What is the Project

We allow the incorporation of arbitrary covariates to improve the estimation of unsupervised topic models.

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.

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 in topics over time, topic prevalence over time and how coverage differs by news source. Existing models can only capture either the temporal structure or the authorship structure.

Our Model

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.

Future Applications

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

References


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.