Topic Models and Structure

Edo Airoldi, Molly Roberts and Brandon Stewart

Departments of Statistics and Government, Harvard University

Summary

1. What is the Project
   We allow the incorporation of arbitrary covariates to improve the estimation of unsupervised topic models.
2. Why Should I Read the Papers?
   There are two papers: one provides a genealogy of topic models aimed at a political science audience about how our model can help improve your project; the other provides the statistical and technical details about the design of the model, its properties and its implementation.
3. Why Should I Use the Model?
   Social scientists often have a great deal of information about their documents beyond the words in them. This meta-data can improve estimation and yield new insights about the nature of the way topics are discussed and their prevalence across different documents.

The Problem

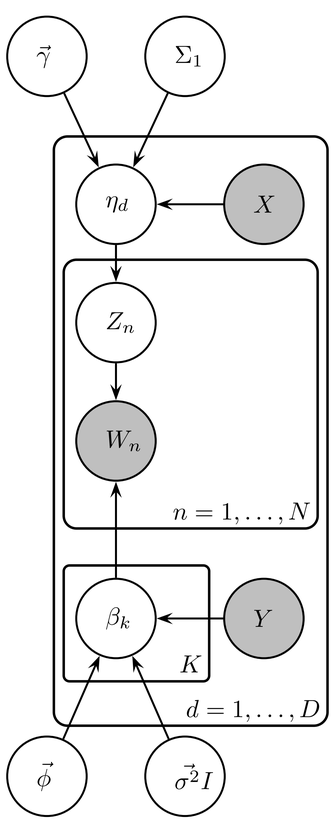
We are often interested in exploring large document corpora using unsupervised learning methods. Latent Dirichlet Allocation (Blei et. al, 2003) assumes exchangeability of documents. However, as political scientists we often know a great deal about the structure of our data in the form of covariates at the document level. We often want to explore how document topics vary over these covariates (e.g. how topic prevalence or topic word-use vary over time and space). Existing models that handle change over time (Quinn et, al 2010, Blei and Lafferty 2007) allow flexible change through evolution equations that require a forward-backward algorithm like the Kalman filter for inference.

Our Solution

We directly parametrize the topic proportions and distributions over vocabulary using a generalized linear model which can contain arbitrary covariates. By placing splines on the continuous covariates we are able to avoid blocking the covariates or making inference sequentially through the documents.

1. Draw $\vec{\eta}_d | X_d\gamma, \Sigma_1 \sim Normal(s(X_d)\gamma, \Sigma_1)$
2. Draw $\vec{\beta}_{d,k} | Y_d\phi, \sigma^2 I \sim Normal(s(Y_d)\phi, \sigma^2 I)$
3. For $n \in 1, \ldots, N_d$:
   - Draw topic assignment $Z_{d,n} | \vec{\eta}_d$ from Multi$(f(\eta_d))$
   - Draw word$W_{d,n} | \{z_{d,n}, \vec{\beta}_{1:K}\}$ from Mult$(\vec{\beta}_{z_{d,n}})$

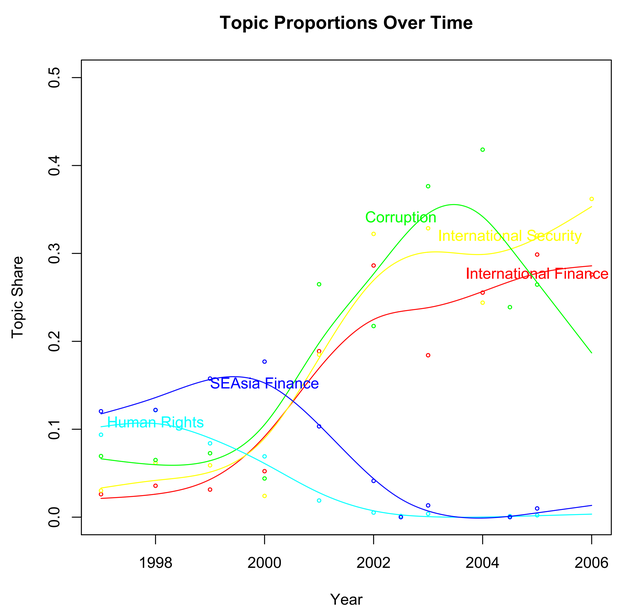
Our Model



An Example Application: The Rise of China

- Corpus: 1.2 million news stories about China dated 1997-2006 from 20 different news sources.
- QOI: Topic evolution over time, topic prevalence over time and how coverage differs by news source.
- Most existing models can only capture either the temporal structure or the authorship structure.

Early Results: Change in Topics Over Time



Human Rights
spiritual, dissident, gong, meditation, qigong, protest, custody, re-education, practitioners, movement, allegedly, arrested, oppose, court

International Security
security, detainees, asia-pacific, cooperation, petroleum, taiwan, scuffle, defending, japanese, north, korea, summit, traffickers, undisputed

International Finance
immigration, development, global, won, asean, cooperation, asia-pacific, venture, market, economy, factories, competitor

SE Asian Finance
confidence, debt, manila, bangkok, stock, recovery, investors, investment, management, traders, unrest, banks

Corruption
officials, bribed, disease,spent, condemning, relationship, patients, lawsuit, fired

Future Applications

- Catalinac: Japanese Party Manifestos (Structure: Year and Political Party)
- Stewart/Young: Constitutional Change in 1950s America (Structure: Day and News Source)

References

- D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993-1022, January 2003.
- D. Blei and J. Lafferty. A correlated topic model of Science. *Annals of Applied Statistics*, 1(1):17-35, 2007.
- J. Grimmer. A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis*, 18(1):1, 2010.
- K.M. Quinn, B.L. Monroe, M. Colaresi, M.H. Crespin D.R. Radev. How to analyze political attention with minimal assumptions and costs.*American Journal of Political Science*,54(1):209-228, 2010.

Future Work on the Model

- Create a fast easy-to-use R package
- Compare performance to alternative models such as the Dirichlet-Multinomial and supervised-LDA
- Conduct further studies on stability and consistency of the model
- Create simple, easy-to-understand interpretations of all model parameters

Acknowledgements

Our thanks to Ken Benoit, Dave Blei, Amy Catalinac, Adam Glynn, Justin Grimmer, and Arthur Spirling for comments, insights and suggestions. Thanks also to the National Science Foundation for a Graduate Research Fellowship to Brandon Stewart.