Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Comment

David M. Blei

To cite this article: David M. Blei (2016) Comment, Journal of the American Statistical Association, 111:516, 1408-1410, DOI: 10.1080/01621459.2016.1245071

To link to this article: https://doi.org/10.1080/01621459.2016.1245071

Published online: 04 Jan 2017.

References

Blei, D. M., Griffiths, T. L., and Jordan, M. I. (2010), "The Nested Chinese Restaurant Process and Bayesian Nonparametric Inference of Topic Hierarchies," *Journal of the ACM (JACM)*, 57, 7. [1407]

Pauca, V. P., Piper, J., and Plemmons, R. J. (2006), "Nonnegative Matrix Factorization for Spectral Data Analysis," *Linear Algebra and Its Applications*, 416, 29–47. [1407]

Pauca, V. P., Shahnaz, F., Berry, M. W., and Plemmons, R. J. (2004), "Text Mining Using Non-Negative Matrix Factorizations," in *SDM* (Vol. 4), SIAM, pp. 452–456. [1407]

Price, B. S., Geyer, C. J., and Rothman, A. J. (2015), "Ridge Fusion in Statistical Learning," *Journal of Computational and Graphical Statistics*, 24, 439–454. [1407]

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2016, VOL. 111, NO. 516, Applications and Case Studies
http://dx.doi.org/10.1080/01621459.2016.1245071

Comment

David M. Blei

Department of Statistics and Department of Computer Science, Columbia University, New York, NY, USA

I congratulate Edo Airoldi and Jonathan Bischof (A&B) on an interesting article. This work brings important new ideas into the field of topic modeling, especially around how to visualize and interpret the topics.

Models. I will first restate the models the authors propose, but in a different order from how they are presented in the article. (I present them in order of simplicity.)

Each document is a p-vector of word counts $\mathbf{w}_d$. Suppose there are K topics. Each count w_{df} comes from a Poisson; its rate is an inner product of per-document topic weights θ_d, which is a point on the $(K-1)$-simplex, and the per-topic word intensities β_f, which is a nonnegative K-vector. The likelihood model is

$$w_{df} \sim \text{Pois}(\theta_d^\top \beta_f). \tag{1}$$

This type of model has been widely studied in machine learning and statistics (Canny 2004; Cemgil 2009; Ball, Karrer, and Newman 2011; Gopalan et al. 2014; Gopalan, Hofman, and Blei 2015; Schein et al. 2015; Zhou and Carin 2015). The formulation here is equivalent to the formulation in Table 1 of the article because the sum of Poissons is a Poisson.

While most previous work uses gamma or Dirichlet priors on the latent weights and components—these facilitate algorithms like Gibbs sampling and mean-field variational inference—A&B use hierarchical log normal priors. They argue and demonstrate that this parameterization regularizes for word "exclusivity," where an exclusive word is one that has higher rate in one or few topics and a nonexclusive word has similar rate in all topics.

This distinguishes their approach from traditional topic modeling (Griffiths and Steyvers 2004; Blei 2012), which places a Dirichlet prior on each topic's distribution over terms. That prior regularizes within a topic, but not across topics. A&B's regularization leads to better approaches to interpreting topics and better model performance at high numbers of topics.

More formally, the flat Poisson deconvolution model is

$$\tau_f^2 \sim \text{Scaled Inv-}\chi^2(\nu, \lambda^2) \tag{2}$$

$$\beta_{fk} \mid \tau_f^2 \sim \text{Log-Normal}(\psi, \tau_f^2) \quad k = 1, \ldots, K \tag{3}$$

$$\theta_d \sim \text{Logistic-Normal}(\eta, \lambda^2 I_K) \tag{4}$$

$$w_{df} \mid \theta_d, \beta_f \sim \text{Pois}(\theta_d^\top \beta_f). \tag{5}$$

The logistic normal distribution, thoroughly described in Aitchison (1982), posits a Gaussian random variable and then transforms it to the simplex via exponentiation and renormalization. It was also used for modeling topic proportions in Blei and Lafferty (2007), though our goals were different and we used a full covariance matrix.

In the next model on their path, we attach a vector of observed labels ℓ_d to each document. (A&B do not exactly consider this model, but it is the natural stepping stone to their more complicated model.) We assume that the topic space is one-to-one with the label space; thus ℓ_d is a K-vector of binary values. We use the observed labels to constrain the topics that the document exhibits, but still vary the strength of those topics. Rewritten, their model begins by generating topics with Equations (2) and (3). Then the documents and their labels are generated by

$$\xi_{kd} \sim \mathcal{N}(\eta_k, \lambda^2) \tag{6}$$

$$\ell_{dk} \mid \xi_k \sim \text{Bernoulli}(\sigma(\xi_k)) \tag{7}$$

$$\theta_{dk} \mid \boldsymbol{\xi} \propto \ell_{dk} \exp\{\xi_k\} \tag{8}$$

$$w_{df} \sim \text{Pois}(\theta_d^\top \beta_f), \tag{9}$$

where $\sigma(\cdot)$ denotes the logistic function. We have expanded out the logistic normal here into its constituent parts—a multivariate Gaussian and a point on the simplex—because of the more elaborate mapping that uses the labels as a "mask." Note the labels are generated by the same variables that determine the

CONTACT David M. Blei david.blei@columbia.edu Columbia University, New York, NY 10027.

topic proportions. This is an interesting detail, which encourages the topics present in the document to have higher probability in its topic proportions.

This model uses observed labels in a novel way. In Ramage et al. (2009), the labels are directly attached to parameter estimates; here they more naturally guide the estimates. In supervised topic models (Blei and McAuliffe 2007; Wang, Blei, and Li 2009), the labels are not one-to-one with topics; supervised topic models focus more on predicting the labels, but lose the direct mapping which A&B require.

Finally, the model A&B present is one where the topics (equivalently, the labels) are organized in a hierarchy, and where documents are labeled with multiple topics at any branch and level. In this model, the number of topics K is the number of topics in the entire tree. The document is generated as in Equations (6) to (9), but the per-term topic variables use the hierarchy: the strength of a word in a topic relates to its strength in the parent topic. Let π_k index the parent of topic k. The generative process of the hierarchy of topic intensities for term f is

$$\mu_{f,0} \sim \mathcal{N}(\psi, \gamma^2) \quad \text{for the root node.} \tag{10}$$

$$\tau_{f,k} \sim \text{Scaled Inv-}\chi^2(\nu, \lambda^2) \quad \text{for each internal node.} \tag{11}$$

$$\mu_{f,k} \mid \mu_{f,\pi_k}, \tau_{f,k-1} \sim \mathcal{N}(\mu_{f,\pi_k}, \tau_{f,k-1}) \quad \text{for each child.} \tag{12}$$

Notably, the intensities for the children of a topic share the same variance parameter around the mean of the parent intensity; that variance determines how exclusive the term is among the children. A term might be exclusive at a higher level of the tree (e.g., "score" to differentiate sports from business) but less exclusive lower down (e.g., "score" occurs equally in baseball, football, and tennis).

Results. With these models in hand, A&B analyze several large corpora of labeled documents and, with the simpler unsupervised model, unlabeled documents. It was gratifying that they treat interpretation and exploration as a first-class activity, accurately reflecting how investigators (especially in the computational social sciences and digital humanities) use topic models. See, for example, Jockers (2013).

A&B found that the FREX measure provides a much more interpretable view of topics as borne out both in their demonstrations and extensive human studies. (Note that FREX is related to the "term score" in Blei and Lafferty (2009), though we did not study it as thoroughly or creatively.) In the unsupervised method, the FREX measure is nearly equal to LDA at lower numbers of topics, indicating that we can improve the results of the simplest model with a better method of visualization. One interesting area of future work would be to embed FREX as a realized discrepancy in a posterior predictive check (Gelman, Meng, and Stern 1996; Mimno and Blei 2011). More generally, FREX is worth exploring as an effective way to visualize topics.

Summary and open problems. Again I congratulate A&B on an interesting article. Regularizing for exclusivity and using metrics like FREX to visualize topics are significant contributions to the growing field of large-scale probabilistic models of discrete data. A&B have opened the door to many avenues of research.

- *Bayesian nonparametrics and combining labeled and unlabeled topics.* Topics serve both to model and to interpret. With labels as part of the distribution, how might we add unlabeled topics to the model? Moreover, can we use new methods in Bayesian nonparametric Poisson factorization (Gopalan et al. 2014; Broderick et al. 2015; Zhou and Carin 2015) in concert with the methods proposed here?
- *Generalization to other types of data.* Poisson models are now used in many settings, such as social network analysis, natural images, computational neuroscience, recommendation systems, and statistical genetics. Can notions of exclusivity—both for regularization and visualization—be adapted to these other settings? Related, can these methods be adapted beyond matrices to large-scale Bayesian models of higher-order tensors (Kolda and Bader 2009; Hoff 2015)?
- *Large vocabularies.* Successful topic modeling requires pruning the vocabulary, and the models are no exception. (For example, A&B use only 3% of the vocabulary in the Reuters corpus.) How can these methods be combined with ideas of semantic dimension reduction (Bengio et al. 2003; Mikolov et al. 2013; Levy and Goldberg 2014) to better handle larger vocabularies?

References

Aitchison, J. (1982), "The Statistical Analysis of Compositional Data," *Journal of the Royal Statistical Society*, Series B, 44, 139–177. [1408]

Ball, B., Karrer, B., and Newman, M. (2011), "Efficient and Principled Method for Detecting Communities in Networks," *Physical Review E*, 84, 036103 (1–13). [1408]

Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003), "A Neural Probabilistic Language Model," *Journal of Machine Learning Research*, 3, 1137–1155. [1409]

Blei, D. (2012), "Probabilistic Topic Models," *Communications of the ACM*, 55, 77–84. [1408]

Blei, D., and Lafferty, J. (2007), "A Correlated Topic Model of Science," *Annals of Applied Statistics*, 1, 17–35. [1408]

——— (2009), "Topic Models," in *Text Mining: Theory and Applications*, eds. A. Srivastava, and M. Sahami, Taylor and Francis, pp. 77–116. [1409]

Blei, D., and McAuliffe, J. (2007), "Supervised Topic Models," in *Neural Information Processing Systems*, pp. 121–128. [1409]

Broderick, T., Mackey, L., Paisley, J., and Jordan, M. (2015), "Combinatorial Clustering and the Beta Negative Binomial Process," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37, 290–306. [1409]

Canny, J. (2004), "GaP: A Factor Model for Discrete Data," in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 122–129. [1408]

Cemgil, A. (2009), "Bayesian Inference for Nonnegative Matrix Factorization Models," *Computational Intelligence and Neuroscience*, 2009, 4:1–4:17. [1408]

Gelman, A., Meng, X., and Stern, H. (1996), "Posterior Predictive Assessment of Model Fitness via Realized Discrepancies," *Statistica Sinica*, 6, 733–807. [1409]

Gopalan, P., Hofman, J., and Blei, D. (2015), "Scalable Recommendation With Hierarchical Poisson Factorization," in *Uncertainty in Artificial Intelligence*, pp. 326–335. [1408]

Gopalan, P., Ruiz, F., Ranganath, R., and Blei, D. (2014), "Bayesian Nonparametric Poisson Factorization for Recommendation Systems," *Artificial Intelligence and Statistics*, 275–283. [1408,1409]

Griffiths, T., and Steyvers, M. (2004), "Finding Scientific Topics," *Proceedings of the National Academy of Science*, 101, 5228–5235. [1408]

Hoff, P. D. (2015), "Multilinear Tensor Regression for Longitudinal Relational Data," *Annals of Applied Statistics*, 3, 1169–1193. [1409]

Jockers, M. (2013), *Macroanalysis: Digital Methods and Literary History*, Champaign, IL: University of Illinois Press. [1409]

Kolda, T. G., and Bader, B. W. (2009), "Tensor Decompositions and Applications," *SIAM Review*, 51, 455–500. [1409]

Levy, O., and Goldberg, Y. (2014), "Neural Word Embedding as Implicit Matrix Factorization," in *Advances in Neural Information Processing Systems*, pp. 2177–2185. [1409]

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013), "Distributed Representations of Words and Phrases and Their Compositionality," in *Neural Information Processing Systems*, pp. 3111–3119. [1409]

Mimno, D., and Blei, D. (2011), "Bayesian Checking for Topic Models," in *Empirical Methods in Natural Language Processing*, pp. 227–237. [1409]

Ramage, D., Hall, D., Nallapati, R., and Manning, C. (2009), "Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-Labeled Corpora," in *Empirical Methods in Natural Language Processing*, pp. 248–256. [1409]

Schein, A., Paisley, J., Blei, D., and Wallach, H. (2015), "Bayesian Poisson Tensor Factorization for Inferring Multilateral Relations From Sparse Dyadic Event Counts," in *Knowledge Discovery and Data Mining*, pp. 1045–1054. [1408]

Wang, C., Blei, D., and Li, F. (2009), "Simultaneous Image Classification and Annotation," in *Computer Vision and Pattern Recognition*, pp. 1903–1910. [1409]

Zhou, M., and Carin, L. (2015), "Negative Binomial Process Count and Mixture Modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37, 307–320. [1408,1409]

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2016, VOL. 111, NO. 516, Applications and Case Studies
http://dx.doi.org/10.1080/01621459.2016.1245070

Rejoinder

Edoardo M. Airoldi

Department of Statistics, Harvard University, Cambridge, MA, USA

We wish to thank David Blei, Aleksandrina Goeva, Eric Kolaczyk, and Matthew Taddy for raising some questions and discussing a number of interesting points. Taken together, the discussions complement the data analysis in our article, explore novel connections with nonnegative matrix factorization techniques, and suggest avenues for further methodological development. While there were no disagreements, we welcome the opportunity to further discuss some of the points that were raised. We also take this opportunity to place the lessons we learned in a broader historical perspective.

1. A Historical Perspective

Quantitative analyses of text, and more specifically statistical analyses of word counts, is an area of methodological research with a long history (e.g., see Zipf 1932; Yule 1944; Miller, Newman, and Friedman 1958; Mosteller and Wallace 1963, 1964, 1984; De Morgan 1872; Efron and Thisted 1976; Mendenhall 1887), which has been quite active at the interface of statistics and the computational and information sciences, in the past decade (e.g., see Blei, Ng, and Jordan 2003; Erosheva, Fienberg, and Lafferty 2004; Airoldi et al. 2006; Blei and Lafferty 2007; Airoldi et al. 2010; Roberts, Stewart, and Airoldi 2016). Today, applications range from biology and medicine to economics and the social sciences, to the political sciences and the digital humanities, and to the IT industry at large.

Recurring elements in these analyses are: a matrix of word counts $\boldsymbol{w}$, whose entries record the number of occurrences of v unique terms (in a prearranged vocabulary) in n documents; the assumption of the existence of k subpopulations, typical of mixture models and mixed membership models (Airoldi et al. 2014); and a matrix of rates of occurrence $\boldsymbol{\beta}$ for v terms in the k subpopulations. The inferential targets of interest are often both the matrix of rates, and n vectors $\boldsymbol{\theta}$ that live in a $(k\text{-}1)$-simplex whose entries capture fractional associations between each of the n documents and the k subpopulations.

Whether indicators for the subpopulations are observed or not depends on the specific applications. For instance, in authorship attribution problems the subpopulations correspond to authors, and author indicators are typically observed for a large fraction of the documents (e.g., see Mosteller and Wallace 1963). In modern analyses of topical content, topic indicators are largely unobserved (e.g., see Blei 2012) with a few exceptions, including the model presented in Section 2 of our article and the corresponding data analyses, up to Section 4.5. Unobserved indicators introduce complications, methodologically and in the data analysis. However, whether they are observed or not is inconsequential for our narrative.

An intriguing difference between statistical and machine learning approaches to the analysis of word counts, relevant to our work, is the way the matrix of rates $\boldsymbol{\beta}$ is regularized.

In statistics, following Mosteller and Wallace (1963, 1964, 1984), the rates are regularized *per word*. For example, consider the word *and* in the analysis of "The Federalist" articles, where the two subpopulations are associated with two authors—Hamilton and Madison. Mosteller and Wallace reparameterized the rates for the word *and*, denoted $(\beta_{\mathrm{H}}, \beta_{\mathrm{M}})$, in terms of a total rate $\sigma = \beta_{\mathrm{H}} + \beta_{\mathrm{M}}$ and a differential rate $\tau_{\mathrm{M}} = \beta_{\mathrm{M}}/\sigma$. This reparameterization leads to the specification of sensible prior distributions for the rates of occurrence of all the terms in the vocabulary. Especially for the differential rates, since $\tau_{\mathrm{M}} \in [0, 1]$, it is

CONTACT Edoardo M. Airoldi *airoldi@fas.harvard.edu* Department of Statistics, Harvard University, 1 Oxford St., Cambridge, MA 02138.