Diagnosing and Improving Topic Models by Analyzing Posterior Variability

Linzi Xing , Michael J. Paul
University of Colorado Boulder
{linzi.xing, mpaul}@colorado.edu

Introduction

Bayesian inference methods have been widely used at obtaining more robust model estimates like LDA[1]. However, little work has explored the possibility that Bayesian inference can also be used to evaluate and understand other characteristics of the topic model. We focus on the variability of posterior samples of topic parameters across Gibbs sampling process and find the fluctuation in parameters can indicate the quality, consistency of topics. When narrow down to the word level, the fluctuation of word probability may improve the final outcome of topic models.

Topic-level Analysis

We newly define a metric named **topic stability** to measure the degree of a topic's parameters change during Gibbs sampling, a posterior inference algorithm..

$$\boldsymbol{stability}(\boldsymbol{\Phi}_k) = \frac{\mathbf{1}}{|\boldsymbol{\Phi}_k|} \sum_{\varphi_k \in \boldsymbol{\Phi}_k} \boldsymbol{sim}(\varphi_k, \bar{\varphi}_k)$$

After experimenting, we select **cosine similarity** as our vector similarity function ***sim()*** since it has better performance than other methods we consider in all cases.

Following the equation we proposed, each topic can be assigned a stability, we then try to align these stabilities with topic quality and consistency to test if it is a effective indicator of topic's quality and consistency[2]. We compare the correlation of topic stability with two other popular metrics – **coherence[3]** and **NPMI[4].**

Word-level Analysis

When focusing on the words within an individual topic, we also investigate the variability of posterior of individual word probability and its capability. We find that words with high posterior variance tend to be less strongly associated with the topic, like common words 'new' and 'said'. Hence, topic word list can be adjusted by variance to reorder the topic words in a better way.

We propose two methods for using the posterior variability to re-rank the top words in a certain topic.

- **Mean/SD:** dividing the mean word probability by the standard deviation across all the samples.
- **Min:** taking the lowest value φ_{kv} (the word probability assigned to topic k), which is the 0th percentile of the value distribution.

Table 1: Three different posterior samples of two topics (highest and lowest stability) from iteration 1000, 1600 and 2000. Words only show in one column are highlighted.

1000		1600		2000	
Topic 6 (News, Stability = 0.9334)					
housing	.027	store	.023	store	.023
stores	.021	stores	.023	stores	.022
store	.019	homeless	.019	homeless	.018
homeless	.018	food	.014	food	.013
home	.015	christmas	.012	christmas	.013
food	.012	market	.011	animals	.010
christmas	.011	clothing	.008	market	.009
animals	.010	animals	.008	video	.008
city	.009	video	.008	bought	.008
shopping	.007	shopping	.008	owner	.007
Topic 52 (Wiki, Stability = 0.9999)					
age	.058	age	.059	age	.059
population	.037	population	.037	population	.037
median	.029	median	.029	median	.029
income	.028	income	.028	income	.028
census	.027	census	.027	census	.027
living	.025	living	.025	living	.025
households	.025	households	.025	households	.025
average	.024	average	.024	average	.024
years	.023	years	.024	years	.024
families	.023	families	.023	families	.023

Topic	Method	Top 10 words
Topic 8	Mean	said ship water coast river boat sea guard island species
	Mean/SD	ship species coast water birds boat sea fish guard ships
	Min	ship water coast boat river sea species island ships fish
Topic 22	Mean	television network cbs nbc news tv abc million broadcast rating
	Mean/SD	cbs nbc network abc rating radio television cable cnn broadcast
	Min	network television cbs nbc tv abc news broadcast rating cable
Topic 74	Mean	house building built castle th tower buildings city hall garden
	Mean/SD	building house built tower buildings garden castle designed hall design
	Min	building house built tower buildings garden castle hall houses site

Table 2: Example of topic representations of three methods, where Mean is the baseline method of using the average sample probability. Highlighted words indicate the words that only appear in the set for that particular method.

	Quality		Consistency	
Metrics	***News***	***Wiki***	***News***	***Wiki***
Stability	.248	.253	.627	.354
Coherence	.198	.040	.456	.298
NPMI	.553	.462	.340	.142

Table 3: Correlation between metrics and topic quality, consistency

	Mean vs Mean/SD		Mean vs Min		Mean/SD vs Min	
News						
3/5	16	**34**	19	**30**	24	**26**
4/5	10	**21**	6	**13**	7	**9**
5/5	0	**1**	1	**3**	0	0
Wiki						
3/5	38	**62**	39	**52**	**56**	44
4/5	16	**35**	17	**23**	**23**	15
5/5	1	**9**	0	**7**	**7**	3

Table 4: Number of times of three methods win majority vote

Experiments

Datasets: All experiments are done on two datasets respectively.

- *News:* 2,243 articles from Associated Press (50 topics)
- *Wiki:* 10,000 articles randomly picked from Wikipedia (100 topics)

LDA settings: 2000 iterations(1000 burn-in iterations), 10-sample lag.

Topic-level Analysis: We collected quality judgments from humans by having people rate topics on a 4-point Likert scale (4-best, 1-worst) through Amazon Mechanical Turk.

- ***Baseline:*** coherence[3], NPMI[4]
- ***Correlation with manually rated quality:*** compute Spearman's rho between human ratings and three metrics on two datasets.
- ***Correlation with consistency across models:*** run LDA four times on each corpus and applied the up-to-one topical alignment process[2], using a cosine similarity threshold of 0.2.

Word-level Analysis:

- ***Baseline:*** simple mean of φkv across all the samples
- ***Comparison on human ratings:*** apply the same 4-point Likert scale on topics before and after adjusting and compute the average scores.
- ***Comparison on human voting:*** pair topics from three different methods and require human to compare and pick the better one, counting the method which wins the majority vote.

Discussion

Topic Level:

- Topic stability is correlated with consistency and quality of topics rated manually. It can beat one of two topic quality evaluation metrics.
- Different from coherence and NPMI, topic stability doesn't use any information about words in a certain topic.

Word Level:

- Variability of words assigned to certain topics is used to adjust the topic word lists by Mean/SD and Min we proposed. Experiments show people prefer our modification more.

Future Work:

- In future, it's worthy to explore the feature of variability at document level[14].

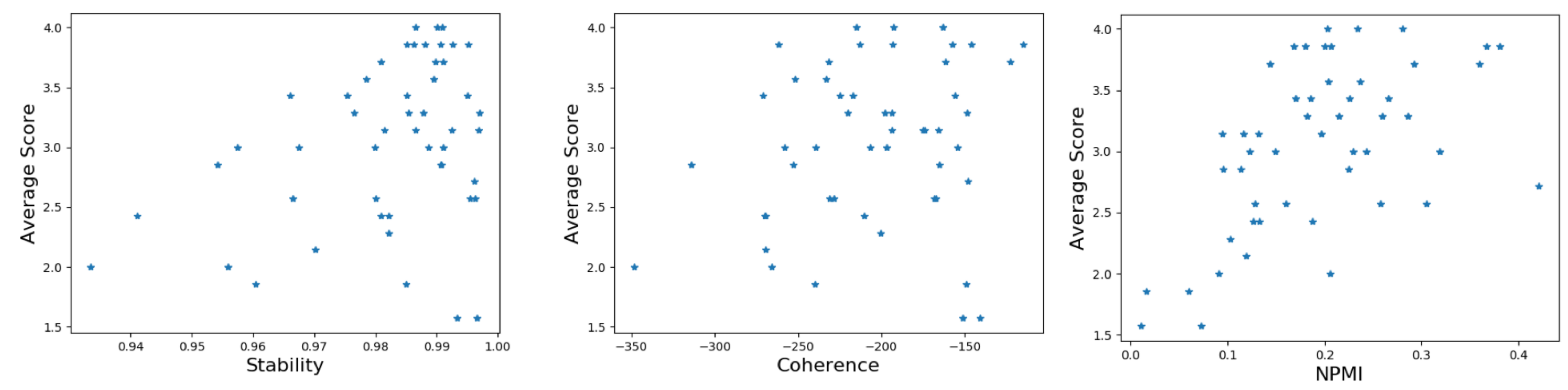

Figure 1: Manually rated topic scores along with three metrics: topic stability, coherence and NPMI on the News corpus.

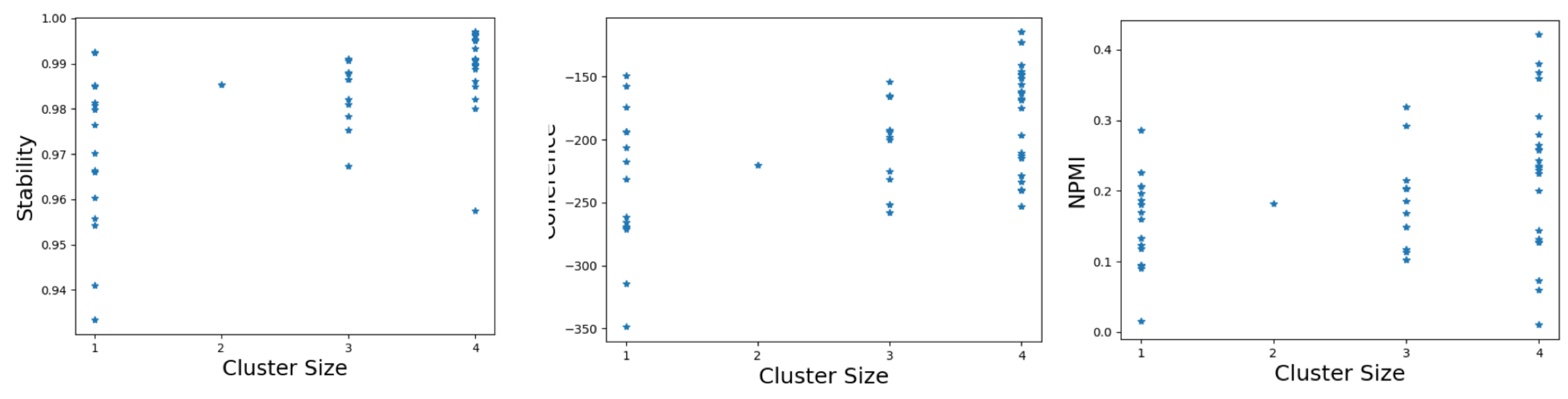

Figure 2: Distribution of the topic stability, coherence and NPMI scores within different sized clusters on the topic alignment task for the News corpus.

References

1. Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet allocation. Journal of Machine Learning Research3:9931022.
2. Chuang, J.; Roberts, M. E.; Stewart, B. M.; Weiss, R.; Tingley, D.; Grimmer, J.; and Heer., J. 2015. TopicCheck: Interactive alignment for assessing topic model stability. In HLT-NAACL, 175–184.
3. Mimno, D., and Blei, D. 2011. Bayesian checking for topic models. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 227–237.
4. Lau, J. H.; Newman, D.; and Baldwin., T. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In Proceedings of EACL 2014, 530–539.