Online appendix for Huang, Perry, and Spirling (2019)

Leslie Huang (corresponding author)
Patrick O. Perry
Arthur Spirling

A Relationship Between our Approach and multinominal naive Bayes

In the conventional multinomial naive Bayes arrangement used for text, interest focuses on placing a document (for us, a speech, i) into one of two classes (for us, having been spoken by a member of parliament, s or t) using information about the terms used (for us, the counts x_{iv} of the tokens in speech i). We would express this as

$$\Pr(t|i) \propto \Pr(t) \prod_{v \in V_c} \Pr(v|t)^{x_{iv}}, \tag{2}$$

where, on the right hand side, $\Pr(t)$ is our prior that speech comes from member t, and the likelihood is the product—over all terms making up speech i—of the probability of observing that word from member t. Recall that the exponent x_{iv} is the number of times a particular v occurs in speech i, and note that if that frequency is zero (i.e. that element of the vocabulary does not occur in the speech), simply multiplies the product by one. The "naive" independence assumption allows us to compose the product this way, while each $\Pr(v|t)$ is assumed to be a multinomial distribution. To reiterate then, this is the "multinomial naive Bayes" approach.

Keeping within the standard set up, consider the ratio of the probability that t spoke speech i relative to s (our only other speaker, initially). That is then the ratio

$$\frac{\Pr(t|i)}{\Pr(s|i)} \propto \frac{\Pr(t)}{\Pr(s)} \frac{\prod_{v \in V_c} \Pr(v|t)^{x_{iv}}}{\prod_{v \in V_c} \Pr(v|s)^{x_{iv}}} \tag{3}$$

and appears in, for example, Hand and Yu (2001). If one assumes equal priors on the members (which we do), this expression is

$$\frac{\Pr(t|i)}{\Pr(s|i)} \propto \frac{\prod_{v \in V_c} \Pr(v|t)^{x_{iv}}}{\prod_{v \in V_c} \Pr(v|s)^{x_{iv}}}. \tag{4}$$

Taking logs, we obtain

$$\log\left(\frac{\Pr(t|i)}{\Pr(s|i)}\right) \propto \sum_{v \in V_c} x_{iv} \log(\Pr(v|t)) - \sum_{v \in V_c} x_{iv} \log(\Pr(v|s)). \tag{5}$$

Collecting terms, and writing $\log(\Pr(v|s))$ as η_{sv}, we have

$$\log\left(\frac{\Pr(t|i)}{\Pr(s|i)}\right) \propto \sum_{v \in V_c} x_{iv}(\eta_{tv} - \eta_{sv}). \tag{6}$$

Recall that our core distinction measure as given in Equation (1) was

$$\frac{1}{n_i} \sum_{v \in V_c} x_{iv}(\eta_{tv} - \eta_{sv}).$$

The resemblance between this and Equation (6) is obvious. Indeed, the only difference is that we divide by the length of the speech, but that does not affect the maximum *a posteriori* class (i.e. speaker) designation for a given speech.

In the traditional arrangement, prediction into multiple classes can be done in various ways. This includes "one-vs-rest" techniques, which make the problem binary, in the sense that one is placing a document into one of two classes: the class of interest (for us, an MP) or an "other" category, composed of all other possible classes (for us, all other MPs). We follow a similar strategy in principle, but in practice, we take average pairwise distances from a given speaker to all others to obtain our estimates (which is not usually done in the "one-vs-rest" arrangement).

B A Reexamination of the *Federalist Papers* Authorship Problem

Recall that in the original Mosteller and Wallace (1963) set-up, the goal was to determine whether Hamilton or Madison wrote the twelve mystery papers. Now we re-examine that question, except that we also consider John Jay as a possible author. We can do this because unlike in Mosteller and Wallace (1963), our method is not restricted to a single pairwise comparison of candidate producers of the text. To keep the problem otherwise similar, we use the same vocabulary as in the earlier presentation (the seventy function words).

Our results are pictured in Figure 5. There each bar sums to one as a total probability of authorship for each disputed Federalist Paper: the light gray part corresponds the probability that Madison was the author, dark gray corresponds to Hamilton and the mid-gray part to Jay. Reassuringly from a validation perspective, we get results essentially identical to Mosteller and Wallace (1963). We mean this in the sense that Madison is by a large margin the most likely author in all cases, with the exception of *Federalist* 55 where—as in the original presentation—the evidence is more equivocal. But again, note that our model is more flexible: we can include Jay (or indeed any other candidates) as a potential author "automatically" in our comparison. It so happens that the estimated probability for Jay is near zero, but that is as we would expect given our priors in this case. In short, our more general model works when taken to this canonical application.

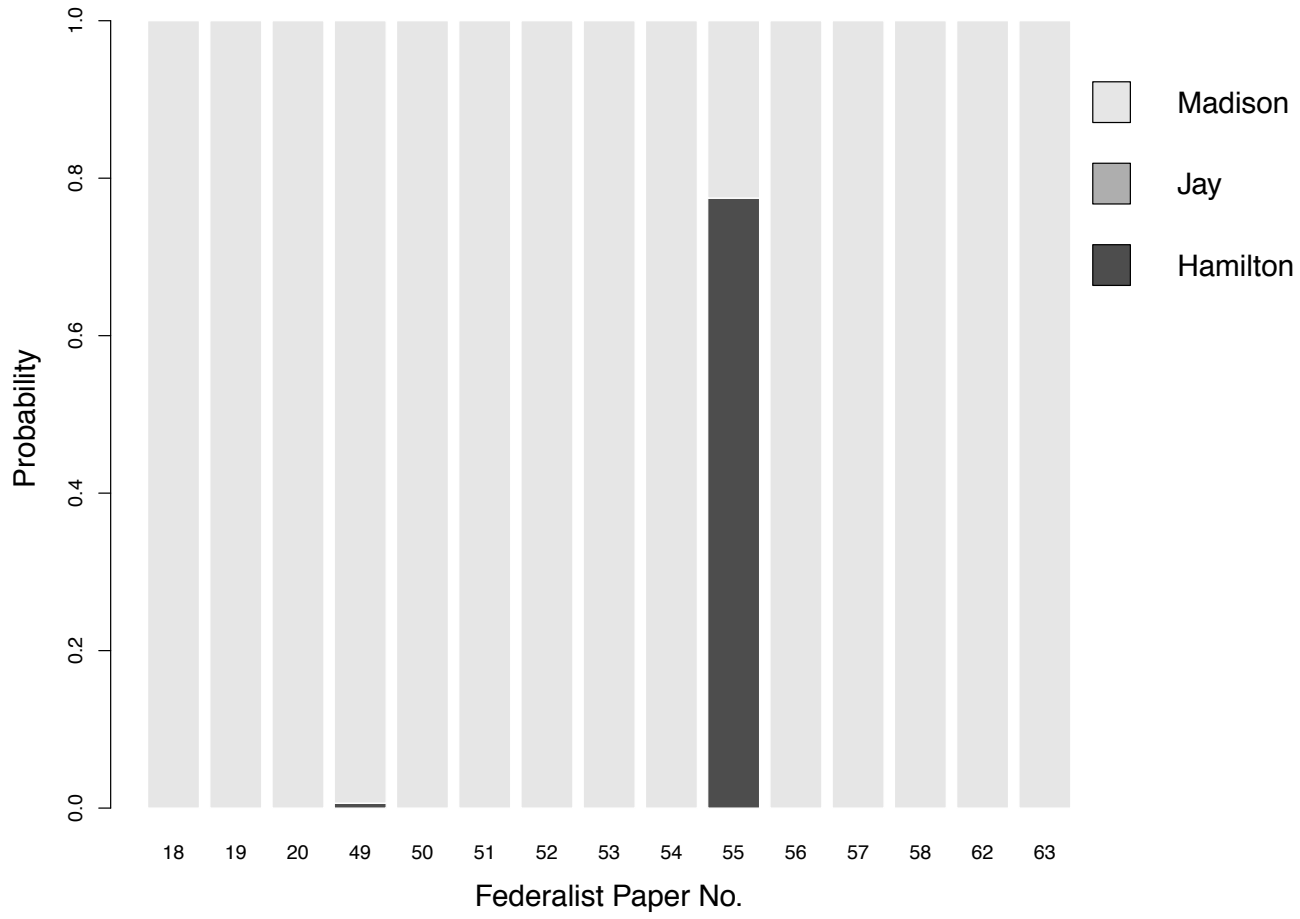

Figure 5: Reexamining the "mystery" papers using a more flexible, three-way model. We find results similar to those in the original Mosteller and Wallace (1963) paper, with Madison almost certainly the author in all cases, bar one. In addition, we have very small but non-zero estimates for Jay in each case.

Table 7: Most influential terms, 1995–1996.

the	of	to	and	that
i	we	in	is	a
it	hon	for	not	my
be	have	bill	are	will
they	was	he	as	labour
people	our	friend	on	local
would	european	has	education	which
committee	council	right	service	government
by	do	defence	schools	or
who	there	about	been	member

Table 8: Most influential terms, 1998–1999.

the	of	to	that	and
i	in	a	we	is
it	not	for	london	have
be	people	was	health	are
they	hon	will	women	my
on	children	bill	local	government
he	as	house	industry	who
by	scottish	our	police	with
their	wales	friend	would	right
but	about	there	has	committee

C Influential Tokens

A nice feature of our approach is that we can inspect the 'most influential' terms, both within certain sessions and over the data as a whole. Term influence is estimated separately for each speaker–term pair in each session, and is calculated for term v and speaker s by multiplying the mean (over speaker s's speeches) term frequency count $\frac{1}{|I_s|}\sum_{i\in I_s} x_{iv}$ with the mean-centered term frequency rate η_{sv} (the latter term is $\eta_{sv} - \frac{1}{|S|}\sum_{s\in S}\eta_{sv}$). Thus, the difference between the speaker's rate of use of word v and the mean rate of use of v is weighted by how many times the speaker actually used the word v. One consequence of this arrangement is that stop words tend to be 'influential' because they are so commonly used, boosting the mean frequency rate across all speakers.

In Table 7 and Table 8 we list the 'top 50' most influential words, in terms of mean term influence, from the pre- and post-Blair election landslide (the time period we used to validate above). For each term, we compute the mean of the absolute value of the term's influence over all speakers in each session. (We use absolute values because it is possible for an influence value to be negative for a term v since it is comprised in part of the difference between the speaker's rate of use of v and the mean rate of use of v.)

As is clear and expected, we indeed see a large number of stop words in both tables: 'the', 'of', 'to' and so on. More compellingly though, we also see words that make sense as impor-

tant or distinctive during this time. For example, Tory backbenchers made 'european' influential, which accords to the protracted and fractious discussion of the various EU treaties at this time. Meanwhile, 'defence'—a common right-wing concern—turns up as an influential word in 1995, but not in 1998 (among Labour MPs). After Blair's election, it is words like 'london', 'scottish' and 'wales' that are most influential. Given that the Labour government was discussing and passing legislation to create new governing institutions in the capital, Scotland and Wales, this makes some sense. We also typical left-wing concerns pertaining to 'women', 'children' and 'health' at this time.

Another nice feature of our model is that we can examine "influential" tokens to get some sense of how their use changes over time, and how that use affects distinctiveness. In Figure 6 we do exactly that, with a lowess imposed on the data for visualization purposes. In particular, we plot the 'average influence'—defined as the mean of a term's influence over all speakers within a session—of the various tokens for every session of parliament under study. Starting at the top left, we see that 'defence' surged in influence during the Cold War, and then fell away after it ended, around 1990. Moving right, we see 'european' coming to a peak during the Tory rebel years of the early 1990s, during the Maastricht negotiations. Moving to the bottom left, 'Scottish' peaks in the late 1970s and late 1990s—the periods in which some devolved rule was offered to Scotland (and referendums took place on the same). Finally, 'health' peaks in the late 1990s/early 2000s as the Blair government abolished the internal market, added large real term budget increases, but allowed independent foundation status for hospitals. This latter move was very unpopular with Labour rebels, and in 2003 they almost defeated the government's bill in the Commons that dealt with this policy change.

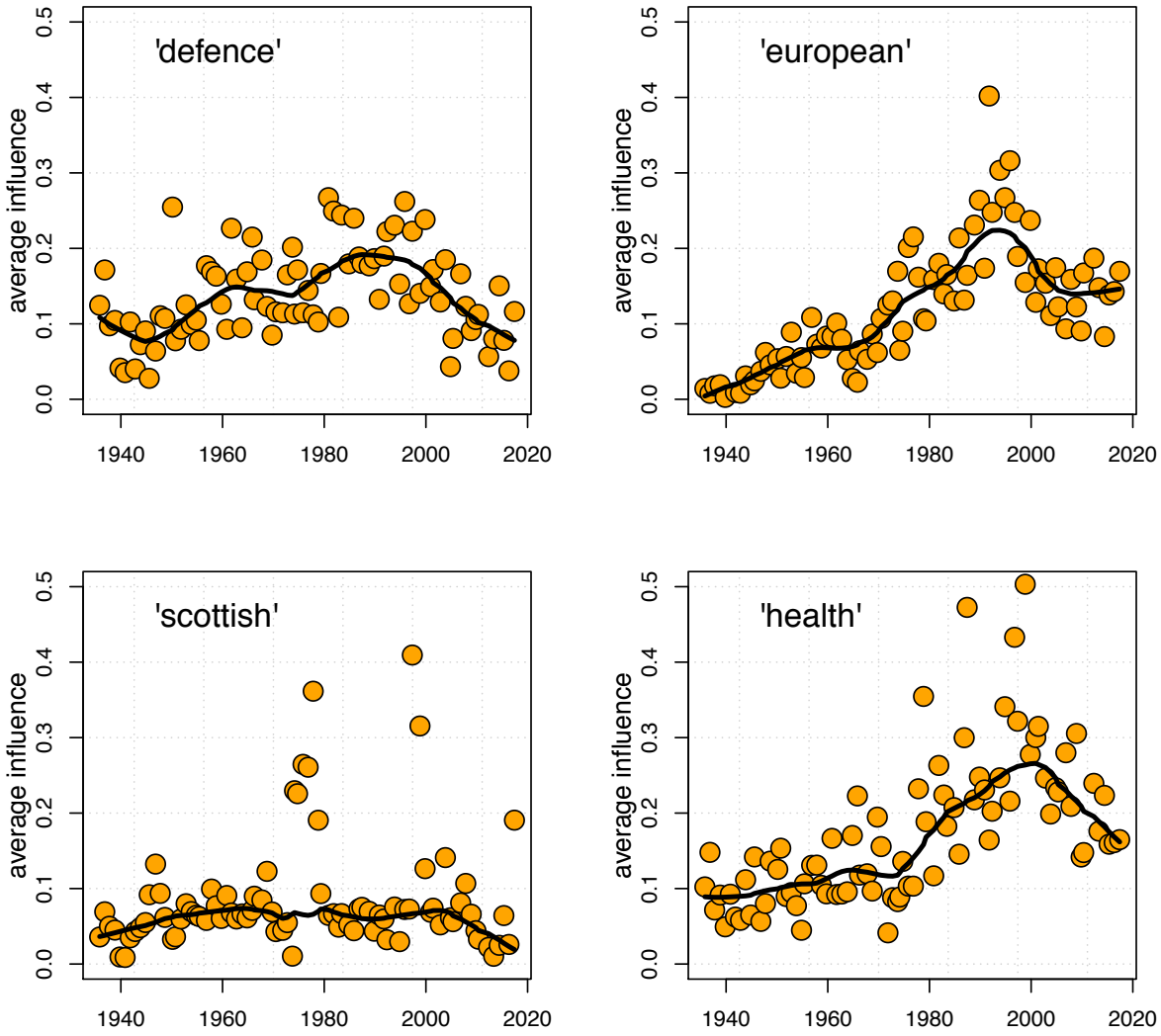

Figure 6: Average influence of some key terms over time. The non-stop word tokens from the sessions of 1995 and 1998 accord with our priors, in terms of their behavior over the entirety of the data. For example, 'defence' ceases to be (as) influential once the Cold War ends.

Online Appendix References

Hand, David J and Keming Yu. 2001. "Idiot's Bayesnot so stupid after all?" *International statistical review* 69(3):385–398.