Journal of the American Statistical Association

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

Improving and Evaluating Topic Models and Other Models of Text

Edoardo M. Airoldi & Jonathan M. Bischof

To cite this article: Edoardo M. Airoldi & Jonathan M. Bischof (2016) Improving and Evaluating Topic Models and Other Models of Text, Journal of the American Statistical Association, 111:516, 1381-1403, DOI: 10.1080/01621459.2015.1051182

To link to this article: https://doi.org/10.1080/01621459.2015.1051182

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2016, VOL. 111, NO. 516, 1381–1412, Applications and Case Studies
http://dx.doi.org/10.1080/01621459.2015.1051182

Taylor & Francis
Taylor & Francis Group

Improving and Evaluating Topic Models and Other Models of Text

Edoardo M. Airoldi[a] and Jonathan M. Bischof[b]

[a]Department of Statistics, Harvard University, Cambridge, MA, USA; [b]Google, San Francisco, CA, USA

ABSTRACT

An ongoing challenge in the analysis of document collections is how to summarize content in terms of a set of inferred *themes* that can be interpreted substantively in terms of topics. The current practice of parameterizing the themes in terms of most frequent words limits interpretability by ignoring the differential use of words across topics. Here, we show that words that are both frequent and exclusive to a theme are more effective at characterizing topical content, and we propose a regularization scheme that leads to better estimates of these quantities. We consider a supervised setting where professional editors have annotated documents to topic categories, organized into a tree, in which leaf-nodes correspond to more specific topics. Each document is annotated to multiple categories, at different levels of the tree. We introduce a hierarchical Poisson convolution model to analyze these annotated documents. A parallelized Hamiltonian Monte Carlo sampler allows the inference to scale to millions of documents. The model leverages the structure among categories defined by professional editors to infer a clear semantic description for each topic in terms of words that are both frequent and exclusive. In this supervised setting, we validate the efficacy of word frequency and exclusivity at characterizing topical content on two very large collections of documents, from *Reuters* and the *New York Times*. In an unsupervised setting, we then consider a simplified version of the model that shares the same regularization scheme with the previous model. We carry out a large randomized experiment on Amazon Mechanical Turk to demonstrate that topic summaries based on frequency and exclusivity, estimated using the proposed regularization scheme, are more interpretable than currently established frequency-based summaries, and that the proposed model produces more efficient estimates of exclusivity than the currently established models.

ARTICLE HISTORY
Received September 2012
Revised January 2015

KEYWORDS
Categorical data; Hamiltonian Monte Carlo; High-dimensional data; Parallel inference; Text analysis

1. Introduction

A recurring challenge in multivariate statistics is how to construct interpretable low-dimensional summaries of high-dimensional data. Historically, simple models based on correlation matrices, such as principal component analysis (Jolliffe 1986) and canonical correlation analysis (Hotelling 1936), have proven to be effective tools for data reduction. More recently, multilevel models have become a flexible and powerful tool for finding latent structure in high-dimensional data (McLachlan and Peel 2000; Blei, Ng, and Jordan 2003; Airoldi et al. 2008, 2014; Sohn and Xing 2009). However, while interpretable statistical summaries are highly valued in applications, dimensionality reduction models are rarely optimized to aid qualitative discovery; there is no guarantee that the optimal low-dimensional projections will be understandable in terms of quantities of scientific interest that can help practitioners make decisions. Here, we design a model with scientific estimands of interest in mind to achieve an optimal balance of interpretability and dimensionality reduction.

We consider a setting in which we observe two sets of categorical data for each unit of observation: $\boldsymbol{w}_{1:V}$, which live in a high-dimensional space, and $\boldsymbol{l}_{1:K}$, which live in a structured low-dimensional space and provide a direct link to information of scientific interest about the sampling units. The goal of the analysis is two-fold. First, we desire to develop a joint model for the observations $\boldsymbol{Y} \equiv \{\boldsymbol{W}_{D\times V}, \boldsymbol{L}_{D\times K}\}$ that can be used to project the data onto a low-dimensional parameter space $\boldsymbol{\Theta}$ in which interpretability is maintained by mapping categories in $\mathcal{L}$ to directions in $\boldsymbol{\Theta}$. Second, we would like the mapping from the original space to the low-dimensional projection to be scientifically interesting so that statistical insights about $\boldsymbol{\Theta}$ can be understood in terms of the original inputs, $\boldsymbol{w}_{1:V}$, in a way that guides future research.

In the application to text analysis that motivates this work, $\boldsymbol{w}_{1:N}$ are the raw word counts observed in each document and $\boldsymbol{l}_{1:K}$ are a set of labels created by professional editors that are indicative of topical content. Specifically, the words are represented as an unordered vector of counts, with the length of the vector corresponding to the size of a known dictionary. The labels are organized in a tree-structured ontology, from the most generic topic at the root of the tree to the most specific topic at the leaves. Each news article may be annotated with more than one label, at the editors' discretion. The number of labels is given by the size of the ontology and typically ranges from tens to hundreds of categories. In this context, the inferential challenge is to discover a low-dimensional representation of topical content, $\boldsymbol{\Theta}$, that aligns with the coarse labels provided by editors while at the same time providing a mapping between the textual content

CONTACT Edoardo M. Airoldi airoldi@fas.harvard.edu Department of Statistics, Harvard University, Cambridge, MA 02138.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.

and directions in $\boldsymbol{\Theta}$ in a way that formalizes and enhances our understanding of how low-dimensional structure is expressed in the space of observed words.

Recent approaches to this problem in the machine learning literature have taken a Bayesian hierarchical approach to this task by viewing a document's content as arising from a mixture of component distributions, commonly referred to as "topics" as they often capture thematic structure (Blei 2012). As the component distributions are almost exclusively parameterized as multinomial distributions over words in the vocabulary, the loading of words onto topics is characterized in terms of the relative frequency of within-component usage. While relative frequency has proven to be a useful mapping of topical content onto words, recent work has documented a growing list of interpretability issues with frequency-based summaries: they are often dominated by contentless "stop" words (Wallach, Mimno, and McCallum 2009), sometimes appear incoherent or redundant (Chang et al. 2009; Mimno et al. 2011), and typically require post-hoc modification to meet human expectations (Hu et al. 2011). Selecting the number of topics is also a challenging problem (e.g., see Airoldi et al. 2010). Here, we propose a new regularization scheme that incorporates how words are used differentially across topics, in Sections 2.1 and 4.6. If a word is common in a topic, it is also important to know whether it is common in many topics or relatively exclusive to the topic in question. Both of these summary statistics are informative: nonexclusive words are less likely to carry topic-specific content, while infrequent words occur too rarely to form the semantic core of a topic; see Sections 4.2 and 4.3. We therefore look for the most frequent words in the corpus that are also preferentially used to write about the topic of interest to summarize content associated with such topic. Our approach borrows ideas from the statistical literature, in which models of differential word usage have been leveraged for analyzing writing styles in a supervised setting (Mosteller and Wallace 1984; Airoldi et al. 2006), and combine them with ideas from the machine learning literature, in which latent variable and mixture models based on frequent word usage have been used to infer structure that often captures topical content (McCallum et al. 1998; Blei, Ng, and Jordan 2003; Canny 2004).

From a statistical perspective, models based on topic-specific distributions over the vocabulary (Blei, Ng, and Jordan 2003) often fail to produce stable estimates of differential word usage since they only model the relative frequency of words within topics. Results in Eisenstein, Ahmed, and Xing (2011) suggest that such a popular parameterization leads to an amplification of the estimated differential word usage rates, especially for rare words, arguably because of the lack of mechanisms to regularize word rates across topics. The trade-off between these two orthogonal regularization strategies (over words within a topic versus over the same word across topics) has been explored in the literature (Mosteller and Wallace 1964, 1984; Canny 2004; Airoldi, Fienberg, and Xing 2007b). To tackle this issue, we introduce the generative framework of hierarchical Poisson convolution (HPC) that parameterizes topic-specific word counts as unnormalized count variates whose rates can be regularized across topics as well as within them, leading to stable inference of both word frequency and exclusivity, as we show in Section 4.6. HPC can be seen as a fully generative extension of sparse topic coding (Zhu and Xing 2012) that emphasizes regularization and interpretability rather than exact sparsity. Additionally, HPC leverages hierarchical systems of topic categories created by professional editors in collections such as *Reuters*, *New York Times*, *Wikipedia*, and *Encyclopedia Britannica* to make focused comparisons of differential use between neighboring topics on the tree and build a sophisticated joint model for topic memberships and labels in the documents. By conditioning on a known hierarchy, we avoid the complicated task of inferring hierarchical structure (Blei et al. 2003; Mimno, Li, and McCallum 2007; Adams, Ghahramani, and Jordan 2010). We introduce a parallelized Hamiltonian Monte Carlo (HMC) estimation strategy that makes full Bayesian inference efficient and scalable.

The proposed model is designed to infer an interpretable description of human-generated labels; thus, we restrict the topic components to have a one-to-one correspondence with the human-generated labels, as in labeled LDA (Ramage et al. 2009). This *descriptive* link between the labels and topics differs from the *predictive* link used in Supervised LDA (Blei and McAuliffe 2010; Perotte et al. 2012), where topics are learned as an optimal covariate space to predict an observed document label or response variable. The more restrictive descriptive link can be expected to limit predictive power, but leads to summaries directly associated with individual labels. We then infer a description of these labels in terms of words that are both frequent and exclusive. We anticipate that learning a concise semantic description for any collection of topics implicitly defined by professional editors is the first step toward the semi-automated creation of domain-specific topic ontologies. Domain-specific topic ontologies may be useful for evaluating the semantic content of *inferred* topics, or for predicting the semantic content of new social media, including Twitter messages and Facebook wall-posts.

2. Hierarchical Poisson Convolution

The HPC model is a data-generating process for document collections whose topics are organized in a hierarchy, and whose topic labels are observed. We refer to the structure among topics interchangeably as a *hierarchy* or *tree* since we assume that each topic has exactly one parent and that no cyclical parental relations are allowed. Each document $d \in \{1, \ldots, D\}$ is a record of counts w_{fd} for every feature in the vocabulary, $f \in \{1, \ldots, V\}$. The length of the document is given by L_d, which we normalize by the average document length L to get $l_d \equiv \frac{1}{L} L_d$ (Mosteller and Wallace 1984; Airoldi et al. 2006; Airoldi, Fienberg, and Skinner 2007a). Documents have unrestricted membership to any combination of topics $k \in \{1, \ldots, K\}$ represented by a vector of labels $\boldsymbol{I}_d$ where $I_{dk} \equiv I\{\text{document } d \text{ associates with topic } k\}$. The HPC model uses the typical "bag of words" representation of text data, whereby information about relative ordering of words in a document is ignored and only strictly positive word counts are retained (McCallum et al. 1998; Nigam et al. 2000), for inference purposes, as detailed in Section 3 and the Appendix. For illustrating the data-generating process, however, we find useful to consider the whole count matrix, $\boldsymbol{W}_{D\times V}$, which reports the counts of each term in the vocabulary, including zero counts, in all documents. This is due to a key feature that distinguishes our model from most of the literature on topic models (e.g., Blei 2012); we avoid conditioning on the total number of words observed in the documents. Instead, document length is part

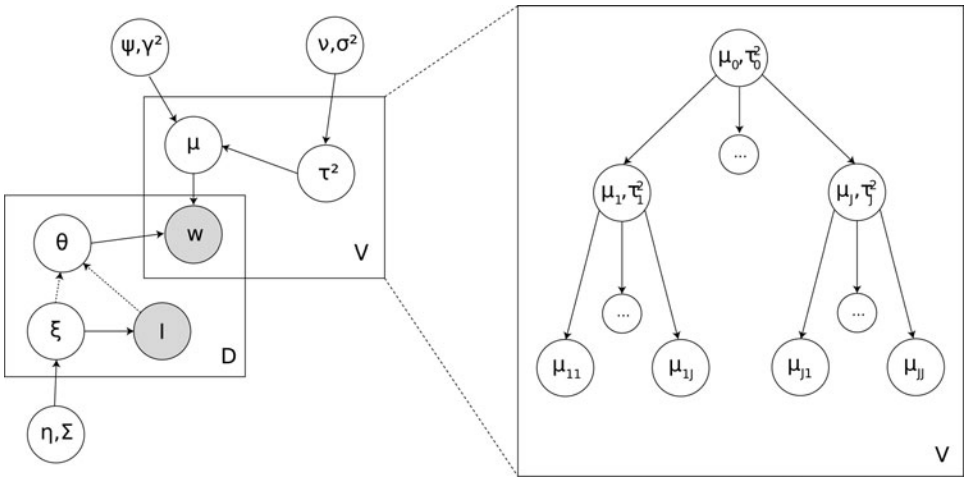

Figure 1. Graphical representation of hierarchical Poisson convolution (left) and detail on tree plate (right).

of the model, controlled by the parameters l_d and L introduced above (also see Mosteller and Wallace 1984; Airoldi et al. 2006).

2.1 Modeling Word Usage Rates on the Hierarchy

The HPC model leverages the known topic hierarchy by assuming that words are used similarly in neighboring topics. Let $\beta_{f,k}$ be the occurrence rate for word f in topic k, and define $\mu_{f,k} \equiv \beta_{f,k}$ for convenience of notation. Specifically, the log rate for a word across topics follows a Gaussian diffusion down the tree. Consider the topic hierarchy presented in the right panel of Figure 1. At the top level, $\mu_{f,0}$ represents the log rate for feature f overall in the corpus. The log rates $\mu_{f,1}, \ldots, \mu_{f,J}$ for first-level topics are then drawn from a Gaussian centered around the corpus rate with dispersion controlled by the variance parameter $\tau^2_{f,0}$. From first-level topics, we then draw the log rates for the second-level topics from another Gaussian centered around their mean $\mu_{f,j}$ and with variance $\tau^2_{f,j}$. This process is continued down the tree, with each parent node having a separate variance parameter to control the dispersion of its children.

The variance parameters τ^2_{fp} directly control the local differential expression in a branch of the tree. Words with high variance parameters can have rates in the child topics that differ greatly from the parent topic p, allowing the child rates to diverge. Words with low variance parameters will have rates close to the parent and so will be expressed similarly among the children. If we learn a population distribution for the τ^2_{fp} that has low mean and variance, it is equivalent to saying that most features are expressed similarly across topics a priori and that we would need a preponderance of evidence to believe otherwise.

Because of the hierarchy on the rates of word occurrence, the typical equivalence between an array of Poisson distributions for topic-specific word counts and a Poisson distribution for the total counts combined with a Multinomial distribution to allocate counts across topics (e.g., see Canny 2004; Buntine and Jakulin 2006; Airoldi, Fienberg, and Xing 2007b; Eisenstein, Ahmed, and Xing 2011) no longer holds.

2.2 Modeling the Topic Membership of Documents

Documents in the HPC model can contain content from any of the K topics in the hierarchy at varying proportions, with the exact allocation given by the vector $\boldsymbol{\theta}_d$ on the $K-1$ simplex. The model assumes that the count for word f contributed by each topic follows a Poisson distribution whose rate is moderated by the document's length and membership to the topic; that is, $w_{fdk} \sim \text{Pois}(l_d \theta_{dk} \beta_{fk})$. The only data we observe is the total word count $w_{fd} \equiv \sum_{k=1}^{K} w_{fdk}$, but the infinite divisibility property of the Poisson distribution gives us that $w_{fd} \sim \text{Pois}(l_d \boldsymbol{\theta}_d^T \boldsymbol{\beta}_f)$. These draws are done for every word in the vocabulary (using the same $\boldsymbol{\theta}_d$) to get the content of the document.[1]

In labeled document collections, human coders provide an extra piece of information for each document, $\boldsymbol{I}_d$, that indicates the set of topics that contributed its content. As a result, we know $\theta_{dk} = 0$ for all topics k where $I_{dk} = 0$, and only have to determine how content is allocated between the set of active topics. The data-generating process in Table 1 leads to well-defined topic proportions $\boldsymbol{\theta}_d$ when at least one element of $\boldsymbol{I}_d$ is positive. This constraint is a nonissue, however, since we are implicitly conditioning on $I_{dk} = 1$ for some k when fitting the model to data.

The HPC model assumes that these two sources of information for a document are not generated independently. A document should not have a high probability of being labeled to a topic from which it receives little content and vice versa. Instead, the model posits a latent K-dimensional topic affinity vector $\boldsymbol{\xi}_d \sim \mathcal{N}(\boldsymbol{\eta}, \boldsymbol{\Sigma})$ that expresses how strongly the document is associated with each topic. The topic memberships and labels of the document are different manifestations of this affinity. Specifically, each ξ_{dk} is the log odds that topic label k is active in the document, with $I_{dk} \sim \text{Bernoulli}(\text{logit}^{-1}(\xi_{dk}))$. Conditional on the labels, the topic memberships are the relative sizes of the document's affinity for the active topics and zero for inactive topics: $\theta_{dk} \equiv e^{\xi_{dk}} I_{dk} / \sum_{j=1}^{K} e^{\xi_{dj}} I_{dj}$. Restricting each document's membership vectors to the labeled topics is a natural and efficient way to generate sparsity in the mixing parameters, stabilizing inference, and reducing the computational burden of posterior simulation.

We outline the generative process in full detail in Table 1, which can be summarized in three steps. First, a set of rate and variance parameters are drawn for each feature in the vocabulary. Second, a topic affinity vector is drawn for each document in the corpus, which generates topic labels. Finally, both sets of

[1] This is where the model's name arises: the observed feature count in each document is the convolution of (unobserved) topic-specific Poisson variates.

Table 1. Generative process for hierarchical Poisson convolution.

Step	Generative process
Tree parameters	For feature $f \in \{1, \ldots, V\}$: • Draw $\mu_{f,0} \sim \mathcal{N}(\psi, \gamma^2)$ • Draw $\tau^2_{f,0} \sim$ Scaled Inv-$\chi^2(\nu, \sigma^2)$ • For $j \in \{1, \ldots, J\}$ (first level of hierarchy): – Draw $\mu_{f,j} \sim \mathcal{N}(\mu_{f,0}, \tau^2_{f,0})$ – Draw $\tau^2_{f,j} \sim$ Scaled Inv-$\chi^2(\nu, \sigma^2)$ • For $j \in \{1, \ldots, J\}$ (terminal level of hierarchy): – Draw $\mu_{f,j1}, \ldots, \mu_{f,jJ} \sim \mathcal{N}(\mu_{f,j}, \tau^2_{f,j})$ • Define $\beta_{f,k} \equiv e^{\mu_{f,k}}$ for $k \in \{1, \ldots, K\}$
Topic membership parameters	For document $d \in \{1, \ldots, D\}$: • Draw $\boldsymbol{\xi}_d \sim \mathcal{N}(\boldsymbol{\eta}, \boldsymbol{\Sigma} = \lambda^2 \boldsymbol{I}_K)$ • For topic $k \in \{1, \ldots, K\}$: – Define $p_{dk} \equiv 1/(1 + e^{-\xi_{dk}})$ – Draw $l_{dk} \sim$ Bernoulli(p_{dk}) – Define $\theta_{dk}(\boldsymbol{l}_d, \boldsymbol{\xi}_d) \equiv e^{\xi_{dk}} l_{dk} / \sum_{j=1}^{K} e^{\xi_{dj}} l_{dj}$, where $l_{dj} = 1$ for some j
Data generation	For document $d \in \{1, \ldots, D\}$: • Draw normalized document length $l_d \sim$ Pois$(\upsilon)/L$ • For every topic k and feature f: – Draw count $w_{fdk} \sim$ Pois$(l_d \boldsymbol{\theta}_{dk} \boldsymbol{\beta}_{fk})$ • Define $w_{fd} \equiv \sum_{k=1}^{K} w_{fdk}$ (observed data)

parameters are then used to generate the words in each document. For simplicity of presentation we assume that each nonterminal node has J children and that the tree has only two levels below the corpus level, but the model can accommodate any tree structure.

2.3 Estimands

To measure topical semantic content, we consider the topic-specific frequency and exclusivity of each word in the vocabulary. These quantities form a two-dimensional summary of each word's relation to a topic of interest, with higher scores in both being positively related to topic-specific content. Additionally, we develop a univariate summary of semantic content that can be used to rank words in terms of their semantic content. These estimands are simple functions of the rate parameters of HPC; the distribution of the documents' topic memberships is a nuisance parameter needed to disambiguate the content of a document between its labeled topics.

A word's topic-specific frequency, $\beta_{fk} \equiv \exp \mu_{fk}$, is directly parameterized in the model and is regularized across words (via hyperparameters ψ and γ^2) and across topics. A word's exclusivity to a topic, $\phi_{f,k}$, is its usage rate relative to a set of comparison topics $\mathcal{S}$: $\phi_{f,k} = \beta_{f,k} / \sum_{j \in \mathcal{S}} \beta_{f,j}$. A topic's siblings are a natural choice for a comparison set to see which words are overexpressed in the topic compared to a set of similar topics. While not directly modeled in HPC, the exclusivity parameters are also regularized by the τ^2_{fp}, since if the child rates are forced to be similar then the $\phi_{f,k}$ will be pushed toward a baseline value of $1/|\mathcal{S}|$. We explore the regularization structure of the model empirically in Section 4. While the set $\mathcal{S}$ can be taken to be the set of all topics, in the analysis we focus on the arguably most difficult task of thematically distinguishing between pairs of closely related topics. Success in this task requires the topical summaries to be descriptive of closely related themes, while being quantitatively and qualitatively indicative of the differences.

Since both frequency and exclusivity are important factors in determining a word's semantic content, a univariate measure of topical importance is a useful estimand for diverse tasks such as dimensionality reduction, feature selection, and content discovery. In constructing a composite measure, we do not want a high rank in one dimension to be able to compensate for a low rank in the other since frequency or exclusivity alone are not necessarily useful. We therefore adopt the harmonic mean to pull the "average" rank toward the lower score. For word f in topic k, we define the FREX$_{fk}$ score as the harmonic mean of the word's rank in the distribution of $\phi_{\cdot,k}$ and $\mu_{\cdot,k}$:

$$\text{FREX}_{fk} = \left(\frac{w}{\text{ECDF}_{\phi_{\cdot,k}}(\phi_{f,k})} + \frac{1-w}{\text{ECDF}_{\mu_{\cdot,k}}(\mu_{f,k})} \right)^{-1},$$

where w is the weight for exclusivity (which we set to 0.5 as a default) and ECDF$_{x_{\cdot,k}}$ is the empirical CDF function applied to the values x over the first index.

3. Scalable Inference via Parallelized HMC Sampler

We use a Gibbs sampler to obtain the posterior expectations of the unknown rate and membership parameters (and associated hyperparameters) given the observed data. Specifically, inference is conditioned on $\boldsymbol{W}$, a $D \times V$ matrix of word counts, $\boldsymbol{I}$, a $D \times K$ matrix of topic labels, $\boldsymbol{l}$, a D-vector of document lengths, and $\mathcal{T}$, a tree structure for the topics.

Creating a scalable inference method is critical since the space of latent variables grows linearly in the number of words and documents, with $K(D + V)$ total unknowns. Our model offers an advantage in that the posterior consists of two groups of parameters whose conditional posterior factors given the other. On one side, the conditional posterior of the rate and variance parameters $\{\boldsymbol{\mu}_f, \boldsymbol{\tau}^2_f\}_{f=1}^V$ factors by word given the membership parameters and the hyperparameters ψ, γ^2, ν, and σ^2. On the other, the conditional posterior of the topic affinity parameters $\{\boldsymbol{\xi}_d\}_{d=1}^D$ factors by document given the hyperparameters $\boldsymbol{\eta}$ and $\boldsymbol{\Sigma}$ and the rate parameters $\{\boldsymbol{\mu}_f\}_{f=1}^V$.

Conditional on the hyperparameters, we are left with two blocks of draws that can be broken into V or D independent threads. Using parallel computing software such as message passing interface (MPI), the computation time for drawing the parameters in each block is only constrained by resources required for a single draw. The total runtime need not significantly increase with the addition of more documents or words as long as the number of available cores also increases.

Both of these conditional distributions are only known up to a constant and can be high dimensional if there are many topics, making direct sampling impossible and random walk Metropolis inefficient. We are able to obtain uncorrelated draws through the use of HMC (Neal 2011), which leverages the posterior gradient and Hessian to find a distant point in the parameter space with high probability of acceptance. HMC works well for log densities that are unimodal and have relatively constant curvature. We give step-by-step instructions for our implementation of the algorithm in the Appendix.

After appropriate initialization, we follow a fixed Gibbs scan where the two blocks of latent variables are drawn in parallel from their conditional posteriors using HMC. We then draw the hyperparameters conditional on all the inputed latent variables.

3.1 Block Gibbs Sampler

To set up the block Gibbs sampling algorithm, we derive the relavant conditional posterior distributions and explain how we sample from each.

3.1.1 Updating Tree Parameters

In the first block, the conditional posterior of the tree parameters factors by word:

$$p(\{\boldsymbol{\mu}_f, \boldsymbol{\tau}_f^2\}_{f=1}^V | \boldsymbol{W}, \boldsymbol{I}, \boldsymbol{l}, \psi, \gamma^2, \nu, \sigma^2, \{\boldsymbol{\xi}_d\}_{d=1}^D, \mathcal{T}) \propto \prod_{f=1}^V \left\{ \prod_{d=1}^D p(w_{fd} | \boldsymbol{I}_d, l_d, \mu_f, \boldsymbol{\xi}_d) \right\} \cdot p(\boldsymbol{\mu}_f, \boldsymbol{\tau}_f^2 | \psi, \gamma^2, \mathcal{T}, \nu, \sigma^2).$$

Given the conditional conjugacy of the variance parameters and their strong influence on the curvature of the rate parameter posterior, we sample the two groups conditional on each other to optimize HMC performance. Conditioning on the variance parameters, we can write the likelihood of the rate parameters as a Poisson regression where the documents are observations, the $\boldsymbol{\theta}_d(\boldsymbol{I}_d, \boldsymbol{\xi}_d)$ are the covariates, and the l_d serve as exposure weights. The model matrix, $\boldsymbol{\Theta}$, has D rows and K columns, with each row containing topic membership proportions for document d across the K topics. Let $S_d = \sum_{j=1}^K e^{\xi_{dj}} I_{dj}$, then we can write more explicitly

$$\boldsymbol{\Theta} = \begin{bmatrix} e^{\xi_{1,1}} I_{1,1}/S_1 & e^{\xi_{1,2}} I_{1,2}/S_1 & \dots & e^{\xi_{1,K}} I_{1,K}/S_1 \\ e^{\xi_{2,1}} I_{2,1}/S_2 & e^{\xi_{2,1}} I_{2,1}/S_2 & \dots & e^{\xi_{2,K}} I_{2,K}/S_2 \\ \vdots & \vdots & \ddots & \vdots \\ e^{\xi_{D,1}} I_{D,1}/S_D & e^{\xi_{D,2}} I_{D,2}/S_D & \dots & e^{\xi_{D,K}} I_{D,K}/S_D \end{bmatrix}.$$

The prior distribution of the rate parameters is a Gaussian graphical model, so a priori the log rates for each word are jointly Gaussian with mean $\psi \mathbf{1}$ and precision matrix $\boldsymbol{\Lambda}(\gamma^2, \boldsymbol{\tau}_f^2, \mathcal{T})$, which has nonzero entries only for topic pairs that have a direct parent–child relationship.[2] The log-conditional posterior is

$$\log p(\boldsymbol{\mu}_f | \boldsymbol{W}, \boldsymbol{I}, \boldsymbol{l}, \{\boldsymbol{\tau}_f^2\}_{f=1}^V, \psi, \gamma^2, \nu, \sigma^2, \{\boldsymbol{\xi}_d\}_{d=1}^D, \mathcal{T}) = -\sum_{d=1}^D l_d \boldsymbol{\theta}_d^T \boldsymbol{\beta}_f + \sum_{d=1}^D w_{fd} \log(\boldsymbol{\theta}_d^T \boldsymbol{\beta}_f) - \frac{1}{2}(\boldsymbol{\mu}_f - \psi \mathbf{1})^T \boldsymbol{\Lambda} (\boldsymbol{\mu}_f - \psi \mathbf{1}).$$

We use HMC to sample from this unnormalized density. Note that the covariate matrix $\boldsymbol{\Theta}_{D \times K}$ is very sparse in most cases, so we speed computation with a sparse matrix representation.

We know the conditional distribution of the variance parameters due to the conjugacy of the Inverse-χ^2 prior with the normal distribution of the log rates. Specifically, if $\mathcal{C}(\mathcal{T})$ is the set of child topics of topic k with cardinality J, then

$$\tau_{fk}^2 | \boldsymbol{\mu}_f, \nu, \sigma^2, \mathcal{T} \sim \text{Inv-}\chi^2 \left(J + \nu, \frac{\nu\sigma^2 + \sum_{j \in \mathcal{C}} (\mu_{fj} - \mu_{fk})^2}{J + \nu} \right).$$

3.1.2 Updating Topic Affinity Parameters

In the second block, the conditional posterior of the topic affinity vectors factors by document:

$$p(\{\boldsymbol{\xi}_d\}_{d=1}^D | \boldsymbol{W}, \boldsymbol{I}, \boldsymbol{l}, \{\boldsymbol{\mu}_f\}_{f=1}^V, \boldsymbol{\eta}, \boldsymbol{\Sigma}) \propto \prod_{d=1}^D \left\{ \prod_{f=1}^V p(w_{fd} | \boldsymbol{I}_d, l_d, \mu_f, \boldsymbol{\xi}_d) \right\} \cdot p(\boldsymbol{I}_d | \boldsymbol{\xi}_d) \cdot p(\boldsymbol{\xi}_d | \boldsymbol{\eta}, \boldsymbol{\Sigma}).$$

We can again write the likelihood as a Poisson regression, now with the rates as covariates. The log-conditional posterior for one document is

$$\log p(\boldsymbol{\xi}_d | \boldsymbol{W}, \boldsymbol{I}, \boldsymbol{l}, \{\boldsymbol{\mu}_f\}_{f=1}^V, \boldsymbol{\eta}, \boldsymbol{\Sigma}) = -l_d \sum_{f=1}^V \boldsymbol{\beta}_f^T \boldsymbol{\theta}_d + \sum_{f=1}^V w_{fd} \log(\boldsymbol{\beta}_f^T \boldsymbol{\theta}_d) - \sum_{k=1}^K \log(1 + e^{-\xi_{dk}}) - \sum_{k=1}^K (1 - I_{dk}) \xi_{dk} - \frac{1}{2}(\boldsymbol{\xi}_d - \boldsymbol{\eta})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\xi}_d - \boldsymbol{\eta}).$$

We use HMC to sample from this unnormalized density. Here the parameter vector $\boldsymbol{\theta}_d$ is sparse rather than the covariate matrix $\boldsymbol{B}_{V \times K}$, with generic entry $B_{fk} \equiv \beta_{fk}$. If we remove the entries of $\boldsymbol{\theta}_d$ and columns of $\boldsymbol{B}$ pertaining to topics k where $I_{dk} = 0$, then we are left with a low-dimensional regression where only the active topics are used as covariates, greatly simplifying computation.

3.1.3 Updating Corpus-Level Parameters

We draw the hyperparameters after each iteration of the block update. We put flat priors on these unknowns so that we can learn their most likely values from the data. As a result, their conditional posteriors only depend on the latent variables they generate.

[2] In practice this precision matrix can be found easily as the negative Hessian of the log-prior distribution.

The log corpus-level rates $\mu_{f,0}$ for each word follow a Gaussian distribution with mean ψ and variance γ^2. The conditional distribution of these hyperparameters is available in closed form:

$$\psi|\gamma^2, \{\mu_{f,0}\}_{f=1}^V \sim \mathcal{N}\left(\frac{1}{V}\sum_{f=1}^V \mu_{f,0}, \quad \frac{\gamma^2}{V}\right),$$

$$\text{and} \quad \gamma^2|\psi, \{\mu_{f,0}\}_{f=1}^V \sim \text{Inv-}\chi^2\left(V, \quad \frac{1}{V}\sum_{f=1}^V (\mu_{f,0}-\psi)^2\right).$$

The discrimination parameters τ_{fk}^2 independently follow an identical scaled inverse-χ^2 with convolution parameter ν and scale parameter σ^2, while their inverse follows a Gamma($\kappa_\tau = \frac{\nu}{2}, \lambda_\tau = \frac{2}{\nu\sigma^2}$) distribution. We use HMC to sample from this unnormalized density. Specifically,

$$\begin{aligned}\log p(\kappa_\tau, \lambda_\tau|\{\boldsymbol{\tau}_f^2\}_{f=1}^V, \mathcal{T}) = (\kappa_\tau - 1)\sum_{f=1}^V \sum_{k\in\mathcal{P}} \log(\tau_{fk}^2)^{-1} \\ -|\mathcal{P}|V\kappa_\tau \log\lambda_\tau - |\mathcal{P}|V\log\Gamma(\kappa_\tau) \\ -\frac{1}{\lambda_\tau}\sum_{f=1}^V\sum_{k\in\mathcal{P}}(\tau_{fk}^2)^{-1},\end{aligned}$$

where $\mathcal{P}(\mathcal{T})$ is the set of parent topics on the tree. Each draw of $(\kappa_\tau, \lambda_\tau)$ is then transformed back to the (ν, σ^2) scale.

The document-specific topic affinity parameters $\boldsymbol{\xi}_d$ follow a multivariate normal distribution with mean parameter $\boldsymbol{\eta}$ and a covariance matrix parameterized in terms of a scalar, $\boldsymbol{\Sigma} = \lambda^2 \boldsymbol{I}_K$. The conditional distribution of these hyperparameters is available in closed form. For efficiency, we choose to put a flat prior on $\log\lambda^2$ rather than the original scale, which allows us to marginalize out $\boldsymbol{\eta}$ from the conditional posterior of λ^2:

$$\lambda^2|\{\boldsymbol{\xi}_d\}_{d=1}^D \sim \text{Inv-}\chi^2\left(DK-1, \quad \frac{\sum_d\sum_k(\xi_{dk}-\bar{\xi}_k)^2}{DK-1}\right),$$

$$\text{and} \quad \boldsymbol{\eta}|\lambda^2, \{\boldsymbol{\xi}_d\}_{d=1}^D \sim \mathcal{N}\left(\bar{\boldsymbol{\xi}}, \quad \frac{\lambda^2}{D}\boldsymbol{I}_K\right).$$

3.2 Estimation

As discussed in Section 2.3, our estimands are the topic-specific frequency and exclusivity of the words in the vocabulary, as well as the frequency-exclusivity (FREX) score that averages each word's performance in these dimensions. We use posterior means to estimate frequency and exclusivity, computing these quantities at every iteration of the Gibbs sampler and averaging the draws after the burn-in period. For the FREX score, we applied the ECDF function to the frequency and exclusivity posterior expectations of all words in the vocabulary to estimate the true ECDF.

3.3 Inference for Unlabeled Documents

To classify unlabeled documents, we need to find the posterior predictive distribution of the membership vector $\boldsymbol{I}_{\tilde{d}}$ for a new document $\tilde{d}$. Inference is based on the new document's word counts $\boldsymbol{w}_{\tilde{d}}$ and the unknown parameters, which we hold constant at their posterior expectation. Unfortunately, the posterior predictive distribution of the topic affinities $\boldsymbol{\xi}_{\tilde{d}}$ is intractable without conditioning on the label vector since the labels control which topics contribute content. We therefore use a simpler model where the topic proportions depend only on the relative size of the affinity parameters

$$\theta_{dk}^*(\boldsymbol{\xi}_d) \equiv \frac{e^{\xi_{dk}}}{\sum_{j=1}^K e^{\xi_{dj}}} \quad \text{and} \quad I_{dk} \sim \text{Bern}\left(\frac{1}{1+\exp(-\xi_{dk})}\right).$$

The posterior predictive distribution of this simpler model factors into tractable components

$$\begin{aligned}p^*(\boldsymbol{I}_{\tilde{d}}, \boldsymbol{\xi}_{\tilde{d}}|\boldsymbol{w}_{\tilde{d}}, \boldsymbol{W}, \boldsymbol{I}) &\approx p(\boldsymbol{I}_{\tilde{d}}|\boldsymbol{\xi}_{\tilde{d}})\ p^*(\boldsymbol{\xi}_{\tilde{d}}|\{\hat{\boldsymbol{\mu}}_f\}_{f=1}^V, \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\Sigma}}, \boldsymbol{w}_{\tilde{d}}) \\ &\propto p(\boldsymbol{I}_{\tilde{d}}|\boldsymbol{\xi}_{\tilde{d}})\ p^*(\boldsymbol{w}_{\tilde{d}}|\boldsymbol{\xi}_{\tilde{d}}, \{\hat{\boldsymbol{\mu}}_f\}_{f=1}^V) \\ &\quad \times p(\boldsymbol{\xi}_{\tilde{d}}|\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\Sigma}}).\end{aligned}$$

It is then possible to find the most likely $\boldsymbol{\xi}_{\tilde{d}}^*$ based on the evidence from $\boldsymbol{w}_{\tilde{d}}$ alone.

4. Results

We analyze the fit of the HPC model to Reuters Corpus Volume I (RCV1), a large collection of newswire stories. First, we demonstrate how the variance parameters τ_{fp}^2 regularize the exclusivity with which words are expressed within topics. Second, we show that regularization of exclusivity has the greatest effect on infrequent words. Third, we explore the joint posterior of the topic-specific frequency and exclusivity of words as a summary of topical content, giving special attention to the upper right corner of the plot where words score highly in both dimensions. We compare words that score highly on the FREX metric to top words scored by frequency alone, the current practice in topic modeling. Finally, we compare the classification performance of HPC to baseline models.

4.1 The Reuters Corpus Dataset

RCV1 is an archive of 806,791 newswire stories from a 12-month period during 1996–1997.[3] As described in Lewis et al. (2004), Reuters staffers assigned stories into any subset of 102 hierarchical topic categories. In the original data, assignment to any topic required automatic assignment to all ancestor nodes, but we removed these redundant ancestor labels since they do not allow our model to distinguish intentional assignments to high-level categories from assignment to their offspring. In our modified annotations, the only documents we see in high-level topics are those labeled to them and none of their children, which maps onto general content. We preprocessed document tokens with the Porter stemming algorithm (getting 300,166 unique stems) and chose the most frequent 3% of stems (10,421 unique stems, over 100 million total tokens) for the feature set.[4]

The Reuters topic hierarchy has three levels that divide the content into finer categories at each cut. At the first level, content is divided between four high-level categories: three that focus on

[3] Available upon request from the National Institute of Standards and Technology (NIST), *http://trec.nist.gov/data/reuters/reuters.html*.

[4] Including rarer features did not meaningfully change the results.

business and market news (Markets, Corporate/Industrial, and Economics) and one grab bag category that collects all remaining topics from politics to entertainment (Government/Social). The second level provides fine-grained divisions of these broad categories and contains the terminal nodes for most branches of the tree. For example, the Markets topic is split between equity, bond, money, and commodity markets at the second level. The third level offers further subcategories where needed for a small set of second-level topics. For example, the Commodity Markets topic is divided between agricultural (soft), metal, and energy commodities. We present a graphical illustration of the Reuters topic hierarchy in Figure 2.

Many documents in the Reuters corpus are labeled to multiple topics, even after redundant ancestor memberships are removed. Overall, 32% of the documents are labeled to more than one node of the topic hierarchy. Fifteen percent of documents have very diverse content, being labeled to two or more of the main branches of the tree (Markets, Commerce, Economics, and Government/Social). Twenty-one percent of documents are labeled to multiple second-level categories on the same branch (e.g., bond markets and equity markets in the Markets branch). Finally, 14% of documents are labeled to multiple children of the same second-level topic (e.g., metals trading and energy markets in the commodity markets branch of Markets). Therefore, a completely general mixed membership model such as HPC is necessary to capture the labeling patterns of the corpus. A full breakdown of membership statistics by topic is presented in Tables 2 and 3.

4.2 How the Differential Usage Parameters Regulate Topic Exclusivity

A word can only be exclusive to a topic if its expression across the sibling topics is allowed to diverge from the parent rate. Therefore, we would only expect words with high differential usage parameters τ^2_{fp} at the parent level to be candidates for highly exclusive expression ϕ_{fk} in any child topic k. Words with child topic rates that cannot vary greatly from the parent should have nearly equal expression in each child k, meaning $\phi_{fk} \approx \frac{1}{C}$ for a branch with C child topics. An important consequence is that, although the ϕ_{fk} are not directly modeled in HPC, their distribution is regularized by positing a prior distribution on the τ^2_{fp}.

This tight relation can be seen in the HPC fit. Figure 3 shows the joint posterior expectation of the differential usage parameters in a parent topic and exclusivity parameters across the child topics. Specifically, the left panel compares the rate variance of the children of Markets from their parent to exclusivity between the child topics; the right panel does the same with the two children of Performance, a second-level topic under the Corporate category. The plots have similar patterns. For low levels of differential expression, the exclusivity parameters are clustered around the baseline value, $\frac{1}{C}$. At high levels of child rate variance, words gain the ability to approach exclusive expression in a single topic.

4.3 How Frequency Modulates Regularization of Exclusivity

One of the most appealing aspects of regularization in generative models is that it acts most strongly on the parameters for which we have the least information. In the case of the exclusivity parameters in HPC we have the most data for frequent words, so for a given topic the words with low rates should be most affected by regularization of their exclusivity parameters—by means of the proposed shrinkage prior on the parent node's variance parameter τ^2_{fp}.

Figure 4 shows for two topics the joint posterior expectation of each word's frequency in that topic and its exclusivity compared to sibling topics (the FREX plot, henceforth). The left panel features the Science and Technology topic, a child in the grab bag Government/Social branch, and the right panel features the Research/Development topic, a child in the Corporate branch. The overall shape of the joint posterior is very similar for both topics. On the left side of the plots, the exclusivity of rare words is unable to significantly exceed the $\frac{1}{C}$ baseline. This is because the model does not have much evidence to estimate usage in the topic, so the estimated rate is shrunk heavily toward the parent rate. However, we see that it is possible for rare words to be underexpressed in a topic, which happens if they are frequent and overexpressed in a sibling topic. Even though their rates are similar to the parent in this topic, sibling topics may have a much higher rate and account for most appearances of the word in the comparison group.

4.4 Frequency and Exclusivity are Two Key Dimensions of Semantic Content

Words in the upper right of the FREX plot—those that are both frequent and highly exclusive—are of greatest interest. These are the most common words in the corpus that are also likely to have been generated from the topic of interest (rather than similar topics). We show words in the upper 5% quantiles in both dimensions for our example topics in Figure 5. In particular, words on the left end of these scatterplots are the least frequent, highly exclusive words, and may not appear in topic summaries based on frequency alone. These high-scoring words can help to clarify content even for labeled topics. In the Science and Technology topic, we see almost all terms are specific to the American and Russian space programs. Similarly, in the Research/Technology topic, almost all terms relate to clinical trials in medicine or to agricultural research.

We also compute the FREX score for each word-topic pair, a univariate summary of topical content that averages performance in both dimensions. In Figure 6 we compare the top FREX words in three topics to a ranking based on frequency alone, which is the current practice in topic modeling. For context, we also show the immediate neighbors of each topic in the tree. The topic being examined is in bolded red, while the borders of the comparison set are solid. The Defense Contracts topic is a special case since it is an only child. In these cases, we use a comparison to the parent topic to calculate exclusivity.

By incorporating exclusivity information, FREX-ranked lists include fewer words that are used similarly everywhere (such as *said* and *would*) and fewer words that are used similarly in a set of related topics (such as *price* and *market* in the Markets branch). One can understand this result by comparing the rankings for known stop words from the SMART list to other words. In Figure 7, we show the maximum ECDF ranking for each word across topics in the distribution of frequency (left panel) and

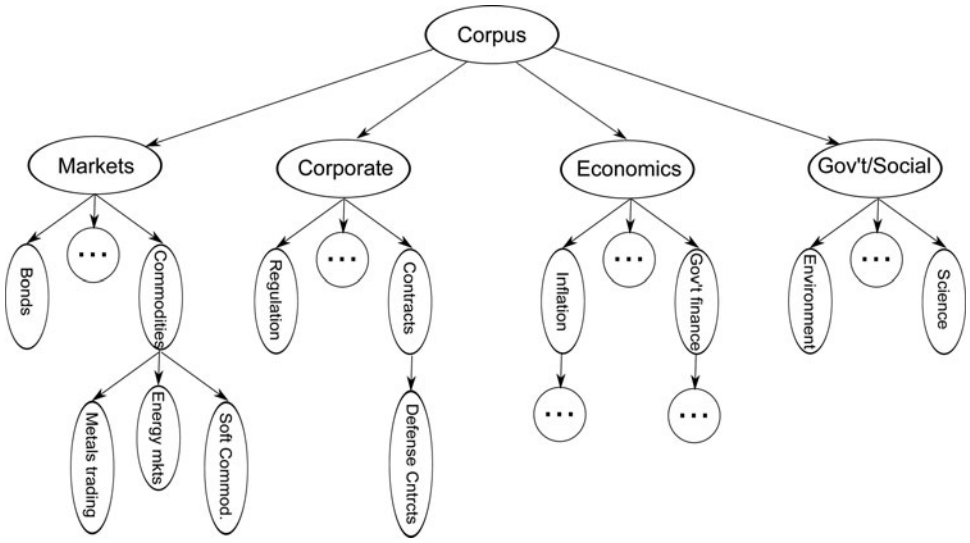

Figure 2. Topic hierarchy of Reuters corpus.

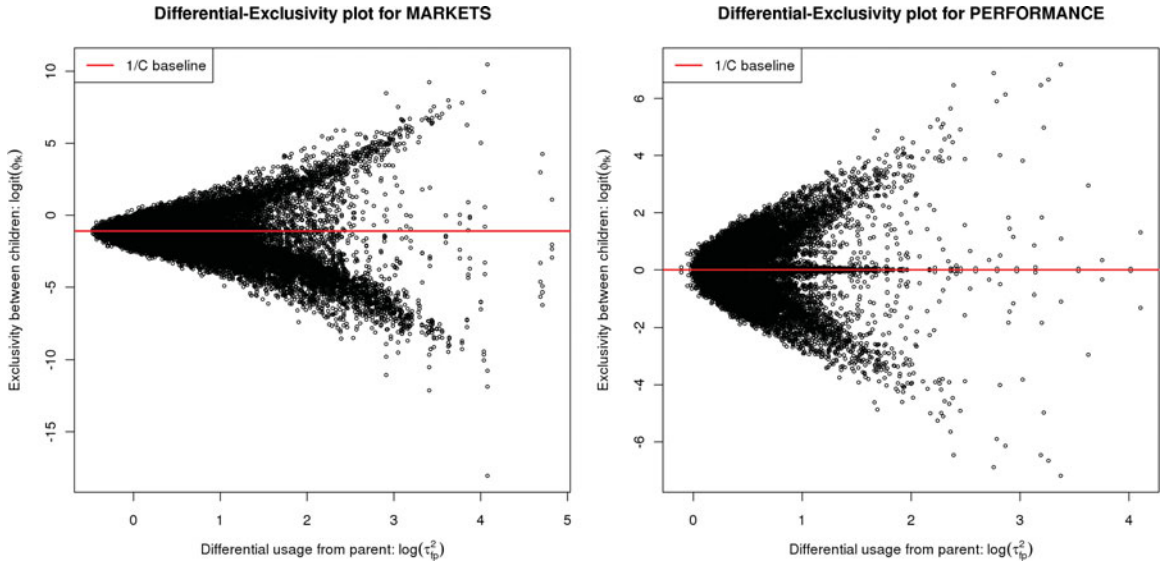

Figure 3. Exclusivity as a function of differential usage parameters.

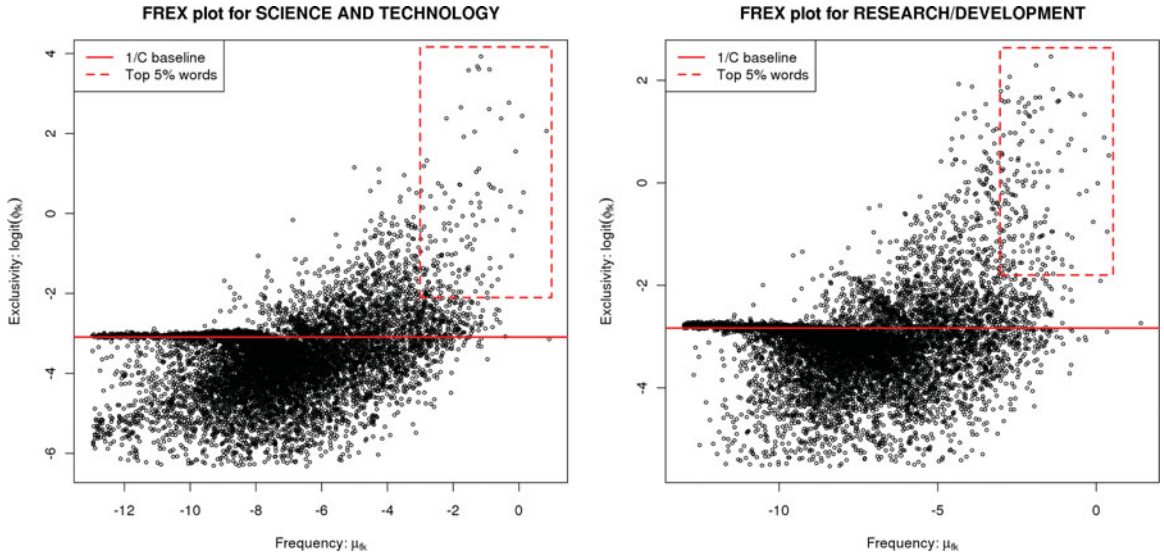

Figure 4. Frequency-exclusivity (FREX) plots. The red boxes were drawn by selecting words that had both a top 5% exclusivity score and a top 5% frequency score, calculated independently.

Table 2. Topic membership statistics.

Topic code	Topic name	# docs	Any MM	CB L1 MM	CB L2 MM	CB L3 MM
CCAT	Corporate/Industrial	2170	79.60%	79.60%	13.10%	0.80%
C11	Strategy/Plans	24,325	51.50	11.50	44.50	4.50
C12	Legal/Judicial	11,944	99.20	98.90	50.20	1.70
C13	Regulation/Policy	37,410	85.90	55.60	61.40	4.50
C14	Share Listings	7410	30.30	7.90	10.30	15.80
C15	Performance	229	82.10	35.80	74.20	1.70
C151	Accounts/Earnings	81,891	7.90	1.30	0.60	6.40
C152	Comment/Forecasts	73,092	18.90	4.80	1.60	13.50
C16	Insolvency/Liquidity	1920	66.70	31.50	54.60	3.60
C17	Funding/Capital	4767	78.10	41.40	67.70	5.00
C171	Share Capital	18,313	44.60	3.20	1.70	41.50
C172	Bonds/Debt Issues	11,487	15.10	5.70	0.30	9.70
C173	Loans/Credits	2636	24.70	8.50	3.60	15.60
C174	Credit Ratings	5871	65.60	59.00	0.50	7.50
C18	Ownership Changes	30	76.70	23.30	76.70	3.30
C181	Mergers/Acquisitions	43,374	34.40	6.50	4.80	26.90
C182	Asset Transfers	4671	28.30	4.70	5.70	21.00
C183	Privatizations	7406	73.70	34.20	6.30	44.10
C21	Production/Services	25,403	76.40	46.50	53.60	0.80
C22	New Products/Services	6119	55.00	15.30	49.10	0.40
C23	Research/Development	2625	77.00	36.40	57.80	0.90
C24	Capacity/Facilities	32,153	72.20	33.60	58.40	0.90
C31	Markets/Marketing	29,073	46.90	25.30	34.60	1.30
C311	Domestic Markets	4299	80.60	73.70	9.50	18.70
C312	External Markets	6648	78.10	70.40	9.60	14.20
C313	Market Share	1115	39.70	10.30	5.10	27.80
C32	Advertising/Promotion	2084	63.80	26.90	52.50	1.40
C33	Contracts/Orders	14,122	48.00	12.60	40.50	0.80
C331	Defense Contracts	1210	68.00	65.50	13.30	3.40
C34	Monopolies/Competition	4835	92.30	54.90	75.70	14.00
C41	Management	1083	75.60	52.10	59.90	2.00
C411	Management Moves	10,272	17.70	9.60	2.40	8.20
C42	Labor	11,878	99.70	99.60	46.50	1.50
ECAT	Economics	621	90.50	90.50	9.70	1.40
E11	Economic Performance	8568	43.00	24.20	29.10	5.10
E12	Monetary/Economic	24,918	81.70	75.40	17.90	13.70
E121	Money Supply	2182	30.50	23.10	0.70	9.20
E13	Inflation/Prices	130	60.00	46.90	28.50	0.80
E131	Consumer Prices	5659	24.70	15.60	6.00	12.00
E132	Wholesale Prices	939	19.00	3.40	0.60	16.90
E14	Consumer Finance	428	73.80	43.20	61.00	1.60
E141	Personal Income	376	75.00	63.80	9.60	22.30
E142	Consumer Credit	200	46.00	30.00	3.50	18.50
E143	Retail Sales	1206	27.50	19.70	2.40	10.20
E21	Government Finance	941	86.70	81.40	53.90	4.00
E211	Expenditure/Revenue	15,768	78.20	72.40	16.10	13.80
E212	Government Borrowing	27,405	32.70	29.60	2.70	4.50
E31	Output/Capacity	591	45.20	18.30	35.20	0.50
E311	Industrial Production	1701	17.70	9.80	3.10	9.30
E312	Capacity Utilization	52	65.40	13.50	3.80	57.70
E313	Inventories	111	26.10	10.80	0.00	16.20
E41	Employment/Labor	14,899	100.00	100.00	49.40	2.20
E411	Unemployment	2136	92.00	90.60	10.40	12.00
E51	Trade/Reserves	4015	85.10	75.50	38.70	1.90
E511	Balance of Payments	2933	63.80	43.70	8.20	25.70
E512	Merchandise Trade	12,634	64.90	59.10	11.50	11.70
E513	Reserves	2290	30.10	22.70	1.30	16.80
E61	Housing Starts	391	51.70	47.80	13.80	0.80
E71	Leading Indicators	5270	2.90	0.60	2.40	0.20

NOTE: MM = Mixed membership; CB Lx = Cross-branch MM at level x.

exclusivity (right panel) estimates. One can see that while stop words are more likely to be in the extreme quantiles of frequency, very few of them are among the most exclusive words. This prevents general and context-specific stop words from ranking highly in a FREX-based index.

4.5 Classification Performance

We compare the classification performance of HPC with a support vector machine (SVM), a L2-regularized logistic regression and labeled LDA (Ramage et al. 2009), on both the Reuters corpus and the New York Times corpus (Sandhaus 2008). All methods were trained on a random sample of 15% of the documents using the 3% most frequent words in the corpus as features. These fits were used to predict memberships in the withheld documents. This out-of-sample prediction experiment was repeated 10 times with a new random sample as a training set. More in detail, we used a stratified sampling technique to get a balanced sample (across topics) for training, validation, and test partitions with a 15/25/60 split, respectively. We fit the

Table 3. Topic membership statistics, continued.

Topic code	Topic name	# docs	Any MM	CB L1 MM	CB L2 MM	CB L3 MM
GCAT	Government/Social	24,546	2.50	2.50	0.50	0.10
G15	European Community	1545	16.10	6.90	14.60	0.00
G151	EC Internal Market	3307	98.00	87.20	10.60	94.30
G152	EC Corporate Policy	2107	96.70	90.70	40.30	50.30
G153	EC Agriculture Policy	2360	96.10	94.20	31.40	27.70
G154	EC Monetary/Economic	8404	98.20	93.00	11.50	43.90
G155	EC Institutions	2124	70.80	42.00	24.30	54.00
G156	EC Environment Issues	260	75.00	57.70	28.80	50.80
G157	EC Competition/Subsidy	2036	100.00	99.80	60.20	32.50
G158	EC External Relations	4300	80.70	62.80	27.00	24.80
G159	EC General	40	47.50	17.50	35.00	2.50
GCRIM	Crime, Law Enforcement	32,219	79.50	41.60	59.40	0.90
GDEF	Defense	8842	93.70	17.20	84.40	0.50
GDIP	International Relations	37,739	73.70	20.50	60.70	0.90
GDIS	Disasters and Accidents	8657	75.70	40.10	52.20	0.20
GENT	Arts, Culture, Entertainment	3801	68.80	29.20	49.60	0.50
GENV	Environment and Natural World	6261	90.20	51.50	72.30	2.50
GFAS	Fashion	313	76.40	45.70	41.50	1.90
GHEA	Health	6030	81.90	56.10	65.00	1.20
GJOB	Labor Issues	17,241	99.60	99.40	44.60	3.30
GMIL	Millennium Issues	5	100.00	100.00	40.00	0.00
GOBIT	Obituaries	844	99.40	15.30	99.40	0.00
GODD	Human Interest	2802	60.70	9.70	55.20	0.10
GPOL	Domestic Politics	56,878	79.60	29.70	63.00	1.80
GPRO	Biographies, Personalities, People	5498	87.50	10.00	84.70	0.10
GREL	Religion	2849	86.10	6.60	84.30	0.10
GSCI	Science and Technology	2410	55.20	22.20	45.10	0.30
GSPO	Sports	35,317	1.30	0.60	0.90	0.00
GTOUR	Travel and Tourism	680	89.60	69.70	34.70	3.40
GVIO	War, Civil War	32,615	67.30	10.10	64.60	0.10
GVOTE	Elections	11,532	100.00	13.30	100.00	1.30
GWEA	Weather	3878	73.90	46.80	46.40	0.10
GWELF	Welfare, Social Services	1869	95.40	75.50	74.10	3.40
MCAT	Markets	894	81.10	81.10	14.50	2.20
M11	Equity Markets	48,700	16.30	12.30	3.90	2.90
M12	Bond Markets	26,036	21.30	15.60	5.20	3.50
M13	Money Markets	447	65.80	51.90	23.30	1.60
M131	Interbank Markets	28,185	15.10	9.40	0.70	6.40
M132	Forex Markets	26,752	36.90	24.70	3.10	16.10
M14	Commodity Markets	4732	18.00	16.70	2.30	0.10
M141	Soft Commodities	47,708	24.10	22.80	5.50	2.00
M142	Metals Trading	12,136	34.70	19.30	4.10	16.10
M143	Energy Markets	21,957	21.10	18.40	4.80	2.90

NOTE: MM = Mixed membership, CB Lx = Cross-branch MM at level x.

four models to each training set and then used the validation set to calibrate a threshold, except for SVM. We used the fit from the training set and the threshold from the validation set to predict topic memberships in the test set. We trained SVM using both the training and validation data, since it does not need a threshold. Table 4 shows the results of these experiments, using both micro averages (every document weighted equally) and macro averages (every topic weighted equally).

HPC compares comparably with SVM on average, dominating on the New York Times corpus, while loosing only to SVM on the Reuters corpus. Labeled LDA displays a better performance than regularized logistic regression, but loses consistently to both HPC and SVM. HPC is not designed for optimizing predictive accuracy out-of-sample, rather it is designed to maximize interpretability of the label-specific summaries, in terms of words that are both frequent and exclusive. Neither is labeled LDA. Additional performance gain in prediction tasks for any generative model may be achieved by training such models discriminatively (Zhu, Ahmed, and Xing 2012). The results offer a quantitative illustration of the trade-off between predictive and explanatory power of statistical models (Breiman 2001). For an additional comparative performance evaluation focused on LDA-based models and SVMs we refer interested readers to Rubin et al. (2012).

4.6 Experiments with Human Evaluators

The data analysis in Sections 4.2–4.4 suggest a few hypotheses that warrant further exploration. First, the results suggest FREX summaries improve the interpretability of the topic summaries specified in terms of lists of words. Second, the results suggest the proposed parameterization and regularization scheme for the rates of word occurrence lead to estimates of frequency and exclusivity that are less affected by sampling variations. These in turn translate into further improvements in the interpretability of topic summaries, thus creating a synergistic effect. In addition, in light of these hypotheses, it is plausible to expect that the variability of the estimates of exclusivity are larger, thus less stable, for models that posit regularization of word rates within a topic than for models in the proposed model, which regularizes the rates of the same word across topics.

Table 4. Classification performance for 10-fold cross-validation. Standard deviation of precision and recall over 10 folds is reported in parenthesis.

Data model	SVM	L2-reg Logit	Labeled LDA	HPC
		Reuters corpus		
Micro-ave precision	0.711 (0.002)	0.195 (0.031)	0.487 (0.027)	0.695 (0.007)
Micro-ave recall	0.706 (0.001)	0.768 (0.013)	0.543 (0.019)	0.589 (0.008)
Macro-ave precision	0.563 (0.002)	0.481 (0.025)	0.344 (0.082)	0.505 (0.094)
Macro-ave recall	0.551 (0.006)	0.600 (0.007)	0.476 (0.078)	0.524 (0.093)
		New York Times corpus		
Micro-ave precision	0.822 (0.001)	0.378 (0.029)	0.670 (0.023)	0.891 (0.003)
Micro-ave recall	0.785 (0.001)	0.732 (0.010)	0.714 (0.015)	0.846 (0.004)
Macro-ave precision	0.657 (0.002)	0.512 (0.022)	0.524 (0.043)	0.729 (0.054)
Macro-ave recall	0.541 (0.004)	0.580 (0.008)	0.638 (0.039)	0.784 (0.043)

In this section, we present the results of a large experiment on Amazon Turk that enlists human evaluators to test the hypotheses we outlined above quantitatively.

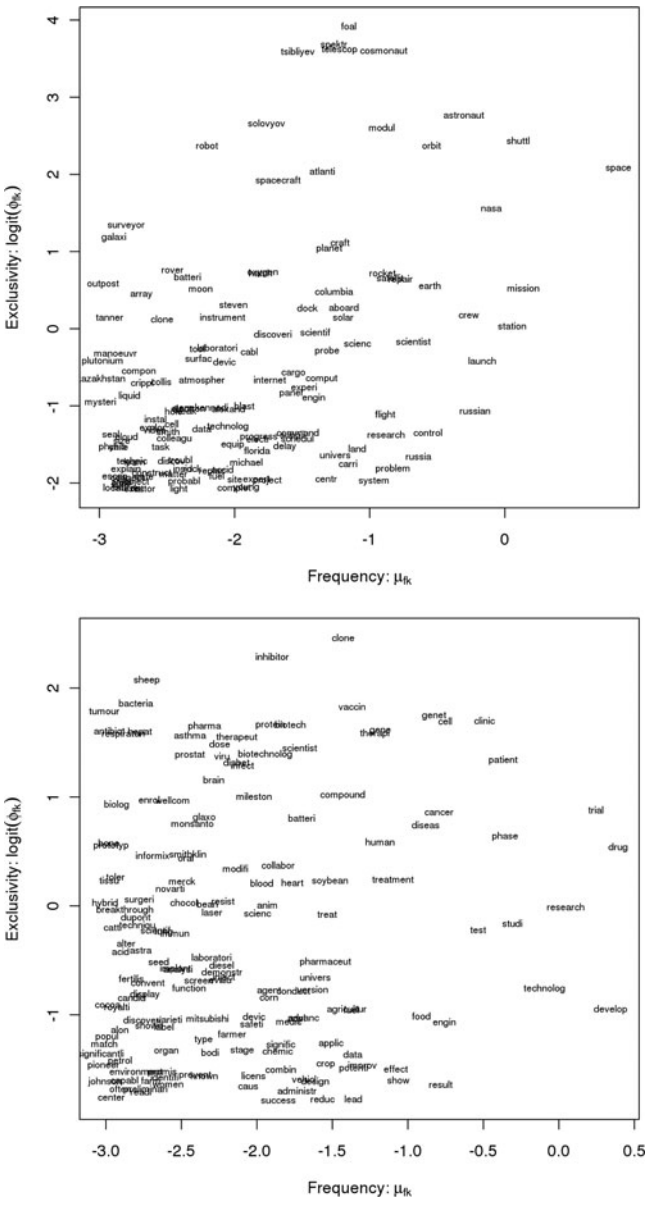

Figure 5. The two panels are zoomed-in versions of the red boxes in the corresponding panels of Figure 4: namely, the upper right corner of FREX plot for Science and Technology (top) and of the Research/Development (bottom).

4.6.1 Design Choices

The goal is to design a randomized experiment with human evaluators to test the hypotheses outlined above, about the implications of the regularization scheme on the rates of occurrence by word on the interpretability of topic summaries and on the estimates of differential usage. To avoid confounding in the results due to the complexity of the model presented in Table 1, we chose to build the experiment around a simplified model that leverages the regularization scheme of interest. We consider two competing models for the randomized experiment: Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003), the simplest model that posits a regularization scheme on the rates of occurrence of *different words within a topic*, and a simpler unsupervised variant of the proposed HPC model that regularizes the rates of occurrence of *the same word across topics*, without the hierarchy on the rates of word occurrence. Thus, the simpler model we are going to use in this section consists of the portion of the HPC model labeled as data generation in Table 1, with the addition of the assumption that $\boldsymbol{\theta}_d \sim$ Dirichlet (α), both to complete the specifications, and to avoid further confounding by using the assumption on the mixed membership vectors $\boldsymbol{\theta}_d$ used by the latent Dirichlet allocation model, used as baseline in the experiment.

We quantify interpretability in two ways: indirectly, in terms of topic diversity, measured by the number of unique words that appear in 5-, 10-, 25-, and 50-word topic summaries, and directly, in terms of evaluator preferences for the topic summary of a model over that of another. We quantified stability of the estimates of exclusivity both directly, in terms of variance of the estimated word rates, and indirectly, in terms of the maximum exclusivity of a word across topics.

We fit both models to 2246 documents from Associated Press corpus (Harman 1992). We performed some preprocessing steps to maximize the interpretability of inferred topics, and to avoid confounding in the results due to the different impact of stop words on the word rates inferred with different models. First, we removed all stop words on the SMART stop list to prevent obvious filler words from dominating topic summaries.[5] We also removed all proper nouns using the part-of-speech tagger in the Python Natural Language Toolkit (NLTK) so that topic summaries do not require encyclopedic knowledge of people and places to be interpretable. For both models we varied the number of topics across a wide range, with $K \in \{10, 25, 50, 100\}$.

[5] This list is available at *www.jmlr.org/papers/volume5/lewis04a/a11-smart-stop-list/english.stop*.

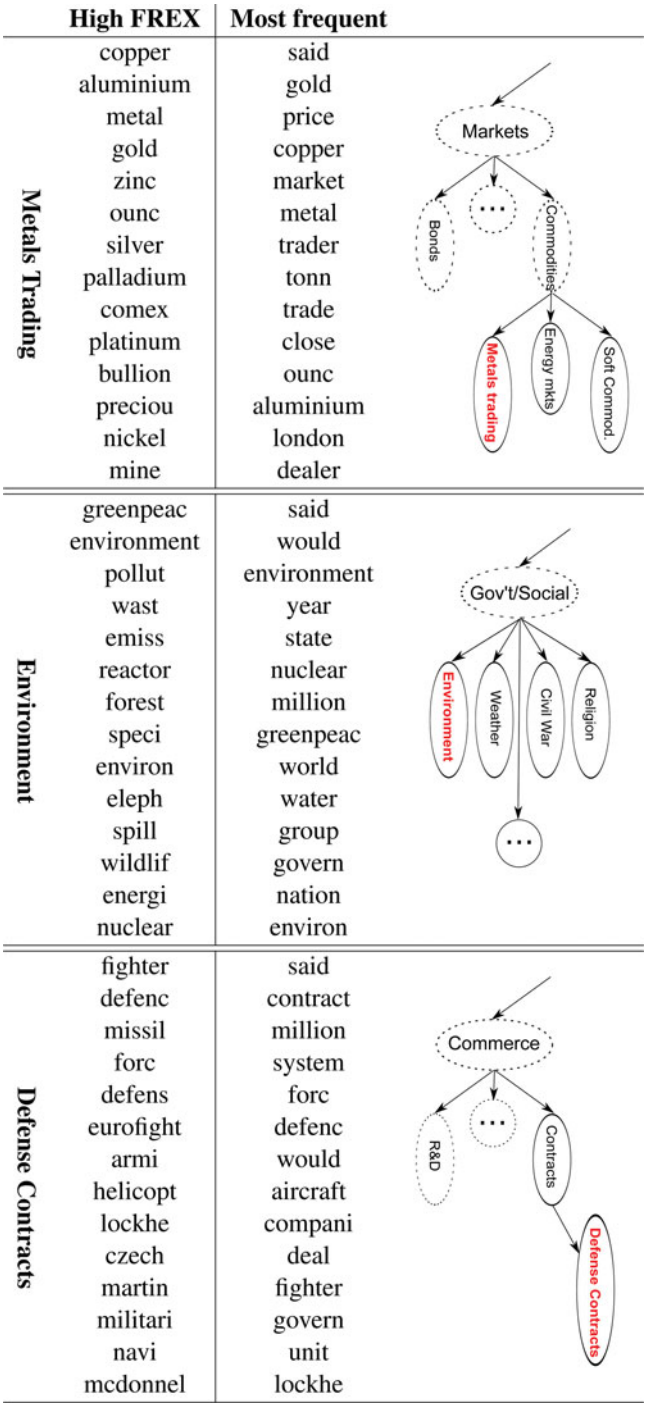

Figure 6. Comparison of high FREX words (both frequent and exclusive) to most frequent words (featured topic name bold red; comparison set in solid ovals).

We sought to compare FREX-based topic summaries obtained with the proposed model, to topic summaries obtained with LDA, to FREX-based topic summaries where the estimates of frequency and exclusivity are obtained by leveraging LDA word rate estimates—thus, effectively using the proposed FREX score to rerank the words associated with each topic according to LDA.

The latent Dirichlet allocation model is parameterized in terms of the probability of a word given a topic, not the topic given a word. To compute the FREX score from LDA word rate estimates, we need to reverse this conditioning. A simple calculation involving the marginal probability of each topic is necessary. Specifically,

$$p(\text{topic } k|\text{word } f) = \frac{p(\text{word } f|\text{topic } k)p(\text{topic } k)}{\sum_{j=1}^{K} p(\text{word } f|\text{topic } j)p(\text{topic } j)}. \quad (1)$$

Since LDA uses a symmetric Dirichlet prior on the topic membership probabilities, the marginal topic probabilities are equal. Therefore, the conditional distributions are equal and no correction is needed. However, for more complicated models where topic probabilities can be unequal (e.g., see Blei 2012), a posterior estimate of this inverse probability is required to get the FREX score.

4.6.2 Diversity in the Inferred Topics

The analysis in Section 4.4 suggested that the FREX score helps produce more diverse topical summaries. A set of topics that do not overlap in their word summaries arguably provides a more interpretable thematic structure underlying a given collection of documents. Here, we compare the diversity of topical summaries obtained with the proposed approach and with LDA.

One simple metric for quantifying the diversity between topic summaries is the proportion of unique words across all the top-word summaries produced from a model fit. For example, five-word summaries from a 100-topic model would have at most 500 unique words, and the proportion of the total achieved is an indication for whether the word lists are presenting diverse information. Table 5 shows this proportion for 5-, 10-, 25-, and 50-word summaries obtained with strategies of interest: ranking words by FREX scores estimated using the proposed Poisson convolution model (PCM FREX in the table), reranking words by FREX score estimated leveraging LDA word rate estimates (LDA FREX in the table), and ranking words by frequency using LDA word rate estimates (LDA FREQ in the table).

The results in Table 5 show that, in the word summaries obtained with the proposed model (PCM) in combination with the FREX score, over 90% the top words are unique, or equivalently, assigned to a single topic. This happens independently of the number of words in the summary, and of the number of topics in the range we considered—which is typical for topic

Table 5. Proportion of unique words in short topic summaries obtained with different strategies.

N topics	10	25	50	100
	(a) 5-word summaries			
PCM FREX	1.000	1.000	1.000	0.998
LDA FREX	1.000	1.000	1.000	0.974
LDA FREQ	0.820	0.752	0.612	0.522
	(b) 10-word summaries			
PCM FREX	1.000	1.000	0.998	0.989
LDA FREX	1.000	1.000	0.990	0.948
LDA FREQ	0.790	0.744	0.594	0.462
	(c) 25-word summaries			
PCM FREX	1.000	0.998	0.978	0.924
LDA FREX	1.000	0.997	0.942	0.846
LDA FREQ	0.744	0.650	0.493	0.384
	(d) 50-word summaries			
PCM FREX	1.000	0.997	0.977	0.907
LDA FREX	1.000	0.985	0.934	0.826
LDA FREQ	0.678	0.553	0.448	0.384

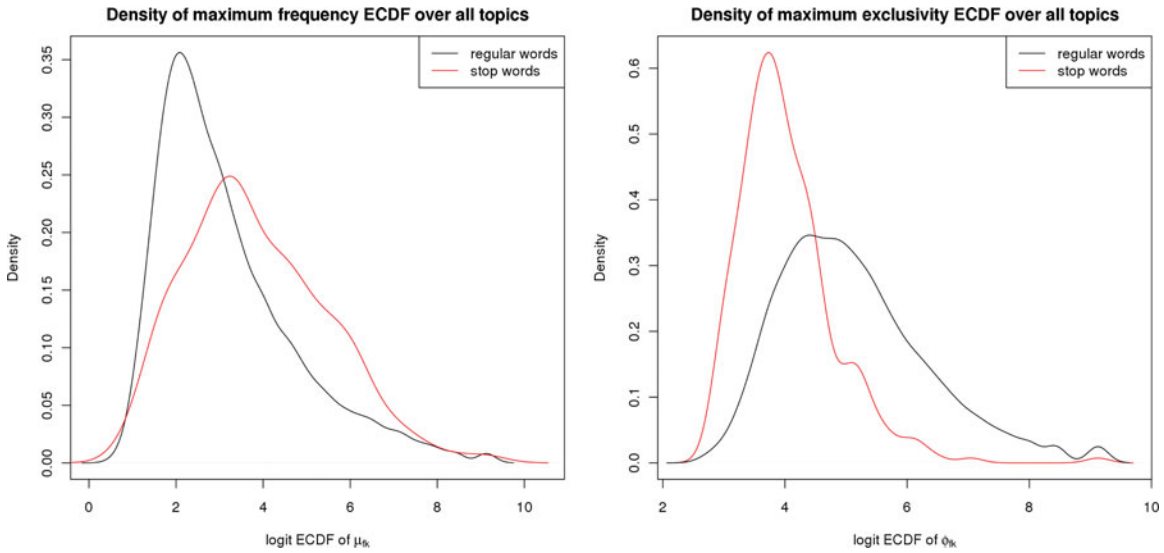

Figure 7. Comparison of FREX score components for SMART stop words versus regular words.

models. When restricting to top-5 and top-10 word summaries, almost all the words are assigned to a single topic. For frequency-based LDA summaries, in contrast, the proportion of unique words drops off quickly as the length of the summaries and number of topics increases. For example, even in 5-word summaries of a 100-topic model, only half the words are unique to any given topic. This repetition makes it more difficult to understand distinct thematic concepts reflected in each topic and may reduce the interpretability of the model fit. Using a FREX-based summary with LDA word rate estimates to rerank topic summaries does increase the proportion of unique words. The gains in topic diversity, however, become less pronounced both as the length of the summaries and the number of topics increase. These results can be explained, in part, by the fact that FREX-based summaries do not share nontopical "filler" words across topics, which can dominate their frequency-based summaries. FREX scores also increase the diversity in the topic summaries by promoting less common words that only occur in a given topic.

In the next section, we use human evaluators to determine the extent to which topic summaries containing a larger fraction of unique words convey more interpretable themes.

4.6.3 *A Randomized Experiment to Compare the Interpretability of Topic Summaries*

The two compelling hypotheses that the previous experiments and analyses suggest fairly strongly are that topic summaries based on the FREX score are more interpretable than currently established frequency-based summaries, and that the proposed model produces estimates of the FREX scores that are superior to those obtained from LDA. However, interpretability is hard to quantify, and it is difficult to develop automated methods that are reliable proxies for human judgment. For instance, recent research has found that out-of-sample likelihood is negatively correlated with human judgments of topic interpretability (Chang et al. 2009). Here we report the results of large randomized experiment we conducted on Amazon Mechanical Turk (aws.amazon.com/mturk) that aims at leveraging human evaluators to execute a comparative analysis of the interpretability of topic summaries, obtained with the three different strategies we have been considering. The experiment consists of evaluation tasks that require participants to interact with the output of different models, in a way that tests their ability to extract coherent themes from the topic summaries these models produce (Newman et al. 2010; Aletras and Stevenson 2013; Jia et al. 2014).

We implement two human evaluation tasks with users from Amazon Turk that both involve comparing the threes strategies of interest for producing topic summaries (PCM FREX, LDA FREQ, and LDA FREX, using abbreviations established in Section 4.6.2) to test the two hypotheses about model interpretability outlined above. We refer to the first task as the "word intrusion" task (Chang et al. 2009). This task measures the coherence of topic summaries by asking users to find which word does not belong in a topic summary; the intruder word is chosen among the words that are highly associated with another topic. Intuitively, intruder words will be easiest to identify in summaries that express clear and distinct themes. We refer to the second task as the "topic coherence" task (Newman et al. 2010). This task involves directly asking users to rate the coherence of a topic summary on a 1–3 scale. In addition, to get a clearer picture of the relative value of the three strategies to produce topic summaries we consider, after asking users to rate summaries from each of the methods, we also ask them to identify the most coherent summary among them, with an option for stating they have no preference.

Figure 8(a) shows an example of a word intrusion task. Each of the questions presents—for a single strategy—the top five scoring words in a random topic and along with an intruder word from the top-20 scoring words in one of the other topics. The order of words in the list is shuffled randomly before being presented to the user, who is asked to identify the intruder. Each task has six questions—exactly two from each strategy—also presented in a random order. All the 5-word topic summaries being compared in a task come from models with the same number of topics. The estimand of interest is the probability of correctly identifying the intruder word associated with each strategy to produce topic summaries. We considered models with 10, 25, 50, and 100 topics. For each model size, we gave

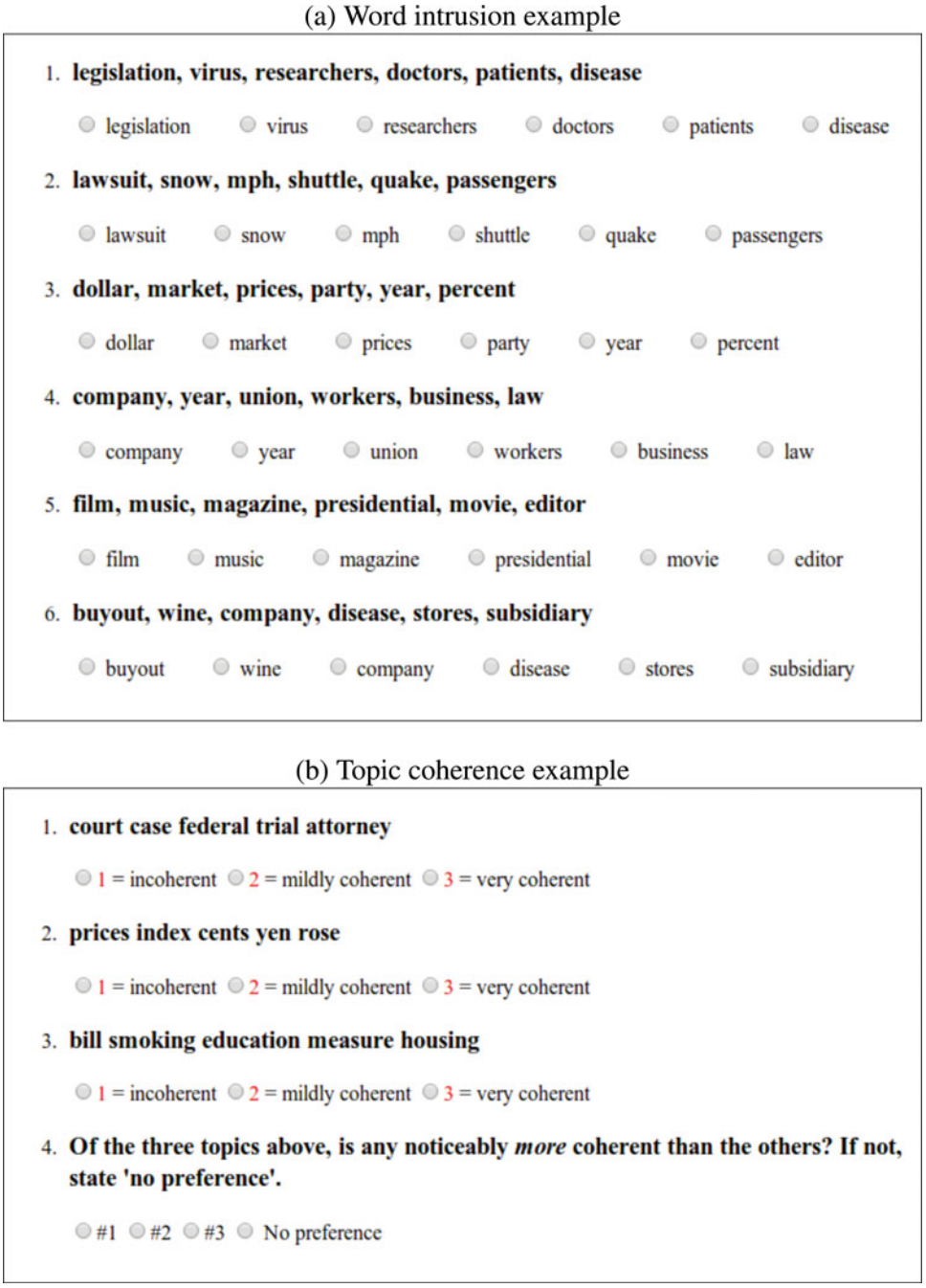

Figure 8. Screenshots of Amazon Turk tasks.

the task to 400 users, resulting in a total 3200 responses for each of the strategy to produce topic summaries.

Figure 8(b) shows an example of a topic coherence task. The first three questions provide a randomly chosen summary from each of the strategies and asks the user to rate it on a 1–3 scale. The order of summaries is randomized. Several examples of coherent and incoherent topics are provided to users in an included rubric. The final question asks the user if any of the three summaries are noticeably more coherent than the others to gauge the relative interpretability of the strategies. Included is an option to express no preference so that users do not choose arbitrarily whenever there is not an obvious top choice. The two estimands of interest are the average rating for each type of strategy, and the probability of each strategy being the most coherent. We considered models with 10, 25, 50, and 100 topics. For each model size, we gave the task to 400 users, resulting in a total 1600 responses about absolute coherence and 1600 responses about relative coherence for each of the strategy to produce topic summaries.

Figure 9 shows the results for the word intrusion task. In the plot we compare the probability of a user finding the intruder word across all three strategies as a function of the number of topics in the model. The performance for frequency-based summaries using LDA is consistently low, with the detection probability at 0.5 for small topic spaces and falling to 0.4 for a 100-topic model. Reranking LDA topic summaries using the FREX scores only improves performance for models with small number of topic, with the probability of finding the intruder

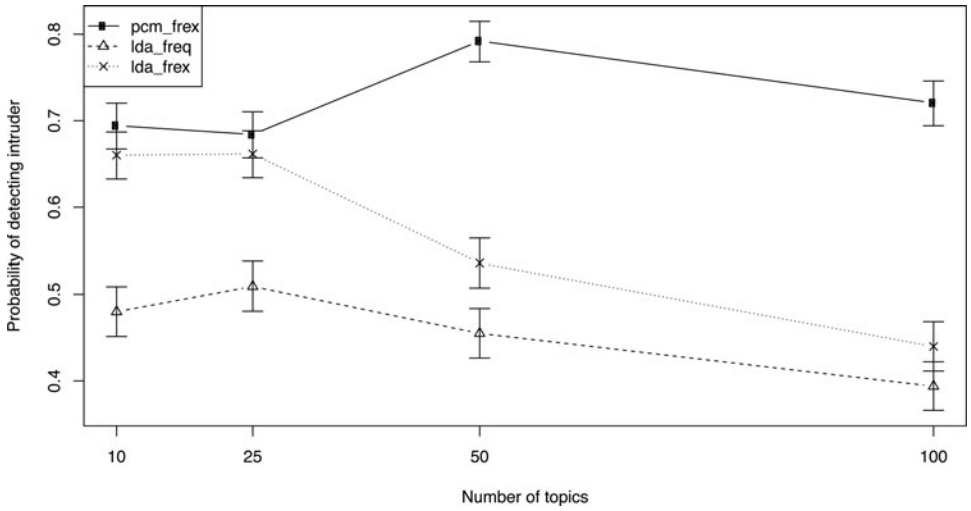

Figure 9. Results from Amazon Turk word intrusion task.

word nearly equal that of the LDA frequency-based summary for models with 50 and 100 topics. In contrast, users consistently detect the intruding words with high probability—between 0.7 and 0.8—when the topic summaries are based on the FREX score estimated using the proposed model. Furthermore, these results indicate that the interpretability of these topic summaries does not degrade as the size of the number of topic in the model increases.

Figure 10 shows the results for the topic coherence task. Panel (a) shows the average absolute coherence ratings of the topic summaries obtained with each of the three strategies. Similar to the word intrusion results, the summaries produced by the CPM FREX strategy maintain consistently high ratings—around 2.6 out of 3—regardless of the size of the model. Interestingly, topic summaries obtained with both strategies based on LDA, whether ranked according to frequency or to FREX, lead to indistinguishable ratings for most model sizes, with high ratings for smaller models that quickly drop as the size of the models increase. For the model with 100 topics, reranking by LDA FREQ topic summaries by FREX scores display the worst performance, with average ratings below two. Panel (b) of Figure 10 shows the relative preferences of human evaluators for the three strategies to produce topic summaries. Again, the preference for the topic summaries produced by the FREX scores estimated with the proposed method (PCM FREX) increases as the size of the models increase, with over 50% of workers choosing that strategy for the largest model we considered, with 100 topics. Interestingly, preference for the topic summaries obtained with both strategies based on LDA, whether ranked according to frequency or to FREX, is consistently low independently of the model size.

A natural question arising from the results in Figures 9 and 10 is whether the average quality degradation we observe for topic summaries based on LDA as the number of topics in the model grows is due to the declining quality of all topics or to the addition of many low-quality topics. To better understand this observed trend in average quality, Figure 11 shows the distribution of average coherence ratings for individual topic summaries obtained with the three strategies. In the figure, the number of topics in the model varies along the rows, while the strategy to obtain topic summaries varies along the columns. The distributions of topic coherence ratings from human evaluators for LDA FREQ (middle column) and LDA FREX (right column) flatten out for larger models, rather than concentrating around a mediocre score. This suggest that while some high-quality topic summaries remain for the lager 50- and 100-topic models, these high-quality topic are outnumbered by a growing number of middle- and low-quality topic summaries. In contrast, the distributions of the ratings for PCM FREX topic summaries remain relatively unchanged as the models grow in size, with most of the topic summaries retaining average ratings above 2.5.

Overall, the results of the randomized experiment on Amazon Mechanical Turk provide strong evidence in support of the two hypotheses that topic summaries based on the FREX score are more interpretable than currently established frequency-based summaries, and that the proposed model produces estimates of the FREX scores that are superior to those obtained from LDA.

4.6.4 *Stability of Exclusivity Estimates*

The experiments above suggest that, while the exclusivity of a word to a topic can be computed from word rate estimates obtained with an LDA-type parameterization, such estimates lead to less interpretable topics, especially in larger models. One plausible explanation for these results is that estimates of exclusivity based on models that regularize word rates within a topic are less stable, in some sense, than estimates obtained with the proposed approach to modeling and regularization. Here, we explore the stability of the exclusivity estimates obtained with both approaches.

Stability is quantified indirectly, in terms of the maximum exclusivity of a word across topics, and directly, in terms of variance of the estimated word rates.

The working hypothesis is that unregularized, or poorly regularized, estimates of exclusivity may promote rare words, which can produce word counts across topics that depart significantly from the uniform vector in a corpus even if their usage across topics is equal in expectation. As a result, exclusivity-based topic summaries might be dominated by rare words, regardless of their topical content.

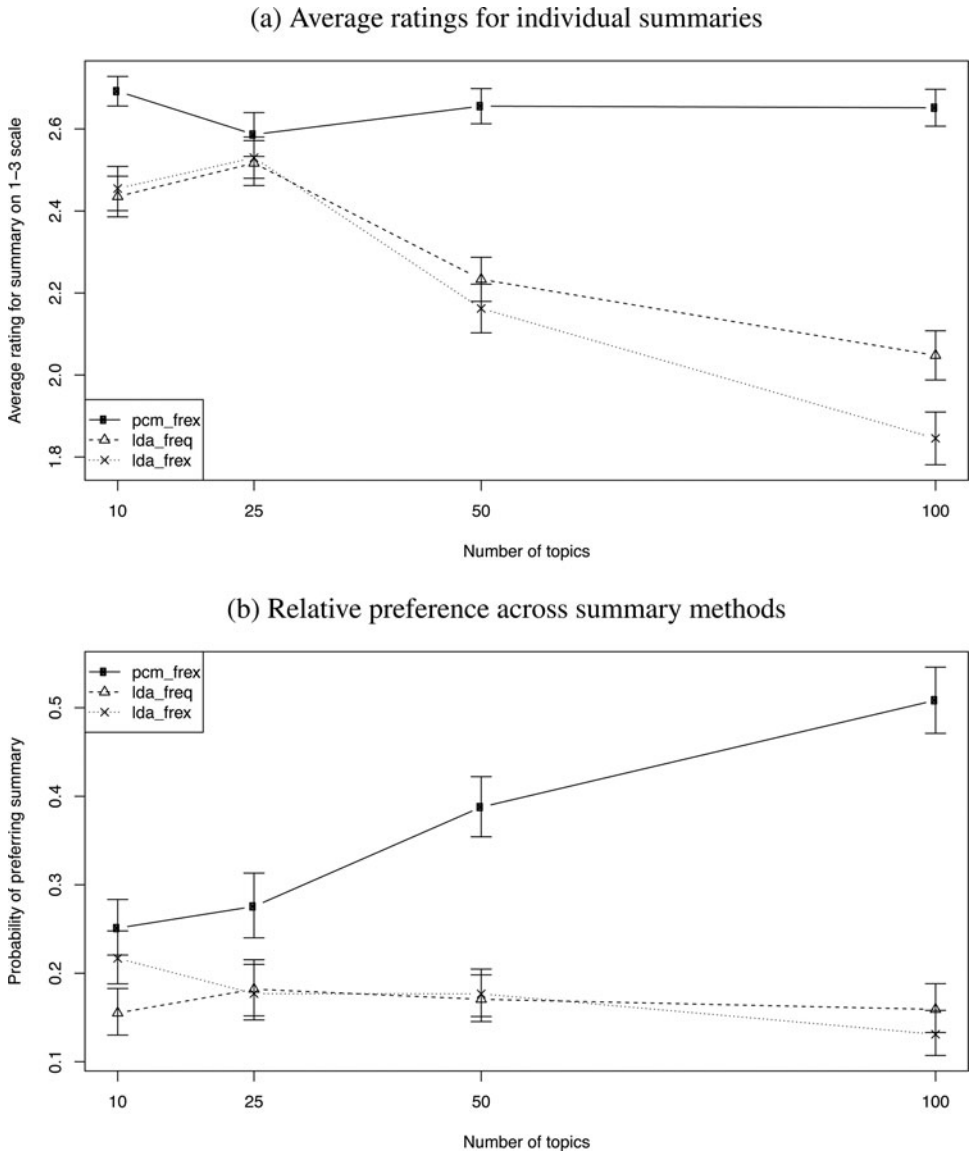

Figure 10. Results from Amazon Turk topic coherence task.

To gauge the severity of this problem, Figures 12 and 13 show scatterplots where the maximum exclusivity of each word across topics (top panels), and the variance of the estimated word rates (bottom panels) are plotted as a function of the marginal word count, for the LDA (left panels) and for proposed model (right panels). Figure 12 refers to model fits with 10 topics. Figure 13 refers to model fits with 100 topics. The top panels show that LDA assigns its highest exclusivity scores to words with less than 100 total occurrences, whose scores dominate those of high-frequency words by several orders of magnitude (on the logit scale). The bottom panels show that LDA assigns the highest variance of word rates across topics to words with less than 100 total occurrences. In contrast, the proposed model reverses these relationships in all these scatterplots, giving the highest maximum exclusivity and variance to the most frequent words. These patterns are consistent across model sizes. The variance results are consistent with previous work showing that LDA leads to highly variable word rates across topics, especially for rare words (Eisenstein, Ahmed, and Xing 2011).

5. Concluding Remarks

The main finding that emerges from our work is the need to quantify how words are used differentially across topics as well as within them to summarize topical content in an interpretable fashion; we refer to these dimensions of content as word exclusivity and frequency. Topical summaries that focus on word frequency alone are often dominated by stop words or other terms used similarly across many topics. Words can be visualized graphically in the exclusivity versus frequency space, or these dimensions can be combined into a scalar quantity, such as the FREX score proposed in Section 2.3, to obtain a univariate measure of the topical content for words in each topic.

Estimates of exclusivity based on rates of word occurrence regularized within a topic, as in LDA, are biased toward rare words due to sensitivity to small differences in estimated use across topics, as shown in Section 4.6.4. Topic models with regularization strategies borrowed from LDA cannot regularize differential use due to topic normalization of usage rates; its symmetric Dirichlet prior on topic distributions regularizes within,

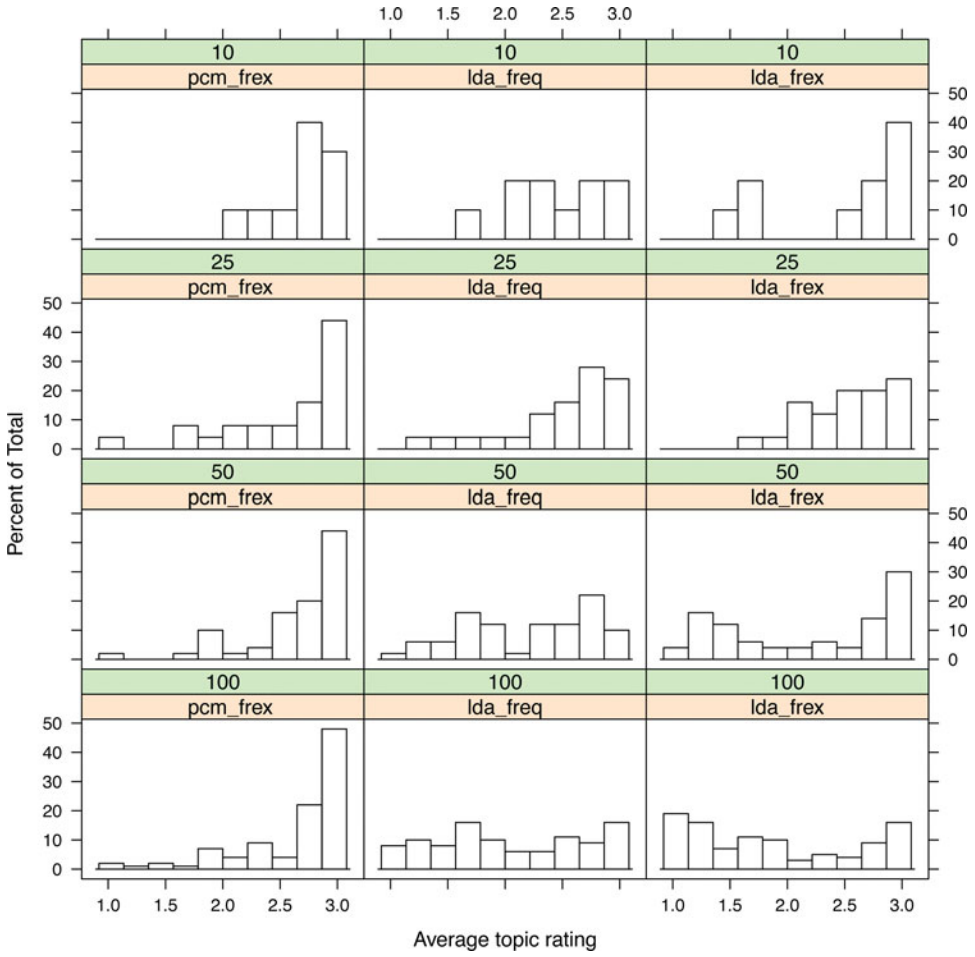

Figure 11. Distribution of topic coherence ratings across number of topics in model (rows) and summary method (columns).

not between, topic usage. While topic-regularized models can capture many important facets of word usage, they are not optimal for the estimands used in our analysis of topical content.

Related issues that affect the interpretability of the output of topic models are the treatment of stop words, and the presence of baseline, or nonsense, topics. In any word summary of a topic, there is an issue of top versus bottom of the list; that is, words down the list are arguably not associated with any one topic. In the proposed approach, stop words get assigned a low FREX score in any one topic; thus, their contribution to the summary becomes negligible. The proposed model reduces the importance of the stop words by design, using regularization induced by sensible priors. For instance, in Figure 4 the stop words lie along a line where exclusivity is constant and frequency varies from high (for noncontextual stop words) to low (for corpus-specific stop words). By contrast, in most models with a within-topic regularization (e.g., Blei, Ng, and Jordan 2003), the emphasis on frequency exacerbates the issue of stop words, artificially promoting them and leading to the appearance of nonsense topics. These issues are well known, and research efforts have proposed ways to mitigate the relevance of stop words and nonsense topics, either at the model level or at pre- or post-processing stages (e.g., see Wallach, Mimno, and McCallum 2009; Mimno et al. 2011). The Amazon Turk experiments in Section 4.6 show that the number of noncoherent topics, which can be taken as a proxy for nonsense topics, is reduced using our model-based estimates of frequency and exclusivity. This evidence supports the argument that our approach leads to a smaller number of nonsense topics. Topic summaries based on the FREX score are more interpretable.

HPC breaks from standard topic models by modeling topic-specific word counts as unnormalized count variates whose rates can be regularized both within and across topics to compute word frequency and exclusivity. It was specifically designed to produce stable exclusivity estimates in human-annotated corpora by smoothing differential word usage according to a semantically intelligent distance metric: proximity on a known hierarchy. This supervised setting is an ideal test case for our framework and will be applicable to many high-value corpora such as the *ACM library*, *IMS* publications, the *New York Times*, and *Reuters*, which all have professional editors and authors and provide multiple annotations to a hierarchy of labels for each document.

HPC offers a complex challenge for full Bayesian inference. To offer a flexible framework for regularization, it breaks from the simple Dirichlet-multinomial conjugacy of traditional models. Specifically, HPC uses Poisson likelihoods whose rates are smoothed across a known topic hierarchy with a Gaussian diffusion and a novel mixed membership model where document label and topic membership parameters share a Gaussian prior. The membership model is the first to create an explicit link between the distribution of topic labels in a document and of the

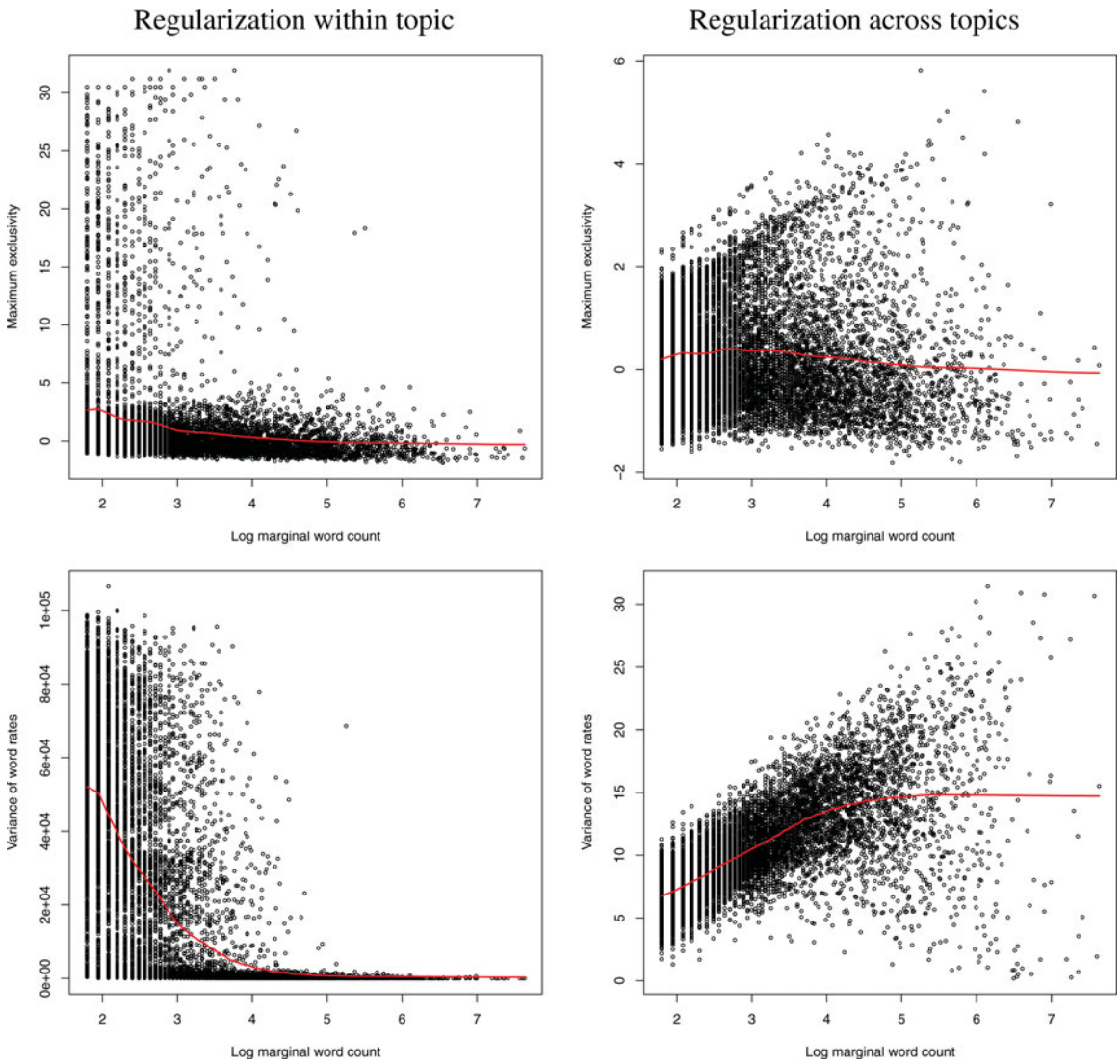

Figure 12. Comparison of word/topic metrics for LDA (left panels) and the proposed model (right panels) fitted with 10 topics. The scatterplots show maximum exclusivity across topics (top panels), and variance of word rates across topics (bottom panels). Constant loess smoother in red.

words that appear in a document and allow for multiple labels. However, the resulting inference is challenging since, conditional on word usage rates, the posterior of the membership parameters involves Poisson and Bernoulli likelihoods of differing dimensions constrained by a Gaussian prior.

We offer two methodological innovations to make inference tractable. First, we design our model with parameters that divide cleanly into two blocks (the tree and document parameters) whose members are conditionally independent given the other block, allowing for parallelized, scalable inference. However, these factorized distributions cannot be normalized analytically and are of the same dimension as the number of topics (102 in the case of *Reuters*). We therefore implement an HMC conditional sampler that mixes efficiently through high-dimensional spaces by leveraging the posterior gradient and Hessian information. This allows HPC to scale to large and complex topic hierarchies that would be intractable for random walk Metropolis samplers.

One unresolved bottleneck in our inference strategy is that the MCMC sampler mixes slowly through the hyperparameter space of the documents—the $\boldsymbol{\eta}$ and λ^2 parameters that control the mean and sparsity of topic memberships and labels. This is due to a large fraction of missing information in our augmentation strategy (Meng and Rubin 1991). Conditional on all the documents' topic affinity parameters $\{\boldsymbol{\xi}_d\}_{d=1}^D$, these hyperparameters index a normal distribution with D observations; marginally, however, we have much less information about the exact loading of each topic onto each document. While we have been exploring more efficient data-augmentation strategies such as parameter expansion (Liu and Wu 1999), we have not found a workable alternative to augmenting the posterior with the entire set of $\{\boldsymbol{\xi}_d\}_{d=1}^D$ parameters.

5.1 Toward Semi-Automated Topic Onthologies

The HPC model can be leveraged to semi-automate the construction of topic ontologies targeted to specific domains, for instance, when fit to comprehensive human-annotated corpora such as *Wikipedia*, *The New York Times*, *Encyclopedia Britannica*, or databases such as *JSTOR* and the *ACM repository*. By learning a probabilistic representation of high-quality topics, HPC output

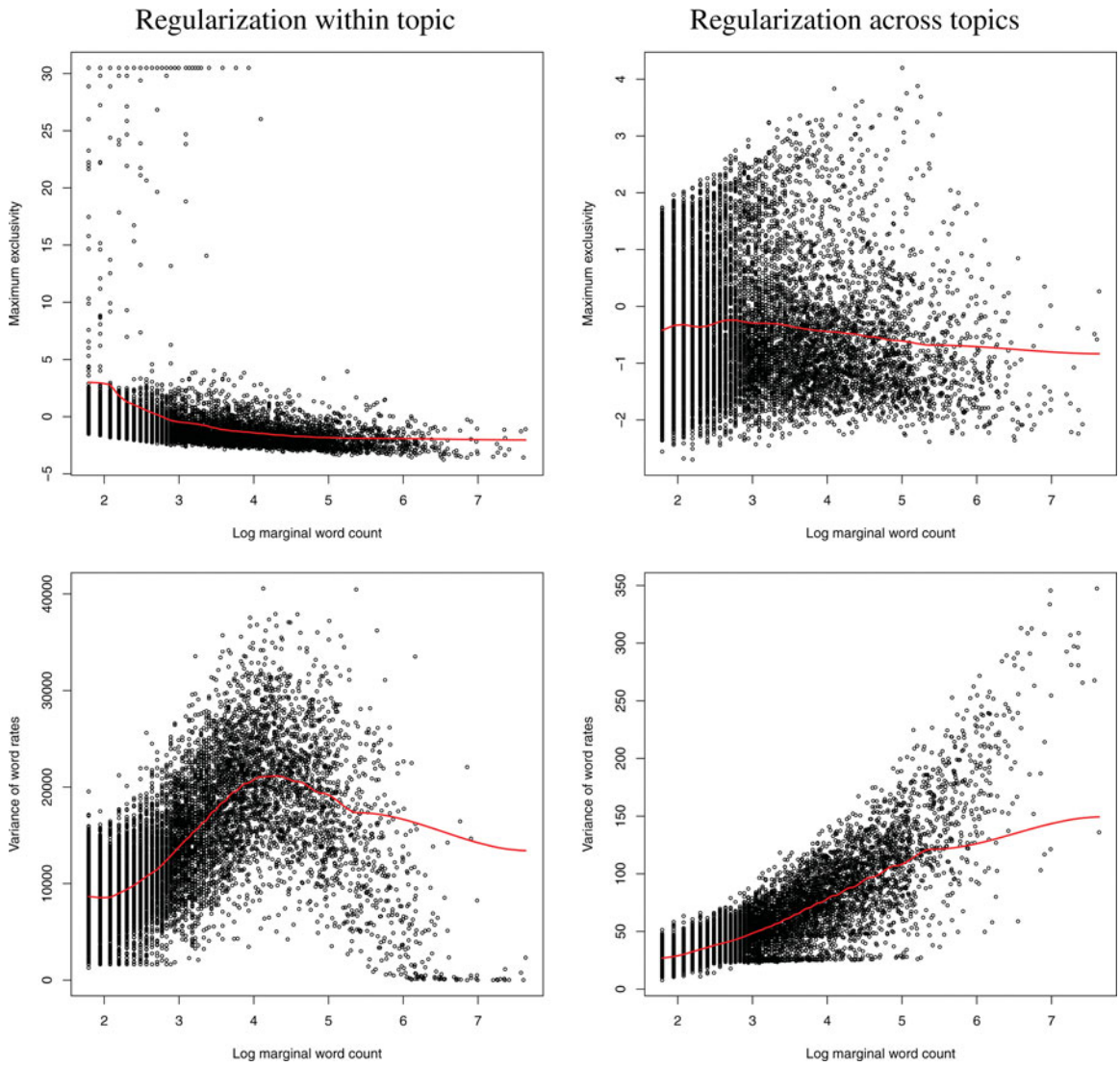

Figure 13. Comparison of word/topic metrics for LDA (left panels) and the proposed model (right panels) fitted with 100 topics. The scatterplots show maximum exclusivity across topics (top panels), and variance of word rates across topics (bottom panels). Constant loess smoother in red.

can be used as a gold standard to aid and evaluate other learning methods (Bakalov et al. 2012).

Targeted ontologies have been a key factor in monitoring scientific progress in biology (Ashburner et al. 2000; Kanehisa and Goto 2000). A hierarchical ontology of topics would lead to new metrics for measuring progress in text analysis. It would enable an evaluation of the semantic content of any collection of inferred topics, thus finally allowing for a *quantitative comparison* among the output of topic models. Current evaluations are qualitative, anecdotal, and unsatisfactory; for instance, authors argue that lists of most frequent words describing an arbitrary selection of topics inferred by a new model make sense intuitively, or that they are better than lists obtained with other models.

In addition to model evaluation, a news-specific ontology could be used as prior to inform the analysis of unstructured text, including Twitter feeds, Facebook wall-posts, and blogs. Unsupervised topic models infer a latent topic space that may be oriented around unhelpful axes, such as authorship or geography. Using a human-created ontology as a prior could ensure that a useful topic space is discovered without being so dogmatic as to assume that unlabeled documents have the same latent structure as labeled examples.

Appendix A. Implementing the Parallelized HMC Sampler

A.1 Hamiltonian Monte Carlo Conditional Updates

HMC is the key tool that makes high-dimensional, nonconjugate updates tractable for our Gibbs sampler. It works well for log densities that are unimodal and have relatively constant curvature. We outline our customized implementation of the algorithm here; a general introduction can be found in Neal (2011).

HMC is a version of the Metropolis-Hastings algorithm that replaces the common Multivariate Normal proposal distribution with a distribution based on Hamiltonian dynamics. It can be used to make joint proposals on the entire parameter space or, as in this article, to make proposals along the conditional posteriors as part of a Gibbs scan. While it requires closed-form calculation of the posterior gradient and curvature to perform well, the algorithm can produce uncorrelated or negatively correlated draws from the target distribution that are almost always accepted.

A consequence of classical mechanics, Hamiltonian's equations can be used to model the movement of a particle along a frictionless surface. The total energy of the particle is the sum of its potential energy (the height of the surface relative to the minimum at the current position) and its kinetic energy (the amount of work needed to accelerate the particle from rest to its current velocity). Since energy is preserved in a closed system, the particle can only convert potential energy to kinetic (or vice versa) as it moves along the surface.

Imagine a ball placed high on the side of the parabola $f(q) = q^2$ at position $q = -2$. Starting out, it will have no kinetic energy but significant potential energy due to its position. As it rolls down the parabola toward zero, it speeds up (gaining kinetic energy), but loses potential energy to compensate as it moves to a lower position. At the bottom of the parabola the ball has only kinetic energy, which it then translates back into potential energy by rolling up the other side until its kinetic energy is exhausted. It will then roll back down the side it just climbed, completely reversing its trajectory until it returns to its original position.

HMC uses Hamiltonian dynamics as a method to find a distant point in the parameter space with high probability of acceptance. Suppose we want to produce samples from $f(\boldsymbol{q})$, a possibly unnormalized density. Since we want high-probability regions to have the least potential energy, we parameterize the surface the particle moves along as $U(\boldsymbol{q}) = -\log f(\boldsymbol{q})$, which is the height of the surface and the potential energy of the particle at any position $\boldsymbol{q}$. The total energy of the particle, $H(\boldsymbol{p}, \boldsymbol{q})$, is the sum of its kinetic energy, $K(\boldsymbol{p})$, and its potential energy, $U(\boldsymbol{q})$, where $\boldsymbol{p}$ is its momentum along each coordinate. After drawing an initial momentum for the particle (typically chosen as $\boldsymbol{p} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{M})$, where $\boldsymbol{M}$ is called the *mass matrix*), we allow the system to evolve for a period of time—not so little that there is negligible absolute movement, but not so much that the particle has time to roll back to where it started.

HMC will not generate good proposals if the particle is not given enough momentum in each direction to efficiently explore the parameter space in a fixed window of time. The higher the curvature of the surface, the more energy the particle needs to move to a distant point. Therefore, the performance of the algorithm depends on having a good estimate of the posterior curvature $\hat{\boldsymbol{H}}(\boldsymbol{q})$ and drawing $\boldsymbol{p} \sim \mathcal{N}(\boldsymbol{0}, -\hat{\boldsymbol{H}}(\boldsymbol{q}))$. If the estimated curvature is accurate and relatively constant across the parameter space, the particle will have high initial momentum along directions where the posterior is concentrated and less along those where the posterior is more diffuse.

Unless the (conditional) posterior is very well behaved, the Hessian should be calculated at the log-posterior mode to ensure positive definiteness. Maximization is generally an expensive operation, however, so it is not feasible to update the Hessian every iteration of the sampler. In contrast, the log-prior curvature is very easy to calculate and well behaved everywhere. This led us to develop the *scheduled conditional HMC sampler* (SCHMC), an algorithm for nonconjugate Gibbs draws that updates the log-prior curvature at every iteration but only updates the log-likelihood curvature in a strategically chosen subset of iterations. We use this algorithm for all nonconjugate conditional draws in our Gibbs sampler.

Specifically, suppose we want to draw from the conditional distribution $p(\boldsymbol{\theta}|\boldsymbol{\psi}_t, \boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\psi}_t)p(\boldsymbol{\theta}|\boldsymbol{\psi}_t)$ in each Gibbs scan, where $\boldsymbol{\psi}$ is a vector of the remaining parameters and $\boldsymbol{y}$ is the observed data. Let $\mathcal{S}$ be the set of full Gibbs scans in which the log-likelihood Hessian information is updated (which always includes the first). For Gibbs scan $i \in \mathcal{S}$, we first calculate the conditional posterior mode and evaluate both the Hessian of the log-likelihood, $\log p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\psi}_t)$, and of the log-prior, $\log p(\boldsymbol{\theta}|\boldsymbol{\psi}_t)$, at that mode, adding them together to get the log-posterior Hessian. We then get a conditional posterior draw with HMC using the negative Hessian as our mass matrix. For Gibbs scan $i \notin \mathcal{S}$, we evaluate the log-prior Hessian at the current location and add it to our last evaluation of the log-likelihood Hessian to get the log-posterior Hessian. We then proceed as before. The SCHMC procedure is described in step-by-step detail in Algorithm 1.

A.2 SCHMC Implementation Details for HPC Model

In the previous section we described our general procedure for obtaining samples from unnormalized conditional posteriors, the SCHMC algorithm. In this section, we provide the gradient and Hessian calculations necessary to implement this procedure for the unnormalized conditional densities in the HPC model, as well as strategies to obtain the maximum of each conditional posterior.

A.1.1 Conditional Posterior of the Rate Parameters

The log-conditional posterior of the rate parameters for one word is

$$\begin{aligned}
&\log p(\boldsymbol{\mu}_f|\boldsymbol{W}, \boldsymbol{I}, \boldsymbol{l}, \{\boldsymbol{\tau}_f^2\}_{f=1}^V, \psi, \gamma^2, \nu, \sigma^2, \{\boldsymbol{\xi}_d\}_{d=1}^D, \mathcal{T}) \\
&= \sum_{d=1}^{D} \log \text{Pois}(w_{fd}|l_d \boldsymbol{\theta}_d^T \boldsymbol{\beta}_f) + \log \mathcal{N}(\boldsymbol{\mu}_f|\psi \boldsymbol{1}, \boldsymbol{\Lambda}(\gamma^2, \boldsymbol{\tau}_f^2, \mathcal{T})) \\
&= -\sum_{d=1}^{D} l_d \boldsymbol{\theta}_d^T \boldsymbol{\beta}_f + \sum_{d=1}^{D} w_{fd} \log(\boldsymbol{\theta}_d^T \boldsymbol{\beta}_f) \\
&\quad -\frac{1}{2}(\boldsymbol{\mu}_f - \psi \boldsymbol{1})^T \boldsymbol{\Lambda}(\boldsymbol{\mu}_f - \psi \boldsymbol{1}).
\end{aligned}$$

Since the likelihood is a function of $\boldsymbol{\beta}_f$, we need to use the chain rule to get the gradient in $\boldsymbol{\mu}_f$ space:

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\mu}_f}\left[\log p(\boldsymbol{\mu}_f|\boldsymbol{W}, \boldsymbol{I}, \boldsymbol{l}, \{\boldsymbol{\tau}_f^2\}_{f=1}^V, \psi, \gamma^2, \{\boldsymbol{\xi}_d\}_{d=1}^D, \mathcal{T})\right] \\
&= \frac{\partial l(\boldsymbol{\beta}_f)}{\partial \boldsymbol{\beta}_f}\frac{\partial \boldsymbol{\beta}_f}{\partial \boldsymbol{\mu}_f} + \frac{\partial}{\partial \boldsymbol{\mu}_f}\left[\log p(\boldsymbol{\mu}_f|\{\boldsymbol{\tau}_f^2\}_{f=1}^V, \psi, \gamma^2, \mathcal{T})\right] \\
&= -\sum_{d=1}^{D} l_d(\boldsymbol{\theta}_d^T \circ \boldsymbol{\beta}_f^T) + \sum_{d=1}^{D}\left(\frac{w_{fd}}{\boldsymbol{\theta}_d^T \boldsymbol{\beta}_f}\right)(\boldsymbol{\theta}_d^T \circ \boldsymbol{\beta}_f^T) - \boldsymbol{\Lambda}(\boldsymbol{\mu}_f - \psi \boldsymbol{1}),
\end{aligned}$$

where $\circ$ is the Hadamard (entrywise) product. The Hessian matrix follows a similar pattern:

$$\begin{aligned}
&\boldsymbol{H}(\log p(\boldsymbol{\mu}_f|\boldsymbol{W}, \boldsymbol{I}, \boldsymbol{l}, \{\boldsymbol{\tau}_f^2\}_{f=1}^V, \psi, \gamma^2, \{\boldsymbol{\xi}_d\}_{d=1}^D, \mathcal{T})) \\
&\quad = -\boldsymbol{\Theta}^T \boldsymbol{W} \boldsymbol{\Theta} \circ \boldsymbol{\beta}_f \boldsymbol{\beta}_f^T + \boldsymbol{G} - \boldsymbol{\Lambda},
\end{aligned}$$

where

$$\boldsymbol{W} = \text{diag}\left(\left\{\frac{w_{fd}}{(\boldsymbol{\theta}_d^T \boldsymbol{\beta}_f)^2}\right\}_{d=1}^D\right)$$

and

$$\boldsymbol{G} = \text{diag}\left(\frac{\partial l(\boldsymbol{\beta}_f)}{\partial \boldsymbol{\beta}_f} \circ \boldsymbol{\beta}_f^T\right) = \text{diag}\left(\frac{\partial l(\boldsymbol{\beta}_f)}{\partial \boldsymbol{\mu}_f}\right).$$

We use the BFGS algorithm with the analytical gradient derived above to maximize this density for iterations where the likelihood Hessian is updated; this quasi-Newton method works well since the conditional posterior is unimodal. The Hessian of the likelihood in $\boldsymbol{\beta}$ space is clearly negative definite everywhere since $\boldsymbol{\Theta}^T\boldsymbol{W}\boldsymbol{\Theta}$ is a positive definite matrix. The prior Hessian $\boldsymbol{\Lambda}$ is also positive definite by definition since it is the precision matrix of a Gaussian variate. However, the contribution of the chain rule term $\boldsymbol{G}$ can cause the Hessian to become indefinite away from the mode in $\boldsymbol{\mu}$ space if any of the gradient entries are sufficiently large and positive. Note, however, that the conditional posterior is still unimodal since the logarithm is a monotone transformation.

A.2.2 Conditional Posterior of the Topic Affinity Parameters

The log-conditional posterior for the topic affinity parameters for one document is

$$\begin{aligned}
&\log p(\boldsymbol{\xi}_d|\boldsymbol{W}, \boldsymbol{I}, \boldsymbol{l}, \{\boldsymbol{\mu}_f, \boldsymbol{\tau}_f^2\}_{f=1}^V, \boldsymbol{\eta}, \boldsymbol{\Sigma}) \\
&\quad = l_d \sum_{f=1}^{V} \log \text{Pois}(w_{fd}|\boldsymbol{\beta}_f^T\boldsymbol{\theta}_d) \\
&\qquad + \log \text{Bernoulli}(\boldsymbol{I}_d|\boldsymbol{\xi}_d) + \log \mathcal{N}(\boldsymbol{\xi}_d|\boldsymbol{\eta}, \boldsymbol{\Sigma}) \\
&= -l_d \sum_{f=1}^{V} \boldsymbol{\beta}_f^T\boldsymbol{\theta}_d + \sum_{f=1}^{V} w_{fd} \log(\boldsymbol{\beta}_f^T\boldsymbol{\theta}_d) - \sum_{k=1}^{K} \log(1+\exp(-\xi_{dk})) \\
&\qquad - \sum_{k=1}^{K} (1-I_{dk})\xi_{dk} - \frac{1}{2}(\boldsymbol{\xi}_d - \boldsymbol{\eta})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\xi}_d - \boldsymbol{\eta}).
\end{aligned}$$

Since the likelihood of the word counts is a function of $\boldsymbol{\theta}_d$, we need to use the chain rule to get the gradient of the likelihood in $\boldsymbol{\xi}_d$ space. This mapping is more complicated than in the case of the $\boldsymbol{\mu}_f$ parameters since each ξ_{dk} is a function of all elements of $\boldsymbol{\theta}_d$:

$$\nabla l_d(\boldsymbol{\xi}_d) = \nabla l_d(\boldsymbol{\theta}_d)^T \boldsymbol{J}(\boldsymbol{\theta}_d \to \boldsymbol{\xi}_d),$$

where $\boldsymbol{J}(\boldsymbol{\theta}_d \to \boldsymbol{\xi}_d)$ is the Jacobian of the transformation from $\boldsymbol{\theta}$ space to $\boldsymbol{\xi}$ space, a $K \times K$ symmetric matrix. Let $S = \sum_{l=1}^{K} \exp \xi_{dl}$. Then

$$\begin{aligned}
&\boldsymbol{J}(\boldsymbol{\theta}_d \to \boldsymbol{\xi}_d) \\
&\quad = S^{-2} \begin{bmatrix} S\exp\xi_{d1} - \exp 2\xi_{d1} & \dots & -\exp(\xi_{dK} + \xi_{d1}) \\ -\exp(\xi_{d1} + \xi_{d2}) & \dots & -\exp(\xi_{dK} + \xi_{d2}) \\ \vdots & \ddots & \vdots \\ -\exp(\xi_{d1} + \xi_{dK}) & \dots & S\exp\xi_{dK} - \exp 2\xi_{dK} \end{bmatrix}.
\end{aligned}$$

The gradient of the likelihood of the word counts in terms of $\boldsymbol{\theta}_d$ is

$$\nabla l_d(\boldsymbol{\theta}_d) = -l_d \sum_{f=1}^{V} \boldsymbol{\beta}_f^T + \sum_{f=1}^{V} \frac{w_{fd}\boldsymbol{\beta}_f^T}{\boldsymbol{\beta}_f^T\boldsymbol{\theta}_d}.$$

Finally, to get the gradient of the full conditional posterior, we add the gradient of the likelihood of the labels and of the normal prior on the $\boldsymbol{\xi}_d$:

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\xi}_d}\left[\log p(\boldsymbol{\xi}_d|\boldsymbol{W}, \boldsymbol{I}, \boldsymbol{l}, \{\boldsymbol{\mu}_f\}_{f=1}^V, \boldsymbol{\eta}, \boldsymbol{\Sigma})\right] \\
&\quad = \nabla l_d(\boldsymbol{\theta}_d)^T \boldsymbol{J}(\boldsymbol{\theta}_d \to \boldsymbol{\xi}_d) + (\mathbf{1} + \exp \boldsymbol{\xi}_d)^{-1} \\
&\qquad -(\mathbf{1} - \boldsymbol{I}_d) - \boldsymbol{\Sigma}^{-1}(\boldsymbol{\xi}_d - \boldsymbol{\eta}).
\end{aligned}$$

The Hessian matrix of the conditional posterior is a complicated tensor product that is not efficient to evaluate analytically. Instead, we compute a numerical Hessian using the analytic gradient presented above at minimal computational cost.

Algorithm 1: Scheduled conditional HMC sampler for iteration i

input : $\tilde{\boldsymbol{\theta}}_{t-1}$, $\boldsymbol{\psi}_t$ (current value of other parameters), $\boldsymbol{y}$ (observed data), L (number of leapfrog steps), ϵ (stepsize), and $\mathcal{S}$ (set of full Gibbs scans in which the likelihood Hessian is updated)
output: $\tilde{\boldsymbol{\theta}}_t$

$\boldsymbol{\theta}_0^* \leftarrow \boldsymbol{\theta}_{t-1}$;

/* Update conditional likelihood Hessian if iteration in schedule */
if $i \in \mathcal{S}$ **then**
  $\hat{\boldsymbol{\theta}} \leftarrow \arg\max_{\boldsymbol{\theta}} \{\log p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\psi}_t) + \log p(\boldsymbol{\theta}|\boldsymbol{\psi}_t)\}$;
  $\hat{\boldsymbol{H}}_l(\boldsymbol{\theta}) \leftarrow \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\left[\log p(\boldsymbol{y}|\hat{\boldsymbol{\theta}}, \boldsymbol{\psi}_t)\right]|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$;
end
/* Calculate prior Hessian and set up mass matrix */
$\hat{\boldsymbol{H}}_p(\boldsymbol{\theta}) \leftarrow \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\left[\log p(\boldsymbol{\theta}|\boldsymbol{\psi}_t)\right]|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0^*}$;
$\hat{\boldsymbol{H}}(\boldsymbol{\theta}) \leftarrow \hat{\boldsymbol{H}}_l(\boldsymbol{\theta}) + \hat{\boldsymbol{H}}_p(\boldsymbol{\theta})$;
$\boldsymbol{M} \leftarrow -\hat{\boldsymbol{H}}(\boldsymbol{\theta})$;

/* Draw initial momentum */
Draw $\boldsymbol{p}_0^* \sim \mathcal{N}(\mathbf{0}, \boldsymbol{M})$;

/* Leapfrog steps to get HMC proposal */
for $l \leftarrow 1$ **to** L **do**
  $\boldsymbol{g}_1 \leftarrow -\frac{\partial}{\partial\boldsymbol{\theta}}\left[\log p(\boldsymbol{\theta}|\boldsymbol{\psi}_t, \boldsymbol{y})\right]|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{l-1}^*}$;
  $\boldsymbol{p}_{l,1}^* \leftarrow \boldsymbol{p}_{l-1}^* - \frac{\epsilon}{2}\boldsymbol{g}_1$;
  $\boldsymbol{\theta}_l^* \leftarrow \boldsymbol{\theta}_{l-1}^* + \epsilon(\boldsymbol{M}^{-1})^T\boldsymbol{p}_{l,1}^*$;
  $\boldsymbol{g}_2 \leftarrow -\frac{\partial}{\partial\boldsymbol{\theta}}\left[\log p(\boldsymbol{\theta}|\boldsymbol{\psi}_t, \boldsymbol{y})\right]|_{\boldsymbol{\theta}=\boldsymbol{\theta}_l^*}$;
  $\boldsymbol{p}_l^* \leftarrow \boldsymbol{p}_{l,1}^* - \frac{\epsilon}{2}\boldsymbol{g}_2$;
end

/* Calculate Hamiltonian (total energy) of initial position */
$K_{t-1} \leftarrow \frac{1}{2}(\boldsymbol{p}_0^*)^T\boldsymbol{M}^{-1}\boldsymbol{p}_0^*$;
$U_{t-1} \leftarrow -\log p(\boldsymbol{\theta}_0^*|\boldsymbol{\psi}_t, \boldsymbol{y})$;
$H_{t-1} \leftarrow K_{t-1} + U_{t-1}$;

/* Calculate Hamiltonian (total energy) of candidate position */
$K^* \leftarrow \frac{1}{2}(\boldsymbol{p}_L^*)^T\boldsymbol{M}^{-1}\boldsymbol{p}_L^*$;
$U^* \leftarrow -\log p(\boldsymbol{\theta}_L^*|\boldsymbol{\psi}_t, \boldsymbol{y})$;
$H^* \leftarrow K^* + U^*$;

/* Metropolis correction to determine if proposal accepted */
Draw $u \sim \text{Unif}[0, 1]$;
$\log r \leftarrow H_{t-1} - H^*$;
if $\log u < \log r$ **then**
  $\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_L^*$
else
  $\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1}$
end

We use the BFGS algorithm with the analytical gradient derived above to maximize this density for iterations where the likelihood Hessian is updated. We have not been able to show analytically that this conditional posterior is unimodal, but we have verified this graphically for several documents and have achieved very high acceptance rates for our HMC proposals based on this Hessian calculation.

A.3.3 Conditional Posterior of the τ^2_{fk} Hyperparameters

The variance parameters τ^2_{fk} independently follow an identical Scaled Inverse-χ^2 with convolution parameter ν and scale parameter σ^2, while their inverse follows a Gamma$(\kappa_\tau = \frac{\nu}{2}, \lambda_\tau = \frac{2}{\nu\sigma^2})$ distribution. The log-conditional posterior of these parameters is

$$\begin{aligned}\log p(\kappa_\tau, \lambda_\tau | \{\boldsymbol{\tau}^2_f\}_{f=1}^V, \mathcal{T}) = (\kappa_\tau - 1) \sum_{f=1}^{V} \sum_{k \in \mathcal{P}} \log (\tau^2_{fk})^{-1} \\ - |\mathcal{P}| V \kappa_\tau \log \lambda_\tau - |\mathcal{P}| V \log \Gamma(\kappa_\tau) \\ - \frac{1}{\lambda_\tau} \sum_{f=1}^{V} \sum_{k \in \mathcal{P}} (\tau^2_{fk})^{-1},\end{aligned}$$

where $\mathcal{P}(\mathcal{T})$ is the set of parent topics on the tree. If we allow $i \in \{1, \ldots, N = |\mathcal{P}|V\}$ to index all the f, k pairs and $l(\kappa_\tau, \lambda_\tau) = p(\{\boldsymbol{\tau}^2_f\}_{f=1}^V | \kappa_\tau, \lambda_\tau, \mathcal{T})$, we can simplify this to

$$\begin{aligned}l(\kappa_\tau, \lambda_\tau) = (\kappa_\tau - 1) \sum_{i=1}^{N} \log \tau_i^{-2} - N \kappa_\tau \log \lambda_\tau - N \log \Gamma(\kappa_\tau) \\ - \frac{1}{\lambda_\tau} \sum_{i=1}^{N} \tau_i^{-2}.\end{aligned}$$

We then transform this density onto the $(\log \kappa_\tau, \log \lambda_\tau)$ scale so that the parameters are unconstrained, a requirement for standard HMC implementation. Each draw of $(\log \kappa_\tau, \log \lambda_\tau)$ is then transformed back to the (ν, σ^2) scale. To get the Hessian of the likelihood in log space, we calculate the derivatives of the likelihood in the original space and apply the chain rule:

$$\begin{aligned}&\boldsymbol{H}\Big(l(\log \kappa_\tau, \log \lambda_\tau)\Big) \\ &= \begin{bmatrix} \kappa_\tau \frac{\partial l(\kappa_\tau, \lambda_\tau)}{\partial \kappa_\tau} + (\kappa_\tau)^2 \frac{\partial^2 l(\kappa_\tau, \lambda_\tau)}{\partial (\kappa_\tau)^2} & \kappa_\tau \lambda_\tau \frac{\partial^2 l(\kappa_\tau, \lambda_\tau)}{\partial \kappa_\tau \partial \lambda_\tau} \\ \kappa_\tau \lambda_\tau \frac{\partial^2 l(\kappa_\tau, \lambda_\tau)}{\partial \kappa_\tau \partial \lambda_\tau} & \lambda_\tau \frac{\partial l(\kappa_\tau, \lambda_\tau)}{\partial \lambda_\tau} + (\lambda_\tau)^2 \frac{\partial^2 l(\kappa_\tau, \lambda_\tau)}{\partial (\lambda_\tau)^2} \end{bmatrix},\end{aligned}$$

where

$$\nabla l(\kappa_\tau, \lambda_\tau) = \begin{bmatrix} \sum_{i=1}^{N} \log \tau_i^{-2} - N \log \lambda_\tau - N \psi(\kappa_\tau) \\ -\frac{N \kappa_\tau}{\lambda_\tau} + \frac{1}{(\lambda_\tau)^2} \sum_{i=1}^{N} \tau_i^{-2} \end{bmatrix}$$

and

$$\boldsymbol{H}\Big(l(\kappa_\tau, \lambda_\tau)\Big) = \begin{bmatrix} -N \psi'(\kappa_\tau) & -\frac{N}{\lambda_\tau} \\ -\frac{N}{\lambda_\tau} & \frac{N \kappa_\tau}{(\lambda_\tau)^2} - \frac{2}{(\lambda_\tau)^3} \sum_{i=1}^{N} \tau_i^{-2} \end{bmatrix}.$$

Following Algorithm 1, we evaluate the Hessian at the mode of this joint posterior. This is easiest to find on original scale following the properties of the Gamma distribution. The first-order condition for λ_τ can be solved analytically:

$$\lambda_{\tau,\mathrm{MLE}}(\kappa_\tau) = \arg\max_{\lambda_\tau} \Big\{ l(\kappa_\tau, \lambda_\tau) \Big\} = \frac{1}{\kappa_\tau N} \sum_{i=1}^{N} \tau_i^{-2}.$$

We can then numerically maximize the profile likelihood of κ_τ:

$$\kappa_{\tau,\mathrm{MLE}} = \arg\max_{\kappa_\tau} \Big\{ l(\kappa_\tau, \lambda_{\tau,\mathrm{MLE}}(\kappa_\tau)) \Big\}.$$

The joint mode in the original space is then $(\kappa_{\tau,\mathrm{MLE}}, \lambda_{\tau,\mathrm{MLE}}(\kappa_{\tau,\mathrm{MLE}}))$. Due to the monotonicity of the logarithm function, the mode in the transformed space is simply $(\log \kappa_{\tau,\mathrm{MLE}}, \log \lambda_{\tau,\mathrm{MLE}})$. We can be confident that the conditional posterior is unimodal: the Fisher information for a Gamma distribution is negative definite, and the log transformation to the unconstrained space is monotonic.

Funding

This work was partially supported by the National Science Foundation under grants CAREER IIS-1149662 and IIS-1409177, by the Army Research Office grant MURI W911NF-11-1-0036, and by the Office of Naval Research under grant YIP N00014-14-1-0485. Edoardo M. Airoldi is an Alfred P. Sloan Research Fellow, and a Shutzer Fellow at the Radcliffe Institute for Advanced Studies.

References

Adams, R. P., Ghahramani, Z., and Jordan, M. I. (2010), "Tree-Structured Stick Breaking for Hierarchical Data," in *Advances in Neural Information Processing Systems (NIPS) 23*, pp. 19–27. [1382]

Airoldi, E. M., Anderson, A. G., Fienberg, S. E., and Skinner, K. K. (2006), "Who Wrote Ronald Reagan's Radio Addresses?" *Bayesian Analysis*, 1, 289–320. [1382]

Airoldi, E. M., Blei, D. M., Erosheva, E. A., and Fienberg, S. E. (eds.) (2014), *Handbook of Mixed Membership Models and Their Applications*, Boca Raton, FL: Chapman & Hall/CRC Press. [1381]

Airoldi, E. M., Blei, D. M., Fienberg, S., and Xing, E. (2008), "Mixed-Membership Stochastic Blockmodels," *Journal of Machine Learning Research*, 9, 1981–2014. [1381]

Airoldi, E. M., Erosheva, E. A., Fienberg, S. E., Joutard, C. J., Love, T. M., and Shringarpure, S. (2010), "Reconceptualizing the Classification of PNAS Articles," *Proceedings of the National Academy of Sciences*, 107, 20899–20904. [1382]

Airoldi, E. M., Fienberg, S. E., and Skinner, K. K. (2007a), "Whose Ideas? Whose Words? Authorship of the Ronald Reagan Radio Addresses," *Political Science & Politics*, 40, 501–506. [1382]

Airoldi, E. M., Fienberg, S. E., and Xing, E. P. (2007b), "Mixed Membership Analysis of Genome-Wide Expression Studies—Attribute Data," arXiv no. 0711.2520. [1382,1383]

Aletras, N., and Stevenson, M. (2013), *Evaluating Topic Coherence Using Distributional Semantics, in IWCS, number 2009*, Shrewsbury, PA: ICWS. [1393]

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubinand, G. M., and Sherlock, G. (2000), "Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium," *Nature Genetics*, 25, 25–29. [1399]

Bakalov, A., McCallum, A., Wallach, H., and Mimno, D. (2012), "Topic Models for Taxonomies," in *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 237–240. [1399]

Blei, D. (2012), "Introduction to Probabilistic Topic Models," *Communications of the ACM*, 55, 77–84. [1382,1392]

Blei, D., Griffiths, T., Jordan, M., and Tenenbaum, J. (2003), "Hierarchical Topic Models and the Nested Chinese Restaurant Process," in *NIPS 16*, Cambridge, MA: MIT Press, pp. 17–24. [1382]
Blei, D., and McAuliffe, J. (2010), "Supervised Topic Models," arXiv:1003.0783. [1382]
Blei, D., Ng, A., and Jordan, M. (2003), "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, 3, 993–1022. [1381,1382,1391,1397]
Breiman, L. (2001), "Statistical Modeling: The Two Cultures," *Statistical Science*, 16, 199–231. [1390]
Buntine, W., and Jakulin, A. (2006), "Discrete Components Analysis," in *Subspace, Latent Structure and Feature Selection, volume 3940 of Lecture Notes in Computer Science*, Berlin: Springer, pp. 1–33. [1383]
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. [1382,1383]
Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., and Blei, D. (2009), "Reading Tea Leaves: How Humans Interpret Topic Models," in *Advances in Neural Information Processing Systems 22*, pp. 288–296. [1382,1393]
Eisenstein, J., Ahmed, A., and Xing, E. P. (2011), "Sparse Additive Generative Models of Text," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 1041–1048. [1382,1383,1396]
Harman, D. (1992), "Overview of the First Text Retrieval Conference (TREC-1)," in *Proceedings of the First Text Retrieval Conference (TREC-1)*, pp. 1–20. [1391]
Hotelling, H. (1936), "Relations Between Two Sets of Variants," *Biometrika*, 28, 321–377. [1381]
Hu, Y., Boyd-Graber, J., and Satinoff, B. (2011), "Interactive Topic Modeling," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pp. 248–257. [1382]
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," *Annals of Applied Statistics*, 8, 499–529. [1393]
Jolliffe, I. T. (1986), *Principal Component Analysis*, New York: Springer-Verlag. [1381]
Kanehisa, M., and Goto, S. (2000), "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research*, 28, 27–30. [1399]
Lewis, D. D., Yang, Y., Rose, T. G., and Li, F. (2004), "RCV1: A New Benchmark Collection for Text Categorization Research," *Journal of Machine Learning Research*, 5, 361–397. [1386]
Liu, J. S., and Wu, Y. N. (1999), "Parameter Expansion for Data Augmentation," *Journal of the American Statistical Association*, 94, 1264–1274. [1398]
McCallum, A., Rosenfeld, R., Mitchell, T., and Ng, A. (1998), "Improving Text Classification by Shrinkage in a Hierarchy of Classes," in *Proceedings of the 15th International Conference on Machine Learning*, pp. 359–367. [1382]
McLachlan, G., and Peel, D. (2000), *Finite Mixture Models*, New York: Wiley. [1381]
Meng, X., and Rubin, D. B. (1991), "Using EM to Obtain Asymptotic Variance-Covariance Matrices: The SEM Algorithm," *Journal of the American Statistical Association*, 86, 899–909. [1398]
Mimno, D., Li, W., and McCallum, A. (2007), "Mixtures of Hierarchical Topics With Pachinko Allocation," in *Proceedings of the 24th International Conference on Machine Learning*, pp. 633–640. [1382]
Mimno, D., Wallach, H., Talley, E., Leenders, M., and McCallum, A. (2011), "Optimizing Semantic Coherence in Topic Models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 262–272. [1382,1397]
Mosteller, F., and Wallace, D. (1964), *Inference and Disputed Authorship: The Federalist*, Reading, MA: Addison-Wesley. [1382]
Mosteller, F., and Wallace, D. (1984), *Applied Bayesian and Classical Inference: The Case of "The Federalist" Papers*, New York: Springer-Verlag. [1382]
Neal, R. (2011), "MCMC using Hamiltonian Dynamics," in *Handbook of Markov Chain Monte Carlo*, eds. S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng, Boca Raton, FL: Chapman & Hall/CRC Press, pp. 113–162. [1385,1399]
Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. (2010), "Automatic Evaluation of Topic Coherence," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 100–108. [1393]
Nigam, K., McCallum, A., Thrun, S., and Mitchell, T. (2000), "Text Classification From Labeled and Unlabeled Documents Using EM," *Machine Learning*, 39, 103–134. [1382]
Perotte, A., Bartlett, N., Elhadad, N., and Wood, F. (2012), "Hierarchically Supervised Latent Dirichlet Allocation," in *Advances in Neural Information Processing Systems 24*, pp. 2609–2617. [1382]
Ramage, D., Hall, D., Nallapati, R., and Manning, C. D. (2009), "Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-Labeled Corpora," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp. 248–256. [1382,1389]
Rubin, T., Chambers, A., Smyth, P., and Steyvers, M. (2012), "Statistical Topic Models for Multi-Label Document Classification," *Machine Learning*, 88, 157–208. [1390]
Sandhaus, E. (2008), *The New York Times Annotated Corpus*, Philadelphia, PA: Linguistic Data Consortium. [1389]
Sohn, K., and Xing, E. P. (2009), "A Hierarchical Dirichlet Process Mixture Model for Haplotype Reconstruction From Multi-Population Data," *Annals of Applied Statistics*, 3, 791–821. [1381]
Wallach, H., Mimno, D., and McCallum, A. (2009), "Rethinking LDA: Why Priors Matter," in *Advances in Neural Information Processing Systems 22*, pp. 1973–1981. [1382,1397]
Zhu, J., Ahmed, A., and Xing, E. P. (2012), "Medlda: Maximum Margin Supervised Topic Models," *Journal of Machine Learning Research*, 13, 2237–2278. [1390]
Zhu, J., and Xing, E. P. (2012), "Sparse Topical Coding," arXiv:1202.3778. [1382]

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2016, VOL. 111, NO. 516, Applications and Case Studies
http://dx.doi.org/10.1080/01621459.2016.1245073

Comment

Matt Taddy

Microsoft Research New England, Cambridge, MA, and The University of Chicago Booth School of Business, Chicago, IL, USA

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 adoption in the social sciences.

CONTACT Matt Taddy taddy@microsoft.com Microsoft Research New England, 1 Memorial Drive, Cambridge MA 02142

In my collaborations with economists and others, much effort has been spent pouring through the lists of words that are "top" or representative within, say, latent topics from LDA (Blei, Ng, and Jordan 2003) 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 stopword 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., those 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 nontext 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 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 (2015) described a scalable distributed version of the MNIR algorithm and illustrated 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 (2015), the word-f count for each document-d is treated as Poisson random variable, using AB's notation,

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

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 [in press]) to apply simple ℓ_1 regularization with Bayesian information criterion (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*.

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).

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).

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

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, for example, 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, 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.

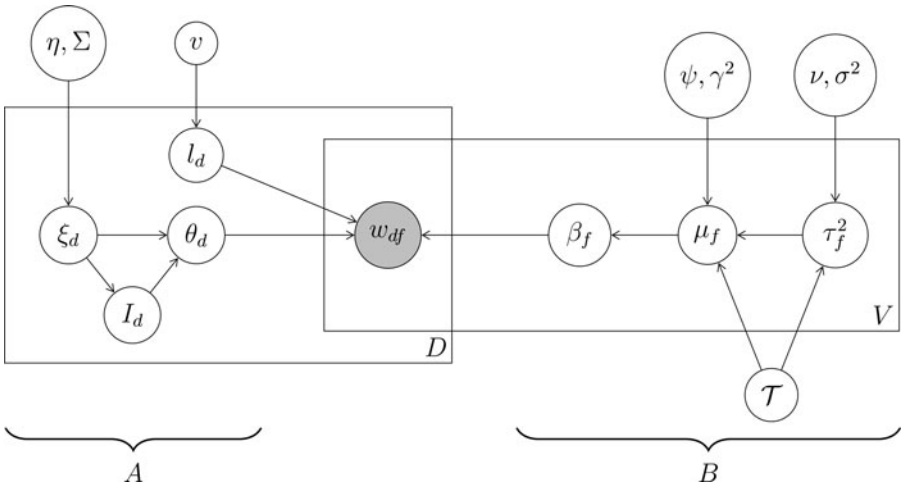

Figure 1. Graphical model diagram of the hierarchical Poisson convolution (HPC) model. Plates indicate replication, outside circles are hyper-parameters for priors, and shading means a quantity is observed. (Note: l_d is not necessarily assumed observed here.)

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 $\boldsymbol{\theta}$.

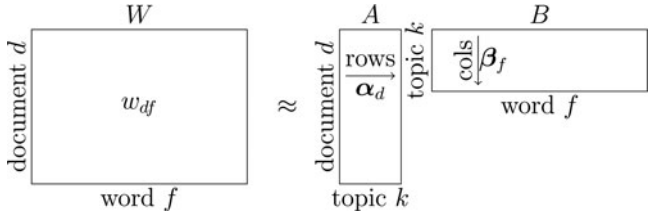

Figure 2. Representation of HPC model likelihood parameterization as a nonnegative matrix factorization (NMF).

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 $\tau_{f,k}^2$ 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

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

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:

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

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 $||B||_{\mathcal{F}}^2$, while Pauca, Piper, and Plemmons (2006) had incorporated both $||A||_{\mathcal{F}}^2$ and $||B||_{\mathcal{F}}^2$. 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 $||B||_{\mathcal{F}}^2$. 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 $\tau_{f,j}^2$ 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.

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.

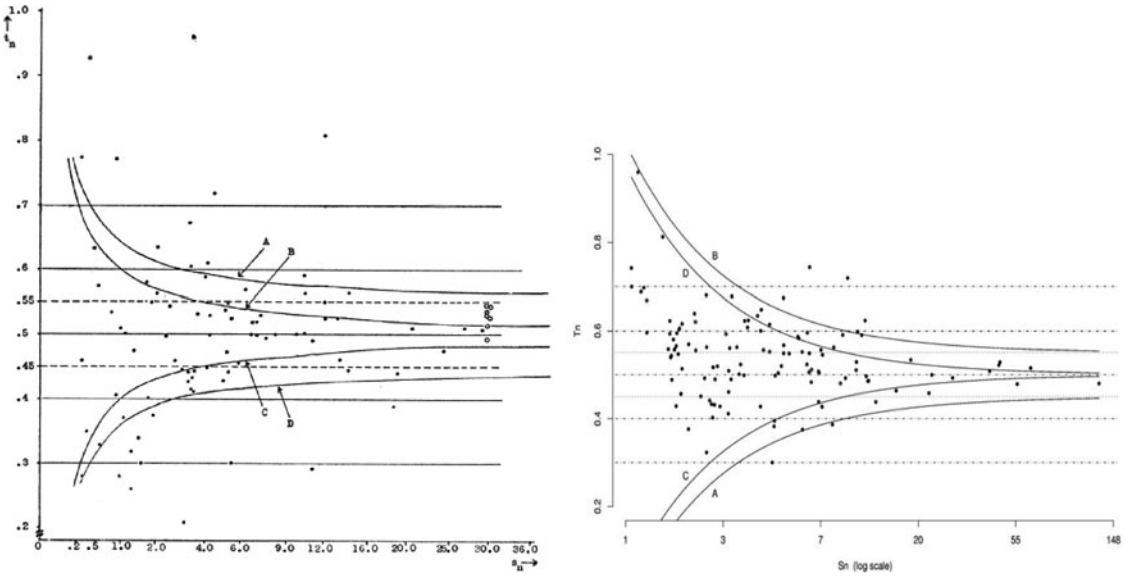

Figure 1. Total rate of word occurrence σ_v, on the *X* axis, versus differential rate tau_v, on the *Y* axis, for two sets words in Mosteller and Wallace (1984) and Airoldi et al. (2006).

reasonable to assume a priori that τ_M is centered around 1/2 (i.e., both authors use words similarly) and that the deviations from 1/2 are less pronounced for words that are used frequently (i.e., with higher total rate). Figure 1 illustrates the prior elicitation carried out by Mosteller and Wallace (left panel) and by Airoldi and Fienberg (right panel) on two different sets of words. For details we refer to the original publications (Mosteller and Wallace 1984; Airoldi et al. 2006). In the case of k subpopulations, this type of prior specification translates into $\boldsymbol{\tau}_v \sim \text{Dirichlet}\,(\alpha_0 + \alpha_1 \cdot \sigma_v)\ \forall v$.

In the computer and information sciences, following Blei, Ng, and Jordan (2003), the rates are regularized *per subpopulation*—or we could also say the rates are regularized *per topic*, because of the desired interpretation of the topics. That is, $\boldsymbol{\beta}_k \sim \text{Dirichlet}\,(\alpha_0)\ \forall k$.

2. Lesson Learned and Further Discussion

While the classic articles above and several other articles have explored modeling and inferential issues, and considered multiple strategies for evaluating the complex output these mixture models of text produce, in-depth, the follow-up literature by and large has not. An almost exclusive focus on word frequency and anecdotal evaluations are the norm. We took issue with the lack of standards for the evaluation of topic models and other models of text, in the literature, and begun to pursue the research presented in the main article as a reaction.

One of the most consequential findings of our research is encapsulated in the claim that subpopulation inference-based regularization *per word* is demonstrably superior to subpopulation inference-based regularization *per topic*, especially in unsupervised analyses of word counts. The randomized experiments we carry out on Amazon Turk, described in Section 4.6 of the main article, directly contrast the parameterization by Mosteller and Wallace and the corresponding regularization *per word*, to the regularization *per subpopulation/topic*, in an unsupervised setting, in which the subpopulation indicators are not observed. These experiments substantiate our claim unambiguously. Regularization *per word* is thus strongly recommended.

The proposed inferential target—the frequency-exclusivity, or FREX, score—is the second aspect of our work that leads to nonnegligible improvements, in terms of stability of the inferences, as reported in the results section of the main article. The FREX score is the harmonic mean of the two ranks of a word induced by frequency and by exclusivity, that is, differential usage. It is meant to combine in a simple way two quantitative aspects of a word: one that is popular and widely used, frequency; and one that was introduced by Mosteller and Wallace, and which we further explore in our work, differential usage. More generally, we believe that a sensible set of new inferential targets may be a simple way to resolve many current issues that plague supervised and unsupervised analyses of word counts. The nearly exclusive focus on frequency is an impediment to methodological progress, and hiders substantive analyses. For example, most analyses begin with a series of preprocessing steps, including the choice of the size of the vocabulary, and of which terms to keep and which to exclude. These choices are largely based on frequency considerations, ad hoc, seldom documented, and yet often have great influence on the results of the analysis. The instability of the inference due to frequent words, the consequential lack of diversity in subpopulations/topics, and the incoherence of a large fraction of the inferred topics are well-known issue in this literature (Wallach, Mimno, and McCallum 2009a; Wallach et al. 2009b; Mimno et al. 2011; Zou and Adams 2012). The FREX score is meant as a possible solution to these issues by bringing key choices back into the inferential framework.

Another important point is about model elicitation. The vast literature on models for unsupervised analyses of word counts often proceeds by making distributional assumptions that are only partly motivated by a real application. We encourage researchers interested in unsupervised analyses to take advantage of the large collections of documents for which subpopulation indicators, be they topics or authors, are observed

(Lewis et al. 2004; Sandhaus 2008). These datasets provide the opportunity to study variation and covariation among variables that are otherwise unobservable in many applications of interest, and to check or validate assumptions that would be impossible to check in an unsupervised analysis.

References

Airoldi, E. M., Anderson, A. G., Fienberg, S. E., and Skinner, K. K. (2006), "Who Wrote Ronald Reagan's Radio Addresses?" *Bayesian Analysis*, 1, 289–320. [1410,1411]

Airoldi, E. M., Blei, D. M., Erosheva, E. A., and Fienberg, S. E. (eds.) (2014), *Handbook of Mixed Membership Models and Their Applications*, Boca Raton, FL: Chapman & Hall / CRC Press. [1410]

Airoldi, E. M., Erosheva, E. A., Fienberg, S. E., Joutard, C. J., Love, T. M., and Shringarpure, S. (2010), "Reconceptualizing the Classification of PNAS Articles," *Proceedings of the National Academy of Sciences*, 107, 20899–20904. [1410]

Blei, D. (2012), "Probabilistic Topic Models," *Communications of the ACM*, 55, 77–84. [1410]

Blei, D., and Lafferty, J. (2007), "A Correlated Topic Model of Science," *Annals of Applied Statistics*, 1, 17–35. [1410]

Blei, D., Ng, A., and Jordan, M. (2003), "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, 3, 993–1022. [1410,1411]

De Morgan, A. (1872), *Budget of Paradoxes*, Green: London Longmans. [1410]

Efron, B., and Thisted, R. (1976), "Estimating the Number of Unseen Species: How Many Words Did Shakespeare Know?" *Biometrika*, 63, 435–447. [1410]

Erosheva, E. A., Fienberg, S. E., and Lafferty, J. (2004), "Mixed-Membership Models of Scientific Publications," *Proceedings of the National Academy of Sciences*, 101, 5220–5227. [1410]

Lewis, D. D., Yang, Y., Rose, T. G., and Li, F. (2004), "RCV1: A New Benchmark Collection for Text Categorization Research," *Journal of Machine Learning Research*, 5, 361–397. [1412]

Mendenhall, T. C. (1887), "The Characteristic Curves of Composition," *Science*, 11, 237–249. [1410]

Miller, G. A., Newman, E. B., and Friedman, E. A. (1958), "Length-Frequency Statistics for Written English," *Information and Control*, 1, 370–389. [1410]

Mimno, D., Wallach, H., Talley, E., Leenders, M., and McCallum, A. (2011), "Optimizing Semantic Coherence in Topic Models," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP),* Edinburgh, UK: Association for Computational Linguistics, pp. 262–272. [1411]

Mosteller, F., and Wallace, D. L. (1963), "Inference in an Authorship Problem," *Journal of the American Statistical Association*, 58, 275–309. [1410]

——— (1964), *Inference and Disputed Authorship: The Federalist*, Reading, MA: Addison-Wesley. [1410]

——— (1984), *Applied Bayesian and Classical Inference: The Case of "The Federalist" Papers*, New York: Springer-Verlag. [1410,1411]

Roberts, M., Stewart, B., and Airoldi, E. M. (2016), "A Model of Text for Experimentation in the Social Sciences," *Journal of the American Statistical Association*, 111, 988–1003. [1410]

Sandhaus, E. (2008), "The New York Times Annotated Corpus," *Linguistic Data Consortium*. Available at *http://catalog.ldc.upenn.edu/LDC2008T19*. [1412]

Wallach, H., Mimno, D., and McCallum, A. (2009a), "Rethinking LDA: Why Priors Matter," in *Proceedings of Advances in Neural Information Processing Systems (NIPS),* eds. Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, Red Hook, NY: Curran Associates, Inc., pp. 1973–1981. [1411]

Wallach, H., Murray, I., Salakhutdinov, R., and Mimno, D. (2009b), "Evaluation Methods for Topic Models," in *Proceedings of the 26th International Conference on Machine Learning* (ICML), New York: Association of Computer Machinery, pp. 1105–1112. [1411]

Yule, U. (1944), *The Statistical Study of Literary Vocabulary*, Cambridge, UK: Cambridge University Press. [1410]

Zipf, G. K. (1932), *Selected Studies of the Principle of Relative Frequency in Language*, Cambridge, MA: Harvard University Press. [1410]

Zou, J. Y., and Adams, R. P. (2012), "Priors for Diversity in Generative Latent Variable Models," *Neural Information Processing Systems*, 25, 2996–3004. [1411]