Partially Collapsed Gibbs Sampling for Latent Dirichlet Allocation

Hongju Park
Department of Statistics, University of Georgia

Contact Information:
Department of Statistics
University of Georgia
310 Herty Drive, GA, USA

Phone: +1 (706) 619 9528
Email: hp97161@uga.edu

Abstract

A latent Dirichlet allocation (LDA) model is a Bayesian hierarchical model that identifies latent topics from text corpora. Current popular inferential methods to fit the LDA model are based on variational Bayesian inference, collapsed Gibbs sampling, or a combination of these. However, these methods can suffer from large bias, particularly when text corpora consist of various clusters with different topic distributions. This research proposes an inferential LDA method to efficiently obtain unbiased estimates under flexible modeling for text corpora by using 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

- **A latent Dirichlet allocation (LDA)** model is a hierarchical Bayesian model used to identify latent topics underlying collections of discrete data as well as text corpora.

- Its current inference procedures, **variational Bayesian (VB)** inference and **collapsed Gibbs (CG)** sampling, suffer from biased parameter estimation caused by hyperparameter values fixed in advance.

- In the current LDA model, we are forced to assume **unimodal distributions** for latent variables $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$.

2 Main objectives

1. To develop an efficient and feasible inference procedure with a **partially collapsed Gibbs sampler (PCG)** yielding unbiased parameter estimation for the enhanced LDA model

2. 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 with **Dirichlet process (DP) prior distributions**

3. To compare the performance of the proposed model with other models in terms of **MSE**, **Likelihood** and **Perplexity**

3 Model description

3.1 Graphical model representation

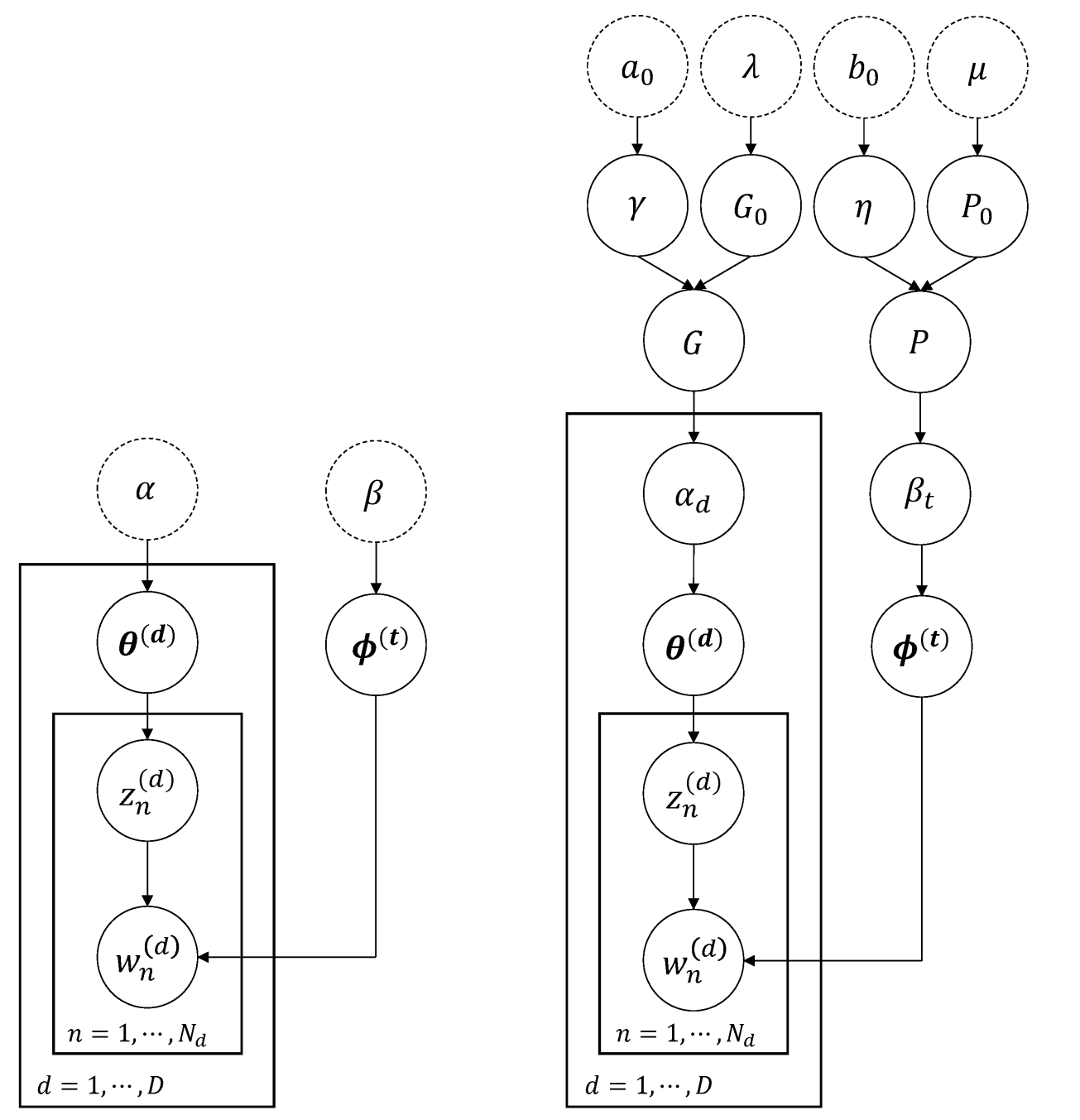

Figure 1: Comparison of (left) probabilistic LDA model and (right) proposed enhanced LDA model.

3.2 Model assumption

$$\begin{aligned}
w_n^{(d)}|(z_n^{(d)}, \boldsymbol{\phi}) &\overset{\text{ind}}{\sim} \text{Discrete}(\boldsymbol{\phi}^{(z_n^{(d)})}),\ n = 1, \ldots, N_d,\ d = 1, \ldots, D \\
z_n^{(d)}|\boldsymbol{\theta} &\overset{\text{ind}}{\sim} \text{Discrete}(\boldsymbol{\theta}^{(d)}),\ n = 1, \ldots, N_d,\ d = 1, \ldots, D \\
\boldsymbol{\phi}^{(t)}|\beta_t &\overset{\text{ind}}{\sim} \text{Dirichlet}(\beta_t),\ t = 1, \ldots, T \\
\beta_t|P &\overset{\text{iid}}{\sim} P,\ t = 1, \ldots, T \\
P|(\eta, \mu) &\sim \text{DP}(\eta P_0(\mu)) \\
\boldsymbol{\theta}^{(d)}|\alpha_d &\overset{\text{ind}}{\sim} \text{Dirichlet}(\alpha_d),\ d = 1, \ldots, D \\
\alpha_d|G &\overset{\text{iid}}{\sim} G,\ d = 1, \ldots, D \\
G|(\gamma, \lambda) &\sim \text{DP}(\gamma G_0(\lambda))
\end{aligned}$$

- $w_n^{(d)}$: the nth word in document d
- $z_n^{(d)}$: the topic index for the nth word in document d
- $\boldsymbol{\phi}^{(t)}$: the probabilities under topic t
- $\boldsymbol{\theta}^{(d)}$: the probabilities in document d
- α_d and β_t : the scalar parameters of symmetric Dirichlet distributions for $\boldsymbol{\theta}^{(d)}$ and $\boldsymbol{\phi}^{(t)}$, respectively
- G and P : the distributions of α_d and β_t drawn from DP with precision parameters γ and η, respectively

3.3 Partially collapsed gibbs sampler for Latent Dirichlet Allocation

Step 1. $z_n^{(d)} \sim p(z_n^{(d)}|\mathbf{Z}_{-(n,d)}, \mathbf{S}, \mathbf{U}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \gamma, \eta, \mathbf{W})$
Step 2. $(\boldsymbol{\theta}, \boldsymbol{\phi}) \sim p(\boldsymbol{\theta}, \boldsymbol{\phi}|\mathbf{Z}, \mathbf{S}, \mathbf{U}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \gamma, \eta, \mathbf{W})$
Step 3. $\mathbf{S} \sim p(\mathbf{S}|\mathbf{Z}, \mathbf{U}, \boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \gamma, \eta, \mathbf{W})$
Step 4. $\mathbf{U} \sim p(\mathbf{U}|\mathbf{Z}, \mathbf{S}, \boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \gamma, \eta, \mathbf{W})$
Step 5. $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \sim p(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*|\mathbf{Z}, \mathbf{S}, \mathbf{U}, \boldsymbol{\phi}, \boldsymbol{\theta}, \gamma, \eta, \mathbf{W})$
Step 6. $(\gamma, \eta) \sim p(\gamma, \eta|\mathbf{Z}, \mathbf{S}, \mathbf{U}, \boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \mathbf{W})$

4 A small example

- The purpose of this example is to check the unbiasedness of estimator of $\boldsymbol{\theta}$ by the proposed model as compared to those of the model with CG via visualized ternary diagrams.

- $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are generated as follows.

$$\begin{aligned}
\boldsymbol{\theta}^{(d)} &\overset{\text{ind}}{\sim} \frac{1}{3} \cdot \text{Dirichlet}(8, 2, 2) + \frac{1}{3} \cdot \text{Dirichlet}(2, 8, 2) \\
&\quad + \frac{1}{3} \cdot \text{Dirichlet}(2, 2, 8),\quad d = 1, \ldots, 100. \\
\boldsymbol{\phi}^{(t)} &\overset{\text{ind}}{\sim} \text{Dirichlet}(0.1, 0.1, 0.1),\quad t = 1,\ 2,\ 3
\end{aligned}$$

- The true and posterior distributions of topic contributions over the documents, where the solid dots represent the expected topic contributions for all documents in a corpus.

- 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.

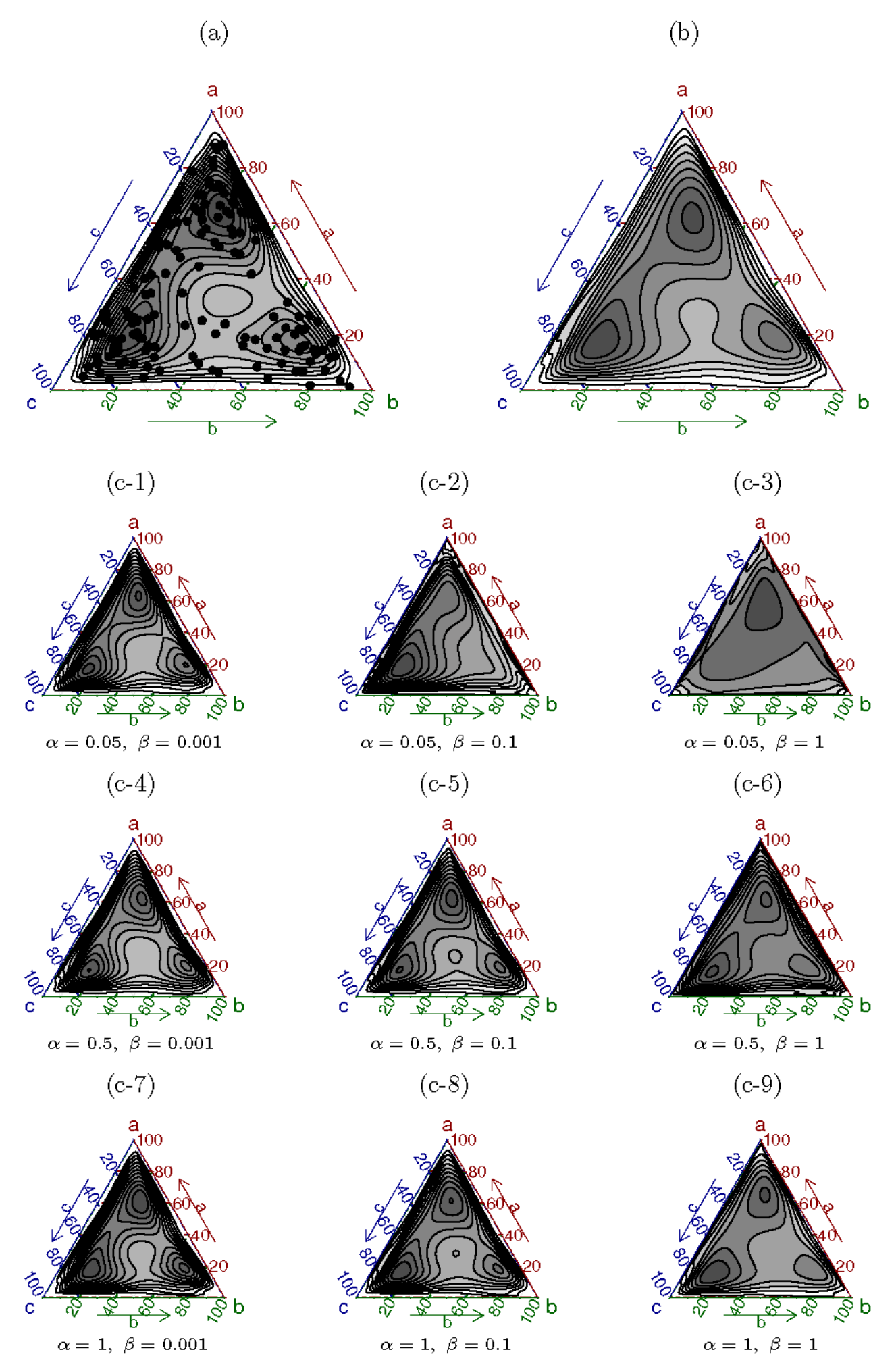

Figure 2: Posterior distribution of topic contributions: (a) true distribution of $\boldsymbol{\theta}$ with expected values represented by solid dots; (b) posterior distribution estimated by PCG, (c) posterior distributions estimated by CG, each with different fixed hyperparameters.

5 A complex example

- This example shows that the proposed model has significant reduced MSE as compared to other models.

- $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are generated from symmetric Dirichelt mixture distributions as follows.

$$\begin{aligned}
\boldsymbol{\theta}^{(d)} &\overset{\text{ind}}{\sim} \frac{1}{3} \cdot \text{Dirichlet}(0.1) + \frac{1}{3} \cdot \text{Dirichlet}(0.3) \\
&\quad + \frac{1}{3} \cdot \text{Dirichlet}(1),\quad d = 1, \ldots, 300. \\
\boldsymbol{\phi}^{(t)} &\overset{\text{ind}}{\sim} \frac{1}{2} \cdot \text{Dirichlet}(0.1) + \frac{1}{2} \cdot \text{Dirichlet}(0.5), \\
&\qquad t = 1, \ldots, 10
\end{aligned}$$

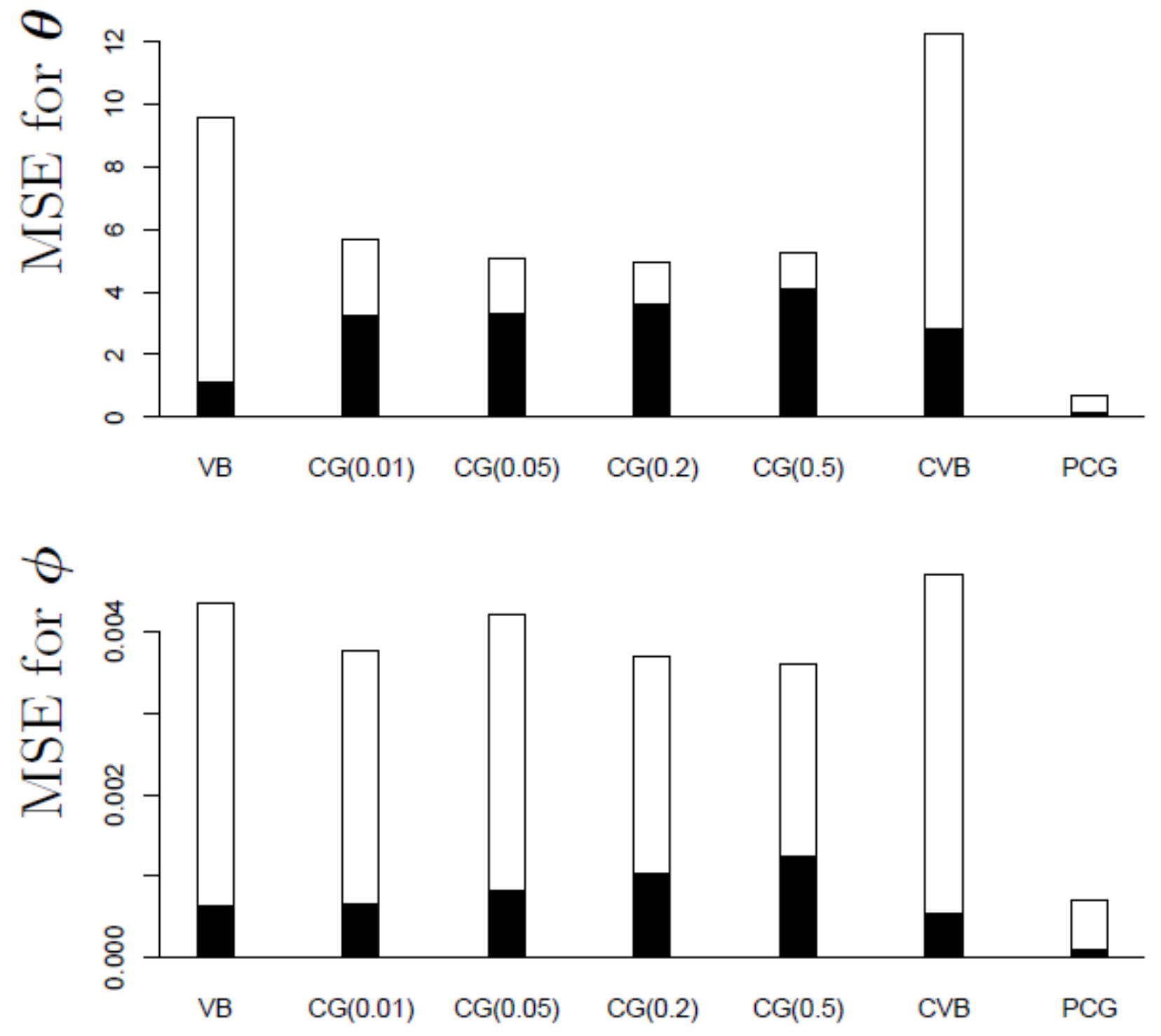

Figure 3: Comparison of model performance in terms of mean squared error (MSE) for $\hat{\theta}$ and $\hat{\phi}$. Black and white bars correspond to bias and variance, respectively.

- The number in the parenthesis CG($\cdot$) means the value of α.

6 Real data analysis

6.1 Data description

- The number of documents : 1300 (388 NIPS articles, 912 BBC news)
- The average length of documents : 250
- The total sort of words : 2135

6.2 Performance measures

- Log-likelihood

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

- Perplexity

$$\text{perplexity} = \exp\left(-\frac{1}{N}\sum_{d=1}^{D_{\text{test}}}\sum_{m=1}^{M} N_d^{(w_m)} \log\left(\boldsymbol{\theta}_{\text{test}}^{(d)} \cdot \boldsymbol{\phi}_{\text{train}}^{(w_m)}\right)\right)$$

  - N_d : the number of words in document d of a training corpus
  - $\boldsymbol{\theta}_{\text{train}}$: topic contributions in a training corpus
  - $\boldsymbol{\phi}_{\text{train}}$: word contributions in a training corpus
  - $N_d^{(w)}$: the number of times word w appears in document d of a test corpus
  - N : the total number of words in the test corpus
  - M : the number of unique words in the entire corpus
  - $\boldsymbol{\theta}_{\text{test}}^{(d)}$: the dth row vector of $\boldsymbol{\theta}$ calculated in the test corpus
  - $\boldsymbol{\phi}_{\text{train}}^{(w_m)}$: the mth column vector of $\boldsymbol{\phi}$ calculated in a training corpus

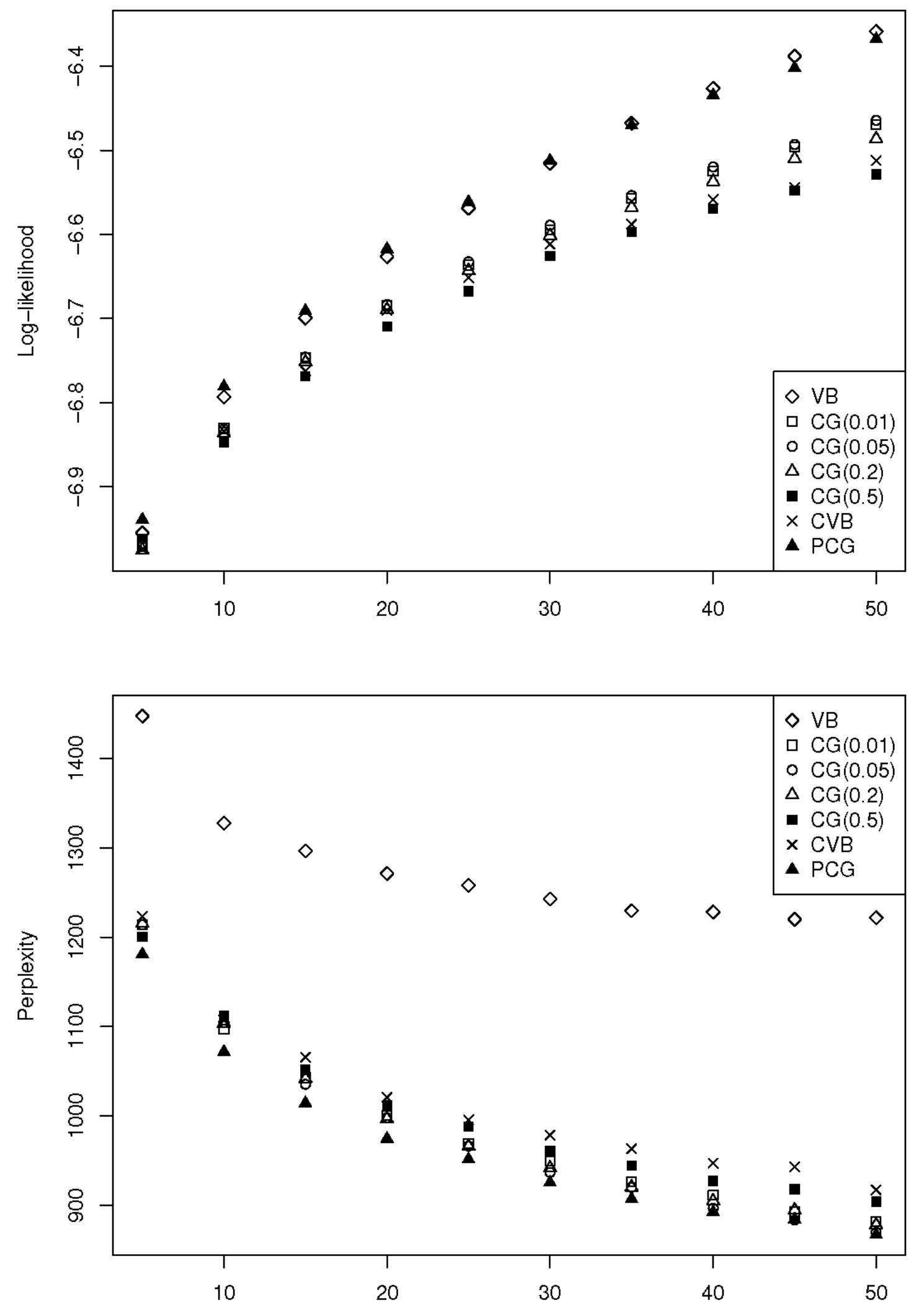

Figure 4: Method performance in terms of log-likelihood (top) and perplexity (bottom) for the NIPS conference and BBC news data.

6.3 Sellected hyperparameters α_d

- Two different resourses constituting the corpus explain why the corpus largely devided into two clusters for all number of topics.
- The documents in the NIPS set largely have a large α_d, while those in the BBC data do smaller ones.

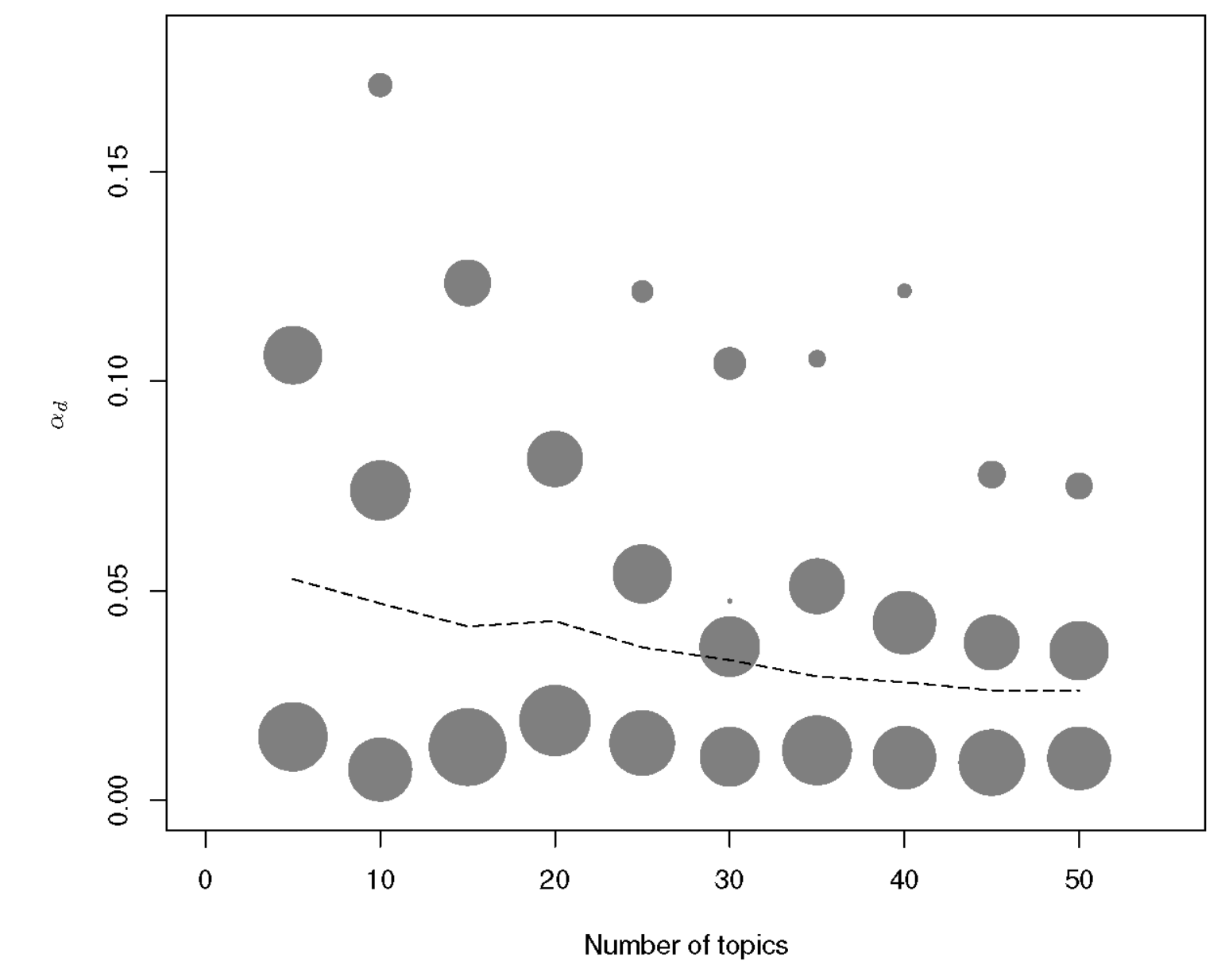

Figure 5: Sampled atoms of α_d for the DP mixture prior for the hyperparameter of the distribution of $\boldsymbol{\theta}^{(d)}$ with the corresponding cluster sizes. The dashed lines represent the weighted average of the sampled atoms of α_d with the weights being cluster sizes.

7 Conclusions

- The proposed enhanced LDA model allows flexible hyperparameter modeling of the distributions of topic contributions to a given document and word contributions to 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.
- 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.