Using Simulations to Study LDA, a Dissertation Proposal

Tommy Jones
Computational and Data Sciences
George Mason University
Fairfax, VA
jones.thos.w@gmail.com

Abstract—This is an abstract

Introduction

Probabilistic topic models are widely used latent variable models of language. Popularized in 2002 by latent Dirichlet allocation (LDA) (Blei, Ng, and Jordan 2003) many related models have been developed, for example (Blei and Lafferty 2007), (Roberts et al. 2013), (Nguyen et al. 2015), and more. These models share common characteristics and estimate the probability of topics within contexts and tokens within topics.[1] Even today LDA remains one of the most popular topic models, and one of the simplest.

Probabilistic topic models have been applied to a variety of tasks. These tasks include information retreival [cite 3], analysis of historical texts [cite 2], machine translation and related tasks [cite 3], and more. In recent years, the machine learning community has focused more on deep architectures typified by text embeddings [cite 2] and pretrain then fine tune transformers [cite 3]. Yet probabilistic topic models have remained popular analytical methods in fields such as computational social science [cite 3] and the digital humanities [cite 3].[2]

In spite of their sustained popularity, probabilistic topic models remain challenging to use. Some of these challenges are conceptual. Probabilistic topic models have user-set tuning parameters, called "hyperparameters" in the machine learning literature, whose optimal settings are not obvious. Moreover, because probabilistic topic models estimate parameters for a process that is not how people write. Because of this, there is no ground truth against which to compare models for a sense of "correctness" that researchers can use to develop modeling strategies and metrics to guard against pathological misspecification.

In some cases, challenges are more practical. Software implementing probabilistic topic models can be challenging to use and offer limited functionality. In particular, those that employ probabilistic topic models in industry often have a need to update models based on new or updated data. To date, there has been little research on transfer learning for probabilistic topic models. No off-the-shelf software implements such a paradigm. The result is that applied practicioners face an unpleasant tradeoff. Either models go stale or they must be re-trained from scratch. In the former case, innacuracies creep in over time. In the latter case, topics are re-initialized at random, breaking continuity with the old model.

What is more, transfer learning is becoming paramount to modern machine learning for natural language. The last few years have seen an explosion of "transformer" models which rely on a paradigm of training a "base" model on an unsupervised or semi-supervised task. These base models tend to use as much language data as possible. Then the base model is transfered to a smaller dataset on a narrow supervised task. The result has been an impressive increase in performance on many standard NLP benchmarks. No such paradigm exists for probabilistic topic models.

In an attempt to address these shortcomings, I propose three research studies, each building on the last. In each, I will focus on Latent Dirichlet Allocation (LDA) for its simplicity and popularity. LDA is closely related to other probabilistic topic models. This enables a natural extension of this research to other probabilistic topic models.

The first study relates some empirical laws of language to LDA as a generative process. This enables a principled method for conducting simulation studies of LDA. Simulation studies are a natural means for imposing a sense of "correctness" in studying statistical models.[cite simulation study tutorial] Afer linking LDA to empirical laws of language, this first study will use a combination of simulations and analytical derivations to address hyperparameter settings for LDA. The objective is not to develop methods for finding the "correct" model on real data, as no such model exists. Rather it is to set up guardrails to avoid models that are pathologically misspecified where an obviously better model does exist.

The second study develops methods for transfer learning in LDA. This enables the applied practicioner to

[1] Technically, probabilistic topic models estimate the probability that any token was sampled from a topic given the context and the probability of sampling each specific token given the topic, respectively.

[2] As we will see in Section 3, probabilistic topic models can share similar conceptual frameworks with newer methods.

update models with new data, preservig continuity with previously-trained models. It also takes a first step extending LDA towards the state of the art "pre-train then fine tune" paradigm currently popular in natural language processing.

The final study introduces tidylda, a software package for the R programming language. tidylda integrates into a wider programming paradigm in the R language known as "tidy" programming. It also implements several novel methods for and related to LDA, including transfer learning.

The remainder of this document is organized as follows:

- Section 2 gives a brief history of embedding models for text, a broader class of models encompassing probabilistic topic models.
- Section 3 re-states the formulation for LDA, compares it to related models, and discusses training algorithms for LDA.
- Section 4 explores current approaches for evaluating and studying probabilistic topic models, with a focus on LDA.
- Section 5 gives an overview of simulation studies in statistics broadly and how they have been applied to probabilistic topic models.
- Section 6 reviews some empirical laws of language that a synthetic data set of language must honor to be considered a valid simulation of natural language.
- Section 7 outlines the proposed study for developing a principled means for simulation studies of LDA.
- Section 8 outlines the proposed study of tranfer learning for LDA.
- Section 9 outlines the proposed study instroducing the tidylda package for the R language.

Paragraph describing the background of the problem: In spite of probabilistic topic models having been around for 20 years, stubborn problems persist that limit their utility.

Paragraph summarizing the (3) studies I plan to do: Demonstrate that

Paragraph describing to whom this dissertation will be of value

Conceptual Framework

A Brief History of Language Embedding Models

LSA and Matrix Factorization Models: Start with LSA (What about term co-occurrence analyses? You have something on that in textmineR)

Probabilistic Topic Models:

- pLSA
- CTM
- STM
- Supervised LDA?
- Hierarchical LDA?
- Dynamic topic models?

Word Embedding Models: word2vec (and skipgram and negative sampling), doc2vec, GloVe

Describe how traditional topic models fit within this framework document = "context"

Word embedding models brought novelty in several ways. First, they introduced the concept of a word being represented as a distribution, rather than a binary presence or absence, or a count of occurence in a downstream model. [cite something] Second, they introduced the concept of "word algebra" where mathematical operations on word vectors in the embedding space seem to have semantic interpretations. [cite] There is some debate about whether this latter phenomenon is real or constructed [cite] but the former has become standard practice for a wide range of language modeling tasks. Finally, and perhaps most interestingly, researchers have explored methods for mapping embeddings in different languages on top of each other allowing two languages to "share" the same semantic space. [cite cross lingual word embeddings]

Transformers and the Pre-Train then Fine-Tune Paradigm: Main point: transformer architecture aside: pre-train then fine-tune is incredibly valuable and maps to a set of problems with traditional topic models: (a) How to update an existing model with new data without having to completely re-train from scratch? (b) How to deliver "big corpus" linguistic structure to niche problems?

Probabilistic Topic Models Today

In spite of the machine learning community shifting its focus neural networks, probabilistic topic models are still in widespread use. Frequent users of probabilistic topic models in recent years have come from political science [cite, cite, cite] and humanities [cite, cite, cite]. These new users, in contrast to the machine learning community, do not necessarily need more complex models pushing the state of the art. Instead they need tools that make these models more accessible and more reliable.

Paragraph emphasizing LDA as an embedding model document = "context" embedding to a probability space

Unresolved Issues with Probabilistic Topic Models: In [cite Boyd-Graber + Mimno] Boyd-Graber et al. cite four areas that need to be addressed for increased accessibility of probabilistic topic models. Paraphrasing, these four areas are:

1. The effects of different preprocessing and vocabulary curation steps on a resulting model,
2. How to think about the different model specification choices a researcher must make and the effects of these choices on the resulting model,
3. Interpreting the results of a probabilistic topic model in a way that is meaningful to humans in the context to which it is applied, and
4. A systematic investigation of the ways in which topic models can assist users' workflows in information oranization and retreival

The research I propose in this document focuses on the second area. In fact, I argue that the third and fourth follow from the first two. If a model is pathalogically misspecified, either from missteps in data curation or in explicit modeling choices, then any interpretation or application of that model is susptect.

Concretely: [Describe 1 and 2 above]

Because probabilistic topic models are latent variable models, and they model a process inconsistent with how humans actually write, there is no "right" model for any set of observed data. [However, we can describe the effects of modeling choices in a more rigorous way. And while there may be no "right" model, there are certainly many "wrong" ones. A better understanding of these issues can guard against pathalogical misspecification.]

Evaluating Probabilistic Topic Models Today:

Studying Complex Models with Stochastic Simulations

Simulation studies involve gnerating pseudo random data using a known stochastic process. Simulation studies have a long history in the field of statistics. [cite Ripley 1987, Hoaglin and Andrews 1975, Feiveson 2002] The purpose is to study data where the data generating mechanism is known. "A key strength of simulation studies is the ability to understand the behavior of statistical methods because some 'truth' [...] is known from teh process of generating the data." [cite morris et al. 2019]

[Establish vocabulary and approach to simulation studies here: estimator vs estimand, properties, etc.]

Probailistic topic models are excellent candidates for study through simulation. They model an explicit data generating process through latent variables. The resulting data are observed, but the generating variables of interest are not. By simulating data through the process modeled by a topic model, one can obtain a "ground truth". One can then interrogate a topic model as an estimator in terms of desireable statistical properties and measure its sensitivity to properties of the input data.

Simulation studies are indeed used to study probabilistic topic models. These studies usually use synthetic data sets to augment study of real data sets, rather than studying the synthetic data primarily. They tend to take one of two forms: either simulated data are drawn from models specified with priors commonly adopted for model fitting [Wallach dissertation] or drawn from the posterior of a model fit on real data [JBG + Mimno].

Yet to be a valid proxy of human language, such simulated data should have the same gross statistical properties of human language. To this end, Zipf's law is paramount.

Simulated Corpora Must Have the Statistical Properties of Human Language:

Zipf's Law: Zipf's law states that the term frequency distribution of any corpus of language in any language or context, follows a power law distribution. [cite Zipf 1949] The conventional wisdom is that this property holds only for documents of sufficient length and that the lowest-frequency words in the corpus do not follow the power law. Yet the inclusion of compound words holds Zipf's law through the lowest frequencies in the corpus.[cite Le Quan Ha et al.] In 2011, Goldwater et al. [cite] point out that "it is important for models used in unsupervised [machine] learning to be able to describe the gross statistical properties of the data that they are intended to learn from. Otherwise, these properties may distort inferences about the parameters of the model." Unfortunately they also note that "[Zipf's law] has been largely ignored in modern machine learning and computational linguistics."

Estimating power laws: Summarize some of the literature here. Lean heavily on the poweRlaw package vignette for source material. (Also cite poweRlaw explicitly)

Zipf's Law in the Context of LDA: Stochastic simulations for LDA have typically used symmetric priors. [cite cite cite] Yet such priors cannot produce corpora that adhere to Zipf's law.

Goldwater et al.

Cite my proof as appendix

Estimation of Zipf's law can be done emprically from data

More than Zipf's law - Sparsity - Distribution of individual words across documents should be similar - Heap's law within documents should be similar(?)

Focus on Latent Dirichlet Allocation: 2nd simplest and most popoular Probabilistic Topic model.

Simplicity makes it a good candidate for study

Simplicity means simulations have known limitations - e.g. documents are independent draws ==> no correlation in topics across documents - topics are independent draws ==> no correlation in tokens across topics

We know and expect that in the real world, certain topics should co-occur across documents and certain tokens should co-occur across topics.

Nevertheless, I hypothesize that studying LDA with simulations in its "gold standard" state may yeild useable insights to guide model specification on real world data.

Methodology

Intro/overall approach

Basic idea: use stochastic simulation to study the idealized properties of LDA for open issues.

First, need to demonstrate that stochastic simulations can produce the gross statistical properties of human language and, thus, be valid.

Study A: Developing a Principled Approach to Performing Simulation Studies of LDA

Hypothesis: Under certain constraints stochastic simulations using the functional form of LDA can produce synthetic corpora with properties of real-world corpora.

Need to identify a "survey" of languages and tasks to encompass "language" broadly.

Goals: 1. Quantify degree and ways that the simulations succeed and fail 2. Look at traditional (and new?) evaluation metrics and their ability to identify the "right" models that generated the data. 3. Intentionally misspecify models. What damage is done? Can we detect it? - Related: Does over-specifying the number of topics and then throwing away "bad" topics lead to the correct number vs. doing grid, random, or optimized search? 4. Post-hoc identification of stop words? 5. Effects of document length, number of documents, sampling tokens, cutting off stop words/infrequent words in model fit

Evaluation metrics to establish simulations: - Correlation and magnitude of zipf curves - Sparsity - Fraction of "failed" KS tests across words ranked by frequency - Note need to adjust for multiple tests/binomial proportion gives single thumbs up/down - i.e. Did we reject more or less based on expected random chance? - Something about Heap's law? Heap distributions across documents?

Evaluation metrics for model specification: 1. Usual suspects (and new?) metrics on both in-sample and held out (i.e. newly generated) data 2. Correlation of learned topics to "true" topics

Approach: 1. Simulations

Approach: 1. Estimate Zipf from MLE and use for shape of eta 2. Use Bayesian HP optimization in SigOpt to identify sets of parameter choices on pareto frontier for each selected real corpus - alpha: shape + magnitude, eta: magnitude, k - doc lengths are empirical 3. Multiple simulations using "optimal" settings and compare how well simulations do to real data 4. Next step: look at simulations and see if there are any patterns to help inform modeling 5. If possible, derive rules to get "best fit" models analytically (i.e. with algebra) 6. Apply rules to real world data - does it look good? What is the counterfactual?

Something else...

Study B: Transfer Learning for LDA - Towards a Pre-train then Fine Tune Paradigm

Hypothesis: None.

Practical: people need to update models in ways that aren't upsetting to users New hotness: Pre-train then fine-tune for LDA

Study C: tidylda - An R Package for LDA Topic Modeling Compatible with Tidy Data Principles

References

Blei, David M., and John D. Lafferty. 2007. "A correlated topic model of Science." The Annals of Applied Statistics 1 (1): 17–35. https://doi.org/10.1214/07-aoas114.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." Journal of Machine Learning Research 3.

Nguyen, Thang, Jordan Boyd-Graber, Jeffrey Lund, Kevin Seppi, and Eric Ringger. 2015. "Is Your Anchor Going Up or Down? Fast and Accurate Supervised Topic Models." In Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the Acl, 746–55. https://doi.org/10.3115/v1/n15-1076.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, and Edoardo M. Airoldi. 2013. "The Structural Topic Model and Applied Social Science." In.