Comment on: *A regularization scheme on word occurrence rates that improves estimation and interpretation of topical content*

Matt Taddy
Microsoft Research and Chicago Booth
faculty.chicagobooth.edu/matt.taddy

This is an interesting and informative article by Bischof and Airoldi, and I welcome the opportunity to comment. The authors are focused on improving *interpretability* in statistical language modeling. This is an undeniably important part of applicability of these techniques, especially in the social sciences. In my collaborations with economists and others, huge effort has been spent pouring through the lists of words that are 'top' or representative within, say, latent topics from LDA (Blei et al., 2003) in order to build a narrative around the fitted language model. Often, these topics are not initially intuitive or self-consistent, and thus one begins iterating through a series of repeated model estimations using different vocabulary sets (e.g., via different stop-words or tokenization rules or minimum occurrence thresholds).

Such labor-intensive narrative cycles are especially frustrating because the goal is unclear. If we want to build lists of words that contain what we already understand as 'representative' of a given topic or feeling, we can simply do so using our own or other's expert opinion (e.g., Tetlock, 2007, is an example of such 'dictionary-based' analysis). If we simply want to choose the best fit for the data, we can use established model selection techniques (e.g., the Bayes factors for LDA in Taddy, 2012). But the desired outcome seems to live somewhere in-between: we want to find topics that are interpretable within existing concepts but which contain application-specific content and whose prevalence within the documents can be described as 'derived from the data'.

These difficulties seem inerrant to truly *unsupervised* topic analysis – when you don't have or are not interested in non-text document attributes, even for a subset of your corpora. But whenever such supervision is available, it can be used to guide estimation. For example, it is

common for topic modeling to be used as an effective tool for dimension reduction in a larger inferential pipeline. Topic weights within each document are downstream inputs to a regression function for some document attributes (this is an especially useful strategy when these attributes are only known for a small subset of your corpora: unsupervised dimension reduction on the full dataset makes it easier to fit regression on the small set of labeled data). In such settings, we can use left-out predictive performance in the downstream task as our arbitrator on topic quality. Alternatively, one can have the document attributes directly inform topic estimation. This is the strategy demonstrated nicely by Airoldi and Bischof in the current article. They use a known document classification as the basis for a hierarchical model of topic generation, specified in such a way that each topic has a well identified and sparse role in language choice. And it works! They provide word lists that are clearly intuitable and self consistent, without any of the usual steps of vocabulary narrowing.

In my own work, I have suggested that there are many scenarios where one can avoid altogether fitting a latent variable topic model and instead make use of standard high-dimensional regression techniques. Taddy (2013) describes a simple framework where the word counts within each document are treated as the response in a multinomial logistic regression onto document attributes. That article focuses on how one can derive sufficient reduction projections from the resulting model fit, and use of these projections in prediction tasks. Taddy (2015) shows how a closely related (and more scalable) version of the same algorithms can be used in a variety of text analysis in addition to the original prediction tasks: identifying words that are indicative of certain sentiment, subject, or even humor; projecting documents into a low dimensional space that represents the amount of funny or useful content; and as the first step in a pipeline where document projections serve as control variables in a causal inference scheme.

References

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent Dirichlet allocation. *the Journal of machine Learning research 3*, 993–1022.

Taddy, M. (2012). On estimation and selection for topic models. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS 2012)*.

Taddy, M. (2013). Multinomial inverse regression for text analysis. *Journal of the American Statistical Association 108*, 755–770.

Taddy, M. (2015). Distributed multinomial regression. *The Annals of Applied Statistics*. To appear.

Tetlock, P. (2007). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *Journal of Finance 62*, 1139–1168.