**DS 6040 | Bayesian Machine Learning Project Report**

Students: Gabby Belmarez, Will Blickle, Patrick Corbett, Nick Keeley

1. **Problem Description**

Latent Dirichlet Allocation is a popular topic model used across many fields, from recommendation algorithms to evolutionary biology. In 2015, the New York Times published a blog post detailing the use of LDA on a new recommendation algorithm for news articles based primarily on topic modeling. One difficulty of implementing Latent Dirichlet Allocation is in integrating over the posterior probability of latent variables, which quickly grows intractable due to the combination of vast word and topic multinomial distributions.

Two approaches for addressing this issue include Markov Chain Monte-Carlo (MCMC) sampling, such as collapsed Gibb's sampling (used in Genesism), and Bayesian Variational Inference (used in SciKit-Learn), which approximates the target posterior distribution through optimization. MCMC approaches approximate the target posterior distribution more accurately but, because they use sampling, they don’t scale well to larger datasets or higher dimensional problems. While Variational Inference (VI) approaches are better-suited for scalability, they are subject to an additional weakness: because VI approaches are based on optimization algorithms, any changes to model assumptions must be reflected via changes in the optimization algorithm – a mathematically arduous process that limits the time and effort researchers can spend on result analysis.

In this report, we utilize Kaggle's News Category Dataset to explore an Autoencoding Variational Inference (AEVI) approach called “ProdLDA” that seeks to address the pitfalls of prevailing approaches to LDA. First, we outline the conceptual mechanics underlying implementation of MCMC, VI, and AEVI approaches. Second, we examine how well AEVI approaches model uncertainty relative to prevailing LDA approaches by comparing estimated by-topic word distributions and by-document topic distributions across approaches. Finally, we outline a path forward for future analysis, as well as providing recommendations for implementing ProdLDA and other AEVI approaches.

1. **Approach**

***Data***

We use Kaggle's News Category Dataset. This dataset consists of data from 200k headlines from the Huffington Post between 2012 and 2018. It consists of 41 different categories. Each element in the dataset has the following fields: category, headline, authors, link, short\_description, and date.

***MCMC Sampling Approach***

Initially, we settled on implementing an MCMC algorithm in pycm3. Unfortunately, we quickly determined that this approach would not be possible given our computational constraints. However, we found a more effective sampling approach to determining the target posterior distribution: a collapsed Gibb’s sampling technique proposed by Griffiths and Steyvers (2004). Their technique involves integrating out the variables ***ϕ***, ***T*** multinomial distributions over ***w*** words, and ***θ***, or ***D*** multinomial distributions over the ***T*** topics. Instead of estimating ***ϕ*** and ***θ***, this collapsed Gibb’s approach approximates the posterior distribution ***P(z|w)*** to make inferences about ***ϕ*** and ***θ*** (Griffiths and Steyvers 2004). It then uses the results of integrating out ***ϕ*** and ***θ***, also known as the full conditional distribution. This conditional distribution is a ratio represented by the probability of a given word in each topic multiplied by the probability of a given topic in a given document, meaning only non-zero counts will need to be calculated (improving performance). The ***P(z|w)*** is then approximated via sampling from the full conditional distribution and ***ϕ*** and ***θ*** are estimated from ***z***.

One of the advantages to this approach is the simplicity of the algorithm and its comparative improvement in performance from a traditional sampling approach. However, collapsed Gibb’s sampling can require a lot of memory and be time consuming.

***Variational Inference Approach to LDA***

We used the sci-kit learn library to compare an approach called online variational Bayes to the results of sampling and ProdLDA. Rather than doing “batch” variational Bayes approaches, which require recomputing parameters for each new observation, online variational Bayes reads each new observation almost as a stream (Hoffman et al. 2010). Similar to batch variational Bayes, the parameters ***γ*** and ***φ*** are optimized to maximize the Evidence Lower Bound (ELBO).

Ultimately, this algorithm has faster run times and lower amounts of memory required than traditional variational Bayes while maintaining high complexity. Online variational Bayes has become relatively ubiquitous for LDA as a result of its inclusion in Sci-Kit learn. It therefore was a good candidate for comparison to the ProdLDA approach.

***Autoencoder Variational Inference Approach to LDA: ProdLDA***

Autoencoder Variational Inference (AEVI) approaches – also referred to as “black box” inference approaches – are attractive because they address the universality problem inherent in standard variational approaches. They require far less algorithmic tweaking than standard VI approaches because AEVI utilizes a neural network to map documents to posterior distributions, coupling (or “encoding”) the variational parameters of *many* documents via the *same* network. By utilizing an unnormalized beta, this process produces a “weighted product of experts” posterior approximation that is smoother than its VI alternative’s “mixture of multinomials” result.

The authors of “Autoencoding Variational Inference for Topic Models” initially attempted to implement AEVI utilizing a common VI algorithm called “Automatic Differentiation Variational Inference” (ADVI) (Srivastava & Sutton, 2017). However, ADVI failed within a neural network framework for two reasons: first, the Dirichlet prior distribution was not a location-scale family, impeding successful reparameterization. Second, the algorithm often arrived at bad local optima, rather than the global optimum.

Here, we implement ProdLDA – Srivastava and Sutton’s (2017) adjustment to the aforementioned problems. ProdLDA utilizes two hidden, fully-connected layers to produce its estimates of latent variable distributions. It utilizes a soft-max/log-normal distribution as a substitute for the Dirichlet prior to successfully achieve reparameterization while avoiding component collapse by implementing an ADAM optimizer/higher learning rate during the optimization process.

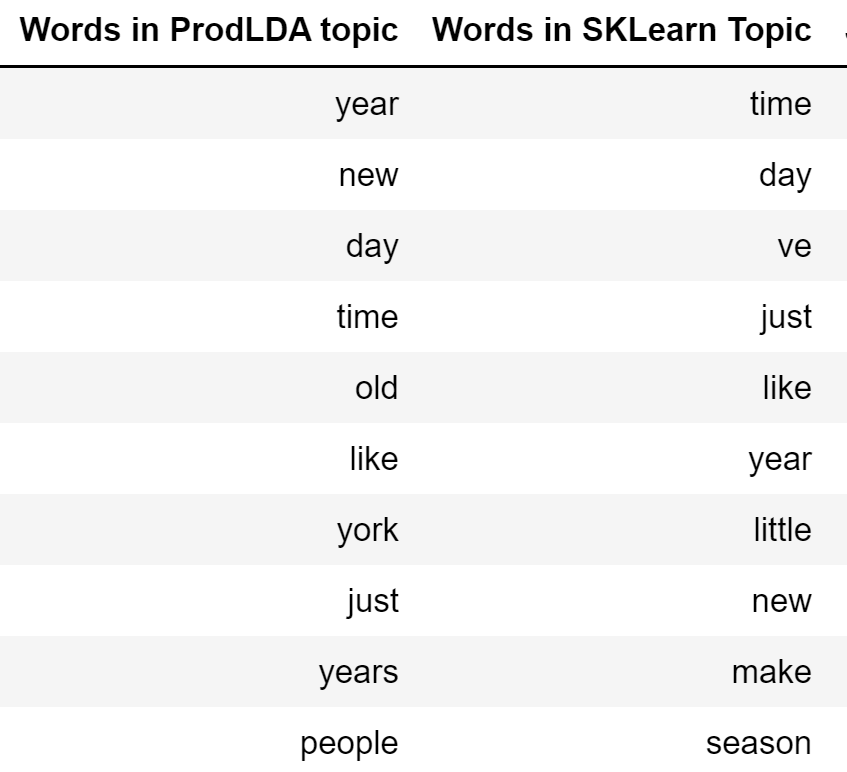
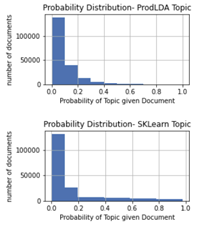
1. **Results**

Using the short\_description field from the Kaggle News dataset, we created a corpus comprising roughly 3500 words and fit topic models using each of these three approaches to produce 10 topics each. To compare the topics produced by these methods, we then calculated Jaccard scores between each set of topics. The Jaccard score is calculated as the number of shared words between 2 topics divided by the total number of distinct words in both topics, with *a higher value indicating more similarity between topics.* For each topic, we then calculated the max Jaccard score between that topic and the 10 topics in the other set, which measures the strength of the best similarity for that topic. We averaged this value across the 10 topics in each set to provide an overall measure of how similar the topics in one set are to topics in the other set (called Overall Similarity Score below):

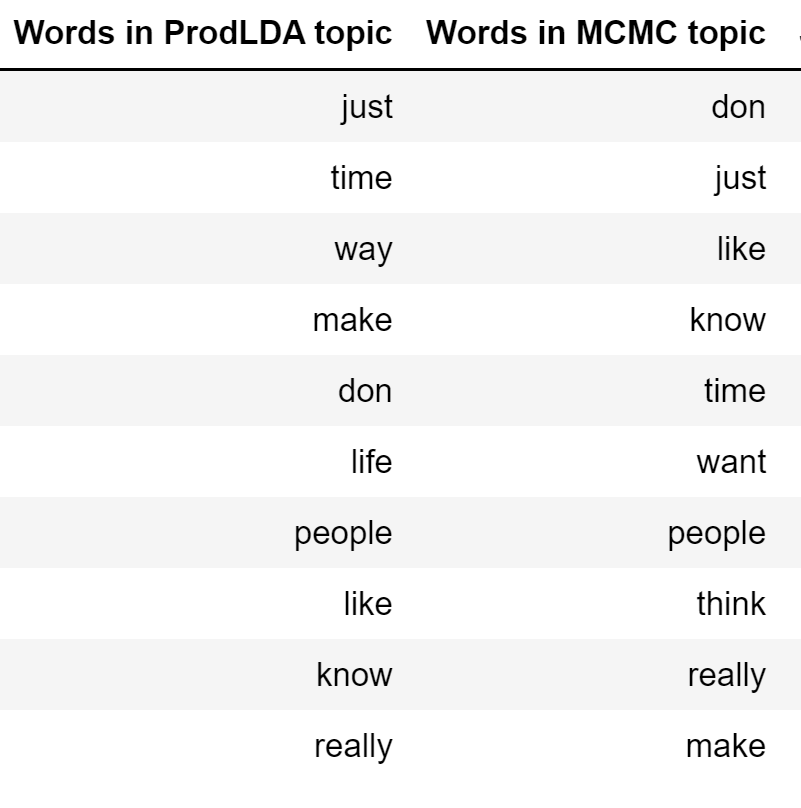
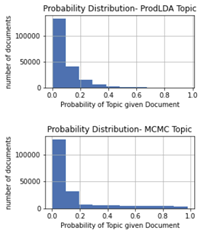
|  |  |  |  |
| --- | --- | --- | --- |
| Approaches Compared | ProdLDA vs. SKLearn | ProdLDA vs MCMC | SKLearn vs MCMC |
| Overall Similarity Score | 0.26 | 0.35 | 0.39 |

These scores indicate that the most similar topics overall are produced by SKLearn and MCMC, while the ProdLDA topics are not quite as close in similarity.

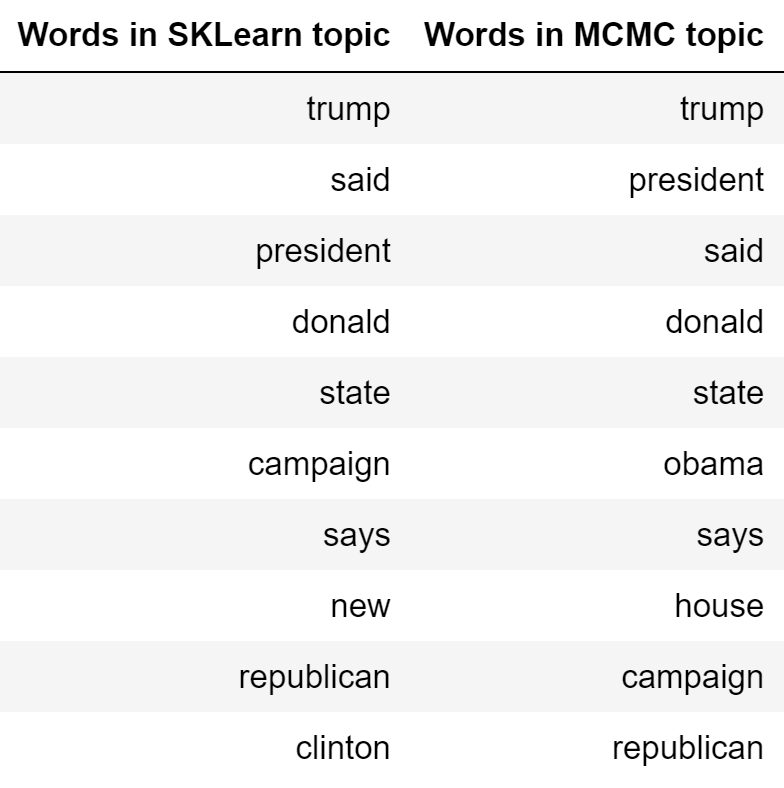
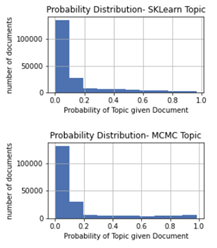
For each pair of methods, we then compared the posterior distribution of the probability of the most similar topics given the documents in our dataset. The results of this comparison for each pair of methods are shown below:

**Jaccard Score: 0.43**

**Jaccard Score: 0.66**

**Jaccard Score: 0.66**

Overall, we observed that compared to either MCMC or VI using SKLearn, the ProdLDA approach produces posterior distributions with lower variance for the most similar topics produced, while the MCMC and SKLearn approaches produce similar variances to each other. We also see that the various approaches do produce some topics with semantic similarity.

1. **Conclusion and Recommendations**

Our work demonstrates the potential utility of using the AEVI approach, as implemented in prodLDA, over other LDA approaches such as sampling and traditional variational inference. Specifically, AEVI allows for simpler mathematical setup than variational inference, while also avoiding the computational expense inherent in most sampling methods.

AEVI accomplishes these advantages by implementing an autoencoding feature and adjusting some distributional assumptions within the traditional LDA approach. The AEVI approach yields topics that are as similar to MCMC-produced topics as those topics produced by variational inference. However, the similarity between AEVI and variational inference topics appears to be slightly lower, which may cause concern when substituting AEVI for variational inference in current topic identification applications.

Future research could explore the relative accuracy of each method (given the constraints inherent in determine ‘accurate’ topics), as well as techniques for generating more similar topics between the different approaches. Both avenues may illuminate how and when AEVI could be substituted for sampling or traditional variational inference, leading to improvements in topic identification speed and accuracy.

1. **References**

Blei, David & Ng, Andrew & Jordan, Michael. (2001). Latent Dirichlet Allocation. The Journal of Machine Learning Research. 3. 601-608. Retrieved 9 December 2021 from <<https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf>>

Brownlee, J. (2019). A Gentle Introduction to Batch Normalization for Deep Neural Networks. Machine Learning Mastery. Retrieved 9 December 2021 from <<https://machinelearningmastery.com/batch-normalization-for-training-of-deep-neural-networks/>>

Griffiths, T.L. and Steyvers, M. (2004) Finding Scientific Topics. Proceedings of the National Academy of Sciences of the United States of America, 101, 5228-5235. Retrieved 12 December 2021 from <<https://doi.org/10.1073/pnas.0307752101>>

Hoffman, M. D., Blei, D. M., & Bach, F. (2010). Online Learning for Latent Dirichlet Allocation. Retrieved 12 December 2021 from <<https://papers.nips.cc/paper/2010/file/71f6278d140af599e06ad9bf1ba03cb0-Paper.pdf>>

Hu, J. (2018). A Simple Starter Guide to Build a Neural Network. Medium. Retrieved 9 December 2021 from <<https://towardsdatascience.com/a-simple-starter-guide-to-build-a-neural-network-3c2cf07b8d7c>>

McGonagle, John & Garcia, Jose Alonso & Mollick, Saruque, et. al. Feedforward Neural Networks. Brilliant. Retrieved 9 December 2021 from <<https://brilliant.org/wiki/feedforward-neural-networks/>>

Misra, R. (2018). News Category Dataset [Data set]. Kaggle. doi:10.13140/RG.2.2.20331.18729. Retrieved 9 December 2021 from <<https://www.kaggle.com/rmisra/news-category-dataset>>.

Naushan, H. (2021). Topic Modeling with Latent Dirichlet Allocation. Medium. Retrieved 8 November 2021, from <<https://towardsdatascience.com/topic-modeling-with-latent-dirichlet-allocation-e7ff75290f8>>

Pyro Contributors. (2020). *Probabilistic Topic Modeling*. Getting Started With Pyro: Tutorials, How-to Guides and Examples. Retrieved December 12, 2021, from <<https://pyro.ai/examples/prodlda.html>>

R, Mageshwaran. (2021). Topic Modeling: Art of Storytelling in NLP. Medium. Retrieved 8 November 2021, from <<https://medium.com/technovators/topic-modeling-art-of-storytelling-in-nlp-4dc83e96a987>>

Spangher, A. (2021). Building the Next New York Times Recommendation Engine. Open Blog. Retrieved 8 November 2021, from <<https://open.blogs.nytimes.com/2015/08/11/building-the-next-new-york-times-recommendation-engine/>>

Srivastava and Sutton, “Autoencoding variational inference for topic models,” conference paper at ICLR 2017, accessed 9 December 2021, <<https://arxiv.org/pdf/1703.01488.pdf>>