Math 640 Final Paper

Max Kearns and Tommy Jones

May 10, 2018

Contents

1 Introduction **2**

2 Methods **2**
2.1 Models 2
2.2 Sampling 4
2.3 Comparison 4

3 Results **4**

4 Conclusion **4**

5 References **4**
5.1 make sure to cite the data source 4
5.2 Cite info on Zipf's law 4

6 Appendix A **4**

7 Appendix B **4**

1 Introduction

The analysis of text data is an area of vital research in both frequentist and Bayesian statistics. Text can, and indeed does, store a vast amount of information that is not easily evaluated with well-understood statistical methods. While text analysis is used throughout our economy, it does not have nearly as much research and knowledge behind it as does numerical data. This paper attempts to slightly further the bank of techniques for text analysis, in the hopes that text data will someday be as understood as numerical data is today.

One method that researchers currently use to model text frequencies is called Latent Dirichlet Allocation. This is an example of a topic model that allows researchers to find word frequencies in documents across various topics. Typically, this model makes use of a multinomial likelihood, with a Dirichlet prior on the probability of the inclusion of each word (θ). The usual model then fixes the Dirichlet parameter (α). In a Bayesian setting, however, this approach seems overly simplistic, and MCMC methods provide a simple solution to sample from a more complex distribution. This research intends to start to answer the question as to whether more uncertainty on α would improve the model. Zipf's law provides a basis for how to vary α in a way that is consistent with knowledge about human language.

Zipf's law is an empirical property of natural language. It states that the word frequencies of any corpus of text follows a power law distribution, regardless of context or language. This means that the most common word will be twice as frequent as the second most common word, and n times as frequent as the nth most common word. (citations needed) Based on what Zipf's law dictates, this research tests the viability of placing a Pareto(1, β) prior on α. This Pareto distribution is a power-law that should provide some uncertainty on α that mirrors the inherent property of language.

In order to determine whether this prior merits further research on more complex models, we will begin by modeling a small data set using a simple model that features a multinomial likelihood, a Dirichlet prior, and a pareto hyper-prior. We will compare this model to a control that does not allow for uncertainty on α The data set is 100 randomly sampled NIH grant proposals from 2014.

2 Methods

2.1 Models

For both the control and the new model, we assume that the word count y is iid multinomial, so has the following likelihood. (Is there a reason we aren't using vector notation here)

$$y \sim multinom(n, \theta) \implies \mathcal{L}(y|\theta, \alpha, \beta) \propto \prod_k \theta_k^{y_k} \tag{1}$$

On θ, the vector of word probabilities, we place a Dirichlet prior.

$$\theta \sim Dir(\vec{\alpha}) \implies \pi(\theta) = \mathcal{B}(\vec{\alpha}) \prod_k \theta_k^{\alpha_k - 1} \tag{2}$$

The control model will stop here, and assume (insert discussion of chosen alpha here). The control model then has the simple form $\theta|y \sim Dir(\vec{y} + \vec{\alpha})$ The candidate model, however, will assume a $Pareto(1, \beta)$ prior on α, with a constant prior on β.

$$\alpha_k \sim Pareto(1, \beta) \implies \pi(\vec{\alpha}) = \prod_k \beta \alpha_k^{-(\beta+1)} \tag{3}$$

$$\pi(\beta) \propto 1 \tag{4}$$

This results in the unknown posterior below, and a full derivation of the model can be found in Appendix B.

$$P(\theta, \alpha, \beta|y) \propto \beta^K \mathcal{B}(\vec{\alpha}) \prod_k \theta_k^{y_k+\alpha_k-1} \alpha_k^{-(\beta+1)} \tag{5}$$

This prior is unrecognizable, so we will proceed be making a Gibbs sampler of the full posteriors, which can be found below.

$$P(\theta|\alpha, \beta, y) \propto \prod_k \theta_k^{y_k+\alpha_k-1} \tag{6}$$

$$P(\alpha|\theta, \beta, y) \propto \mathcal{B}(\vec{\alpha}) \prod_k \theta_k^{y_k+\alpha_k-1} \alpha_k^{-(\beta+1)} \tag{7}$$

$$P(\beta|\theta, \alpha, y) \propto \beta^K (\prod_k \alpha_k)^{-\beta} \tag{8}$$

Of these conditional posteriors, only $\theta|\alpha, \beta, y$ is a known distribution; $Dir(\vec{y} + \vec{\alpha})$.

The other two conditionals will be sampled from using a Metropolis-Hastings algorithm with (insert proposals).

2.2 Sampling

2.3 Comparison

3 Results

4 Conclusion

5 References

5.1 make sure to cite the data source

5.2 Cite info on Zipf's law

6 Appendix A

7 Appendix B

$$P(\theta, \alpha, \beta|y) \propto \left[\prod_k \theta_k^{y_k}\right]\left[\mathcal{B}(\vec{\alpha})\prod_k \theta_k^{\alpha_k - 1}\right]\left[\prod_k \beta\alpha_k^{-(\beta+1)}\right] \tag{9}$$

$$= \beta^K \mathcal{B}(\vec{\alpha})\prod_k \theta_k^{y_k+\alpha_k-1}\alpha_k^{-(\beta+1)} \tag{10}$$

$$P(\theta|\alpha, \beta, y) \propto \prod_k \theta_k^{y_k+\alpha_k-1} \tag{11}$$

$$\implies \theta|\alpha, \beta, y \sim Dir(\vec{y} + \vec{\alpha}) \tag{12}$$

$$P(\alpha|\theta, \beta, y) \propto \mathcal{B}(\vec{\alpha})\prod_k \theta_k^{y_k+\alpha_k-1}\alpha_k^{-(\beta+1)} \tag{13}$$

$$\implies \text{unknown distribution} \tag{14}$$

$$P(\beta|\theta, \alpha, y) \propto \beta^K(\prod_k \alpha_k)^{-(\beta+1)} \tag{15}$$

$$\propto \beta^K(\prod_k \alpha_k)^{-\beta} \tag{16}$$

$$\implies \text{unknown distribution} \tag{17}$$