LDA Model Selection: A Marginal Likelihood Method[*] [†]

Naijia Liu[‡]
naijial@princeton.edu

August 30, 2018

Abstract

Latent Dirichlet Allocation (LDA) topic model enables researchers to detect hidden structures within text corpus. Due to model complexity, the marginal likelihood of LDA model is intractable for direct calculation. Both Gibbs sampler and variational method enable researchers to estimate LDA model. However, very little attention is paid on selecting the important parameters in LDA model. The most commonly adopted methods for selecting the topic number are cross-validation and computing the log-likelihood for the held-out data. Due to the sparsity of text data, these methods often tend to overfit the number of topics in a given corpus. This paper proposes an alternative method to estimate the number of topics by approximating marginal likelihood of Latent Dirichlet Allocation topic model, both under the estimation regime of variational EM and Gibbs sampling. Furthermore, this paper offers discussion over the difference between Gibbs and variational EM and why Gibbs sampler is superior in terms of model fitting. This paper also presents simulated comparison results in favour of the marginal likelihood approximation methods, and also an application on Supreme court data. The R packeage `TMMarginal` is available upon request to implement all methods in the paper.

[*] I thank Prof. John B. Londregan for all of his help and support since the start of this project.

[†] R-package `TMMarginal` to implement the proposed method is available upon request.

[‡] PhD candidate, Department of Politics, Princeton University.

1 Introduction

1.1 Text Analysis in Social Science

Topic models under Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is an unsupervised way to analyse text data. By running the LDA model under a specified prior, one gets the relative topic weights and word distribution under each topic within each document. It has been widely adopted in political science since its introduction. Mueller and Rauh (2017) run LDA models on a newspaper corpus to predict political events based on variations among topic weights over time. Kim (2017) studies firm-level lobbying based on the disclosed reports and claims that certain topics correlate with a higher probability of lobbying. Lauderdale and Clark (2014) use the LDA model to study relative issue importance in Supreme Court cases, to aid the voting data for scaling purposes. The LDA topic model enables computers to learn about and describe text data with a minimal human prior, and thus yielding a more convincing result.

However, all of the aforementioned study results rely heavily on an accurate specification of the LDA model. In other words, the study results may change based on different prior parameters authors choose to use. In fact, the LDA model offers high flexibility in terms of prior selection for researchers, such as the number of topics in the corpus. Therefore, the ability to assess the model performance is essential to validate the result produced by the topic model.

Little guidance has been provided to empirical researchers when choosing the number of topics under LDA. This paper proposes a means to provide a stable and accurate estimation of the most probable number of dimensions, compared to existing methods. This is an important but understudied area of topic models. The ability of correctly specifying the number of topics does not only cast more confidence in the model result, it also potentially provides a new way to study and compare texts for social scientists. One may argue that, knowing the true number of topics for one text dataset does not provide much additional information to human knowledge, especially in the cases where the length is reasonable for a human expert to read. However, this paper aims to propose a method for the situations in which researchers are dealing with many lengthy datasets. It can estimate the number of topic without human experts' reading, which could be time-consuming as the document length increases. Furthermore, it can compare number of topics between these datasets that might be interesting to many social science research questions.

1.2 Related Literature

The most commonly adopted model selection method is to split the text corpus into a training set and a held-out set. After training the LDA model using the training set, we calculate the likelihood of this model using the held-out dataset (also called the model perplexity) (Griffiths and Steyvers, 2004; Cao et al., 2009; Arun et al., 2010; Deveaud et al., 2014; Roberts et al., 2014).

$$\text{Perplexity}(D_{\text{heldout}}) = \exp \frac{-\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{M} N_d} \tag{1}$$

Due to the sparsity of text data, it is rarely the case that training and testing sets are similar to each other. As a result, perplexity measure often overestimates number of topics to compensate for the difference between the two sets. The simulation results shown later in this paper confirms this shortcoming. Wallach et al. (2009) also present evidence that this empirical likelihood method tends not to offer accurate estimation of held out documents and propose instead a more efficient method called document completion, which“completes” the document by predicting second half of the same document. However, this method shares substantial similarity with perplexity method. Instead of splitting the whole dataset, splitting each individual document mitigates the bias of perplexity, but does not solve the problem from its root.

Finally, Taddy (2012) proposed a marginal likelihood-based model selection method that facilitates choosing the number of latent topics, which maximizes the marginal posterior probability. In line with Taddy (2012)'s work, this paper proposes an alternative method to approximate the marginal likelihood of LDA topic model, both under the estimation regime of variational EM and Gibbs sampling (Blei et al., 2003). This method is an application of the estimation technique proposed by Chib (1995); Basu and Chib (2003) on the LDA topic model with modifications tailored towards it. Instead of only using the estimation on marginal posterior probability, this method utilizes more information from the model by "exploiting the fact that the marginal density can be expressed as the prior times the likelihood function over the posterior density"(Chib, 1995). Furthermore, accuracy can be improved by choosing a high density point. In this LDA setting, I use the maximum a posteriori point estimated by the model.

Paper Layout

In section 2.1.1, I will introduce the basic setup of the LDA topic model by Blei et al. (2003) and in section 2.1.2 the marginal likelihood method by Chib (1995). Then, in section 2.2, I will demonstrate briefly how to apply the Chib (1995) method to approximate marginal likelihood for the LDA model. In section 3, I present simulation results to compare some of the aforementioned model selection methods with the proposed one. In section 4, I demonstrate that marginal likelihood method provides a better model fit using Supreme Court opinion texts and voting data Lauderdale and Clark (2014).

2 Proposed Method

The proposed method aims to estimate the marginal likelihood of topic models under LDA. Model selection then can be conducted by maximizing the marginal likelihood. I will talk in detail about LDA model and how to estimate the marginal likelihood of it in this section.

2.1 Model Setup

2.1.1 LDA Topic Model

The LDA topic model takes on a generative view of text data. The model consists of three layers: A corpus D consists of several documents of a certain average length; A document consists of K latent topics; A topic consists of N words. Each layer is characterized by a distributional assumption, which is shown below:

1. Choose $N \sim \text{Poisson}(\xi)$
2. Choose $\theta \sim \text{Dir}(\alpha)$
3. For each of the N words w_n
    (a) Choose a topic $Z_n \sim \text{Mult}(\theta)$
    (b) Choose a word $W_n \sim \text{Mult}(W_n|Z_n, \beta)$

The length of corpus D (in other words, the size of vocabulary) is determined by a Poisson distribution. For each document, we use a Dirichlet distribution with concentration parameter α to determine the topic weight θ. Finally, each word has a topic assignment Z_n, which is determined by a multinomial distribution given β.

If each parameter were to be accurately estimated, the model would return a "correct" and comprehensive description of the data. We would be able to know the weights of each topics and the probability of each word under a given topic. Unfortunately, the updated marginal likelihood (which is shown by equation (2)) is intractable due to the coupling of θ and β (Blei et al., 2003).

$$\begin{aligned} p(w|\alpha, \beta) &= \int p(\theta|\alpha) \prod_{n=1}^{N} \sum_{Z_n} p(Z_n|\theta) p(w_n|Z_n, \beta) d\theta \\ &= \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \int \prod_{i=1}^{K} \theta_i^{\alpha_i - 1} \prod_{n=1}^{N} \sum_{i=1}^{K} \prod_{j=1}^{V} (\theta_i \beta_{ij})^{w_n^j} d\theta \end{aligned} \tag{2}$$

2.1.2 Model Selection

Instead of dealing with the intractable integral in equation (2), Chib (1995) proposes to choose a point θ^* in the posterior.

$$\log \widehat{m}(y) = \log f(y|\theta^*) + \log \pi(\theta^*) - \log \widehat{\pi}(\theta^*|y) \tag{3}$$

In equation (3), $\widehat{m}(y)$ denotes the marginal likelihood of the dataset y. $f(y|\theta^*)$ is the likelihood of the model. $\pi(\theta^*)$ denotes the prior of the model. And $\widehat{\pi}(\theta^*|y)$ denotes the posterior probability of the model, which can be obtained by taking the means of MCMC draws. Chib (1995) provided a detailed proof that this method provides stable estimation for the marginal likelihood. Furthermore, for a given number of posterior draws, the density is likely to be more accurately estimated at a high density point for the estimation of $\pi(\theta^*|y)$. As a result, I use the maximum a posteriori (MAP) point of the model.

2.2 Marginal Likelihood for Topic Models

The goal (stated in equation (4)) of the model is to estimate the best number of topics $\tilde{K}$ for a given text corpus D, by maximizing its marginal likelihood $\hat{m}(D)$. By the assumption in section 2.1.1, a Poisson distribution over the size of corpus enables us to conduct the calculation in equation (4): summing up log-likelihood of each document d to get the log-likelihood of the corpus D. I assume conditional independence among each document given the Poisson distribution, but penalize certain topics in each document with almost zero posterior.

Finally, (α^*, θ^*) is the posterior point estimated by LDA model. Throughout the paper, I assume this is estimated from variational EM by Blei et al. (2003). The accompanying R code can also deal with LDA estimates from Gibbs sampler. However, due to the nature of Gibbs sampling, the estimation is not as stable as the one obtained by variational EM.

$$\widetilde{K} = \arg\max_k \sum_{d=1}^{D} \log \hat{m}(d) \tag{4}$$

$$\log \hat{m}(d) = \log p(d|\alpha^*, \theta^*, k) + \log \pi(\alpha_0, \theta_0, k) - \log \hat{\pi}(\alpha^*, \theta^*, k|d) \tag{5}$$

Prior probability $\pi(\alpha_0, \theta_0, k)$ is calculated by equation (2), for a given α and θ. I adopt the common practice of assigning a symmetric prior to the sparsity parameter, where for $\vec{\alpha}_0 : \alpha_1 = \ldots = \alpha_k$, and for $\vec{\theta}_0 : \theta_1 = \ldots = \theta_k = \frac{1}{k}$.

$$
\begin{aligned}
\pi(\alpha_0, \theta_0, k) &= \frac{\Gamma(k\alpha_0)}{(\Gamma\alpha_0)^k}\left(\frac{1}{k}\right)^{k(\alpha_0-1)} k(\theta_0\beta)^{n_k} \\
&= \frac{\Gamma(k\alpha_0)}{(\Gamma\alpha_0)^k}\left(\frac{1}{k}\right)^{k(\alpha_0-1)} k\left(\frac{1}{kn_k}\right)^{kn_k}
\end{aligned}
$$

The posterior at the chosen point is estimated by taking the mean of Monte Carlo draws. For each document with each topic k with n_k words:

$$
\hat{\pi}(\alpha^*, \theta^*, k|d) = \frac{\Gamma k\alpha^*}{(\Gamma\alpha^*)^k}\prod_{i=1}^{K}\hat{\boldsymbol{\theta}}_i^{*\alpha^*-1}\sum_{i=1}^{K}\sum_{j=1}^{n_k}(\hat{\boldsymbol{\theta}}_i^*\beta_{ij}^*)^{n_k}
$$

where $\hat{\boldsymbol{\theta}}^* = \frac{1}{G}\sum_{g=1}^{G}\pi(\theta^g)$ is obtained by G Monte Carlo multinomial Dirichlet draws with a starting value at θ^* and parameter of α^* (Martin et al., 2011). Furthermore, β_{ij}^* is truncated given words per topic n_k, such that $\vec{\beta^*} = \beta_1, \beta_2, \cdots, \beta_{n_k}, 0, \cdots$. This truncation prevents the posterior estimation from being too close to zero.

3 Simulation and Comparison

The simulation dataset is generated under the distribution of LDA, with the number of topics $K = 5$, words per topic $n_k = 50$, number of documents $D = 20$ and a vocabulary size of $N = 200$.

The simulated dataset is a good way to test the model performance, since the number of topics is known. However, it is generated to be much less sparse than the real-life text data. This is a relatively easy test for the model since, with fewer zero entries in the document term matrix, it is easier to learn the pattern. As a result, the same comparison is done with the well-known `AssociatedPress` data from the `topicmodels` package (Hornik and Grün, 2011). This dataset consists of 2246 documents (short news coverage from the Associated Press), with a vocabulary size of 10473.

I compare the proposed method with three other R-packages commonly adopted by social scientists: topicmodels, `stm` (Roberts et al., 2016) [1], and `ldatuning` (Murzintcev, 2015). [2] The summary result is presented in table 1. I also visualized the comparison results to show the convergence pattern of each method.

[1] Heldout likelihood method follows Wallach et al. (2009)'s work, residual method follows Taddy (2012)'s work, which were both discussed earlier in the paper. Semantic coherence by Mimno et al. (2011) aims to maximize the co-occurrence of words in a given topic frequency. Lower bound shows the convergence of stm model.

[2] Arun et al. (2010) aims to minimize a distance measure (KL divergence) between word distributions, claiming that the distance tends to be higher with a non-optimal topic number. Griffiths and Steyvers (2004) advocates for the use of log-likelihoods based on full set of the data. Cao et al. (2009) propose a density based method to select the number of topic.

Table 1: Comparison Results

Method	K=5 Simulated	Associated
Marginal Likelihood	**5**	**15**
`topicmodels` Perplexity	6	30
`stm` Heldout Likelihood	15	2
`stm` Residual	9	5
`stm` Semantic Coherence	2	30
`stm` Lower Bound	15	30
`ldatuning` Arun2010	15	30
`ldatuning` CaoJuan2009	15	30
`ldatuning` Griffiths2004	15	5

4 Application

Lauderdale and Clark (2014) developed a method to combine votes and opinion texts from the U.S. Supreme Court. They adopt LDA to first uncover the relative importance of topics for each dimension d in document j λ_{jd}, then fit an IRT model using the weights of topics. Rest of the model setup follows conventional IRT model (See equation (6)). α_j, β_j are discrimination parameters for each vote j. θ_{id} are issue specific ideal point for each voter. Vote $y_{ij} = 1$ if y^*_{ij} is greater than 0, and $y_{ij} = 0$ otherwise.

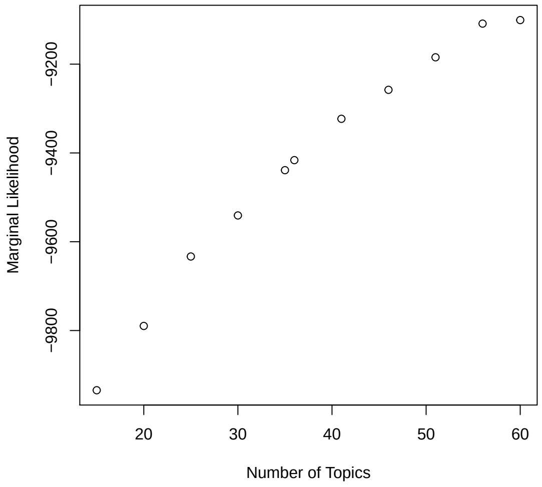
$$\begin{aligned} y^*_{ij} &\sim N(\alpha_j + \sum_{d=1}^{D} \beta_{jd}\theta_{id}, 1) \\ \theta_{ij} &= \sum_{d=1}^{D} \lambda_{jd}\theta_{id} \end{aligned} \tag{6}$$

In their paper, Lauderdale and Clark (2014) calculated DIC on only a subset of the full data. And they concluded that $k = 24$ is a reasonable number for the dataset. This practice, though time-saving, potentially introduces problematic estimation. First of all, as discussed in the method section, randomly sampling text data hardly ever gives researchers a representative subset of the full dataset. Secondly, as shown in their paper, DIC keeps fluctuating up and down even at $k = 40$. I argue that the choice of $k = 24$ both lack theoretical and empirical foundation.

To determine an appropriate number of topics for LDA model in the first step, I apply the marginal likelihood method on the full dataset. Figure 1 presents the full dataset marginal likelihood calculated using `TMMarginal`. This is marginal likelihood estimated using the full dataset, hence one should not be worried about the sampling quality. The result suggests a much higher k than 24. I will proceed with the IRT model using $k = 60$.

With more topics in LDA model, more issue areas are discovered from the dataset. As shown

Figure 1: Marginal Likelihood of Full Dataset



in figure 2, LDA topics are mapped with the Supreme Court Database issue area coding (Spaeth et al., 2014). Darker color indicates a stronger correlation between the two. For example, ”search, warrant, forth” is strongly correlated with ”criminal procedure” in Spaeth issue. All 60 topics have correlations with one or multiple of the 12 Spaeth issues. This is evidence showing that all 60 of them are relevant and distinct dimensions of the dataset.

LDA topic model reveals that some of the more complicated Spaeth issues bear multiple dimensions. For example, ”criminal procedure” is strongly correlated with 12 topics, keywords including ”jury, appeal, petition, wiretap, search warrant, trial, offense, afdc” and more. ”Economic activity” is strongly correlated with more than 15 topics, keywords including ”bank, tax, EPA, FDA, commerce, congress, land, immunity, bankruptcy, insurance, antitrust, gas” and more. These findings indicate that Spaeth issues only capture the grand topics and tend to ignore the more detailed divisions.

Figure 3 show the comparative stance between Justice Kennedy and Justice O’Connor. Both of them were swing medians in the Supreme Court and this graph shows the variation by issue. Most of the scaling remains the same from original study. Justice Kennedy is more conservative in issues like education, religion and discrimination, while he is more liberal in issues like commerce, jury and labor union.

One difference between $k = 60$ LDAIRT model and $k = 24$ model is the widened confidence interval. This is inevitable as we add more topics into the model. The decreasing statistical power of the scale also matches with the reality that two of them were the swing median of the Supreme Court.

Figure 2: Spaeth Issue Areas Versus LDA Issues

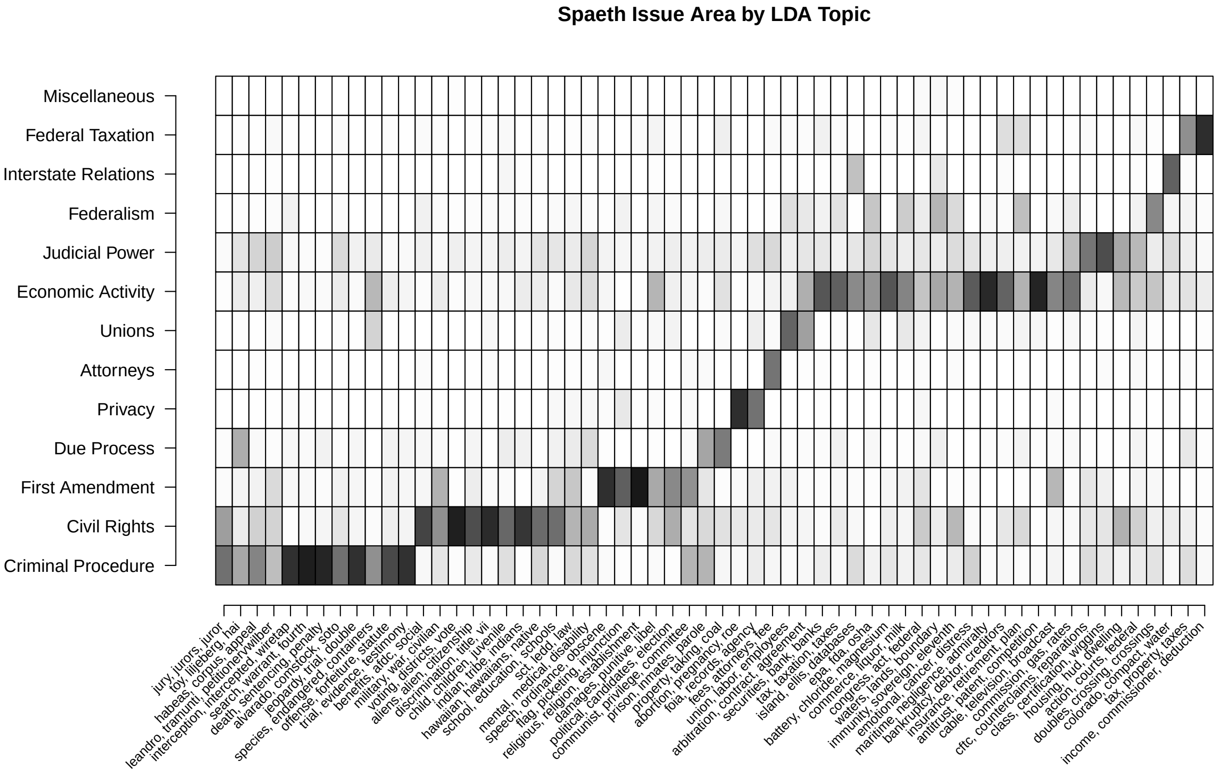
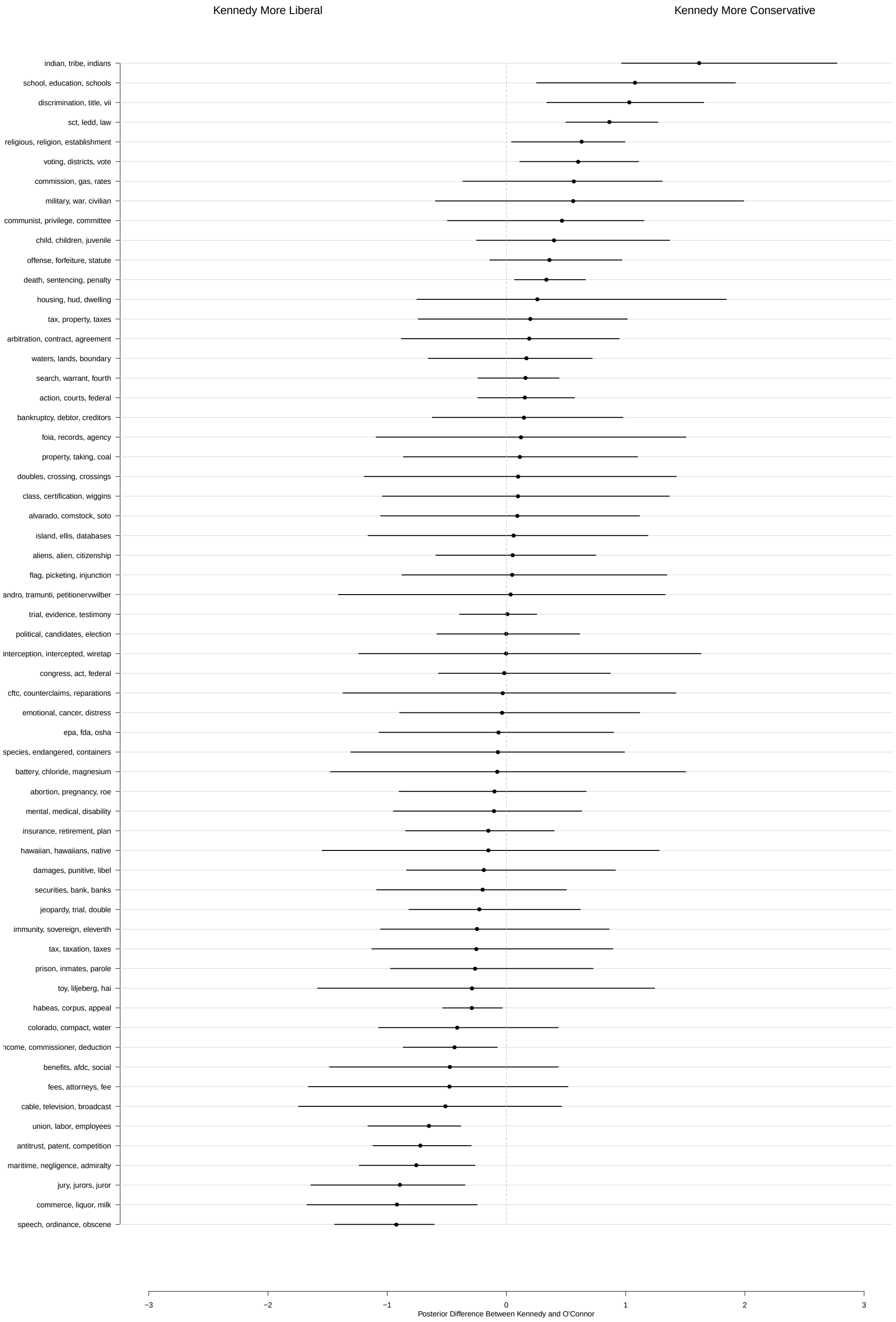


Figure 3: Comparison of Justice Kennedy's and Justice O'Connor's Ideal Points on Each Issue Dimension



5 Discussion

More attention should be given to LDA topic model specification, with an increasing number of social science studies adopting the model. This paper proposes a new method of estimating dimensionality of the model by maximizing marginal likelihood of LDA. The application on Supreme court opinion text and voting data shows that a correct specification of LDA model can uncover more information from the given dataset.

One future step is to address the robustness of LDA model. Miller and Harrison (2016) point out that the distributional assumption of LDA topic model may not be accurate at first place, thus making the estimation of number of topics less meaningful. However, simply ignoring model selection for this reason is also a questionable practice

Both Gibbs sampler and variational EM provide an approximated log likelihood of the model. The marginal likelihood method, in other words, is a weighted version of this log likelihood. The weight is determined by how far prior and posterior draws are from each other. This weighted likelihood incorporates more information from both prior and posterior of the model. This approach leads to a more accurate estimation of marginal likelihood, in trade off with a longer computational time. As what is shown in figure 4, the models fitted by Gibbs in general have a higher likelihood than VEM method. Both method select the number of topics to be around 10, which is the simulated truth. However, Gibbs sample is sensitive to small changes in number of topics and the loglikelihood fluctuates accordingly. Due to the nature of variational method, once the KL-divergence reaches a certain threshold (Blei et al., 2003) the results tend to be stable.

When dealing with small size of text data, Gibbs sampler is guaranteed to reach the true value after certain iterations. It is worth noting that, in general, Gibbs sampler provides a better model fit then VEM method (See figure 4). For larger size of dataset, VEM enables a shorter computing time. However, it is a trade off for researchers to choose between computing time and finding out a more accurate model selection.

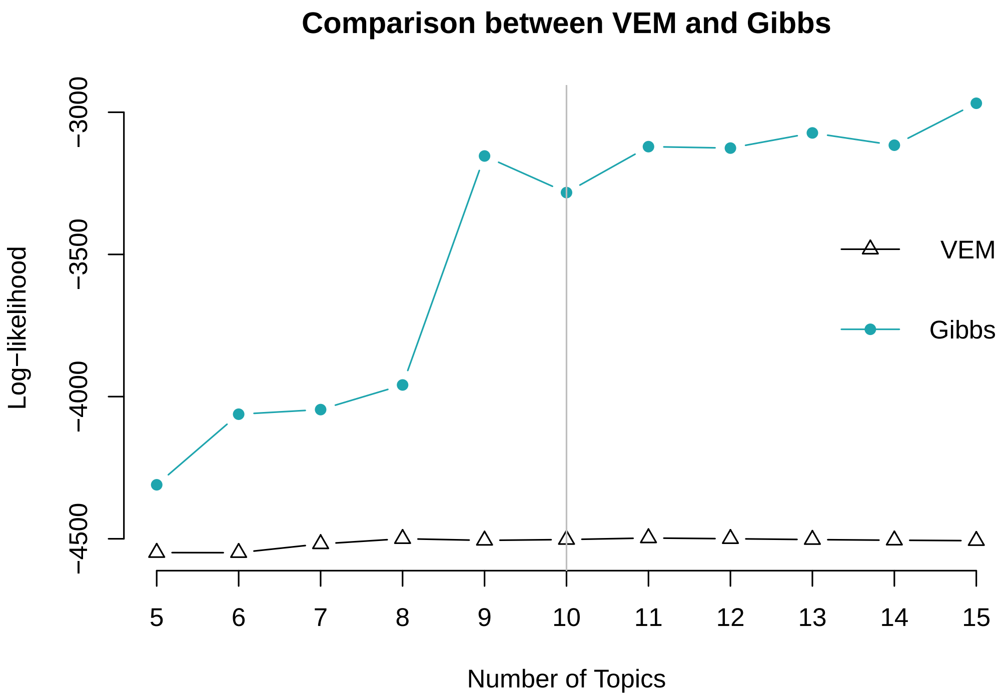

Comparison between VEM and Gibbs
Log−likelihood
−3000
−3500
−4000
−4500
VEM
Gibbs
5
6
7
8
9
10
11
12
13
14
15
Number of Topics

Figure 4

References

Arun, R., Suresh, V., Madhavan, C. V., and Murthy, M. N. (2010). On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 391–402. Springer.

Basu, S. and Chib, S. (2003). Marginal likelihood and bayes factors for dirichlet process mixture models. *Journal of the American Statistical Association*, 98(461):224–235.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Cao, J., Xia, T., Li, J., Zhang, Y., and Tang, S. (2009). A density-based method for adaptive lda model selection. *Neurocomputing*, 72(7-9):1775–1781.

Chib, S. (1995). Marginal likelihood from the gibbs output. *Journal of the american statistical association*, 90(432):1313–1321.

Deveaud, R., SanJuan, E., and Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique*, 17(1):61–84.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235.

Hornik, K. and Grün, B. (2011). topicmodels: An r package for fitting topic models. *Journal of Statistical Software*, 40(13):1–30.

Kim, I. S. (2017). Political cleavages within industry: firm-level lobbying for trade liberalization. *American Political Science Review*, 111(1):1–20.

Lauderdale, B. E. and Clark, T. S. (2014). Scaling politically meaningful dimensions using texts and votes. *American Journal of Political Science*, 58(3):754–771.

Martin, A. D., Quinn, K. M., and Park, J. H. (2011). MCMCpack: Markov chain monte carlo in R. *Journal of Statistical Software*, 42(9):22.

Miller, J. W. and Harrison, M. T. (2016). Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, 113(521).

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., and McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the conference on empirical methods in natural language processing*, pages 262–272. Association for Computational Linguistics.

Mueller, H. and Rauh, C. (2017). Reading between the lines: Prediction of political violence using newspaper text. *American Political Science Review*, page 118.

Murzintcev, N. (2015). ldatuning: Tuning of the latent dirichlet allocation (lda) models prameters. *R package.*

Roberts, M. E., Stewart, B. M., and Tingley, D. (2016). Stm: R package for structural topic models, 2014. *URL http://www. structuraltopicmodel. com. R package version*, 1(8).

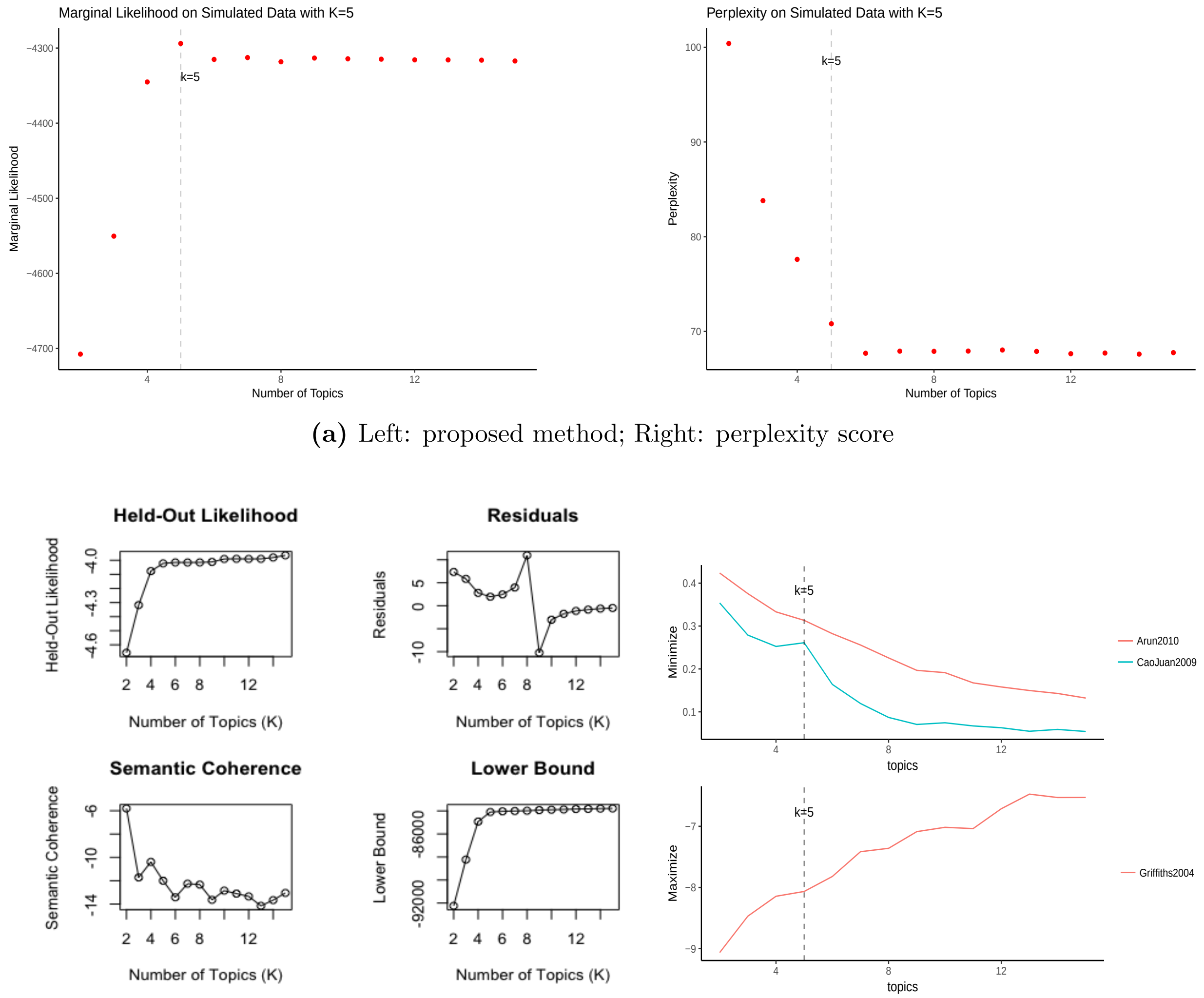
Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., and Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4):1064–1082.

Spaeth, H. J., Benesh, S. C., Epstein, L., Martin, A. D., Segal, J. A., and Ruger, T. J. (2014). U.s. supreme court judicial database: St. louis, washington university. *Supreme Court Database.*

Taddy, M. (2012). On estimation and selection for topic models. In *Artificial Intelligence and Statistics*, pages 1184–1193.

Wallach, H. M., Murray, I., Salakhutdinov, R., and Mimno, D. (2009). Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning*, pages 1105–1112. ACM.

Figure 5: Simulated Dataset



(a) Left: proposed method; Right: perplexity score

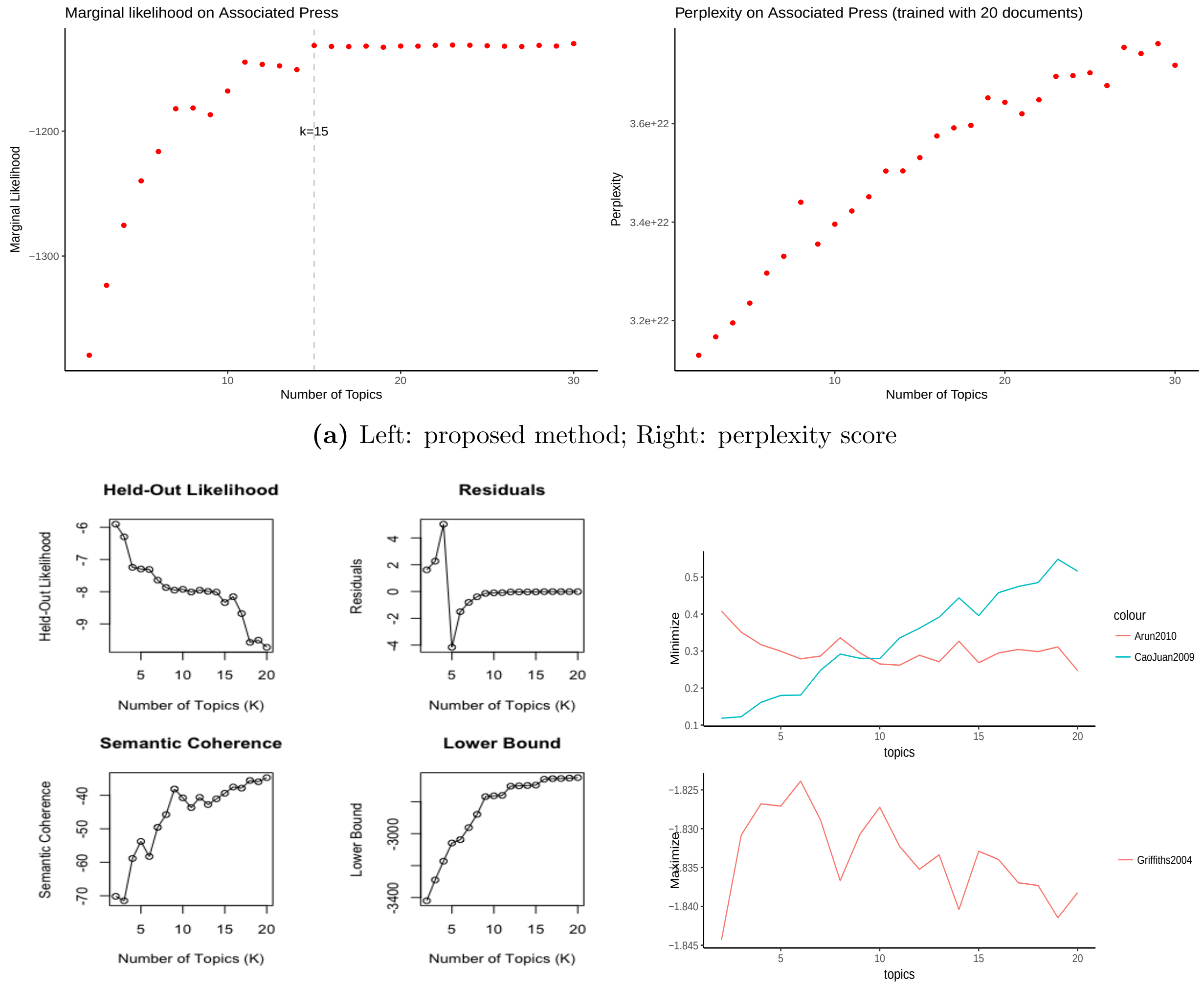
(b) Left: stm; Right: ldatuning

A Appendix

As expected, in figure 5, perplexity score does a decent job since that dataset is strictly simulated with the same data generating process. This is the same with the held-out likelihood by the `stm` package, since they adopt almost the same method (held-out portion might differ). The `ldatuning` package results are plotted in two groups depending on if the goal is to minimize or to maximize the value. Note that all three estimators by `ldatuning` indicate that 15 or more topics exist in the data. The proposed method recovers $K = 5$ exactly, while others fail to provide us with stable estimates.

Figure 6 presents the comparison between these methods using the Associated Press dataset. Due to computing limitations, I used a subset of 11 documents from the whole dataset, with a

Figure 6: Associated Press Data



(a) Left: proposed method; Right: perplexity score

(b) Left: stm; Right: ldatuning

sparsity score of 99%.

Hornik and Grün (2011) conducted a 10-fold cross-validation on the whole dataset to test their model performance. They concluded that "40 topics are suggested as optimal for both VEM and Gibbs methods." After some approximation, I conclude that the 11 documents contain 10-15 topics, with each document consisting of fewer than (or equal to) 2 topics.

As shown in figure 6, both `stm` and `ldatuning` package gave self-conflicting estimations, while the perplexity measure fails to converge.[3] The proposed method offers a reasonable estimation at $K = 15$.

[3] Note that for the perplexity score here, the LDA model is tested on another 20 documents from the same Associated Press dataset, and then the perplexity score is calculated for the 11 documents.