# 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 fitting a latent variable topic model and instead make use of standard high-dimensional regression techniques. Taddy (2013) describes the multinomial inverse regression (MNIR) framework wherein 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 (2015a) shows how a closely related (and more scalable) version of the MNIR algorithm 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.

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

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

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

This is a standard generalized linear model (up to a $\log L_d$ mean 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, such as $|\varphi_{jf}|$, on the elements of $\boldsymbol{\varphi}_f$. We use the `gamlr` R package implementing the POSE algorithms of Taddy (2015b), which includes simple $\ell_1$ regularization (we use this specification, with BIC selection for the penalty magnitude). We control for the document's geographic subject matter (as classified by Reuters) by including these tags in our $\mathbf{V}_d$ vectors. We use the term tokenization supplied in Lewis et al. (2004). All of the code for our analysis is in `https://github.com/TaddyLab/reuters`.

Looking to summarize our MNIR fit leads to an important question: what do we report? The quality of the topic word-lists will be determined by our 'top word' ranking function: do we rank by topic-word loading $\varphi$? Or by loading weighted by word occurence, or some other combination? Similarly, although AB fit more interpretable word-topic loadings in a sparse supervised model, their FREX summarization seems essential for the successful interpretability. More thoughtful topic summarization is also an easy way to get improved interpretability for standard LDA-type schemes, without requiring any new modeling. For example, Taddy (2012) describes topics in terms of the top words by topic 'lift' (word probability within topic over the aggregate word occurence rate) and this gives more coherent word lists.

For the MNIR Reuters data fit, we are interested in reporting the top words as a function of the corresponding $\varphi_{fk}$ loadings, for each topic $k$. A ranking of top words by $\varphi_{fk}$ alone (the increase in log word intensity for that topic) yields rare terms; e.g., names of companies or individuals. On the other extreme, ranking words by $\bar{w}_f\varphi_{fk}$ – the occurence weighted loading – yields top-word lists with noticable overlap across topics (i.e., the top words are too generic). Inspired by Airoldi and Bischoff, I decided to use a criteria that can be tuned within a given application: $\varphi_{fk}\bar{w}_f^q$, where $q$ varies between 0 and 1. In Table 1, apply this criterion with $q = 0.6$ to obtain lists of top 10 words for a selection of topics. To my eye, this listing 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 for *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 can be interpreted as those which differentiate monetary topics from others *within* economics.

| | |
|---|---|
| **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: The top 10 words in a selection of topics, ranked by $\varphi_{fk}\bar{w}_f^{0.6}$. These are expanded from the stemmed tokens in Lewis et al. (2004).

Finally, a question: what are the lessons from this work towards more interpretable *unsupervised* modeling? The Reuters annotations are clearly of huge value for building an interpretable model. In HPC or MNIR or another analysis framework, 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 the LDA of Blei (2012), is massively useful and popular. The authors outline in Section 3.3 a procedure for estimating the topics associated with new unlabelled documents, but there doesn't seem to be a pathway for these documents to inform model estimation. That is, like MNIR and others , your scheme is inherrently 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

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

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

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*. To appear.

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.