Math 640 Final Paper

Max Kearns and Tommy Jones

May 10, 2018

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 with a Dirichlet prior on a Multinomial likelihood. The prior provides uncertainty on $\vec{\theta}$, which is the vector of word probabilities. The usual model then assumes a non-informative uniform prior on the Dirichlet parameter ($\vec{\alpha}$). 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 $\vec{\alpha}$ would improve the model. Zipf's law provides a basis for how to vary $\vec{\alpha}$ 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(γ, β) prior on $\vec{\alpha}$, where γ is fixed at a sufficiently small value. This Pareto distribution is a power-law that should provide some uncertainty on $\vec{\alpha}$ that mirrors this 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, a Pareto hyper-prior, with a non-informative Jeffrey's hyper-prior . We will compare this model to a control that does not allow for uncertainty on $\vec{\alpha}$. The data set is 100 randomly sampled NIH grant abstracts from 2014.

2 Methods

2.1 Models

For both the control and the experimental model, we assume that the word count $\vec{y}$ is Multinomial, so has the following likelihood.

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

On $\vec{\theta}$, the vector of word probabilities, we place a Dirichlet prior.

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

The control model will stop here, and assume a uniform distibution on α_k. The control model then has the simple form $\vec{\theta}|\vec{y}, \vec{\alpha} \sim Dir(\vec{y} + \vec{\alpha})$, and an unknown distribution for $\vec{\alpha}|\vec{y}, \vec{\beta}$; shown below (the full derivation can be found in Appendix B). We will sample from $\vec{\alpha}|\vec{y}, \vec{\beta}$ with an Inverse-Gaussian proposal.

$$p(\alpha_k|\theta_k, y_k) \propto \theta_k^{\alpha_k} \tag{3}$$

The candidate model, however, will assume a $Pareto(\gamma, \beta)$ prior on α_k, with Jeffrey's prior on β.

$$\alpha_k \sim Pareto(\gamma, \beta) \implies \pi(\vec{\alpha}) = \prod_k \gamma^\beta \beta \alpha_k^{-(\beta+1)} \tag{4}$$

$$\pi(\beta) \propto \frac{1}{\beta} \tag{5}$$

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

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

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

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

$$P(\vec{\alpha} | \vec{\theta}, \beta, \vec{y}) \propto \mathcal{B}(\vec{\alpha}) \prod_k \theta_k^{y_k + \alpha_k - 1} \alpha_k^{-(\beta+1)} \tag{8}$$

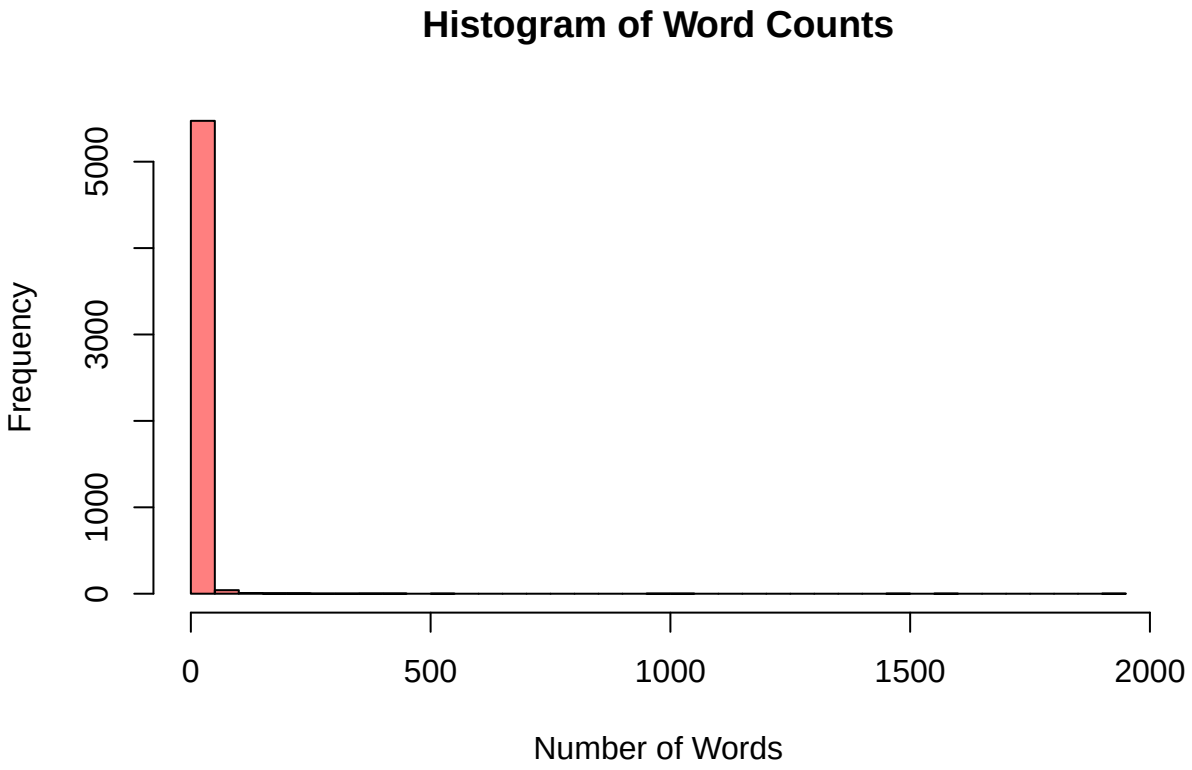
$$P(\beta | \vec{\theta}, \vec{\alpha}, \vec{y}) \propto \beta^{K-1} \exp\left[-\beta \left(\sum_k \log(\alpha_k) - k \log(\gamma) \right) \right] \tag{9}$$

$$\tag{10}$$

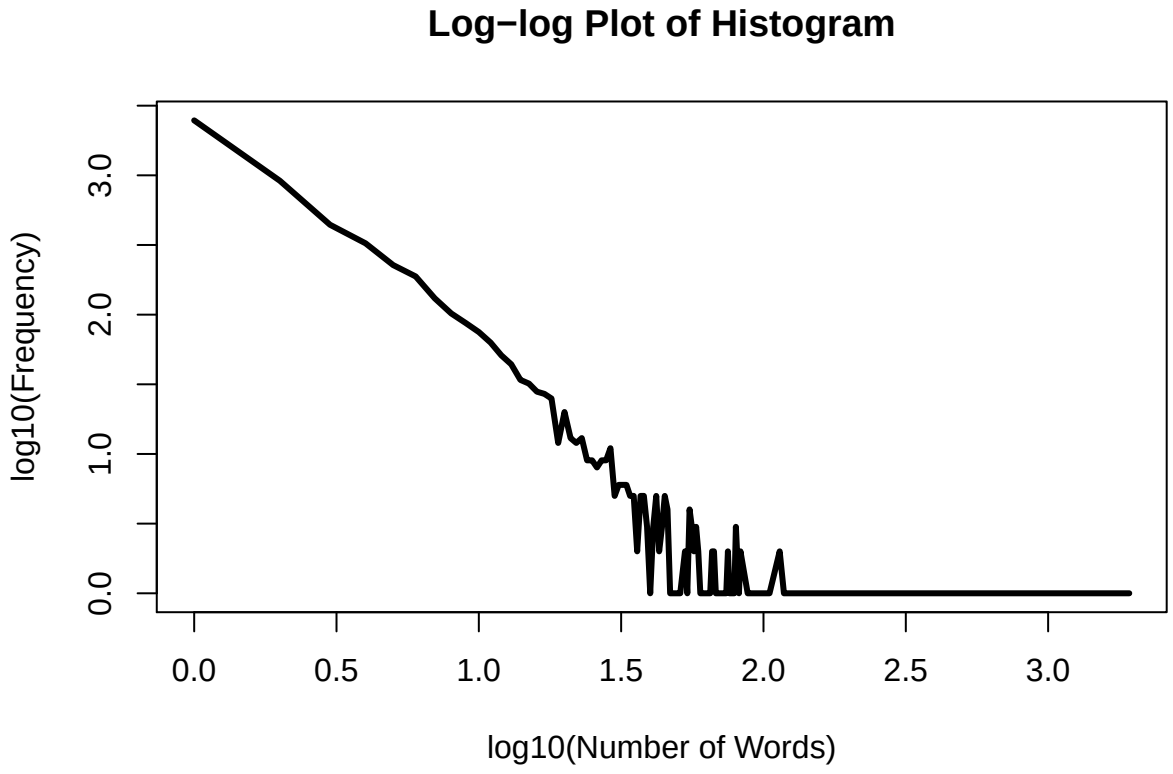
Of these conditional posteriors, only $\vec{\alpha} | \vec{\theta}, \beta, \vec{y}$ is an unknown distribution. $\vec{\theta} | \vec{\alpha}, \beta, \vec{y}$ is a $Dir(\vec{y} + \vec{\alpha})$, and $\beta | \vec{\theta}, \vec{\alpha}, \vec{y}$ is a Gamma$\left(k, \sum_k \log(\alpha_k) - k \log(\gamma) \right)$. The other two conditionals will be sampled using a Metropolis-Hastings algorithm with another Inverse-Gaussian proposal.

2.2 Sampling

2.2.1 Description of data



Describe Zipf's law plot here

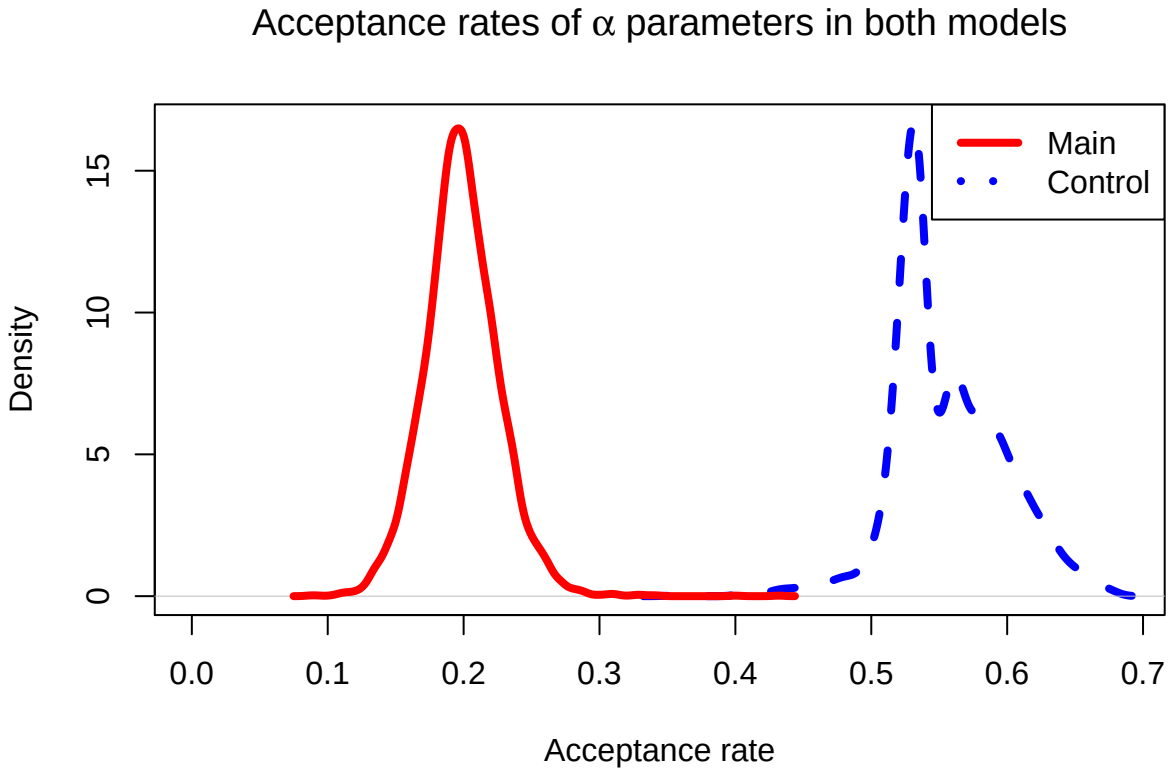


2.2.2 Sampling the control model

2.2.3 Sampling the main model

2.3 Comparison

We should put some posterior plots here, maybe



	Main	Control
min.	0.09	0.35
25%	0.18	0.53
50%	0.20	0.55
75%	0.22	0.58
max.	0.43	0.68

Convergence is based on log likelihood of both models

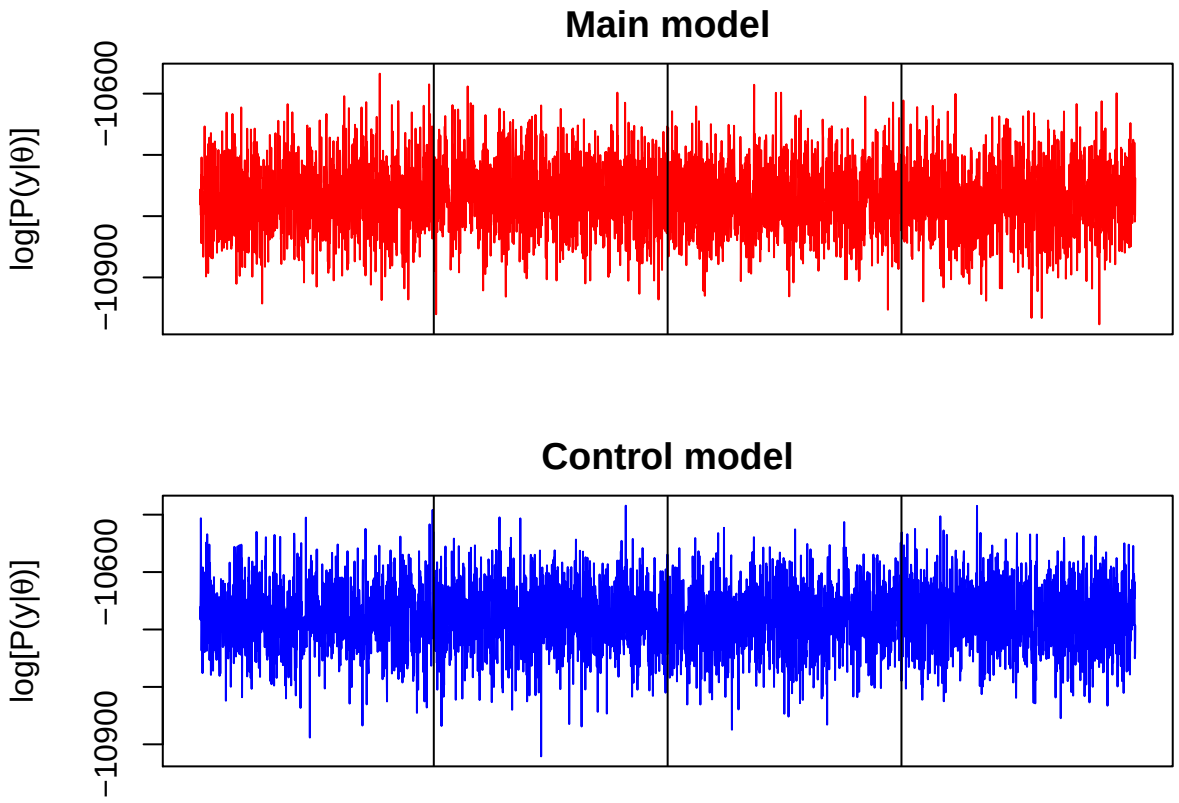
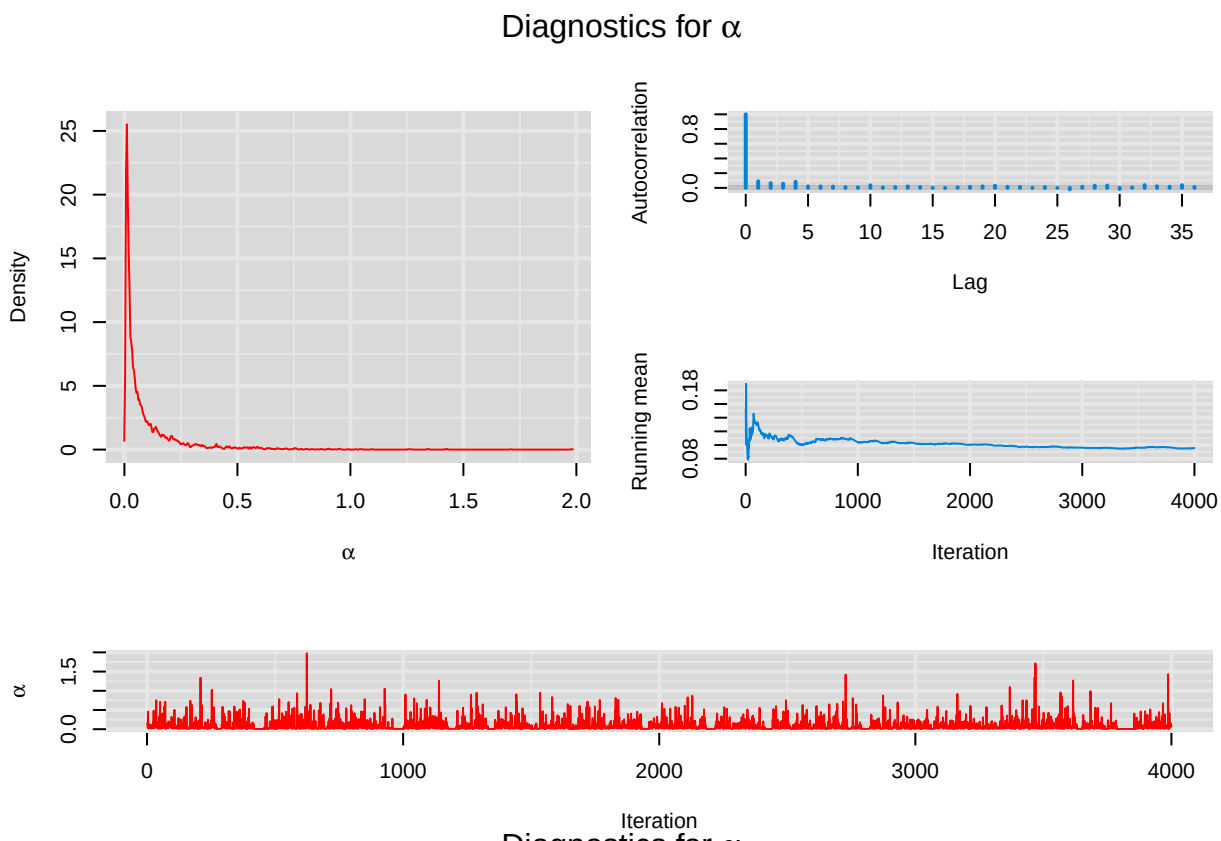
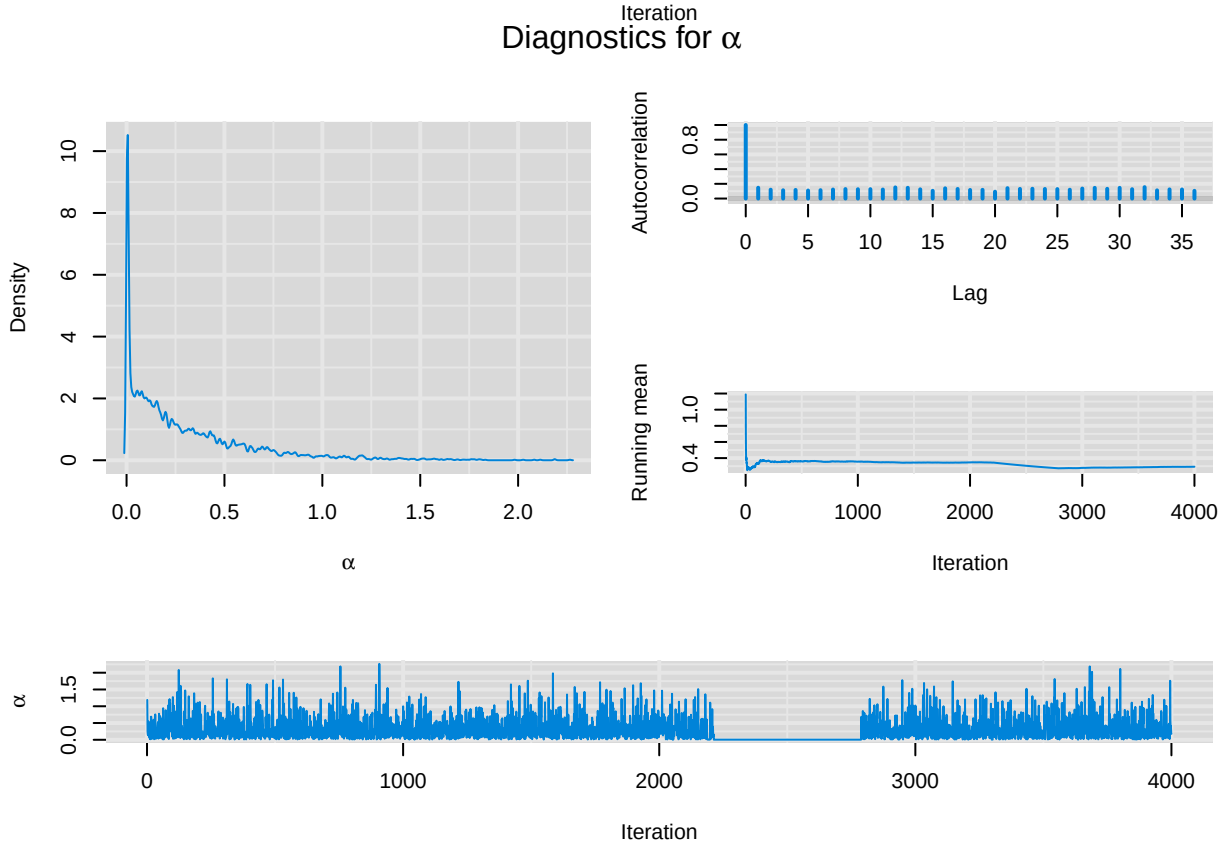


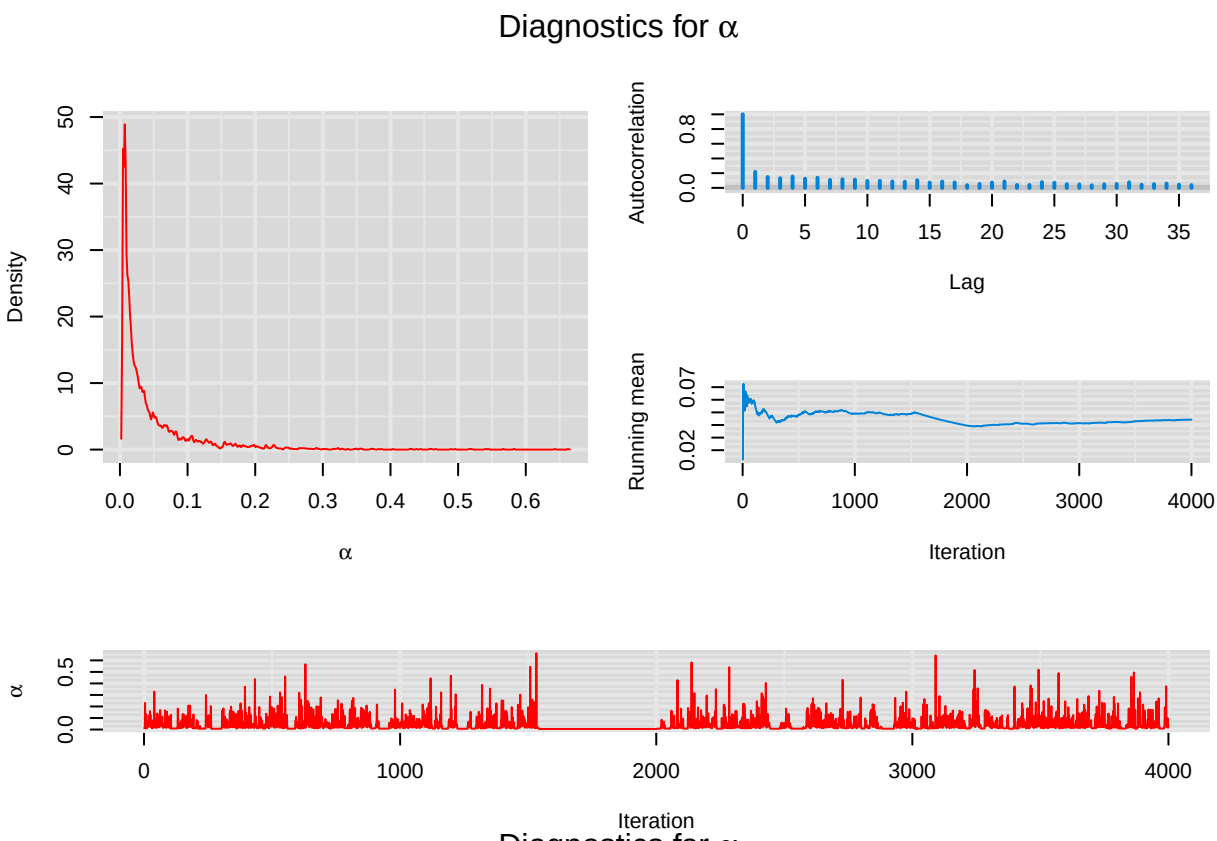
Table 2: Geweke statistic of log likelihoods

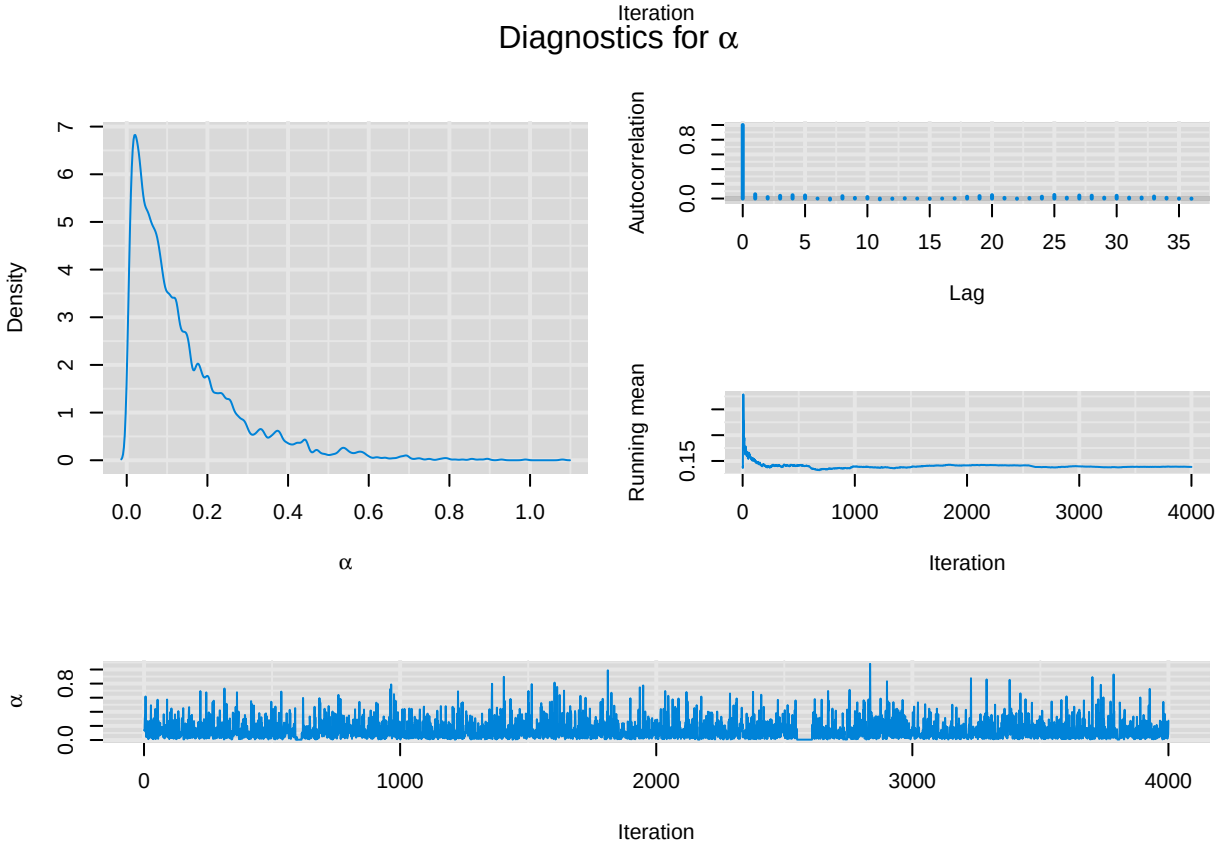
Main	Control
0.26	-0.18

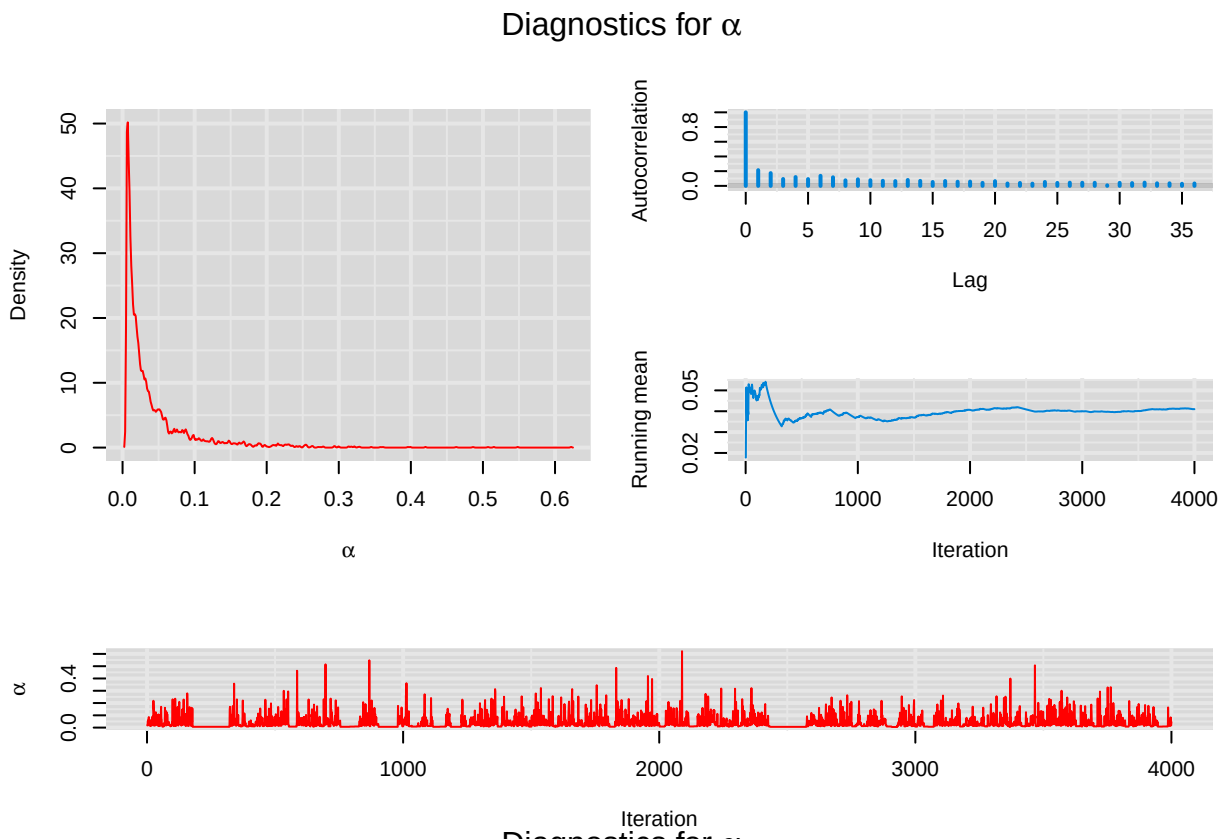
MCMC plots of selected words

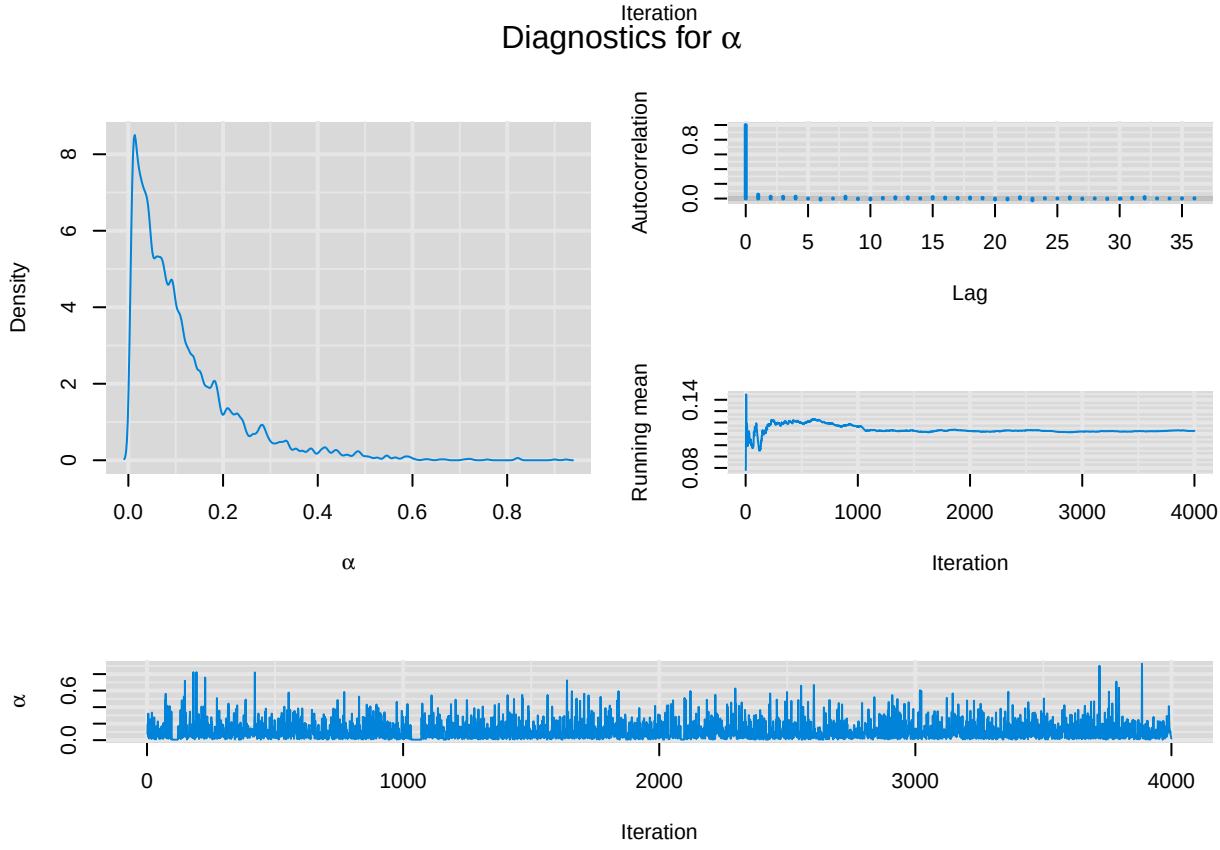

Diagnostics for α
Density
α
Autocorrelation
Lag
Running mean
Iteration
α
Iteration

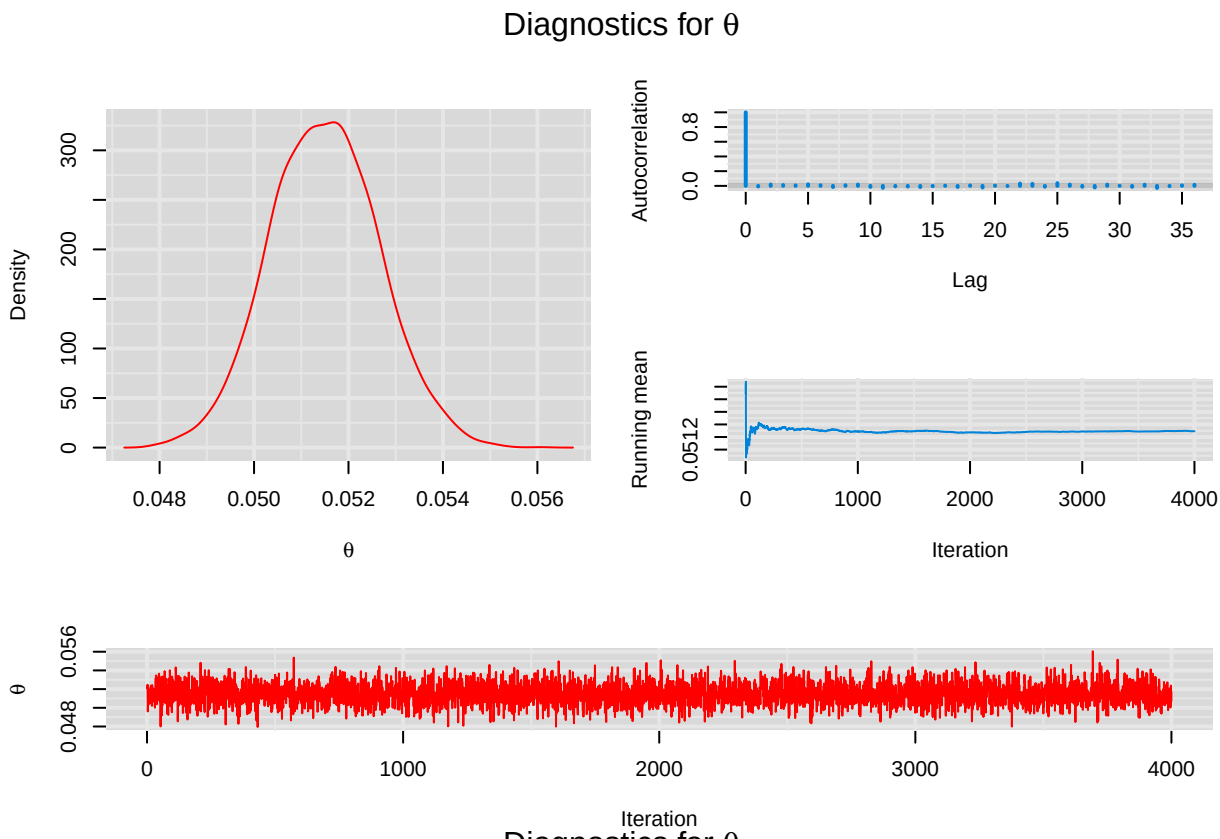

Diagnostics for α
Density
α
Autocorrelation
Lag
Running mean
Iteration
α
Iteration

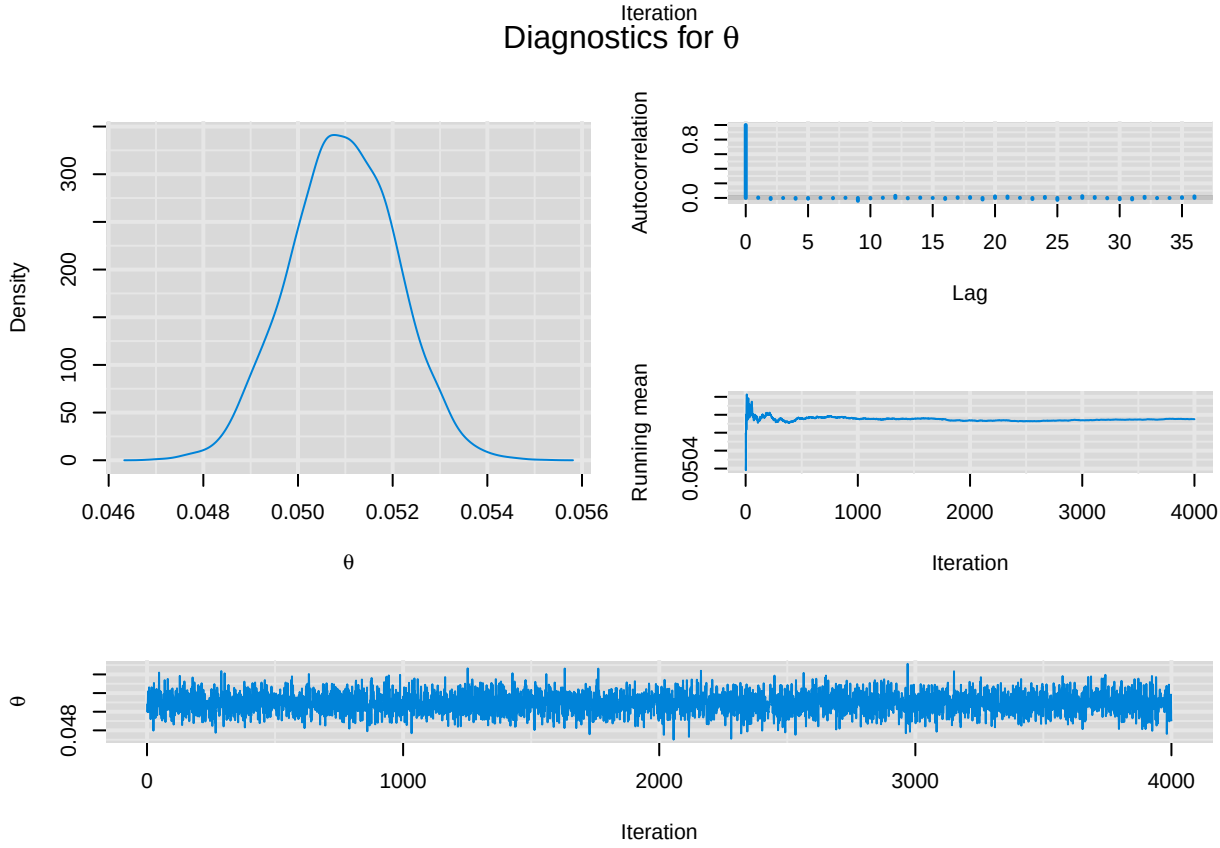

Diagnostics for α
Density
α
Autocorrelation
Lag
Running mean
Iteration
α
Iteration

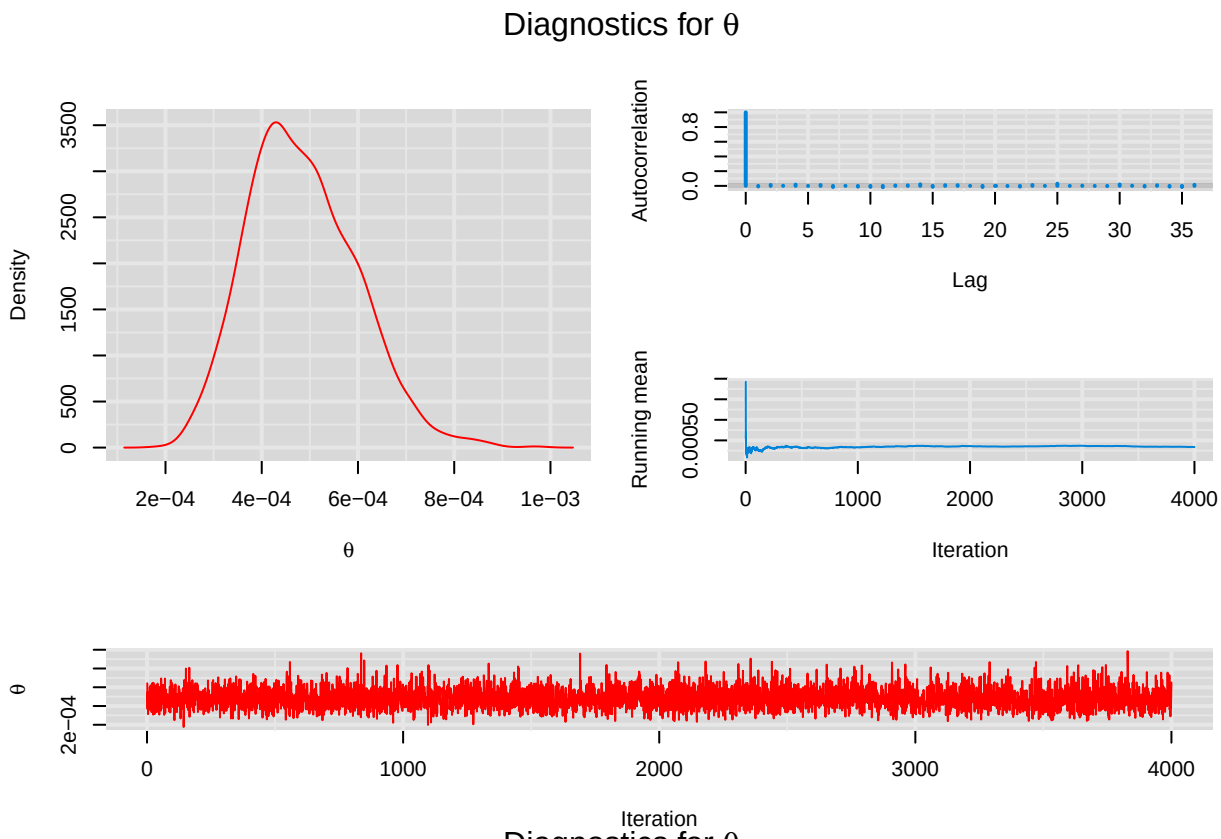

Diagnostics for α
Density
α
Autocorrelation
Lag
Running mean
Iteration
α
Iteration

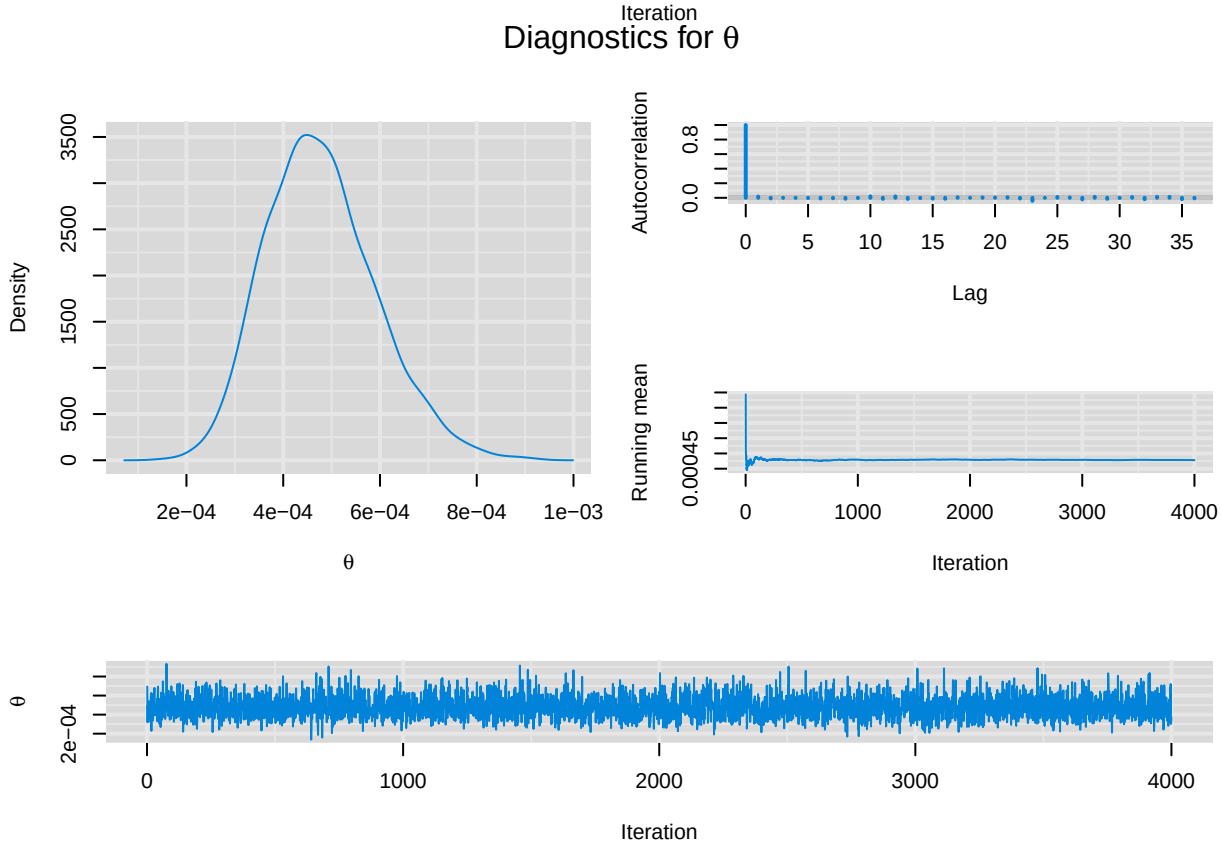

Diagnostics for α
Density
α
Autocorrelation
Lag
Running mean
Iteration
α
Iteration

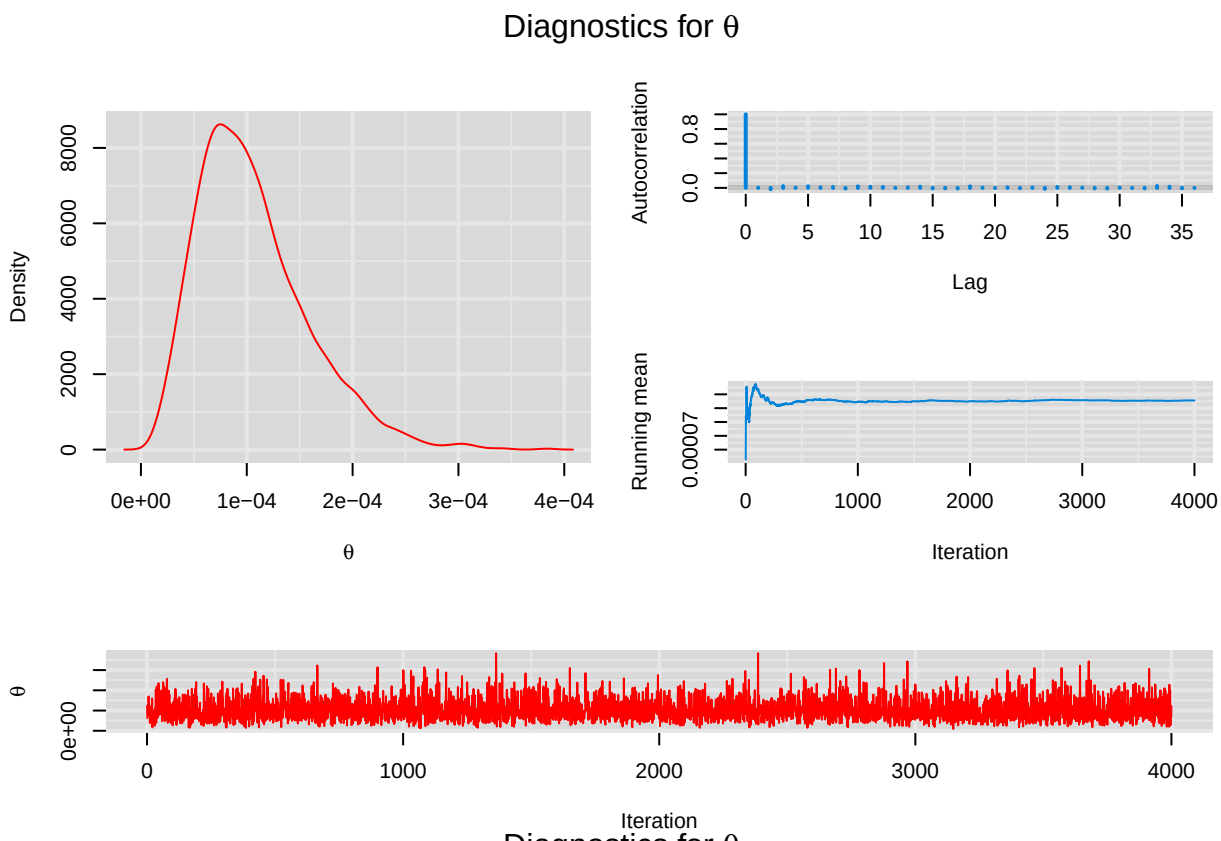

Diagnostics for α
Density
α
Autocorrelation
Lag
Running mean
Iteration
α
Iteration

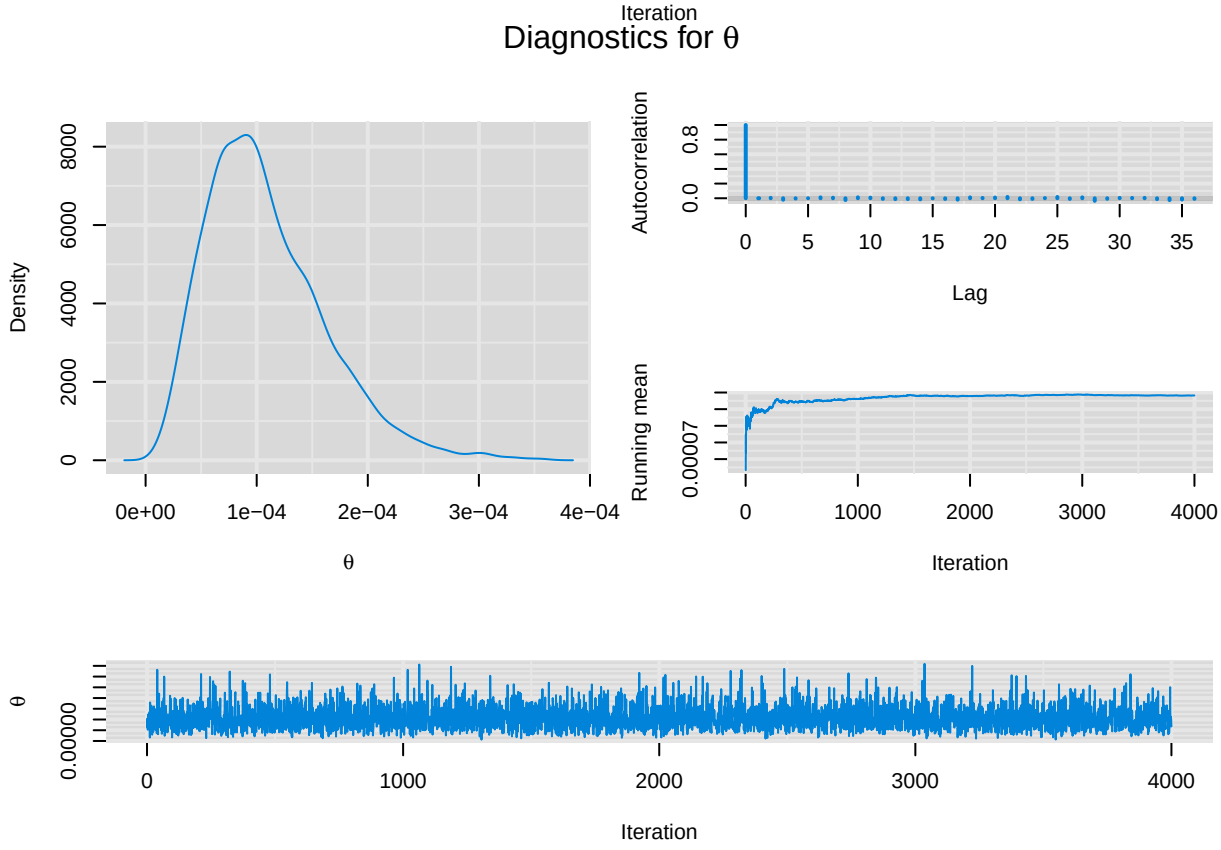

Diagnostics for θ
Density
θ
Autocorrelation
Lag
Running mean
Iteration
θ
Iteration

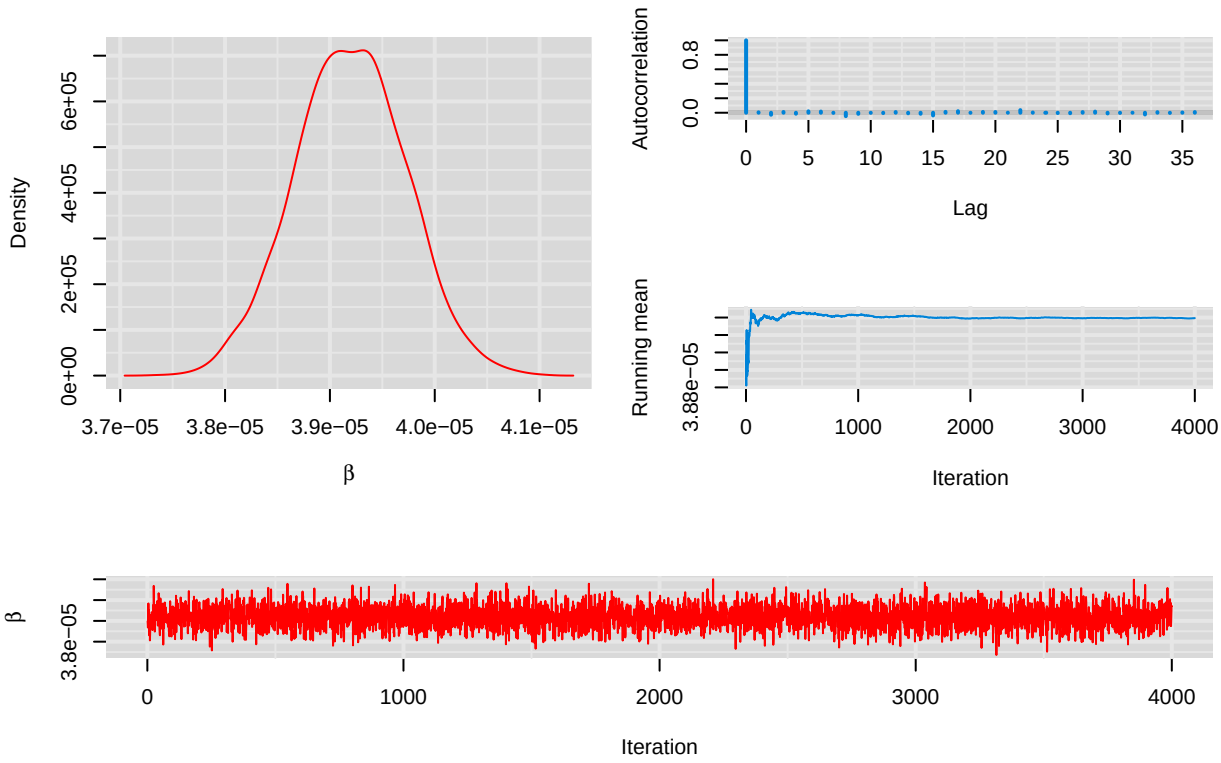

Diagnostics for θ
Density
θ
Autocorrelation
Lag
Running mean
Iteration
θ
Iteration


Diagnostics for θ
Density
θ
Autocorrelation
Lag
Running mean
Iteration
θ
Iteration

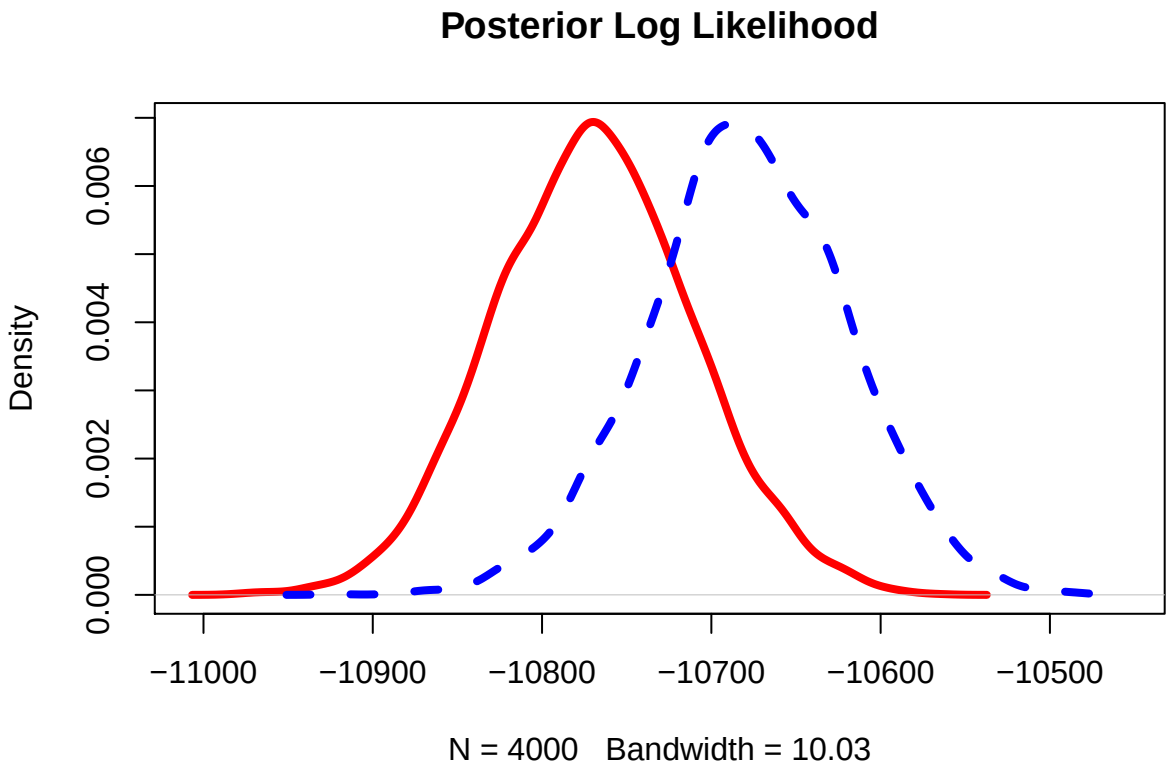

Diagnostics for θ
Density
θ
Autocorrelation
Lag
Running mean
Iteration
θ
Iteration


Diagnostics for θ
Density
θ
Autocorrelation
Lag
Running mean
Iteration
θ
Iteration

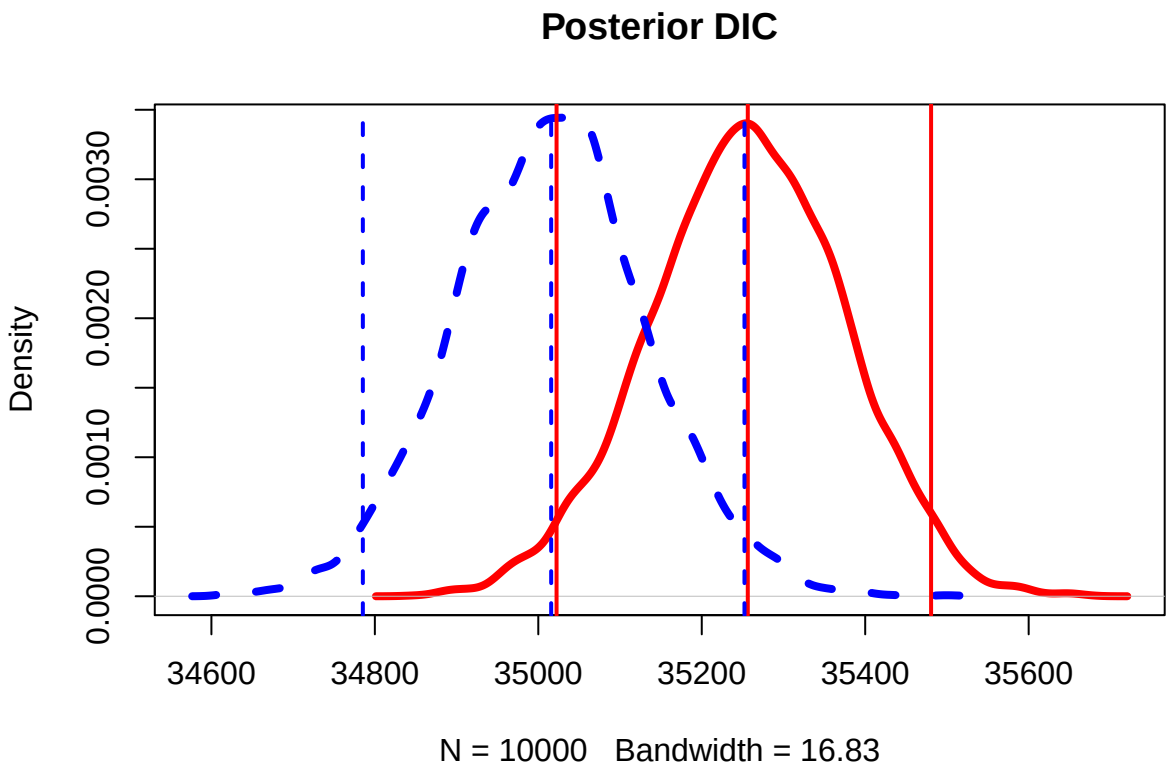

Diagnostics for θ
Density
θ
Autocorrelation
Lag
Running mean
Iteration
θ
Iteration



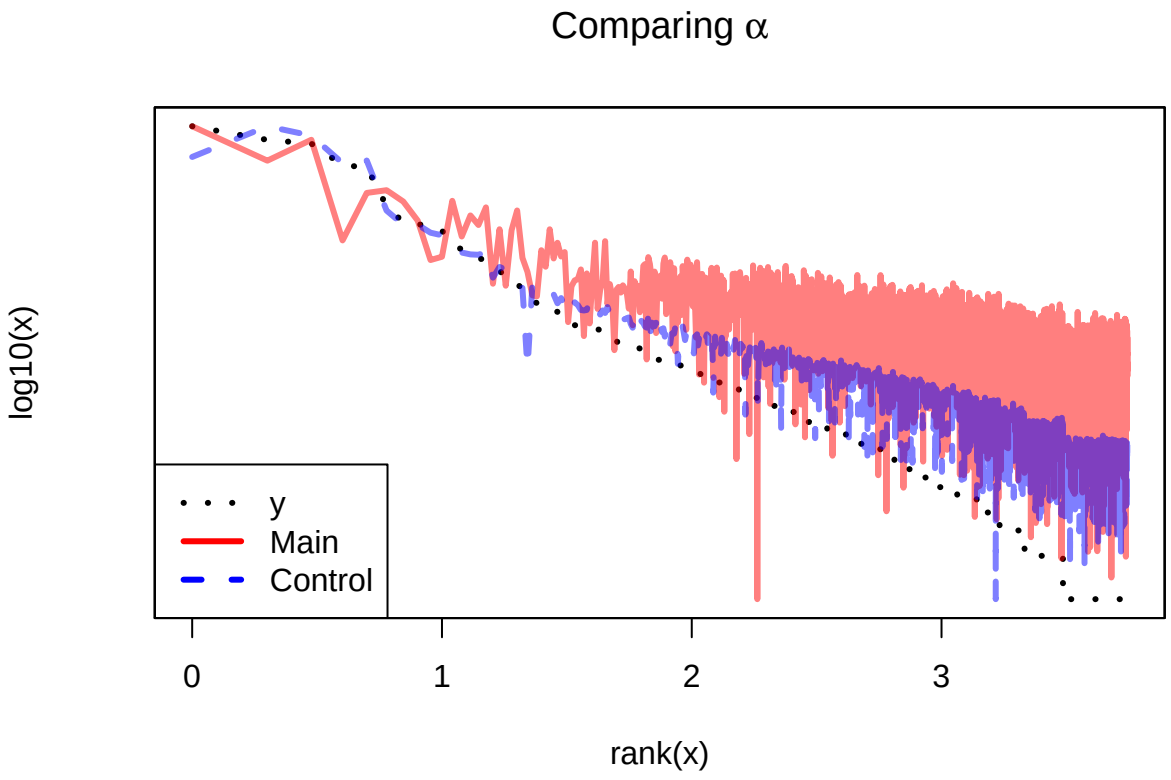
log posterior

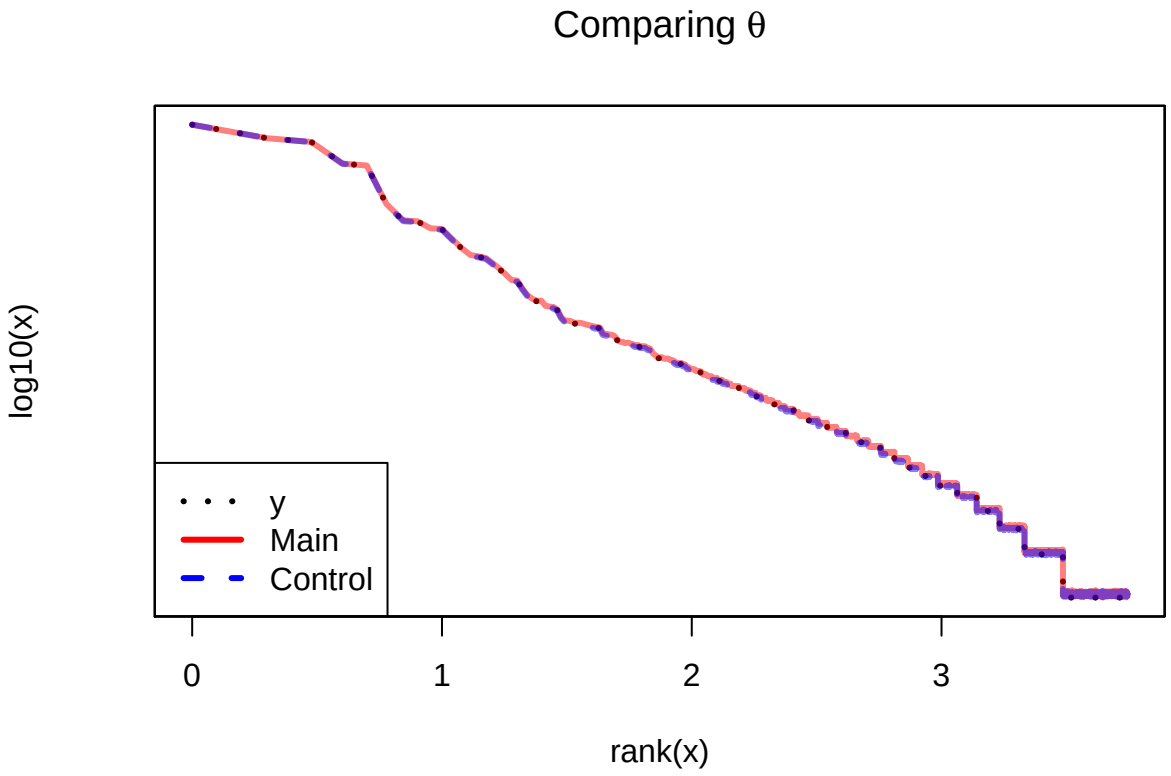


DIC, now in distribution form!



full circle: zipfs law and our models




Comparing θ
log10(x)
y
Main
Control
0
1
2
3
rank(x)

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

7.1 Control Posterior Derivations

$$P(\vec{\theta}, \vec{\alpha}|\vec{y}) \propto \left[\prod_k \theta_k^{y_k}\right]\left[\mathcal{B}(\vec{\alpha})\prod_k \theta_k^{\alpha_k - 1}\right] \times 1 \tag{11}$$

$$= \left[\prod_k \theta_k^{y_k}\right]\left[\mathcal{B}(\vec{\alpha})\prod_k \theta_k^{\alpha_k - 1}\right] \tag{12}$$

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

$$\implies \vec{\theta}|\vec{\alpha}, \vec{y} \sim Dir(\vec{y} + \vec{\alpha}) \tag{14}$$

$$P(\vec{\alpha}|\vec{\theta}, \vec{y}) \propto \mathcal{B}(\vec{\alpha})\prod_k \theta_k^{\alpha_k} \tag{15}$$

$$P(\alpha_k|\theta_k, y_k) \propto \theta_k^{\alpha_k} \tag{16}$$

7.2 Main Posterior Derivations

$$P(\vec{\theta}, \vec{\alpha}, \beta|\vec{y}) \propto \left[\prod_k \theta_k^{y_k}\right]\left[\mathcal{B}(\vec{\alpha})\prod_k \theta_k^{\alpha_k-1}\right]\left[\prod_k \gamma^{\beta}\beta\alpha_k^{-(\beta+1)}\right] \tag{17}$$

$$= \beta^{K-1}\gamma^{\beta k}\mathcal{B}(\vec{\alpha})\prod_k \theta_k^{y_k+\alpha_k-1}\alpha_k^{-(\beta+1)} \tag{18}$$

$$P(\vec{\theta}|\vec{\alpha}, \beta, \vec{y}) \propto \prod_k \theta_k^{y_k+\alpha_k-1} \tag{19}$$

$$\implies \vec{\theta}|\vec{\alpha}, \beta, y \sim Dir(\vec{y}+\vec{\alpha}) \tag{20}$$

$$P(\vec{\alpha}|\vec{\theta}, \beta, \vec{y}) \propto \mathcal{B}(\vec{\alpha})\prod_k \theta_k^{y_k+\alpha_k-1}\alpha_k^{-(\beta+1)} \tag{21}$$

$$\implies \text{unknown distribution} \tag{22}$$

$$P(\beta|\vec{\theta}, \vec{\alpha}, \vec{y}) \propto \beta^{K-1}\gamma^{\beta k}(\prod_k \alpha_k)^{-(\beta+1)} \tag{23}$$

$$\propto \beta^{K-1}\gamma^{\beta k}(\prod_k \alpha_k)^{-\beta} \tag{24}$$

$$\propto \beta^{K-1}\exp\left[-\beta\Big(\sum_k \log(\alpha_k) - k\log(\gamma)\Big)\right] \tag{25}$$

$$\implies \beta|\vec{\theta}, \vec{\alpha}, \vec{y} \sim Gamma\Big(k, \sum_k \log(\alpha_k) - k\log(\gamma)\Big) \tag{26}$$

7.3 Jeffrey's prior on β

$$p(\beta) \propto \prod_k \beta\alpha_k^{-(\beta+1)} \tag{27}$$

$$p(\beta) \propto \beta^k(\prod_k \alpha_k)^{-(\beta+1)} \tag{28}$$

$$\log\big(p(\beta)\big) \propto k\log(\beta) - \beta\log(\prod_k \alpha_k) - \log(\prod_k \alpha_k) \tag{29}$$

$$\frac{\partial}{\partial\beta}\log\big(p(\beta)\big) \propto \frac{k}{\beta} - \log(\prod_k \alpha_k) \tag{30}$$

$$\frac{\partial^2}{\partial\beta^2}\log\big(p(\beta)\big) \propto \frac{-k}{\beta^2} \tag{31}$$

$$-E\Big[\frac{\partial^2}{\partial\beta^2}\log\big(p(\beta)\big)\Big] \propto \frac{k}{\beta^2} \tag{32}$$

$$\pi(\beta) \propto \frac{1}{\beta} \tag{33}$$