The Structural Topic Model and Applied Social Science*

Margaret E. Roberts†
Department of Government
Harvard University
roberts8@fas.harvard.edu

Brandon M. Stewart†
Department of Government
Harvard University
bstewart@fas.harvard.edu

Dustin Tingley
Department of Government
Harvard University
dtingley@gov.harvard.edu

Edoardo M. Airoldi
Department of Statistics
Harvard University
airoldi@fas.harvard.edu

Abstract

We develop the Structural Topic Model which provides a general way to incorporate corpus structure or document metadata into the standard topic model. Document-level covariates enter the model through a simple generalized linear model framework in the prior distributions controlling either topical prevalence or topical content. We demonstrate the model's use in two applied problems: the analysis of open-ended responses in a survey experiment about immigration policy, and understanding differing media coverage of China's rise.

1 Topic Models and Social Science

Over the last decade probabilistic topic models, such as *Latent Dirichlet Allocation* (LDA), have become a common tool for understanding large text corpora [1].[1] Although originally developed for descriptive and exploratory purposes, social scientists are increasingly seeing the value of topic models as a tool for measurement of latent linguistic, political and psychological variables [2]. The defining element of this work is the presence of additional document-level information (e.g. author, partisan affiliation, date) on which variation in either topical *prevalence* or topical *content* is of theoretic interest.[2] As a practical matter, this generally involves running an off-the-shelf implementation of LDA and then performing a post-hoc evaluation of variation with a covariate of interest.

A better alternative to post-hoc comparisons is to build the additional information about the structure of the corpus into the model itself by altering the prior distributions to partially pool information amongst similar documents. Numerous special cases of this framework have been developed for particular types of corpus structure affecting both topic prevalence (e.g. time [3], author [4]) and topical content (e.g. ideology [5], geography [6]). Applied users have been slow to adopt these models because it is often difficult to find a model that exactly fits their specific corpus.

We develop the *Structural Topic Model* (STM) which accommodates corpus structure through document-level covariates affecting topical prevalence and/or topical content. The central idea is to

*Prepared for the NIPS 2013 Workshop on Topic Models: Computation, Application, and Evaluation. A forthcoming R package implements the methods described here.

† These authors contributed equally.

[1] We assume a general familiarity with LDA throughout (see [1] for a review)

[2] By "topical prevalence" we mean the proportion of document devoted to a given topic. By "topical content" we mean the rate of word use within a given topic.

specify the priors as generalized linear models through which we can condition on arbitrary observed data. This allows us to directly estimate the quantities of interest in applied problems. The model generalizes several existing approaches in the literature and, in conjunction with our forthcoming `R` package, allows users to incorporate the specific structure of their corpus without developing new models from scratch.

After describing the model, we demonstrate the use of STM to analyze two social science questions using open-ended responses from a survey experiment and an international newswire corpus.

2 The Structural Topic Model

The model (Figure 1) combines and extends three existing models: the correlated topic model (CTM) [7], the Dirichlet-Multinomial Regression (DMR) topic model [8] and the Sparse Additive Generative (SAGE) topic model [9]. The logistic normal prior on topical prevalence in the standard CTM is replaced by a logistic-normal linear model. The design matrix for the covariates X allows for arbitrarily flexible functional forms of the original covariates using radial basis functions (our `R` package also provides B-splines). The distribution over words is replaced with a multinomial logit such that a token's distribution is the combination of three effects (topic, covariates, topic-covariate interaction) operationalized as sparse deviations from a baseline word frequency (m). Our software provides the analyst with a choice of regularizing priors for the GLM coefficients (κ, γ) with defaults: Normal-Gamma prior pooled by topic for γ and the "Gamma Lasso" prior [10] for κ.

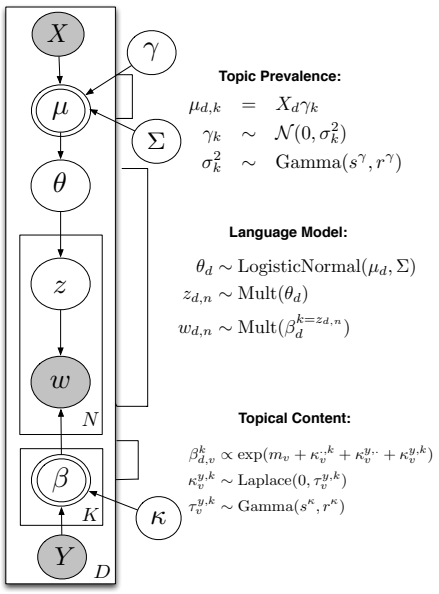

Figure 1: Plate Diagram for the Structural Topic Model

2.1 Posterior Inference and Quantities of Interest

We estimate the model using a fast semi-collapsed variational EM algorithm. For the nonconjugate logistic normal variables in the E-step we use a Laplace approximation [11]. We also integrate over the token-level latent variable z in order to speed convergence. The model directly estimates covariate effects which are analogous to GLM coefficients familiar to social scientists. To improve interpretation of the topics themselves, we assign labels using a variation of the Frequency-Exclusivity approach developed in [12] which we describe in [13].

2.2 Related Work and Other Ways to Include Information

Our approach to including corpus structure reflects our interest in making inference about *observed* covariates rather than predicting covariate values in unseen text. We briefly overview three other approaches to including document information, highlighting their different strengths. Supervised LDA [14] is designed towards a prediction task and assumes a generative model for a document-level variable. This finds a low-dimensional representation that both predicts words and the covariate. Partially Labeled LDA [15] allows the user to include prior information that particular documents are at least somewhat about certain topics. Finally Factorial LDA [16] has a similar mathematical setup to our model but focuses on latent covariates with an emphasis on interpretation.

3 Applications

In lieu of standard demonstrations based on predictive measures such as held-out likelihood or recovery of simulated parameters, we provide a short discussion of two research driven applications of our work. Simulations and held-out likelihood comparisons can be found in [13].

3.1 The STM for Open-ended Questions in Survey Experiments

Survey researchers often face a difficult to decision about whether to employ open-ended responses which can be costly to analyze by humans, or close-ended responses which require the *a priori* specification of possible responses. We argue that STM can substantially lower the costs of analyzing open-ended responses. We use open-ended survey responses collected by [17] who study how negatively valanced emotions affect political attitudes. In one design, they use a survey experiment to study how encouraging subjects to be worried about immigration influences their reaction to immigration policy. Using a $K = 3$ topic model we estimate the influence of the encouragement on the open-ended responses, conditioning on treatment and the stated party of the respondent. We find that the treatment makes respondents talk more about security and welfare concerns, while the control group stresses citizenship and the challenges immigrants face (Figure 2). The treatment effect is greater for Republicans than for Democrats, indicating that Republicans are more likely to respond to fear-provoking encouragements about immigration than Democrats. In the parlance of standard terminology, partisan ID moderates the effect of the treatment.

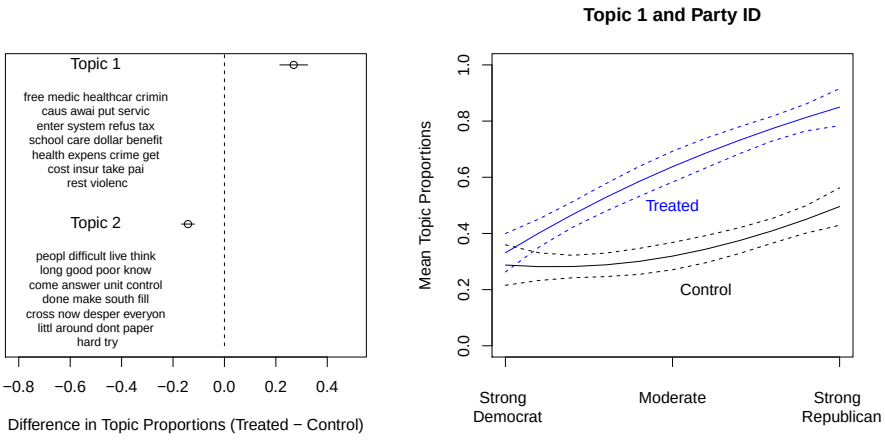

Figure 2: Party ID, Treatment, and the Predicted Proportion in Fear Topic (1 of 3)

3.2 How News Wires Describe China's Rise, 1997-2006

Next we demonstrate how STM can be applied to capture a source-time structure. Specifically, we investigate how China's rise is talked about differently by news wire services around the word. Comparing how the world views China to how the Chinese government presents itself to its own people is important in understanding patterns in public opinion in China versus the world at large and the Chinese government's own intentions [18]. We first collected news wire stories that included the word "China" from 1997 to 2006 from five major newswire services. We fit an $K = 80$ topic model allowing topic prevalence to vary by year and news wire source, and topical content to vary by news wire source. We find that the model captures important events and differences between newspapers' depictions of these events. For example, the model shows (Figure 3) increases in topic prevalence about the topic "Taiwan" during 2000 and 2004, capturing the Taiwanese presidential elections in these years. It also shows how the Associated Press describes these elections differently than the Chinese state-owned Xihua news source: while Xinhua uses words such as "one-china", "province", and "reunif" when talking about Taiwan, the AP uses words such as "elect", "democrat", and "vote". While the outside world views the Taiwanese elections through the lens of electoral competition, China presents the election in terms of its own interest: reunifying with Taiwan.

4 Discussion

We have developed the Structural Topic Model as a tool for applied social scientists to incorporate document meta-data into the topic modeling process. Using a familiar GLM framework the model

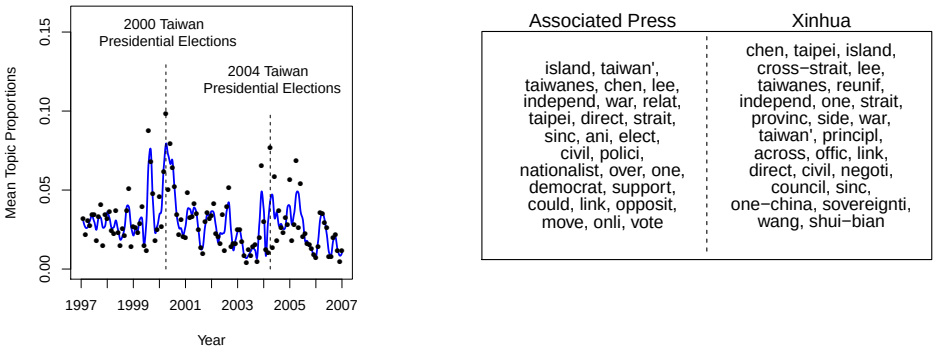

Figure 3: Taiwanese Presidential Election Topic (1 of 80) with news-source specific content (2 of 5)

allows for the incorporation of quite general corpus structure. This combined with our forthcoming `R` package, allows users to forgo the development of application-specific models. To illustrate the model's advantages we provide a sample of results from our existing work demonstrating the use of the model in analyzing surveys and news media.

References

[1] D. M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[2] J. Grimmer and B. M. Stewart. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297, 2013.

[3] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *ICML*, 2006.

[4] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *UAI*, 2004.

[5] A. Ahmed and E.P. Xing. Staying informed: supervised and semi-supervised multi-view topical analysis of ideological perspective. In *EMNLP*, pages 1140–1150, 2010.

[6] J. Eisenstein, B. O'Connor, N.A. Smith, and E.P. Xing. A latent variable model for geographic lexical variation. In *EMNLP*, pages 1277–1287, 2010.

[7] D. M. Blei and J. D. Lafferty. A correlated topic model of science. *AAS*, 1(1):17–35, 2007.

[8] D. Mimno and A. McCallum. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *UAI*, 2008.

[9] J. Eisenstein, A. Ahmed, and E. P. Xing. Sparse additive generative models of text. In *ICML*, pages 1041–1048, 2011.

[10] M. Taddy. Multinomial inverse regression for text analysis. *JASA*, 108(503):755–770, 2013.

[11] C. Wang and D. M. Blei. Variational inference in nonconjugate models. *JMLR*, 14:1005–1031, 2013.

[12] J. Bischof and E. M. Airoldi. Summarizing topical content with word frequency and exclusivity. In *ICML*, pages 201–208, 2012.

[13] M. E. Roberts, B. M. Stewart, D. Tingley, C. Lucas, J. Leder-Luis, S. Gadarian, B. Albertson, and D. Rand. Structural topic models for open-ended survey responses. *Am. Journal of Political Science*, Forthcoming.

[14] D. M. Blei and J. D. McAuliffe. Supervised topic models. In *NIPS*, 2007.

[15] D. Ramage, C. D. Manning, and S. Dumais. Partially labeled topic models for interpretable text mining. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 457–465. ACM, 2011.

[16] M. Paul and M. Dredze. Factorial lda: Sparse multi-dimensional text models. In *NIPS*, pages 2591–2599, 2012.

[17] S.K. Gadarian and B. Albertson. Anxiety, immigration, and the search for information. *Political Psychology*, 2013.

[18] G. King, J. Pan, and M. E. Roberts. How censorship in china allows government criticism but silences collective expression. *American Political Science Review*, 107:1–18, 2013.