Journal of the American Statistical Association

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

Comment

Aleksandrina Goeva & Eric D. Kolaczyk

To cite this article: Aleksandrina Goeva & Eric D. Kolaczyk (2016) Comment, Journal of the American Statistical Association, 111:516, 1405-1408, DOI: 10.1080/01621459.2016.1245072

To link to this article: http://dx.doi.org/10.1080/01621459.2016.1245072

Published online: 04 Jan 2017.

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, for example, for standard LDA, Taddy (2012) ranked 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 toward 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 are annotated with only a small number of labels of low relevance. This is why unsupervised topic modeling, especially LDA from Blei, Ng, and Jordan (2003) and its extensions, is massively useful and popular (and it is why advice such as that in Wallach, Mimno, and McCallum 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 does not 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., Erosheva, E. A., Fienberg, S. E., Joutard, C., Love, T., and Shringarpure, S. (2010), "Reconceptualizing the Classification of PNAS Articles," *Proceedings of the National Academy of Sciences*, 107, 20899–20904. [1404]

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003), "Latent Dirichlet Allocation," *The Journal of Machine Learning Research*, 3, 993–1022. [1404,1405]

Jia, J., Miratrix, L., Yu, B., Gawalt, B., El Ghaoui, L., Barnesmoore, L., and Clavier, S. (2014), "Concise Comparative Summaries (ccs) of Large Text Corpora With a Human Experiment," *The Annals of Applied Statistics*, 8, 499–529. [1404]

Lewis, D. D., Yang, Y., Rose, T. G., and Li, F. (2004), "Rcv1: A New Benchmark Collection for Text Categorization Research," *The Journal of Machine Learning Research*, 5, 361–397. [1404]

Taddy, M. (2012), "On Estimation and Selection for Topic Models," in *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS 2012)*. [1404]

——— (2013), "Multinomial Inverse Regression for Text Analysis," *Journal of the American Statistical Association*, 108, 755–770. [1404]

——— (2015), "Distributed Multinomial Regression," *The Annals of Applied Statistics*, 9, 1394–1414. [1404]

——— (in press), "One-Step Estimator Paths for Concave Regularization," *Journal of Computational and Graphical Statistics*. [1404]

Tetlock, P. (2007), "Giving Content to Investor Sentiment: The Role of Media in the Stock Market," *Journal of Finance*, 62, 1139–1168. [1404]

Wallach, H. M., Mimno, D., and McCallum, A. (2009), "Rethinking LDA: Why Priors Matter," *Advances in Neural Information Processing Systems*, 22. [1405]

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2016, VOL. 111, NO. 516, Applications and Case Studies
http://dx.doi.org/10.1080/01621459.2016.1245072

Comment

Aleksandrina Goeva and Eric D. Kolaczyk

Department of Mathematics & Statistics, Boston University, Boston, MA, USA

1. Introduction

We congratulate the authors on an impressive piece of work. At the heart of this work, as indicated in the title, is a novel regularization scheme, which is intended to address certain shortcomings identified by the authors in the literature on dimensionality reduction principles and techniques for topic modeling in document analysis. This regularization is carefully motivated, and its effectiveness is demonstrated empirically with a thoroughness that should serve as a model for the field. At the same time, it can be said that the regularization is rather complex and, as a result, interpretation arguably suffers to some extent, particularly at a first reading. Accordingly, we have taken as our modest goal in this discussion to attempt to lend further insight into the nature of the regularization proposed here. Toward this end, while the authors take a formally Bayesian approach to modeling and estimation, here we adopt for our purpose the perspective of complexity-penalized regularization, as an alternative lens through which to view the authors' contributions. Throughout we consider certain simplifications of the assumptions of the proposed model, where we feel doing so lends additional insight, hopefully without excessive loss of fidelity to the original.

The authors' hierarchical Poisson convolution (HPC) model, conditional on the topic hierarchy tree, can be summarized by the graphical model diagrammed in our Fig. 1. As indicated in the authors' own Fig. 1, in the article itself, structure on the word frequency matrix W is provided by imposing structure on documents (left) and words (right). Let $\boldsymbol{\beta}_f$ be a $K \times 1$ vector of occur-

CONTACT Eric Kolaczyk *kolaczyk@bu.edu* Department of Mathematics and Statistics, Boston University, 111 Cummington Street, Boston, MA 02215.

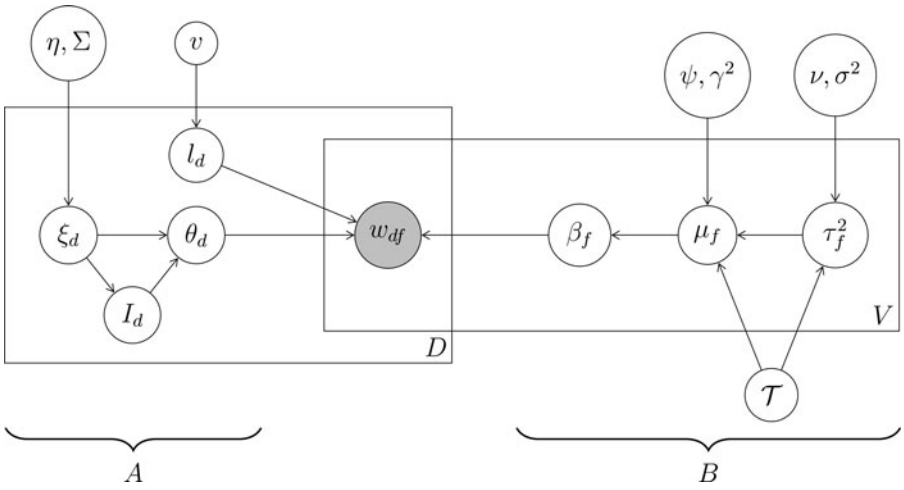

Figure 1. Graphical model diagram of the hierarchical Poisson convolution (HPC) model. Plates indicate replication, outside circles are hyper-parameters for priors, and shading means a quantity is observed. (Note: l_d is not necessarily assumed observed here.)

rence rates for word $f \in \{1, \ldots, V\}$, across all K topics in the topic hierarchy. Define $\boldsymbol{\alpha}_d = l_d \boldsymbol{\theta}_d$, where l_d is a scalar and $\boldsymbol{\theta}_d$ is a $K \times 1$ vector containing the proportion with which document $d \in \{1, \ldots, D\}$ belongs to each one of the K topics. According to the HPC generative process, $w_{df} \sim \text{Poisson}(\boldsymbol{\alpha}_d^T \boldsymbol{\beta}_f)$. Therefore, $\mathbb{E}[W] = AB$, where the dth row of A is $\boldsymbol{\alpha}_d$ and the fth column of B is $\boldsymbol{\beta}_f$. Hence, ignoring the (important) scaling inherent in the parameters l_d, the proposed model can be viewed usefully as constraining a certain nonnegative matrix factorization (NMF), that is, $\Rightarrow W \approx AB$. This factorization is reflected at the bottom of Figure 1 here, and shown explicitly in Figure 2.

Now consider the structure lent to the matrices A and B in this NMF, through the priors adopted by the authors in their HPC model. We connect our NMF approach to the original parameterization through a rederivation of the log-posterior distribution of A and B given the observed word count matrix W, with the goal of producing a complexity-penalized formulation of the optimization problem underlying the authors' proposed HPC-based estimation of these two matrices.

Writing the log-posterior as

$$\log \mathbb{P}(A, B|W) \approx \log \mathbb{P}(W|A, B) + \log \mathbb{P}(A) + \log \mathbb{P}(B),$$

we begin with the likelihood. Formally, the likelihood is Poisson. However, in the literature on NMF, various error functions have been proposed, with the most widely used arguably being squared-error loss. This suggests approximating the log-likelihood $\log \mathbb{P}(W|A, B)$ by the quantity $||W - AB||_{\mathcal{F}}^2$, where $||\cdot||_{\mathcal{F}}$ denotes the Frobenius norm.

Next consider the priors on A and B. Beginning with $\mathbb{P}(A)$, and treating the document lengths l_d as fixed and known, we write

$$\log \mathbb{P}(A) = \log \mathbb{P}\left(\{\boldsymbol{\alpha}_d\}_{d=1}^D\right) = \sum_{d=1}^D \log \mathbb{P}(l_d \boldsymbol{\theta}_d)$$
$$= \sum_{d=1}^D \log \mathbb{P}(\boldsymbol{\theta}_d) + c,$$

where here and elsewhere c denotes an arbitrary constant (not necessarily the same). Now

$$\mathbb{P}(\boldsymbol{\theta}_d) = \sum_{\boldsymbol{I}_d} \int_{\boldsymbol{\xi}_d} \mathbb{P}(\boldsymbol{\theta}_d|\boldsymbol{I}_d, \boldsymbol{\xi}_d) \mathbb{P}(\boldsymbol{I}_d|\boldsymbol{\xi}_d) \mathbb{P}(\boldsymbol{\xi}_d) m(\boldsymbol{I}_d, \boldsymbol{\xi}_d). \quad (1)$$

But note that

$$\mathbb{P}(\boldsymbol{\theta}_d|\boldsymbol{I}_d, \boldsymbol{\xi}_d) = \begin{cases} 1, & \text{iff } \text{supp}(\boldsymbol{\theta}_d) = \text{supp}(\boldsymbol{I}_d) \text{ and } \boldsymbol{\xi}_d \in \mathcal{A} \\ 0, & \text{otherwise} \end{cases},$$

where

$$\mathcal{A} = \left\{ \boldsymbol{\xi}_d : \text{ for } k \in \text{supp}(\boldsymbol{\theta}_d), \theta_{dk} = \frac{e^{\xi_{dk}}}{\sum_k e^{\xi_{dk}}} := f(\xi_{d|k}) \right\}.$$

Furthermore, for any $\boldsymbol{\xi}_d$ there is only one $\boldsymbol{I}_d$ that satisfies $\text{supp}(\boldsymbol{\theta}_d) = \text{supp}(\boldsymbol{I}_d)$. Finally, for $k \notin \text{supp}(\boldsymbol{\theta}_d)$, $\boldsymbol{\xi}_d$ can take on any value. Combining these observations and simplifying the resulting expressions, we obtain that

$$\log \mathbb{P}(A) = \sum_{d=1}^D -\frac{1}{2\lambda^2} \sum_{k \in \text{supp}(\boldsymbol{\theta}_d)} \left(f^{-1}(\boldsymbol{\theta}_d)[k] - \boldsymbol{\eta}[k]\right)^2 + c, \quad (2)$$

where $[k]$ indicates the kth entry of a vector and λ is the scale parameter for the (conditional) normal prior on $\boldsymbol{\theta}$.

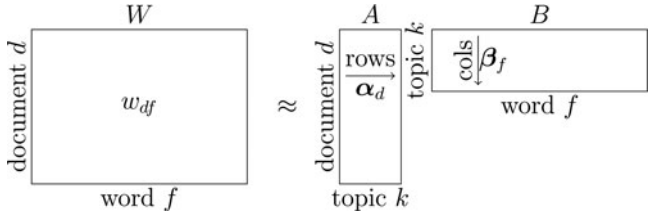

Figure 2. Representation of HPC model likelihood parameterization as a nonnegative matrix factorization (NMF).

For $\mathbb{P}(B)$, we can write

$$\log \mathbb{P}(B) = \sum_{f=1}^{V} \log \mathbb{P}\left(\boldsymbol{\beta}_f\right) = \sum_{f=1}^{V} \left(-\mathbf{1}^T \boldsymbol{\mu}_f + \log \mathbb{P}\left(\boldsymbol{\mu}_f\right)\right),$$

where $\boldsymbol{\mu}_f = \log(\boldsymbol{\beta}_f)$ is the collection of all log-rates in the tree for word f. Now suppose the dispersion parameters $\tau^2_{f,k}$ are treated as fixed and known. In the HPC model, the elements $\mu_{f,k}$ of $\boldsymbol{\mu}_f$ are then conditionally independent normal in a Markov fashion down the topic hierarchy tree, from root to leaves. So, ignoring the contribution of the corpus-level term in the prior, $\log \mathbb{P}(B)$ can be expressed as

$$\sum_{f=1}^{V} \left(-\mathbf{1}^T \boldsymbol{\mu}_f - \sum_{j \in \text{int}(\mathcal{T})} \frac{1}{2\tau^2_{f,j}} ||\boldsymbol{\mu}_{f,\text{ch}(j)} - \mu_{f,j}\mathbf{1}||_2^2 \right), \quad (3)$$

where $\text{int}(\mathcal{T})$ is the set of interior nodes (i.e., nonleaves) of the topic tree $\mathcal{T}$ and $\text{ch}(j)$ denotes the children of node j in $\mathcal{T}$.

Combining the above arguments, we arrive at the following complexity-penalized NMF problem as an approximation of the posterior maximization posed in the article:

$$\min_{A,B} \left\{ ||W - AB||^2_{\mathcal{F}} + \underbrace{\lambda_1 \sum_{d=1}^{D} \|\boldsymbol{\xi}_d(A) - \boldsymbol{\eta}\|^2}_{\substack{\text{over active topics} \\ \text{regularization} \\ \text{on rows of } A}} + \underbrace{\sum_{f=1}^{V} \left(\mathbf{1}^T \boldsymbol{\mu}_f(B) + \sum_{j \in \text{int}(\mathcal{T})} \frac{1}{2\tau^2_{f,j}} ||\boldsymbol{\mu}_{f,\text{ch}(j)}(B) - \mu_{f,j}(B)\mathbf{1}||_2^2 \right)}_{\text{regularization on cols of } B} \right\} \quad (4)$$

The representation in (4) allows us finally to make several observations.

1. The posterior-based estimation strategy associated with the HPC model can be viewed, to a reasonable extent, as being in the family of NMF solutions with ℓ_2-based penalties. Previously, for example, Pauca et al. (2004) had applied a penalty proportional to $||B||^2_{\mathcal{F}}$, while Pauca, Piper, and Plemmons (2006) had incorporated both $||A||^2_{\mathcal{F}}$ and $||B||^2_{\mathcal{F}}$. However, in the current article there are at least three key differences: (a) the nonlinear and atomized fashion (i.e., over active topics only) in which A enters the penalty; (b) the hierarchical nature of the ℓ_2 penalty for B; and (c) the addition of the linear term $\mathbf{1}^T\boldsymbol{\mu}_f(B)$. Furthermore, we note that B is penalized on a logarithmic scale (i.e., since $\boldsymbol{\mu}_f = \log(\boldsymbol{\beta}_f)$) and that the penalty on the log-rates of words in columns of B differs markedly from $||B||^2_{\mathcal{F}}$. The regularization on the columns of B that we arrive at combines the use of hierarchies, which is popular in topic modeling (e.g., Blei, Griffiths, and Jordan 2010), with principles of ℓ_2 penalties. The manner in which children log-rates are shrunk toward their parents can be interpreted as a variant of the ridge fusion penalty, discussed in Price, Geyer, and Rothman (2015), along paths from root to leaves. Note too that, where the $\boldsymbol{\mu}_f$ are positive, we have $\mathbf{1}^T\boldsymbol{\mu}_f(B) = ||\boldsymbol{\mu}_f||_1$, in which case it is perhaps tempting to think of the penalty on B in the spirit of a convex combination of ℓ_1 and ℓ_2 norms.
2. From a computational perspective, the optimization in (4) is somewhat nonstandard. Suppose the elements ξ are unconstrained. The last two terms of the objective function (i.e., deriving from $P(A)$ and $P(B)$) are convex in the ξ and μ parameterizations. And the elements of the product AB in the first term are sums of products of exponential functions applied to the ξ and μ, albeit with a renormalization in the ξ variable and an unbounded domain for both variables. So it seems possible that convex optimization procedures could be used to solve this problem, with appropriate care. However, the atomization implicit in the role the set $\mathcal{A}$ plays in the penalty on the ξ (and hence A) arising through the use of multinomial sampling of word-topic associations in the prior on A, requires thought. It might be possible to relax the problem to a more tractable variant. Alternatively, one might focus on the supervised version of the unsupervised posterior optimization we consider here, as the authors do in their applications, replacing $\mathbb{P}(A, B|W)$ by $\mathbb{P}(A, B|W, I)$ throughout, which simplifies away this challenge. In any event, from the computational perspective, a strength of the probabilistic approach adopted by the authors in formulating their regularization is readily apparent—the resulting optimization problem becomes primarily a problem of designing an appropriate Monte Carlo sampler.
3. There are several parameters in the HPC model that we have assumed here to be fixed and known. Our treatment of the document lengths l_d (important to the authors' formulation of the problem and a key way in which their work differs from much of that in the literature on topic models) is equivalent to conditioning on $\boldsymbol{l} \equiv \{l_d\}$, as the authors do as well. On the other hand, our treatment of the variances $\tau^2_{f,j}$ is analogous to needing to set the regularization parameter(s) in a ridge regression. The probabilistic perspective adopted in the article facilitates an inferential approach to setting these parameters.
4. The manner in which the authors' regularization is reexpressed in (4) is useful in helping to further highlight a central feature of their approach: the regularization is across topics over words (i.e., within columns of B, over rows) rather than the converse. It is this feature that appears to facilitate gains in interpretability.

Again, we congratulate the authors on a very interesting article. The work not only makes important inroads in and of itself in the area of document analysis, but, moreover, can be viewed as suggesting interesting future directions from the perspective of complexity-penalized NMF methods in this area.

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.