Diagnosing and Improving Topic Models by Analyzing Posterior Variability

Linzi Xing
Department of Computer Science
University of Colorado, Boulder, CO 80309
linzi.xing@colorado.edu

Michael J. Paul
Department of Information Science
University of Colorado, Boulder, CO 80309
mpaul@colorado.edu

Abstract
Bayesian inference methods for probabilistic topic models can quantify uncertainty in the parameters, which has primarily been used to increase the robustness of parameter estimates. In this work, we explore other rich information that can be obtained by analyzing the posterior distributions in topic models. Experimenting with latent Dirichlet allocation on two datasets, we propose ideas incorporating information about the posterior distributions at the topic level and at the word level. At the topic level, we propose a metric called topic stability that measures the variability of the topic parameters under the posterior. We show that this metric is correlated with human judgments of topic quality as well as with the consistency of topics appearing across multiple models. At the word level, we experiment with different methods for adjusting individual word probabilities within topics based on their uncertainty. Humans prefer words ranked by our adjusted estimates nearly twice as often when compared to the traditional approach. Finally, we describe how the ideas presented in this work could potentially applied to other predictive or exploratory models in future work.

Introduction
Topic models, which extract themes from text datasets, have been widely used for large-scale corpus analysis with diverse applications including the study of topics in scientific articles (Hall, Jurafsky, and Manning 2008), finding patterns in classic literature (Jockers and Mimno 2013), understanding news media coverage (Roberts et al. 2013), and detecting population activities in online data (Paul and Dredze 2014). Of course, using topic models in these ways requires the belief that the topics meaningfully correspond to real concepts in a dataset. When topic models produce nonsensical topics or inconsistent outputs, it is difficult for a user to reliably use these models as a method of scientific inquiry. These challenges have motivated work on evaluating and understanding what it is that topic models discover (Chang et al. 2009; Mimno et al. 2011; Chuang et al. 2015).

In some sense, the lack of certainty about topics is already built into many commonly used models and inference algorithms. As a Bayesian model, the popular latent Dirichlet allocation (LDA) (Blei, Ng, and Jordan 2003) affords the possibility of inferring a posterior distribution over its parameters, which can provide an (approximate) measure of confidence of different parameter estimates. However, while Bayesian inference methods have been successful at obtaining more robust model estimates, they have mostly not been used for evaluating and understanding topic models.¹

In this work, we investigate what kinds of characteristics about topics can be learned from the posterior of the parameters, focusing on variability in the posterior. Specifically, we look at how LDA topic parameters change across different posterior samples during Gibbs sampling. We find that the level of fluctuation in the parameter estimates can provide insights into the quality, consistency, and salience of the topics and their word probabilities.

This paper is divided into two main sections:

• Topic-level analysis: We explore how the variability in a topic’s word distribution can be indicative of the quality and consistency of the topic. We propose a metric called topic stability that we show is correlated with human judgments of topic quality as well as with the consistency of the topic across multiple models.

• Word-level analysis: Within a topic, we examine how the probabilities of individual words vary, and find that words whose posteriors have high variability tend to be less salient and representative of the topic. We propose modifications to the ranking of words within a topic to adjust for this characteristic, and we show that people consistency prefer our modified rankings in experiments.

Our proposed methods are quite different from existing approaches. This leaves open a variety other possibilities for applying these ideas beyond what we investigated here, which we discuss after presenting our findings.

Topic Modeling
We use latent Dirichlet allocation (LDA) (Blei, Ng, and Jordan 2003) as our topic model. LDA has two types of parameters. Each document $d$ has a probability distribution over topics, $\theta_d$. Each topic $k$ is associated with a distribution over

¹An exception is work on posterior predictive checking of topic models (Mimno and Blei 2011), which diagnoses models by looking at discrepancies in how the model fits the data, but not variability of the posterior as proposed here.
words, $\phi_k$. Additionally, the parameters $\phi$ and $\theta$ have priors defined by the Dirichlet distribution. The number of topics $K$ must be specified, while the topic variables and parameters are unknown and must be inferred from the data.

Many algorithms existing for maximizing or inferring a posterior distribution over the latent variables and parameters (Blei 2012). We use Gibbs sampling (Geman and Geman 1984) as our posterior inference algorithm (Griffiths and Steyvers 2004) in this work. A Gibbs sampler generates samples of variable configurations from the posterior distribution. Each sample can provide a snapshot of the parameters, and our experiments explore how the topic model parameters, specifically the word distributions $\phi$, vary across the different samples.

While topics are defined by an entire distribution over the vocabulary, they are usually presented to humans by displaying the most probable words, usually a fixed number of words. We represent topics as a ranked list of 10 words in this work. Many of our experiments will focus on how these 10 words are perceived by people, and when we refer to a word being “in” a topic, we mean that the word is in the set of 10 most probable words.

### Evaluating Topic Models

A variety of work has developed methods to evaluate and characterize the quality and behavior of topic models (Boyd-Graber, Mimno, and Newman 2014). While it is possible to evaluate a topic model in a quantifiable predictive task (Wallach et al. 2009), it is non-trivial to evaluate the intrinsic quality of topics will be perceived by humans in exploratory tasks. Human feedback can be collected in the form of numeric ratings of quality (Newman et al. 2010) or by voting on different topic variants (Li and McCallum 2006). “Intruder” tasks conduct experiments to more objectively judge topic coherence by requiring people to identify out of place words or topics (Chang et al. 2009). Some “human-in-the-loop” systems are designed to interactively bring people into the modeling process, which is one way to help people diagnosis and improve their models (Hu et al. 2014; Chuang et al. 2015).

Due to the difficulty of obtaining human feedback, particularly when a large number of models are being considered and tuned, a number of automated alternatives to evaluation have been proposed. Most automated evaluations of quality focus on the semantic coherence of a topic—do the words form a cohesive group of related words? This is done by measuring the semantic similarity of the pairs of top words in the topic, usually using various co-occurrence statistics to estimate semantic similarity (Lau, Newman, and Baldwin. 2014; Roder, Both, and Hinneburg 2015). We will use two such metrics in this work:

The semantic coherence metric proposed by Mimno et al. (2011) is related to the sum of each conditional probability of each word in the topic given all other words, defined as:

$$C(t; V^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(v^{(t)}_m, v^{(t)}_l)}{D(v^{(t)}_l)} + 1$$  \hspace{1cm} (1)$$

where $t$ represents topic $t$ and $V^{(t)} = (v^{(t)}_1, \cdots, v^{(t)}_M)$

represents the $M$ most probable words assigned to this topic. $D(v^{(t)}_m, v^{(t)}_l)$ is the document frequency of word $v^{(t)}_m$ and $D(v^{(t)}_m, v^{(t)}_l)$ is the co-document frequency of both words.

The NPMI metric used by Lau et al. (2014) uses the normalized pointwise mutual information (Bouma 2009) of unique word pairs:

$$NPMI(t; V^{(t)}) = \sum_{m=1}^{M-1} \sum_{l=m+1}^{M} \frac{\log \frac{P(v^{(t)}_m, v^{(t)}_l)}{P(v^{(t)}_m)P(v^{(t)}_l)}}{\log P(v^{(t)}_m)P(v^{(t)}_l)}$$  \hspace{1cm} (2)$$

where $P(v^{(t)}_m)$ is marginal probability of word $v^{(t)}_m$ and $P(v^{(t)}_m, v^{(t)}_l)$ represents joint probability of both words.

### Experimental Setting

We now describe the datasets and experimental details that are used in both our topic-level and word-level analyses.

We experiment with two datasets. The News corpus contains 2,243 articles from the Associated Press. The Wiki corpus contains 10,000 articles from Wikipedia. We removed stop words and low-frequency words (appearing in fewer than five documents) from both datasets. Additionally, we removed proper nouns from the Wiki articles, following Chang et al. (2009), so that the topic model discovers more general concepts across the corpus. Statistics for the pre-processed datasets are provided in Table 1.

We set the number of topics to 50 for News and 100 for Wiki. We ran the Gibbs samplers for a burn-in period of 1,000 iterations, during which we also optimized the hyper-parameters of the Dirichlet priors, before freezing the hyper-parameters and collecting 100 samples, each separated by a 10-sample lag, running for a total of 2,000 iterations.$^2$

### Topic-Level Analysis

In this section, we explore what the posterior variability of a topic’s word distribution can tell us about the topic. We hypothesize that topics whose distributions fluctuate between samples are more likely to contain ambiguous words and less likely to be topics that consistently represent the corpus. We will test this in two ways: comparing the variability to quality ratings provided by humans, and comparing the variability to how consistently topics appear in multiple models.

To measure posterior variability, we define a metric called topic stability which measures the degree to which a topic’s parameters change during sampling. Examples of topics with high and low stability are show in Table 2.

---

$^2$We also experimented with a total of 6,000 and 11,000 iterations, which made little difference in the results.

---

<table>
<thead>
<tr>
<th>Dataset</th>
<th># Docs</th>
<th># Vocab</th>
<th># Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>News</td>
<td>2,243</td>
<td>24,578</td>
<td>436,252</td>
</tr>
<tr>
<td>Wiki</td>
<td>10,000</td>
<td>80,011</td>
<td>7,383,116</td>
</tr>
</tbody>
</table>

Table 1: Statistics for the two document collections used in our experiments after pre-processing.
Table 2: Examples of different topic samples. The columns correspond to different Gibbs sampling iterations (from 1000 to 2000), with the 10 most probable words shown based on the estimate from that specific sample. We show two topics with relatively low stability and two with relatively high stability. The high stability topics do not vary as much across samples.

The parameters for a topic \( k \) are its word distribution vector, denoted \( \phi_k \). Each sample from the Gibbs sampler can be used to obtain an estimate of \( \phi_k \). Let \( \Phi_k \) denote the set of estimates of \( \phi_k \) from different samples, and \( \bar{\phi}_k \) is the mean of those samples. We define the stability of the set of sample estimates as:

\[
\text{stability}(\Phi_k) = \frac{1}{|\Phi_k|} \sum_{\phi_k \in \Phi_k} \text{sim}(\phi_k, \bar{\phi}_k)
\]  

for a vector similarity function \( \text{sim} \). We initially experimented with four similarity (or distance) metrics: cosine similarity, Euclidean distance, KL-divergence, and Jaccard similarity. For Jaccard similarity, we took the 10 most probable words in a sample \( \phi_k \) as the set for comparison, as this is the set of words shown to humans in our experiments. Table 3 shows the performance of these four metrics on our two tasks (described in the two subsections below). We found that cosine similarity worked substantially better than the others in all cases, so this is what we use as our stability in the rest of this section.

Baseline In the experiments in this section, we will compare our proposed stability metric to two commonly used measurements of topic quality: the coherence metric of (Mimno et al. 2011) and the related NPMI metric (Lau, Newman, and Baldwin. 2014), both defined in the “Evaluating Topic Models” section above.

Comparison with Quality Rated by Humans We explore whether our topic stability metric can indicate if a topic will be perceived as a high or low quality topic.
Comparison with Consistency across Models

The notion of “stability” in topic models was previously described by Chuang et al. (2015) in the context of comparing multiple models. Their study investigated variations of different trials of LDA, as different Gibbs sampling runs will result in different output each time due to randomness in the inference procedure. This behavior can be problematic for potential users of topic models, such as social scientists, who are not sure how to interpret a topic that only sometimes appears in topic model output.

Chuang et al. (2015) studied this by aligning topics from different modeling runs and quantifying how consistently a topic is discovered by LDA. Some topics, stable topics, are always inferred by LDA, while others may be one off topics that cannot be replicated. This work used an up-to-one alignment algorithm for topical alignment. Pairs of topics from differently trained topic models are merged together if they meet a similarity threshold. The number of topics that are cluster indicates how stable or consistent the topic is.

Compared to quality judgments, we consider this an orthogonal approach to understanding topic models, as low-quality topics may consistently appear across models, while high-quality topics may appear inconsistently. However, we hypothesize that the consistency of a topic across models may be related to the consistency of a topic across its posterior distribution, and so we separately experiment to see if automated metrics applied to one model, including our proposed topic stability metric, can predict consistency across multiple models.

We implemented this approach and applied it to our two datasets. We ran LDA four times on each corpus and then applied the up-to-one topical alignment process, using a cosine similarity threshold of 0.2.

After getting topic clusters, we calculate the average value of each metric (topic stability, coherence, NPMI) in each cluster. Figure 2 shows the distribution of average cluster values for different sizes of clusters, with correlations shown in Table 4. Stability is most correlated with cluster size on both News and Wiki datasets. NPMI, which has the best correlation with topic quality, has a poor correlation with cluster size, suggesting that topic consistency is not necessarily related with topic quality, although there appears to at least be a small relation.

Overall, topic stability measured through posterior variability appears to be a good indicator of the consistency of topics across multiple models, although this finding is stronger in the News corpus than the larger Wiki corpus.

Word-Level Analysis

When focusing on the words within an individual topic, we also investigate how the variability of the posterior of individual word probabilities can be informative. Anecdotally, we find that words with high posterior variance tend to be less strongly associated with the topic, often common words like “said” and “new” that might be considered stop words, but were not in our stop word list during preprocessing. See Figure 3 for an example of this phenomenon.

3Indeed, the two measurements do not appear strongly related. The alignment cluster size has a low rank correlation with human ratings: .129 (News) and .110 (Wiki).

<table>
<thead>
<tr>
<th>Metric</th>
<th>Quality</th>
<th>Consistency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cosine similarity</td>
<td>0.249</td>
<td>0.315</td>
</tr>
<tr>
<td>KL-divergence</td>
<td>0.016</td>
<td>0.254</td>
</tr>
<tr>
<td>Euclidean distance</td>
<td>0.013</td>
<td>0.152</td>
</tr>
<tr>
<td>Jaccard similarity</td>
<td>0.108</td>
<td>0.058</td>
</tr>
</tbody>
</table>

Table 3: The rank correlation between each potential metric of posterior variability compared to two topic-level metrics: mean human rating (quality) and model alignment cluster size (consistency).

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Quality</th>
<th>Consistency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stability</td>
<td>0.248</td>
<td>0.253</td>
</tr>
<tr>
<td>Coherence</td>
<td>0.198</td>
<td>0.040</td>
</tr>
<tr>
<td>NPMI</td>
<td>0.553</td>
<td>0.462</td>
</tr>
</tbody>
</table>

Table 4: The rank correlation between each potential topic-level metric and the quality ratings or consistency.

to humans. We collected quality judgments from humans by having people rate topics (the 10 most probable words in the topic) on a 4-point Likert scale, with a 4 meaning that all words in the topic are related to each other, and a 1 meaning that most of the words are unrelated.

We collected ratings through Amazon Mechanical Turk. We collected ratings from seven different workers for each topic, in order to construct more robust estimates given the variability in human judgments. We removed and re-collected ratings from workers who completed the ratings were outliers in time to complete the tasks or in similarity to the ratings from other workers. We took the average of the seven ratings to produce a final rating for each topic. The average score across all topics is 3.07.

Figure 1 shows the topic ratings along with each of the three metrics (topic stability, coherence, NPMI), with correlations (Spearman’s rank correlation ρ) in Table 4. On both datasets, topic stability is more correlated with quality than topic coherence, but NPMI has a higher correlation than either. Even though stability does not have the highest correlation, it is still noteworthy that it has a significant correlation with quality ratings. While coherence and NPMI both attempt to directly measure the relatedness of the words in the topic, the stability metric uses no information about the words that are in the topic, only the certainty of the parameters under the posterior. That topics with less certain parameters have a tendency to be judged as lower quality topics is an interesting finding that can be explored more in future work.
In this section, we propose two methods for taking the posterior variability into consideration when ranking the top words in a topic, based on the hypothesis that words whose topic probabilities fluctuate highly are less likely to be salient representations of the topic. We obtain human judgments to evaluate different methods, and find that our adjustments result in higher quality topic representations.

Adjusted Word Scores

The conditional probability of word \( v \) in topic \( k \) is denoted \( \phi_{kv} \). We propose two methods of adjusting the values of \( \phi_{kv} \) based on the set of posterior samples, \( \Phi_k \).

In the first method (referred to as Mean/SD), we divide the mean word probability \( \bar{\phi}_{kv} \) by the standard deviation of \( \phi_{kv} \) across the samples (the inverse coefficient of variation). This has the effect of downweighting the score of words whose posterior value of \( \phi_{kv} \) fluctuates more, i.e., the value is less certain.

In the second method (referred to as Min), we take the lowest value of \( \phi_{kv} \) that was observed in any of the samples. In other words, this is the empirical 0th percentile of the distribution of values.

In both methods, we obtain human judgments to evaluate different methods, and find that our adjustments result in higher quality topic representations.

### Experiments and Results

We conducted two experiments with human feedback collected through Amazon Mechanical Turk.

In the first experiment, we obtained quality ratings on the same 4-point Likert scale used in the previous section, but on topic representations where the top 10 words were ranked by our adjusted metrics. As before, we obtained ratings from seven different workers per topic.

The average topic ratings are shown in Table 5. Ratings under both adjusted scores are better than the baseline on both datasets. The difference in average ratings between the two adjusted methods (Mean/SD vs Min) is smaller, with inconsistent results between datasets.

### Baseline

Our baseline comparison is the sample mean of \( \phi_{kv} \), typically used as an estimate of \( E[\phi_{kv}] \).

### Table 5: The average topic ratings (standard error in parentheses) on each dataset when the top 10 words are ranked by the mean probability (the baseline method), as well as our two proposed adjustments.

<table>
<thead>
<tr>
<th>Version</th>
<th>News</th>
<th>Wiki</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>3.04 (.10)</td>
<td>3.09 (.06)</td>
</tr>
<tr>
<td>Mean/SD</td>
<td>3.12 (.10)</td>
<td>3.25 (.06)</td>
</tr>
<tr>
<td>Min</td>
<td>3.19 (.10)</td>
<td>3.18 (.06)</td>
</tr>
</tbody>
</table>
Figure 2: Distribution of topic stability, coherence and NPMI scores within different sized clusters on the topic alignment task for the News (top) and Wiki (bottom) datasets. The topics in larger cluster sizes indicate they are easier to be replicated across multiple runs of LDA. On both datasets, topic stability is aligned relatively well with the cluster size. The number of topics within each cluster size (1, 2, 3, and 4, respectively) are 16, 1, 11, and 22 on News and 26, 15, 20, and 39 on Wiki.

<table>
<thead>
<tr>
<th>Cluster Size</th>
<th>News</th>
<th>Wiki</th>
</tr>
</thead>
<tbody>
<tr>
<td>3/5</td>
<td>Mean</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>Mean/SD</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>Min</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>Mean/SD vs Min</td>
<td>26</td>
</tr>
<tr>
<td>4/5</td>
<td>10</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>Min</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Mean/SD vs Min</td>
<td>23</td>
</tr>
<tr>
<td>5/5</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Min</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Mean/SD vs Min</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 6: The number of times each word ranking method won a majority of votes in pairwise comparisons of different representations (Mean vs Mean/SD, Mean vs Min, Mean/SD vs Min), when the majority vote had at least 3, 4, or 5 votes.

Discussion and Future Directions

We have presented a new way of characterizing Bayesian topic models based on the variability of the posterior distr-
Table 7: Example of 10-word topic representations using three different methods, where Mean is the baseline method of using the average sample probability. Highlighted words indicate words that only appear in the set for that particular method.

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

Figure 3: Different percentiles of probabilities throughout sampling of five words within the same topic (top 10 words: drug, attorney, office, investigation, said, case, charges, department, fbi, documents). From the word list, we might infer this topic is about justice and law enforcement. The word “said,” which is not particularly related to this theme, fluctuates highly across different samples. The word “fbi,” which is strongly related, fluctuates the least.

We explored these ideas in two ways.

At the topic level, we introduced a metric called topic stability and showed that it is correlated with the consistency of topics across models and with the quality of topics rated by humans. Even though our proposed topic stability metric did not achieve state-of-the-art performance in the majority of cases, it always outperformed one of the two baseline methods. We argue that it is a strong and surprising result that topic stability is as highly correlated with other topic quality measures, given that it does not use any information about the words in the topic. This is a novel way of diagnosing and characterizing topics, which may end up being complementary to other types of metrics, and can likely be improved in future work.

At the word level, we found that words with high variability tend to be unrelated to the topic, and we proposed two ways of adjusting the weighting of words based on this information. In direct comparisons, people preferred our proposed modifications by a factors 1.6, 1.8, and 10.0 times compared to the standard way of ranking words in topics. These ideas can thus potentially provide a way to identify more salient representations of topics.

Because this work explored ideas that were quite different from prior approaches, there are many potential directions for future research. With respect to the tasks in this paper, future work could search for better definitions of topic stability and better methods for adjusting word scores. Additionally, this work only focused on the word distributions of topics, and not the topic distributions in documents. The latter is also important, and automatic evaluations have recently been proposed for these parameters as well (Bhatia, Lau, and Baldwin. 2017). It would likely be beneficial to explore the posterior variability at the document level. Beyond topic models, the idea of stability could be applied to other predictive models that would benefit from more interpretable parameters (Paul 2016). New metrics and methods using posterior variability could motivate performing Bayesian inference in models where this is not commonly done.

References


