Math 640 Final Presentation

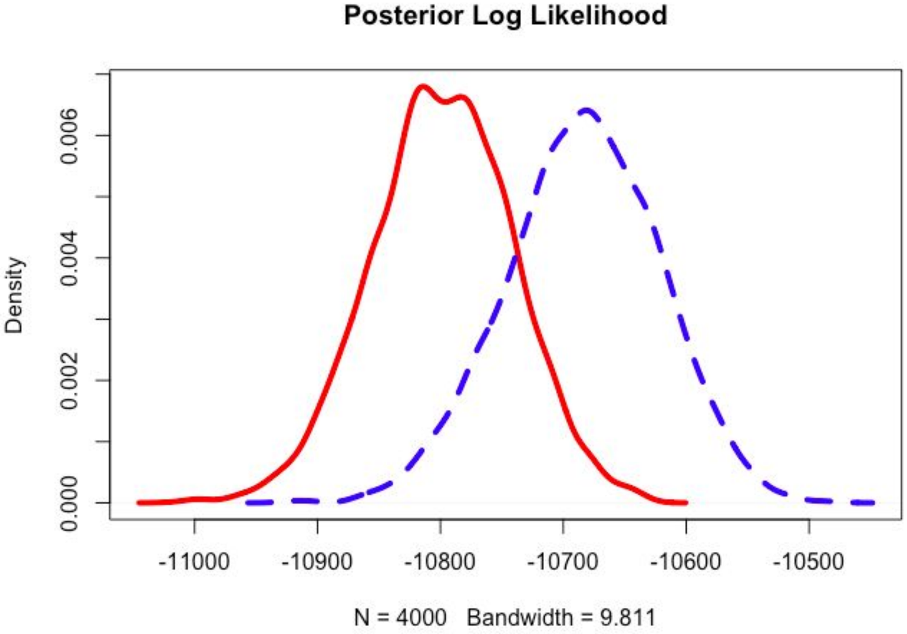
Tommy Jones and Max Kearns

Goals

- Better understand the analysis of text data
- Improve the analysis of text Data in a Bayesian Context
- Build a better model from an existing framework

BLUF: More complicated models aren’t always better.

https://imgflip.com/memetemplate/10089132/Sad-kitten



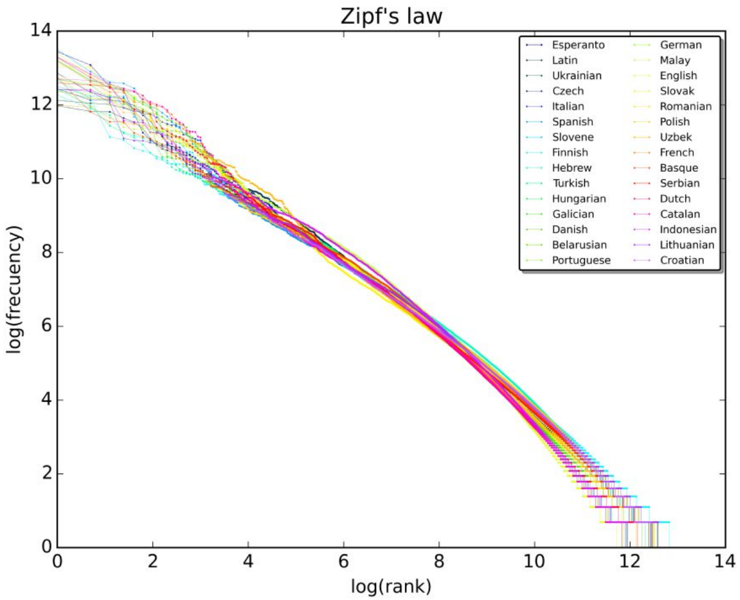
Existing Models for Text Analysis

- y is a vector of word counts (length=k)
- θ is a vector of word probabilities
- α is a vector of length k

- $y \sim \text{Multinom}(n, \theta)$
- $\theta \sim \text{Dir}(\alpha)$
- $\alpha \sim \text{Unif}$

Zipf’s Law

- Empirical law that holds for *all languages*
- When ordered by rank, frequency follows a harmonic series
- Language models may benefit from this prior knowledge



Putting Zipf's law into a Hyperprior

$$E[\vec{y}] = E\left[E\left[\vec{y}|\vec{\theta}\right]\right]$$

$$E[\vec{y}] = E\left[n\vec{\theta}\right]$$

$$E[\vec{y}] = nE[\vec{\theta}]$$

$$E[\vec{y}] = n\frac{\vec{\alpha}}{\sum_k \alpha_k}$$

$$E[\vec{y}] \propto \vec{\alpha}$$

- May get better estimates of α with more uncertainty
- Can create a prior based on inherent properties of language
- Zipf's law provides a framework for a prior on α

Main Model

- y ~ Multinom(n, θ)
- θ ~ Dir(α)
- α ~ Pareto(γ, β)
- β ~ 1/β

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

Data

- 100 randomly sampled NIH grant abstracts from 2014
- 5,542 unique words
- Most common word (‘the’) appears 1928 times
- 2,479 words appear one time
- ‘toxicology’ appears 7 times

Conditional Posterior Distributions

Main Model

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

$$\vec{\alpha}|\vec{\theta}, \vec{\beta}, \vec{y} \sim Unknown$$

$$\beta|\vec{\theta}, \vec{\alpha}, \vec{y} \sim Gamma\left(k, \sum_k log(\alpha_k) - klog(\gamma)\right)$$

Control Model

$$\vec{\theta}|\vec{\alpha}, \vec{y} \sim Dir(\vec{\alpha} + \vec{y})$$

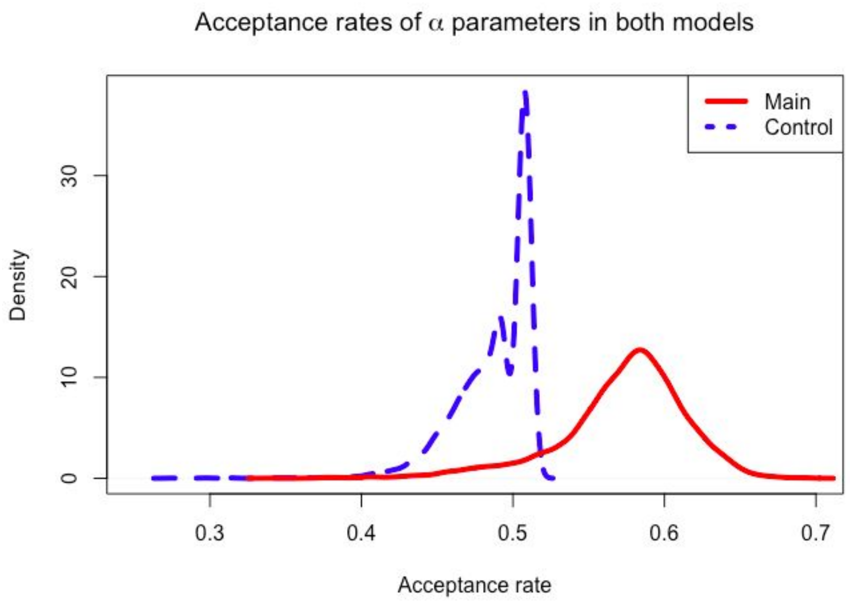
$$\vec{\alpha}|\vec{\theta}, \vec{y} \sim Unknown$$

Sampling

- Sampled 4 chains of 20,000 iterations
- 4,000 samples remain after the 50% burn-in and 10% thinning
- Proposal distribution for α is Inverse-Gaussian(0.1, 0.01)

MCMC Diagnostics

Acceptance Rates

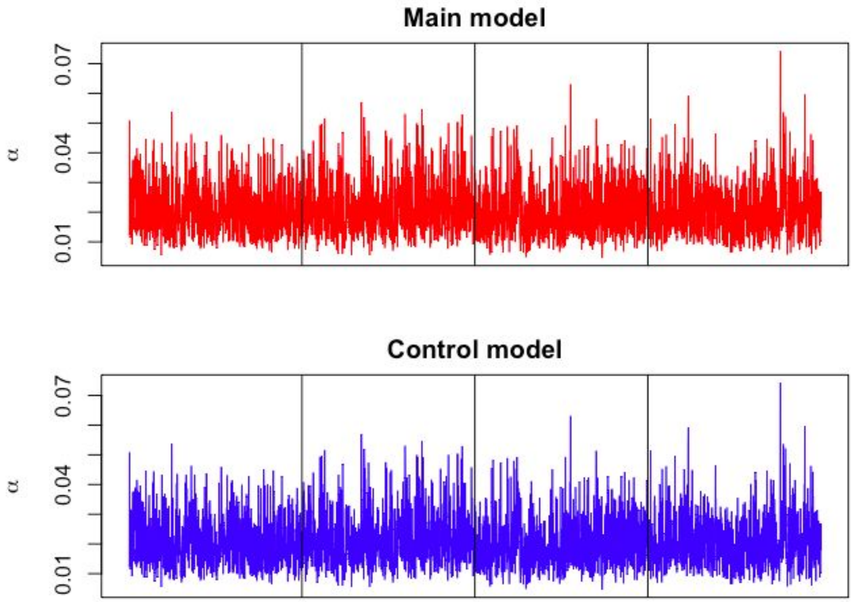


	Main	Control
Minimum	35.15%	27.46%
First Quartile	55.4%	47.38%
Median	57.89%	49.26%
Third Quartile	59.94%	50.73%
Maximum	69.51%	51.46%

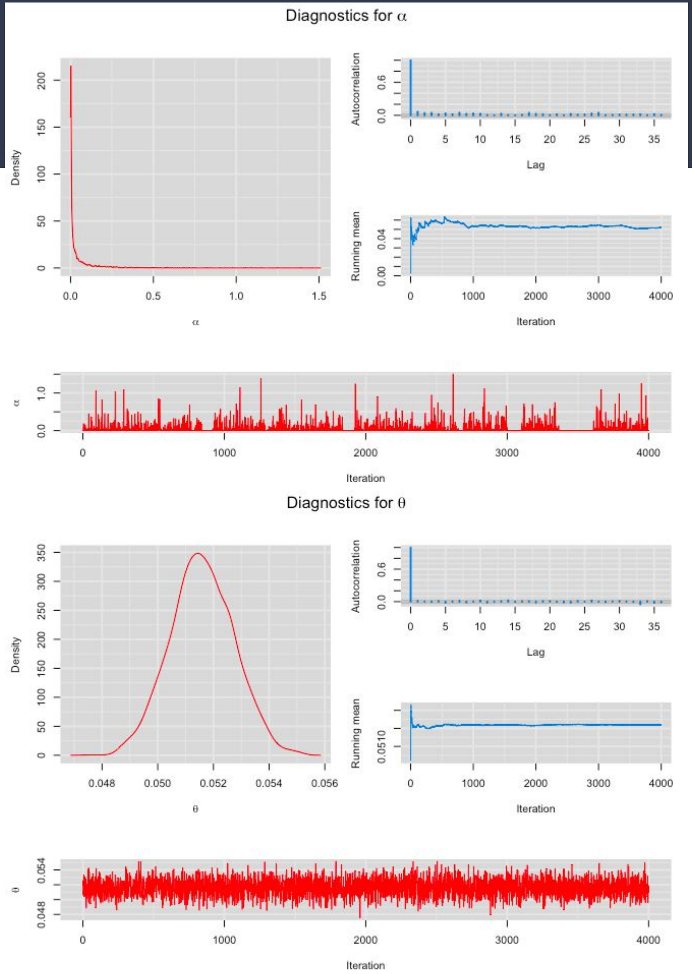
Convergence

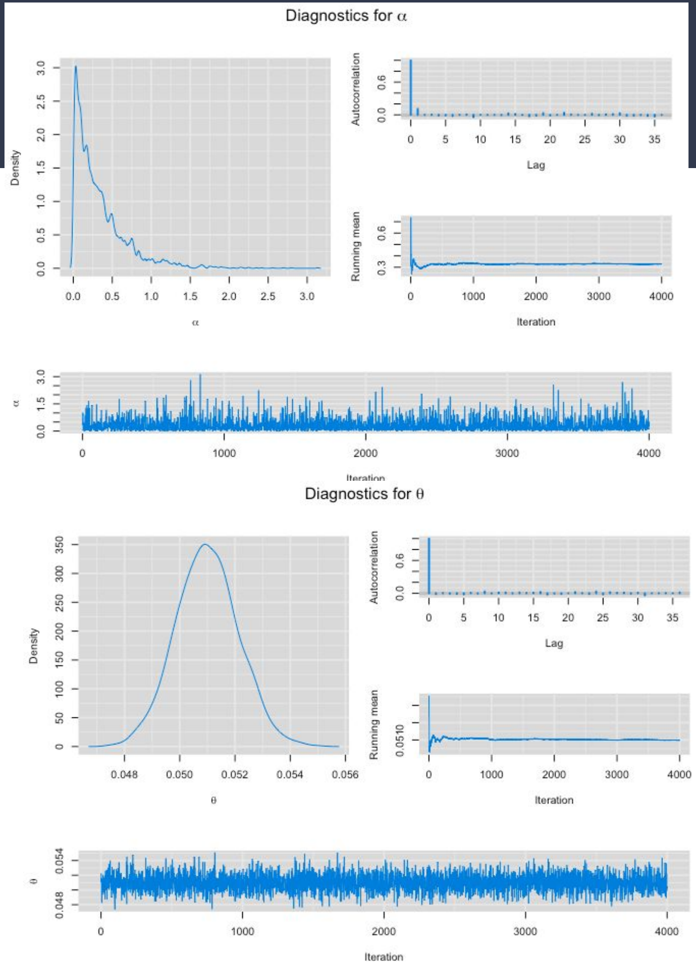
Main θ	Main α	Control θ	Control α
5.1	38.9	5.8	7.6
5.0	26.3	4.9	6.1
5.7	36.3	5.6	6.7
5.5	38.3	5.8	6.9

Percent of Geweke statistics greater than 1.96 in absolute value

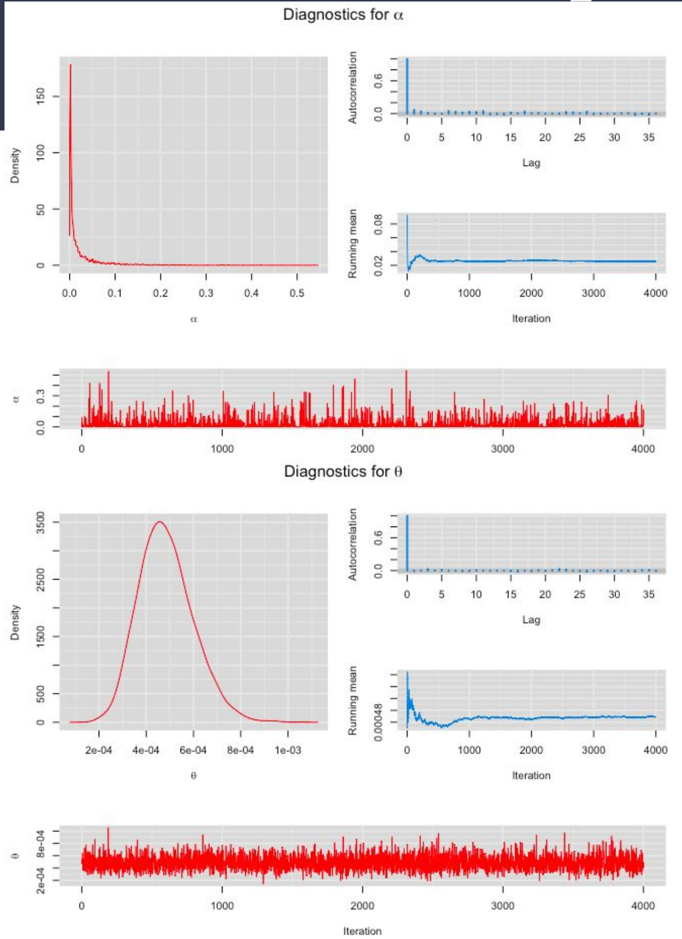


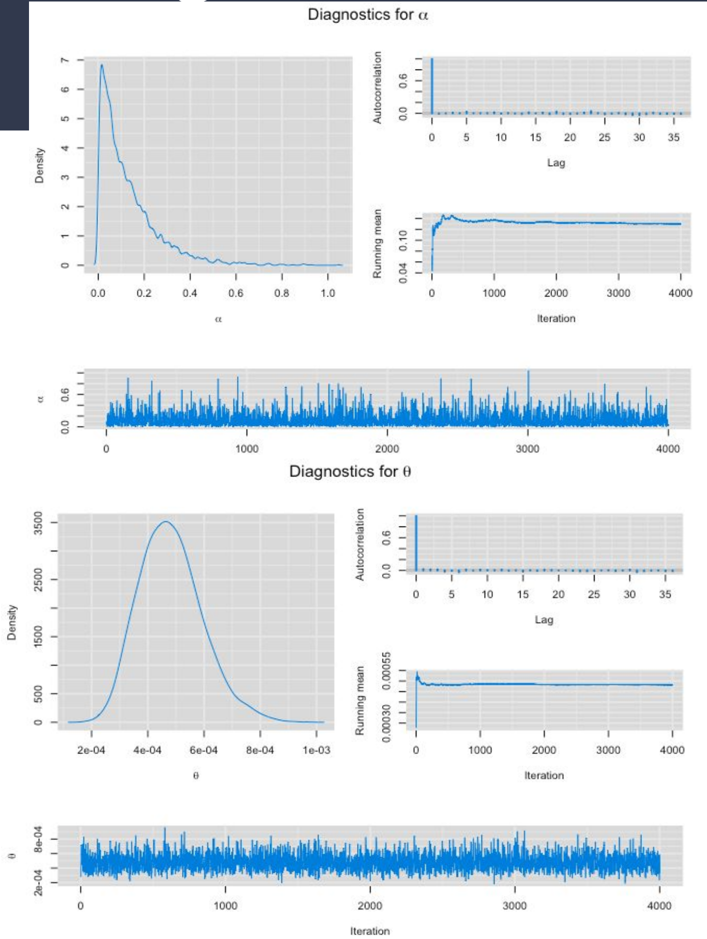
Most Common Word: ‘the’



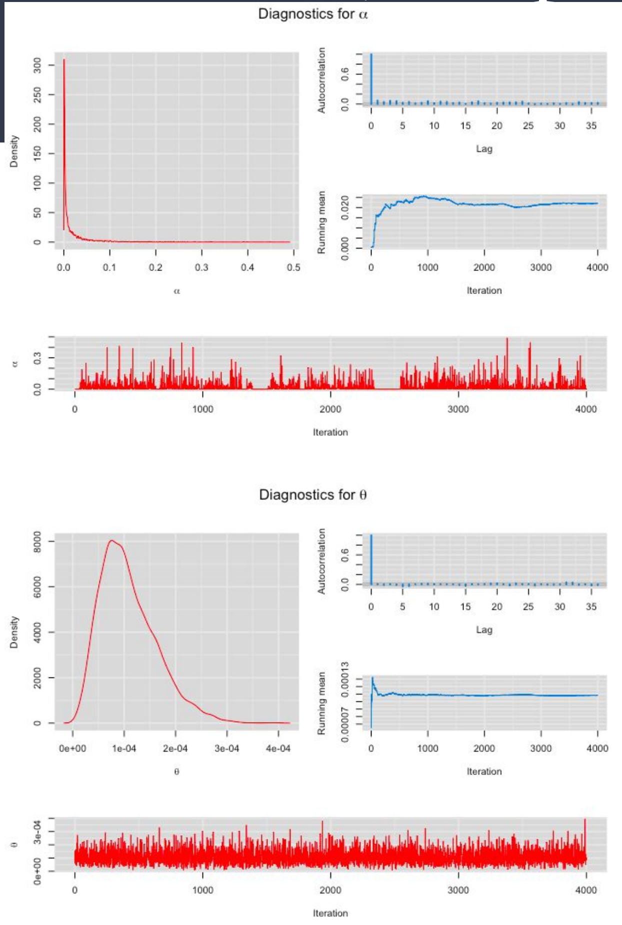


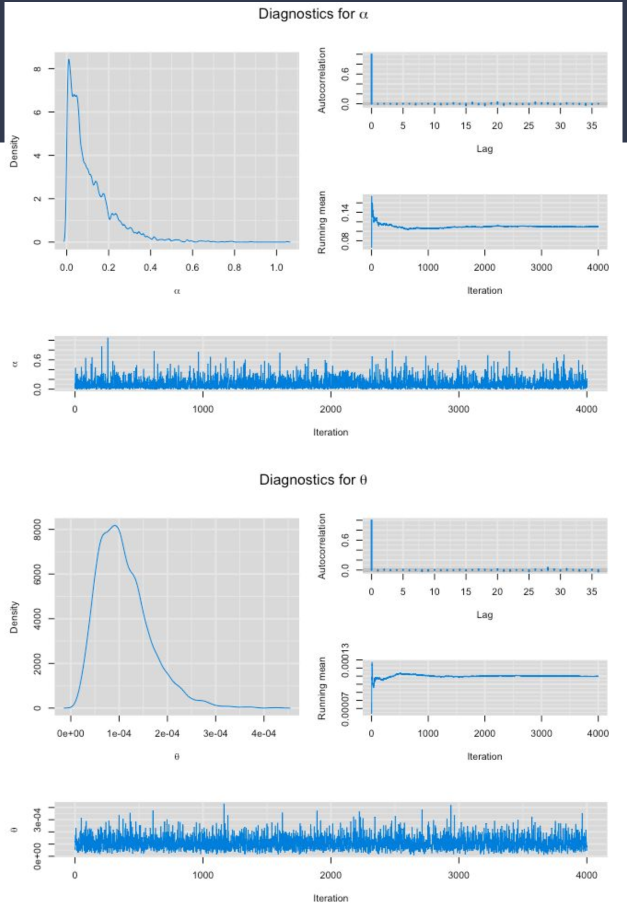
Word at the 95th percentile: ‘regulation’



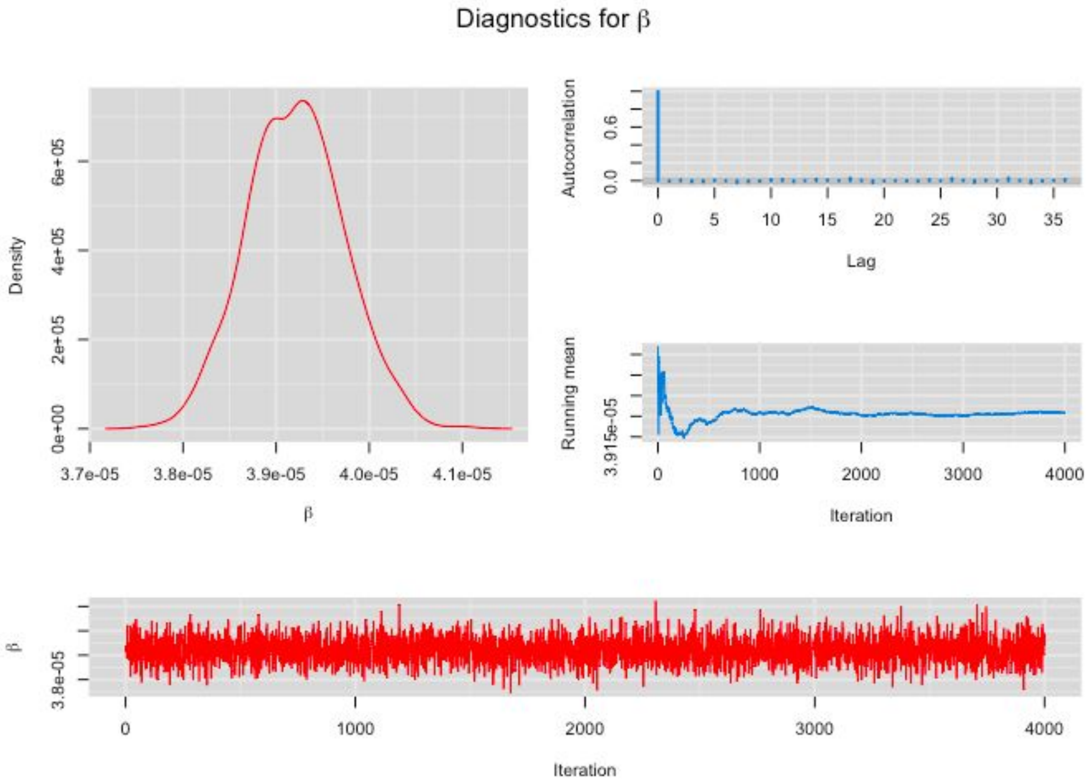


Word at the 75th percentile: ‘men’





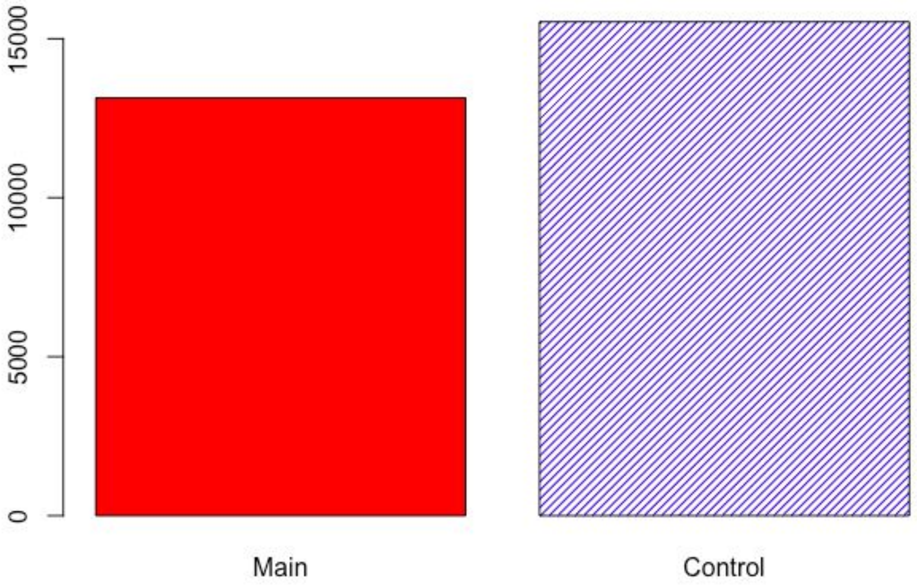
Beta



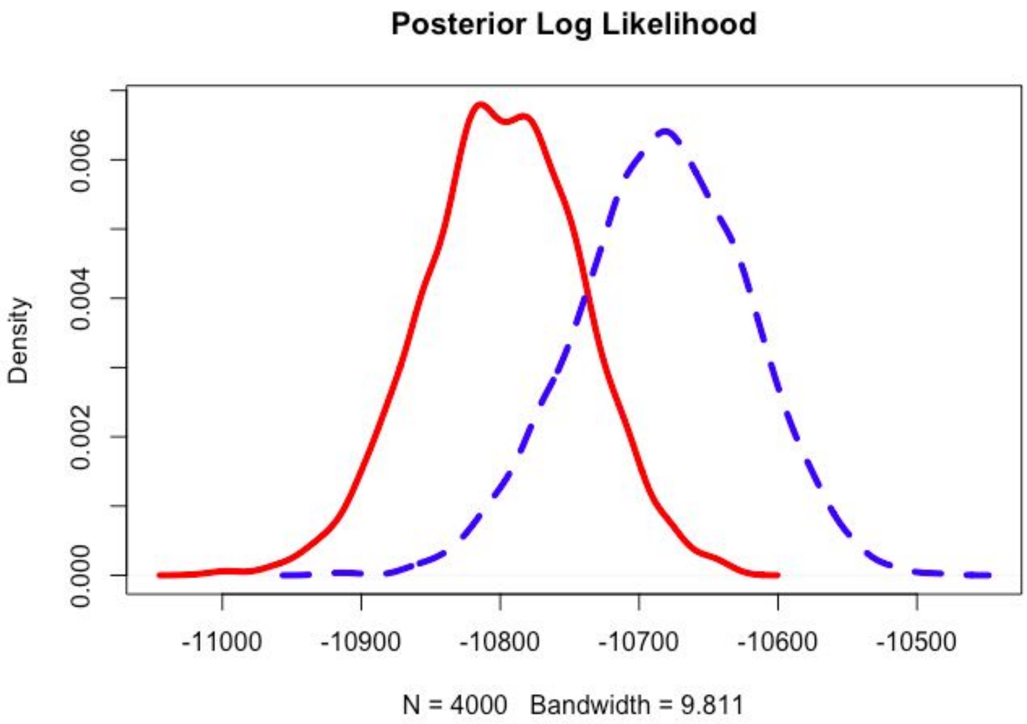
Diagnostics for β
Density
0e+00
2e+05
4e+05
6e+05
3.7e-05
3.8e-05
3.9e-05
4.0e-05
4.1e-05
β
Autocorrelation
0.0
0.6
0
5
10
15
20
25
30
35
Lag
Running mean
3.915e-05
0
1000
2000
3000
4000
Iteration
β
3.8e-05
0
1000
2000
3000
4000
Iteration

Model Comparison

DIC

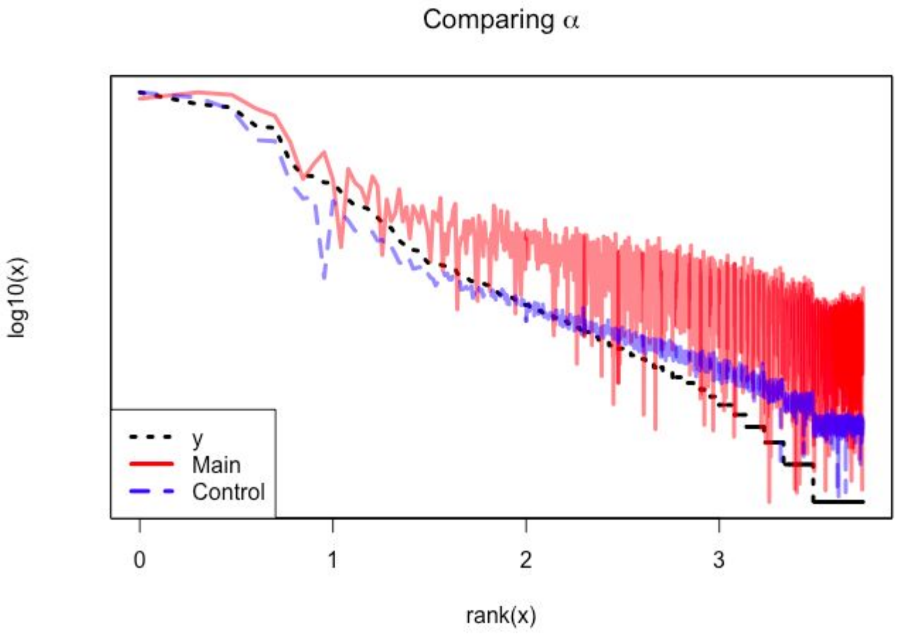


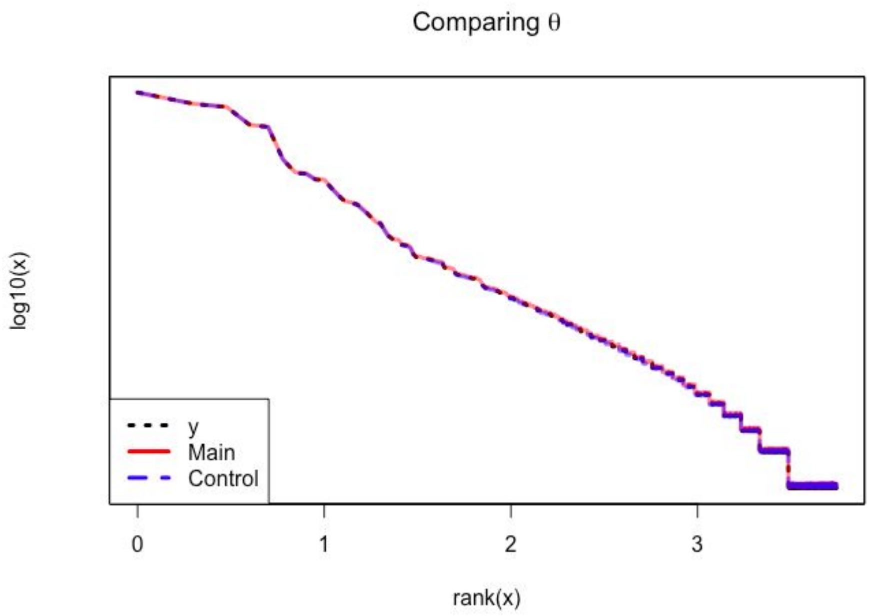
Log-likelihood



https://imgflip.com/memetemplate/10089132/Sad-kitten

Remember our core observation: $E[\vec{y}] \propto \vec{\alpha}$





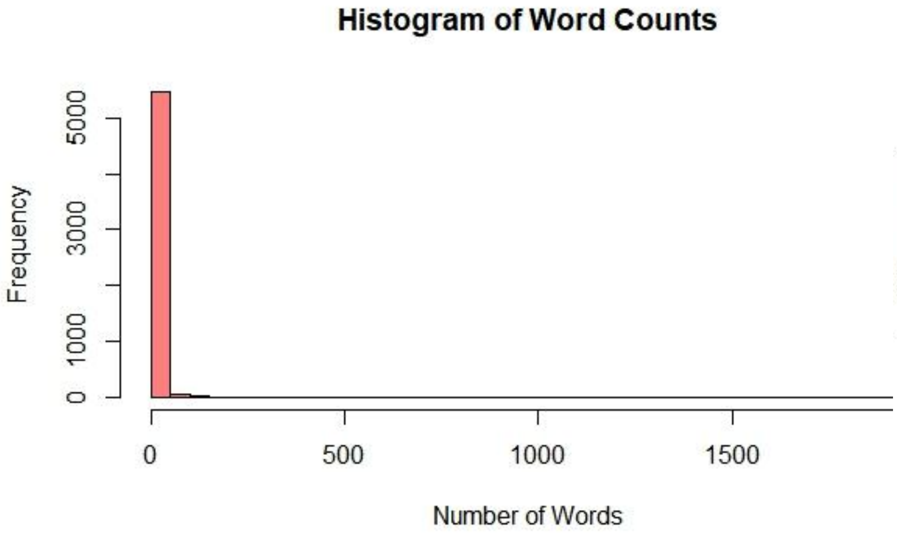
Discussion/Questions

References

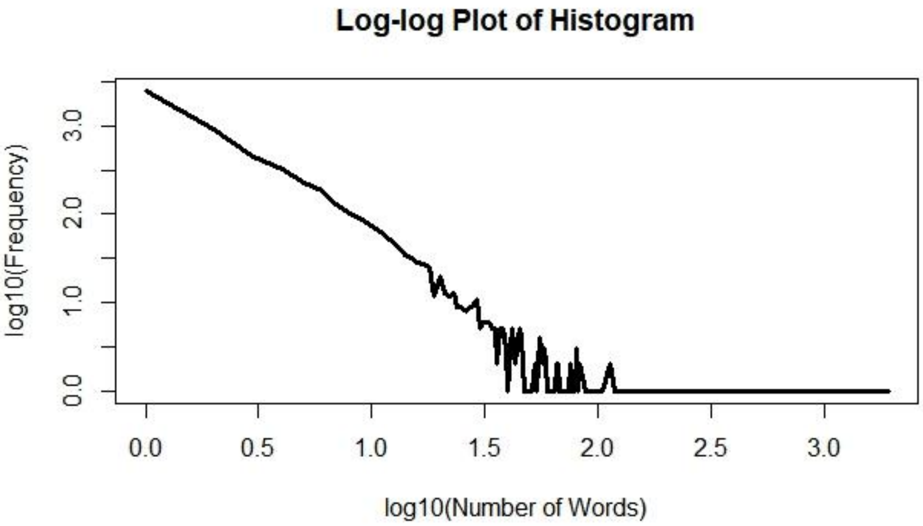
- Meyers, G. (1994). Quantifying the Uncertainty in Claim Severity Estimates for an Excess Layer When Using the Single Parameter Pareto. In Proceedings of the Casualty Actuarial Society (Vol. 81, pp. 91-122).
- Zipf, G. K. (2016). Human behavior and the principle of least effort: An introduction to human ecology. Ravenio Books.
- Manaris, B. Z., Pellicoro, L., Pothering, G., & Hodges, H. (2006, February). Investigating Esperanto's Statistical Proportions Relative to other Languages using Neural Networks and Zipf's Law. In Artificial Intelligence and Applications (pp. 102-108).
- National Institutes of Health. (2014). NIH ExPORTER. Retrieved from http://exporter.nih.gov/ExPORTER_Catalog.aspx (January 2015)

Backup

Data



Histogram of Word Counts
Frequency
0
1000
3000
5000
0
500
1000
1500
Number of Words



Log-log Plot of Histogram
log10(Frequency)
0.0
1.0
2.0
3.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
log10(Number of Words)

Conditional Distributions: Control

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

$$= \left[\prod_k \theta_k^{y_k} \right] \left[\mathcal{B}(\vec{\alpha}) \prod_k \theta_k^{\alpha_k - 1} \right]$$

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

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

$$P(\vec{\alpha} | \vec{\theta}, \vec{y}) \propto \mathcal{B}(\vec{\alpha}) \prod_k \theta_k^{\alpha_k}$$

$$P(\alpha_k | \theta_k, y_k) \propto \theta_k^{\alpha_k}$$

Conditional Distributions: Main

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

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

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

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

Conditional Distributions: Main, cont.

$$
\begin{aligned}
P(\beta|\vec{\theta}, \vec{\alpha}, \vec{y}) &\propto \beta^{K-1}\gamma^{\beta k}(\prod_k \alpha_k)^{-(\beta+1)} \\
&\propto \beta^{K-1}\gamma^{\beta k}(\prod_k \alpha_k)^{-\beta} \\
&\propto \beta^{K-1}\exp\left[-\beta\left(\sum_k \log(\alpha_k) - k\log(\gamma)\right)\right] \\
&\implies \beta|\vec{\theta}, \vec{\alpha}, \vec{y} \sim Gamma\left(k, \sum_k \log(\alpha_k) - k\log(\gamma)\right)
\end{aligned}
$$