Comment: *Coauthorship and citation networks for statisticians*

Mladen Kolar – University of Chicago Booth School of Business
Matt Taddy – Microsoft Research and Chicago Booth

This article by Ji and Jin (JJ throughout) provides a network analysis, by two experts in the field, on the connections between statistics research papers. This is not just an exercise in navel-gazing. It gives us the opportunity to compare results from the methods we advocate to our intuition in an area we know well: our profession. And we think that the article contains some lessons for how we use network models and what data we choose to analyze. First, one can gain insight into network

Topic analysis

In our first study, we consider decomposing the abstracts of the articles into latent 'topics', e.g., as in the LDA of Blei et al. (2003). The properties of the citation network can then be considered in light of the topical *content* of the articles.

We use the `maptpx` R package (Taddy, 2012) to obtain posterior maximizing point estimates for LDA topics. The `maptpx` package applies the Bayes factors of Taddy (2012) in model selection, and for this data we find that a 15 topic decomposition is optimal. The full code to fit this model (and all of our other analyses) and to generate summaries for each topic are at `https://github.com/TaddyLab/statsArticles`.

We focus on topics that have seen their usage change over time – their mean proportions within documents during the first and last five years differ by more than 0.01. Most topics were stable over time, so that only three meet this threshold. These three topics are shown in the left panel of Figure 1. The list of words given for each topic are those with the highest *lift*: within-topic probability over the average corpus-wide occurrence rate.

Topic 1 seems to contain traditional mathematical statistics content, especially for non and semi parametric analysis. The three articles most representative of this topic (i.e., have the highest estimated usage) are

Backfitting and smooth backfitting for additive quantile models (Lee et al., 2010)

Depth weighted scatter estimators (Zuo and Cui, 2005), and

Estimating invariant laws of linear processes by U-statistics (Schick and Wefelmeyer, 2004).

This topic remains relatively popular, but its usage is decreasing over time. The other topic that is decreasing in popularity, topic 4, appears to represent Bayesian analysis. Its three most representative articles are

A conjugate prior for discrete hierarchical log-linear models (Massam et al., 2009)

Bayesian analysis of variable-order, reversible Markov chains (Bacallado, 2011)

Recapture models under equality constraints for the conditional capture probabilities (Farcomeni, 2011)

Over the same time period, we see a dramatic rise in the prevalence of topic 8, which appears to represent the material related to the popular penalized-deviance estimation framework. Its three most representative articles are

The adaptive lasso and its oracle properties (Zou, 2006)

Regularization and variable selection via the elastic net (Zou and Hastie, 2005)

On the adaptive elastic-net with a diverging number of parameters (Zou and Zhang, 2009)

Thus, our topic model shows an increased interest in penalized deviance estimation (and techniques popular in both statistics and machine learning). This rise in popularity for topic 8 has come at the expense of Bayesian analysis and another topic that is roughly interpretable as a different flavor of flexible analysis. Certainly, the authors' personal experience supports the loss of market share for fully Bayesian analysis; as datasets have grown in size we are often maximizing posteriors (which yields penalized deviance estimators) rather than averaging over them (Taddy, 2012, 2015, are personal examples).

This topic decomposition can be related back to the document networks. For example, we might be interested in which topics tend to generate higher citations. A linear regression of citations per year (the total count divided by the years since publication) onto a year indicator (i.e, a publication-date-specific intercept) and the document-topic weights yields an estimated extra 0.1 citations for per extra 0.1 weight on topic 8, our penalized deviance topic. This has a z-value of around 100, and it is the *only* topic that has a significant effect on citations. Note that this is only the 8th most common topic in our sample, so the higher cite count is not a function of prevalence alone. It is impossible to tell whether the topic is gaining popularity because it tends to lead to citations, or (what seems more likely) if these papers are cited often because the topic is becoming popular.

This analysis is a simple two stage procedure: we first estimate the topics and then relate them to network statistics (cite counts, i.e., degree in the citation network). A more complex procedure, such as the recent work by Tan et al. (2015), could be used to jointly estimate

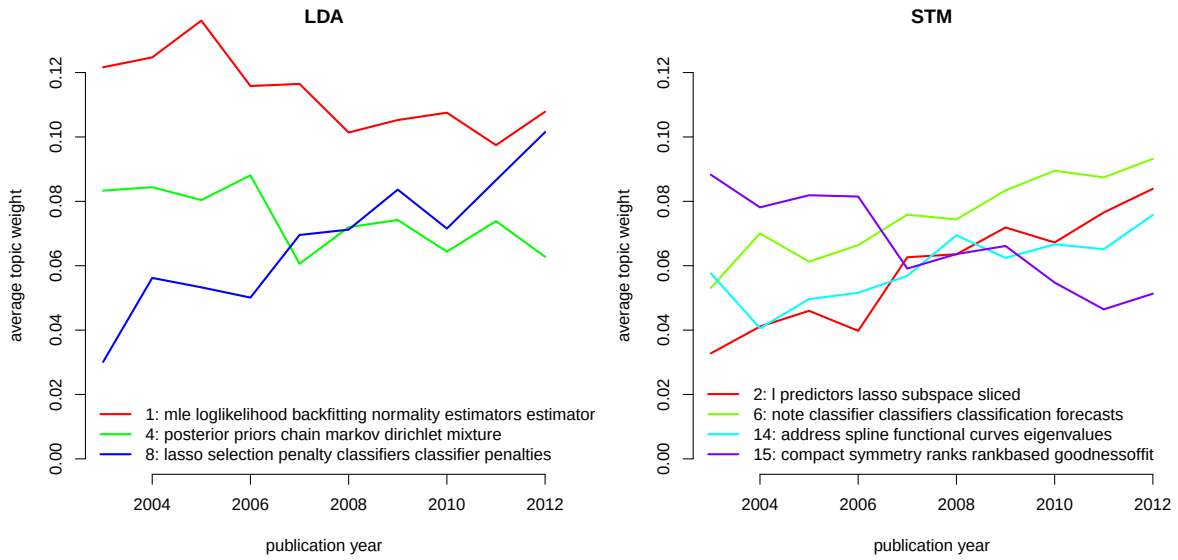

Figure 1: Annual average document-topic weights (that proportion of document devoted to each topic) for topics that saw their usage change by more than 0.1 between the first and last five years of our sample. The left panel shows topics from LDA via `maptpx` and the right shows topics from STM via `stm`.

network communities and text topics (e.g., to see if authors tend to cite within rather than across topics). As a partial step in this direction, we also applied the structural topic model (STM) of Roberts et al. (2013) as implemented in the `stm` package for R (Roberts et al., 2014). STM assumes that document-topic proportions are generated around a linear function of document attributes. Applying STM with both citations-per-year and publication date (annual indicators) as inputs yields the topics in the right panel of Figure 1. We used 15 topics, as in our earlier LDA. Again, a topic that focuses on penalized deviance estimation shows the biggest gains in usage over time. Two other topics are also gaining in popularity: one focused on classification (perhaps including much of the machine learning material that was also included in LDA's topic 8) and another that is tough for us to interpret. There is only one big loser over time: topic 15, which is also tough to interpret. Across both models, STM and LDA, the point of clear agreement is that penalized deviance estimation is a well-defined topic that is gaining in popularity. As we found in our two stage estimation, this topic is associated with higher citations: the fitted STM has a posterior mean effect of an extra 0.5 citations-per-year for an extra 0.1 usage of the penalized deviance topic.

Sensitivity of the citation networks to journal choice

One issue that we have with JJ's analysis is that it includes only a small portion of the statistics publication venues. Excluded are any applied journals (except for the Applications and Case Studies section of JASA), journals from other fields that use statistics and cite our work, and the entire machine learning literature. Certain network statistics, such as the betweenness scores that JJ report, are especially sensitive to any missingness in the network. But we also find that even simple degree statistics (i.e., citation counts) are sensitive to the set of journals considered.

References

Bacallado, S. (2011). Bayesian analysis of variable-order, reversible Markov chains. *The Annals of Statistics 39*, 838–864.

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

Farcomeni, A. (2011). Recapture models under equality constraints for the conditional capture probabilities. *Biometrika 98*(1), 237–242.

Lee, Y. K., E. Mammen, B. U. Park, et al. (2010). Backfitting and smooth backfitting for additive quantile models. *The Annals of Statistics 38*(5), 2857–2883.

Massam, H., J. Liu, A. Dobra, et al. (2009). A conjugate prior for discrete hierarchical log-linear models. *The Annals of Statistics 37*, 3431–3467.

Roberts, M. E., B. M. Stewart, and D. Tingley (2014). stm: R package for structural topic models. *R package vignette*.

Roberts, M. E., B. M. Stewart, D. Tingley, E. M. Airoldi, et al. (2013). The structural topic model and applied social science. In *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation*.

Schick, A. and W. Wefelmeyer (2004). Estimating invariant laws of linear processes by u-statistics. *Annals of statistics*, 603–632.

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

Tan, L. S. L., A. H. Chan, and T. Zheng (2015). Topic-adjusted visibility metric for scientific articles. *arXiv 1502.07190*.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association 101*, 1418–1429.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*, 301–320.

Zou, H. and H. H. Zhang (2009). On the adaptive elastic-net with a diverging number of parameters. *Annals of statistics 37*(4), 1733.

Zuo, Y. and H. Cui (2005). Depth weighted scatter estimators. *Annals of statistics*, 381–413.