

Set-up

The Federalist Papers were a collection of 85 articles written in the late 18th century by Alexander Hamilton, James Madison and John Jay. They were written to influence voters to ratify the United States Constitution.

Unlike good group projects, not all articles were written collaboratively. That is, some of the 85 articles was authored by either Hamilton, Madison, or Jay alone. The authorship of 73 of the 85 articles are mostly agreed upon, but there is still dispute about who wrote the remaining 12 articles. As it turns out, the scholarly debate is only about Madison or Hamilton for the disputed articles (scholars agree that Jay did not write them). There have been many (statistical) analyses trying to identify the authorship. In this case study, you will carry out an adjacent analysis that doesn't quite answer the question of who authored which disputed paper, but does set you up well for answering that question in a more complex model should you choose.

In particular, we will examine the frequencies of word counts in the articles.

Data

For each article, we have data on the count of each unique word (i.e. frequencies) in that article, along with the total number of words in the article, who has been attributed authorship to that article, and whether or not that attribution is disputed. The dataset can be found on the website under `federalist_papers.csv`. The variables are:

- **Article:** name of the Federalist article
- **Total:** total number of words in the particular **Article**
- **word:** the particular word that is counted
- **N:** number of times the **word** is found in the particular **Article**
- **Authorship:** author of paper (Hamilton, Jay, or Madison)
- **Disputed:** whether or not the author listed in **Authorship** is disputed

Task

The idea behind this analysis is the theory that one author may use certain words more frequently than another author. However, due to the topic, a particular article might use a certain word (e.g. “liberty”) a lot that won’t show up in other articles, which doesn’t make it useful for discerning authorship in other articles. Therefore, we will focus on “filler” words that are almost guaranteed to show up in all the articles. Examples of filler words include “the”, “of”, and “in”. See the end of this document for a list.

Carry out a Bayesian analysis that attempts to answer the following question: Which (if any) filler words might be useful for helping solve the authorship problem? Which (if any) filler words are not helpful? Given the following sampling model, this analysis is quite

constrained. However, you will have to implement everything by hand in R. The first step is to make sure you understand the sampling model below, identify the parameters of interest, and then think about ways to use them to answer the questions above.

Sampling model set-up

In this write-up, we will focus on one particular author and a particular filler word. Let Y_i be number of times we see the word in Article i , for $i = 1, \dots, n$ where n is the total number of articles written by that particular author. Because Y_i is a count and we'd probably like to allow for differences between mean and variance, we will use the Negative Binomial sampling model (finally!):

$$Y_i | \alpha, p_i \stackrel{\text{indep}}{\sim} \text{NegBinom}(\alpha, p_i) \quad i = 1, \dots, n$$

We will re-parameterize/re-represent p in terms of another parameter β (you'll see why later on): define $p_i = \frac{\beta}{\beta + n_i/1000}$ where n_i is the total number of words in article i . Note that once you have β , you get p_i "for free" (deterministic given β). So really, the unknown parameters here are α and β . But I use p_i notation above to match the way we've seen the Negative Binomial PMF before.

Use this sampling model to try and answer the research question using 6 filler words of your choice. *You will almost surely have to play around with/stare at the Negative Binomial a bit.*

Some questions to consider

1. What is your model? What assumptions are you making? Are the assumptions justified? Are you using all the articles? In this case study, think about the different words and different authors. If you decide to go hierarchical, explain why you think you need to.
2. What are the parameter(s) in your model(s)? What prior distribution(s) do you use?
3. How will you use the parameters and their posteriors to answer the research question? What "derived quantities" (i.e. functions of the parameters) are of interest? Which filler words will you focus on? *Please don't examine all of them!*
4. What is your computation scheme? If you use Metropolis: what are your proposal distribution(s)? What is/are the (log) acceptance ratio(s) equal to?
5. What results would you like to report, and how will you report them effectively? In your results, you should include at least one thoughtful, well-designed, and ideally aesthetic plot.
6. What is your conclusion? Are there any limitations/unreasonable assumptions you made? How might you envision using your model to do analysis to predict the authorship in the disputed articles?

Requirements/General rubric

- Clearly state your model with correct and appropriate notation (address questions 1 and 2 above). (5 pts)
- Clearly describe how you will use the model to address the research question. This includes providing some derivations or explanation for the reader (addresses question 3 above). (5 pts)
- Clearly state how you will implement your model (address question 4 above). You should provide enough detail such that someone else could implement your analysis. (5 pts)
- Code that correctly implements your MCMC. Don't forget MCMC diagnostics to make sure your chain has converged. If you run Metropolis, also report the acceptance probabilities. **I almost guarantee you that you will have some funky traceplots and some errors. These are learning moments; this case study is very educational!** (10 pts)
- Provide results *with interpretation* (address question 5 above). (7 pts)
- Give a conclusion that answers the research question and expands on the findings (address question 6 above). (3 pts)
- Your code should be clean. That means only code necessary for implementing your sampler and obtaining your results should be included in your final document. Your code should also be as reproducible as possible. *While not necessary, after you feel confident in your code, you may want to consider creating a function in R to keep your code tidy.* (5 pts)
- Your report should be well-written. This means correct spelling, complete sentences, a nice flow to the report, etc. You should only provide plots/results that you actually intend to discuss in words. Think about this as a *professional* report. I estimate that the final document to be somewhere between 5-10 pages, depending on figure sizes and length of code, but feel free to write more/less if you think that is what the report needs. (5pts)

Filler words

- a
- also
- an
- and
- any
- be
- by
- can
- from
- his
- in
- is
- it
- may
- of
- on
- the
- that
- there
- to
- this
- upon
- which