Measuring Group Differences in High-Dimensions

An Empirical Exercise Following Gentzkow et al. 2019

1 Setting

The paper by Gentzkow et al. (2019) aims to estimate the degree of partisanship in the US Congress over the last century. Specifically, they use US speech data to measure to what extent politicians from the two parties (Democratic and Republican) choose different words when giving a speech. In their model, giving a speech basically means drawing a series of words from a probability distribution over a large choice set of different options. If these probability distributions are the same (i.e., if the probability of choosing any word is the same for a Republican or a Democrat), we would have no partisanship. To have a measure of partisanship with an intuitive interpretation, they define it as the probability of an observer correctly guessing a politician's party affiliation after hearing a single phrase from a speech. This measure ranges from 0.5 (no partisanship, 50% chance of guessing correctly) to 1 (maximum partisanship).

The empirical challenge stems from the high dimensionality of the choice set: because the number of words drawn (the length of observed speeches) is small relative to the size of the choice set (all possible words that could be used), it is very unlikely that the observed occurrence of a word is the same for both parties, even if there is no partisanship. To illustrate this, think of the choice set consisting of 1000 words and speeches being 100 words long. Even if the probability of drawing any word is uniform over the choice set for both parties (i.e., a Republican and a Democrat would both choose each word with a probability of 0.1%), we expect the vector of 100 words chosen by each party to be very different just due to randomness. Therefore, we need to find a way to remove as much of this high-dimensionality bias from our measure as possible.

2 Data Simulation

Since Gentzkow et al. (2019) use speech data from over 100 years, the original dataset is very large and estimations are difficult to run on a standard computer. For that reason, we have to simulate data ourselves. For that, we make use of a multinomial logit choice model, which is also the theoretical underpinning of the original analysis. Specifically, we derive choice probabilities from utilities

$$u_{jit} = \pi_t \cdot \alpha_{j,t} \cdot P_i + \beta_{1,t} \cdot A_i + \beta_{2,t} \cdot O_i + \epsilon_{it}$$

with P_i as a politician's party affiliation, A_i as their age, and O_i as their origin. We draw the $\alpha_{j,t}$ from a constant Gamma-distribution and let π_t increase in the last 30 periods to simulate increasing true partisanship.

To replicate the key aspects of the paper, our simulated data has the following characteristics:

- 100 politicians with different demographics make speeches (draw words) over 100 periods
- Verbosity (the number of words drawn) rises over time from 100 to 250
- The choice set (number of words to choose from) increases over time from 1000 to 1200
- Partisanship increases only in the last 30 periods

We have to limit the number of draws and size of the choice set to ensure that computations do not take too long. At the same time, we need the data to still be high-dimensional so that we can illustrate the core estimation challenge that Gentzkow et al. (2019) address. Because in our simulated data the ratio of draws to choices is much higher than in the original data (for details, see the R Notebook), the high-dimensional problem is smaller but still very much present, as we discuss below.

3 Estimation

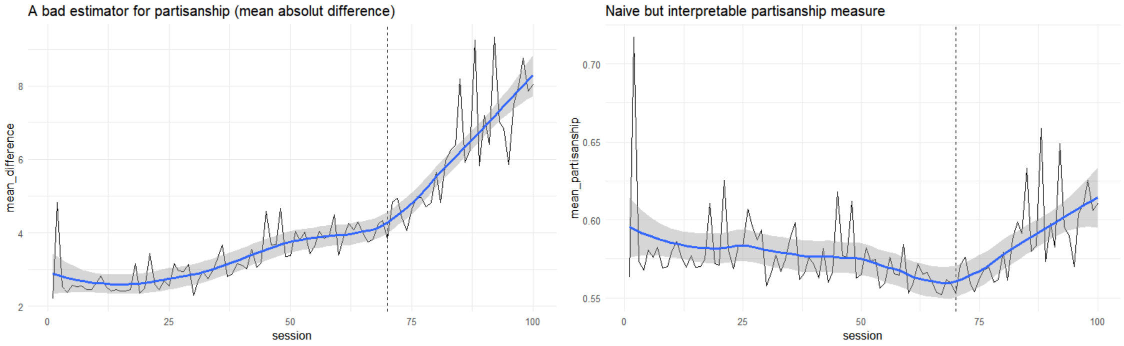
3.1 Naive estimators

Without thinking about the high-dimensionality problem described above, a first naive estimator that may come to mind would be to measure the mean absolute difference in the $J \times 1$ vector of draws per phrase for each party. Figure 1a shows the behavior of this estimator using our simulated data. Unsurprisingly, it is directly affected by changes in verbosity, and consequently indicates an increase in partisanship even before true partisanship rises, which simply stems from higher verbosity.

To address the direct influence of verbosity and to give an intuitive measure of partisanship p, Gentzkow et al. (2019) propose using the probability of guessing party affiliation after hearing a phrase j. Not taking into account non-party speaker characteristics, it is calculated as the sum over all j of

$$p_j = \frac{1}{2} \cdot Pr(j|P_i = 1) \cdot Pr(P_i = 1|j) + \frac{1}{2} \cdot Pr(j|P_i = 0) \cdot Pr(P_i = 0|j)$$

i.e., as the probability that a member of party 1 uses phrase j multiplied by the probability that they are actually a member of party 1 given that j was being said, plus the same for a member of party 0. Not accounting for the high-dimensionality problem, we could just use the empirical analogues of the probabilities.



(a) Mean difference

(b) Interpretable measure (plug-in estimator)

Figure 1: Naive estimators of partisanship

For example, we would replace $Pr(j|P_i = 1)$ by the number of times phrase j was used by members of party 1 in period t divided by their total speech length. Figure 1b shows the results of such an estimator. We identify two problems with this estimator:

- There is substantial bias (estimates above 0.5 even when there is no true partisanship)
- The bias varies over time (we do not pick up the true trend of partisanship)

Both of these problems stem from the high-dimensionality problem described in the Introduction. The bias varies over time because the size of and number of draws from the choice set itself vary. Intuitively, if the number of draws increased to ∞ for a constant choice set, we would eliminate the high-dimensionality problem (empirical frequencies would match true choice probabilities). On the contrary, a larger choice set weakens the link between the observed choice frequencies and theoretical choice probabilities.

3.2 Leave-one-out

One estimator that solves this problem is the leave-one-out estimator. As shown in Gentzkow et al. (2019), the bias described above can be decomposed into two parts: a) the fundamental problem of not being able to reliably estimate choice probabilities from empirical estimates, plus b) the correlation between the estimation error for $Pr(j|P_i = 1)$ and $Pr(P_i = 1|j)$. The intuition becomes clear from an extreme example: if we observe some phrase j only once in party 1 and never by party 0, we would classify it as a very partisan phrase ($\widehat{p}_j = 1$ because $\widehat{Pr}(j|P_i = 1) = 1$ and $\widehat{Pr}(P_i = 1|j) = 1$). However, the estimates are driven by a single observation. If we removed from the sample the politician using phrase j, we would estimate $\widehat{Pr}(j|P_i = 1) = 0$ and $\widehat{Pr}(P_i = 1|j) = 0.5$.

To remove this correlation in estimates for the two probabilities, we can simply estimate them from different samples. In the leave-one-out estimation, we estimate $\widehat{Pr}(j|P_i = 1)$ using a single observation and $\widehat{Pr}(P =$

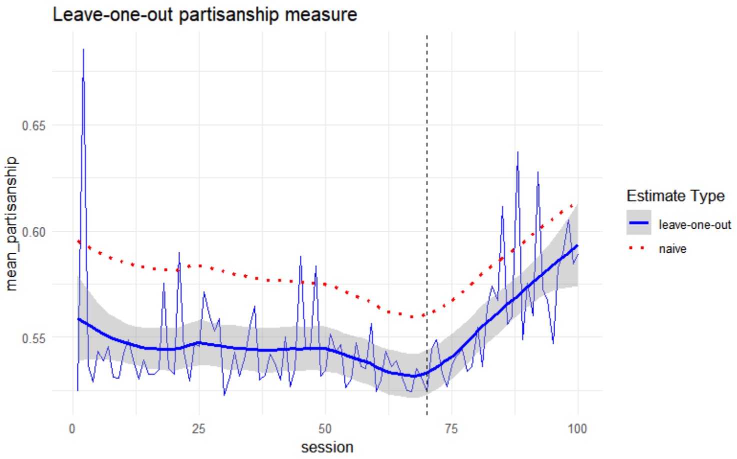


Figure 2: Leave-one-out estimator

$1|j$) from the rest of the sample. Figure 2 shows the outcome of this procedure and compares it to the naive estimate from Figure 1b. We see that it has some success in reducing the bias (recall that the first bias component is still present), so that the trend in the true partisanship is more visible. However, the estimate performs worse than in the original paper, likely due to our limited data.

3.3 Penalized estimator

Because the leave-one-out estimator does not allow the incorporation of control variables, Gentzkow et al. (2019) also propose a penalized estimator for partisanship. The basic idea is that the high-dimensionality problem causes an overestimation of partisanship (i.e., an overestimation of the parameters α_j). Therefore, this new estimator tries to find parameter values α and β in the utility specification such that the choice probabilities are close to the observed choice frequencies, while adding a penalty for larger values of α. A weakness of this estimator is its computational intensity, since we need to do $J \times T$ optimizations to find the three parameters for all phrases in all periods. Therefore, we illustrate the estimation on a much smaller dataset (as such, this estimate cannot be compared to those reported in the figures below) in the R Notebook, and do not implement the selection of the optimal penalty.

3.4 Our approach

From a Bayesian perspective, the penalized estimator simply adds a prior belief about the distribution of α to derive its posterior estimate. We can follow the same logic by constructing an estimator based on the beta distribution, which is the conjugate prior probability distribution of the binomial distribution. Essentially, we include a prior k to the plug-in estimator described above by adding k to all empirical choice frequencies.

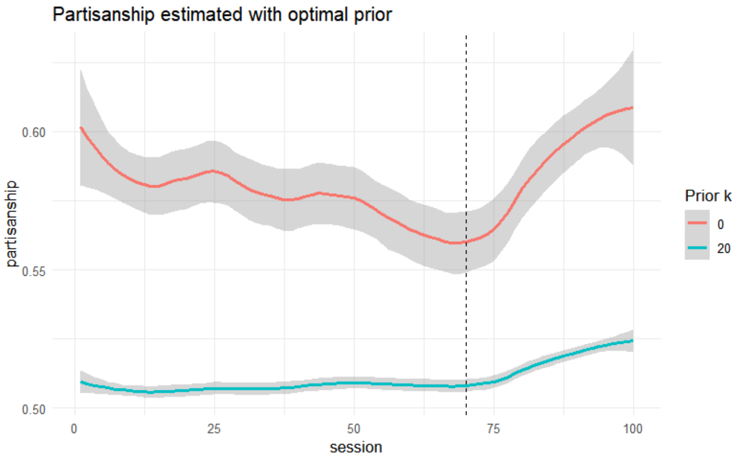


Figure 3: Own corrected estimator

To get the intuition, recall the extreme example of observing a phrase only once by party 1 and never by party 0. Without any prior, we estimate maximum partisanship. By adding $k = 10$, we treat this phrase as if it had been said 11 times by party 1 and 10 times by party 0, which clearly leads to much lower partisanship.

Of course, selecting k is crucial. We suggest two simple approaches. The first and preferred approach uses 5-fold cross validation. It splits the original data into 5 subsamples, estimates annual partisanship in each of these using different k, and selects the k that leads to an estimator with the lowest variance over the different subsamples. The second approach comes from the idea that a good estimator estimates partisanship close to 0.5 in simulated data without true partisanship ($\alpha = 0$), but sensitively detects increasing partisanship in data with $\alpha > 0$. After translating these desired properties into a loss function, we simulate the two types of data for different parameters specifying draws and the choice set, then select the k that performs best on average. However, as long as the loss function is not grounded in theory, the chosen k performs worse than the k selected using 5-fold cross validation.

As shown in Figure 3, the partisanship estimates for the optimal k are much better than the naive estimator in detecting changes in true partisanship. Additionally, it saves a lot of computation time compared to the penalized estimator in Gentzkow et al. (2019). However, as it cannot estimate the taste parameters α and β, it sheds less light onto the question of *why* partisanship changes. Only if demographics stay roughly constant (which is the case in our simulated data, and can also be confirmed for the real-world data), we know that changes in party differences are actually linked to the parameter α, i.e., political partisanship.