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 Airoldi and Bischof (AB), and I am grateful for the opportunity to comment. The authors are focused on improving *interpretability* of statistical language models, which is essential for their applicability in the social sciences.

In my collaborations with economists and others, much effort has sometimes 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, the topics are not initially intuitive or self-consistent, and hence one begins iterating through a series of repeated model estimations using different specifications (e.g., number of topics) or vocabulary sets (e.g., via stop-word removal or minimum word-occurrence thresholds).

This labor-intensive iterative model building is especially frustrating when the goals are unclear. If we want lists of words that we already understand as 'representative' of a given topic or feeling, we can simply select these words using our own or other's expert opinion (e.g., Tetlock, 2007). If we just want the best fit for the data, we can use established model selection techniques (e.g., techniques in Airoldi et al., 2010 or the Bayes factors for LDA in Taddy, 2012). However, the desired outcome often lives somewhere in-between: we want 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'.

Difficulties in model interpretation are inherent to truly *unsupervised* topic analysis – when you are without relevant 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 to use unsupervised topic modeling as a dimension reduction step in a larger pipeline. Document-topic weights are downstream inputs to a function that predicts some attributes (this is an especially useful strategy when these attributes are known only for a small subset of your corpora). In such settings, we can use out-of-sample predictive performance in the downstream task as the arbitrator on topic quality. Alternatively, one can specify and estimate models that force document attributes to directly inform the topics. This is the strategy demonstrated nicely in AB's work here: 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 role in language choice. And it works! AB provide word lists that are intuitive and self consistent, without any of the usual steps of vocabulary narrowing.

In my own work, I have suggested that when document attributes are available you can often avoid latent variable models altogether and instead make use of standard high-dimensional regression techniques (others have made this point; e.g., Jia et al. 2014). In the multinomial inverse regression (MNIR) framework of Taddy (2013), word counts are treated as the response in a multinomial logistic regression onto document attributes. That article emphasizes the derivation of sufficient projections from the model and use of these projections in prediction. Taddy (2015a) describes a scalable distributed version of the MNIR algorithm and illustrates its use in a variety of additional tasks: identifying words that are indicative of a certain sentiment or subject; projecting documents into a low-dimensional space that quantifies, say, funny or useful content; and in constructing text-based control variables for a causal inference scheme.

We can apply these ideas on the Reuters dataset studied by AB. In the distributed version of MNIR from Taddy (2015a), the word-f count for each document-d is treated as Poisson random variable, using AB's notation,

$$w_{fd} \sim \text{Pois}\left(L_d \exp\left[\alpha_{0f} + \mathbf{I}_d' \boldsymbol{\varphi}_f + \mathbf{V}_d' \boldsymbol{\gamma}_f\right]\right) \tag{1}$$

where $L_d = \sum_f w_{fd}$ is the document length and $\mathbf{I}_d$ contains topic membership information. The extra attribute vector $\mathbf{V}_d$ can include any other conditioning information, such as the *region* or *industry* tags supplied by Reuters. The inferred topic loadings – elements of each $\boldsymbol{\varphi}_d$ – are then interpretable as topic effects on word choice *after controlling for* the characteristics in $\mathbf{V}_d$.

This is a standard generalized linear model (up to a $\log L_d$ shift). It can be estimated using any of the many available methods for such models, in particular regularized regression estimators that avoid overfit by placing penalties on the elements of $\boldsymbol{\varphi}_f$. I use the `gamlr` R package (implementing the POSE algorithms of Taddy, 2015b) to apply simple ℓ_1 regularization with BIC selection for the penalty magnitude. Everything is run 'out-of-the-box' without careful tuning, and regressions for different words are distributed across many compute nodes via the `distrom` package. I control for each document's geographic focus (as classified by Reuters) by including these tags in our $\mathbf{V}_d$ vectors. I took the tokenization supplied in Lewis et al. (2004) and all of the analysis code is in `https://github.com/TaddyLab/reuters`.

Metals	*gold, LME, copper, metal, COMEX, palladium, silver, aluminum, bullion, platinum*
Environment	*EPA, pollution, sulphur, environment, wildlife, emitter, soot, soybean, dioxide, species*
Defense	*Aberdeen, ldd, uld, chemical, defend, base, force, military, army, arms*
Economics	*nondurable, adjusted, unadjusted, percent, year, economy, statistics, month, growth, billion*
Monetary Econ	*policy, market, interest, bank, cent, rate, governor, make, meet, share*

Table 1: Top 10 words in a selection of topics, ranked by $\varphi_{fk}\bar{w}_f^{0.6}$ for φ_{fk} estimated in the MNIR specification of (1). These words are expanded from the stemmed tokens of Lewis et al. (2004).

Table 1 shows lists of top-10 words for a selection of topics. These words are 'top' as ranked by their corresponding MNIR loading, φ_{fk} for each topic k, multiplied by a measure of word prevalence. The analysis has done a good job of selecting words that are uniquely associated with those given topics but are not so rare as to be unrecognizable (except *ldd* and *uld* for defense). Note that *monetary economics* is a sub-topic within *economics*; due to the hierarchical nature of topic membership, the words in the last line of our table are those which differentiate monetary topics from others *within* economics. This happens naturally when hierarchical information is encoded in the regression design (i.e., in $\mathbf{I}_d$). Indeed, given that simple log-linear regressions can be used to resolve complex collinear effects of topics and other attributes, I would like to hear from AB what they see as the advantages of instead building and inferring a full generative model (which is more computationally expensive, even with nice HMC).

The quality of the word-lists in Table 1 is dependent upon the choice of 'top word' ranking function. Ranking by loading φ_{fk} alone yields mostly rare terms; e.g., names of companies or individuals. On the other hand, ranking by $\bar{w}_f\varphi_{fk}$, where $\bar{w}_f = \frac{1}{D}\sum_d w_{fd}$, leads to noticeable overlap across topics (i.e., the top words are too generic). I use a criteria that can be tuned between these two extremes: $\varphi_{fk}\bar{w}_f^q$, where $q \in [0, 1]$. Table 1 uses $q = 0.6$. I was inspired here by the example of AB's FREX, which similarly balances between topic specificity and usage probability via a tuning parameter. FREX seems to be a key ingredient in AB's framework, so that both my lists and AB's are the results of strategic model summarization. Careful summarization can also bring intuition to less obviously interpretable models; e.g., for standard LDA, Taddy (2012) ranks words by their topic 'lift' (word probability within topic over the aggregate word rate) for more coherent word lists than from the usual within-topic probability ranking.

Finally, a question: what are the lessons from AB's work towards more interpretable *unsupervised* modeling? The Reuters annotations are clearly of huge value for building an interpretable model. In HPC or MNIR, this supervision allows us to avoid the difficult task of topic interpretation and labeling. However, most available text data is annotated with only a small number of labels of low relevance. This is why unsupervised topic modeling, especially LDA from Blei et al. (2003) and its extensions, is massively useful and popular (and it is why advice such as that in Wallach et al. 2009, on more interpretable *unsupervised* modeling, is important). AB outline in Section 3.3 a procedure for estimating the topics associated with new unlabeled documents, but there doesn't seem to be a pathway for these documents to inform model estimation. That is, like MNIR, AB's scheme is inherently supervised. It would be great if there are lessons in this article that apply when we need to tell stories with little or no supervision.

References

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

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

Jia, J., L. Miratrix, B. Yu, B. Gawalt, L. El Ghaoui, L. Barnesmoore, S. Clavier, et al. (2014). Concise comparative summaries (ccs) of large text corpora with a human experiment. *The Annals of Applied Statistics 8*, 499–529.

Lewis, D. D., Y. Yang, T. G. Rose, and F. Li (2004). Rcv1: A new benchmark collection for text categorization research. *The Journal of Machine Learning Research 5*, 361–397.

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. (2015a). Distributed multinomial regression. *The Annals of Applied Statistics 9*, 1394–1414.

Taddy, M. (2015b). One-step estimator paths for concave regularization. *arXiv:1308.5623*.

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

Wallach, H. M., D. Mimno, and A. McCallum (2009). Rethinking LDA: Why Priors Matter. In *Neural Information Processing Systems*.