Expert Systems With Applications 131 (2019) 208–218

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Partially collapsed Gibbs sampling for latent Dirichlet allocation

Hongju Park[a,1], Taeyoung Park[b,1], Yung-Seop Lee[c,*]

[a] Department of Statistics, University of Georgia, Athens, GA 30602, USA
[b] Department of Applied Statistics, Yonsei University, Seoul 03722, Korea
[c] Department of Statistics, Dongguk University-Seoul, Seoul 04620, Korea

ARTICLE INFO

Article history:
Received 12 September 2018
Revised 14 April 2019
Accepted 14 April 2019
Available online 17 April 2019

MSC:
62-XX
97-XX

Keywords:
Bayesian analysis
Latent Dirichlet allocation
Dirichlet process mixture
Partial collapse
Machine learning
Natural language processing

ABSTRACT

A latent Dirichlet allocation (LDA) model is a machine learning technique to identify latent topics from text corpora within a Bayesian hierarchical framework. Current popular inferential methods to fit the LDA model are based on variational Bayesian inference, collapsed Gibbs sampling, or a combination of these. Because these methods assume a unimodal distribution over topics, however, they can suffer from large bias when text corpora consist of various clusters with different topic distributions. This paper proposes an inferential LDA method to efficiently obtain unbiased estimates under flexible modeling for heterogeneous text corpora with the method of partial collapse and the Dirichlet process mixtures. The method is illustrated using a simulation study and an application to a corpus of 1300 documents from neural information processing systems (NIPS) conference articles during the period of 2000–2002 and British Broadcasting Corporation (BBC) news articles during the period of 2004–2005.

1. Introduction

There is increasing availability of collections of discrete data, such as text corpora, in many fields. A latent Dirichlet allocation (LDA) model (Blei, Ng, & Jordan, 2003) is a hierarchical Bayesian model used to identify latent topics underlying collections of discrete data. In the context of text modeling, LDA represents each text corpus, called a document set, as mixtures of latent topics that generate words with certain probabilities. Thus, LDA is very useful for document clustering, information retrieval, etc. The LDA model can be also used to identify latent characteristics among collections of general discrete data. For example, White, Harwin, Holderbaum, and Johnson (2015) used LDA to extract hidden patterns in eating behaviors, which can be regarded as topics, by treating eating events and food names as documents and words, respectively. Based on 7500 OSAKA university students, Fujimoto, Etoh, Kinno, and Akinaga (2011) used LDA to extract their university website interests, where students and web searching behaviors corresponded to documents and words, respectively. The results showed that student interests consisted of topics such as technology oriented, job hunting, and SNS addict.

To date, though many language models and the corresponding computational techniques such as deep learning have been invented, LDA has been functioning as a representative model for text and other categorical data analysis. This is because LDA not only has a great performance and a flawless probabilistic structure under the bag of words assumption, but gives intuitively interpretable output for analysis. For these reasons, many language models have been developed by referring to LDA rather than replacing it. Due to the bag of words assumption, however, it has inherent limitation for word embedding problems where the order of words matters. Instead, other language models take advantage of LDA to embed words. Mikolov and Zweig (2012) proposed a language model with recurrent neural networks (RNN). This model computes a context vector on the basis of the sentence history combined with the information given by LDA. In addition, Liu, Liu, Chua, and Sun (2015) suggested a framework for topical word embeddings utilizing results from LDA. This framework enhances its prediction performance by making the best use of topical information from LDA.

The LDA model has been most commonly used in topic modeling, where each document is represented as a probabilistic distribution over latent topics, but its current inference procedures, variational Bayesian (VB) inference and Gibbs sampling, suffer from

* Corresponding author.
E-mail addresses: hp97161@uga.edu (H. Park), tpark@yonsei.ac.kr (T. Park), yung@dongguk.edu (Y.-S. Lee).

[1] Hongju Park and Taeyoung Park contributed equally to this paper.

https://doi.org/10.1016/j.eswa.2019.04.028
0957-4174/

two major limitations: computational speed and biased parameter estimation. Many studies have investigated methods to mitigate these limitations.

Regarding computational speed, Newman, Smyth, Welling, and Asuncion (2008) and Newman, Asuncion, Smyth, and Welling (2009) proposed a distributed algorithm for fast LDA model computation by partitioning data across separate processors and allowing processors to concurrently perform Gibbs sampling over local data, followed by a global update of topic counts. Wang, Bai, Stanton, Chen, and Chang (2009) advanced this approach by implementing parallel computation for the LDA model using Message-passing interface (MPI) (Gropp, Gropp, Lusk, & Skjellum, 1999) and MapReduce (Dean & Ghemawat, 2008).

Regarding large bias in parameter estimation, Griffiths and Steyvers (2004) proposed a probabilistic inference procedure using the strategy of a collapsed Gibbs (CG) sampler (Liu, 1994) to overcome problems due to using an approximate posterior distribution for inference. Compared to previous inference procedures, the CG method provided superior accuracy for large corpora that required a large number of topics. However, despite its accuracy and simplicity by marginalizing over latent parameters, CG is not flexible enough to deal with multimodal structures of a topic distribution due to assuming that the topic distribution follows a single symmetric Dirichlet distribution and fixing topic distribution parameters to some constants in advance. Hence, the probabilistic LDA method is restrictive and can result in biased estimators with low accuracy. To capitalize on the decisiveness of VB and the accuracy of CG, Teh, Newman, and Welling (2007) proposed a collapsed variational Bayesian (CVB) inference algorithm for LDA that reduced large bias incurred by VB and provided an efficient and deterministic inference procedure in a collapsed space. However, CVB still suffers from poor accuracy because it is based on an approximate posterior distribution and fixes the values of hyperparameters in advance. Thus, it is critical to provide unbiased and efficient parameter estimators even when the true topic distribution is high-dimensional unimodal.

To reduce bias in latent parameter estimation of LDA, Blei et al. (2003) suggested to use the Newton's method to optimize hyperparameters. Because bias in latent parameters estimated by VB causes unstable inference on hyperparameters, however, the values of hyperparameters are typically fixed as constants rather than estimated. This strategy for hyperparameters is also used in CG (Griffiths & Steyvers, 2004), so that the performance of CG naturally depends on the values of hyperparameters. The hierarchical distributed (HD) latent Dirichlet allocation (Newman et al., 2009; Newman et al., 2008) utilized the Minka's method (Minka, 2000) to estimate hyperparameters in the Monte Carlo EM (MCEM) algorithm (Wei & Tanner, 1990). The HD method is, however, not a fully probabilistic model and its estimation is not reliable.

With these motivations, we focus on unbiased parameter estimation for the latent topic distribution with exact estimation of hyperparameters by using an iterative algorithm, and tried to reduce total computation time by improving the convergence characteristics of the iterative algorithm. Such estimation should accommodate documents with a mixture of topic distributions, and various document types. To achieve this, we capitalize on a Dirichlet process (DP) mixture model to flexibly represent latent topic distributions among documents. Some related models that use the DP mixture for the LDA model have been proposed previously. Teh, Jordan, Beal, and Blei (2006) developed a hierarchical Dirichlet process that could be applied to topic modeling, where the number of topics need not be set in advance and documents could be clustered with shared topic atoms. Rodriguez, Dunson, and Gelfand (2008) proposed a nested Dirichlet process that allowed document level distributions to be shared across documents, simultaneously clustering the documents. Blei, Griffiths, Jordan, and Tenenbaum (2004) proposed a nested Chinese restaurant process, extending a Chinese restaurant process for hierarchical topic modeling and allowing a tree structure for underlying topics. Paisley, Wang, Blei, and Jordan (2015) proposed a nested hierarchical Dirichlet process that generalizes the nested Chinese restaurant process using nonparametric clustering with a single path for a document in a shared tree, which allows documents to straddle multiple word specific paths.

Although these methods can effectively deal with a mixture of topic distributions present in a large corpus, interpretation of the results may not be simple compared the original LDA model outputs. Thus, we keep the topic model simple and add a hierarchical DP structure to the original LDA structure. Hence, we modify the probabilistic LDA model to effectively deal with highly multimodal topic distributions and to avoid the need to fix hyperparameter values in advance even when a topic distribution is unimodal. This strategy may appear simple to implement, but the enhanced LDA model cannot fit with current inference procedures and requires the method of partial collapse (van Dyk & Park, 2008; Park & van Dyk, 2009) to deal with the functionally incompatible set of conditional distributions. Thus, the main goal of this study is to develop an efficient and feasible inference procedure to yield unbiased parameter estimation for the enhanced LDA model to address highly multimodal latent topic distributions as well as unimodal ones, without the need to fix topic distribution parameters to some constants in advance.

The remainder of this article is organized as follows. Section 2 describes the proposed enhanced LDA model and its efficient inference procedure using the method of partial collapse. Section 3 details a simulation study to validate the proposed method performance compared with current inference procedures for an LDA model. Section 4 applies the proposed methodology to a corpus of 1300 documents from the neural information processing systems (NIPS) conference articles during the period of 2000–2002 and British Broadcasting Corporation (BBC) news articles during the period of 2004–2005, compared with current inference procedure performance in terms of likelihood and perplexity. Section 5 summarizes and concludes the paper.

2. Model description

2.1. Model specification

For clarity, we specify our proposed model in a newspaper setting. A *word* represents the basic unit of discrete data, a *news article* is a sequence of words, and a *newspaper* has various *sections* each containing a number of news articles. The section a news article belongs to is a natural candidate for the article topic. Thus, if all news articles are well separated by their sections, a single distribution of topic contributions to a news article is sufficient to identify how a news article is related to a set of topics. This assumption is made by the probabilistic LDA model (Griffiths & Steyvers, 2004), where both topic contributions for a given document and word contributions for a given topic are assumed to follow symmetric Dirichlet prior distributions, each with a single fixed constant hyperparameter. However, a news article in one section could be linked to an article in another section. For example, a news article about Brexit in a politics section could be related to an article about the global economy affected by Brexit in an economy section. Hence the two articles share the Brexit topic, although their natural topics, politics and economy from their sections, are different. This suggests that a single prior distribution of topic contributions to a document in an LDA model can be restrictive and should be generalized to a mixture prior distribution. In this way, we can investigate how a news article is related to a set of topics in several different ways.

In addition, the values of hyperparameters in the probabilistic LDA model are fixed in advance rather than estimated from data (Griffiths & Steyvers, 2004). Because the performance of the probabilistic LDA model depends on the choice of hyperparameters and failure to find the optimal value can produce large bias in parameter estimation, hyperparameters play a key role. To circumvent the problem, hyperparameters can be treated as random and then hyperprior distributions are imposed on the hyperparameters. In this way, we allow data to inform about the optimal choice of hyperparameters for robust results.

Specifically, the probabilistic LDA model is expressed as

$$\begin{aligned} w_n^{(d)}|(z_n^{(d)}, \boldsymbol{\phi}) &\overset{\text{ind}}{\sim} \text{Discrete}(\boldsymbol{\phi}^{(z_n^{(d)})}),\ n = 1, \ldots, N_d,\ d = 1, \ldots, D, \\ z_n^{(d)}|\boldsymbol{\theta} &\overset{\text{ind}}{\sim} \text{Discrete}(\boldsymbol{\theta}^{(d)}),\ n = 1, \ldots, N_d,\ d = 1, \ldots, D, \\ \boldsymbol{\theta}^{(d)}|\alpha_d &\overset{\text{ind}}{\sim} \text{Dirichlet}(\alpha_d),\ d = 1, \ldots, D, \\ \boldsymbol{\phi}^{(t)}|\beta_t &\overset{\text{ind}}{\sim} \text{Dirichlet}(\beta_t),\ t = 1, \ldots, T, \end{aligned} \tag{1}$$

where a corpus $\mathbf{W} = \{\mathbf{W}^{(d)}\}_{d=1}^{D}$ of D documents each with N_d words has M unique words and T latent topics, with $\mathbf{W}^{(d)} = \{w_n^{(d)}\}_{n=1}^{N_d}$, $w_n^{(d)}$ denotes the nth out of N_d words in document d, $z_n^{(d)}$ denotes the topic index for the nth word among T latent topics in document d, with $\mathbf{Z}^{(d)} = \{z_n^{(d)}\}_{n=1}^{N_d}$ and $\mathbf{Z} = \{\mathbf{Z}^{(d)}\}_{d=1}^{D}$, $\phi_n^{(t)}$ denotes the probability of word n under topic t, with $\boldsymbol{\phi}^{(t)} = (\phi_1^{(t)}, \ldots, \phi_M^{(t)})^\top$ and $\boldsymbol{\phi} = (\boldsymbol{\phi}^{(1)}, \ldots, \boldsymbol{\phi}^{(T)})^\top$, $\theta_t^{(d)}$ denotes the probability of topic t in document d, with $\boldsymbol{\theta}^{(d)} = (\theta_1^{(d)}, \ldots, \theta_T^{(d)})^\top$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(D)})^\top$, α_d and β_t are the scalar parameters of symmetric Dirichlet distributions for $\boldsymbol{\theta}^{(d)}$ and $\boldsymbol{\phi}^{(t)}$, respectively. In Griffiths and Steyvers (2004), $\boldsymbol{\theta}^{(d)}$ and $\boldsymbol{\phi}^{(t)}$ are assumed to follow symmetric Dirichlet distributions with hyperparameters fixed at $\alpha_d = 50/T$ and $\beta_t = 0.1$, respectively, hence the shape of each prior distribution was assumed to be a symmetric multidimensional parabola. When documents are from different categories, however, some categories may have underlying characteristics in common and such a unimodal prior distribution may not be appropriate to fully identify how a document is related to a set of topics. The fixed choice of hyperparameters can prevent from flexibly investigating the parameter space of $\boldsymbol{\theta}^{(d)}$ and $\boldsymbol{\phi}^{(t)}$.

To effectively deal with potentially multimodal topic distributions and avoid bias in parameter estimation, we propose to impose an additional probabilistic structure on the hyperparameters by using a DP mixture model. Specifically, we have

$$\begin{aligned} \alpha_d|G &\overset{\text{iid}}{\sim} G,\ d = 1, \ldots, D, \\ G|(\gamma, \lambda) &\sim \text{DP}(\gamma G_0(\lambda)), \\ \beta_t|P &\overset{\text{iid}}{\sim} P,\ t = 1, \ldots, T, \\ P|(\eta, \mu) &\sim \text{DP}(\eta P_0(\mu)), \end{aligned} \tag{2}$$

where the α_d's are independent and identically distributed according to an unknown prior distribution G, which are generated from DP with base distribution G_0 on $\Re_+$ and precision parameter γ, and the β_t's are independent and identically distributed according to an unknown prior distribution P, which are generated from DP with base distribution P_0 on $\Re_+$ and precision parameter η. We let G_0 and P_0 follow independent exponential distributions with rates λ and μ, respectively.

The stick-breaking representation (Sethuraman, 1994) of DP implies that G and P can be expressed as

$$\begin{aligned} G(\cdot) &= \sum_{i=1}^{\infty} \pi_i^{\alpha}(\mathbf{U}^{\alpha}) \delta_{\alpha_i^*}(\cdot), \quad \alpha_i^* \overset{\text{iid}}{\sim} G_0(\lambda) \\ P(\cdot) &= \sum_{j=1}^{\infty} \pi_j^{\beta}(\mathbf{U}^{\beta}) \delta_{\beta_j^*}(\cdot), \quad \beta_j^* \overset{\text{iid}}{\sim} P_0(\mu), \end{aligned} \tag{3}$$

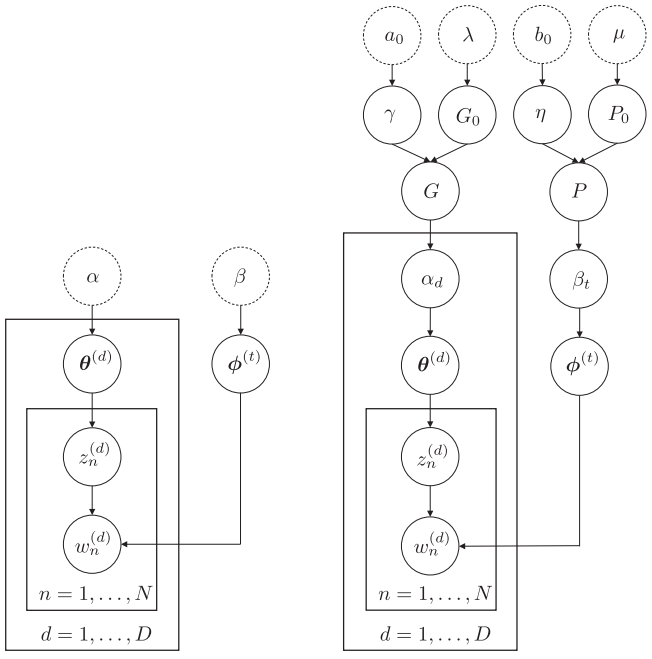

Fig. 1. Comparison of (left) probabilistic LDA model and (right) proposed enhanced LDA model.

where $\pi_i^{\alpha}(\mathbf{U}^{\alpha}) = U_i^{\alpha} \prod_{l<i}(1 - U_l^{\alpha})$, with $U_i^{\alpha}|\gamma \overset{\text{iid}}{\sim} \text{Beta}(1, \gamma)$, represents a probability weight assigned to atom α_i^* generated from an independent and identical distribution G_0, $\pi_j^{\beta}(\mathbf{U}^{\beta}) = U_j^{\beta} \prod_{l<j}(1 - U_l^{\beta})$, with $U_j^{\beta}|\eta \overset{\text{iid}}{\sim} \text{Beta}(1, \eta)$, represents a probability weight assigned to atom β_j^* generated from an independent and identical distribution P_0, and $\delta_{\xi}(\cdot)$ is a probability measure concentrated on ξ.

Based on the constructive stick-breaking representation for the unknown prior distribution G, the resulting prior distribution of topic contributions $\boldsymbol{\theta}^{(d)}$ for a given document can be expressed as a DP mixture of symmetric Dirichlet distributions, i.e.,

$$\begin{aligned} p(\boldsymbol{\theta}^{(d)}) &= \int p(\boldsymbol{\theta}^{(d)}|\alpha_d) dG(\alpha_d) \\ &= \sum_{i=1}^{\infty} \pi_i^{\alpha}(\mathbf{U}^{\alpha}) p(\boldsymbol{\theta}^{(d)}|\alpha_i^*). \end{aligned} \tag{4}$$

Hence, the topic contribution distribution for a given document is flexibly modeled as a mixture distribution. Similarly, the prior distribution of word contributions $\boldsymbol{\phi}^{(t)}$ for a given topic can be represented as a DP mixture of symmetric Dirichlet distributions, thereby adding flexibility to the probabilistic LDA model. The graphical representation of the probabilistic LDA model and the proposed enhanced LDA model is shown in Fig. 1, where solid circles represent random variables and dashed circles represent constants.

To facilitate estimating the enhanced LDA model, we introduce a set of latent variables, $\mathbf{S}^{\alpha} = \{S_d^{\alpha}\}_{d=1}^{D}$ and $\mathbf{S}^{\beta} = \{S_t^{\beta}\}_{t=1}^{T}$, such that $S_d^{\alpha} = i$ indicates that α_d belongs to document cluster i and $S_t^{\beta} = j$ indicates that β_t belongs to topic cluster j. Applying an almost sure truncation (Ishwaran & James, 2001) to the stick-breaking representations of G and P, i.e., $U_l^{\alpha} = 1$ and $U_J^{\beta} = 1$, respectively, the enhanced LDA model can be expressed as

$$\begin{aligned} w_n^{(d)}|(z_n^{(d)}, \boldsymbol{\phi}) &\overset{\text{ind}}{\sim} \text{Discrete}(\boldsymbol{\phi}^{(z_n^{(d)})}),\ n = 1, \ldots, N_d,\ d = 1, \ldots, D, \\ z_n^{(d)}|\boldsymbol{\theta} &\overset{\text{ind}}{\sim} \text{Discrete}(\boldsymbol{\theta}^{(d)}),\ n = 1, \ldots, N_d,\ d = 1, \ldots, D, \end{aligned}$$

$$
\begin{aligned}
\boldsymbol{\phi}^{(t)}|(\boldsymbol{\beta}^*, \mathbf{S}^\beta) &\overset{\text{ind}}{\sim} \text{Dirichlet}(\beta^*_{S_t^\beta}),\ t = 1, \ldots, T,\\
S_t^\beta|\mathbf{U}^\beta &\overset{\text{iid}}{\sim} \textstyle\sum_{j=1}^{J} \pi_j^\beta(\mathbf{U}^\beta)\delta_j(S_t^\beta),\ t = 1, \ldots, T,\\
U_j^\beta|\eta &\overset{\text{iid}}{\sim} \text{Beta}(1, \eta),\ j = 1, \ldots, J-1,\\
\beta_j^*|\mu &\overset{\text{iid}}{\sim} \text{Exp}(\mu),\ j = 1, \ldots, J,\\
\boldsymbol{\theta}^{(d)}|(\boldsymbol{\alpha}^*, \mathbf{S}^\alpha) &\overset{\text{ind}}{\sim} \text{Dirichlet}(\alpha^*_{S_d^\alpha}),\ d = 1, \ldots, D,\\
S_d^\alpha|\mathbf{U}^\alpha &\overset{\text{iid}}{\sim} \textstyle\sum_{i=1}^{I} \pi_i^\alpha(\mathbf{U}^\alpha)\delta_i(S_d^\alpha),\ d = 1, \ldots, D,\\
U_i^\alpha|\gamma &\overset{\text{iid}}{\sim} \text{Beta}(1, \gamma),\ i = 1, \ldots, I-1,\\
\alpha_i^*|\lambda &\overset{\text{iid}}{\sim} \text{Exp}(\lambda),\ i = 1, \ldots, I,\\
(\gamma, \eta) &\sim \text{Exp}(a_0)\text{Exp}(b_0),
\end{aligned}
\tag{5}
$$

where the values of $(\lambda,\ \mu,\ a_0,\ b_0)$ are chosen as $\lambda = 1,\ \mu = 10,\ a_0 = 1$, and $b_0 = 1$.

To identify latent topics, we need to estimate the parameters, $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, given the observed data $\mathbf{W}$. This goal can be achieved by simulating the target posterior distribution $p(\mathbf{Z}, \mathbf{S}, \mathbf{U}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \gamma, \eta|\mathbf{W})$, where $\mathbf{S} = \{\mathbf{S}^\alpha, \mathbf{S}^\beta\}$, $\mathbf{U}^\alpha = \{U_d^\alpha\}_{d=1}^D$, $\mathbf{U}^\beta = \{U_t^\beta\}_{t=1}^T$, $\mathbf{U} = \{\mathbf{U}^\alpha, \mathbf{U}^\beta\}$, $\boldsymbol{\alpha}^* = (\alpha_1^*, \ldots, \alpha_I^*)^\top$, and $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_J^*)^\top$. To simulate the target posterior distribution, however, a Gibbs sampler cannot be implemented because the conditional distribution of $z_n^{(d)}$ given the remaining model components including $\boldsymbol{\phi}$ becomes degenerate and thus the implementation of the sampling step is infeasible. Thus the apparently simple concept of imposing additional probabilistic structure on the hyperparameters is difficult to implement in current inference procedures, such as CG, and the method of partial collapse should be devised (van Dyk & Park, 2008; Park & van Dyk, 2009). This is why $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are collapsed out of the conditional distribution of $\mathbf{Z}$ in a collapsed Gibbs sampler, which makes the corresponding sampling step feasible. Unfortunately, however, the resulting set of conditional distributions used to simulate the target posterior distribution under the proposed enhanced LDA model remains functionally incompatible, which is not allowed in the collapsed Gibbs sampler (van Dyk & Park, 2008; Park & van Dyk, 2009). In addition, $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are estimated by using point estimation, which ignores their uncertainties. A typical concern about iterative sampling algorithms is slow convergence, requiring many iterations to attain convergence. Although a relatively large number of iterations is inevitable for simulating the target posterior distribution, we capitalize on a functionally incompatible set of conditional distributions for quicker convergence characteristics, thereby alleviating the concern. That is, we devise an efficient iterative sampling scheme using the method of partial collapse, resulting in the partially collapsed Gibbs (PCG) sampler (van Dyk & Park, 2008; Park & van Dyk, 2009).

2.2. Partially collapsed Gibbs sampler for LDA

Given the proposed model in (5), the target posterior distribution is

$$p(\mathbf{Z}, \mathbf{S}, \mathbf{U}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \gamma, \eta|\mathbf{W}),$$

based on which we can construct a PCG sampler using three basic tools (van Dyk & Park, 2008) because $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ can be partially collapsed out of a conditional distribution for $\mathbf{Z}$ that is proportional to

$$
\begin{aligned}
&p(\mathbf{Z}, \mathbf{S}, \mathbf{U}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \gamma, \eta|\mathbf{W})\\
&= \int\int p(\mathbf{Z}, \mathbf{S}, \mathbf{U}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \gamma, \eta|\mathbf{W}) d\boldsymbol{\theta} d\boldsymbol{\phi}.
\end{aligned}
\tag{6}
$$

Specifically, the pseudocode of the PCG sampler for efficient Bayesian nonparametric inference on the enhanced LDA model is given below; details of the PCG sampler are shown in Appendix.

Algorithm 1 PCG for LDA.

1: Generate initial values
2: **repeat**
3: Step 1. Draw $z_n^{(d)}$ from $p(z_n^{(d)}|\mathbf{Z}_{-(n,d)}, \mathbf{S}, \mathbf{U}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \gamma, \eta, \mathbf{W})$
4: Step 2. Draw $(\boldsymbol{\theta}, \boldsymbol{\phi})$ from $p(\boldsymbol{\theta}, \boldsymbol{\phi}|\mathbf{Z}, \mathbf{S}, \mathbf{U}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \gamma, \eta, \mathbf{W})$
5: Step 3. Draw $\mathbf{S}$ from $p(\mathbf{S}|\mathbf{Z}, \mathbf{U}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \gamma, \eta, \mathbf{W})$
6: Step 4. Draw $\mathbf{U}$ from $p(\mathbf{U}|\mathbf{Z}, \mathbf{S}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \gamma, \eta, \mathbf{W})$
7: Step 5. Draw $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ from $p(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*|\mathbf{Z}, \mathbf{S}, \mathbf{U}, \boldsymbol{\theta}, \boldsymbol{\phi}, \gamma, \eta, \mathbf{W})$
8: Step 6. Draw (γ, η) from $p(\gamma, \eta|\mathbf{Z}, \mathbf{S}, \mathbf{U}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \mathbf{W})$
9: **until** convergence criterion satisfied

The derived PCG sampler is composed of a functionally incompatible set of conditional distributions, thereby converging to an unknown stationary distribution unless the sampler is run with a specific sampling order (van Dyk & Park, 2008; Park & van Dyk, 2009). The proposed PCG sampler not only makes it possible to jointly sample $\mathbf{Z}$ and $\boldsymbol{\phi}$ in an iterative sampling scheme, but also provides a computationally efficient algorithm with better convergence characteristics. The advantages of the PCG sampler have been illustrated in many previous studies (van Dyk & Jiao, 2015; Jeong, Park, & Park, 2017; Jeong & Park, 2016; Min & Park, 2017; Park, 2011; Park, Jeong, & Lee, 2012; Park & Jeong, 2015; 2018; Park, Krafty, & Sanchez, 2012; Park & Min, 2016).

When the enhanced LDA model is dealt with, the CG method (Griffiths & Steyvers, 2004) would collapse both $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ as in Step 1 of the PCG sampler. Thus, both CG and PCG have the same sampling step for $\mathbf{Z}$, except that CG fixes the other model components as constants and PCG iteratively samples them. By implementing additional sampling steps, PCG requires more total computation time than CG, but PCG can provide unbiased parameter estimation unlike CG.

The proposed PCG sampler is different from the *partially collapsed scheme* devised in the HD method (Newman et al., 2009; Newman et al., 2008) in that the former is a fully iterative algorithm and the latter estimates some model components during iterations. Both algorithms, however, take advantage of the method of partial collapse to improve convergence characteristics of an iterative algorithm. It is known beneficial to partially collapse more components out of a model; see van Dyk and Park (2008) for theoretical justification. Thus, PCG is expected to outperform the fully iterative version of the partially collapsed scheme for an identical LDA model. This is because PCG partially collapses both $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ but the partially collapsed scheme $\boldsymbol{\theta}$ only. In addition, the fully iterative version of the partially collapsed scheme is reduced to the collapsed Gibbs sampler (Liu, 1994) and can be viewed as a special case of the partially collapsed Gibbs sampler (van Dyk & Park, 2008).

3. Simulation study

We conducted a simulation study to compare the proposed model and methodology with current popular methods. We present a small example to demonstrate how existing methods can fail even in simple situations, and a more complex example to compare current and proposed methods for a significant task.

3.1. A small example

We consider a simple case with three topics. Hence, the distribution of topic contributions to each document can be visualized in a ternary diagram. The standard LDA model assumes that both

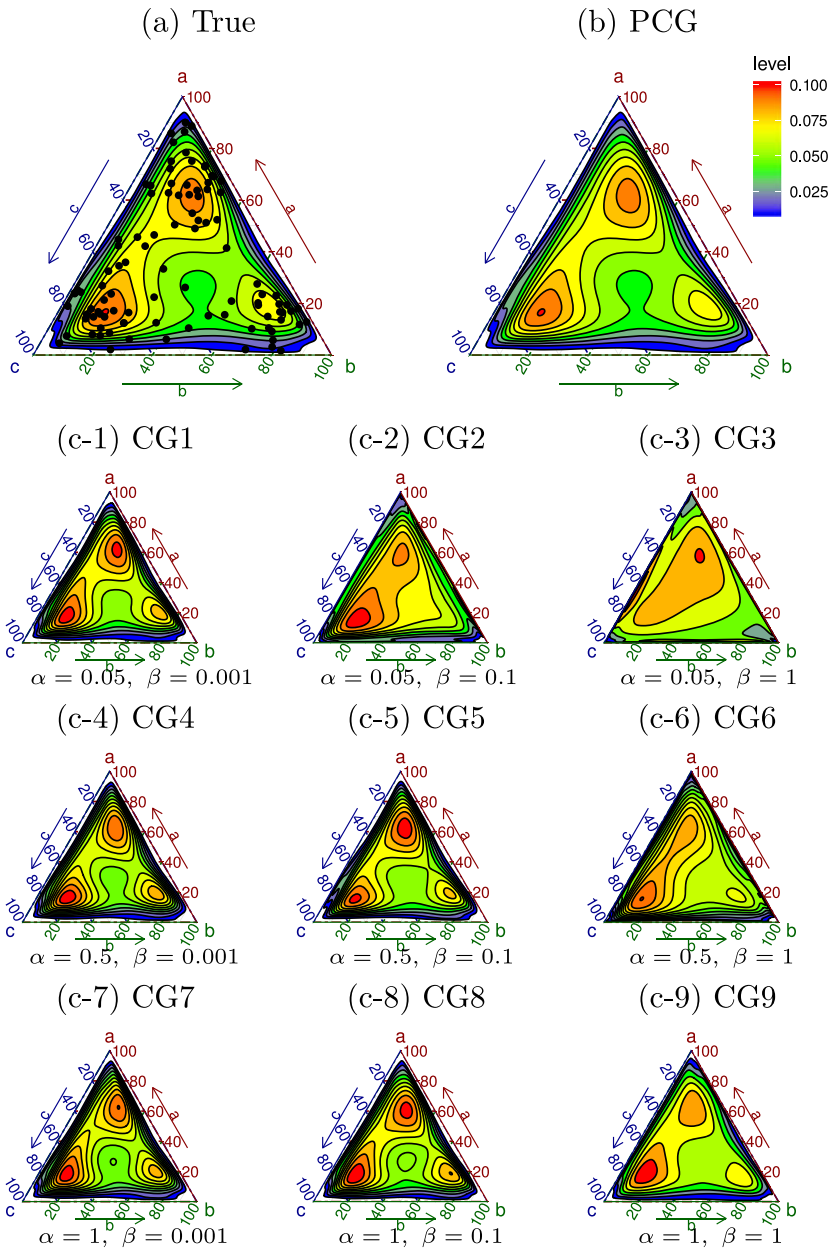

Fig. 2. Posterior distribution of topic contributions: (a) true distribution of $\boldsymbol{\theta}$ with expected values with solid dots that are the values of $\boldsymbol{\theta}$; (b) posterior distribution estimated by PCG, (c) posterior distributions estimated by CG, each with different fixed hyperparameters. Lower posterior densities are represented by blue colors and higher posterior densities by red colors.

topic contributions for a given document and word contributions for a given topic follow symmetric Dirichlet prior distributions, each with a single hyperparameter that is determined before running the CG sampler. In contrast, the proposed LDA model does not fix hyperparameter values for the symmetric Dirichlet prior distributions, but flexibly models them using a hierarchical structure. We simulate a corpus of $D = 100$ documents each with a sequence of words generated from a Poisson distribution with mean of 300, where we have $M = 100$ unique words and $T = 3$ topics. The true distribution of topic contributions for a given document was set to a mixture of three *asymmetric* Dirichlet distributions,

$$\boldsymbol{\theta}^{(d)} \overset{\text{iid}}{\sim} \frac{1}{3} \cdot \text{Dirichlet}(8,2,2) + \frac{1}{3} \cdot \text{Dirichlet}(2,8,2) + \frac{1}{3} \cdot \text{Dirichlet}(2,2,8), \quad d = 1, \ldots, 100, \tag{7}$$

which represents a highly multimodal structure. By contrast, the true distribution of word contributions for a given topic was generated from a single *symmetric* Dirichlet distribution,

$$\boldsymbol{\phi}^{(t)} \overset{\text{iid}}{\sim} \text{Dirichlet}(0.1), \quad t = 1, 2, 3, \tag{8}$$

which is unimodal.

In such a simple setting where the true topic distribution is multimodal and the true word distribution is unimodal, CG can fail to recover the true multimodal structure of the topic distribution, depending on the fixed values of hyperparameters. The proposed enhanced LDA model shown in the right panel of Fig. 1 misspecifies the true topic distribution in the small example because of using a mixture of *symmetric* Dirichlet distributions. Unlike CG, however, PCG can recover the true multimodal structure of the topic distribution by letting data determine the values of hyperparameters for the mixture distribution. Fig. 2 compares between

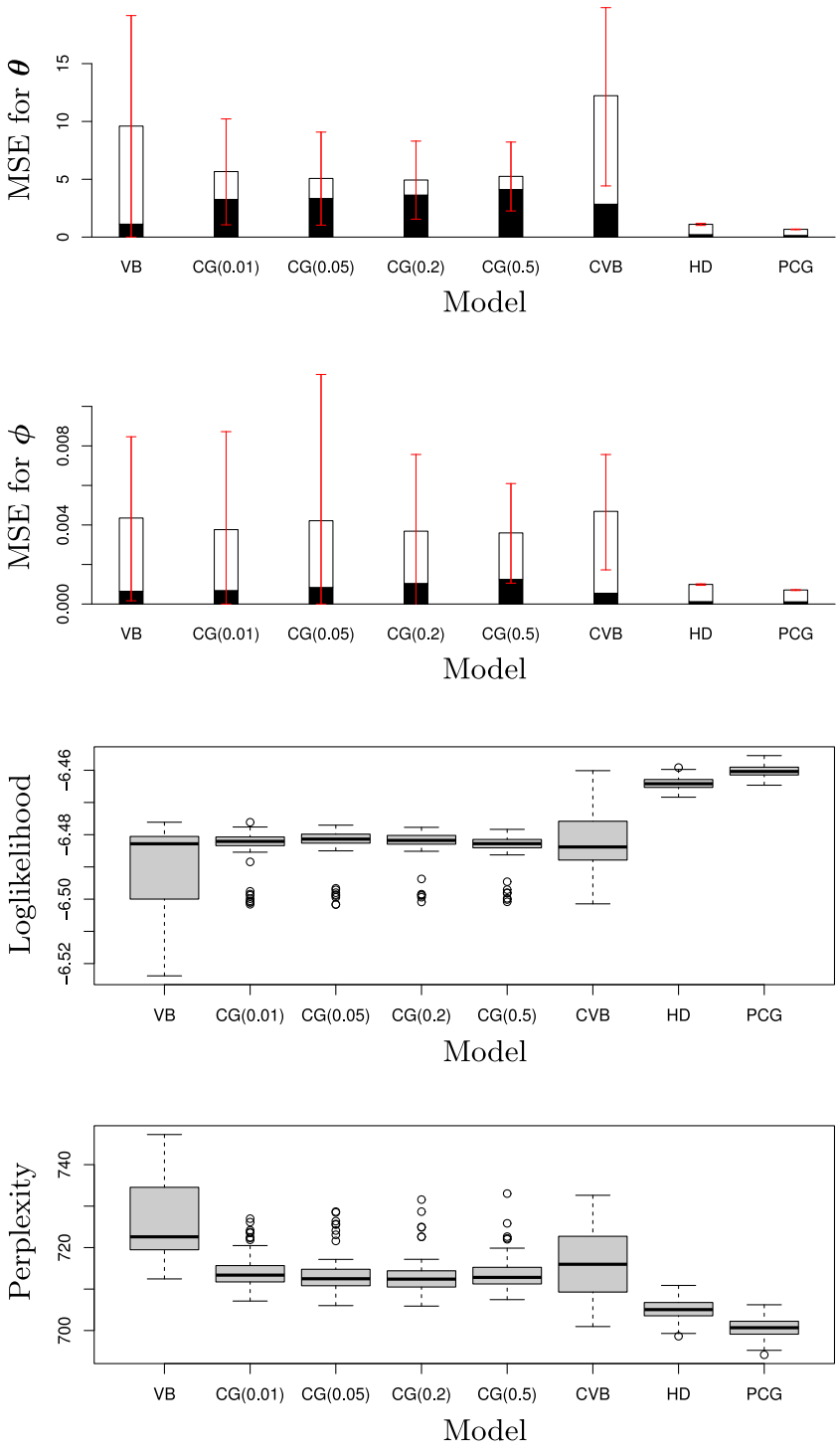

Fig. 3. Comparison of model performance in terms of the mean squared error (MSE) of $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\phi}}$, log-likelihood, and perplexity. In the top two panels, white and black bars represent variance and bias squared, respectively, and red lines represent the error bars (mean $\pm$ standard deviation) of MSE.

CG and PCG in terms of how the posterior distributions of topic contributions over documents are close to the true one. Fig. 2(a) shows the true topic distribution in equation (7), which is highly multimodal; solid dots represent the true values of $\boldsymbol{\theta}$. In Fig. 2(b), the true topic distribution is well estimated by the posterior distribution of topic contributions over documents, computed by the proposed model and methodology based on PCG. Fig. 2(c) shows the posterior distributions of topic contributions as a function of the fixed values of hyperparameters, α and β, in the probabilistic LDA model shown in the left panel of Fig. 1. Because a contour plot with $\alpha = 0.5$ and $\beta = 0.001$ is the most similar to the true one among contour plots in Fig. 2(c), the corresponding CG method produces seemingly unbiased estimates for the true topic contributions. Other choices would be predominantly unimodal, causing large posterior estimation bias, e.g., $\alpha = 0.05$ and $\beta = 1$ in Fig. 2(c-3). Thus, optimal hyperparameter choice depends on observed data, and the true multimodal topic contribution distribution may not be properly mapped by fixing hyperparameters a priori. That is, the performance of CG depends on the fixed hyperparameter choice, which may incur large bias. In contrast, the proposed PCG method produces robust results by flexibly modeling the distribution of topic contributions, adding flexibility for the distribution of word contributions, and allowing data to automatically estimate hyperparameter values.

3.2. A complex example

We considered a more complex case with ten topics, simulating a corpus of $D = 390$ documents each with a sequence of words generated from a Poisson distribution with mean of 1000, where we have $M = 1200$ unique words and $T = 10$ topics. The true distribution of topic contributions for a given document was set to a mixture of three symmetric Dirichlet distributions,

$$\boldsymbol{\theta}^{(d)} \overset{\text{iid}}{\sim} \frac{1}{3}\cdot \text{Dirichlet}(0.1) + \frac{1}{3}\cdot \text{Dirichlet}(0.3) + \frac{1}{3}\cdot \text{Dirichlet}(1), \quad d = 1, \ldots, 390, \tag{9}$$

while the true distribution of word contributions for a given topic was set to a mixture of two symmetric Dirichlet distributions,

$$\boldsymbol{\phi}^{(t)} \overset{\text{iid}}{\sim} \frac{1}{2}\cdot \text{Dirichlet}(0.1) + \frac{1}{2}\cdot \text{Dirichlet}(0.5), \quad t = 1, \ldots, 10. \tag{10}$$

These simulation settings represent a multimodal structure for both $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$.

We divided the simulated data into 300 training and 90 test documents. The simulation was repeated 100 times, with current and proposed methods applied for each of the training documents. The VB and CVB methods were run until convergence, where VB computed hyperparameter values using the Newton–Raphson method and CVB used $\alpha = 0.05$ and $\beta = 0.005$ that are typical choices in practice. The CG and PCG methods were run for 2000 iterations with 500 burn-in iterations. Based on empirical analyses, CG is more susceptible to the choice of α than β. Therefore, we set $\beta = 0.005$ and considered $\alpha = 0.01$, 0.05, 0.2, and 0.5, represented as CG(0.01), CG(0.05), CG(0.2), and CG(0.5), respectively. The HD method was run until convergence based on the values of likelihood and perplexity with the setting in the original text (Newman et al., 2009; Newman et al., 2008). The proposed PCG method used DP mixture priors for the hyperparameters of $\boldsymbol{\theta}^{(d)}$ and $\boldsymbol{\phi}^{(t)}$, where base distributions, G_0 and P_0, were set to exponential distributions with rates 1 and 10, respectively.

We compared VB, CG, CVB and HD methods with PCG in terms of MSE (mean squared error) of topic contributions and word contributions in a training corpus, the average of marginal log-likelihood for both topic contributions and word contributions in a training corpus, and perplexity per a word in a test corpus. The MSE of topic contributions estimated in a training corpus is defined as the sum of the average of squared errors between each element of the topic contributions and the corresponding true value, i.e.,

$$\text{MSE}(\widehat{\boldsymbol{\theta}}) = \sum_{d=1}^{D}\sum_{t=1}^{T} E\left[\left(\widehat{\theta}_t^{(d)} - \theta_t^{(d)}\right)^2\right], \tag{11}$$

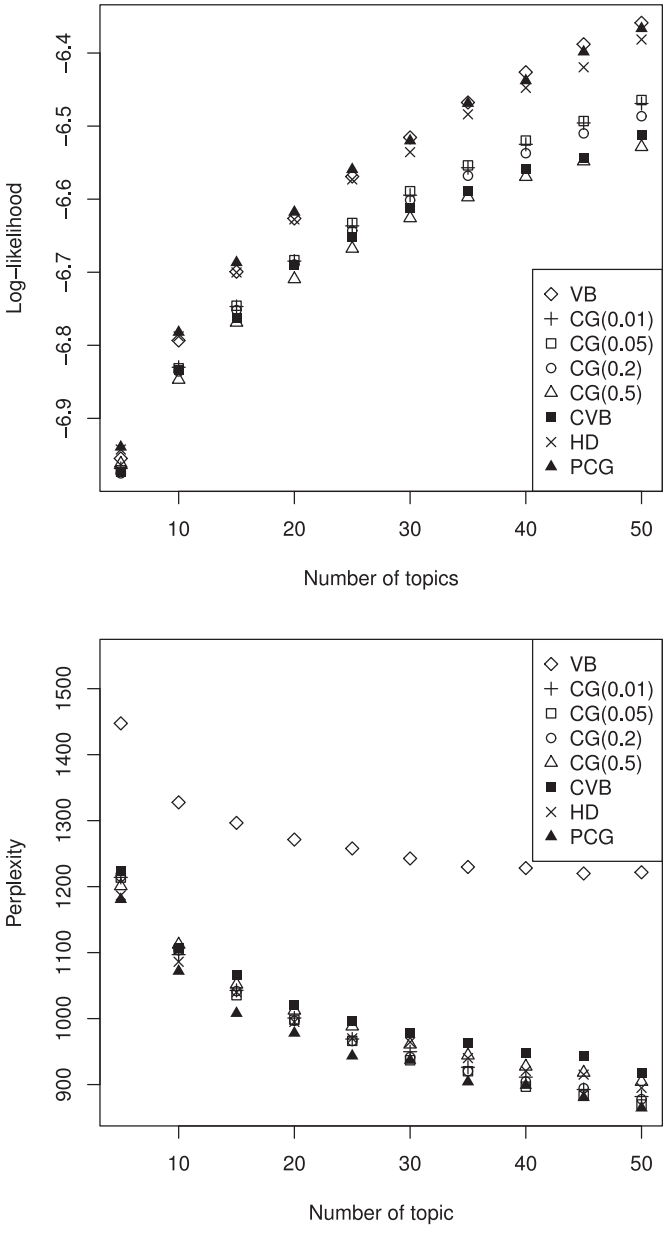

Fig. 4. Method performance in terms of log-likelihood (top) and perplexity (bottom) for the NIPS conference and BBC news data.

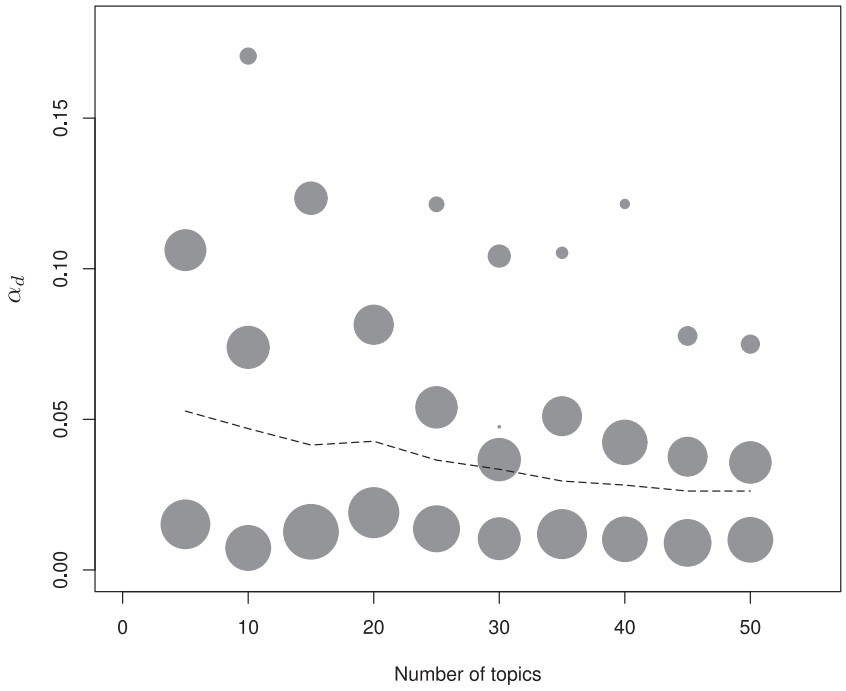

Fig. 5. Sampled atoms of α_d for the DP mixture prior for the hyperparameter of the distribution of $\boldsymbol{\theta}^{(d)}$ with the corresponding cluster sizes. The dashed lines represent the weighted average of the sampled atoms of α_d with the weights being cluster sizes.

where $\theta_k^{(d)}$ is the true value used in the simulation study. Likewise, given the true value $\boldsymbol{\phi}_m^{(k)}$, the MSE of word contributions estimated in a training corpus is defined as

$$\text{MSE}(\widehat{\boldsymbol{\phi}}) = \sum_{t=1}^{T}\sum_{m=1}^{M} E\left[\left(\widehat{\boldsymbol{\phi}}_m^{(t)} - \boldsymbol{\phi}_m^{(t)}\right)^2\right]. \tag{12}$$

The MSE can be decomposed into the variance of an estimator and the squared bias of the estimator. The average of marginal log-likelihood for both topic contributions and word contributions in a training corpus is computed as

$$\ell(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\phi}}|\mathbf{W}) = \frac{\sum_{d=1}^{D}\sum_{n=1}^{N_d} \log p(w_n^{(d)}|\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\phi}})}{\sum_{d=1}^{D} N_d} \tag{13}$$

and the perplexity measure in a test corpus is defined as

$$\text{perplexity} = \exp\left(-\frac{1}{N^*}\sum_{d=1}^{D^*}\sum_{n=1}^{N_d^*} \log p(w_n^{(d)*}|\widehat{\boldsymbol{\theta}}^*, \widehat{\boldsymbol{\phi}})\right), \tag{14}$$

where $p(w_n^{(d)} = w_m|\boldsymbol{\theta}, \boldsymbol{\phi}) = \sum_{t=1}^{T}\theta_t^{(d)}\phi_m^{(t)}$ for $m = 1, \ldots, M$ and * indicates quantities in a test corpus. Because the perplexity measure represents the level of uncertainty in the prediction of each word in a test corpus and its log value is equivalent to the test set cross-entropy (Chen, 2009), lower perplexity indicates better performance. The log-likelihood and perplexity represent performance in a training corpus and a test corpus, respectively. For fair performance comparison, we controlled total computational time for all methods but CG. In the case of CG, several model parameters are fixed to some constants in advance, which reduces total computation time compared to PCG. Thus, in this simulation study, PCG pays approximately 20% more computation time per iteration than CG. Although HD has an advantage of maximizing its performance in multiple processors environment, we did not consider multiple processing in order to consider comparison in the same environment.

There is a label switching problem for topics when calculating the MSE. We thus select permutations for the elements of $\boldsymbol{\theta}^{(d)}$ and the row vectors of $\boldsymbol{\phi}$ by minimizing the Euclidean distance between the true and generated values. This label switching problem only matters when we are interested in MSE in a simulation setting. The first two plots of Fig. 3 shows the MSE of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, respectively, where MSE is decomposed into variance and bias squared, and an error bar (mean $\pm$ standard deviation) is indicated on the top of each bar. The MSE of the PCG method is the smallest among the current methods, with the shortest error bars. This illustrates that PCG clearly outperforms current methods in terms of the MSE of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, properly accounting for the multimodality of the topic contribution distribution. The last two plots of Fig. 3 compare the performance of proposed and existing methods in terms of log-likelihood and perplexity. As compared existing methods, the PCG method has significantly high log-likelihood and low perplexity, thereby explaining a training corpus well and achieving good predictability in a test corpus. As the number of topic increases, the likelihood of VB is increasingly better than that of PCG, but the perplexity of VB get much worse. This is because VB overfits a training corpus and thus yields poor performance in a test corpus. Overall, the PCG method significantly outperforms existing methods, illustrating the validity and advantages of the proposed method.

4. Application to the NIPS conference and BBC news data

We analyzed a corpus of documents that vary in topic and size. To compose such a corpus of documents, we collected 388 documents from the neural information processing systems (NIPS)[2] conference articles during the period of 2000–2002 and 912 documents from the British Broadcasting Corporation (BBC)[3] sports and technology news articles during the period of 2004–2005. The NIPS conference addresses various subjects related to human and artificial intelligence, and the dataset contained 9 sections: algorithms and architectures (AA), applications (AP), cognitive science

[2] http://books.nips.cc
[3] http://mlg.ucd.ie/datasets/bbc.html

(CS), control and navigation (CN), implementations (IM), learning theory (LT), neuroscience (NS), signal processing (SP), and vision sciences (VS) for the 2000 conference, 2000–2002 sections were slightly different year to year. The same text data were previously analyzed by Teh et al. (2006). The BBC news articles are composed of 5 sections: business, entertainment, politics, sports, and technology. Of these five sections, we collected 912 articles from the sections of sports (511) and technology (401). The same text data were previously analyzed by Greene and Cunningham (2006).

Since the corpus contained an extremely large number of words, we excluded standard stop words and words occurring more than 500 times or fewer than 50 times in the corpus. Thus the number M of unique words were reduced to from 35,064 to 2,135, and the average number of words in each document decreased from 679 to 250. Each document in the NIPS and BBC corpus has 665 and 100 words on average, respectively. For validation, $D = 1300$ documents were analyzed using 5-fold cross-validation. The number of topics was set to 5–50 with increments of 5.

We calculated log-likelihood and perplexity in the same way as the simulation study and compared PCG, VB, CG with different hyperparameters, CVB and HD methods. The top panel of Fig. 4 shows that the PCG and VB methods provided significantly high log-likelihood, compared the other existing methods. While the bottom panel of Fig. 4 shows that PCG also had significantly low perplexity, however, the perplexity of VB is the highest among all methods, implying very poor performance in terms of perplexity. Thus, the PCG method clearly outperforms all existing methods in terms of the combined measure of log-likelihood and perplexity, although the overall performance of CG(0.05) is comparable to that of PCG.

Fig. 5 illustrates that the optimal choice of a hyperparameter for the CG method changes with the number of topics. However, the optimal hyperparameter values for the CG method were arrived at by trial and error, whereas the PCG method does let data choose hyperparameter values for robust results. In Fig. 5, circles for each number of topics indicate the clusters of documents that have the same α_d, and their sizes correspond to cluster sizes. Hyperparameter clustering in Fig. 5 shows that the sampled atoms of α_d in a DP mixture prior are multiples with comparable cluster sizes, i.e., the estimated distribution of $\boldsymbol{\theta}^{(d)}$ tends to be multimodal with a mixture of several symmetric Dirichlet distributions. Because the weighted average of the sampled atoms of α_d is between 0.01 and 0.1, optimal α from the CG method is close to the weighted average from PCG. Thus, PCG is consistent with CG, and PCG automatically estimates optimal hyperparameters from the data.

5. Discussion

We developed an enhanced LDA model that effectively deals with highly multimodal topic distributions. The proposed enhanced LDA model allows flexible hyperparameter modeling of the distributions of topic contributions for a given document and word contributions for a given topic, rather than fixing them before analysis. The enhanced LDA model has advantages over probabilistic LDA models because optimal hyperparameters are automatically derived from the data. The enhanced LDA model inference procedure is, however, hampered by functional incompatibility for the resulting set of conditional distributions, and hence current inference procedures are infeasible. Therefore, we proposed an inference procedure based on the method of partial collapse. The resulting PCG sampler not only makes inference on the proposed method feasible, but also provides unbiased parameter estimation for highly multimodal latent topic distributions with quick convergence. Simulation studies and a real data application verified that the proposed method outperforms current methods in terms of MSE, likelihood, and perplexity. Because of approximating a target posterior distribution with samples from a Markov chain, the proposed method cannot easily scale to massive data. Slow total computation time is thus inherent in a fully iterative algorithm such as the proposed PCG sampler. Nonetheless, such a fully iterative sampler would be useful for the precise analysis of a relatively small size and can complement other existing methods that are easier to scale. Between a quick-and-dirty solution and a slow-and-perfect solution, our paper thus focused on a slow-and-perfect solution, and improved practicality by reducing its total computation time via the method of partial collapse.

Future research on the proposed topic modeling will incorporate word and topic selection. The curse of dimensionality is also a critical problem for topic modeling because a corpus often consists of more than ten thousand words. Thus, word selection to filter unnecessary words would be useful to devise an informative method with a simpler structure. The proposed method currently assumes a fixed number of topics, which should be generalized to allow the input data to find an unknown number of topics.

Conflict of interest

None.

Credit authorship contribution statement

Hongju Park: Data curation, Formal analysis, Software, Writing - original draft. **Taeyoung Park:** Data curation, Investigation, Methodology, Writing - review & editing. **Yung-Seop Lee:** Conceptualization, Funding acquisition, Supervision, Writing - review & editing.

Acknowledgements

T. Park's research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2017R1D1A1B03033536). Y.-S. Lee's research was supported by the Korea Meteorological Administration Research and Development Program under Grant KMIPA 2015-1110.

Appendix

The proposed PCG sampler iterates among the following steps:

Step 1. Draw $z_n^{(d)}$ from $p(z_n^{(d)}|\mathbf{Z}_{-(n,d)}, \mathbf{S}, \mathbf{U}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \gamma, \eta, \mathbf{W})$, which is discrete with probabilities,

$$p(z_n^{(d)} = t^*|\mathbf{Z}_{-(n,d)}, \mathbf{S}, \mathbf{U}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \gamma, \eta, \mathbf{W}) \propto \frac{C^{(w_n^{(d)})}_{-(n,d),t^*} + \beta^*_{S_t^\beta}}{C^{(+)}_{-(n,d),t^*} + M\beta^*_{S_t^\beta}} \cdot \frac{C^{(d)}_{-(n,d),t^*} + \alpha^*_{S_d^\alpha}}{C^{(d)}_{-(n,d),+} + T\alpha^*_{S_d^\alpha}}, \quad t^* = 1, \ldots, T, \quad (15)$$

where $\mathbf{Z}_{-(n,d)} = \mathbf{Z} \setminus \{z_n^{(d)}\}$ denotes a set of topic indices except for the nth word in document d, $C^{(w)}_{-(n,d),t}$ denotes the number of times word w is assigned to topic t in $\mathbf{Z}_{-(n,d)}$, $C^{(d)}_{-(n,d),t}$ denotes the number of times a word in document d is assigned to topic t, except for $w_n^{(d)}$, and

$$C^{(+)}_{-(n,d),t} = \sum_{m=1}^{M} C^{(w_m)}_{-(n,d),t} \quad (16)$$

$$C^{(d)}_{-(n,d),+} = \sum_{t=1}^{T} C^{(d)}_{-(n,d),t}. \quad (17)$$

Step 2. Draw $(\boldsymbol{\theta}, \boldsymbol{\phi})$ from $p(\boldsymbol{\theta}, \boldsymbol{\phi}|\mathbf{Z}, \mathbf{S}, \mathbf{U}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \gamma, \eta, \mathbf{W})$, which is a product of D independent T-dimensional Dirichlet distributions,

$$\boldsymbol{\theta}^{(d)}|(\mathbf{Z}, \mathbf{S}, \mathbf{U}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \gamma, \eta, \mathbf{W}) \overset{\text{ind}}{\sim} \text{Dirichlet}\left(C_1^{(d)} + \alpha^*_{S_d^\alpha}, \ldots, C_T^{(d)} + \alpha^*_{S_d^\alpha}\right), \ d = 1, \ldots, D, \tag{18}$$

and T independent M-dimensional Dirichlet distributions,

$$\boldsymbol{\phi}^{(t)}|(\mathbf{Z}, \mathbf{S}, \mathbf{U}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \gamma, \eta, \mathbf{W}) \overset{\text{ind}}{\sim} \text{Dirichlet}\left(C_t^{(w_1)} + \beta^*_{S_t^\beta}, \ldots, C_t^{(w_M)} + \beta^*_{S_t^\beta}\right), \ t = 1, \ldots, T, \tag{19}$$

where $C_t^{(d)}$ denotes the number of times a word in document d is assigned to topic t and $C_t^{(w)}$ denotes the number of times word w is assigned to topic t in a set of topic indices $\mathbf{Z}$.

Step 3. Draw $\mathbf{S}$ from $p(\mathbf{S}|\mathbf{Z}, \mathbf{U}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \gamma, \eta, \mathbf{W})$, which is a product of D independent discrete distributions, $p(S_d^\alpha|\mathbf{Z}, \mathbf{U}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \gamma, \eta, \mathbf{W})$ for $d = 1, \ldots, D$, each with probabilities,

$$p(S_d^\alpha = i|\mathbf{Z}, \mathbf{U}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \gamma, \eta, \mathbf{W}) \propto U_i^\alpha \prod_{l<i}(1 - U_l^\alpha) \cdot \frac{\Gamma(T\alpha_i^*)}{\Gamma(\alpha_i^*)^T} \prod_{t=1}^{T} \left(\theta_t^{(d)}\right)^{\alpha_i^* - 1}, \ i = 1, \ldots, I, \tag{20}$$

and T independent discrete distributions, $p(S_t^\beta|\mathbf{Z}, \mathbf{U}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \gamma, \eta, \mathbf{W})$ for $t = 1, \ldots, T$, each with probabilities,

$$p(S_t^\beta = j|\mathbf{Z}, \mathbf{U}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \gamma, \eta, \mathbf{W}) \propto U_j^\beta \prod_{l<j}(1 - U_l^\beta) \cdot \frac{\Gamma(M\beta_j^*)}{\Gamma(\beta_j^*)^M} \prod_{m=1}^{M} (\phi_m^{(t)})^{\beta_j^* - 1}, \ j = 1, \ldots, J. \tag{21}$$

Step 4. Draw $\mathbf{U}$ from $p(\mathbf{U}|\mathbf{Z}, \mathbf{S}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \gamma, \eta, \mathbf{W})$, which is a product of I independent beta distributions,

$$U_i^\alpha|(\mathbf{Z}, \mathbf{S}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \gamma, \eta, \mathbf{W}) \overset{\text{ind}}{\sim} \text{Beta}\left(1 + \sum_{d=1}^{D} \delta_i(S_d^\alpha), \gamma + \sum_{d=1}^{D} \sum_{k=i+1}^{I} \delta_k(S_d^\alpha)\right), \ i = 1, \ldots, I-1, \tag{22}$$

and J independent beta distributions,

$$U_j^\beta|(\mathbf{Z}, \mathbf{S}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \gamma, \eta, \mathbf{W}) \overset{\text{ind}}{\sim} \text{Beta}\left(1 + \sum_{t=1}^{T} \delta_j(S_t^\beta), \eta + \sum_{t=1}^{T} \sum_{k=j+1}^{J} \delta_k(S_t^\beta)\right), \ j = 1, \ldots, J-1, \tag{23}$$

where $U_I^\alpha = 1$ and $U_J^\beta = 1$.

Step 5. Draw $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ from $p(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*|\mathbf{Z}, \mathbf{S}, \mathbf{U}, \boldsymbol{\theta}, \boldsymbol{\phi}, \gamma, \eta, \mathbf{W})$, which is

$$p(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*|\mathbf{Z}, \mathbf{S}, \mathbf{U}, \boldsymbol{\theta}, \boldsymbol{\phi}, \gamma, \eta, \mathbf{W}) = \prod_{i=1}^{I} p(\alpha_i^*|\mathbf{Z}, \mathbf{S}, \mathbf{U}, \boldsymbol{\theta}, \boldsymbol{\phi}, \gamma, \eta, \mathbf{W}) \prod_{j=1}^{J} p(\beta_j^*|\mathbf{Z}, \mathbf{S}, \mathbf{U}, \boldsymbol{\theta}, \boldsymbol{\phi}, \gamma, \eta, \mathbf{W}), \tag{24}$$

where the conditional distribution of α_i^* is proportional to

$$\left(\frac{\Gamma(T\alpha_i^*)}{\Gamma(\alpha_i^*)^T}\right)^{n_i^\alpha} \exp\left(-\alpha_i^*\left(\lambda - \sum_{d:S_d^\alpha = i} \sum_{t=1}^{T} \log \theta_t^{(d)}\right)\right), \tag{25}$$

and the conditional distribution of β_j^* is proportional to

$$\left(\frac{\Gamma(M\beta_j^*)}{\Gamma(\beta_j^*)^M}\right)^{n_j^\beta} \exp\left(-\beta_j^*\left(\mu - \sum_{t:S_t^\beta = j} \sum_{m=1}^{M} \log \phi_m^{(t)}\right)\right), \tag{26}$$

where $n_i^\alpha = \sum_{d=1}^{D} \delta_i(S_d^\alpha)$ and $n_j^\beta = \sum_{t=1}^{T} \delta_j(S_t^\beta)$. Since the conditional distributions can be shown to be log-concave, we use Metropolized independent sampling to simulate $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ by using a normal proposal distribution truncated to be positive, with mode and curvature matched to the corresponding conditional distribution.

Step 6. Draw (γ, η) from $p(\gamma, \eta|\mathbf{Z}, \mathbf{S}, \mathbf{U}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \mathbf{W})$, which is a product of independent Gamma distributions,

$$\gamma|(\mathbf{Z}, \mathbf{S}, \mathbf{U}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \mathbf{W}) \sim \text{Gamma}\left(I, a_0 - \sum_{i=1}^{I-1} \log(1 - U_i^\alpha)\right), \tag{27}$$

and

$$\eta|(\mathbf{Z}, \mathbf{S}, \mathbf{U}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \mathbf{W}) \sim \text{Gamma}\left(J, b_0 - \sum_{j=1}^{J-1} \log(1 - U_j^\beta)\right). \tag{28}$$

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.eswa.2019.04.028

References

Blei, D. M., Griffiths, T. L., Jordan, M. I., & Tenenbaum, J. (2004). Hierarchical topic models and the nested chinese restaurant process. In *Advances in neural information processing systems* (pp. 17–24).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Chen, S. F. (2009). Performance prediction for exponential language models. In *Proceedings of human language technologies: The 2009 annual conference of the north american chapter of the association for computational linguistics* (pp. 450–458).

Dean, J., & Ghemawat, S. (2008). Mapreduce: Simplified data processing on large clusters. *Communications of the ACM, 51*, 107–113.

van Dyk, D. A., & Jiao, X. (2015). Metropolis–Hastings within partially collapsed Gibbs samplers. *Journal of Computational and Graphical Statistics, 24*(2), 301–327.

van Dyk, D. A., & Park, T. (2008). Partially collapsed Gibbs samplers: Theory and methods. *Journal of the American Statistical Association, 193*, 790–796.

Fujimoto, H., Etoh, M., Kinno, A., & Akinaga, Y. (2011). Topic analysis of web user behavior using LDA model on proxy logs. In *Pacific-Asia conference on knowledge discovery and data mining* (pp. 525–536).

Greene, D., & Cunningham, P. (2006). Practical solutions to the problem of diagonal dominance in kernel document clustering. In *Proceedings of the 23rd international conference on machine learning* (pp. 377–384).

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences, 101*, 5228–5235.

Gropp, W. D., Gropp, W., Lusk, E., & Skjellum, A. (1999). *Using MPI: Portable parallel programming with the message-passing interface*: 1. MIT press.

Ishwaran, H., & James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association, 96*(453), 161–173.

Jeong, S., Park, M., & Park, T. (2017). Analysis of binary longitudinal data with time–varying effects. *Computational Statistics and Data Analysis, 112*, 145–153.

Jeong, S., & Park, T. (2016). Bayesian semiparametric inference on functional relationships in linear mixed models. *Bayesian Analysis, 11*(4), 1137–1163.

Liu, J. S. (1994). The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association, 89*, 958–966.
Liu, Y., Liu, Z., Chua, T.-S., & Sun, M. (2015). Topical word embeddings.. In *Proceedings of the twenty-ninth AAAI conference on artificial intelligence* (pp. 2418–2424).
Mikolov, T., & Zweig, G. (2012). Context dependent recurrent neural network language model.. *Spoken Language Technology Workshop (SLT), 2012 IEEE, 12*(234–239), 8.
Min, S., & Park, T. (2017). Bayesian variable selection in Poisson change-point regression analysis. *Communications in Statistics - Simulation and Computation, 46*(3), 2267–2282.
Minka, T. (2000). Estimating a Dirichlet distribution. *Technical report*. MIT.
Newman, D., Asuncion, A., Smyth, P., & Welling, M. (2009). Distributed algorithms for topic models. *Journal of Machine Learning Research, 10*, 1801–1828.
Newman, D., Smyth, P., Welling, M., & Asuncion, A. U. (2008). Distributed inference for latent Dirichlet allocation. In *Advances in neural information processing systems* (pp. 1081–1088).
Paisley, J., Wang, C., Blei, D. M., & Jordan, M. I. (2015). Nested hierarchical Dirichlet processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 37*, 256–270.
Park, T. (2011). Bayesian analysis of individual choice behavior with aggregate data. *Journal of Computational and Graphical Statistics, 20*(1), 158–173.
Park, T., & van Dyk, D. A. (2009). Partially collapsed Gibbs samplers: Illustrations and applications. *Journal of Computational and Graphical Statistics, 18*, 283–305.
Park, T., Jeong, J.-H., & Lee, J. W. (2012). Bayesian nonparametric inference on quantile residual life function: Application to breast cancer data. *Statistics in Medicine, 31*(18), 1972–1985.
Park, T., & Jeong, S. (2015). Efficient Bayesian analysis of multivariate aggregate choices. *Journal of Statistical Computation and Simulation, 85*(16), 3352–3366.
Park, T., & Jeong, S. (2018). Analysis of Poisson varying-coefficient models with autoregression. *Statistics, 52*(1), 34–49.
Park, T., Krafty, R. T., & Sanchez, A. I. (2012). Bayesian semi-parametric analysis of Poisson change-point regression models: Application to policy-making in Cali, Colombia. *Journal of Applied Statistics, 39*(10), 2285–2298.
Park, T., & Min, S. (2016). Partially collapsed Gibbs sampling for linear mixed-effects models. *Communications in Statistics - Simulation and Computation, 45*(1), 165–180.
Rodriguez, A., Dunson, D. B., & Gelfand, A. E. (2008). The nested Dirichlet process. *Journal of the American Statistical Association, 103*, 1131–1154.
Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica, 4*(2), 639–650.
Teh, Y., Newman, D., & Welling, M. (2007). A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. *Advances in Neural Information Processing Systems*, 1353–1360.
Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association, 101*, 1566–1581.
Wang, Y., Bai, H., Stanton, M., Chen, W.-Y., & Chang, E. Y. (2009). PLDA: Parallel latent Dirichlet allocation for large-scale applications.. *Algorithmic Aspects in Information and Management, 9*, 301–314.
Wei, G. C., & Tanner, M. A. (1990). A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American statistical Association, 85*(411), 699–704.
White, R., Harwin, W. S., Holderbaum, W., & Johnson, L. (2015). Investigating eating behaviours using topic models. In *Proceedings of the IEEE 14th international conference on machine learning and applications* (pp. 265–270).