Evaluation Methods for Topic Models

Hanna M. Wallach* **Iain Murray†** **Ruslan Salakhutdinov†** **David Mimno***

* Department of Computer Science
University of Massachusetts Amherst
Amherst, MA 01003
{wallach,mimno}@cs.umass.edu

† Department of Computer Science
University of Toronto
Toronto, Ontario. M5S 3G4
{murray,rsalakhu}@cs.toronto.edu

Statistical topic modeling has become a popular tool for analyzing large, unstructured text collections. There is a significant body of work developing sophisticated topic models and their applications. To date, however, the task of evaluating topic models has not been specifically addressed. Evaluation is an important issue: the unsupervised nature of topic models makes model selection difficult. For some applications there may be extrinsic tasks, such as information retrieval or document classification, for which performance can be evaluated. More universally, however, a topic model's ability to generalize can be measured by computing the probability of held-out documents under the model, which is independent of any specific application. Here, we consider the simplest topic model, latent Dirichlet allocation (LDA) [**?**], and compare a number of methods for estimating the probability of held-out documents given a trained model. Our empirical results on synthetic and real-world data sets show that the estimators currently used in the topic modeling literature are much less accurate and have higher variance than two proposed alternative methods. These proposed alternatives are also applicable to more complicated topic models.

Evaluating LDA: LDA generates a new document $\boldsymbol{w}$ by drawing

$$\begin{aligned}
\boldsymbol{\theta} &\sim \mathrm{Dir}\,(\boldsymbol{\theta}; \alpha\boldsymbol{m}) && \text{a document-specific distribution over topics,}\\
\boldsymbol{z} &\sim P(\boldsymbol{z} \,|\, \boldsymbol{\theta}) = \textstyle\prod_n \theta_{z_n} && \text{a latent topic assignment for each token,}\\
\boldsymbol{w} &\sim P(\boldsymbol{w} \,|\, \boldsymbol{z}, \Phi) = \textstyle\prod_n \phi_{w_n|z_n} && \text{and finally the observed tokens.}
\end{aligned}$$

The "topic" parameters, Φ, and hyperparameters, $\alpha\boldsymbol{m}$, are shared by all documents. Variational methods [**?**] and MCMC methods [**?**] are effective at marginalizing out the topic assignments and document-specific topic distributions associated with training data to infer Φ and $\alpha\boldsymbol{m}$. The latent variables associated with an individual held-out document are harder to identify. We focus on marginalizing out the latent variables for a held-out document $\boldsymbol{w}$ to evaluate its probability:

$$P(\boldsymbol{w} \,|\, \Phi, \alpha\boldsymbol{m}) = \sum_{\boldsymbol{z}} \int \mathrm{d}\boldsymbol{\theta}\; P(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\theta} \,|\, \Phi, \alpha\boldsymbol{m}). \tag{1}$$

Although it is possible to perform the integral over $\boldsymbol{\theta}$ analytically, the resulting sum over latent topic assignments must be approximated for documents longer than a few words.

Evaluation Methods: Document probabilities (1) are usually approximated using Monte Carlo methods. Previous work has favored simple methods based on existing Gibbs sampling code used for other inference tasks. Little attention, however, has been paid to the quality of these evaluations. We aim to identify simple methods that also work well and carry some guarantees.

The probability in (1) can be approximated by importance sampling. We compared two simple importance sampling methods: the "empirical likelihood" method provided by MALLET [**?**] and a method that uses a fully-factored approximation to the posterior over $\boldsymbol{z}$ as the sampling distribution.

Presenter: HMW **Category:** graphical models **Preference:** Oral

We also implemented two more advanced importance samplers that use Gibbs sampling code: annealed importance sampling (AIS) [**?**] and a recently-proposed sampler inspired by Chib's method [**?**]. Although some importance samplers have better properties than others, all of them yield results that can be used to obtain a probabilistic lower bound on the correct answer [**?**]. Worse samplers, although unbiased, provide underestimates with high-probability. Given several importance samplers, the one with the largest estimate is usually the most accurate.

The harmonic mean method [**?**] is an importance sampling estimator for $1 \,/\, P(\boldsymbol{w} \,|\, \Phi, \alpha\boldsymbol{m})$, combined with a Gibbs sampling approximation. This estimator has very poor properties, as discussed by Neal [**?**]. The underlying importance sampler will frequently underestimate $1 \,/\, P(\boldsymbol{w} \,|\, \Phi, \alpha\boldsymbol{m})$, leading to overestimates of document probabilities. Furthermore, the Gibbs sampling approximation makes it hard to construct any useful formal guarantees. Despite these problems, the harmonic mean method has been used in previous work to evaluate several topic models [**? ? ?**].

Another way of evaluating document probabilities is to predict each word in turn, "left-to-right" [**?**]: $P(\boldsymbol{w} \,|\, \Phi, \alpha\boldsymbol{m}) = \prod_n P(w_n \,|\, \boldsymbol{w}_{<n}, \Phi, \alpha\boldsymbol{m})$. Each conditional probability can then be estimated using Gibbs sampling. Although is hard to provide guarantees for this approximation, the conditionals define a generative procedure that achieves exactly the test performance reported.

All of the above methods are simple to implement and generalize readily to more complicated topic models, including non-parametric versions based on hierarchical Dirichlet processes [**?**].

Results: We compared all the methods described above using synthetic and real-world textual corpora. Our experiments with synthetic data suggested that AIS gives accurate answers if run for long enough. On real-world data, the harmonic mean method consistently gave much larger estimates than AIS. On synthetic data, we confirmed that the harmonic mean method always over-estimated the true test probabilities. As expected, all the importance sampling methods gave smaller estimates whenever they differed from a long AIS run. Of the importance sampling methods, only the Chib-style method gave good results. This method performed better than AIS when allowed the same amount of computer time. The left-to-right algorithm usually, but not always, out-performed the Chib-style method, given the same amount of computer time. Unless a stochastic bound on the model's true probability is required, the left-to-right algorithm should be used. Figure 1 shows example results on a single corpus with bootstrap-based error bars.

Discussion: Estimating the probability of held-out documents provides an interpretable metric for evaluating the performance of topic models. We found empirically, however, that the evaluation methods currently used in the topic modeling community, including the harmonic mean method and the "empirical likelihood" method provided by MALLET, are generally inaccurate. Even if these methods do result in a correct ranking of different models, the relative advantage of one model over another may be incorrectly represented. In contrast, the Chib-style method and left-to-right algorithm offer much better ways of accurately assessing and selecting topic models.

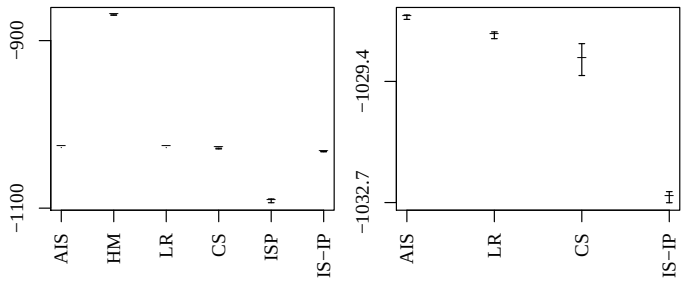

Figure 1: Average log probability per held-out document (20 Newsgroups corpus). The right-hand figure shows the best methods only. **AIS:** annealed importance sampling. **HM:** harmonic mean method. **LR:** "left-to-right" method. **CS:** Chib-style method. **ISP:** "empirical likelihood" method from MALLET. **IS-IP:** importance sampling using an approximate posterior over $\boldsymbol{z}$. LR, CS, ISP and IS-IP were all parameterized to process a 200 word document in at most 3.2 seconds; AIS and HM were given much more time.