Evaluating Topic Stability and Variability for Variational Autoencoder Topic Models

Kenny Chiu
Department of Statistics
University of British Columbia
Vancouver, BC, Canada, V6T 1Z4
kenny.chiu@stat.ubc.ca

Abstract

Many co-occurrence based metrics have been proposed for evaluating probabilistic topic models. Recently, metrics that instead use the variability of the model posterior distribution have been shown to be more correlated with human judgment than the co-occurrence based metrics. However, these metrics rely on the availability of Gibbs samples and hence do not directly apply to neural topic models. We propose several translations of these posterior based metrics to the neural setting and evaluate them against existing baseline metrics. The results of our experiment involving two variational autoencoder topic models and two datasets suggest that intermediate estimates obtained after training epochs do not substitute for Gibbs samples, but random samples drawn from the posterior distribution after the network has been trained may have potential.

1 Introduction

The outputs of probabilistic topic models such as latent Dirichlet allocation (LDA) (Blei et al., 2003) have found applications in further downstream natural language processing tasks. Thus, evaluating the quality of a topic model is necessary for determining whether the model produces reasonable or noisy results. As manual human evaluation of topics is inefficient and expensive, many numerical metrics have been proposed for evaluating the quality of topics with the aim of mimicking human intuition and judgment. Most of these metrics are *co-occurrence based* and use statistics computed from co-occurrences between pairs of words. Recently, *posterior based metrics* (Xing and Paul, 2018; Xing et al., 2019) have been proposed for LDA topic models and were found to be more correlated with human judgment than the co-occurrence based metrics. These posterior based metrics measure the variability of the posterior topic and word distributions, where a smaller variability indicates a better quality. Variability is estimated using posterior samples obtained from the Gibbs sampling procedure which is commonly used for inference in LDA topic models (Steyvers and Griffiths, 2007).

While posterior based metrics appear promising for topic model evaluation, not all probabilistic topic models use Gibbs sampling for inference. In particular, the rise of neural networks in recent years have led to new neural approaches to topic modeling. Variational autoencoder (VAE) topic models (Miao et al., 2015; Srivastava and Sutton, 2017) are one of these new approaches that have the advantage of being fast and reusable for new documents once trained. VAE topic models are typically trained through stochastic gradient optimization, and so the posterior based metrics computed from Gibbs samples do not directly apply to these models.

Our objective for this project is therefore to explore possible translations of the posterior based metrics to the neural setting. While the new metrics we consider in this project do not appear to outperform existing metrics in term of correlation with human judgment, they provide some insight that may guide future development of new metrics for evaluating the quality of VAE topic models.

1.1 Contribution

The main contributions of our work are the following:

1. Evaluations of several new metrics against existing metrics using two different VAE topic models trained on two different datasets.

2. Insights gained from the results of an experiment for guiding future work.

3. Publicly available Python 3 implementations for the NVLDA and ProdLDA topic models that are

adapted from the original Python 2 implementations. A link to the GitHub repository is provided in Appendix A.

2 VAE Topic Models

VAE topic models come in various flavours but ultimately have the same approach to topic modeling. VAEs (Kingma and Welling, 2013) consist of an *inference network* (i.e., encoder) that maps observed data to some latent space and a *generative network* (i.e., decoder) that reconstructs the observations from values of the latent space. In the context of topic modeling, the observations are the documents (typically in bag-of-words representation) and the latent space describes the topic and word distributions. The inference network is the main advantage that VAE topic models have over other topic models as once the network has trained, the network can be reused on new documents to infer their topic distributions without needing additional training. Some VAE topic models still operate in the LDA setting, where the network approximates a LDA topic model and the latent space is the space of Dirichlet document-topic and topic-word probability distributions (Srivastava and Sutton, 2017). Other VAE topic models forgo LDA all together and model the topics through other representations, such as through a Gaussian topic distribution as in the Neural Variational Document Model (Miao et al., 2015). In this project, we consider two closely related VAE-LDA topic models.

2.1 NVLDA

The Dirichlet document-topic and topic-word distributions are challenging to deal with in VAEs due to the distributions living on a simplex and requiring constrained optimization. The NVLDA topic model (Srivastava and Sutton, 2017) avoids this issue by using a Laplace approximation to the Dirichlet priors where distribution estimates are projected onto a simplex by doing a softmax operation. The posterior over the simplex basis is then approximated using a logistic normal distribution with mean and covariance parameters.

The inference network in NVLDA consists of two fully connected-softplus layers that feed into two separate layers that output the estimated mean and covariance parameters of the logistic normal. The generative network is a simple stochastic layer that reconstructs documents based on samples drawn from the approximated logistic normal distribution. The networks are trained using the variational evidence lower bound objective as is standard in VAEs.

2.2 ProdLDA

The ProdLDA topic model (Srivastava and Sutton, 2017) is identical to the NVLDA model in concept and in architecture except that the unconstrained estimates rather than the softmaxed estimates are used implicitly. This is described as LDA in an unconstrained space and allows new documents to be mapped to distributions outside the simplex spanned by the topic distributions of the posterior. Produced distribution estimates are still fed through a softmax operation when a distribution is needed for interpretation or for metric calculations.

3 Co-occurrence Based Metrics

We consider several existing co-occurrence based metrics to use as baselines in the evaluation of our proposed metrics.

3.1 Coherence and Generalized Coherence

Let $D(w)$ be the document frequency of word w (number of documents that w appears in) and $D(w, w')$ be the co-document frequency of words w and w' (number of documents that both w and w' appear in). Let $W^{(t)} = \left(w_1^{(t)}, ..., w_M^{(t)}\right)$ be the list of M most probable words in topic t (we use $M = 10$ for all the relevant metrics considered in this project). Then the *topic coherence* (Mimno et al., 2011) is defined as

$$C\left(t; W^{(t)}\right) = \sum_{i=2}^{M} \sum_{j=1}^{i-1} \log \frac{D\left(w_i^{(t)}, w_j^{(t)}\right) + 1}{D\left(w_j^{(t)}\right)}$$

We also consider the generalized topic coherence metric that was found to improve coherence scores in certain cases (Stevens et al., 2012). The metric (which we refer to as *topic coherence*$_\epsilon$) allows different smoothing factors and is defined as

$$C\left(t, \epsilon; W^{(t)}\right) = \sum_{i=2}^{M} \sum_{j=1}^{i-1} \log \frac{D\left(w_i^{(t)}, w_j^{(t)}\right) + \epsilon}{D\left(w_j^{(t)}\right)}$$

where we use $\epsilon = 10^{-12}$ following the original authors.

3.2 Pointwise Mutual Information (PMI)

Let $P(w)$ be the marginal probability of word w and $P(w, w')$ be the joint probability of words w and w'. Then *topic PMI* (Lau et al., 2014) is defined as

$$\text{PMI}\left(t, \epsilon; W^{(t)}\right) = \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \log \frac{P\left(w_i^{(t)}, w_j^{(t)}\right) + \epsilon}{P\left(w_i^{(t)}\right) P\left(w_j^{(t)}\right)}$$

where a smoothing factor ϵ is again added to avoid taking the log of 0 in the case that $w_i^{(t)}$ and $w_j^{(t)}$ do not appear in the same documents.

3.3 Normalized PMI (NPMI)

NPMI is closely related to PMI except that it is normalized to be on the interval $[-1, 1]$. In this project, we scale this interval to $[0, 1]$ for ease of interpretation. *Topic NPMI* (Lau et al., 2014) is then defined as

$$\text{NPMI}\left(t; W^{(t)}\right) = \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \frac{\log \frac{P\left(w_i^{(t)}, w_j^{(t)}\right)}{P\left(w_i^{(t)}\right) P\left(w_j^{(t)}\right)}}{-\log P\left(w_i^{(t)}, w_j^{(t)}\right)}$$

where NPMI is taken to be -1 (or 0 after scaling) if a pair of words do not co-occur. Because the metric is bounded, a smoothing factor is unneeded here.

4 Posterior Based Metrics

We introduce the two posterior based metrics that our proposed metrics for VAE topic models are based off of.

4.1 Stability

For topic t, let $\phi^{(t)}$ be its topic-word distribution. Let $\Phi^{(t)} = \left(\phi_1^{(t)}, ..., \phi_K^{(t)}\right)$ be the sample of K Gibbs estimates of the distribution. Denote the mean of the sample as mean $\left(\Phi^{(t)}\right)$. Then *topic stability* (Xing and Paul, 2018) is defined as

$$S\left(t; \Phi^{(t)}\right) = \frac{1}{K} \sum_{i=1}^{K} \text{sim}\left(\phi_i^{(t)}, \text{mean}\left(\Phi^{(t)}\right)\right)$$

where sim$(\cdot)$ is some vector (dis)similarity function.

4.2 Variability

For a document d, let $\theta^{(d)}$ be its document-topic distribution. Let $\Theta^{(d)} = \left(\theta_1^{(d)}, ..., \theta_K^{(d)}\right)$ be the sample of K Gibbs estimates of the distribution. For a topic t, denote the sample of K probability estimates as $\Theta_t^{(d)}$. Let std$(\cdot)$ be the standard deviation function. The *coefficient of variance* for topic t and document d is defined as

$$\text{cv}_d^{(t)} = \frac{\text{std}\left(\Theta_t^{(d)}\right)}{\text{mean}\left(\Theta_t^{(d)}\right)}$$

Let $\Sigma^{(t)} = \left(\text{cv}_1^{(t)}, ..., \text{cv}_D^{(t)}\right)$ be the set of coefficient of variance for topic t across D documents. Then *topic variability* (Xing et al., 2019) is defined as

$$V\left(t; \Sigma^{(t)}\right) = \text{std}\left(\Sigma^{(t)}\right)$$

5 Proposed Metrics for VAE Topic Models

We propose several new metrics inspired by the posterior based metrics described in Section 4. Note that in this project, we scale the metrics such that they take on values in $[0, 1]$ if the original interval is bounded.

5.1 VAE-Stability and VAE-Variability

The VAE translations of topic stability (*VAE-Stability*) and variability (*VAE-Variability*) are identical to the original formulations except that rather than $\Phi^{(t)}$ and $\Theta^{(d)}$ being the sample of Gibbs estimates, we use the intermediate estimates of the logistic normal mean obtained after each epoch during the training phase of the network. The initial idea comes from treating the epoch estimates as random samples drawn from a distribution induced by the stochastic gradient optimization. This is partially inspired by stochastic gradient Langevin dynamics (Welling and Teh, 2011) where by injecting random noise into the stochastic gradient iterations and applying a diminishing learning rate, the training procedure automatically transitions from performing optimization to drawing samples from the posterior. However, due to the scope of this project, we do not inject random noise during training and only use the epoch estimates as is.

For the vector (dis)similarity function sim$(\cdot)$ in VAE-Stability, we consider cosine similarity, symmetric KL divergence, and Euclidean distance. We also consider Jaccard similarity, which we define as

$$J(\phi, \phi') = \frac{1}{M^2} \sum_{i=1}^{M} \sum_{j=1}^{M} \frac{D(w_i, w_j')}{D(w_i) + D(w_j') - D(w_i, w_j')}$$

where w and w' correspond to the M most probable words in word distributions ϕ and ϕ' respectively. Note that our definition makes VAE-Stability with Jaccard similarity both a co-occurrence and posterior based metric.

5.2 VAE-Posterior Variability

Another VAE translation of topic variability we propose, *VAE-Posterior Variability*, considers the variability of the posterior itself. The variability is estimated using a random sample $\Theta^{(d)}$ of K document-topic estimates generated for each training document after the network has been trained. The random samples are obtained as a natural output of the VAE due to the *reparameterization trick* (Kingma and Welling, 2013) that outsources the randomness in the inputs to an implicit distribution within the network. The network generates random samples from this implicit distribution and transforms them into samples of the target latent distribution given a document. The same computation for variability is then applied over the samples $\Theta^{(d)}$ of each document.

Note that this metric is computationally time-intensive even with a GPU. Given the timeframe of the project, we only use a randomly sampled 10% of the training set rather than the entire training set to calculate the metric.

6 Evaluation

We evaluate our proposed metrics against the existing metrics in terms of correlation with human judgment.

	Pearson's r					Spearman's ρ				
	NVLDA		ProdLDA			NVLDA		ProdLDA		
Metric	20NG	NYT	20NG	NYT	**Mean**	20NG	NYT	20NG	NYT	**Mean**
Coherence	0.274	0.006	0.128	0.157	0.141	0.262	0.048	0.252	0.128	0.173
Coherence$_\epsilon$	0.151	-0.003	0.334	0.369	0.213	0.272	0.029	0.264	0.118	0.171
PMI	0.276	0.397	0.330	0.404	0.352	0.237	0.387	0.182	**0.376**	0.296
NPMI	0.260	**0.400**	**0.352**	**0.413**	**0.356**	0.249	**0.390**	0.184	0.374	**0.299**
Cosine Stability	0.358	-0.085	-0.062	-0.082	0.032	0.419	-0.070	-0.048	-0.037	0.066
KL Stability	**-0.380**	0.087	0.034	0.069	-0.048	-0.416	0.079	0.008	0.023	-0.077
Euclid. Stability	-0.350	0.080	0.092	0.079	-0.025	**-0.433**	0.063	0.061	0.045	-0.066
Jaccard Stability	0.065	-0.310	0.205	-0.116	-0.039	0.081	-0.211	**0.271**	-0.005	0.034
Variability	-0.093	-0.216	0.097	0.059	-0.038	-0.089	-0.228	0.070	0.066	-0.045
Post. Variability	-0.190	-0.085	-0.237	-0.237	-0.187	-0.196	-0.043	-0.195	-0.299	-0.183

Table 1: Measures of correlation between human judgment and the metrics.

6.1 NVLDA and ProdLDA Implementation

We use modified versions of the original TensorFlow implementations of NVLDA and ProdLDA in our experiment. The original code was written in Python 2 and older package versions, while our version works with Python 3 and newer package versions. Our version also adds code for computing the metrics discussed in this report. The GitHub repository containing our implementation is linked in Appendix A.

We keep the default settings of the networks. For NVLDA, the number of epochs is 300 and the learning rate is 0.005. For ProdLDA, the number of epochs is 200 and the learning rate is 0.002. For both NVLDA and ProdLDA, the number of topics is fixed to 50 and the hidden dimension of the fully connected layers are set to 100. For computing VAE-Posterior Variability, 200 samples are generated for each sampled training document.

6.2 Datasets

The two datasets that we use in our experiment are the *20 Newsgroups* (20NG) dataset and the *New York Times articles* (NYT) dataset.

The 20NG dataset was included with the original NVLDA and ProdLDA implementations and is already preprocessed. The preprocessing involved tokenization, removal of some non UTF-8 characters, and English stop word removal (Srivastava and Sutton, 2017). The training set includes 11,258 newsgroup posts from 20 different newsgroups and has a vocabulary size of 1,995.

The NYT dataset was obtained from Kaggle (link provided in Appendix B), and the training set includes 7,112 New York Times articles. We apply the same preprocessing that was done to the 20NG dataset and also remove words with a document frequency of less than 1% of the dataset. The vocabulary size is 5,212.

6.3 Experimental Design

Our experimental design is similar to the one by Xing et al. (2019). We train the two VAE topic models on the two datasets and calculate the discussed metrics for each of the returned topics. We also manually annotate each topic with a (human) score on a 4-point scale based on the 10 most probable of words of the topic where 1 indicates poor quality and 4 indicates good quality. Indicators of poor quality include the presence of words that do not appear related to the others and the presence of what appears to be multiple topics.

To measure the strength of correlation between the computed metrics and the human judgment, we use the Pearson correlation coefficient that measures linear association and the Spearman's rank correlation coefficient that measures monotonic association. Both correlation measures take on values between $[-1, 1]$ where 0 indicates no association and ± 1 indicates perfect association.

6.4 Results

The correlation results are shown in Table 1. NPMI and PMI appear to be metrics that are most correlated with human judgment out of all the considered metrics under both the Pearson correlation coefficient and Spearman's rank correlation coefficient. Our proposed metrics that use the epoch estimates are noisy, which suggests that the variability of the epoch estimates does not capture any signals that are meaningful indicators of topic quality. Closer inspection of the epoch metrics shows that the variability across estimates is very small. The mean scores for VAE-Stability with Euclidean distance shown in Table 2 suggest that the epoch estimates are practically identical up to a few significant digits. Thus, the variability originating from the stochastic optimization appears to be negligible. The learning rate parameter of the network may be a relevant factor here, but our experiment does not investigate the effect of this parameter.

Metric	NVLDA 20NG	 NYT	ProdLDA 20NG	 NYT
Human	2.22	2.56	2.72	2.44
Coherence	-78.63	-62.97	-73.95	-61.48
Coherence$_\epsilon$	-88.68	-65.44	-117.72	-84.83
PMI	34.28	20.95	15.70	45.64
NPMI	0.52	0.51	0.51	0.53
Cosine	0.97	0.99	1.00	1.00
KL	0.03	0.01	0.00	0.00
Euclidean	0.01	0.00	0.00	0.00
Jaccard	0.10	0.19	0.11	0.17
Variability	0.00	0.00	0.00	0.00
Post. Vari.	0.11	0.08	0.11	0.07

Table 2: Mean scores for human judgment and metrics.

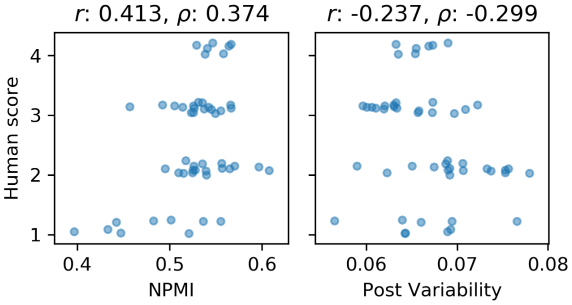

Figure 1: Human judgment against NPMI and VAE-Posterior Variability for ProdLDA trained on NYT. Pearson's r and Spearman's ρ are shown at the top. Small noise is added to the human scores for visual clarity.

In contrast, the results for VAE-Posterior Variability suggest that there is merit to using the variability of the posterior distribution. This is somewhat expected as VAE-Posterior Variability is conceptually the most similar to the original stability and variability based on Gibbs sampling. However, VAE-Posterior Variability is procedurally very different from the original metrics in that the samples are all obtained only after the model has been trained. Each topic model-dataset run in the experiment was able to train the network and compute the metrics in approximately one hour on a single GPU, but the calculations for VAE-Posterior Variability contributed to approximately half of that time even when using only 10% of the training documents to calculate the metric. Computation aside, we expect that posterior variability has potential to be an informative topic model metric. Possible improvements include collecting more estimates and calculating the metric differently. A possible solution for avoiding the post hoc calculations may be to use stochastic gradient Langevin dynamics (Welling and Teh, 2011) to collect samples directly during training.

We also note the unusually small correlation scores for all metrics in Table 1 where for example, NPMI has been shown to achieve a Pearson correlation coefficient of greater than 0.6 with human judgment (Xing et al., 2019). This is likely due to the quality of the human scores that were contributed by a single untrained annotator given the short timeframe of the project. A visual example of the correlation between human scores and the NPMI and VAE-Posterior Variability metrics is shown in Figure 1.

7 Project Summary

We reflect on the status of this project. The project was successful in that some results were produced with meaningful insights. While our proposed metrics did not perform well against the existing metrics, they provide valuable information as to what directions may be worth pursuing for developing VAE topic model metrics that are more correlated with human judgment.

Due to the timeframe of the project however, we were unable to assess whether our results generalize across different corpora and different topic models. The two corpora considered in this project were of similar sizes. It would be of interest to know whether our results still hold with larger corpora such as Wikipedia. We also only managed to get two working VAE topic model implementations (both originally by the same author). We were able to get a TensorFlow implementation for the briefly mentioned NVDM running, but the model trained much slower than NVLDA and ProdLDA and we also did not have time to learn the details of the code to properly implement our metric calculations.

Another factor that may affect the validity of our results are the quality of the human scores that we produced. While we tried to be objective by using fixed criteria when evaluating the topics, it was difficult to maintain consistency as ultimately the evaluation was subjective.

7.1 Learning and Development

The key concepts learned in this project include the different approaches that neural and non-neural topic models take, as well as the variety of existing metrics used to evaluate topic models. The main challenge encountered in this project was finding runnable implementations of the VAE topic models. Third party PyTorch implementations were first considered but later abandoned due to issues with downgrading and incompatible package versions. The TensorFlow implementations by the original authors are a few years old and are written in Python 2 and older package interfaces. Modifications to the original code eventually allowed the implementations to be run with Python 3 and newer package versions, but the process ultimately involved much trial and error due to the code not being heavily documented.

8 Conclusions and Future Work

In this project, we propose and compare several new metrics against human judgment for evaluating VAE topic models. Our results show that metrics based on intermediate epoch estimates collected during the training phase do not provide meaningful information, while the metric based on variability of the posterior may have potential. Possible improvements to the posterior variability metric may include using a bigger sample in the calculations, modifying how the metric is calculated, and applying stochastic gradient Langevin dynamic techniques to avoid post hoc sample collection. Other directions for future work include repeating our experiment with larger datasets as well as using other VAE topic models. Our experiment may also be further improved by crowdsourcing the human annotation of the topics to obtain higher quality scores contributed from more than one annotator.

Acknowledgements

We thank Professor Giuseppe Carenini and Linzi Xing from the University of British Columbia for their ideas and feedback on this project. We also thank Akash Srivastava and Charles Sutton from the University of Edinburgh for making their NVLDA and ProdLDA implementations publicly available.

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational Bayes. In *the 2nd International Conference on Learning Representations*.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539.

Yishu Miao, Lei Yu, and Phil Blunsom. 2015. Neural variational inference for text processing. In *Proceedings of the 33nd International Conference on Machine Learning*, pages 1727–1736.

David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 262–272.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *the 5th International Conference on Learning Representations*.

Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. 2012. Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 952–961.

Mark Steyvers and Tom Griffiths. 2007. Probabilistic topic models. In T. K. Landauer, D. S. McNamara, S. Dennis, and W. Kintsch (Eds.), *Handbook of latent semantic analysis*, pages 427–448. Lawrence Erlbaum Associates Publishers.

Max Welling and Yee Whye Teh. 2011. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 681–688.

Linzi Xing and Michael J. Paul. 2018. Diagnosing and improving topic models by analyzing posterior variability. In *AAAI Conference on Artificial Intelligence*, pages 6005–6012.

Linzi Xing, Michael J. Paul, and Giuseppe Carenini. 2019. Diagnosing and improving topic models by analyzing posterior variability. In *2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing*.

Appendix

A VAE Topic Model Implementations

Our Python 3 TensorFlow implementations of NVLDA and ProdLDA can be found on our GitHub.

B New York Times Dataset Source

The NYT dataset was downloaded from Kaggle.