Learning* (Vol. 29), pp. 201–208. [993]

Blei, D. M. (2012), "Probabilistic Topic Models," *Communications of the ACM*, 55, 77–84. [988,1000]

Blei, D. M., and Lafferty, J. D. (2006), "Dynamic Topic Models," in *ICML*, 113–120. [990]

——— (2007), "A Correlated Topic Model of Science," *Annals of Applied Statistics*, 1, 17–35. [989,991,992]

——— (2009), "Topic Models," in *Text Mining: Classification, Clustering, and Applications*, eds. A. N. Srivastava and M. Sahami, Boca Raton, FL: CRC Press, pp. 71–94. [993]

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2001), "Latent Dirichlet allocation," in *Advances in Neural Information Processing Systems*, pp. 601–608. [988]

——— (2003), "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, 3, 993–1022. [988,992,993]

Braun, M., and McAuliffe, J. (2010), "Variational Inference for Large-Scale Models of Discrete Choice," *Journal of the American Statistical Association*, 105, 324–335. [993]

Buot, M., and Richards, D. (2006), "Counting and Locating the Solutions of Polynomial Systems of Maximum Likelihood Equations, I," *Journal of Symbolic Computation*, 41, 234–244. [993]

Chaney, A., and Blei, D. (2012), *Visualizing Topic Models*, Princeton, NJ: Department of Computer Science, Princeton University. [1000]

Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S., and Blei, D. M. (2009), "Reading Tea Leaves: How Humans Interpret Topic Models," in *NIPS*, 288–296. [993,1000]

Chen, J., Zhu, J., Wang, Z., Zheng, X., and Zhang, B. (2013), "Scalable Inference for Logistic-Normal Topic Models," in *Advances in Neural Information Processing Systems*, pp. 2445–2453. [992,1000]

Chuang, J., Manning, C., and Heer, J. (2012a), "Termite: Visualization Techniques for Assessing Textual Topic Models," in *Proceedings of the International Working Conference on Advanced Visual Interfaces*, ACM, pp. 74–77. [1000]

Chuang, J., Ramage, D., Manning, C., and Heer, J. (2012b), "Interpretation and Trust: Designing Model-Driven Visualizations for Text Analysis," in *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, ACM, pp. 443–452. [1000]

De Boor, C., De Boor, C., De Boor, C., and De Boor, C. (1978), *A Practical Guide to Splines* (Vol. 27), New York: Springer-Verlag. [991]

Dempster, A. P., Laird, N., and Rubin, D. B. (1977), "Maximum Likelihood Estimation From Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Association*, 39, 1–38. [989]

Doyle, G., and Elkan, C. (2009), "Accounting for Burstiness in Topic Models," in *ICML*, 281–288. [997]

Eisenstein, J., Ahmed, A., and Xing, E. (2011), "Sparse Additive Generative Models of Text," in *Proceedings of ICML*, pp. 1041–1048. [990,991]

Eisenstein, J., O'Connor, B., Smith, N., and Xing, E. (2010), "A Latent Variable Model for Geographic Lexical Variation," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, pp. 1277–1287. [990]

Erosheva, E. A., Fienberg, S. E., and Lafferty, J. (2004), "Mixed-Membership Models of Scientific Publications," *Proceedings of the National Academy of Sciences*, 101, 5220–5227. [988,993]

Fang, Y.-J. (2001), "Reporting the Same Events? A Critical Analysis of Chinese Print News Media Texts," *Discourse & Society*, 12, 585–613. [996]

Ferguson, N. (2010), "Complexity and Collapse: Empires on the Edge of Chaos," *Foreign Affairs*, 89, 18–32. [996]

Friedman, J., Hastie, T., and Tibshirani, R. (2010), "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of statistical software*, 33, 1–22. [992]

Gelman, A., and Hill, J. (2007), *Data Analysis Using Regression and Multilevel/Hierarchical Models* (Vol. 3), New York: Cambridge University Press. [991]

Gelman, A., Meng, X.-L., and Stern, H. (1996), "Posterior Predictive Assessment of Model Fitness via Realized Discrepancies," *Statistica Sinica*, 6, 733–760. [989,997]

Genkin, A., Lewis, D. D., and Madigan, D. (2007), "Large-Scale Bayesian Logistic Regression for Text Categorization," *Technometrics*, 49, 291–304. [988]

Gerrish, S., and Blei, D. (2012), "How They Vote: Issue-Adjusted Models of Legislative Behavior," in *Advances in Neural Information Processing Systems*, 25, 2762–2770. [989]

Griffiths, T. L., and Steyvers, M. (2004), "Finding Scientific Topics," *Proceedings of the National Academy of Sciences of the United States of America*, 101, 5228–5235. [988,992]

Grimmer, J. (2010), "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases," *Political Analysis*, 18, 1–35. [988,989]

Grimmer, J., and King, G. (2011), "General Purpose Computer-Assisted Clustering and Conceptualization," *Proceedings of the National Academy of Sciences*, 108, 2643–2650. [1000]

Grimmer, J., and Stewart, B. M. (2013), "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts," *Political Analysis*, 21, 267–297. [988,997,1000]

Groenewald, P. C., and Mokgatlhe, L. (2005), "Bayesian Computation for Logistic Regression," *Computational Statistics & Data Analysis*, 48, 857–868. [992]

Hastie, T., and Tibshirani, R. (1990), *Generalized Additive Models* (Vol. 43), Boca Raton, FL: Chapman & Hall/CRC. [991]

Holmes, C. C., and Held, L. (2006), "Bayesian Auxiliary Variable Models for Binary and Multinomial Regression," *Bayesian Analysis*, 1, 145–168. [992]

Hornik, K. (2005), "A CLUE for CLUster Ensembles," *Journal of Statistical Software*, 14, 1–25. [994]

Hu, Y., Boyd-Graber, J. L., and Satinoff, B. (2011), "Interactive Topic Modelingin," in *ACL*, pp. 248–257. [1000]

Hurley, J., and Cattell, R. (1962), "The Procrustes Program: Producing Direct Rotation to Test a Hypothesized Factor Structure," *Behavioral Science*, 7, 258–262. [993]

Ikenberry, G. J. (2008), "The Rise of China and the Future of the West: Can the Liberal System Survive?" *Foreign Affairs*, 87, 23–37. [996]

Jeske, D. R., and Liu, R. Y. (2007), "Mining and Tracking Massive Text Data: Classification, Construction of Tracking Statistics, and Inference Under Misclassification," *Technometrics*, 49, 116–128. [988]

Jia, J., Miratrix, L., Yu, B., Gawalt, B., El Ghaoui, L., Barnesmoore, L., and Clavier, S. (2014), "Concise Comparative Summaries (CCS) of Large Text Corpora With a Human Experiment," *Annals of Applied Statistics*, 8, 499–529. [988]

Johnston, A. I., and Stockmann, D. (2007), *Chinese Attitudes Toward the United States and Americans*, Ithaca, NY: Cornell University Press, pp. 157–195. [996]

Khan, M. E., and Bouchard, G. (2009), "Variational EM Algorithms for Correlated Topic Models," Technical Report, University of British Columbia. [993]

Knowles, D. A., and Minka, T. (2011), "Non-Conjugate Variational Message Passing for Multinomial and Binary Regression," in *Advances in Neural Information Processing Systems*, pp. 1701–1709. [992]

Liu, J. S. (1994), "The Collapsed Gibbs Sampler in Bayesian Computations With Applications to a Gene Regulation Problem," *Journal of the American Statistical Association*, 89, 958–966. [989]

Lucas, C., Nielsen, R., Roberts, M. E., Stewart, B. M., Storer, A., and Tingley, D. (2015), "Computer Assisted Text Analysis for Comparative Politics," *Political Analysis*, 23, 254–277. [989,1000]

Manning, C. D., Raghavan, P., and Schütze, H. (2008), *An Introduction to Information Retrieval*, Cambridge, England: Cambridge University Press. [988]

Mei, Q., Shen, X., and Zhai, C. (2007), "Automatic Labeling of Multinomial Topic Models," in *KDD*, pp. 490–499. [993]

Mendenhall, T. C. (1887), "The Characteristic Curves of Composition," *Science*, 11, 237–249. [988]

Meng, X.-L., and Van Dyk, D. (1997), "The EM Algorithm—An Old Folk-Song Sung to a Fast New Tune," *Journal of the Royal Statistical Society*, Series B, 59, 511–567. [989]

Miller, G. A., Newman, E. B., and Friedman, E. A. (1958), "Length-Frequency Statistics for Written English," *Information and Control*, 1, 370–389. [988]

Mimno, D., and Blei, D. (2011), "Bayesian Checking for Topic Models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, pp. 227–237. [997,1000]

Mimno, D., and McCallum, A. (2008), "Topic Models Conditioned on Arbitrary Features With Dirichlet-Multinomial Regression," in *UAI*, pp. 411–418. [990]

Mimno, D., Wallach, H., and McCallum, A. (2008), "Gibbs Sampling for Logistic Normal Topic Models With Graph-Based Priors," in *NIPS Workshop on Analyzing Graphs*, pp. 1–8. [992]

Mimno, D., Wallach, H., Talley, E., Leenders, M., and McCallum, A. (2011), "Optimizing Semantic Coherence in Topic Models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, pp. 262–272. [1000]

Minka, T., and Lafferty, J. D. (2002), "Expectation-Propagation for the Generative Aspect Model," in *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 352–359. [988]

Mosteller, F., and Wallace, D. (1963), "Inference in an Authorship Problem," *Journal of the American Statistical Association*, 58, 275–309. [988]

Mosteller, F., and Wallace, D. L. (1964), *Inference and Disputed Authorship: The Federalist*, Reading, MA: Addison-Wesley. [988]

——— (1984), *Applied Bayesian and Classical Inference: The Case of The Federalist Papers*, New York: Springer-Verlag. [988,991]

Mukherjee, I., and Blei, D. M. (2009), "Relative Performance Guarantees for Approximate Inference in Latent Dirichlet Allocation," in *Neural Information Processing Systems*, pp. 1129–1136. [993]

Nguyen, X. (2015), "Posterior Contraction of the Population Polytope in Finite Admixture Models," *Bernoulli*, 21, 618–646. [993]

Paisley, J., Wang, C., and Blei, D. (2012), "The Discrete Infinite Logistic Normal Distribution," *Bayesian Analysis*, 7, 235–272. [996]

Papadimitriou, C. H., and Steiglitz, K. (1998), *Combinatorial Optimization: Algorithms and Complexity*, Mineola, NY: Courier Dover Publications. [994]

Polson, N. G., Scott, J. G., and Windle, J. (2013), "Bayesian Inference for Logistic Models Using Polya-Gamma Latent Variables," *Journal of the American Statistical Association*, 108, 1339–1349. [992,1000]

Pritchard, J. K., Stephens, M., Rosenberg, N. A., and Donnelly, P. (2000), "Association Mapping in Structured Populations," *American Journal of Human Genetics*, 67, 170–181. [993]

Quinn, K., Monroe, B., Colaresi, M., Crespin, M., and Radev, D. (2010), "How to Analyze Political Attention With Minimal Assumptions and Costs," *American Journal of Political Science*, 54, 209–228. [988,989]

Ramage, D., Hall, D., Nallapati, R., and Manning, C. D. (2009), "Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-Labeled Corpora," in *EMNLP*, pp. 248–256. [1000]

Reich, J., Tingley, D., Leder-Luis, J., Roberts, M. E., and Stewart, B. M. (2015), "Computer Assisted Reading and Discovery for Student Generated Text in Massive Open Online Courses," *Journal of Learning Analytics*, 2, 156–184. [1000]

Roberts, M. E., Stewart, B. M., and Tingley, D. (2014a), "stm: R Package for Structural Topic Models," Technical Report, Harvard University. [989,997]

——— (2016), "Navigating the Local Modes of Big Data: The Case of Topic Models," *Computational Social Science: Discovery and Prediction*, ed. R. M. Alvarez, New York: Cambridge University Press, pp. 51–97. [993]

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S., Albertson, B., and Rand, D. (2014b), "Structural Topic Models for Open-Ended Survey Responses," *American Journal of Political Science*, 58, 1064–1082. [989,996,1000]

Rosenbaum, P. R. (1987), "Model-Based Direct Adjustment," *Journal of the American Statistical Association*, 82, 387–394. [996]

Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P. (2004), "The Author-Topic Model for Authors and Documents," in *UAI*, pp. 487–494. [990]

Roy, D. (1996), "The "China Threat" Issue: Major Arguments," *Asian Survey*, 36, 758–771. [996]

Socher, R., Gershman, S., Perotte, A., Sederberg, P., Blei, D., and Norman, K. (2009), "A Bayesian Analysis of Dynamics in Free Recall," *Advances in Neural Information Processing Systems*, 22, 1714–1722. [988]

Stephens, M. (2000), "Bayesian Analysis of Mixtures With an Unknown Number of Components—An Alternative to Reversible Jump Methods," *Annals of Statistics*, 28, 40–74. [993]

Taddy, M. (2013), "Multinomial Inverse Regression for Text Analysis," *Journal of the American Statistical Association*, 108, 755–770. [988]

——— (2015), "Distributed Multinomial Regression," *Annals of Applied Statistics*, 9, 1394–1414. [990]

Tang, J., Meng, Z., Nguyen, X., Mei, Q., and Zhang, M. (2014), "Understanding the Limiting Factors of Topic Modeling via Posterior Contraction Analysis," *Journal of Machine Learning Research*, 32, 190–198. [993]

Toulis, P., and Airoldi, E. M. (2014), "Implicit Stochastic Gradient Descent," *arxiv no. 1408.2923*. Available at *https://arxiv.org/abs/1408.2923* [1000]

——— (2015), "Scalable Estimation Strategies Based on Stochastic Approximations: Classical Results and New Insights," *Statistics and Computing*, 25, 781–795. [1000]

Turney, P. D., and Pantel, P. (2010), "From Frequency to Meaning: Vector Space Models of Semantics," *Journal of Artificial Intelligence Research*, 37, 141–188. [988]

Wallach, H., Mimno, D., and McCallum, A. (2009a), "Rethinking LDA: Why Priors Matter," in *NIPS*, pp. 1973–1981. [1000]

Wallach, H., Murray, I., Salakhutdinov, R., and Mimno, D. (2009b), "Evaluation Methods for Topic Models," in *ICML*, pp. 1105–1112. [993,996,1000]

Wang, C., and Blei, D. (2013), "Variational Inference in Nonconjugate Models," *Journal of Machine Learning Research*, 14, 1005–1031. [989,992]

Yule, U. (1944), *The Statistical Study of Literary Vocabulary*, Cambridge, UK: Cambridge University Press. [988]

Zipf, G. K. (1932), *Selected Studies of the Principle of Relative Frequency in Language*, Cambridge, MA: Harvard University Press. [988]

Zou, J., and Adams, R. (2012), "Priors for Diversity in Generative Latent Variable Models," *Advances in Neural Information Processing Systems*, 25, 3005–3013. [1000]

Online supplement for: "A model of text for experimentation in the social sciences"

A Estimation of topic prevalence and content

Optimization of coefficients governing the topical content and prevalence models are dependent on the choice of priors.

Topic prevalence coefficients are given a Gaussian prior with topic-specific variance. These variances are given conjugate priors such that,

$$\gamma_{p,k} \sim \text{Normal}(0, \sigma_k^2) \tag{1}$$

$$\sigma_k^2 \sim \text{Inverse-Gamma}(1, 1) \tag{2}$$

where p indexes the covariates in the design matrix X. We leave the intercept unpenalized and compute the posterior mean of $\gamma_{p,k}$ using standard variational linear regression (Drugowitsch 2013, e.g.). The $\boldsymbol{\gamma}_k$ update for a given penalty parameter takes the form of a penalized linear regression

$$\hat{\gamma}_k = \left(X^T X + \text{diag}(1/\sigma_k^2)\right)^{-1} X^T \lambda_k \tag{3}$$

and the update for the variance is

$$\hat{\sigma_k^2} = \left(.5 + \sum_p \gamma_{p,k}^2\right) / (.5 + p)$$

We iterate between the coefficients and the shared variance until convergence.

In our software we also provide the option to estimate γ using the `glmnet` package (Friedman et al. 2010) which allows for penalized estimation using the lasso or elastic-net. This can be particularly useful for high-dimensional factor variables where many of the topic-effects are expected to be exactly zero. In this case the full regularization path is estimated and the value of the penalty parameter is selected using a modified information criterion (Taddy 2013b).

Estimation of the topical content covariate coefficients κ is equivalent to a multinomial logistic regression on the token level latent variables ϕ. In order to induce sparsity in κ we assign Laplace priors and compute the posterior mode using the lasso. In order to make the procedure more computationally efficient we adopt the distributed multinomial regression approach of Taddy (2013a). The idea is to use a plugin estimator for the document fixed effects which decouples the parameters of the single multinomial logistic regression into independent poisson regressions (one for each element of the vocabulary). This approach is not only faster but also allows for the operations to be parallelized over the vocabulary. The regularization parameter controlling sparsity is chosen using a modified information criterion as described in Taddy (2013b,a).

We also provide the option of estimating the topical content coefficients using a Normal-Jeffreys prior such that $\kappa_{k,v}$ $\sim$Normal$(0, \tau_{k,v})$ and $\tau_{k,v} \sim 1/\tau_{k,v}$. The prior for τ is the improper Jeffreys prior. Here estimation involves alternating between maximization of κ and τ. Following Eisenstein et al. (2011) we use a block relaxation approach where each V-length vector in κ_k is updated using quasi-Newton methods, followed by an update of the variances. Taking as an example the vector for topic k we obtain the objective and gradient

$$\mathcal{L}_{\boldsymbol{\kappa}_k} = \langle \mathbf{c_k} \rangle \boldsymbol{\kappa}_k - \langle C_k \rangle \log \sum_v \exp(\kappa_{k,v} + m_v)) - \frac{1}{2} \boldsymbol{\kappa}_k^2 / \boldsymbol{\tau_k} \tag{4}$$

$$\nabla \mathcal{L}_{\boldsymbol{\kappa}_k} = \langle \mathbf{c_k} \rangle - \sum_j \langle C_{jk} \rangle \boldsymbol{\beta_{jk}} - \boldsymbol{\kappa_k} / \boldsymbol{\tau_k} \tag{5}$$

where $\langle c_k \rangle$ is V-length vector of expected counts for each term in the vocabulary for topic k. C_k is the summation over that vector producing a scalar equal to the expected number of words assigned to topic k. Updates for each covariate and interaction proceed analogously.

After each update of $\boldsymbol{\kappa}$ we update the corresponding penalty vector $\boldsymbol{\tau}$ with its variational expectation $\hat{\tau}_{v,k} = \kappa_{v,k}^2$.

References

Drugowitsch, J. (2013). Variational bayesian inference for linear and logistic regression. *arXiv preprint arXiv:1310.5438*.

Eisenstein, J., Ahmed, A., and Xing, E. (2011). Sparse additive generative models of text. In *Proceedings of ICML*, pages 1041–1048.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

Taddy, M. (2013a). Distributed multinomial regression. *arXiv preprint arXiv:1311.6139*.

Taddy, M. (2013b). The gamma lasso. *arXiv preprint arXiv:1308.5623*.