Finding Better Topics: Features, Priors and Constraints

Xiaona Wu, Jia Zeng, Jianfeng Yan, and Xiaosheng Liu

School of Computer Science and Technology, Soochow University,
Suzhou 215006, China
zengja@gmail.com

Abstract. Latent Dirichlet allocation (LDA) is a popular probabilistic topic modeling paradigm. In practice, LDA users usually face two problems. First, the common and stop words tend to occupy all topics leading to bad topic interpretability. Second, there is little guidance on how to improve the low-dimensional topic features for a better clustering or classification performance. To find better topics, we re-examine LDA from three perspectives: continuous features, asymmetric Dirichlet priors and sparseness constraints, using variants of belief propagation (BP) inference algorithms. We show that continuous features can remove the common and stop words from topics effectively. Asymmetric Dirichlet priors have substantial advantages over symmetric priors. Sparseness constraints do not improve the overall performance very much.

Keywords: Latent Dirichlet allocation, belief propagation, continuous features, asymmetric Dirichlet priors, sparseness constraints.

1 Introduction

Latent Dirichlet allocation (LDA) [1] is a widely-used probabilistic topic modeling paradigm, which has found many important applications in natural language processing and computer vision areas. LDA represents documents as mixtures over latent topics, where each topic is a distribution over a fixed vocabulary. Using approximate inference techniques like variational Bayes (VB) [1], Gibbs sampling (GS) [2] or belief propagation (BP) [3], LDA automatically learns the topic-word and document-topic distributions from a large collection of documents. In practice, LDA users usually encounter two problems. First, the common and stop words tend to occupy all topics. For example, if we use LDA to extract topics from a machine learning corpus like NIPS, we find that the common words "learning" and "model" dominate (having very high likelihood) almost all topic-word distributions. This phenomenon makes the interpretability of topics undesirable [4]. Second, there is relatively little guidance on how to improve the lower-dimensional topic features for a better retrieval, clustering and classification performance. Therefore, we explore LDA from three perspectives: continuous features, asymmetric Dirichlet priors and sparseness constraints to find better topics.

V.S. Tseng et al. (Eds.): PAKDD 2014, Part II, LNAI 8444, pp. 296–310, 2014.

LDA has long been used for discrete features such as word tokens and counts. Continuous features or term weighting schemes have been rarely discussed such as term frequency-inverse document frequency (TF-IDF) [5] and LTC [6]. One major concern is that LDA cannot generate continuous observations in its probabilistic modeling process. So, in practice users have to manually remove stop words having little contribution to the meaning of the text [7]. But, removing common words requires contextual knowledge of the entire corpus, which is often a big challenge to users without prior knowledge. Recently, continuous features for LDA have gained intensive research interests. A simple term-frequency feature scheme [8] has been used for tagged document within the framework of LDA. Point-wise mutual information (PMI) features [9] have been incorporated into the GS inference algorithm referred to as pmiGS. The PMI feature gives common and stop words some lower weights. Then, pmiGS infers topic-word distributions from weighted word counts. The results show that the PMI feature not only lowers the likelihood of common and stop words in the topic-word distribution, but also gains a no-trivial improvement in cross-language retrieval tasks. This line of research inspires us to consider continuous features for LDA to improve the topic interpretability.

Most LDA algorithms [2,3,7] consider fixed symmetric Dirichlet priors over document-topic and topic-word distributions for simplicity. Although it is possible to automatically learn Dirichlet hyperparameters from training data according to the maxumum-likelihood criterion [10], the extensive empirical studies [11] confirm that the inferred symmetric priors do not significantly improve the topic modeling performance than the fixed ones. However, asymmetric Dirichlet priors over document-topic and symmetric Dirichlet priors over topic-word distributions have substantial advantages on removing the common words and choosing the number of topics [12]. The asymmetric prior over document-topic distribution can guide common or stop words to be grouped into a few topics with higher likelihoods because these words often occupy the larger proportion of each document. So, asymmetric priors are also effective in finding better topics.

If we can control the sparseness of document-topic and topic-word distributions, we can possibly control the quality and interpretability of lower-dimensional topic features. Sparse topic coding (STC) [13] can directly control the sparsity of the inferred representations by relaxing the normalization constraint, which can be integrated with any convex loss function. STC identifies sparse topic meanings of words and improves time efficiency and classification accuracy. Also, sparse coding can be directly combined with LDA's extensions [14] for computer vision applications. In sparse coding, each document or word only has a few salient topical meanings or senses. Sparse distributions carry salient information for a better interpretability, so that the low-dimensional sparse topic features may be more distinguishable. Therefore, we will consider adding sparse constrains [15] on LDA's document-topic and topic-word distributions.

Although continuous features, asymmetric priors and sparseness constraints for LDA have been studied either by GS [2] or by VB [1] inference algorithms, we re-examine these three perspectives within the novel BP inference framework [3], which is very competitive in both speed and accuracy. As a result, we incoporate continuous features, asymmetric Dirichlet priors and sparseness constraints into BP algorithms to find better topics than traditional GS and VB algorithms. Besides, most of previous studies focus only on one of three aspects, and lack a comprehensive comparison in terms of generalization performance, document clustering/classification and topic interpretability. Here, we compare these three aspects on different data sets, and provide evidence on which one can produce high-quality topics.

2 Background

We begin by reviewing batch BP algorithms for learning collapsed LDA [3,16,17]. The probabilistic topic modeling task can be interpreted as a labeling problem, in which the objective is to assign a set of thematic topic labels, $\mathbf{z}_{W \times D} = \{z^k_{w,d}\}$, to explain the observed elements in document-word matrix, $\mathbf{x}_{W \times D} = \{x_{w,d}\}$. The notations $1 \le w \le W$ and $1 \le d \le D$ are the word index in vocabulary and the document index in corpus. The notation $1 \le k \le K$ is the topic index. The nonzero element $x_{w,d} \neq 0$ denotes the number of word counts at the index $\{w,d\}$. For each word token $x_{w,d,i} = \{0,1\}, 1 \le i \le x_{w,d}$, there is a topic label $z^k_{w,d,i} = \{0,1\}, \sum_{k=1}^K z^k_{w,d,i} = 1, 1 \le i \le x_{w,d}$, so that the soft topic label for the word index $\{w,d\}$ is $z^k_{w,d} = \sum_{i=1}^{x_{w,d}} z^k_{w,d,i}/x_{w,d}$.

The collapsed LDA [18] has joint probability $p(\mathbf{x}, \mathbf{z}|\alpha v_k, \beta u_w)$, where the Dirichlet hyperparameters $\{\alpha v_k, \beta u_w\}, \sum_k v_k = 1, \sum_w u_w = 1, \alpha, \beta > 0$. In practice, we may use the fixed symmetric hyperparameters $\{v_k = 1/K, u_w = 1/W\}$ and the concentration parameters $\{\alpha, \beta\}$ are provided by users for simplicity [2]. To maximize the joint probability in terms of $\mathbf{z}$, the BP algorithm [3] computes the posterior probability, $\mu_{w,d}(k) = p(z^k_{w,d,i} = 1|\mathbf{z}^k_{-(w,d,i)}, \mathbf{x})$, called *message*, which can be normalized by local computation, i.e., $\sum_{k=1}^K \mu_{w,d}(k) = 1$. The approximate message update equation is

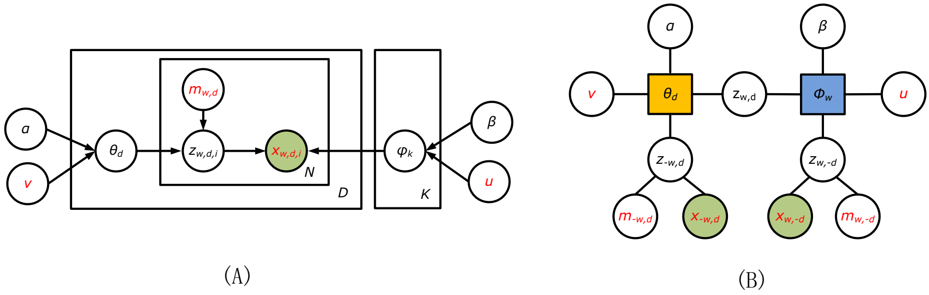
$$\mu_{w,d}(k) \propto \frac{[\hat{\theta}_{-w,d}(k) + \alpha v_k] \times [\hat{\phi}_{w,-d}(k) + \beta u_w]}{[\sum_w x_{w,d} + \alpha] \times [\hat{\phi}_{-(w,d)}(k) + \beta]}, \tag{1}$$

where the *sufficient statistics* for LDA model are

$$\hat{\theta}_{-w,d}(k) = \sum_{-w} x_{w,d}\mu_{w,d}(k), \tag{2}$$

$$\hat{\phi}_{w,-d}(k) = \sum_{-d} x_{w,d}\mu_{w,d}(k), \tag{3}$$

where $-w$ and $-d$ denote all word indices except w and all document indices except d. Obviously, the message update equation (1) depends on all other neighboring messages $\boldsymbol{\mu}_{-(w,d)}$ excluding the current message $\mu_{w,d}$. Two multinomial parameters, the document-topic distribution θ and the topic-word distribution



(A)

(B)

Fig. 1. (A)Generative graphical representation of LDA based on continuous features, asymmetric Dirichlet priors and sparseness constraints, (B)Factor graph and message passing

ϕ, can be calculated from *sufficient statistics* $\hat{\theta}_d(k)$ and $\hat{\phi}_w(k)$ by normalization. Message passing process will iterate Eqs. (1), (2) and (3) until all messages converge to a local stationary point [3].

As mentioned in Section 1, LDA users often use the document-topic distribution in (2) as the lower-dimensional features for document retrieval, clustering and classification. The word-topic distribution in (3) is used to find the hot words in each topic. Usually, users will inspect the hot words with higher likelihood in each topic to understand the topic's semantic meaning. Observing (2) and (3), we find that these two distributions are determined by three factors:

1. The features or observations: the word counts $x_{w,d}$.
2. The Dirichlet priors or hyperparameters: the base vectors $\{v_k, u_w\}$ and the concentration parameters $\{\alpha, \beta\}$ in Eq. (1).
3. The message: the K-tuple vector $\mu_{w,d}(k)$ for the topic likelihood at index $\{w, d\}$.

In this paper, we will regulate these three factors to find better topics including document-topic (2) and topic-word distributions (3).

3 Finding Better Topics

The major reason that the common and stop words occupy almost all topics is that LDA uses word counts as features. The bigger the word counts, the higher the influence to the topic distributions. In Eqs. (2) and (3), the normalized message $\mu_{w,d}(k)$ is multiplied by the nonzero word count $x_{w,d}$. Thus, $x_{w,d}$ can be regarded as the weight of $\mu_{w,d}(k)$ in estimating document-topic and topic-word distributions. In this way, the topics may be dominated by those high-frequent common and stop words. We see that the bigger word count $x_{w,d}$ corresponds to the greater influence of the estimated distributions in (2) and (3). This phenomenon motivates us to use the continuous features such as TF-IDF or LTC to lower the weights of common and stop words during message passing.

As far as Dirichlet priors are concerned, if we use the symmetric priors $\{v_k = 1/K, u_w = 1/W\}$, the common and stop words have equal likelihoods to be assigned to all topics in Eq. (1). However, if we use the asymmetric priors, words will have higher likelihood to be assigned to the topic with higher priors. In this way, most common and stop words may be assigned to a few topic groups with higher priors [12]. This phenomenon motivates us to incorporate the asymmetric Dirichlet prior learning into the message passing process (1), (2) and (3).

The message $\mu_{w,d}(k)$ represents the topic likelihood for each word token $x_{w,d,i}$. If the message is not sparse, the word token may have multiple topic meanings leading to unclear explanations. So, we encourage passing those sparse messages by adding a weight proportional to the sparseness of the message. This weighted message passing strategy can strengthen the sparseness of document-topic and topic-word distributions in (2) and (3). According to [13] and [14], the sparseness will make the lower-dimensional topic features more distinguishable for clustering or classification purposes. This motivates us to add sparseness constraints on messages during their passing process.

Fig. 1(A) shows the continuous features, asymmetric Dirichlet priors and sparseness constraints denoted by red colors in the generative graphical representation of LDA. The asymmetric Dirichlet priors are divided into the connection parameters $\{\alpha, \beta\}$ and the base measure vectors $\{\mathbf{v}, \mathbf{u}\}$, and $m_{w,d}$ is the sparseness constraints for the message $\mu_{w,d}(k) \sim z^k_{w,d}$. Note that if $x_{w,d} = \sum_i x_{w,d,i}$ becomes continuous observations like TF-IDF, the generative model in Fig. 1(A) cannot generate such observations. However, the factor graph representation of the collapsed LDA [3] shows that it is possible to describe the continuous features using the undirected factor graph, which does not need to encode the generative relations between variables. In this way, we may think that the factor graph is a close approximation to LDA [3]. Fig. 1(B) shows the factor graph representation and the message passing process based on continuous features, asymmetric Dirichlet priors and sparseness constraints. We see that the message $\mu_{w,d}(k) \sim z^k_{w,d}$ can be inferred by its neighboring messages including $\{(\mathbf{x}_{-w,d}, \mathbf{z}^k_{-w,d}, \mathbf{m}_{-w,d}), \alpha v_k\}$ and $\{(\mathbf{x}_{w,-d}, \mathbf{z}^k_{w,-d}, \mathbf{m}_{w,-d}), \beta u_w\}$ via factor nodes θ_d and ϕ_w, respectively. We group the variables $(\mathbf{x}_{w,d}, \mathbf{z}^k_{w,d}, \mathbf{m}_{w,d})$ together because they work together to influence the neighboring messages according to (1). From the message passing over factor graphs, we can derive the similar message update equation to (1) that considers continuous features, asymmetric priors and sparseness constraints within the unified BP framework.

3.1 Continuous Features

In linguistics, the high frequent stop words like "the, and, of" which occur in most of the documents do not contribute to the topic formation. To avoid stop words dominating every topic, we have to remove stop words before running LDA according to a corpus-specific stop word list. However, even if the stop words have been removed, there still are many common words such as "model, learning, data" in the machine learning corpus. In such cases, we may use the continuous features such as TF-IDF [5] and LTC [6] that give the lower weights to the "common word" messages in (1). Let $x_{w,d}/\sum_w x_{w,d}$ be the frequency of word w in document d, and $\sum_d x_{w,d}$ be the total number of times that the word w occurs in all documents. We get the continuous TF-IDF feature as

$$x_{w,d}^{tfidf} = \frac{x_{w,d}}{\sum_w x_{w,d}} \times \log\left(\frac{D}{\sum_d x_{w,d}}\right), \tag{4}$$

and the LTC feature as

$$x_{w,d}^{ltc} = \frac{\log(\frac{x_{w,d}}{\sum_w x_{w,d}}+1) \times \log\left(\frac{D}{\sum_d x_{w,d}}\right)}{\sqrt{\sum_{d=1}^{D}\left[\log(\frac{x_{w,d}}{\sum_w x_{w,d}}+1) \times \log\left(\frac{D}{\sum_d x_{w,d}}\right)\right]^2}}. \tag{5}$$

The difference between (5) and (4) is that (5) uses the logarithm of word frequency and is normalized by the geometric mean of the numerator. This normalization makes LTC features more distinguishable than TF-IDF features.

We simply replace the discrete word count feature $x_{w,d}$ by the continuous features $x_{w,d}^{tfidf}$ and $x_{w,d}^{ltc}$ in Eqs. (2), (3) and (1). Without loss of generality, we focus on LTC features for topic modeling. We refer to the message passing algorithms for LTC feature as ltcBP. Obviously in (2) and (3), the higher TF-IDF and LTC values will have the bigger influence to the topic formation. Generally, the stop and common words have lower TF-IDF and LTC weights, so that they will be automatically removed from hot word list in each topic during the message passing process.

3.2 Asymmetric Priors

There are several approaches to learn Dirichlet priors from training data. Here, we choose to place Gamma priors on the hyperparameters $\alpha \sim G[C, S]$, where C and S are shape and scale parameters of Gamma distribution. Generally, these parameters are fixed by users during learning Dirichlet priors. We adopt the improved method of Minka's fixed point iteration [10,12]. However, this method is based on discrete counts on topic labels rather than messages in BP (1). To solve this problem, we sample the topic label $z_{w,d,i}^k$ for each word token $x_{w,d,i}$ from the conditional probability $\mu_{w,d}(k)$. From the sampled $[z_{w,d,i}^k = 1]$, we get two topic count matrices

$$\gamma_d(k) = \sum_{w=1}^{W}\sum_{i=1}^{x_{w,d}}[z_{w,d,i}^k = 1], \tag{6}$$

$$\eta_w(k) = \sum_{d=1}^{D}\sum_{i=1}^{x_{w,d}}[z_{w,d,i}^k = 1]. \tag{7}$$

```
input : $\mathbf{x_{W \times D}}, \mathbf{K}, \mathbf{T}, \alpha\mathbf{v}, \beta\mathbf{u}, \mathbf{C}, \mathbf{S}$.
output: $\theta_d$,$\phi_w$.
$\mu^1_{w,d}(k) \longleftarrow$ initialization and normalization;
$\hat{\theta}^1_{-w,d}(k) \leftarrow \sum_{-w} x_{w,d}\mu^1_{w,d}(k)$;
$\hat{\phi}^1_{w,-d}(k) \leftarrow \sum_{-d} x_{w,d}\mu^1_{w,d}(k)$;
$\alpha \leftarrow 50, v_k \leftarrow 50/K, \beta u_w \leftarrow 0.01, C \leftarrow 1.001, S \leftarrow 1$;
for $t \leftarrow 1$ to $T$ do
    $\mu^{t+1}_{w,d}(k) \propto \frac{[\hat{\theta}^t_{-w,d}(k)+\alpha v^t_k]\times[\hat{\phi}^t_{w,-d}(k)+\beta u^t_w]}{[\sum_w x_{w,d}+\alpha^t]\times[\hat{\phi}^t_{-(w,d)}(k)+\beta^t]}$;
    $\hat{\theta}^{t+1}_{-w,d}(k) \leftarrow \sum_{-w} x_{w,d}\mu^{t+1}_{w,d}(k)$;
    $\hat{\phi}^{t+1}_{w,-d}(k) \leftarrow \sum_{-d} x_{w,d}\mu^{t+1}_{w,d}(k)$;
    sampling $z$ from $\mu^{t+1}_{w,d}(k)$;
    $\gamma_d(k) \leftarrow \sum_{w=1}^{W} \sum_{i=1}^{x_{w,d}} z^k_{w,d,i}$;
    $\eta_w(k) \leftarrow \sum_{d=1}^{D} \sum_{i=1}^{x_{w,d}} z^k_{w,d,i}$;
    $\alpha v^{t+1}_k \leftarrow \alpha v^t_k \frac{\sum_{n=1}^{b1} I_{n,k} \sum_{f=1}^{n} \frac{1}{f-1+\alpha v^t_k} + C}{\sum_{n=1}^{b2} I_n \sum_{f=1}^{n} \frac{1}{f-1+\alpha^t} - \frac{1}{S}}$;
    using $\eta_w(k)$ to learn symmetric $\beta^{t+1}$;
    $\beta u^{t+1}_w \leftarrow \beta^{t+1}/W$;
end
$\theta_d(k) \leftarrow \frac{\hat{\theta}_d(k)+\alpha v_k}{\sum_k \hat{\theta}_d(k)+\alpha}$; $\phi_w(k) \leftarrow \frac{\hat{\phi}_w(k)+\beta u_w}{\sum_w \hat{\phi}_w(k)+\beta}$.
```

Fig. 2. The asBP algorithm for LDA

Based on these two count matrices, we can directly use the Minka's fixed point iteration

$$\alpha v_k \leftarrow \alpha v_k \frac{\sum_{n=1}^{b1} I_{n,k} \sum_{f=1}^{n} \frac{1}{f-1+\alpha v_k} + C}{\sum_{n=1}^{b2} I_n \sum_{f=1}^{n} \frac{1}{f-1+\alpha} - \frac{1}{S}}, \tag{8}$$

where

$$I_{n,k} = \sum_{d=1}^{D} \delta(\gamma_d(k) - n), \tag{9}$$

$$I_n = \sum_{d=1}^{D} \delta(len(d) - n), \tag{10}$$

where $b1 = \max_d \gamma_d(k), b2 = \max_d len(d)$, and $len(d)$ is the total number of observations in document d, and n and f are positive integers. The value αv_k acts as an initial set for the topic k in all documents. $I_{n,k}$ is the number of documents in which the topic k has been seen exactly n times. I_n is the number of documents that contain a total of n observations. $I_n(\cdot) = \sum_{k=1}^{K} I_{n,k}$ is the total number of documents whose topics $(1, \ldots, K)$ has been seen exactly n times. For the symmetric Dirichlet priors, the base measure is fixed as $v_k = 1/K$ and the concentration parameter α is updated as

$$\alpha v_k \leftarrow \frac{\alpha}{K} \times \frac{\sum_{n=1}^{b3} I_n(\cdot) \sum_{f=1}^{n} \frac{1}{f-1+\alpha/K}}{\sum_{n=1}^{b2} I_n \sum_{f=1}^{n} \frac{1}{f-1+\alpha}}, \tag{11}$$

where $b3 = max_{d,k}\gamma_d(k)$. It is the same way to learn asymmetric or symmetric βu_w according to the count matrix $\eta_w(k)$.

Symmetric and asymmetric Dirichlet priors over $\{\theta, \phi\}$ play different roles in topic modeling. Similar to [12], we implement an asymmetric prior over θ and a symmetric prior over ϕ, which is referred to as the asBP algorithm. In practice, this implementation performs the best than other combinations of priors [12]. Fig. 2 summaries the asBP algorithm for learning LDA, where T is the total number of learning iterations. The asymmetric prior αv_k can be learned by Eqs. (9), (10), (8). At the first $t \le 100$ iterations, asBP is the same with the batch BP which updates and normalizes all messages for all topics. For $t > 100$, we learn the asymmetric prior αv_k and the symmetric prior βu_w every 20 iterations.

3.3 Sparseness Constraints

In addition to the continuous features and asymmetric Dirichlet priors, sparseness constraints over messages also has an effect on the topic interpretability. In this paper, we adopt a sparseness measure based on the L_1 norm and the L_2 norm [15],

$$m_{w,d} = \frac{\sqrt{K} - (\sum_k |\mu_{w,d}(k)|)/\sqrt{\sum_k [\mu_{w,d}(k)]^2}}{\sqrt{K} - 1}, \qquad (12)$$

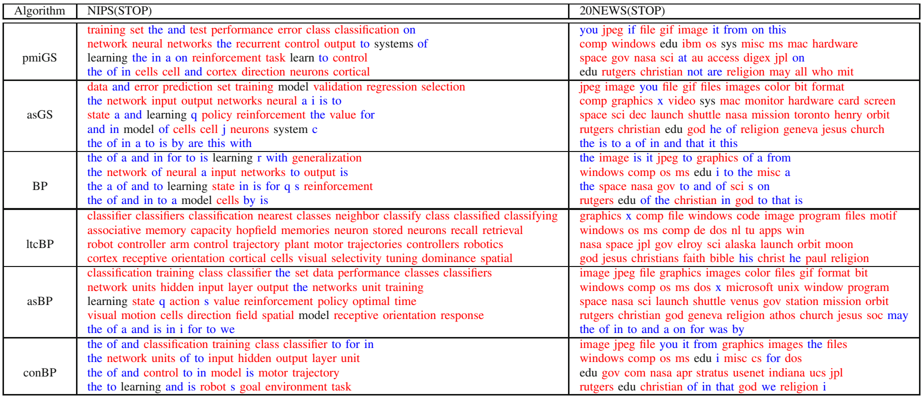
where K is the number of topics and the dimensionality of $\mu_{w,d}(k)$. The quantity $m_{w,d}$ is the sparseness of $\mu_{w,d}$. Usually, the messages of stop and common words have relatively lower sparseness because they often occupy many topics for a lower interpretability. For example, when the number of topics is 10 in CORA data set, the meaningful words such as "reinforcement", "Bayesian" have relatively higher sparseness values 0.9999 and 0.9615 than 0.8663 and 0.8417 of the common words such as "learning" and "model". Our intuition is that we need to encourage passing those messages with higher sparseness values, so we use the sparseness value (12) as the weight of message during message update (1). More specifically, we simply use the weighted sum $m_{w,d}x_{w,d}\mu_{w,d}(k)$ in Eqs. (2), (3) and (1). Such a weighted message passing strategy will encourage sparse messages with higher weights in topic formation. We refer this message passing algorithm as conBP. If all sparseness constraints $m_{w,d} = 1$, conBP will become the standard BP algorithm for learning LDA [3].

4 Experiments

In this section, we evaluate the effectiveness of the proposed ltcBP, asBP, and conBP algorithms on six publicly available data sets. Table 1 summarizes the statistics of six data sets, where D is the total number of documents, $\overline{N}_d$ the average document length, N the total number of tokens, W the vocabulary size, and "stop" indicates whether there are stop words. All algorithms are evaluated by five performance metrics. Lower perplexity [3, 11] indicates better generalization performance. The lower-dimensional document-topic distributions can be fed into standard SVM classifiers for document classification. The higher classification accuracy implies the more distinguishable ability of the lower-dimensional

Table 1. Data set statistics

Data sets	D	$\overline{N}_d$	N	W	$STOP$
CORA	2410	57	136394	2961	no
WEK	2785	127	352647	7061	no
NIPS	1740	1323	2301375	13649	no
20NEWS	2000	200	399669	36863	no
NIPS (STOP)	1740	2939	5114634	70629	yes
20NEWS (STOP)	2000	372	743180	37370	yes



Algorithm	NIPS(STOP)	20NEWS(STOP)
pmiGS	training set the and test performance error class classification on network neural networks the recurrent control output to systems of learning the in a on reinforcement task learn to control the of in cells cell and cortex direction neurons cortical	you jpeg if file gif image it from on this comp windows edu ibm os sys misc ms mac hardware space gov nasa sci at au access digex jpl on edu rutgers christian not are religion may all who mit
asGS	data and error prediction set training model validation regression selection the network input output networks neural a i is to state a and learning q policy reinforcement the value for and in model of cells cell j neurons system c the of in a to is by are this with	jpeg image you file gif files images color bit format comp graphics x video sys mac monitor hardware card screen space sci dec launch shuttle nasa mission toronto henry orbit rutgers christian edu god he of religion geneva jesus church the is to a of in and that it this
BP	the of a and in for to is learning r with generalization the network of neural a input networks to output is the a of and to learning state in is for q s reinforcement the of and in to a model cells by is	the image is it jpeg to graphics of a from windows comp os ms edu i to the misc a the space nasa gov to and of sci s on rutgers edu of the christian in god to that is
ltcBP	classifier classifiers classification nearest classes neighbor classify class classified classifying associative memory capacity hopfield memories neuron stored neurons recall retrieval robot controller arm control trajectory plant motor trajectories controllers robotics cortex receptive orientation cortical cells visual selectivity tuning dominance spatial	graphics x comp file windows code image program files motif windows os ms comp de dos nl tu apps win nasa space jpl gov elroy sci alaska launch orbit moon god jesus christians faith bible his christ he paul religion
asBP	classification training class classifier the set data performance classes classifiers network units hidden input layer output the networks unit training learning state q action s value reinforcement policy optimal time visual motion cells direction field spatial model receptive orientation response the of a and is in i for to we	image jpeg file graphics images color files gif format bit windows comp os ms dos x microsoft unix window program space nasa sci launch shuttle venus gov station mission orbit rutgers christian god geneva religion athos church jesus soc may the of in to and a on for was by
conBP	the of and classification training class classifier to for in the network units of to input hidden output layer unit the of and control to in model is motor trajectory the to learning and is robot s goal environment task	image jpeg file you it from graphics images the files windows comp os ms edu i misc cs for dos edu gov com nasa apr stratus usenet indiana ucs jpl rutgers edu christian of in that god we religion i

Fig. 3. Top ten words of four topics when $K = 50$. Blue and black colors denote stop and common words, respectively. Red color denotes meaningful key words in each topic.

topic features. We can also use the document-topic distribution as the soft document clustering results. Normalized mutual information (NMI) [19] evaluates the performance of clustering by comparing predicted clusters with true class labels of a corpus. When displaying topics to users, each topic is generally represented as a list of the most probable words (for example, top ten hot words in each topic). Topic "coherence" [20] evaluates the topic quality. Point-wise mutual information (PMI) [21] is very similar to coherence. The higher coherence and PMI values correspond to the better topic interpretability.

For a fair comparison, we implement all algorithms using the MATLAB C/C++ MEX platform publicly available at [22] and run experiments on the Sun fire X4270 M2 server. The initial hyperparameters is set as $\alpha = 50/K, \beta = 0.01$, where K is the number of topics. We use the same $T = 1000$ training iterations for all algorithms. We compare our algorithms with the four benchmark topic modeling algorithms such as BP [3], asGS [12], pmiGS [9] and STC [13]. Since STC outputs the word-topic distribution containing negative values, we only compare our algorithms with STC in terms of document clustering and classification tasks.

Fig. 3 shows the top ten words of four topics when $K = 50$. The meaningful key words of each topic are highlighted with the red color, and the stop and

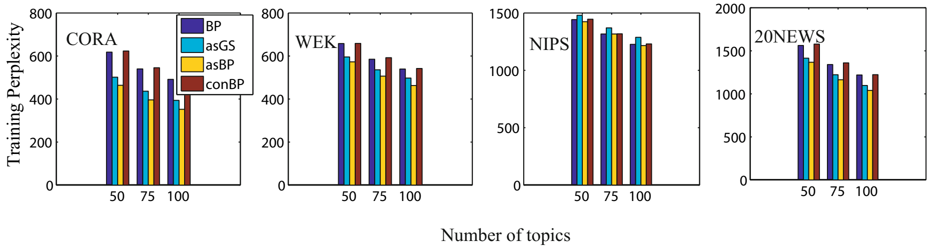


Fig. 4. Training perplexity as a function of the number of topics

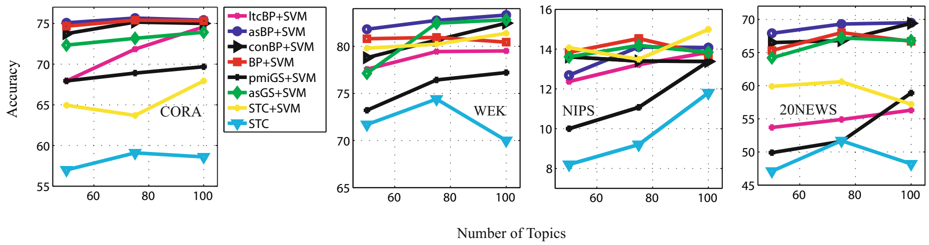


Fig. 5. Document classification accuracy as a function of the number of topics

common words are highlighted with blue and black colors, respectively. We use the subjective “word intrusion” [4] to evaluate the topic interpretability, i.e., the number of conflict stop and common words in each topic. It is easy to see that ltcBP performs the best to remove almost all stop and common words in each topic, which demonstrates the effectiveness of the continuous LTC features in topic modeling. Note that asBP can also remove the most stop words by clustering them such as “the of a and is in i for we” in a separate topic on both NIPS (STOP) and 20NEWS (STOP). This result shows that the asymmetric prior has an effect on allocating the most frequent stop words to a specific topic with a higher prior value v_k. But asBP still has difficulty in handling some common words like “learning” and “model”. Note that asGS can also cluster stop words in one topic, but some topics contain more common words than those of asBP. BP performs the worst since its extracted topics are influenced by those high-frequent stop and common words. Although pmiGS uses the continuous PMI feature in topic modeling, it performs significantly worse than ltcBP because it cannot remove most stop and common words in each topic. The underlying reason is that LTC features are more effective in lowering the weights of stop and common words in topic modeling. We see that using sparseness constraints cannot effectively remove stop and common words from each topic. The conBP is only slightly better than BP, but significantly worse than both asBP and ltcBP. So, to find more interpretable topic-word distributions, the continuous features and asymmetric priors provide the best performance.

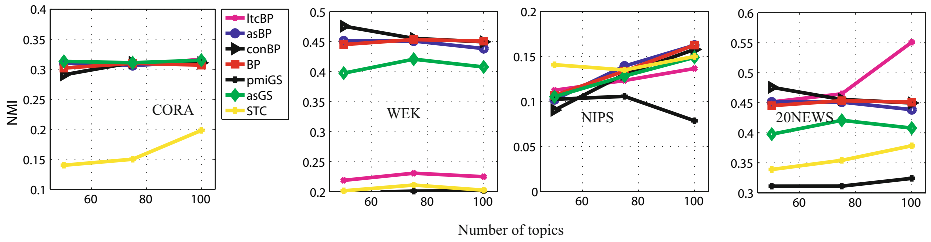


Fig. 6. The NMI as a function of the number of topics

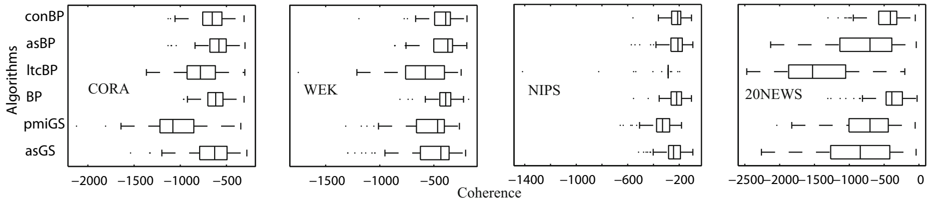


Fig. 7. The coherence of CORA, WEK, NIPS and 20NEWS datasets when $K = 100$

Fig. 4 shows the training perplexity as a function of the number of topics on CORA, WEK, NIPS and 20NEWS for $K = \{50, 75, 100\}$. Note that ltcBP, pmiGS and STC do not describe how to generate word tokens, so that they cannot be measured by the perplexity metric. Except on NIPS, asGS yields a lower perplexity value than BP. We see that conBP has almost the same perplexity of BP, which implies that sparseness constraints do not improve the likelihood of word generation. On all data sets, we see that the training perplexity of asBP is the lowest, showing the highest topic modeling accuracy. The result shows that learning asymmetric Dirichlet prior of αv_k and the symmetric prior βu_w can improve the topic modeling accuracy. The training perplexity has a smaller difference on the NIPS data set. One possible reason is that each document in NIPS contains more word tokens, so that the prior has a smaller impact on the message update (1). To summarize, learning an asymmetric Dirichlet prior over the document-topic distributions and an symmetric Dirichlet prior over the topic-word distributions still has substantial advantages on improving the document-topi and topic-word distributions to generate word tokens.

Fig. 5 shows the document classification accuracy as a function of the number of topics on CORA, WEK, NIPS and 20NEWS for $K = \{50, 75, 100\}$. In our experiments, we randomly divide each data set into half as training and test sets. Then, we use the standard linear SVM classifier to classify the lower-dimensional document-topic features produced by the topic modeling algorithms. As far as STC is concerned, it can directly output the class predictions. Also, we can use STC to generate lower-dimensional topic features and use SVM to do the classification.

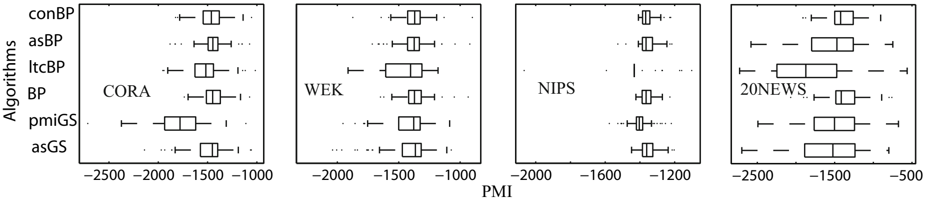


Fig. 8. The PMI of CORA, WEK, NIPS and 20NEWS datasets when $K = 100$

Table 2. Performance on CORA, WEK, NIPS and 20NEWS datasets when $K = 100$

Datasets	CORA					WEK				
	Perplexity	Accuracy	NMI	PMI	Coherence	Perplexity	Accuracy	NMI	PMI	Coherence
ltcBP	–	74.58	**0.3168**	−1536.1	−781.42	–	79.51	0.2251	−1458.8	−620.45
asBP	**352.29**	**75.42**	0.3150	−1447.6	**−609.72**	**462.66**	83.32	0.2469	−1380.6	−426.72
conBP	496.16	75.00	0.3107	−1466.7	−673.98	541.94	**84.46**	0.2347	−1381.4	−428.86
BP	491.31	75.33	0.3069	**−1444.5**	−631.86	539.41	80.45	0.2297	**−1370.9**	**−403.68**
pmiGS	–	69.68	0.2161	−1788.9	−1075.10	–	77.21	0.2015	−1418.9	−553.56
asGS	393.79	73.92	0.2852	−1485.5	−677.81	497.56	82.82	**0.2532**	−1425.0	−526.96
STC	–	67.94	0.1981	–	–	–	81.38	0.2027	–	–

Datasets	NIPS					20NEWS				
	Perplexity	Accuracy	NMI	PMI	Coherence	Perplexity	Accuracy	NMI	PMI	Coherence
ltcBP	–	13.84	0.1365	**−1254.1**	−415.10	–	56.30	**0.5511**	−1837.3	−1461.1
asBP	**1215.72**	14.07	**0.1632**	−1357.2	−227.49	**1039.56**	**69.50**	0.4386	−1549.2	−793.0
conBP	1230.30	13.38	0.1577	−1357.5	**−225.55**	1222.99	69.40	0.4498	−1386.9	−474.63
BP	1226.43	13.73	0.1626	−1358.3	−226.54	1219.04	66.70	0.4511	**−1374.8**	**−411.8**
pmiGS	–	13.38	0.0785	−1389.7	−279.01	–	58.90	0.3242	−1518.1	−772.23
asGS	1288.02	13.84	0.1489	−1349.9	−249.15	1096.89	66.80	0.4079	−1604.7	−906.59
STC	–	**14.99**	0.1449	–	–	–	57.20	0.3785	–	–

We see that BP and asBP performs comparably, and outperform other methods. Their classification performance is relatively stable as the number of topics changes. Although ltcBP can effectively remove stop and common words, it does not perform the best in document classification. On possible reason is that the distributions of stop and common words also provide useful information for classification. Surprisingly, STC cannot predict the class label very well when compared with other methods. But STC works well on the lower-dimensional topic features. As we see, conBP works slightly better than BP on classification when $K = 100$, which implies that sparseness constraints do not provide useful information in this task. Overall, asBP performs the best in document classification. For example, asBP outperforms BP and asGS by around 0.6% and 3.7% on CORA for $K = 50$, and by around 4.0% and 3.9% on 20NEWS data set for $K = 100$ in terms of classification accuracy. This result shows that the asymmetric priors play an important role in regulating document-topic features for classification. When the dimensionality of latent space is small, learning an asymmetric Dirichlet prior over the document-topic distributions and symmetric Dirichlet prior over the topic-word distributions is worse than heuristically set symmetric Dirichlet priors on NIPS. One reason is that the Dirichlet prior have more effects on shorter documents than longer documents.

Fig. 6 shows the document clustering results measured by NMI. This result confirms that STC and pmiGS often predict the wrong clusters of documents on all data sets. All BP-based algorithms perform equally well but conBP performs slightly better when $K = 100$. It is interesting to see that the performance of document clustering is not consistent with that of document classification in Fig. 5. One possible reason is the unknown number of clusters in the clustering task.

Fig. 7 shows the coherence on all data sets when $K = 100$. Because STC has no topic-word distributions, it cannot be measured by the coherence metric. The plot produces a separate box for $K = 100$ coherence values of each algorithm. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually by the black dot sign. We see that asBP and conBP have higher coherence median values with smaller variances. BP also yields a stable coherence value. However, ltcBP and pmiGS have lower coherence values. The major reason is that they remove most common words, which contribute much to the coherence metric.

Fig. 8 shows the PMI values of all algorithms when $K = 100$. Because STC has no topic-word distributions, it cannot be measured by the PMI metric. The plot produces a separate box for $K = 100$ PMI values of each algorithm. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually by the black dot sign. We see that most results are consistent with those of Fig. 7. For example, asBP, conBP and BP have relatively smaller variances and median values, while ltcBP and pmiGS have relatively bigger variances and median values. Both Fig. 7 and 8 confirm that asBP provide more coherent and related word groups. Note that asBP clusters stop and common words in a separate topic, which enhances coherence and PMI when compared with ltcBP.

Table 2 summarizes the overall performance of all algorithms on four data sets when $K = 100$. We mark the best performance by the bold face. We see that asBP wins 8/20 columns and all variants of BP win around 18/20 columns. This result confirms that BP and its variants find better document-topic and topic-word distributions. As far as perplexity is concerned, asBP is always the best method, which means that it is very likely to recover the observed words from the document-topic and topic-word distributions. We see that ltcBP and asBP learns better document-topic distributions for soft document clustering with relatively higher NMI values. Moreover, both ltcBP and asBP can effectively remove stop and common words as shown in Fig. 3. Although STC uses sparse coding for document classification, it performs relatively worse than conBP partly because conBP incorporates the sparseness constraints naturally. Note that conBP often provides a stable clustering and classification performances though it is not the best. On CORA and 20NEWS, conBP outperforms BP with a large margin, which reflects that sparseness constraints can improve clustering and classification performance. When compared with pmiGS, ltcBP wins all columns, confirming the effectiveness of LTC features for topic modeling as well as BP framework for learning LDA. Form Table 2, we suggest continuous features and asymmetric priors for topic modeling because sparseness constraints do not provide significant improvement. The underlying reason is that the estimated document-topic and topic-word distributions are already very sparse so that any sparseness constraints can give only marginal improvement.

5 Conclusions

In this paper, we extensively explore three factors to find better topics: continuous features, asymmetric priors, and sparseness constraints within the unified BP framework. We develop several novel BP-based algorithms to study the three perspectives. Through extensive experiments, we advocate asymmetric priors for topic modeling because they can enhance the overall performance in terms of several metrics. Also, the continuous features can improve the interpretability of topic-word distributions by effectively remove almost all stop and common words. Finally, we find that sparseness constraints do not improve the topic modeling performance very much, partly because the sparse nature of document-topic and topic-word distributions of LDA.

Acknowledgements. This work is supported by NSFC (Grant No. 61003154, 61373092, 61033013, 61272449 and 61202029), Natural Science Foundation of the Jiangsu Higher Education Institutions of China (Grant No. 12KJA520004), Innovative Research Team in Soochow University (Grant No. SDT2012B02), and Guangdong Province Key Laboratory Project(Grant No. SZU-GDPHPCL-2012-09).

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022 (2003)
2. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proc. Natl. Acad. Sci. 101, 5228–5235 (2004)
3. Zeng, J., Cheung, W.K., Liu, J.: Learning topic models by belief propagation. IEEE Trans. Pattern Anal. Mach. Intell. 33(5), 1121–1134 (2013)
4. Chang, J., Boyd-Graber, J., Gerris, S., Wang, C., Blei, D.: Reading tea leaves: How humans interpret topic models. In: NIPS, pp. 288–296 (2009)
5. Salton, G., McGill, M.J.: Introduction to modern information retrieval. McGraw-Hill, New York (1983)
6. Buckley, C.: Automatic query expansion using SMART: Trec 3. In: Proceedings of The Third Text REtrieval Conference (TREC-3), pp. 69–80 (1994)
7. Hoffman, M., Blei, D., Bach, F.: Online learning for latent Dirichlet allocation. In: NIPS, pp. 856–864 (2010)
8. Ramage, D., Heymann, P., Manning, C.D., Garcia-Molina, H.: Clustering the tagged web. In: Web Search and Data Mining, pp. 54–63 (2009)

9. Wilson, A.T., Chew, P.A.: Term weighting schemes for latent Dirichlet allocation. In: North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 465–473 (2010)
10. Minka, T.P.: Estimating a Dirichlet distribution. Technical report, Microsoft Research (2000)
11. Asuncion, A., Welling, M., Smyth, P., Teh, Y.W.: On smoothing and inference for topic models. In: UAI, pp. 27–34 (2009)
12. Wallach, H., Mimno, D., McCallum, A.: Rethinking LDA: Why priors matter. In: NIPS, pp. 1973–1981 (2009)
13. Zhu, J., Xing, E.P.: Sparse topical coding. In: UAI (2011)
14. Zhu, W., Zhang, L., Bian, Q.: A hierarchical latent topic model based on sparse coding. Neurocomputing 76(1), 28–35 (2012)
15. Hoyer, P.O.: Non-negative matrix factorization with sparseness constraints. Journal of Machine Learning Research 5, 1457–1469 (2004)
16. Zeng, J., Cao, X.-Q., Liu, Z.-Q.: Residual belief propagation for topic modeling. In: Zhou, S., Zhang, S., Karypis, G. (eds.) ADMA 2012. LNCS, vol. 7713, pp. 739–752. Springer, Heidelberg (2012)
17. Zeng, J., Liu, Z.Q., Cao, X.Q.: A new approach to speeding up topic modeling, arXiv:1204.0170 [cs.LG] (2012)
18. Heinrich, G.: Parameter estimation for text analysis. Technical report, University of Leipzig (2008)
19. Zhong, S., Ghosh, J.: Generative model-based document clustering: A comparative study. Knowl. Inf. Syst. 8(3), 374–384 (2005)
20. Mimno, D.M., Wallach, H.M., Talley, E.M., Leenders, M., McCallum, A.: Optimizing semantic coherence in topic models. In: EMNLP, pp. 262–272 (2011)
21. Newman, D., Karimi, S., Cavedon, L.: External evaluation of topic models. In: Australasian Document Computing Symposium, pp. 11–18 (2009)
22. Zeng, J.: TMBP: A topic modeling toolbox using belief propagation. J. Mach. Learn.Res. 13, 2233–2236 (2012)