J Sign Process Syst (2014) 74:375–389
DOI 10.1007/s11265-013-0759-x

Bayesian Sparse Topic Model

Jen-Tzung Chien · Ying-Lan Chang

Received: 1 October 2012 / Revised: 15 March 2013 / Accepted: 24 April 2013 / Published online: 30 June 2013
© Springer Science+Business Media New York 2013

Abstract This paper presents a new Bayesian sparse learning approach to select salient lexical features for sparse topic modeling. The Bayesian learning based on latent Dirichlet allocation (LDA) is performed by incorporating the *spike-and-slab priors*. According to this sparse LDA (sLDA), the spike distribution is used to select salient words while the slab distribution is applied to establish the latent topic model based on those selected relevant words. The variational inference procedure is developed to estimate prior parameters for sLDA. In the experiments on document modeling using LDA and sLDA, we find that the proposed sLDA does not only reduce the model perplexity but also reduce the memory and computation costs. Bayesian feature selection method does effectively identify relevant topic words for building sparse topic model.

Keywords Bayesian sparse learning · Feature selection · Topic model

1 Introduction

The goal of feature selection aims to select a subset of relevant features for building robust learning machine. By discarding the outlier features from raw data, feature selection can be employed to increase the generalization capability, speed up the learning process, and improve the model interpretability [7]. The curse of dimensionality can be alleviated as well. Feature selection has been known as one of the most challenging issues in pattern recognition and machine learning. An attractive approach to this issue is to conduct *Bayesian variable selection* [18] where *a priori* knowledge about a relatively small proportion of influential features was considered. Also, in the view of unsupervised learning, there should be only a few features which contribute for learning model structure. Many other features may be redundant or even harmful for structural learning. For instance, in gene mapping problem, it is assumed that there are only a small number of genes that have substantial effect on trait, while most of genes have little or even no effect. The underlying biological structure is sparse, i.e. only a few factors have influence on the trait. Sparse representation is meaningful for this problem. In addition, Bayesian sparse learning [22] was performed through Bayesian treatment by introducing the prior of weight parameter that expressed the uncertainty in heterogeneous data and enforced the sparsity of representation. A sparse prior distribution was introduced. Typically, sparsity-favoring prior is defined as any distribution with excess of kurtosis, indicating that it is highly peaked with heavy tails or it has a delta-mass at zero.

In general, prior distributions that favor sparsity fall into two categories: continuous sparsity-favoring priors and spike-and-slab priors. Laplace distribution [1] and Student's t-distribution [22] correspond to the first category of priors where high kurtosis is measured and the resulting Bayesian models for continuous variables are prone to be sparse. On the other hand, the spike-and-slab prior model [13, 16, 17] is formed by a discrete mixture of a point mass at zero, which is referred to as the 'spike', and any other distribution, which is known as the 'slab'. This distribution allows Bayesian inference with exact zeroes in the posterior samples for

J.-T. Chien (✉) · Y.-L. Chang
Department of Electrical and Computer Engineering,
National Chiao Tung University,
Hsinchu, Taiwan 30010, Republic of China
e-mail: jtchien@nctu.edu.tw

Y.-L. Chang
e-mail: ylchang@chien.cm.nctu.edu.tw

discrete variables, thereby enforcing true sparsity. Bayesian sparse learning has been attracting many researchers in the communities of signal processing and machine learning and has been developed for speech recognition [19], image reconstruction [1], document representation [5, 23], choice modeling [10], and many others.

In the application of document modeling, latent Dirichlet allocation (LDA) [3] was developed to build latent topic model from observed documents and then extended for gene clustering, document clustering, document summarization [4], and language modeling [6]. Using LDA, each word in a document is viewed as a feature which is represented by a fixed set of topic mixtures. The mixture weights are used to build coordinate vector of a word in semantic or topic space. However, some words or features are noisy, irrelevant or redundant and shall result in a poor model. How to select semantically significant features becomes a key issue in LDA-based topic model.

Recently, Wang and Blei [23] proposed a sparse representation based on the hierarchical Dirichlet process (HDP) [20]. This work decoupled the sparsity and smoothness in HDP and achieved a sparse topic model by introducing a Bernoulli distribution to detect whether each feature appears in the topic or not. Conditioned on these variables, each topic is represented by a multinomial distribution over its subset of vocabulary words. More recently, a focused topic model (FTM) [24] was exploited to learn the sparse topic mixture patterns from documents. This method integrated the desirable features through HDP and Indian buffet process (IBP) [9, 11] and allowed sparse representation over different topics. IBP is an exchangeable distribution over binary matrices for implementing the Bayesian nonparametric feature model [21]. A variational inference algorithm based on a truncated stick-breaking approximation [8] was developed. Furthermore, a sparse exponential family [17] was proposed to fulfill sparse representation of latent variable model based on the exponential family distributions. In [15], a Bayesian topic model was established by combining the efficiency of sparse Gibbs sampling with the scalability of online stochastic inference.

This paper proposes a new Bayesian sparse topic model where sparse LDA (sLDA) is implemented via Bayesian feature selection by using spike-and-slab prior distribution. The semantically-significant features are selected to assure model fitness and generalization. An indicator variable with Bernoulli distribution is adopted for feature selection [14]. Using this method, the memory and computation requirements are significantly reduced. Sparse topic model is established for document representation. Experiments on datasets of Wall Street Journal (WSJ) and Associated Press newswire (AP) show effectiveness and efficiency by using sLDA compared to LDA. The organization of this paper is arranged as follows. First of all, we survey LDA model and address some issues in LDA. Then, Bayesian sparse topic model is introduced. The spike-and-slab distribution is surveyed. The resulting model construction and inference based on sLDA are described. Several related methods are compared. Next, the experiments on document modeling using different methods are evaluated. Sparsity of the estimated parameters and hyperparameters is analyzed. Finally, the conclusions drawn from this study are given.

2 Topic Model

2.1 Latent Dirichlet Allocation

Blei et al. [3] introduced the LDA for topic-based document representation where the documents are treated as 'a bag of words'. LDA is known as an extension of topic model based on probabilistic latent semantic analysis (PLSA) [12]. LDA improves PLSA by generalizing for unseen documents through the shared topic information expressed by Dirichlet prior. LDA has been recognized as the representative topic model. Figure 1 shows the graphical representation of LDA. There are N words in a document, V vocabulary words, K latent topics and M documents in the corpus. Each word w in a document d is associated with a hidden variable z, which denotes the latent topic. Variable z is sampled from a multinomial distribution with parameter θ indicating the generating probabilities of latent topics. The prior density of multinomial parameter θ is given by a Dirichlet distribution with K-dimensional hyperparameter $\boldsymbol{\alpha} = \{\alpha_k\}$. The $K \times V$ parameter matrix $\boldsymbol{\beta} = \{\beta_{kv}\}$ denotes the topic-dependent word probability. LDA outperformed PLSA and other topic models in evaluation of document modeling [3]. LDA was also extended for document summarization [4] and speech recognition [6]. A Dirichlet class language model was developed for speech recognition by considering the word order into LDA-based class or topic prediction. In general, LDA is constructed for document representation as follows:

1. For each document $d \in \{1, \cdots, M\}$
   Draw a K-dimensional topic mixture vector by $\boldsymbol{\theta}_d \sim \mathrm{Dir}(\boldsymbol{\alpha})$

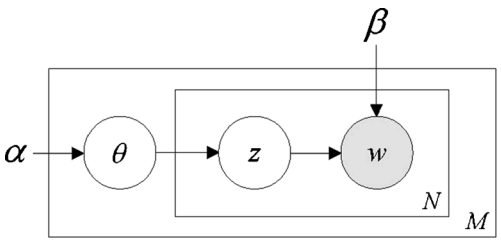


Figure 1 Graphical model for LDA.

2. For each word w_n in document d where $n \in \{1, \cdots, N\}$

   (a) Choose a topic by $z_{dn} \sim \text{Mult}(\boldsymbol{\theta}_d)$
   (b) Choose a word by $w_{dn}|\{z_{dn} = k\} \sim \text{Mult}(\boldsymbol{\beta}_k)$

The LDA parameters $\{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$ are estimated by maximizing the marginal likelihood $p(\mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})$ from a set of text documents $\mathbf{w} = \{w_{dn}\}$

$$p(\mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{d=1}^{M} \int p(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) \times \left[\prod_{n=1}^{N} \sum_{z_{dn}} p(w_{dn}|z_{dn}, \boldsymbol{\beta}) p(z_{dn}|\boldsymbol{\theta}_d) \right] d\boldsymbol{\theta}_d \tag{1}$$

where the marginalization is operated over Dirichlet parameter $\boldsymbol{\theta}_d$ and latent topics $\mathbf{z} = \{z_{dn} = k\}$ of the words in corpus $\mathbf{w}$. However, direct optimization of Eq. 1 is intractable. The variational inference is applied to estimate LDA parameters by maximizing a lower bound of Eq. 1. Considering the factorized variational inference where latent variables $\mathbf{z}$ and $\boldsymbol{\theta} = \{\boldsymbol{\theta}_d\}$ are conditionally independent, a variational distribution $q(\mathbf{z}, \boldsymbol{\theta}|\boldsymbol{\phi}, \boldsymbol{\gamma})$ of $\{\mathbf{z}, \boldsymbol{\theta}\}$ is formed to approximate the true posterior probability $p(\mathbf{z}, \boldsymbol{\theta}|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})$. By maximizing the lower bound, the optimal variational parameters $\left\{\hat{\boldsymbol{\phi}} = \left\{\hat{\phi}_{nk}\right\}, \hat{\boldsymbol{\gamma}} = \left\{\hat{\gamma}_k\right\}\right\}$ and LDA parameters $\left\{\hat{\boldsymbol{\alpha}} = \left\{\hat{\alpha}_k\right\}, \hat{\boldsymbol{\beta}} = \left\{\hat{\beta}_{kv}\right\}\right\}$ are estimated by

$$\hat{\phi}_{nk} \propto \beta_{kw_n} \exp\left\{ \Psi(\gamma_k) - \Psi\left(\sum_{j=1}^{K} \gamma_j \right) \right\} \tag{2}$$

$$\hat{\gamma}_k = \alpha_k + \sum_{n=1}^{N} \phi_{nk} \tag{3}$$

$$\hat{\beta}_{kv} \propto \sum_{d=1}^{M} \sum_{n=1}^{N} \phi_{dnk} \delta(w_{dn}, v) \tag{4}$$

$$\hat{\boldsymbol{\alpha}}^{(t+1)} = \boldsymbol{\alpha}^{(t)} - \mathbf{H}_{lda}\left(\boldsymbol{\alpha}^{(t)}\right)^{-1} \mathbf{g}_{lda}\left(\boldsymbol{\alpha}^{(t)}\right) \tag{5}$$

where $\Psi(\cdot)$ denotes the first derivative of log gamma function $\log\Gamma(\cdot)$, $\delta(\cdot)$ denotes the Kronecker delta function, t denotes the iteration index in decent algorithm, $\mathbf{H}_{lda}(\cdot)$ and $\mathbf{g}_{lda}(\cdot)$ denote the Hessian matrix and the gradient vector of the lower bound with respect to $\boldsymbol{\alpha}$, respectively. The estimated variational distribution $\hat{q}\left(\mathbf{z}, \boldsymbol{\theta}|\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\gamma}}\right)$ approximates the true distribution $p(\mathbf{z}, \boldsymbol{\theta}|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ with the smallest Kullback–Leibler (KL) divergence.

2.2 Some Issues in LDA

Some issues exist in LDA and could be tackled to improve topic-based document representation. From Eqs. 2–4, we find that the variational parameters ϕ_{nk} and γ_k act as sufficient statistics for calculating topic-dependent word probability $\hat{\beta}_{kv}$ for topic k and word $v = w_{dn}$. The probability $\hat{\beta}_{kv}$ is seen as the proportion or mixture weight of a word v assigned by different topic or mixture component k. Considering a case study with three topics $\{k_1, k_2, k_3\}$ as illustrated in Fig. 2, the topic assignment of a word w_{dn} is determined by finding topic k with the highest probability $\hat{\beta}_{kv}$ among K topics. LDA parameter $\hat{\beta}_{kv}$ is calculated by summing up the variational topic proportions $\left\{\hat{\phi}_{nk}\right\}$ corresponding to word v in different training documents. Typically, all words in a document are fully connected to different topics as marked by yellow in the figure. The issues in LDA are twofold. First, the computation of this fully-connected network between words and topics is proportionally increased by the number of topics. Second, some words or topics are noisy and irrelevant for model construction. In this study, we select the informative features and prune the redundant features for compact topic modeling. A sparse LDA (sLDA) is proposed. The fully-connected network is simplified to a partially-connected network where irrelevant words and topics are automatically detected and disconnected via sparse Bayesian learning by using spike-and-slab priors. The computation and memory requirements are alleviated accordingly.

3 Bayesian Sparse Topic Model

Real-world text documents data are usually contaminated with noisy and redundant words. Generalization of a trained model to new data is not assured. Selecting the informative features becomes a crucial issue for model construction. This paper proposes a Bayesian feature selection for topic-

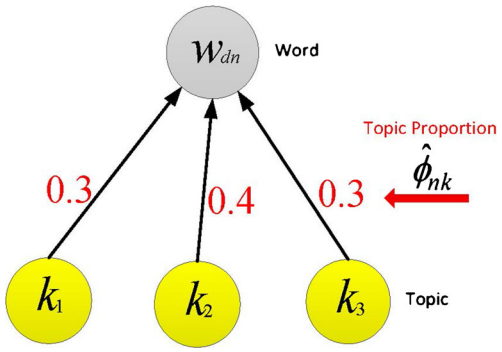


Figure 2 Illustration for LDA in case that a word w_{dn} is assigned by three topics $\{k_1, k_2, k_3\}$. All three topics are active (and *marked by yellow*) for prediction of word w_{dn}.

based document representation. Only the selected features are employed in construction of topic model. All documents are generated by a shared topic mixture model where a basis of topic-dependent word distributions with topic proportions is introduced. Each word in a document is sampled according to this distribution basis. Importantly, we disregard the uncertain features and select the fitted features for text modeling. The conceptual difference between LDA and sLDA is demonstrated by Fig. 3. Typically, sLDA adopts the spike-and-slab distribution to pursue sparsity when estimating the word distributions for different topics. In what follows, the spike-and-slab prior distribution for Bayesian sparse learning is addressed.

3.1 Spike-and-Slab Distribution

Spike-and-slab distribution is formed as a mixture model of an impulse mass at zero referred to as 'spike' and any other distribution known as 'slab' [17]. The original distribution was expressed in [16]. This distribution was seen as a discrete mixture of two prior distributions [13, 17] and acted as a type of prior for linear regression. In [17], spike-and-slab distribution was explored for unsupervised learning of sparse exponential family. Here, we present a new sLDA document model by introducing spike-and-slab model for Bayesian feature selection and topic modeling. Bayesian inference with exact zeroes in posterior distributions is performed to achieve true sparsity. The 'spike' model is realized through a binary indicator matrix indicating whether a latent variable contributes to generating an observation sample or not. Each observation has a corresponding Bernoulli indicator variable. The beta prior and its hyperparameters are incorporated in Bayesian learning [2]. On the other hand, the 'slab' model is implemented by either discrete or continuous variable which is expressed by exponential family distribution and its conjugate prior. In general, the use of multinomial and Gaussian distributions for discrete and continuous slab variables is beneficial to facilitate rapid Gibbs sampling of the posterior, respectively. This makes spike-and-slab variable selection computationally attractive and extensively popular. In this study, a Bayesian sparse

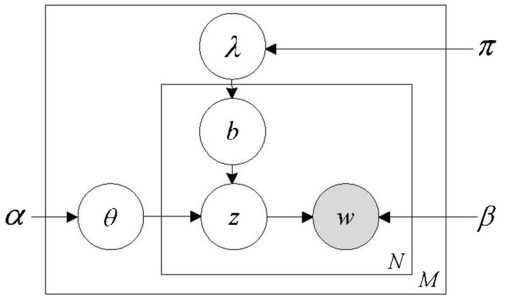


Figure 4 Graphical model for sLDA.

topic model is proposed for document representation where the text documents are represented by a mixture of latent variables which is extended from topic model based on LDA. The semantically-rich words in documents are automatically selected and merged into construction of a precise and reliable topic model.

3.2 Model Construction

Bernoulli distribution is introduced as a spike model to judge whether the target feature is informative or not. The beta distribution is used as conjugate prior [2] for spike model. The standard LDA can be seen as a slab model consisting of those features which are selected by spike model. Figure 4 depicts the graphical representation for sLDA. Model construction of sLDA is addressed as follows:

1. For each document $d \in \{1, \cdots, M\}$
   (a) Draw a proportion by $\lambda_{dk} \sim \text{Beta}(\boldsymbol{\pi})$
   (b) Draw a K-dimensional topic mixture vector by $\boldsymbol{\theta}_d \sim \text{Dir}(\boldsymbol{\alpha})$
2. For each word w_n in document d where $n \in \{1, \cdots, N\}$
   (a) Choose an indicator by $b_{dnk} \sim \text{Bern}(\lambda_{dk})$
   (b) Choose a topic by $z_{dn} \sim \text{Mult}(\boldsymbol{\theta}_d)$
   (c) Choose a word by $w_{dn} | \{b_{dnk} = 1, z_{dn} = k\} \sim \text{Mult}(\boldsymbol{\beta}_k)$

In this procedure, the indicator variable b_{dnk} is introduced for each latent variable z_{dn}. This indicator is governed

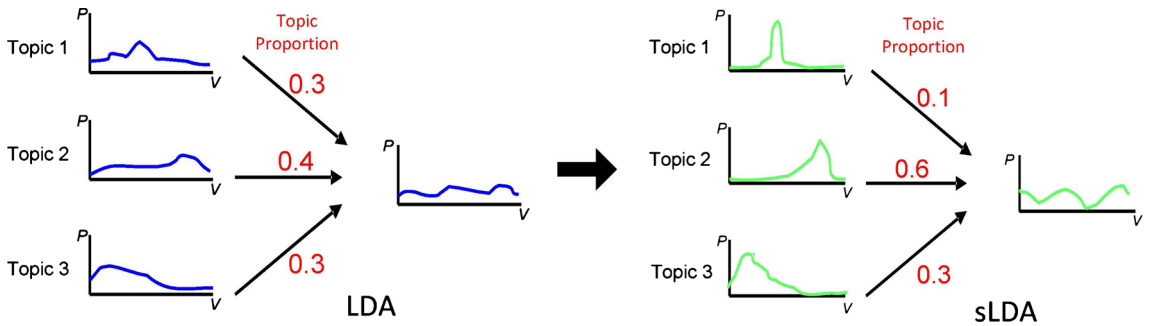


Figure 3 Comparison of LDA and sLDA for generation of word distributions.

by a document-dependent Bernoulli parameter λ_{dk} which is drawn from a beta prior distribution with parameter $\boldsymbol{\pi}$. The topic label $z_{dn} \in \Omega_k$ and its associated word distribution $\boldsymbol{\beta}_k$ are used to generate word w_{dn}. The indicator b_{dnk} determines whether word w_{dn} is relevant to topic k or not. Given this specialized sLDA, the marginal likelihood of training documents $\mathbf{w} = \{w_{dn}\}$ by using slab model with $b_{dnk} = 1$ is calculated by

$$p(\mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi}) = \prod_{d=1}^{M} \int p(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) \times \left[\prod_{n=1}^{N} \sum_{z_{dn} \in \Omega_k} p(w_{dn}|z_{dn} = k, b_{dnk} = 1, \boldsymbol{\beta}) \times p(z_{dn} = k|b_{dnk} = 1, \boldsymbol{\theta}_d) \times \int p(b_{dnk} = 1|\lambda_{dk}) p(\lambda_{dk}|\boldsymbol{\pi}) d\lambda_{dk} \right] d\boldsymbol{\theta}_d. \quad (6)$$

In case of spike model with $b_{dnk} = 0$, the word w_{dn} for topic $z_{dn} = k$ is viewed as a redundancy and is excluded from building topic model. In this sLDA, there are one observation variable $\mathbf{w}$ and four latent variables $\{\mathbf{z}, \boldsymbol{\theta}, \mathbf{b} = \{b_{dnk}\}, \boldsymbol{\lambda} = \{\lambda_{dk}\}\}$. Model parameters $\{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi}\}$ are estimated by maximizing the marginal likelihood as given in Eq. 6.

3.3 Model Inference

However, the exact solution to direct maximization of marginal likelihood in Eq. 6 does not exist due to the coupling effect of latent variables $\mathbf{z}$, $\boldsymbol{\theta}$, $\mathbf{b}$ and $\boldsymbol{\lambda}$ in posterior distribution $p(\mathbf{z}, \boldsymbol{\theta}, \mathbf{b}, \boldsymbol{\lambda}|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi})$. We develop a new variational Bayesian expectation maximization (VB-EM) procedure for the proposed sLDA. Instead of direct maximization, we apply the factorized variational inference and maximize the lower bound $\mathcal{L}(\cdot)$ of the logarithm of marginal likelihood, i.e.

$$\begin{aligned} &\log p(\mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi}) \\ &= \log \int\int\int \sum_{\mathbf{z}} \frac{p(\mathbf{w}, \mathbf{z}, \boldsymbol{\theta}, \mathbf{b}, \boldsymbol{\lambda}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi}) q(\mathbf{z}, \boldsymbol{\theta}, \mathbf{b}, \boldsymbol{\lambda})}{q(\mathbf{z}, \boldsymbol{\theta}, \mathbf{b}, \boldsymbol{\lambda})} d\boldsymbol{\theta} d\mathbf{b} d\boldsymbol{\lambda} \\ &\geq \int\int\int \sum_{\mathbf{z}} q(\mathbf{z}, \boldsymbol{\theta}, \mathbf{b}, \boldsymbol{\lambda}) \log p(\mathbf{w}, \mathbf{z}, \boldsymbol{\theta}, \mathbf{b}, \boldsymbol{\lambda}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi}) d\boldsymbol{\theta} d\mathbf{b} d\boldsymbol{\lambda} \\ &\quad - \int\int\int \sum_{\mathbf{z}} q(\mathbf{z}, \boldsymbol{\theta}, \mathbf{b}, \boldsymbol{\lambda}) \log q(\mathbf{z}, \boldsymbol{\theta}, \mathbf{b}, \boldsymbol{\lambda}) d\boldsymbol{\theta} d\mathbf{b} d\boldsymbol{\lambda} \\ &\triangleq \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\psi}, \boldsymbol{\eta}; \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi}) \end{aligned} \quad (7)$$

which is derived by applying the Jensen's inequality. Here, we introduce the hyperparameters or variational parameters $\boldsymbol{\phi}$, $\boldsymbol{\gamma}$, $\boldsymbol{\psi}$ and $\boldsymbol{\eta}$ corresponding to the latent variables $\mathbf{z}$, $\boldsymbol{\theta}$, $\mathbf{b}$ and $\boldsymbol{\lambda}$, respectively, and determine the variational distribution $q(\mathbf{z}, \boldsymbol{\theta}, \mathbf{b}, \boldsymbol{\lambda}|\boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\psi}, \boldsymbol{\eta})$. This distribution is used to approximate the true posterior distribution $p(\mathbf{z}, \boldsymbol{\theta}, \mathbf{b}, \boldsymbol{\lambda}|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi})$. The lower bound in Eq. 7 is expanded as

$$\begin{aligned} \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\psi}, \boldsymbol{\eta}; \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi}) &= E_q[\log p(\mathbf{w}|\mathbf{z}, \mathbf{b}, \boldsymbol{\beta})] \\ &+ E_q[\log p(\mathbf{z}|\mathbf{b}, \boldsymbol{\theta})] + E_q[\log p(\boldsymbol{\theta}|\boldsymbol{\alpha})] \\ &+ E_q[\log p(\mathbf{b}|\boldsymbol{\lambda})] + E_q[\log p(\boldsymbol{\lambda}|\boldsymbol{\pi})] \\ &- E_q[\log q(\mathbf{z})] - E_q[\log q(\boldsymbol{\theta})] \\ &- E_q[\log q(\mathbf{b})] - E_q[\log q(\boldsymbol{\lambda})]. \end{aligned} \quad (8)$$

Notably, latent variables in variational distribution are assumed to be conditionally independent as $q(\mathbf{z}, \boldsymbol{\theta}, \mathbf{b}, \boldsymbol{\lambda}|\boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\psi}, \boldsymbol{\eta}) = q(\mathbf{z}|\boldsymbol{\phi}) q(\boldsymbol{\theta}|\boldsymbol{\gamma}) q(\mathbf{b}|\boldsymbol{\psi}) q(\boldsymbol{\lambda}|\boldsymbol{\eta})$ where $\boldsymbol{\phi}$, $\boldsymbol{\gamma}$, $\boldsymbol{\psi}$ and $\boldsymbol{\eta}$ denote the variational parameters for multinomial, Dirichlet, Bernoulli and beta distributions, respectively. The expectation functions in Eq. 8 are calculated individually for latent variables $\mathbf{z}$, $\boldsymbol{\theta}$, $\mathbf{b}$ and $\boldsymbol{\lambda}$. The graphical representation for variational sLDA is displayed in Fig. 5. We accordingly establish the relation [3]

$$\begin{aligned} \log p(\mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi}) &= \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\psi}, \boldsymbol{\eta}; \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi}) \\ &+ \mathcal{D}(q(\mathbf{z}, \boldsymbol{\theta}, \mathbf{b}, \boldsymbol{\lambda}|\boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\psi}, \boldsymbol{\eta}) \| p(\mathbf{z}, \boldsymbol{\theta}, \mathbf{b}, \boldsymbol{\lambda}|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi})) \end{aligned} \quad (9)$$

where $\mathcal{D}(\cdot)$ denotes the KL divergence between variational posterior and true posterior. Maximizing the lower bound $\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\psi}, \boldsymbol{\eta}; \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi})$ with respect to $\{\boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\psi}, \boldsymbol{\eta}\}$ is equivalent to minimizing the KL divergence, namely finding new variational posterior distribution $\hat{q}\left(\mathbf{z}, \boldsymbol{\theta}, \mathbf{b}, \boldsymbol{\lambda}|\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\eta}}\right)$ or its corresponding variational parameters $\left\{\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\eta}}\right\}$. VB-E step is performed. The updated variational distribution is then substituted into lower bound in Eq. 8. In VB-M step, the updated lower bound $\mathcal{L}\left(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\eta}}; \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi}\right)$ is further maximized with respect to $\{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi}\}$ to estimate new sLDA parameters $\left\{\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\pi}}\right\}$. The updating formulas for new sLDA variational parameters and model parameters are derived by

$$\hat{\phi}_{nk} \propto \beta_{kw_n}^{\psi_{nk(b=1)}} \exp\left\{\Psi(\gamma_k) - \Psi\left(\sum_{j=1}^{K} \gamma_j\right)\right\} \quad (10)$$

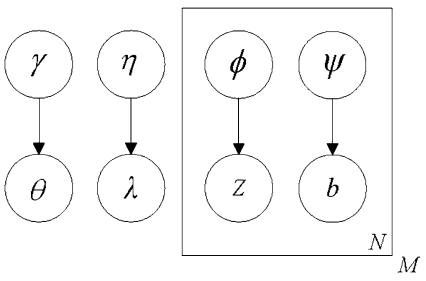


Figure 5 Graphical model for variational sLDA.

$$\hat{\gamma}_k = \alpha_k + \sum_{n=1}^{N} \phi_{nk} \tag{11}$$

$$\hat{\psi}_{nkb} \propto \beta_{kw_n}^{\phi_{nk}} \exp\left\{\Psi(\eta_{kb}) - \Psi\left(\sum_{s=1}^{2} \eta_{ks}\right)\right\} \tag{12}$$

$$\hat{\eta}_{kb} = \pi_b + \sum_{n=1}^{N} \psi_{nkb} \tag{13}$$

$$\hat{\beta}_{kv} \propto \sum_{d=1}^{M} \sum_{n=1}^{N} \psi_{dnk(b=1)} \phi_{dnk} \delta(w_{dn}, v) \tag{14}$$

$$\hat{\boldsymbol{\alpha}}^{(t+1)} = \boldsymbol{\alpha}^{(t)} - \mathbf{H}_{slda}\left(\boldsymbol{\alpha}^{(t)}\right)^{-1} \mathbf{g}_{slda}\left(\boldsymbol{\alpha}^{(t)}\right) \tag{15}$$

$$\hat{\boldsymbol{\pi}}^{(t+1)} = \boldsymbol{\pi}^{(t)} - \mathbf{H}_{slda}\left(\boldsymbol{\pi}^{(t)}\right)^{-1} \mathbf{g}_{slda}\left(\boldsymbol{\pi}^{(t)}\right) \tag{16}$$

where $\mathbf{H}_{slda}(\cdot)$ and $\mathbf{g}_{slda}(\cdot)$ denote the Hessian matrix and gradient vector of the lower bound $\mathcal{L}(\cdot)$ with respect to $\{\boldsymbol{\alpha}, \boldsymbol{\pi}\}$, respectively. After several VB-EM iterations, the variational posterior $\hat{q}\left(\mathbf{z}, \boldsymbol{\theta}, \mathbf{b}, \boldsymbol{\lambda} | \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\eta}}\right)$ is estimated and converged to true posterior $p(\mathbf{z}, \boldsymbol{\theta}, \mathbf{b}, \boldsymbol{\lambda} | \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi})$ with the smallest KL divergence.

3.4 Model Interpretation

In the proposed sLDA, beta prior distribution for parameter λ_{dk} is controlled by hyperparameter π. This distribution serves as conjugate prior to combine with Bernoulli distribution for indicator variable b_{dnk}. Similar to interpretation of LDA, Fig. 6 illustrates how a word w_{dn} is predicted through sLDA. An additional layer consisting of Bernoulli indicators $b_{dnk} = 1$ and $b_{dnk} = 0$ for individual topics k is introduced. The variational parameters $\hat{\psi}_{nk(b=1)}$ and $\hat{\psi}_{nk(b=0)}$ are seen as spike probabilities for identifying relevant topics (topics k_2 and k_3) with $b_{dnk} = 1$ and irrelevant topics $b_{dnk} = 0$ (topic k_1) for prediction of word w_{dn}, respectively. The variational parameters or topic proportions $\left\{\hat{\phi}_{nk}\right\}$ for each word w_{dn} are acted as slab probabilities. The active or relevant topics and their slab probabilities are considered for prediction of word w_{dn}. The inactive or irrelevant topic k_3 is disregarded for word prediction. We are equivalent to perform Bayesian feature selection where the semantically-meaningful words are selected as mass points to contribute for topic-dependent word probabilities $\{\hat{\beta}_{kv}\}$. The redundant features are pruned for Bayesian sparse coding. Bayesian sparse topic model is constructed accordingly. In this manner, sLDA does not only reduce redundant connection between topic k and word w_{dn} but also save considerable memory and computation costs. The estimated probability parameter $\hat{\boldsymbol{\beta}} = \left\{\hat{\beta}_{kv}\right\}$ turns out to be a sparse matrix containing quite several components with near zero values. For those components with sufficiently small values, we force them zero so that true sparsity can be achieved for the corresponding words $\{w_{dn} = v\}$ in the corresponding topics $\{k\}$. Since noisy features are removed, the ill-posed condition in data modeling is alleviated. Bayesian regularization [2] is assured for the estimated document model.

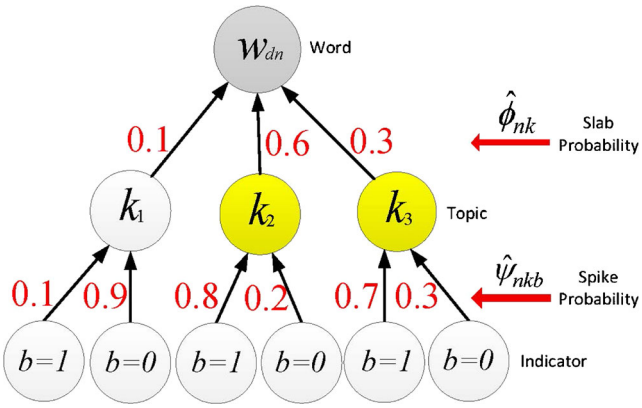


Figure 6 Illustration for sLDA in case that a word w_{dn} is assigned by three topics $\{k_1, k_2, k_3\}$. The active topics k_2 and k_3 (*marked by yellow*) and inactive topic k_1 are determined by spike probabilities $\{\hat{\psi}_{nkb}\}$. Only the active topics and their corresponding slab probabilities $\{\hat{\phi}_{nk}\}$ are considered for prediction of word w_{dn}.

3.5 Comparison with Other Methods

We compare the proposed sLDA with two related works [23, 24]. In [23], a sparse representation of documents was implemented by using finite spike distributions. The resulting sparse topic model was built according to a HDP where the approximate inference using collapsed Gibbs sampling was performed. A finite binary-valued matrix was introduced to attain sparse representation. The model complexity was reduced as well. This model is an extension of HDP based topic model by additionally decoupling sparsity and smoothness based on spike-and-slab prior. In [24], the IBP compound Dirichlet process was proposed. The infinite spike distributions were allowed for flexible document model where the number of topics was unfixed and the sparse representation was based on HDP. The infinite binary-valued matrix was estimated. The resulting topic distribution is focused on a finite subset of topics. The number of topics within a single document was practically finite although total number of topics was released to be unbounded. Using this FTM, IBP was applied to simulate the spike model and HDP was acted as the slab model. In this study, the sLDA with finite and fixed number of topics is proposed. Using sLDA, it is not necessary to additionally

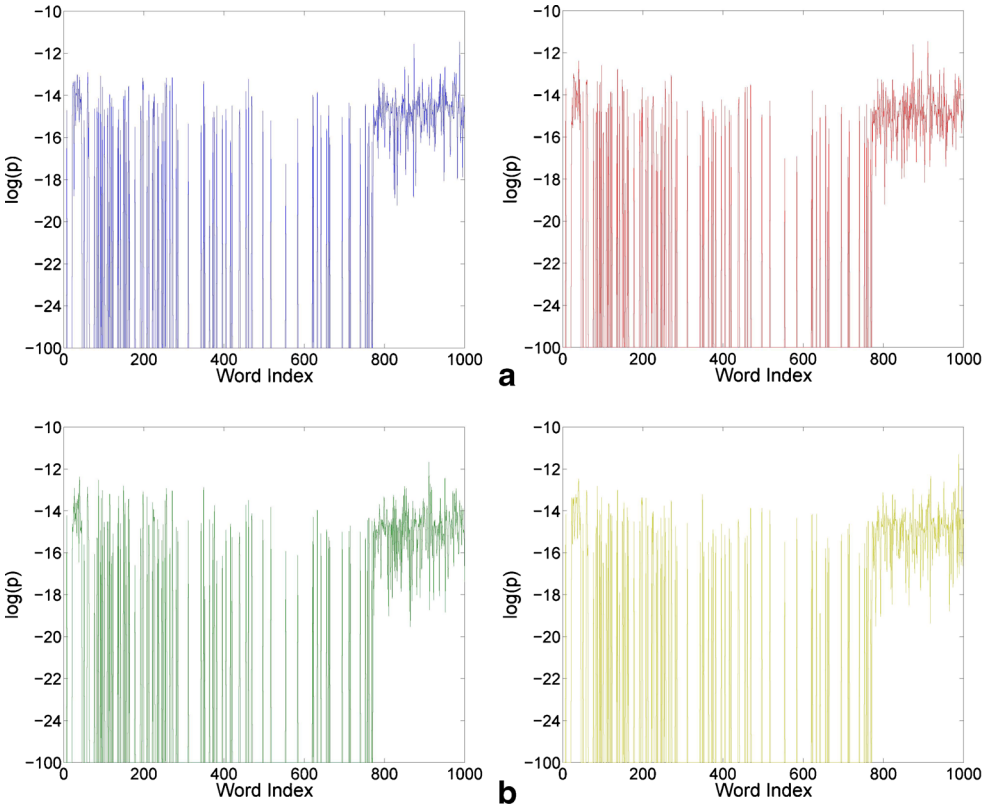


Figure 7 Comparison of logarithms of topic-dependent word probabilities for different words and two selected topics. The cases of LDA trained with (**a**) 20 topics and (**b**) 100 topics are examined.

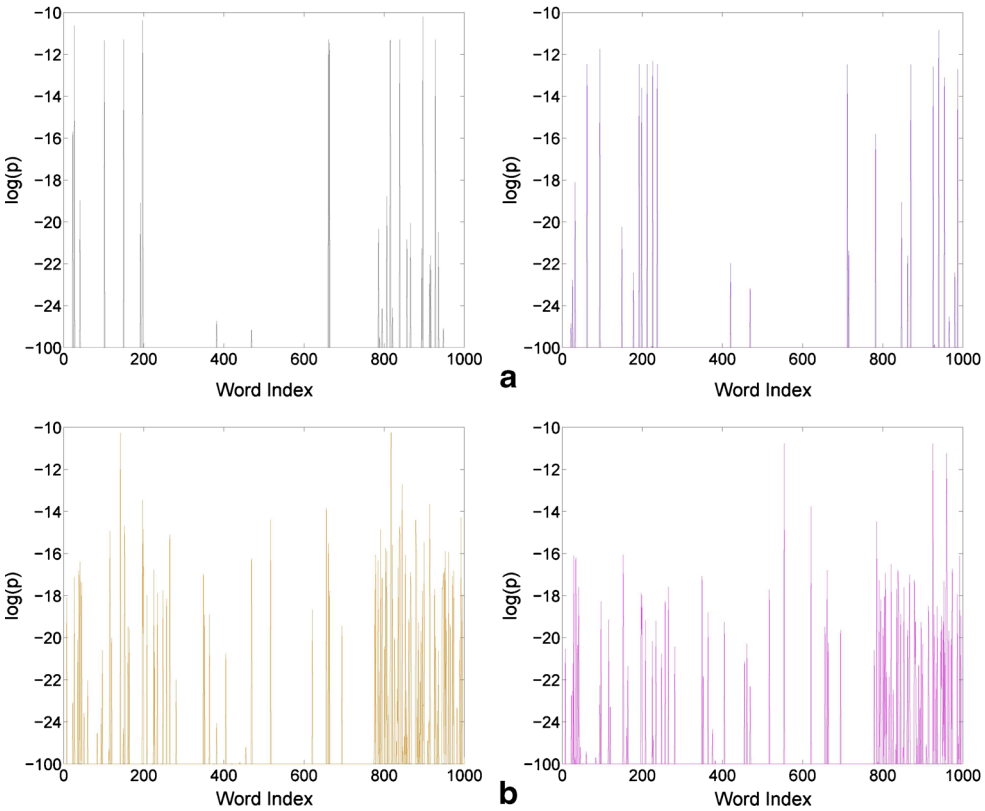


Figure 8 Comparison of logarithms of topic-dependent word probabilities for different words and two selected topics. The cases of sLDA trained with (**a**) 20 topics and (**b**) 100 topics are examined.

calculate and store a large binary-valued matrix parameter for sparse document representation. Also, instead of using Gibbs sampling procedure, the variational inference procedure based on VB-EM algorithm is implemented for sLDA. The selected features provide salient mass points to estimate the topic-dependent word distribution $\hat{\boldsymbol{\beta}}$ from training documents.

4 Experiments

4.1 Experimental Setup

In the experiments, the Associated Press newswire (AP) dataset and the Wall Street Journal (WSJ) dataset from TREC collection were used to evaluate document modeling based on LDA [3] and the proposed sLDA. The evaluation was conducted by considering three cases; AP88-90, WSJ87-89 and WSJ87-92. In AP88-90 dataset, there were 10,411 training documents randomly selected from years 1988 to 1990. Test data contained 2,000 documents. In WSJ87-89 and WSJ87-92 datasets, there were 5,075 and 15,201 training documents randomly selected from years 1987 to 1989 and 1987 to 1992, respectively. A common test set consisting of 2,008 documents was collected for evaluation using WSJ dataset. The performance of document modeling was investigated for different amount of training documents. All the documents were preprocessed by performing stemming and stop word removal. The vocabulary size V in WSJ dataset and in AP dataset was 55,106 and 73,794, respectively. In implementation of sLDA, the sparsity is controlled according to the estimated variational parameters $\hat{\boldsymbol{\psi}} = \left\{\hat{\psi}_{dnkb}\right\}$ which are associated with Bernoulli indicator variables $\{b_{dnk}\}$. The performance of different document models is evaluated by the metric of

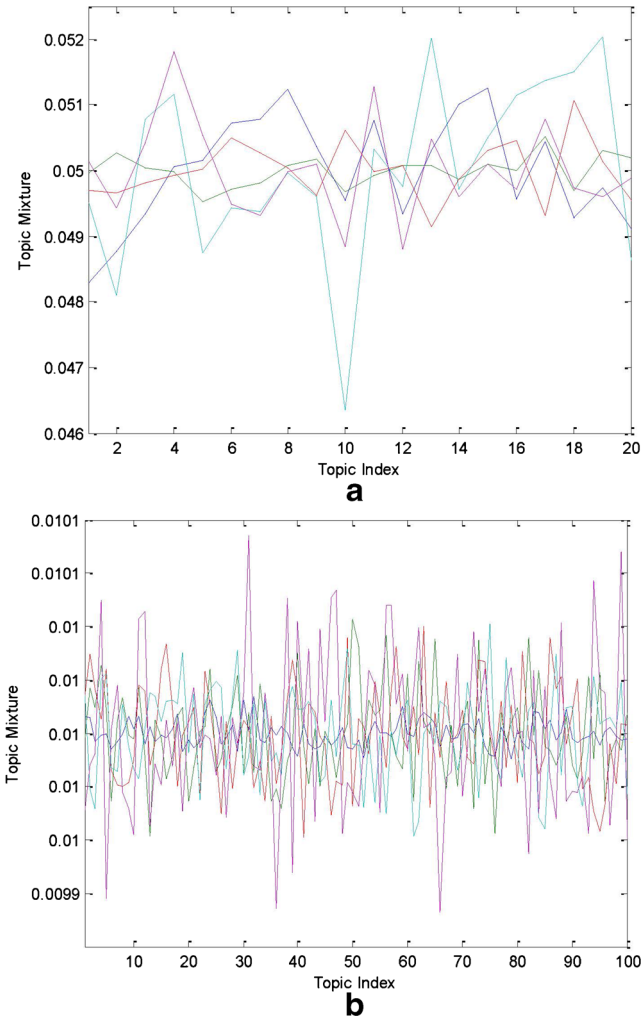


Figure 9 Comparison of topic mixtures for five selected documents (*denoted by different colors*). The LDA trained with (**a**) 20 topics and (**b**) 100 topics are examined.

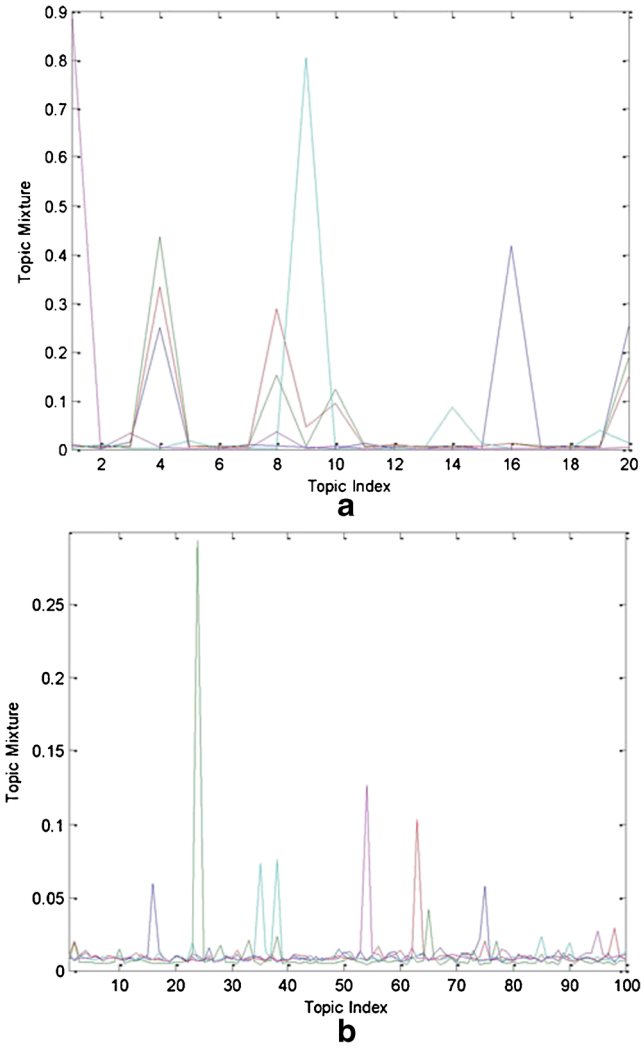


Figure 10 Comparison of topic mixtures for five selected documents (*denoted by different colors*). The sLDA trained with (**a**) 20 topics and (**b**)100 topics are examined.

perplexity which is measured from test documents $\{\mathbf{w}_d\}$ according to

$$\text{perplexity} = \exp\left\{-\frac{\sum_{d=1}^{M} \log\ p(\mathbf{w}_d))}{\sum_{d=1}^{M} N_d}\right\} \tag{17}$$

where N_d denotes the number of words in document $\{\mathbf{w}_d\}$. A lower perplexity corresponds to less confusion in the prediction of the words in text documents, or equivalently implies a better generative model. The initial values in $\boldsymbol{\beta}$ were set to be $1/V$. It was found that the perplexity was insensitive to the initial values of $\boldsymbol{\alpha}$ and $\boldsymbol{\pi}$. The initial values of the entries of $\boldsymbol{\alpha}$ and $\boldsymbol{\pi}$ were empirically set to be uniform as 100 and 10, respectively. In implementation of LDA, the source code from http://www.cs.princeton.edu/~blei/lda-c/ was referred. The memory costs for the estimated LDA parameters $\left\{\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}\right\}$ are measured as 16.37 MB and 81.83 MB for the case of 20 topics and 100 topics, respectively. Memory requirement is proportional to the number of topics.

4.2 Evaluation for the Estimated Parameters

First of all, we evaluate the estimated parameters and variational parameters of LDA and sLDA by using WSJ87-92 dataset. Figure 7a and b show the logarithms of the estimated topic-dependent word probabilities $\left\{\hat{\beta}_{kv}\right\}$ by using LDA with 20 topics and 100 topics, respectively. Only the values of the parameters from 1000 selected words and two selected topics are displayed. The words in these four subfigures are identical. The floor value of -100 is set. The values of the parameters $\left\{\hat{\beta}_{kv}\right\}$ corresponding to the same words in different topics look similar. No much discrimination is seen in the estimated parameters for different topics. Figure 8a and b show the logarithms of topic-dependent word probabilities for sLDA under 20 topics and 100 topics, respectively. The 1000 selected words are identical for eight sub-figures in Figs. 7 and 8. Comparing the estimated word probabilities from LDA (Fig. 7a and b) and sLDA (Fig. 8a and b), we find that there are many words with very low value of logarithm of topic-dependent word probability (< -100) by using sLDA. But, using LDA, only a few words have low word probabilities. The estimated word probabilities of sLDA are sparser than those of LDA. In contrast with LDA, the word probabilities using sLDA are distinct for different topics. This is because sLDA is feasible to characterize the relevance between words and topics.

Figure 9 displays the normalized values of variational parameters $\boldsymbol{\gamma} = \{\hat{\gamma}_k\}$ which correspond to the topic mixtures $\boldsymbol{\theta}$ for different topics k. LDA is applied. Each curve expresses the topic mixtures corresponding to a document and is marked by different color. These topic mixtures are used to form a coordinate vector to represent a document in topic space. There are five randomly-selected documents in this evaluation. Figure 10 shows the corresponding curves of topic mixtures where the sLDA is applied. These five documents are identical to those in Fig. 9 for LDA. Again, comparing the estimated topic mixtures using LDA and sLDA, it is obvious that the range of the estimated topic mixtures using sLDA is larger than that using LDA. Using LDA, the style of curves of different documents looks similar while that using sLDA is distinct for different documents. The topic mixtures of sLDA are sharper and sparser than those of LDA. The discrimination using sLDA is much better than that using LDA. This property is consistent for different number of topics. sLDA does effectively select relevant features for Bayesian sparse topic modeling. With the improved parameter discrimination, sLDA shall perform better than LDA for document clustering and classification.

4.3 Evaluation for the Selected Topic Words

In what follows, we examine the selected topic words in corpus level as well as in document level. In this set of evaluation, WSJ87-89 dataset was used. The case of sLDA trained with 100 topics was investigated. Table 1 lists the corpus-level topic words for each of six randomly-selected

Table 1 A list of topic words from six selected topics by using sLDA.

Topic 2	Topic 33	Topic 39	Topic 62	Topic 78	Topic 90
Stock	Car	Japan	Couple	Game	Tax
Share	Sale	Govern	Work	Team	Rate
Market	Ford	Industry	Lawyer	Sport	Card
Investor	GM	Japanese	Husband	New	Credit
Trade	Motor	Economy	Juridical	Player	Gain
Average	Auto	Country	Trial	Time	Interest
Dow	Plant	Foreign	Right	Family	Capital
Jones	Maker	Growth	Judge	Basketball	Company
Gain	Chrysler	Export	Consult	Season	Consume
Analyst	Model	Germanic	Litigant	TV	Revenue
Shearson	Vehicle	Increase	Wife	Network	Bank
Big	Truck	Minister	Open	Win	Hong-Kong
Company	Product	World	Marriage	Show	Australian
Rise	Increase	Member	Juror	Home	American
Fall	Dealer	Trade	Find	Second	New
Earn	Price	Ministry	People	National	Hold
Sell	Market	Private	Select	Coach	State
Price	Consumer	Support	Help	Program	Charge
Volume	Toyota	Action	Home	Young	Raise
Yesterday	Automotive	Policy	Job	Play	Capital-gain

topics. Top 20 words with high topic-dependent word probabilities $\left\{\hat{\beta}_{kv}\right\}$ are selected as topic words. As we can see, the 2nd topic is related to the news of stock market, the 33rd topic is related to the news of motor market, and the 78th topic is related to news of TV sports. It is obvious that the topic words within a topic are closely related and those across topics are significantly apart from each other. Unsupervised learning of topic words is well done by using sLDA model.

Next, the document-level evaluation is performed by investigating the effect of spike-and-slab model in topic-based document representation. Typically, the proposed sLDA conducts Bayesian sparse topic modeling through feature selection based on spike-and-slab model. This model

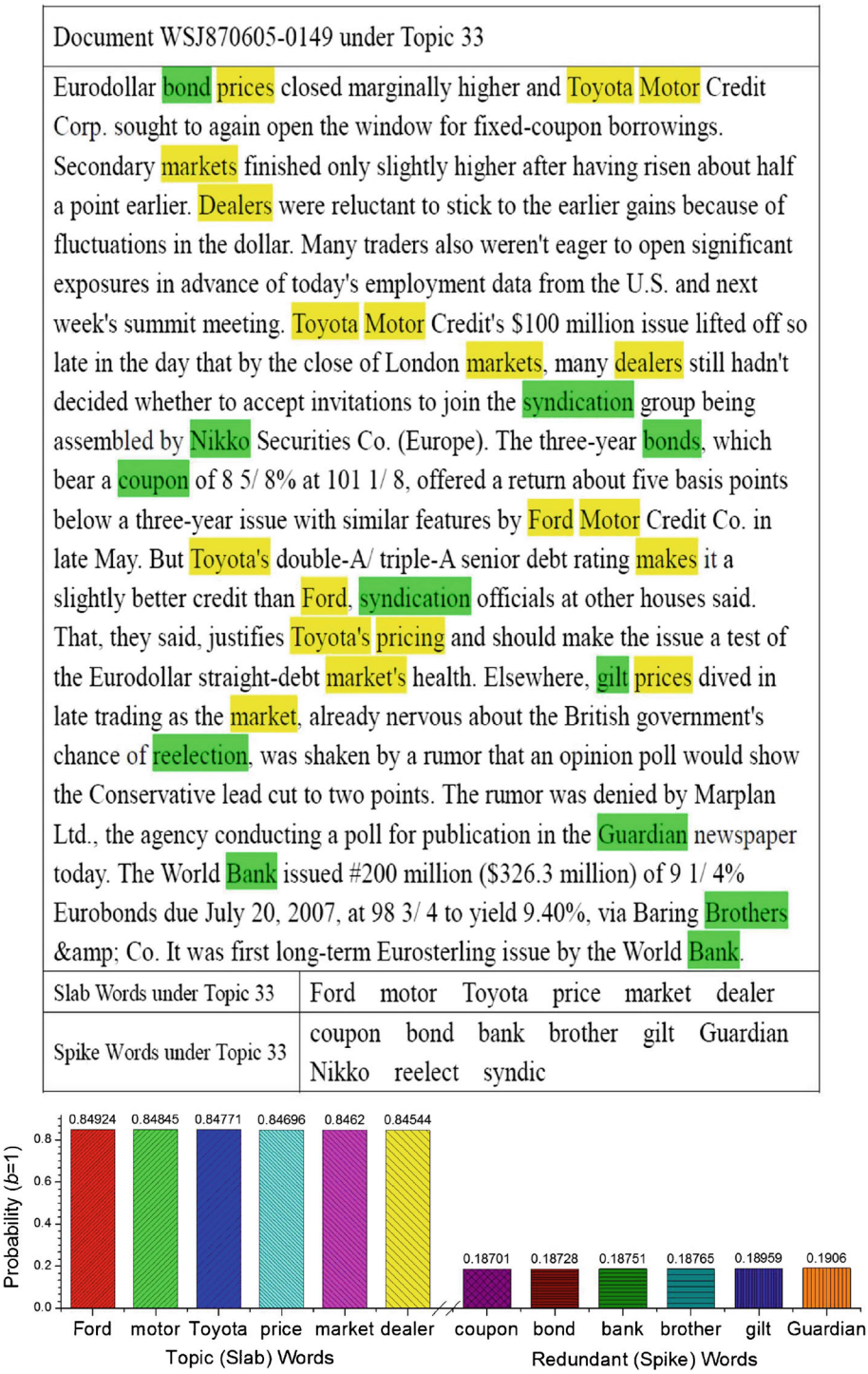


Figure 11 An example of WSJ document with topic or 'slab' words (*yellow*) and redundant or 'spike' words (*green*) under Topic 33. The estimated variational parameter $\hat{\psi}_{dnk(b=1)}$ is shown in the bottom.

is designed to select the slab or topic words with indicator variable $b_{dnk} = 1$ and prune the spike or redundant words with indicator variable $b_{dnk} = 0$. This judgement is made according to the estimated variational parameters $\left\{\hat{\psi}_{dnkb}\right\}$. Figures 11, 12 and 13 illustrate three WSJ documents and their corresponding topic words and redundant words selected by the proposed sLDA under Topics 33, 78 and 90, respectively. The estimated variational parameters $\hat{\psi}_{dnk(b=1)}$ of these words are displayed in the bottom of each figure. For example, in case of document WSJ870605-

Figure 12 An example of WSJ document with topic or 'slab' words (*yellow*) and redundant or 'spike' words (*green*) under Topic 78. The estimated variational parameter $\hat{\psi}_{dnk(b=1)}$ is shown in the bottom.

Document WSJ870604-0099 under Topic 78	
Reinhard Schmuck is getting more attention these days than during his entire college basketball career. Then again, that's not too hard. The 21-year-old accounting major played only 10 games in two seasons for Baruch College in New York City. But last month, as a lark he applied for early entry into the National Basketball Association through its June 22 draft. Only eight other players who haven't fulfilled their college eligibility did likewise. Those other players have stacked up generally impressive credentials. As for Mr. Schmuck, the 6-foot-4, 225-pound forward, who is nicknamed "Rhino," averaged only eight points a game and was sidelined repeatedly with ankle injuries. After he was hurt, he lost his mobility," says Ray Rankis, the basketball coach for Baruch. We used him for giving fouls." Mr. Schmuck says the NBA "has always been a dream and I figured this was the closest I could get." But his dream has soured slightly. He expected his draft application to draw some attention, but not the flood of phone calls he and his family have received from reporters and basketball fans. And a few wisenheimers can't resist poking fun at the family name. (He pronounces it "Schmook.") In the beginning it was humorous, but I wish all the phone calls would end," he says. Also, Mr. Schmuck had hoped his bid for the NBA would somehow help Baruch. CBS did put him on its half-time show during a recent NBA playoff game, but Mr. Schmuck didn't like the way the announcer called Baruch's basketball program "tiny." (Still, the announcer was fairly accurate; the college has had only six winning basketball teams in 19 years and its gym is so small that it must play home games at nearby Xavier High School.) To no one's surprise, none of the phone calls Mr. Schmuck received were from NBA scouts. I'd never heard of him," says Stan Kasten, general manager for the Atlanta Hawks. Mr. Schmuck concedes he never really held out much hope anyway. I'm going to be an accountant," he says. But just in case a basketball scout is wondering, Mr. Schmuck says he was "a decent high school player." And how many basketball players are called Rhino?	
Slab Words under Topic 78	basketball NBA coach college family game National play player program season win
Spike Words under Topic 78	case humorous mobility phone impress nickname

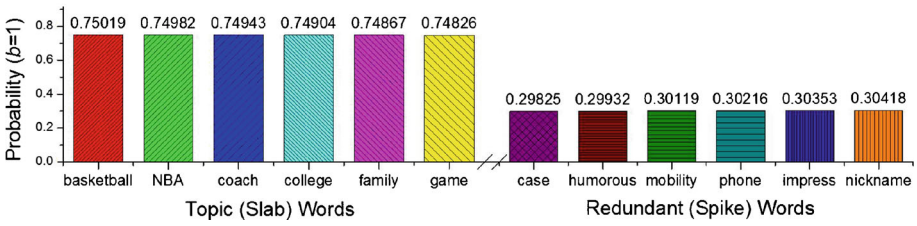


Figure 13 An example of WSJ document with topic or 'slab' words (*yellow*) and redundant or 'spike' words (*green*) under Topic 90. The estimated variational parameter $\hat{\psi}_{dnk(b=1)}$ is shown in the bottom.

Document WSJ861205-0137 under Topic 90	
Metropolitan Life Insurance Co. said it added $25 million in capital to its Metropolitan Reinsurance Co. unit. The injection boosted the property-casualty reinsurance unit's capital to $91 million. Officials said the funds will support the growth and strength of the unit. Metropolitan Reinsurance wrote policies that generated net premiums of $103 million for the first nine months, and is expected to report a profit for 1986. A number of insurers have bolstered the capital of their reinsurance units recently. Losses over the past few years cut into their capital and now, with property-casualty insurance rates having risen, they need additional capital to support their increased volume. A reinsurer participates in policies written by other insurance companies, sharing part of the risk in return for part of the premium.	
Slab Words under Topic 90	rate capital profit increase company month
Spike Words under Topic 90	bolster inject metropolitan net strength property-casualty

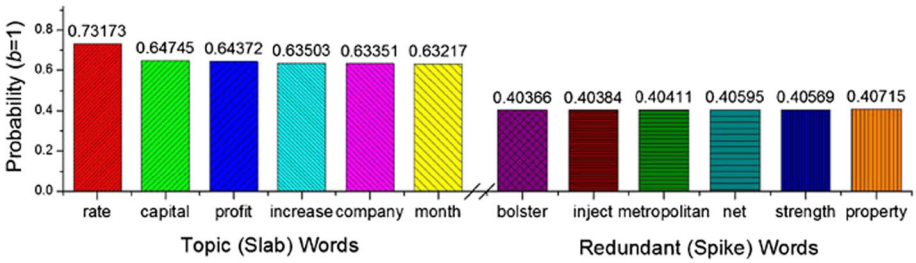
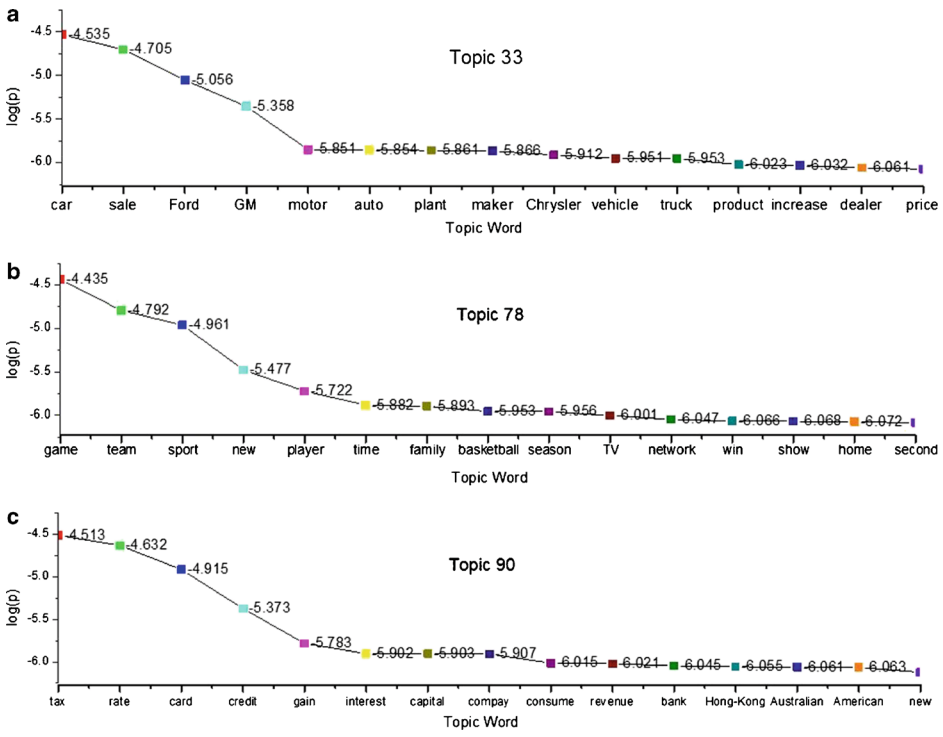


Figure 14 The logarithm of the estimated model parameter $\hat{\beta}_{kv}$ for 15 topic words in (**a**) Topic 33, (**b**) Topic 78, and (**c**) Topic 90.



0149 under Topic 33, it is meaningful that the words 'Toyota', 'market', 'dealer', 'motor', 'Ford' and 'price' are selected as topic words due to high value of $\hat{\psi}_{dnk(b=1)}$, and the words 'bond', 'syndication', 'Nikko', 'coupon', 'gilt', 'reelection', 'Guardian', 'bank' and 'Brothers' are seen as redundant words in construction of sLDA model due to low value of $\hat{\psi}_{dnk(b=1)}$. The irrelevant words are pruned by spike-and-slab model in these documents. Further, we report the quantitative measure of how topic words are related to the corresponding topics. Figure 14a–c display the logarithm of the estimated model parameter $\hat{\beta}_{kv}$ for 15 topic words in Topics 33, 78 and 90, respectively. This is a general measure which is evaluated over all training documents. The degree of relevance between individual words and the corresponding topics is clearly reflected by this measure. From these figures, we confirm that sLDA could extract semantically-rich words and avoid the interference of noise words when constructing topic model. The performance of Bayesian feature selection using sLDA is assured.

4.4 Evaluation for Perplexity versus Number of Topics

In this set of experiments, the perplexities of LDA and sLDA are carried out for different number of topics. Figures 15, 16 and 17 display the comparison of perplexity versus topic size on using AP88-90, WSJ87-89 and WSJ87-92, respectively. The cases of topic size K being 10, 20, 40, 60, 80, 100 are considered. It is consistent to see that sLDA obtains lower perplexity than LDA for different number of topics and different amount of training data in different datasets. The model perplexity is increased with more topics because more noisy document model is calculated accordingly. Selecting relevant features and removing redundant features are helpful to achieve desirable document representation. Moreover, the perplexities of LDA and sLDA of using WSJ87-92 are smaller than those of using WSJ87-89. This is because that the modeling performance of using larger dataset is better than that of using smaller dataset.

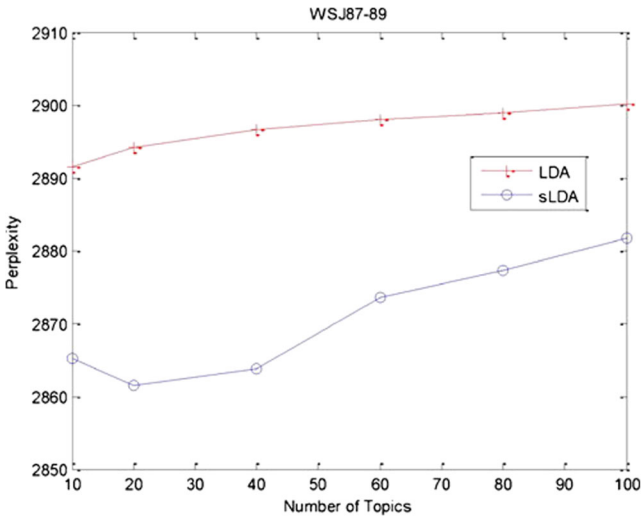


Figure 16 Comparison of perplexities of LDA and sLDA trained with different number of topics. WSJ87-89 is used.

4.5 Evaluation for Sparsity and Memory Cost

To evaluate the performance of sparse modeling, we compare the sparsity of the topic-dependent word probabilities $\hat{\boldsymbol{\beta}} = \left\{\hat{\beta}_{kv}\right\}$ estimated by LDA and sLDA. Here, the sparsity

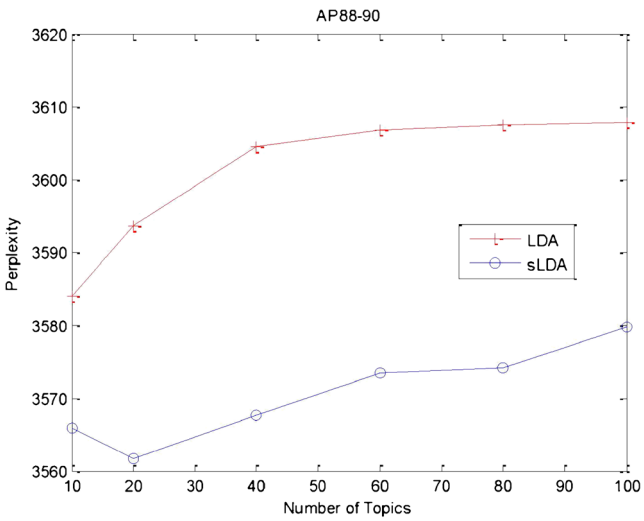


Figure 15 Comparison of perplexities of LDA and sLDA trained with different number of topics. AP88-90 is used.

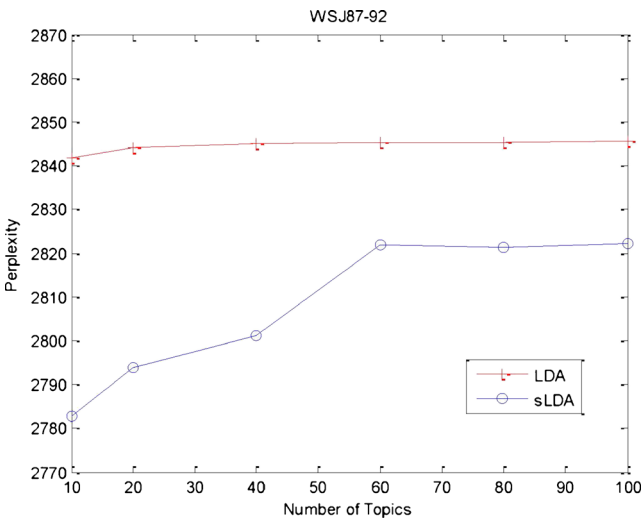


Figure 17 Comparison of perplexities of LDA and sLDA trained with different number of topics. WSJ87-92 is used.

is calculated as a ratio of the number of zero entries out of the number of total entries in $\hat{\boldsymbol{\beta}} = \left\{\hat{\beta}_{kv}\right\}$ [23]

$$\text{sparsity} = \frac{\sum_{k=1}^{K}\sum_{v=1}^{V} I\left(\hat{\beta}_{kv} = 0\right)}{V \cdot K} \tag{18}$$

where $I(\cdot)$ denotes an indicator function for counting the number of zero entries. WSJ87-89 is used. The higher the sparsity is measured, the more zero entries happen in topic-dependent word probability matrix or equivalently the smaller the computation and memory costs are spent. Figure 18 displays the sparsity of LDA and sLDA under different number of topics. Attractively, sLDA significantly increase sparsity by 81.9 % through the proposed Bayesian feature selection. The zero entries in $\hat{\boldsymbol{\beta}} = \left\{\hat{\beta}_{kv}\right\}$ using LDA are due to the unseen words while those using sLDA are not only caused by the *unseen words* but also by the *words which are pruned by feature selection procedure*. Compared to LDA, sLDA reduces the model perplexity and simultaneously improve the model sparsity.

We also examine the model complexity of sLDA according to a metric of memory cost reduction (MCR)

$$\text{MCR} = \frac{2\left[\max_{k}\sum_{v=1}^{V} I\left(\hat{\beta}_{kv} \neq 0\right)\right] \cdot K}{V \cdot K} \tag{19}$$

where the denominator denotes the total memory allocated to store all entries of $\hat{\boldsymbol{\beta}} = \left\{\hat{\beta}_{kv}\right\}$ and the numerator determines the consumption of memory cost for storing nonzero entries. The number in quotation of numerator of Eq. 19 is computed by finding the maximum number of nonzero entries in $\hat{\boldsymbol{\beta}} = \left\{\hat{\beta}_{kv}\right\}$ among different latent topics. Since we do not only store the values but also the addresses of nonzero $\hat{\beta}_{kv}$, the numerator is calculated by multiplying 2. If MCR < 1, the memory cost is reduced. Otherwise, memory cost is increased due to the additional storage of addresses of nonzero entries. Applying this scheme to store $\hat{\boldsymbol{\beta}} = \left\{\hat{\beta}_{kv}\right\}$ for LDA causes the value of MCR larger than 1 and even near 2. This is because that only few entries in LDA parameter $\hat{\boldsymbol{\beta}} = \left\{\hat{\beta}_{kv}\right\}$ are zero. But, this circumstance is changed by using sLDA. Figure 19 illustrates the memory cost reduction based on sLDA. The reduction is elevated by increasing number of topics. This phenomenon is obvious. From these results, we confirm the superiority of sLDA to LDA in terms of perplexity, sparsity and memory cost.

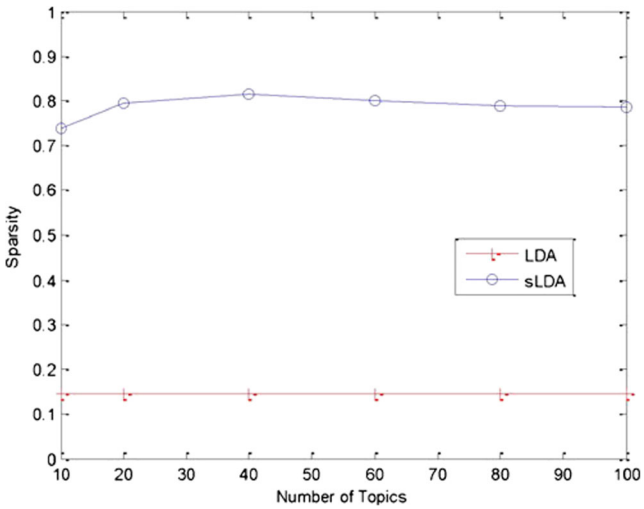


Figure 18 Comparison of sparsity of topic-dependent word probability matrices estimated by LDA and sLDA for different number of topics.

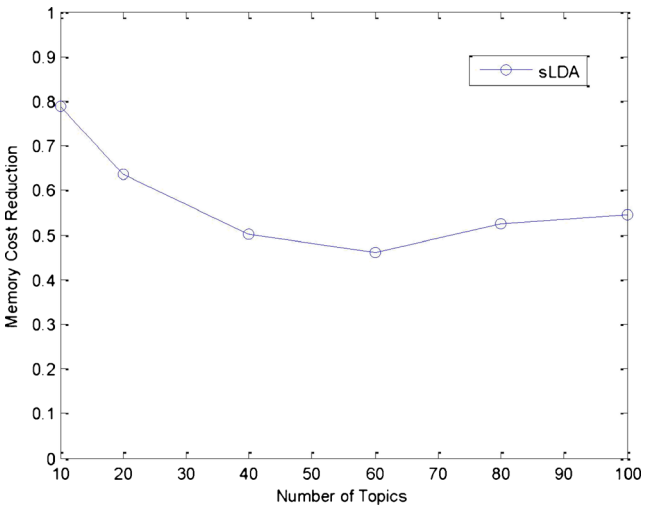


Figure 19 Memory cost reduction versus number of topics by using sLDA.

5 Conclusions

This paper developed a new framework of Bayesian feature selection for building sparse topic model. The spike-and-slab distribution was introduced to select informative features and accordingly improve the document representation and simultaneously reduce the memory and computation costs. The estimated sLDA parameters and hyperparameters were illustrated to be sparse in document level as well as in corpus level. The experiments on text modeling over different number of topics and different amount of training data in different datasets demonstrated that the proposed sLDA attained lower perplexity, higher sparsity and larger memory cost reduction compared to standard LDA. The topic words and redundant words were properly captured by using sLDA. In the future, exploring the suitable number of topics or controlling the model structure from training data shall be investigated for sLDA.

Acknowledgments This work was supported in part by the National Science Council, Taiwan, under Contract NSC 100-2221-E-009-153-MY3.

References

1. Babacan, S.D., Molina, R., Katsaggelos, A.K. (2010). Bayesian compressive sensing using Laplace priors. *IEEE Transactions on Image Processing*, *19*(1), 53–63.
2. Bishop, C.M. (2006). *Pattern recognition and machine learning*. New York: Springer Science.
3. Blei, D., Ng, A., Jordan, M.I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*(5), 993–1022.
4. Chang, Y.-L., & Chien, J.-T. (2009). Latent Dirichlet learning for document summarization. In *Proceedings of international conference on acoustics, speech, and signal processing (ICASSP)* (1689–1692).
5. Chang, Y.-L., Lee, K.-F., Chien, J.-T. (2011). Bayesian feature selection for sparse topic model. In *Proceedings of IEEE international workshop on machine learning for signal processing* (pp. 1–6).
6. Chien, J.-T., & Chueh, C.-H. (2011). Dirichlet class language models for speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, *19*(3), 482–495.
7. Chueh, C.-H., & Chien, J.-T. (2008). Reliable feature selection for language model adaptation. In *Proceedings of international conference on acoustics, speech, and signal processing (ICASSP)* (pp. 5089–5092).
8. Doshi-Velez, F., Miller, K.T., Van Gael, J., Teh, Y.W. (2009). Variational inference for the Indian buffet process. In *Proceedings of artificial intelligence and statistics*.
9. Ghahramani, Z., Griffiths, T.L., Sollich, P. (2007). Bayesian nonparametric latent feature models. *Bayesian Statistics*, *8*, 201–225.
10. Gorur, D., Jakel, F., Rasmussen, C.E. (2006). A choice model with infinitely many latent features. In *Proceedings of the international conference on machine learning* (pp. 361–368).
11. Griffiths, T.L., & Ghahramani, Z. (2005). Infinite latent feature models and the Indian buffet process. In *Advances in neural information processing systems (NIPS)* (Vol. 18).
12. Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of international ACM SIGIR conference on research and development in information retrieval* 50–57.
13. Ishwaran, H., & Rao, J.S. (2005). Spike and slab variable selection: frequentist and Bayesian strategies. *Annals of Statistics*, *33*(2), 730–773.
14. Meeds, E., Ghahramani, Z., Neal, R.M., Roweis, S.T. (2007). Modeling dyadic data with binary latent factors. In *Advances in neural information processing systems (NIPS)* (Vol. 19).
15. Mimno, D., Hoffman, M.D., Blei, D.M. (2012). Sparse stochastic inference for latent Dirichlet allocation. *International Conference on Machine Learning*.
16. Mitchell, T.J., & Beauchamp, J.J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, *83*(404), 1023–1032.
17. Mohamed, S., Heller, K., Ghahramani, Z. (2010). Sparse exponential family latent variable models. In *Proceedings of NIPS workshop on practical applications of sparse modeling: Open issues and new directions*.
18. O'Hara, R.B., & Sillanpaa, M.J. (2009). A review of Bayesian variable selection methods: what, how and which. *Bayesian Analysis*, *4*(1), 85–118.
19. Saon, G., & Chien, J.-T. (2012). Bayesian sensing hidden Markov models. *IEEE Transactions on Audio, Speech and Language Processing*, *20*(1), 43–54.
20. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, *101*(476), 1566–1581.
21. Teh, Y.W., & Gorur, D. (2009). Indian buffet processes with power-law behavior. In *Advances in neural information processing systems* (Vol. 22, pp. 1838–1846).
22. Tipping, M.E. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, *1*, 211–244.
23. Wang, C., & Blei, D.M. (2009). Decoupling sparsity and smoothness in the discrete hierarchical Dirichlet process. In *Advances in neural information processing systems (NIPS)* (Vol. 22).
24. Williamson, S., Wang, C., Heller, K.A., Blei, D.M. (2010). The IBP compound Dirichlet process and its application to focused topic modeling. In *Proceedings of international conference on machine learning*.

Jen-Tzung Chien received his Ph.D. degree from the National Tsing Hua University, Hsinchu, Taiwan, in 1997. During 1997-2012, he was with the National Cheng Kung University, Taiwan. Since 2012, he has been with the Department of Electrical and Computer Engineering, National Chiao Tung University, Hsinchu, where he is currently a Distinguished Professor. He held the Visiting Researcher positions at the Panasonic Technologies Inc., Santa Barbara, CA, the Tokyo Institute of Technology, Tokyo, Japan, the Georgia Institute of Technology, Atlanta, GA, the Microsoft Research Asia, Beijing, China, and the IBM T. J. Watson Research Center, Yorktown Heights, NY. His research interests include machine learning, information retrieval, speech recognition, blind source separation and face recognition.

Dr. Chien is an IEEE senior member. He served as the associate editor of the IEEE Signal Processing Letters, in 2008-2011, and the tutorial speakers of the ICASSP, in 2012 and the Interspeech, in 2013. He is appointed as the APSIPA Distinguished Lecturer for 2012-2013. He was a co-recipient of the Best Paper Award of the IEEE Automatic Speech Recognition and Understanding Workshop in 2011. He received the Ta-You Wu Memorial Award from the National Science Council (NSC), Taiwan, in 2003, the Research Award for Junior Research Investigators from Academia Sinica, Taiwan, in 2004, and the NSC Distinguished Research Awards, in 2006 and 2010.

Ying-Lan Chang received the M.E. degree from the National University of Tainan, Tainan, Taiwan. She is currently pursuing the Ph.D. degree in the Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan. Her research interests include machine learning, Bayesian inference and information retrieval.