Summarizing topical content with word frequency and exclusivity

Jonathan M. Bischof JBISCHOF@FAS.HARVARD.EDU
Edoardo M. Airoldi AIROLDI@FAS.HARVARD.EDU
Department of Statistics; Harvard University; 1 Oxford Street; Cambridge, MA 02138 USA

Abstract

Recent work in text analysis commonly describes topics in terms of their most frequent words, but the exclusivity of words to topics is equally important for communicating content. We introduce Hierarchical Poisson Convolution (HPC), a model which infers regularized estimates of the differential use of words across topics as well as their frequency within topics. HPC uses known hierarchical structure on human-labeled topics to make focused comparisons of differential usage within each branch of the hierarchy of labels. We then infer a summary for each topic in terms of words that are both frequent and exclusive. We develop a parallelized Hamiltonian Monte Carlo sampler that allows for fast and scalable computation.

1. Introduction

Modern text analysis research has focused on discovering latent structure in the content of document collections to assist in critical tasks such as topical content exploration, dimensionality reduction, and classification. Most recently, topic models such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) have taken a probabilistic approach to this task by viewing a document's content as arising from a mixture of component distributions. Inferred components, referred to as "topics", as they often capture thematic structure, characterize content in terms of the relative frequency of within-component word usage (Blei., 2012). While inferred topics have proven to be a useful low-dimensional summary of a corpus' content, recent work has documented a growing list of interpretability issues: they are often dominated by contentless "stop" words (Wallach et al., 2009), are sometimes incoherent or redundant (Mimno et al., 2011; Chang et al., 2009), and typically require post hoc modification to meet human expectations (Hu et al., 2011).

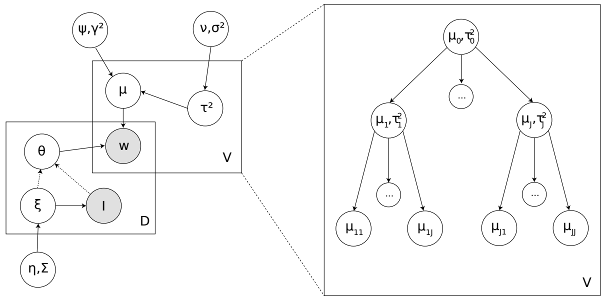
While most attempts to improve topical summaries to date involve changes to the models used to estimate relative frequency, we propose instead a new definition of topical content that incorporates how words are used differentially across topics. If a word is common in a topic, it is also important to know whether it is common in many topics or relatively exclusive to the topic in question. Both measurements 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. We therefore look for the most frequent words in the corpus that are also likely to have been generated from the topic of interest to summarize its content. In this approach we borrow ideas from the statistical literature, in which models of differential word usage have been leveraged for analyzing writing styles in a supervised setting (Mosteller & 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 et al., 2003; Canny, 2004; Ramage et al., 2009).

Models based on topic-specific distributions over the vocabulary (such as LDA) cannot produce stable estimates of differential usage since they only model the relative frequency of words within topics. They cannot regularize usage across topics and naively infer the greatest differential usage for the rarest features (Eisenstein et al., 2011). We introduce the generative framework of *word rate models* that parameterizes topic-specific word counts as unnormalized count variates whose rates can be regularized across topics as well as within them, making stable inference of both word frequency and exclusivity possible. Word rate models can be seen as a fully generative interpretation of Sparse Topic Coding (Zhu & Xing, 2011) that emphasizes regularization and interpretability rather than exact sparsity. We introduce a parallelized Hamiltonian Monte Carlo (HMC) estimation strategy that makes full Bayesian inference efficient and scalable.

In this paper we focus on the case of document corpora for which meaningful topical structure is already avail-

Appearing in *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012.

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



able, avoiding ambiguities about summarizing a topic space that is not semantically meaningful. We utilize large, annotated collections such as *Reuters*, *New York Times*, *Wikipedia*, and *Encyclopedia Britannica* for which human coders have already created a hierarchical system of categories for end users. Working within the framework of word rate models, we develop Hierarchical Poisson Convolution (HPC), a generative model for labeled corpora that exploits the known topic hierarchy in these collections to make focused comparisons of differential use between neighboring topics on the tree and incorporates a sophisticated joint model for topic memberships and labels in the documents. Since HPC is designed to infer an interpretable description of human-generated labels rather than find optimally predictive covariates as with Supervised LDA (Perotte et al., 2012), we restrict the topic components to have a one-to-one correspondence with the human-generated labels. We then infer a clear semantic description of these labels in terms of words that are both frequent and exclusive.

2. Hierarchical Poisson Convolution

The Hierarchical Poisson Convolution model is a generative story for document collections whose topics are organized in a hierarchy. We refer to this structure 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 $d \in \{1, \dots, D\}$ is a record of counts w_{fd} for every feature in the vocabulary, $f \in \{1, \dots, V\}$. The length of the document is given by L_d, which we normalize by the average document length L to get $l_d \equiv \frac{1}{L} L_d$. Documents have unrestricted membership to any combination of topics $k \in \{1, \dots, K\}$ represented by a vector of labels $\boldsymbol{I}_d$ where $I_{dk} \equiv I\{\text{doc } d \text{ belongs to topic } k\}$.

Table 1. Generative process for HPC

Tree parameters: For feature $f \in \{1, \dots, V\}$:

- Draw $\mu_{f,0} \sim \mathcal{N}(\psi, \gamma^2)$
- Draw $\tau^2_{f,0} \sim$ Scaled Inv-$\chi^2(\nu, \sigma^2)$
- For $j \in \{1, \dots, J\}$ (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 $j \in \{1, \dots, J\}$ (terminal level of hierarchy):
  - Draw $\mu_{f,j1}, \dots, \mu_{f,jJ} \sim \mathcal{N}(\mu_{f,j}, \tau^2_{f,j})$
- Define $\beta_{f,k} \equiv e^{\mu_{f,k}}$ for $k \in \{1, \dots, K\}$

Topic parameters: For document $d \in \{1, \dots, D\}$:

- Draw $\boldsymbol{\xi}_d \sim \mathcal{N}(\boldsymbol{\eta}, \boldsymbol{\Sigma} = \lambda^2 \boldsymbol{I}_K)$
- For topic $k \in \{1, \dots, K\}$:
  - Define $p_{dk} \equiv 1/(1 + e^{-\xi_{dk}})$
  - Draw $I_{dk} \sim$ Bernoulli(p_{dk})
  - Define $\theta_{dk}(\boldsymbol{I}_d, \boldsymbol{\xi}_d) \equiv e^{\xi_{dk}} I_{dk} / \sum_{j=1}^{K} e^{\xi_{dj}} I_{dj}$

Data generation: For document $d \in \{1, \dots, D\}$:

- Draw normalized document length $l_d \sim \frac{1}{L}$Pois(υ)
- For every topic k and feature f:
  - Draw count $w_{fdk} \sim$ Pois$(l_d \boldsymbol{\theta}_d^T \boldsymbol{\beta}_f)$
- Define $w_{fd} \equiv \sum_{k=1}^{K} w_{fdk}$ (observed data)

The HPC model leverages the known topic hierarchy by assuming that words are used similarly in neighboring topics. 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 $\mu_{f,1}, \dots, \mu_{f,J}$ for high 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 high level topics, we then draw the log rates for their children 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 child 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.

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 give us 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 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 for 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 generate topic labels. Finally, both sets of parameters are then used to generate the words in each document. For simplicity of presentation we assume that each non-terminal node has J children and that the tree has only two levels below the corpus level, but the model can accommodate any tree structure.

2.1. Estimands

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_{fp}^2, 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.

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 FE_{fk} score as the harmonic mean of the word's rank in the distribution of $\phi_{.,k}$ and $\mu_{.,k}$:

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

where w is the weight for exclusivity (which we set to 0.5 as a default) and $\text{ECDF}_{x_{.,k}}$ 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 $\{\boldsymbol{\mu}_f, \boldsymbol{\tau}_f^2\}_{f=1}^{V}$ 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 $\{\boldsymbol{\mu}_f\}_{f=1}^{V}$.

[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.

Conditional on the hyperparameters, therefore, 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 Hamiltonian Monte Carlo (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 Supplemental Material.[2]

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:

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

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 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 non-zero entries only for topic pairs that have a direct parent-child relationship.[3] 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 density.

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:

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

We can write the likelihood of the word counts as a Poisson regression (now with the rates as covariates), with an independent contribution from the labels. 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 density.

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.

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:

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

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

[2] Available at `http://www.fas.harvard.edu/~airoldi/`

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

The variance parameters τ_{fk}^2 independently follow an identical Scaled Inverse-χ^2 with convolution parameter ν and scale parameter σ^2. The exact form of the conditional posterior of these hyperparameters is unknown, so we use HMC to sample from this density.

The document-specific topic affinity parameters $\boldsymbol{\xi}_d$ follow a 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:

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

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

3.2. Inference for unlabeled documents

In order to classify unlabeled documents, we need to find the 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 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 predictive distribution of this simpler model factors into tractable components:

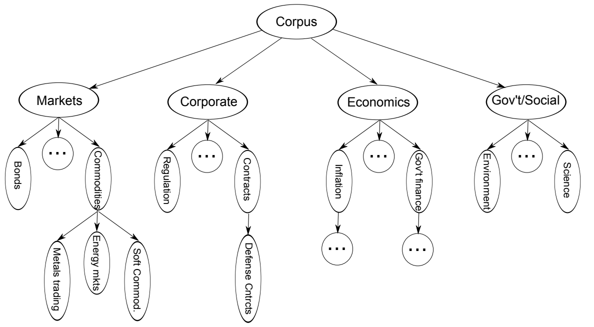
$$\begin{aligned} p^*(\boldsymbol{I}_{\tilde{d}}, \boldsymbol{\xi}_{\tilde{d}} | \boldsymbol{w}_{\tilde{d}}, \boldsymbol{W}, \boldsymbol{I}) &\propto p(\boldsymbol{I}_{\tilde{d}} | \boldsymbol{\xi}_{\tilde{d}})\ p^*(\boldsymbol{\xi}_{\tilde{d}} | \{\hat{\boldsymbol{\mu}}_f\}_{f=1}^V, \hat{\boldsymbol{\eta}},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

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.

Figure 2. Topic hierarchy of Reuters corpus



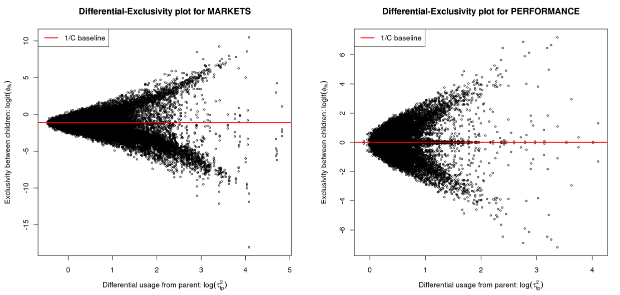
RCV1 is an archive of 806,791 newswire stories from a twelve-month period in 1996-1997.[4] As described in Lewis et al. (2004), Reuters staffers assigned stories into any subset of 102 hierarchical topic categories. In the original data, assignment to any topic required automatic assignment to all ancestor 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. 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.[5]

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 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 com-

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

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

Figure 3. Exclusivity as a function of variance parameters



modities. We present a graphical illustration of the Reuters topic hierarchy in Figure 2.

4.1. 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 learning 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.

4.2. 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 least able to escape regularization of their exclusivity parameters by our shrinkage prior on the parent's τ^2_{fp}.

Figure 4. Frequency-Exclusivity (FREX) plots

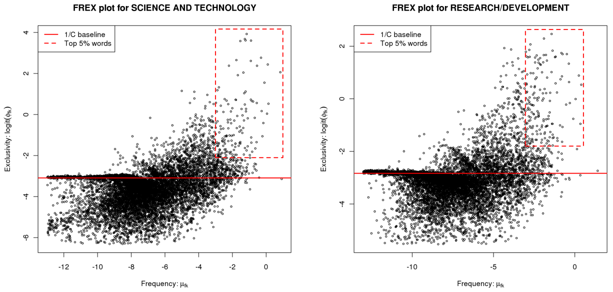


Figure 5. Upper right corner of FREX plot for SCIENCE AND TECHNOLOGY

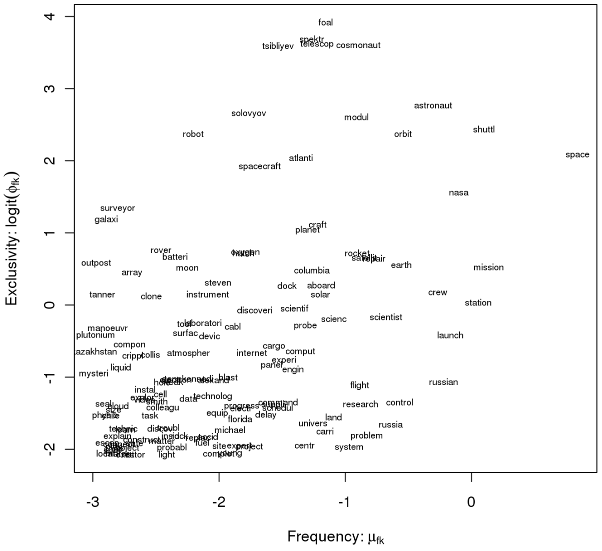
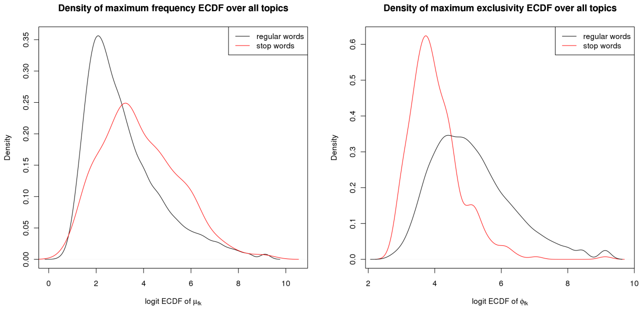


Figure 4 shows for two topics the joint posterior expectation of each word's frequency in that topic and its exclusivity compared to sibling topics (the FREX plot). The left panel features the Science and Technology topic, a child in the grab bag Government/Social branch, and the right panel features 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.

Figure 6. Comparison of FREX score components for SMART stop words vs. regular words



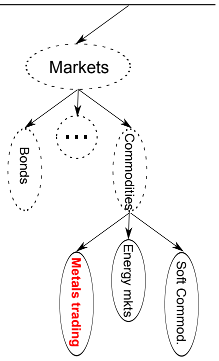
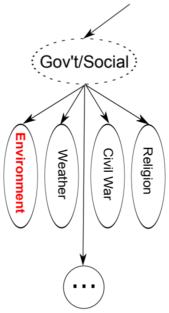
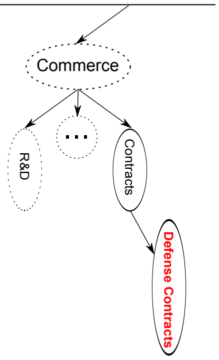
4.3. Frequency and Exclusivity as a two dimensional summary of topical content

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). These high-scoring words can help to clarify content even for labeled topics. For example, in the Science and Technology topic (shown in detail in Figure 5), we see almost all terms are specific to the American and Russian space programs.

We also compute the Frequency-Exclusivity (FREX) score for each word-topic pair, a univariate summary of topical content that averages performance in both dimensions. In Table 2 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 6, we show the maximum ECDF ranking for each word across topics in the distribution of frequency (left panel) and 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.

Table 2. Comparison of High FREX words to most frequent words (comparison set in solid ovals)

	High FREX	Most freq.	
Metals Trading	copper	said	 Markets ... Bonds Commodities Metals trading Energy mkts Soft Commod
	aluminium	gold	
	metal	price	
	gold	copper	
	zinc	market	
	ounc	metal	
	silver	trader	
	palladium	tonn	
	comex	trade	
	platinum	close	
	bullion	ounc	
	preciou	aluminium	
	nickel	london	
	mine	dealer	
Environment	greenpeac	said	 Gov't/Social Environment Weather Civil War Religion ...
	environment	would	
	pollut	environment	
	wast	year	
	emiss	state	
	reactor	nuclear	
	forest	million	
	speci	greenpeac	
	environ	world	
	eleph	water	
	spill	group	
	wildlif	govern	
	energi	nation	
	nuclear	environ	
Defense Contracts	fighter	said	
	defenc	contract	
	missil	million	
	forc	system	
	defens	forc	
	eurofight	defenc	
	armi	would	
	helicopt	aircraft	
	lockhe	compani	
	czech	deal	
	martin	fighter	
	militari	govern	
	navi	unit	
	mcdonnel	lockhe	

4.4. Classification performance

We compare the classification performance of HPC with SVM and a LDA+logit classifier, which fits a logistic regression using LDA topic loadings as covariates. 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, an experiment we repeated ten times with a new random sample as the training set. 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 three models to each training set and then used

Table 3. Classification performance

	SVM	LDA+Logit	HPC
Micro-ave Precision	0.711 (0.01)	0.596 (0.09)	0.695 (0.01)
Micro-ave Recall	0.706 (0.01)	0.594 (0.01)	0.589 (0.01)
Macro-ave Precision	0.563 (0.01)	0.372 (0.01)	0.505 (0.09)
Macro-ave Recall	0.551 (0.06)	0.332 (0.01)	0.524 (0.01)

Standard deviation of performance over ten folds in parenthesis.

the validation set to calibrate a threshold (except for SVM). Finally, we used the fit from the training set and the threshold from the validation set to predict topic memberships in the test set. Table 3 shows the results of our experiment, using both micro averages (every document weighted equally) and macro averages (every topic weighted equally). HPC performs better than LDA+logit on most metrics but does not dominate SVM, suggesting that there is a tradeoff between predictive performance and interpretability.

5. Concluding remarks

While HPC was developed for the specific case of hierarchically labeled document collections, the framework of word rate models can be readily extended to other types of document corpora. For labeled corpora where no hierarchical structure is available, one can use a flat hierarchy that treats all topics as siblings. For corpora where no labeled examples are available, a simple word rate model with a flat hierarchy and dense topic membership structure could be employed to get more informative summaries of inferred topics. In either case, this framework could be combined with non-parameteric Bayesian models that infer hierarchical structure on the topics (Adams et al., 2010). We expect that models based on rates will play an important role in future work on text summarization.

The HPC model can also 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 can be used as a gold standard to aid and evaluate other learning methods.

Targeted ontologies have been a key factor in monitoring scientific progress in biology (Ashburner et al., 2000; Kanehisa & 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 topical 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 then lists obtained with other models.

This work was partially supported by the National Science Foundation under grants DMS-1106980, IIS-1017967, and CAREER award IIS-1149662, and by the Army Research Office under grant 58153-MA-MUR, all to Harvard University. EMA is an Alfred P. Sloan Research Fellow.

References

Adams, R. et al. Tree-structured stick breaking for hierarchical data. NIPS 2010.

Airoldi, E. M., et al. Who wrote Ronald Reagan's radio addresses? *Bayesian Analysis*, 1(2):289–320, 2006.

Ashburner, M. et al. Gene ontology: Tool for the unification of biology. The gene ontology consortium. *Nature Genetics*, 25(1):25–29, 2000.

Blei., D. Introduction to probabilistic topic models. *Communications of the ACM*, 2012. In press.

Blei, David et al. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 2003b.

Canny, J. GAP: A Factor Model for Discrete Data. SIGIR, 2004.

Chang, J. et al. Reading tea leaves: How humans interpret topic models. NIPS, 2009.

Eisenstein, J. et al. Sparse Additive Generative Models of Text. ICML, 2011.

Hu, Y. et al. Interactive Topic Modeling. ACL, 2011.

Kanehisa, M. and Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1): 27–30, 2000.

Lewis, D. et al. RCV1: A New Benchmark Collection for Text Categorization Research. *Journal of Machine Learning Research*, 5:361–397, 2004.

McCallum, A. et al. Improving text classification by shrinkage in a hierarchy of classes. ICML, 1998.

Mimno, D. et al. Optimizing Semantic Coherence in Topic Models. EMNLP, 2011.

Mosteller, F. and Wallace, D.L. *Applied Bayesian and Classical Inference: The Case of "The Federalist" Papers*. Springer-Verlag, 1984.

Neal, R. MCMC using Hamiltonian dynamics. In Brooks, S., et al. (eds.), *Handbook of Markov Chain Monte Carlo*. Chapman & Hall / CRC Press, 2011.

Perotte, A. et al. Hierarchically Supervised Latent Dirichlet Allocation. NIPS, 2012.

Ramage, D. et al. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. EMNLP, 2009.

Wallach, H. et al. Rethinking LDA: Why Priors Matter. NIPS, 2009.

Zhu, J. and Xing, E. P. Sparse Topical Coding. UAI, 2011.