Parsimonious Topic Models with Salient Word Discovery
Hossein Soleimani, and David J. Miller

Abstract—We propose a parsimonious topic model for text corpora. In related models such as Latent Dirichlet Allocation (LDA), all words are modeled topic-specifically, even though many words occur with similar frequencies across different topics. Our modeling determines salient words for each topic that have topic-specific probabilities, with the rest explained by a universal shared model. Further, in LDA all topics are in principle present in every document. By contrast our model gives sparse topic representation, determining the (small) subset of relevant topics for each document. We derive a Bayesian Information Criterion (BIC), balancing model complexity and goodness of fit. Here, interestingly, we identify an effective sample size and corresponding penalty specific to each parameter type in our model. We minimize BIC to jointly determine the topic-specific words, document-specific topics, to estimate model parameters, and to estimate the total number of topics in a wholly unsupervised fashion. Results on three text corpora and an image dataset show that our model achieves higher test set likelihood and better agreement with ground-truth class labels, compared to LDA and to a model designed to incorporate sparsity.

Index Terms—Bayesian Information Criterion (BIC), Model selection, Parsimonious models, Topic models, Unsupervised feature selection

1 INTRODUCTION

PARSIMONIOUS models have been shown to perform better than parameter-rich models in various statistical modeling problems [1], [2]. The latter are more likely to overfit to the limited set of training samples, but parsimonious models can achieve higher posterior probability and better performance on unseen data. Document modeling is a good “target” for parsimonious representation, as the “bag-of-words” model [3] introduces one parameter per word, per topic, for each word in a given lexicon. This can amount to tens of thousands of word probability parameters for every topic in the model. It is expected that some economization on this model representation should be possible.

Recently, topics models such as Latent Dirichlet Allocation (LDA) [4] have been widely used to extract topics from collections of documents. Each topic is a probability mass function defined over a set of given vocabulary. LDA is an extension of standard mixtures of unigrams, with the latter assuming each document is generated by a single topic randomly selected based on corpus-level topic proportions [5]. By contrast, individual documents in LDA are modeled as mixtures of the extracted topics, with each document distinctly characterized by its topic proportions.

In LDA, each word has a freely chosen probability parameter under every topic. This entails a huge set of free parameters to learn and hence makes the model prone to over-fitting. Beyond arguing from a statistical modeling standpoint, we can also argue intuitively that many words do not have context-specific characteristics; i.e. they are used with roughly the same frequency under different topics. Thus, only a subset of the words should be modeled in a context-specific (topic-specific) fashion, with the rest reasonably explained by a universal shared model across all topics.

Similarly, in LDA every topic is in principle present in every document, by having non-zero proportion and contributing to the generation of at least one word. This seems implausible since each document is expected to have a main theme covered by a modestly sized subset of related topics. Allowing all topics to have nonzero proportions in every document again increases the complexity of the model’s representation of the data. In this paper we propose a parsimonious topic model to address this issue. Specifically, our proposed method identifies a sparse set of topics present in each document.

Our model is in fact sparse in two senses. First, we follow a parsimonious approach in modeling words under different topics. Here, a shared feature representation is introduced to model many words in a universal fashion. Under a given topic, each word may be a topic-specific feature with its own probability parameter or a shared feature, using a parameter shared with other topics. Secondly, for each document a sparse set of “occurring” topics – those with non-zero proportions – is identified.

Learning in our proposed model consists of two main parts that are alternately, iteratively applied: structure learning and parameter learning. Assuming that the number of topics is known and given current parameter estimates, structure learning involves determining the set of topic-specific words for each topic as well as the topics that have non-zero proportions in each document.

* H. Soleimani (email: hsoleimani@psu.edu) and D. J. Miller (email: djmiller@engr.psu.edu) are with the Department of Electrical Engineering, Pennsylvania State University, University Park, PA, 16802.
In the parameter learning phase given a fixed structure, all “active” free parameters (document-specific topic proportions and topic-specific word probabilities), are estimated. In the following sections we introduce a joint objective function and propose a computationally efficient procedure to jointly perform both the structure and parameter learning phases, consistent with minimization of this objective.

Our objective function is derived from the model posterior probability \[ \pi \] [6,7]. One criterion derived by approximating the model posterior is the Bayesian Information Criterion (BIC) [8]. BIC, which is a widely accepted criterion for model comparison, is the negative logarithm of the Bayes marginal likelihood and is composed of two main terms: the data log-likelihood and the model complexity cost terms. Accordingly, BIC achieves a balance between goodness of fit and model complexity.

However, BIC has some deficiencies which limit its applicability. The parameter penalty in BIC (\( \frac{1}{2} \log(\text{sample size}) \)) is the same for all parameters in the model [9], while obviously different parameter types introduce different degrees of complexity to the model and contribute unequally to the data goodness of fit. Moreover, the Laplace’s approximation used in deriving BIC is only valid when the sample size is large compared to the size of the feature space. More specifically, validity of this criterion is doubtful when the sample size is less than 5 times the number of free parameters [5]. A small sample size to parameters’ size ratio is often the case in applications such as document modeling, where, nominally, the number of free parameters for each topic is equal to the dictionary size.

In this paper, we derive an approximate objective function starting from the model posterior, which improves on the naïve form of BIC in two aspects: 1) Our proposed form of BIC has differentiated cost terms, based on different effective sample sizes, for the different parameter types in our model. 2) Making use of a shared feature representation essentially increases the sample size to parameters’ size ratio, allowing the sample size to be large enough to provide a valid approximation of the model posterior.

Our proposed framework also provides, in a wholly unsupervised fashion, a direct estimate of the total number of topics present in the corpus. The number of topics (i.e. model order) is a hyper-parameter in topic models which is usually determined by evaluating model performance on a secondary task such as classification, using a validation dataset. Here, solely using the (unlabeled) document corpus, we select the model order that has the highest approximate model posterior (minimum BIC), i.e. model structure, parameter values, and model order are jointly chosen to minimize BIC.

1.1 Related Work
The concepts of “common” and “specific” words, sparsity in topic proportions and word probabilities, as well as estimation of the number of topics have been the subject of some previous studies. [9] introduced asymmetric Dirichlet priors over topic distributions and word probabilities to control skewness in word and topic frequency distributions. Asymmetric priors were shown to be effective in preventing common words from dominating all topics and also in achieving sparser topic presence in documents. However, similar to LDA, this approach is not parsimonious. All topics have nonzero proportion in every document and all words are modeled in a topic-specific fashion.

[10] introduced a spike and slab model to control sparsity in word probabilities. Unlike our approach, this method does not use a shared distribution to model words that are not topic-specific. Moreover, it does not provide the subset of relevant topics for each document. A similar approach based on the Indian Buffet Process was used in [11] to address sparsity only in topic proportions in a non-parametric topic model.

[12] proposed a combination of background, general, and document-specific topics to improve information retrieval. The authors argued that LDA “over-generalizes” and is thus not effective for matching queries that contain both high-level semantics as well as word-level keywords. [12] introduced a huge set of new free parameters by adding a document-specific topic for every document. Accordingly, similar to LDA, each word, under every topic, has a free parameter that must be estimated separately. By contrast, in our model, the shared model is heavily used by all topics, with each topic possessing relatively few topic-specific words. Also unlike [12], our model is sparse in topic proportions and we determine the subset of occurring topics for each document.

[13] presented Sparse Topical Coding (STC), a non-probabilistic topic model which only gives parsimony in topic-proportions. Unlike our approach, this model does not have sparsity in word probabilities and models all words in a topic-specific fashion. Moreover, this method has three hyper-parameters which must be determined by cross-validation, while our model has none.

Non-parametric topic models have been proposed that relax the requirement of specifying the number of topics [14]. However similar to LDA, these methods do not exhibit parsimony in their modeling.

Our approach can also be viewed from the standpoint of unsupervised feature selection. For each topic we select a set of salient features in an unsupervised fashion and model the rest of the features using a universal shared model. Unsupervised feature selection and shared feature representations have been considered in some prior works. [15] used a minimum message length criterion to find salient features in a mixture model. Features in [15] were tied across all components; i.e. each feature is either salient or shared in all components. A Bayesian framework for a similar shared feature space was presented in [16] for Gaussian mixture models.

The concept of shared feature space for standard mixture models was further improved in [17] by proposing
a component-specific feature space; i.e. a feature can be salient in some components but represented via a shared model in others. [17] used the Minimum Description Length (MDL) [18] and standard mixture of unigrams for modeling text documents. [19] performed unsupervised feature selection by minimizing the message length of the data, considering mixtures of generalized Dirichlet distributions. This model was then optimized in a Bayesian framework via variational inference in [20].

 Contributions of this paper: Compared to previous works, the main contributions of this paper are:

1) We extend the concept of shared feature space from standard mixture models to more general topic models, which allows presence of multiple topics in documents. In doing so, we achieve two sources of parsimony in our model: sparsity in topic proportions as well as in topic-specific words. Prior works at best achieve sparsity in one of these two dimensions.

2) Unlike many previous works, our model allows the subset of salient words to be topic-specific. This follows the intuition that, in document modeling applications, some words may have a common meaning (common frequency of occurrence) under some subset of topics, but unique meanings (different frequencies of occurrence) under other individual topics. For example, the word “component” has different meanings under “statistics” and “machine learning” topics than under other topics, and could have higher frequency of occurrence for these specialized topics.

3) We derive a BIC criterion specific to our model. This criterion is our overall objective function that we minimize in performing integrated structure and parameter learning and in determining the model order. Unlike the naive form, satisfyingly, our derived BIC objective function has distinct penalty terms for the different parameter types in our model, which can be interpreted vis a vis different effective sample sizes for the different parameter types.

The rest of the paper is organized as follows: Section 2 introduces the notation we use throughout. In section 3 we briefly review LDA and introduce a criterion to measure sparsity of topics for LDA. Next, in section 4 we present our parsimonious model. Section 5 derives our BIC objective function. Then, in section 6 we develop the joint structure and parameter learning algorithm for our model, which locally minimizes our BIC objective. Experimental results on three text corpora and an image dataset are reported in section 7. Finally, concluding remarks are reported in section 8.

2 NOTATION AND TERMINOLOGY

A corpus $D$ is a collection of $D$ documents and a dictionary is a set of $N$ unique words. We index unique documents in the corpus and unique words in the dictionary by $d \in \{1,\ldots,D\}$ and $n \in \{1,\ldots,N\}$, respectively. Also, $j \in \{1,\ldots,M\}$ is a model’s topic index, $M$ the total number of topics.

We define the following terms specific to our modeling of each document $d$:

- $L_d$ is the number of words in document $d$.
- $w_{id} \in \{1,\ldots,N\}$, $i = 1,\ldots,L_d$ is the $i$-th word in document $d$.
- $M_d \in \{1,\ldots,M\}$ is the number of topics present in document $d$.
- $\alpha_{jd}$ is the topic proportion of topic $j$ in document $d$.

The following quantities are specific to each topic:

- $p_j(n)$ is the probability of word $n$ under topic $j$.
- $N_j$ is the number of topic-specific words in topic $j$.
- $L_j$ is the sum of the lengths of the documents for which topic $j$ is present.

Finally, $p_0(n)$ is the probability of word $n$ under the shared model.

3 LATENT DIRICHLET ALLOCATION

LDA is a generative model originally proposed for extracting topics and organizing text documents [4]. Its generative process for a document $d$ is as follows:

1) Choose topic proportions $\theta_d$ from a Dirichlet distribution with hyper-parameter $\eta$, i.e. $\theta_d \sim \text{Dir}(\eta)$.
2) For each of the $L_d$ words $w_{id}$:
   a) Choose a topic $z_{id} \sim \text{Multinomial}(\theta_d)$.
   b) Choose word $w_{id}$ according to the multinomial distribution for topic $z_{id}$, i.e. $\{p_{z_{id}}(n), n = 1,\ldots,N\}$.

3.1 Parameter Estimation

The main inference in LDA is computing the posterior probability of hidden variables given a document. This involves computing the likelihood of a document (i.e. the normalization constant in the posterior) by marginalizing out hidden variables, which is computationally intractable. A variety of approximate algorithms such as variational inference [4] and Markov Chain Monte Carlo [21] have been used to address this problem. Here we briefly review mean-field variational inference.

Approximate inference in this method is essentially done by obtaining a lower bound on the log-likelihood. A new family of variational distributions is defined by changing some of the statistical dependencies in the original model:

$$q(\theta_d, z_{id} | \gamma^{(d)}, \phi^{(d)}) = q(\theta_d | \gamma^{(d)}) \prod_{i=1}^{L_d} q(z_{id} | \phi_i^{(d)}), \forall d.$$  \hspace{1cm} (1)

Here, $z_{id}$ is the set of hidden variables $\{z_{1id},\ldots,z_{L_d id}\}$ and $q(\theta_d | \gamma^{(d)})$ is a Dirichlet distribution with hyperparameter $\gamma^{(d)}$. Also, $q(z_{id} | \phi_i^{(d)})$ is a multinomial distribution on the $M$ topics, with hyper-parameters
The parameters $\phi^{(d)}_{1}, \ldots, \phi^{(d)}_{M}$. The values of the variational parameters ($\gamma$ and $\phi$) for each document are determined by minimizing the Kullback-Leibler (KL) divergence between $q$ and the posterior distribution of hidden variables, which gives a lower bound on the single-document log-likelihood. Update equations of the variational parameters are:

$$\phi^{(d)}_{ij} \propto p_j(w_{id}) \exp \left( \psi(\gamma^{(d)}_{ij}) - \Psi(\sum_{j'=1}^{M} \gamma^{(d)}_{j'}) \right)$$

$$\gamma^{(d)}_{j} = \eta_{j} + \sum_{i=1}^{L_{d}} \phi^{(d)}_{ij}$$

where $\Psi(\cdot)$ is the first derivative of the log-gamma function. Also the $\phi$ parameters are constrained so that $\sum_{j=1}^{M} \phi^{(d)}_{ij} = 1$, $i = 1, \ldots, L_{d}$, $d = 1, \ldots, D$.

In the next step, after optimizing the variational parameters, the lower bound is optimized with respect to parameters of the model (i.e. the word probabilities under each topic and the Dirichlet hyper-parameter, $\eta$). For the word probabilities, this minimization is achieved via the closed form updates:

$$p_j(n) = \frac{\sum_{d=1}^{D} \sum_{i=1}^{L_{d}} \phi^{(d)}_{ij}}{\sum_{d=1}^{D} \sum_{i=1}^{L_{d}} \phi^{(d)}_{ij}}, \forall n, j.$$ (4)

There is no closed-form update for the Dirichlet hyper-parameter $\eta$. It is updated using Newton-Raphson. This process is iteratively repeated until a termination criterion is met.

### 3.2 Sparsity in LDA

The sparsity of topic proportions is controlled by the hyper-parameter of the Dirichlet distribution, which is a corpus-level parameter and is optimized along with all other parameters of the model. Values of $\eta$ smaller than 1 lead to sparser topic proportions, which indicate that fewer topics have significant presence in a document.

In order to compare sparsity in LDA with our proposed model, we use a criterion to measure the actual number of topics present in each document, as estimated by LDA. The variational parameter $\phi^{(d)}_{ij}$ is essentially the probability that word $i$ in document $d$ is generated by topic $j$. To estimate the number of topics present, we hard-assign each word in a document to the topic that has the maximum $\phi^{(d)}_{ij}$, $j = 1, \ldots, M$. In doing so, we determine the set of topics used by LDA in modeling at least one word in a given document. This is how we measure topic sparsity for LDA.

### 4 Parsimonious Topic Model

In this section we propose our parsimonious model and its associated parameter estimation. For clarity’s sake, in this section, we will focus only on parameter estimation given the total number of topics and the model structure assumed known, i.e. the subset of topics present in each document and the topic-specific words under each topic. Structure learning and model order selection will be discussed later in section 5.

In our model, unlike LDA, we do not treat model parameters as random variables, within a Bayesian setting, but rather as deterministic variables, to be estimated within a maximum likelihood setting. While the Bayesian setting is a natural alternative to our approach, our approximation of the model posterior (developed in section 5), which involves an approximate “integrating out” of parameters, provides similar benefits of a fully Bayesian framework in avoiding overfitting. Moreover, unlike a fully Bayesian approach, our approach gives a computationally tractable way for learning parsimonious models, and thus for avoiding overfitting.

The generative process for a document in our model is similar to LDA. For each word position $i$ in a document, we randomly select one of the topics that has nonzero proportion, and then randomly choose a word $w_{id}$ from that topic. All $D$ documents in the corpus are generated i.i.d using the same process. The likelihood of the corpus $D$ under our parsimonious model is thus:

$$P(D | \Theta, \mathcal{H}) = \prod_{d=1}^{D} \prod_{i=1}^{L_{d}} \sum_{j=1}^{M} \alpha_{jd} v_{jd} \left( \frac{p_j(w_{id})^{v_{jd}}}{p_0(w_{id})^{1-v_{jd}}} \right).$$ (5)

Here, the binary switches $u = \{u_{jn}\}$ indicate topic-specific or shared words, with $(u_{jn} = 1)$ if word $n$ is specific to topic $j$, in which case probability parameter $p_j(n)$ is used, or shared $(u_{jn} = 0)$, in which case probability parameter $p_0(n)$ is used. The set of topic-specific and shared word probabilities under each topic are constrained to give a valid probability mass function (pmf) i.e. $(\sum_{n=1}^{N} u_{jn} p_j(n) + (1 - u_{jn}) p_0(n)) = 1$, $\forall j$. Similarly, the binary switches $v = \{v_{jn}\}$ are defined by $(v_{jd} = 1)$ if topic $j$ is present in document $d$, with proportion $\alpha_{jd}$; otherwise $(v_{jd} = 0)$. The topic proportions must satisfy the pmf constraints $\sum_{j=1}^{M} \alpha_{jd} v_{jd} = 1$, $\forall d$.

The set of binary switches $u$ and $v$ together with the total number of topics, $M$, thus specify the model structure or family $\mathcal{H}(M, v, u)$. In the parameter learning step of our algorithm we estimate the model parameters assuming the model structure is known, i.e. $\Theta = \{p_0(\cdot)\}, \{p_j(\cdot)\}, \{\alpha_{jd}\}$.

#### 4.1 Parameter Estimation

The parameters $\Theta$ are chosen to maximize the log-likelihood of the training data $D$, via the Expectation Maximization (EM) framework [22], which achieves locally optimal parameter estimates. Here the topic of origin for each word in every document is chosen as the hidden data $(C_{id})$ within EM. Given the model structure and the current estimate of the parameters $\Theta^{(t)}$, in the E-step we compute the expected values of the hidden
variables, i.e. the probabilities:

\[
P(C_{id} = j|w_{id}; \Theta^{(i)}, \mathcal{H}) = \frac{\alpha_{jd} v_{jd} p_j(w_{id}) u_{jd} p_0(w_{id})}{\sum_{i=1}^M \alpha_{id} v_{id} p_i(w_{id}) u_{id} p_0(w_{id})}, \quad \forall i, d.
\]  

(6)

Adding normalization constraints to the expected value of the complete data log-likelihood measured with respect to \(P\), we construct the Lagrangian at the current parameter set estimate \(\Theta^{(i)}\):

\[
E[\log(P(D|\Theta, \mathcal{H}))|\Theta^{(i)}, \mathcal{H}] = \sum_{d=1}^D \sum_{i=1}^L_d \sum_{j=1}^M v_{jd} 
\cdot P(C_{id} = j|w_{id}; \Theta^{(i)}, \mathcal{H}) \left( \log(\alpha_{jd}) + u_{jd} \log(p_j(w_{id})) \right) 
+ (1 - u_{jd}) \log(p_0(w_{id})) \right) - \sum_{j=1}^M \lambda_d \left( \sum_{j=1}^M \alpha_{jd} v_{jd} - 1 \right) 
- \sum_{j=1}^M \sum_{n=1}^N \left( u_{jn} p_j(n) + (1 - u_{jn}) p_0(n) \right) - 1.
\]

(7)

5 BIC DERIVATION

In this section we derive our BIC objective function, with respect to which we jointly optimize the model structure \(\mathcal{H}(M, v, u)\) and the model parameters, \(\Theta\). At fixed order \(M\), given fixed model parameters, we will minimize this objective function with respect to \((v, u)\), thus optimizing the model structure. Alternately, given fixed model structure, minimizing this objective function with respect to the model parameters is equivalent to maximizing the data log-likelihood (and is thus achieved as described in section 4.1).

The posterior probability of the model family \(\mathcal{H}\) is proportional to \(p(D|\mathcal{H})p(\mathcal{H})\), \(p(D|\mathcal{H})\) the marginal likelihood and \(p(\mathcal{H})\) the prior probability on the structure. The marginal likelihood is computed by taking expectation of the data likelihood with respect to the prior distribution on the parameters:

\[
I = p(D|\mathcal{H}) = \int p(D|\mathcal{H}, \Theta)p(\Theta|\mathcal{H})d\Theta,
\]

(12)

where \(p(\Theta|\mathcal{H})\) and \(p(D|\mathcal{H}, \Theta)\) are, respectively, the prior on parameters for model structure \(\mathcal{H}\) and the data likelihood under the model. We use Laplace’s method to approximate (12). This is based on the assumption that \(p(D|\mathcal{H}, \Theta)p(\Theta|\mathcal{H})\) is peaked around its maximum (the posterior mode \(\tilde{\Theta}\)) which in fact is valid for large sample sizes [6].

We approximate \(\log(p(D|\mathcal{H}, \Theta)p(\Theta|\mathcal{H}))\) around the posterior mode using a Taylor’s series expansion. Note that the first Taylor term is constant with respect to the variables of integration and the term corresponding to first derivatives is zero. Thus, exponentiating the Taylor’s series expansion at second order, we obtain the product,

\[
\hat{I} = p(D|\mathcal{H}, \tilde{\Theta})p(\tilde{\Theta}|\mathcal{H}) \cdot \int \exp \left( -\frac{1}{2} (\Theta - \tilde{\Theta})^T \Sigma(\Theta - \tilde{\Theta}) \right) d\Theta,
\]

(13)

where \(\Sigma\) is the Hessian matrix.

The integrand in (13) is a scaled normal distribution with mean \(\tilde{\Theta}\) and covariance \(\tilde{\Sigma}\), which when integrated evaluates to \((2\pi)^{k/2}|\Sigma|^{-1/2}\). Here, \(k\) is the number of parameters in \(\Theta\) which is equal to \(\sum_{d=1}^D M_d + \sum_{j=1}^M N_j\), where \(M_d = \sum_{j=1}^M v_{jd}\) is the number of active topics in document \(d\) and \(N_j = \sum_{n=1}^N u_{jn}\) is the number of topic-specific words under topic \(j\). Also, since the shared model parameters are estimated in a universal fashion, once, and then fixed (11), their descriptive complexity is assumed to be fixed, irrespective of \(\mathcal{H}(M, v, u)\) and \(\Theta\). Thus, this (assumed constant) term need not be included in our objective function, and the shared model parameters need not be counted as free parameters in \(\Theta\).

We thus have the following approximation for the marginal likelihood [6]:

\[
\hat{I} = (2\pi)^{k/2}|\tilde{\Sigma}|^{-1/2} p(D|\mathcal{H}, \tilde{\Theta})p(\tilde{\Theta}|\mathcal{H}).
\]

(14)
Based on (14), an approximate negative log-model posterior (BIC cost) is:

\[
BIC = -\log(\hat{I}) - \frac{k}{2} \log(2\pi) + \frac{1}{2} \log(|\hat{\Sigma}|)
\]

By minimizing BIC, a tradeoff is achieved between the data log-likelihood, \(\log(p(\mathcal{D}|\mathcal{H}, \Theta))\), and the other terms, interpreted as the model “complexity” cost.

The Hessian matrix \(\hat{\Sigma}\) in (14), neglecting \(p(\Theta|\mathcal{H})\), is:

\[
[\hat{\Sigma}]_{ij} = \left. \frac{\partial^2 \log(p(\mathcal{D}|\mathcal{H}, \Theta))}{\partial \Theta_i \partial \Theta_j} \right|_{\Theta = \hat{\Theta}}, i, j = 1, 2, \ldots k. \tag{16}
\]

The diagonal elements corresponding to topic proportions are:

\[
\frac{\partial^2 \log(P(\mathcal{D}|\mathcal{H}, \Theta))}{\partial \alpha_j^2} \bigg|_{\Theta = \hat{\Theta}} = -v_j \sum_{l=1}^{L_d} \frac{\partial^2 \log(P(w_{ld}|\mathcal{H}, \Theta))}{\partial \alpha_j^2} \bigg|_{\Theta = \hat{\Theta}}, \tag{17}
\]

where

\[
P(w_{ld}|\mathcal{H}, \Theta) = \sum_{j=1}^M \alpha_j v_{ldj} p_{0j} (w_{ld})^{1-v_{ldj}}. \tag{18}
\]

By invoking the weak law of large numbers, we have with probability 1 that

\[
\frac{\partial^2 \log(P(\mathcal{D}|\mathcal{H}, \Theta))}{\partial \alpha_j^2} \bigg|_{\Theta = \hat{\Theta}} = -v_j \sum_{l=1}^{L_d} \frac{\partial^2 \log(P(w_{ld}|\mathcal{H}, \Theta))}{\partial \alpha_j^2} \bigg|_{\Theta = \hat{\Theta}} = -v_j L_d \frac{\partial^2 E[\log(P(w_{ld}|\mathcal{H}, \Theta))]}{\partial \alpha_j^2} \bigg|_{\Theta = \hat{\Theta}} = -v_j L_d \hat{I}_{jd}^{(\alpha)}, \tag{19}
\]

where \(\hat{I}_{jd}^{(\alpha)}\) is the corresponding element in the Fisher information matrix. Similarly, diagonal elements of the Hessian matrix corresponding to \(p_j(n)\) are

\[
\frac{\partial^2 \log(P(\mathcal{D}|\mathcal{H}, \Theta))}{\partial p_j(n)^2} =
\]

\[
- u_j \sum_{d=1}^D \sum_{l=1: w_{ld}=n} v_{ld} \frac{\partial^2 \log(P(w_{ld}|\mathcal{H}, \Theta))}{\partial p_j(n)^2} \bigg|_{\Theta = \hat{\Theta}} = u_j \sum_{d=1}^D L_d v_{ldj} \hat{I}_{jd}^{(p)} , \tag{20}
\]

where \(I_{jd}^{(p)}\) is the Fisher information matrix element corresponding to the probability of word \(n\) under topic \(j, p_j(n)\). Moreover, we define \(\hat{L}_j = \sum_{d=1}^D L_d v_{ldj}\).

Following the common approach in deriving BIC [23], we neglect the off-diagonal elements and write \(\hat{\Sigma}\) as a block-diagonal matrix:

\[
\hat{\Sigma} = \begin{bmatrix} \hat{\Sigma}_{ff} & 0 \\ 0 & \hat{\Sigma}_{p} \end{bmatrix}, \quad \text{where} \quad \hat{\Sigma}_{ff} \triangleq \text{Diag}[L_d \hat{I}_{jd}^{(\alpha)}], \quad \hat{\Sigma}_{p} \triangleq \text{Diag}[L_j \hat{I}_{jn}^{(p)}]. \tag{21}
\]

As the sample size grows, the Fisher matrix terms \(I_{jd}^{(p)}\) and \(I_{jn}^{(p)}\) become negligible [23]. Thus, the log-determinant of \(\hat{\Sigma}\) is:

\[
\log(|\hat{\Sigma}|) \approx \sum_{d=1}^D (M_d - 1) \log(L_d) + \sum_{j=1}^M N_j \log(\hat{L}_j). \tag{22}
\]

This expression is in fact reminiscent of the penalty term in the naïve form of BIC where, for each parameter, one “pays” \(\frac{1}{2} \log(\text{sample size})\) But here in (22), we identify specific cost terms for each of the two different parameter types in our model. That is, the effective sample size in our BIC model cost is \(L_d\) for the \(\alpha\) parameters and \(\hat{L}_j\) for topic-specific word probabilities \(p_j(\cdot)\).

The other important term in (15) is the prior probability of the model family \(p(\mathcal{H})\), defined based on the configurations of \(u\) and \(v\) switches for a given number of topics \(M\). Here, we assume that priors on \(u\) and \(v\) switches are independent.

Our general principle in defining these prior distributions is to invoke uninformativeness (uniformity). We parameterize the prior distribution on \(u\) switches as a function of the total number of topic-specific words across all topics, \(N_0\), and (uninformatively) assume all configurations with at most \(N_0\) “on” switches \(\sum_{n=1}^N (w_{ln})\) are equally likely. However, since our belief is that in practice a small number of words across all topics will be topic-specific \((N_0 \ll M N)\), the total number of these configurations can be well-approximated by [24]:

\[
\sum_{n=1}^N \binom{M N}{n} \approx \frac{2^{M N H(\frac{N}{M})}}{\sqrt{MN}}, \tag{23}
\]

where \(H(\cdot)\) is Shannon’s entropy for a Bernoulli random variable with parameter \(\frac{N}{M}\) [24]. Accordingly, the corresponding term in BIC is:

\[
- \log(p(u)) = M N H(\frac{\bar{N}}{N}) \log(2) - \frac{1}{2} \log(M N), \tag{24}
\]

where \(\bar{N} \triangleq \frac{1}{M} \sum_{j=1}^M N_j\) is the average number of topic-specific words across all topics.

To define the prior on \(v\) configurations, we consider a two stage process. First, the number of topics in a document \(d\) \((M_d)\) is selected from a uniform distribution \(\binom{M}{d}\), and then a switch configuration is selected from a uniform distribution over all \(\binom{M_d}{d} \) configurations with \(M_d\) “on” switches. Therefore,

\[
- \log(p(v)) = D \log(M) + \sum_{d=1}^D \log \binom{M}{M_d}. \tag{25}
\]
By substituting (22), (24), and (25) into (15), our overall BIC expression becomes:

\[
\begin{align*}
BIC &= D\log(M) + \sum_{d=1}^{D} \log \left( \frac{M}{M_d} \right) + MNH \left( \frac{\bar{N}}{N} \right) \log(2) \\
&\quad - \frac{1}{2} \log(MN) + \frac{1}{2} \sum_{d=1}^{D} (M_d - 1) \log\left( \frac{L_d}{2\pi} \right) \\
&\quad + \frac{1}{2} \sum_{j=1}^{M} N_j \log\left( \frac{\bar{L}_j}{2\pi} \right) - \log(p(D|H, \Theta)).
\end{align*}
\]

(26)

6 INTEGRATED MODEL SELECTION AND PARAMETER ESTIMATION

In this section we develop our algorithm to jointly estimate the structure \(H(M, v, u)\) and the parameters \(\Theta\) of the model. This integrated learning is performed by minimizing the BIC objective (25).

Supposing, for the moment, that the number of topics, \(M\), is fixed, there are \(2^{MD+NM}\) different possible \((u, v)\) switch configurations –minimization of BIC over this search space is a formidable task. However, using a generalized Expectation Maximization (GEM) algorithm (22), (26), we can break up this problem into a series of simple update steps. Each of these steps is decreasing in BIC, with the overall algorithm thus guaranteed to converge to a local minimum.

More specifically, given a fixed model order \(M\), our GEM algorithm consists of the following steps, iterated until convergence:

1) \textbf{E-step}: Following section 4.1, the topic of origin for each word in each document comprise the hidden variables in the EM framework. Expected values of the hidden variables are given in (5). The only term in (26) that depends on these hidden variables is the negative log-likelihood term. Thus, the expected value of the complete BIC is achieved simply by replacing the incomplete data log-likelihood term in (26) by the expected complete data log-likelihood term from (7).

2) \textbf{Generalized M-step}:

a) Estimation of \(\Theta\) given fixed structure: Since the model complexity terms in BIC have no dependence on \(\Theta\), minimization of the expected complete data BIC with respect to \(\{\alpha_{jd}\}\) and \(\{p_j(n)\}\) given fixed structure is equivalent to maximizing the expected complete data log-likelihood, with closed form updates as described in section 4.1.

b) Optimization of \((u, v)\) given fixed \(\Theta\).

Updating the \((u, v)\) switches given fixed parameters \((step\ 2(b))\) is done via an iterative loop in which switches are cyclically visited one by one. Here at each step, all parameters and switches but one are kept fixed and the effect of flipping one switch on BIC is investigated. If a change in the current switch reduces BIC (or expected complete data BIC), that change is accepted. This process is repeated in a cycle over all switches until no further decrease in BIC occurs or a predefined maximum number of cycles is reached. We update the \(u\) and \(v\) switches separately, first performing the update cycles over all \(u\) switches until convergence and then updating the \(v\) switches until convergence. We then go back to the E-step.

A change in one switch affects both the model cost \((\Delta Cost)\) and log-likelihood \((\Delta L)\) terms in BIC. In the following sections, we determine \((\Delta BIC = \Delta Cost - \Delta L)\) associated with switch updates. To ensure that updates are decreasing in BIC, changes are accepted only if \((\Delta BIC < 0)\).

6.1 Updating \((u_{jn})\)

Here, we trial-flip each switch \(u_{j'n'}\), \(\forall n' = 1,...,N\) in every topic \(j' = 1,...,M\), one by one. In performing these updates, each update is constrained to have at least one topic-specific word.

In the model cost terms of BIC, a change in the value of \(u_{j'n'}\) only affects the number of topic-specific words in that topic, \(N_{j'}\). Thus,

\[
\Delta Cost(u_{j'n'}) = \frac{1}{2} \left( -1 \right)^{u_{j'n'}} \log\left( \frac{\bar{L}_{j'}}{2\pi} \right) + N M \log(2) \left( H\left( \frac{\bar{N}^+}{N} \right) - H\left( \frac{\bar{N}^-}{N} \right) \right),
\]

(27)

where \(-\) and \(+\) superscripts denote the current and new values of functions of the switch, respectively.

In computing the change in the log-likelihood, we need the corresponding word probability under the new and old values of the switch. Introducing the variables \(x_{jn}\) and \(\bar{x}_j\) from (9) and (10) allows us to compute \(\mu_{j'}\) efficiently (without performing the E-step and recomputing the required statistics). The effect of changing \(u_{j'n'}\) on the expected value of the complete data log-likelihood is:

\[
\begin{align*}
\Delta L(u_{j'n'}) &= \left( -1 \right)^{u_{j'n'}} x_{j'n'} \cdot \left( \log\left( \frac{\bar{x}_{j'n'}}{\mu_{j'}} \right) - \log(p_0(n')) \right) - \bar{x}_j \log\left( \frac{\mu_{j'}^+}{\mu_{j'}} \right),
\end{align*}
\]

(28)

where:

\[
\mu_{j'}^+ = \frac{\bar{x}_j - \left( -1 \right)^{u_{j'n'}} x_{j'n'}}{1 \sum_{n=1}^{N} \left( 1 - u_{j'n} \right)p_0(n) + \left( -1 \right)^{u_{j'n'}} p_0(n')},
\]

(29)

2. For computational efficiency, optimization over \(u\) is with respect to the expected complete data BIC. However for \(v\), we directly minimize BIC since there is no computational advantage in minimizing with respect to expected BIC in this case. Since optimizing expected BIC ensures improvement in BIC, both \(u\) and \(v\) update steps descend in the BIC objective.
6.2 Updating \( \{v_{jd}\} \)

We update the \( v \) switches by sequentially visiting each topic \( v_{j'd'} \) \( j' = 1, ..., M \) in every document \( d' = 1, ..., D \). Obviously, each document should have at least one active topic. Moreover, every topic is constrained to be used in at least one document in the corpus.

Trial-flipping switch \( v_{j'd'} \) changes \( M_{d'} \) and \( \bar{L}_{j'} \). Therefore,

\[
\Delta \text{Cost}(v_{j'd'}) = \log\left( \frac{M_{d'}}{M_d} \right) + \frac{1}{2} (-1)^{v_{j'd'}} \cdot \left[ \log \left( \frac{L_{d'}}{2\pi} \right) + \log \left( \frac{\bar{L}_{j'} + (-1)^{v_{j'd'}} L_{d'}}{L_{j'}} \right) \right].
\]

Unlike the \( u \) updates, since expected values of hidden variables are zero for topics \( j' \) such that \( v_{j'd'} = 0 \), we need to recompute (16) for document \( d' \) and trial-update \( \alpha_{j'd'} \forall j = 1, ..., M \) using (3). Then, we compute \( \Delta L(v_{j'd'}) \) based on the incomplete data log-likelihood:

\[
\Delta L(v_{j'd'}) = \sum_{i=1}^{L_{d'}} \log \left( \sum_{j=1}^{M} \alpha_{j'd'} v_{jd}' p_j(w_{id'}) u_{jm} p_0(w_{id'})^{-1} u_{jm} \right).
\]

Note that since all topic proportions in document \( d' \) change when flipping one switch, unlike computing \( \Delta L(u_{j'n}) \), there is no computational benefit in using the expected value of the complete data log-likelihood. Thus, we evaluate the incomplete data log-likelihood for these updates.

6.3 Initialization

It is important to initialize the model sensibly to avoid poor local minima. Here, we use a simple but pragmatic two-stage initialization process. First to initialize each topic, we randomly select \( D_{init} \) documents. Only the words that occur in these documents are initially chosen as topic-specific, and their probabilities are initialized via frequency counts. The shared model, however, is estimated via global frequency counts \( \{1\} \) and kept fixed.

In the second stage, based on this initial model, we use a maximum likelihood decision rule to hard-assign each document in the corpus to a single topic. Finally, we repeat the same initialization process as in the first step, but now using this more “refined” set of documents for each topic.

6.4 Computational Complexity

For a fixed number of topics, \( M \), the computational complexity of the parameter learning part of our proposed algorithm is \( O(M_D DL_D) \) where \( M_D \) and \( L_D \) are the number of topics present in a document and the length of a document, respectively. Since in LDA all topics are potentially present in each document, its computational complexity is \( O(M M_D D L_D) \), which is higher than for our model. However, structure learning in our model imposes further complexity.

Updates of the word switches involve an iterative loop over all words in all topics. But the trial-flip of a single switch only requires scalar computation. Thus, the complexity of this part of the structure learning is \( O(M N) \).

Experimentally, we find that the heaviest part of our algorithm is updating the \( v \) switches, where at each step we trial-estimate topic proportions and evaluate the incomplete data log-likelihood for the current document under consideration. We need computations of order \( O(M D L_D) \) for trial-update of each switch. Thus the total computational complexity for this part, including the loop over all \( v \) switches, is \( O(M M_D D L_D) \).

Overall, computational complexity of our model, \( O(M M_D D L_D) \), is higher than LDA by a factor of \( M_D \). This higher complexity may be an issue for a corpus with a large number of active topics across documents. However, we note that we have also investigated complexity experimentally, via recorded execution times. Experimentally, we have found that our method and LDA require comparable execution times. (cf. Table 5).

6.5 Model Order Selection

We compare the BIC values of the estimated models at different orders and choose the one with the minimum BIC as the optimal order for our model, \( M^* \). Three different strategies are conceivable for learning at different model orders. First, somewhat naively, we could choose a set of models with “plausible” numbers of topics for our problem and learn each one of them starting from a separate initialization. Such an approach, however, will entail a large computational burden.

A more sensible approach is to learn a model at some order and use it for initializing new models at other orders. In doing so, we can sweep a range of model orders in either a bottom-up or top-down fashion. In the bottom-up approach, we initialize our model with a predefined low number of topics \( M_{min} \) and perform both the structure and parameter learning parts. Then for a model with one higher order, we keep the topics from the previously learned model and add a new topic, initializing it based on the set of documents with largest number of active topics. Intuitively, the documents with largest number of topics present in them may contain new content that is not fully covered by the current topics.

Alternatively in the top-down approach, we initialize our model with a specified “ceiling” number of topics \( M_{max} \) and reduce the number of topics by a predefined step. In this method we remove the least “plausible” topics from the existing model and apply the learning algorithm to minimize BIC at the now reduced model order, using the current set of model parameters as initialization. Here, as “least plausible”, we simply remove
the topics with the smallest aggregate mass across the document corpus. This procedure is applied repeatedly until a minimum number of topics is reached. Experimentally, this method has been found to be superior to the two other approaches, discovering the best set of topics with reasonable computational complexity. Thus, this is the approach we have taken in our experiments.

7 Experimental Results

In this section we report results of our model on the Ohsumed, Reuters-21578, and 20-Newsgroup corpora as well as a subset of the LabelMe image dataset. For each dataset, we compared performance against LDA [8] with respect to model fit (training and test log-likelihood), sparsity, and a class label consistency measure. We also compare against STC [13] with respect to class label consistency. C implementation of our model is available from http://www.personal.psu.edu/hus152/parstm.tar.gz.

Topic models can also be compared with respect to “quality” of extracted topics. From a computational linguistics standpoint, topics are expected to exhibit a coherent semantic theme rather than covering loosely related concepts. In this paper, we use the quantitative coherence criterion proposed in [28] to evaluate coherence of learned topics for LDA and our model on the text corpora. This measure has been shown to be in agreement with experts’ coherence judgments [28]. The topic coherence of topic j, which is based on word co-occurrence frequencies, is computed as:

\[
C(j; N^{(j)}) = \sum_{k=2}^{T} \sum_{l=1}^{k-1} \log \frac{S(n^{(j)}_k, n^{(j)}_l) + 1}{S(n^{(j)}_k) + S(n^{(j)}_l)},
\]

where \( N^{(j)} = (n^{(j)}_1, \ldots, n^{(j)}_T) \) is the list of the T most probable topic-specific words under topic j and \( S(n, n') \) is the number of documents in the corpus containing both words n and n'. Similarly, \( S(n) \) represents the number of documents with word n. Here, we use the top 2 percent of topic-specific words in each topic for our model to compute coherence. We use the same number of top words to compute coherence for LDA topics. In our experiments, we report the average coherence over all topics.

Results reported here are averaged based on 10 different initializations. For STC, we determined the hyperparameters by a validation set approach, working on the model with highest number of topics, and then kept them fixed for all other model orders. This makes STC complexity manageable and is also reasonable because we observe that the best STC accuracies are anyway achieved at the highest model order. The validation set was created by randomly choosing 20% of the documents in the training set. For inference on test documents, for our model, we allow all topics to be active in each document and only optimize topic proportions using \( \alpha \), given a fixed number of topics. We have chosen this approach rather than taking a transductive inference approach.

7.1 Ohsumed Corpus

The Ohsumed corpus [4] consists of 34389 documents, each assigned to one or multiple labels of the 23 MeSH diseases categories. Each document has on average 2.26 (std. dev. = 0.84) labels. The dataset was randomly divided into 24218 training and 10171 test documents. There are 12072 unique words in the corpus after applying standard stopword removal.

For all three methods, models were initialized with 150 topics and at each step the five least massive topics were removed. Fig. 1 shows the BIC curve for our model and the training and test log-likelihood of our model compared to LDA. The minimum of the BIC curve (i.e. estimated number of topics) is on average \( M^* = 105 \) (std. dev. = 5.98). The figure shows that LDA achieves higher log-likelihood on the training set for models with more than 120 topics but by controlling the likelihood-complexity trade-off, our model achieves higher log-likelihood on the test set at all model orders. We also note that LDA’s test likelihood peaks at \( \sim 140 \) topics (it will not be improved on Ohsumed by using even more topics).

We also used the class labels provided with the dataset to evaluate a class label consistency measure. We first associated to each topic a multinomial distribution on the class labels. We learned these label distributions for each topic by frequency counting over the ground-truth class labels of all documents, weighted by topic proportions:

\[
p_j(c) = \frac{\sum_{d=1}^{D} C_d^{(c,d)}}{\sum_{d=1}^{D} \sum_{j=1}^{C_d \alpha_j}} \alpha_j, \quad \forall j, c,
\]
where \( l_{id} \) is the \( i \)-th class label in document \( d \) and \( |C_d| \) and \( |C| \) are, respectively, the number of labels for document \( d \) and the total number of class labels. For labeling a test document, we then compute the probability of each class label based on the topic proportions in that document; i.e. \( \sum_{j=1}^{M} \alpha_{jd} p_j(c) = 1, ..., |C| \), and assign the labels that have probability higher than a threshold value \( \nu \):

\[
\hat{C}_d = \left\{ c : \sum_{j=1}^{M} \alpha_{jd} p_j(c) > \nu \right\}
\] (34)

As label consistency criteria, we measure precision and recall on the test set. Precision is the number of true discovered labels divided by the total number of ground-truth labels. Recall is the number of correctly classified labels divided by the total number of labels assigned to documents by our classifier. We measure these criteria for different threshold values \( \nu \) and report the area under the precision/recall curve (AUC) as the final measure of performance. For LDA we used the normalized Dirichlet variational parameters \( \gamma_j^{(d)} \) as topic \( j \)'s proportion for document \( d \). Similarly, for STC we used the normalized document codes \( \theta_d \) for unsupervised classification. For both LDA and STC we set \( \psi_{jd} = 1 \) \( \forall j, d \).

Fig. 2a shows the AUC for our model, LDA, and STC. Our model has consistently better label consistency over the entire range of model orders. Also, the highest AUC for our model occurs near \( M^* = 105 \), which is the minimum of the BIC curve.

Average coherence over all topics for our model and LDA are plotted in Fig. 2b. This figure shows that the average coherence in our model is higher than LDA for a wide range of model orders.

Table 1 compares sparsity in our model against LDA and STC at \( M = 105 \). As the measure for topic proportion sparsity, we report the average number of occurring topics in documents (\( \overline{M} \)). In STC and LDA, we considered those topics which dominantly contributed in generating at least one word in a document as “occurring topics”. We also report average number of topic-specific words per topic (\( \overline{N} \)), and the total number of topic-specific words over all topics (\( N_{\text{unique}} \)) for our model.

Average number of occurring topics in our model is 1.85 which compared to LDA and STC is much closer to the average number of labels per document in this corpus (2.26). This suggests our topics better resemble the ground-truth classes than those of LDA and STC. Also, the shared model is widely used in our model and only a relatively small number of words are specific for each topic. Average number of unique topic-specific words over all topics is much larger than \( \overline{N} \) which indicates that there is not great overlap between the set of topic-specific words across topics. In fact, each topic-specific word, on average, is salient in 8.10 topics (out of 105) and is modeled by the shared representation in other topics.

### 7.2 20-Newsgroups Corpus

In this section we report the results of our comparison on the 20-Newsgroup Corpus. This dataset consists of, respectively, 11293 and 7527 documents in the training and test sets, with 20 designated newsgroups. Each document is labeled to a single newsgroup. After standard stopword removal and stemming there are 54520 unique words in the corpus.

Models were initialized with 100 topics and two topics were removed at each elimination step. Fig. 3 shows the performance of our model relative to LDA. The minimum of the BIC curve is on average \( M^* = 36 \) (std. dev. = 3.79). Although LDA achieves higher likelihood on the training data, our parsimonious model has consistently better performance on the test set for all model orders.

Learning the label distributions of topics is done similar to the procedure explained for the Ohsumed corpus. But since documents in this corpus have single labels, we compute the probability over the class labels for each document, \( \sum_{j=1}^{M} \alpha_{jd} p_j(c) = 1, ..., |C| \), and assign only the label with highest probability. Class consistency on the test set is reported in Fig. 4a. We can see that the BIC-chosen model order \( M^* = 36 \) achieves good class consistency. Also, our model achieves better consistency compared to LDA and STC for models with less than 70 topics. Coherence of the discovered topics in LDA

---

7. http://people.csail.mit.edu/jrennie/20Newsgroups/
and our model is plotted in Fig. 4b. The curves are very similar.

Sparsity measures of our model on this dataset are compared with LDA and STC in Table 2. Only a small fraction of the unique words are used as topic-specific words in our model. Also, the average number of occurring topics in our model (1.14) is smaller than for LDA and STC. Note again that documents in this corpus are single-labeled. On average, the 22070 unique topic-specific words in this dataset are context-specific in only 2.67 topics, and shared in the others.

The top 10 words (with highest probabilities) from some of the topics discovered by LDA and by our model are reported in Table 3. For our model, we separately report the top 10 topic-specific and shared words. We can see that some high probability words in LDA topics (“write” and “article” in the first, and “system” in the third topic), are shared words under our model.

### 7.3 Reuters Corpus

We compared our model against LDA and STC on a subset of the Reuters dataset-21578 consisting of documents from 35 classes. There are respectively 6454 and 2513 documents in the training and test sets, each labeled with a single class. The documents include 16000 unique words after applying standard stopword removal and stemming.

We initialized the models with 100 topics and removed two topics at each elimination step. Average BIC for our model as well as the training and test data log-likelihood of our model and LDA are shown in Fig. 5. The minimum of the BIC curve is on average at $M^* = 40$ (std. dev. = 5.75). Fig. 5 shows that our model at $M = 40$ achieves higher log-likelihood on the test set compared to the LDA models at all orders. Moreover, this in spite of the fact that the test likelihood for our model increases modestly for orders above $M = 40$.

Class label consistency on this dataset is evaluated based on the procedure described for the 20-Newsgroup corpus. Average test set classification accuracy of our model, LDA, and STC for this dataset are shown in Fig. 6a. Our model has better performance than LDA and STC across all model orders. Also, the minimum of the BIC curve, $M^* = 40$, is consistent with high class label consistency.

We plotted the average coherence of topics for our model and LDA in Fig. 6b. This figure shows that for model orders less than 50, topics discovered by our model have on average higher coherence than LDA.

Table 4 reports the sparsity measures for our model, LDA, and STC. The topic-specific words in our model are a small subset of the dictionary size. Also, our model achieves sparser topic presence compared to LDA and STC.

Table 5 shows the top 10 words from three sample topics extracted by our model and LDA. For our model,
TABLE 3: Top 10 words for sample topics extracted from the 20-Newsgroup and the Reuters corpora by our method and LDA. “Specific” and “shared” denote, respectively, the top topic-specific and shared words of the topics in our model.

| Model     | 20-Newsgroup Topics                                                                 | Reuters Topics                                                                 |
|-----------|-------------------------------------------------------------------------------------|--------------------------------------------------------------------------------|
| LDA       | medic disease article write patent pitt bank health gordon doctor                    | gold mine oceon miner ton dome south year feet                                 |
| Specific  | doctor patient disease medic pain peopl pitt treatment gordon bank                   | gold mine oceon year feet pct reserv min dlz                                   |
| Shared    | write article don time work year good even problem thing                              | compani oper corp thi two unit includ expect report plan                        |
| LDA       | god christian jesu israel church peopl bibl christ come law                           | oil opec ga price energi dlz barrel mln petroleum crude                        |
| Specific  | godchristian exist peopl don atheist question thing mean reason                       | oil opec price bpd mln saudi barrel crude dlz product                          |
| Shared    | write article time good even apr come find sinc gener                                | stock last two march industri expect quarter per month report                   |
| LDA       | drive problem disk window system work hard run driver printer                        | dividend april record split stock march declar payabl set share                 |
| Specific  | printer font print window problem driver deskjet laser file artic                     | dividend stock share split april record compani common declar sharehold         |
| Shared    | write work good system want apr two question post program                             | corp offer unit quarter plan exchang invest acquir propos pai                  |

Fig. 6: Class label consistency and coherence on the Reuters corpus; (a) Average classification accuracy of our model (solid red), LDA (blue dashed), and STC (green dotted) as a function of number of topics. (b) Average coherence over all topics for our model (solid red) and LDA (blue dashed).

TABLE 4: Sparsity measure for the Reuters corpus; $M$: Average number of topic-specific words per topic; $N_{unique}$: Average number of unique topic-specific words over all topics; $N$: Average number of topics present in a document; $M=40, N=16000$, (std. dev.)

| Dataset       | $M$ (std. dev.) | $N$ (std. dev.) | $N_{unique}$ |
|---------------|----------------|----------------|--------------|
| Our Model     | 1.14 (0.01)    | 306.4 (4.31)   | 7149.8 (55.24) |
| LDA           | 4.97 (0.13)    | -              | -            |
| STC           | 3.01 (0.20)    | -              | -            |

we separately report the top 10 topic-specific and shared words.

7.4 LabelMe Image Dataset

In this section we report the results of our comparison on a subset of the LabelMe image dataset. The dataset which we downloaded[9] consists of 1600 images from 8 classes. Unique codewords were extracted from the images using the method described in [29, 30]. First, 128-dimensional SIFT vector descriptors [31] were generated by $5 \times 5$ sliding grids for each image in the training set. Then, K-means clustering was performed on the collection of SIFT descriptors, giving learned cluster centers. Finally, each SIFT descriptor in every image was assigned to its nearest cluster [32]. There are 158 clusters in this dataset after performing K-means clustering, merging close clusters, and pruning clusters with small number of members. The cluster index set $\{1, 2, \ldots, 158\}$ is effectively the dictionary of words. Each image is represented by a sequence of these cluster indices, of length 2401.

Since the number of codewords in this dataset is very small relative to the text corpora, the approximation used in [23] is not valid anymore. Accordingly, we considered any configuration with at most $(N_0 = NM)$ “on” switches, equally likely and used the exact form of (23). Thus, the cost of topic-specific words in BIC is $NM \log(2)$ in this case.

We initialized the models with 80 topics and reduced the number of topics by one at each step. Fig. 7 shows performance of our model compared with LDA. The minimum of BIC is on average at $M^* = 46$ (std. dev. $= 3.4$). Test set log-likelihood is higher in our model compared to LDA, across all model orders.

We also performed single-label classification, similar to the procedure described for the single-labeled text corpora. Fig. 7 shows the class label consistency of our model compared to LDA and STC. We can see that by taking a parsimonious approach for both topic proportions and word probabilities, our model achieves better label consistency than LDA and STC.

9. http://www.cs.cmu.edu/~chongw/slda/
TABLE 5: Sparsity measure for the LabelMe dataset; \( N \): Average number of topic-specific words per topic; \( N_{\text{unigue}} \): Average number of unique topic-specific words over all topics; \( M \): Average number of topics present in a document; \( M=46, N=158 \), (std. dev.)

| Dataset  | \( M \) | \( N \) | \( N_{\text{unigue}} \) |
|----------|--------|--------|-------------------------|
| Our Model| 7.4 (0.14) | 95.73 (2.31) | 158 (0) |
| LDA      | 37.15 (1.03) | - | - |
| STC      | 1.08 (0.04) | - | - |

Table 5 compares the sparsity measure in our model with LDA and STC for this dataset at \( M^* = 46 \). At this model order, the average number of occurring topics for LDA is 37.15, which seems implausible, but our model and STC provide more reasonable sparsity in topic proportions. However, unlike STC, our model exhibits parsimony in word probabilities, with on average only 95.73 words modeled in a topic-specific fashion.

**Discussion:** On all four data sets, our models achieve higher test likelihoods than LDA, at all model orders. Even stronger, our model evaluated at the single (BIC minimum) order achieves higher test likelihood than the LDA models evaluated at all orders. With respect to class label consistency, our BIC-minimum model gives accuracy better than LDA evaluated at nearly all orders. However, it is seen on all the data sets that LDA’s class consistency improves with increasing order, even exceeding our model’s on 20-Newsgroups for \( M > 80 \). In general, our model has much greater topic sparsity than LDA. Moreover, our model’s average number of topics per document has good agreement with the average number of class labels per document. In this sense, our learned topics better correspond to individual classes than LDA’s. From the curves in Figs. 2, 4, and 6, it may be possible for LDA to achieve greater class label consistency than our model on some of the data sets (not LabelMe) by using a very large number of topics. Two points weigh against such choice. First, at least on Ohsumed, such choice would cause overfitting (decrease in test likelihood). Second, one must recognize that topic models perform unsupervised clustering, aiming to capture the most salient content and to provide human-interpretable data groupings. Choosing a huge number of topics may improve LDA’s class label consistency, but such solutions would defy human interpretation. Our BIC-minimizing solutions, on the other hand, are more interpretable due to their topic sparsity, their word sparsity and, as noted, due to topics better corresponding to ground-truth classes. Moreover, this parsimony and interpretability are achieved without giving up performance (test set likelihood and class consistency).

**Execution Time:** We report the execution time of our model, LDA, and STC in minutes (std. dev.).

| Dataset   | Our Model | LDA | STC |
|-----------|-----------|-----|-----|
| Ohsumed   | 2529 (97) | 818 (231) | 243 (11) |
| 20-Newsgroup | 727 (46) | 1514 (94) | 522 (110) |
| Reuters   | 395 (31)  | 746 (52)  | 140 (9)   |
| LabelMe   | 209 (7)   | 124 (6)   | 280 (30)  |

\( M = 1.85 \) and larger typical document lengths, running time in our model is three times that of LDA. Also, STC has the smallest overall running time across these datasets.

8 Conclusion

We have proposed a parsimonious model well-suited for estimating topics and their salient words, given a database of unstructured documents. Unlike models such as LDA, our model has a sparse representation both in topic presence in documents and in word probabilities under different topics. For each topic, our method selects a set of salient words that have probability parameters specific to that topic, with other words explained by a universal shared model. Furthermore, we determine a sparse set of topics occurring in each document with nonzero topic proportions. We have derived a BIC objective function specific to our model which balances goodness of fit and model complexity. We minimize this objective to jointly determine the set of topic-specific words and active topics in each document, to estimate the parameters of the model, and to estimate the total number of topics present in the corpus in a wholly unsupervised fashion. Experimental results show that our proposed model outperforms LDA and a sparsity-based topic model [13] with respect to several clustering performance measures, achieving both higher test set log-likelihood and better agreement with ground-truth class labels.

REFERENCES

[1] W. H. Jefferys and J. O. Berger, “Ockham’s razor and bayesian analysis,” American Scientist, vol. 80, no. 1, pp. 64–72, 1992.
[2] D. J. C. MacKay, Information Theory, Inference and Learning Algorithms. Cambridge University Press, 2003.
[3] C. D. Manning and H. Schütze, Foundations of Statistical Natural Language Processing. MIT Press, 1999.
[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” JMLR, vol. 3, pp. 993–1022, 2003.
[5] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell, “Text classification from labeled and unlabeled documents using EM,” Machine learning, vol. 39, no. 2-3, pp. 103–134, 2000.
[6] R. E. Kass and A. E. Raftery, “Bayes Factors,” Journal of the American Statistical Association, vol. 90, no. 430, pp. 773–795, 1995.
[7] C. M. Bishop, Pattern Recognition and Machine Learning. Springer, 2006.
[8] G. Schwarz, “Estimating the dimension of a model,” Annals of Statistics, vol. 6, no. 2, pp. 461–464, 1978.
[9] H. M. Wallach, D. M. Mimno, and A. McCallum, “Rethinking LDA: Why Priors Matter.” NIPS, pp. 973–981, 2009.
[10] C. Wang and D. M. Blei, “Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process,” in NIPS, 2009, pp. 1982–1989.
[11] S. Williamson, C. Wang, K. Heller, and D. M. Blei, “The IBP compound Dirichlet process and its application to focused topic modeling,” in ICML, 2010.

[12] C. Chemudugunta and S. M. S. Steyvers, “Modeling general and specific aspects of documents with a probabilistic topic model,” in NIPS, 2007, pp. 241–248.

[13] J. Zhu and E. P. Xing, “Sparse topical coding,” in 27th Conference on Uncertainty in Artificial Intelligence (UAI), 2011.

[14] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, “Hierarchical Dirichlet Processes,” Journal of the American Statistical Association, vol. 101, no. 476, pp. 1566–1581, Dec. 2006.

[15] M. H. C. Law, M. A. T. Figueiredo, and A. K. Jain, “Simultaneous feature selection and clustering using mixture models.” IEEE Trans. Pattern Anal. Mach. Intell., vol. 26, no. 9, pp. 1154–66, Sep. 2004.

[16] C. Constantinopoulos, M. K. Titsias, and A. Likas, “Bayesian feature and model selection for Gaussian mixture models.” IEEE Trans. Pattern Anal. Mach. Intell., vol. 28, no. 6, pp. 1013–8, Jun. 2006.

[17] M. W. Graham and D. J. Miller, “Unsupervised learning of parsimonious mixtures on large spaces with integrated feature and component selection,” IEEE Trans. Signal Process., vol. 54, no. 4, pp. 1289–1303, 2006.

[18] P. D. Grünwald, The minimum description length principle. MIT Press, 2007.

[19] S. Boutemedjet, N. Bouguila, and D. Ziou, “A hybrid feature extraction selection approach for high-dimensional non-Gaussian data clustering.” IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 8, pp. 1429–43, Aug. 2009.

[20] W. Fan, N. Bouguila, and D. Ziou, “Selection for high-dimensional non-Gaussian data clustering with variational inference.” IEEE Trans. Knowl. Data Eng., vol. 25, no. 7, pp. 1670–1685, 2013.

[21] T. L. Griffiths and M. Steyvers, “Finding scientific topics,” Proceedings of the National Academy of Sciences of the United States of America, vol. 101 Suppl, pp. 5228–35, Apr. 2004.

[22] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” Journal of the Royal Statistical Society., vol. 39, no. 1, pp. 1–38, 1977.

[23] J. K. Ghosh, M. Delampady, and T. Samanta, An Introduction to Bayesian Analysis: Theory and Methods. Springer New York, 2006.

[24] R. L. Graham, D. E. Knuth, and O. Patashnik, Concrete Mathematics: A Foundation for Computer Science, 2nd ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., Jan. 1994.

[25] T. M. Cover and J. A. Thomas, Elements of Information Theory, 2nd ed. Wiley, 2006.

[26] X.-L. Meng and D. Van Dyk, “The EM algorithm—an old folk-song sung to a fast new tune,” Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 59, no. 3, pp. 511–567, 1997.

[27] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, “LabelMe: a database and web-based tool for image annotation,” International Journal of Computer Vision, vol. 77, no. 1-3, pp. 157–173, 2008.

[28] D. Mimno, H. M. Wallach, and E. Talley, “Optimizing semantic coherence in topic models,” in Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, no. 2, 2011, pp. 262–272.

[29] L. Fei-Fei and P. Perona, “A Bayesian hierarchical model for learning natural scene categories,” in IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05), vol. 2, 2005, pp. 524–531.

[30] C. Wang, D. M. Blei, and F.-F. Li, “Simultaneous image classification and annotation,” in 2009 IEEE Conference on Computer Vision and Pattern Recognition, Jun. 2009, pp. 1903–1910.

[31] D. G. Lowe, “Object recognition from local scale-invariant features,” in IEEE International Conference on Computer Vision, 1999, pp. 1150–1157.

[32] T. Leung and J. Malik, “Representing and recognizing the visual appearance of materials using three-dimensional textons,” International Journal of Computer Vision, vol. 43, no. 1, pp. 29–44, 2001.