Research Question 2 (Bayesian Hierarchical Modeling)

Does there exist a discrepancy between a person's self-reported political leaning and their true political leaning as defined by their opinions on political issues?

Alan Jian, Yannie Li, Matilda Ju, Shreya Chowdhury

December 2021

j	Question	$y_{ij} = 0$	$y_{ij} = 1$
1	Do we spend too little protecting the environment?	No	Yes
2	Do we spend too little on national defense?	No	Yes
3	Do we spend too little on welfare?	No	Yes
4	Do we spend too little on social security?	No	Yes
5	Would you favor a law requiring a permit before a person could buy a gun?	Oppose	Favor
6	Should a pregnant woman be able to obtain a legal abortion for any reason?	No	Yes
7	Do you think sexual relations between two adults of the same sex is wrong at all?	No	Yes
8	On the average (Blacks/African-Americans) have worse jobs, income, and housing than white people. Do you think these differences are mainly due to discrimination?	No	Yes

Table 1: Binarized Political Questions

1 Methods

To find a solution to this research question, we first began with a simple idea; an individual's self-reported leaning was a noisy estimate of his/her true political leaning. Given $\{1, ..., m\}$ individuals with self-reported political leanings $\alpha_i \in \{-3, ..., 3\}$ corresponding to the ith individual, we modeled individuals' true political leanings with the expression $\vec{\alpha} + \epsilon 1$, where $\vec{\alpha}, 1 \in \mathbb{R}^m$, and ϵ is some random variable that represents a perception gap between actual and perceived political leaning. Since individuals' true political leanings seem to be reasonable predictors of how they might respond to $\{1, ..., n\}$ political, ethical, and philosophical questions (see Table 1 for mapping), we used individuals' observed binary responses to give us information about their (hidden) true political leaning. As a result, our problem required a blend of inference and regression in which each ith individual's observed response to each question j could be modeled by $y_{ij} \sim Bernoulli(\sigma(\beta_j(\alpha_i + \epsilon)))$, where each jth element of n-vector $\vec{\beta}$ represents the weight $\alpha_i + \epsilon$ has on predicting the value y_{ij} (as shown in Figure 1). Given a matrix of observed answers $Y \in \mathbb{R}^{m*n}$ from m individuals across n questions, we couched the problem as the following *maximum a*

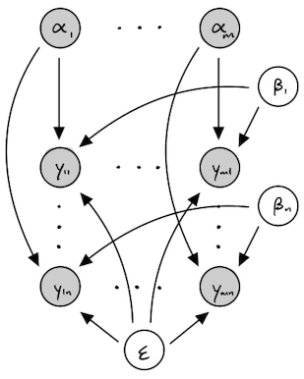

Figure 1: Graphical Model

posteriori estimate:

$$\max_{\epsilon,\vec{\beta}} \mathbb{P}(Y|\vec{\alpha},\vec{\beta},\epsilon)\mathbb{P}(\vec{\beta},\epsilon)$$

where

$$\mathbb{P}(\epsilon) \sim Normal(0, 0.0001)$$

and $\vec{\beta}$ starts with a flat prior. The prior on ϵ was chosen to encourage model convergence (even small amounts of variance caused the model to diverge), and to reflect the fact that we expect an individual's self-reported leaning to be an unbiased estimator of true leaning. After updating priors on ϵ and $\vec{\beta}$, we analyzed the distribution of ϵ to better understand the perception gap between an individual's response and their actual political leaning, and visualize the elements of $\vec{\beta}$ as a sanity check to make sure its values match our expectations.

2 Results

After running our model, our results suggest that there is a real, consistent perception gap between an individual's real and perceived political leaning. When viewing the distribution of samples from the posterior distribution of ϵ, we can see that over 97% of the distribution is greater than or equal to zero, and that an individual's perceived political leaning is off by an average of -0.07 from their true political leaning. Assuming that our data are representative of the U.S population, and that each individual's value of epsilon is independent and identically distributed, this data indicates that an overwhelming majority of Americans misrepresent (either intentionally or unintentionally) their political leaning as being more liberal than they truly are. This has major implications for government officials, and how they choose to best serve their constituents.

Furthermore, upon investigating the elements of $\vec{\beta}$ as shown in Figure 2, we see that the ethical and political questions that we used to build our model's intuition about each individual's true political leaning were informative in the right ways. Being liberal was positively associated with beliefs that the national government was not spending enough on protecting the environment or expanding welfare, while being conservative was associated with the belief that felt the government was not spending enough on national defense. Additionally, we can see that being conservative also made you more likely to oppose homosexuality and oppose abortion rights, beliefs that are consistent with politics today. The only variable that was (correctly) identified to be evenly supported by both liberals and conservatives was national spending on social security. In this case, both liberals and conservatives felt that national spending on social security is not enough, with the value of $\beta_{socialsecurity}$ hovering very close to zero.

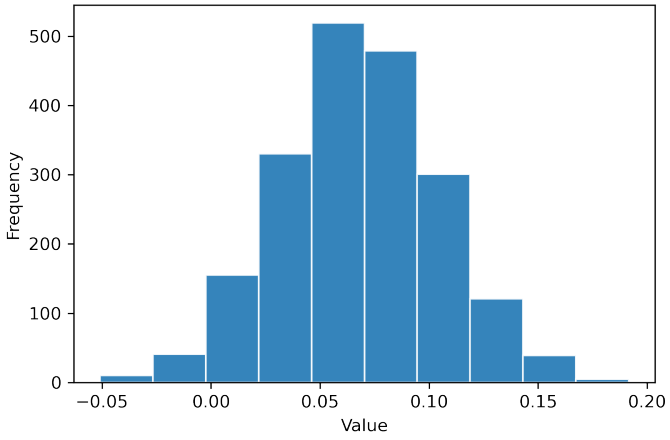

Figure 2: Posterior Distribution of ϵ

3 Discussion

For all this models benefits, it also came with some major limitations. Firstly, due to the setup of our model, it easily diverged when initializing ϵ with a Gaussian who's $\sigma > 0.05$. When this occurs, the posterior of

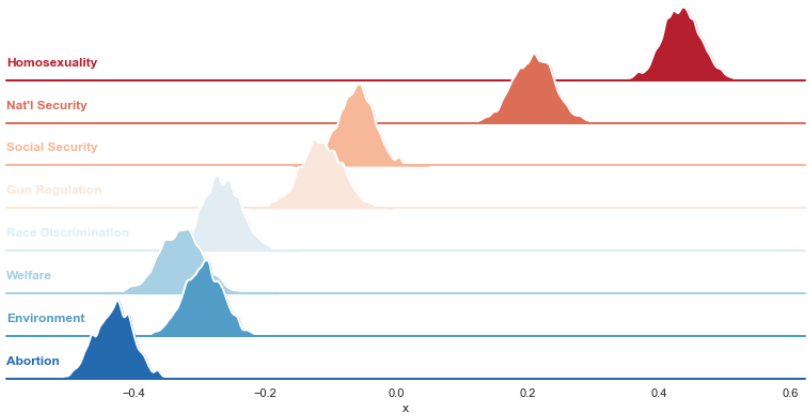

Figure 3: Posterior Distribution of the elements of β

ϵ becomes split, literally diverging towards two different ends of the number line. Based on our tuning, it seems like this problem arises due to lack of informative features. Even though this collection of survey questions helped the model get a better idea of the relationship between true political leaning and opinion on these issues, the model clearly needed more questions to improve accuracy, or more respondents to improve confidence (both of these are equivalent to expanding Y in our graphical model), something that we tried to solve via imputation. Rather than drop all data entries that were not asked every single question of interest, we tried using PyMC3's automatic value imputation via posterior predictive distribution to avoid throwing away a treasure trove of information. Given that our data was missing completely at random (MCAR) via study design, we knew that this imputation would not bias our results. Though this idea sounded great in theory, it proved intractable in practice. The problem was the sheer scale of value imputation required in order for this to work. Internally, when PyMC3 encounters non-response, it treated it like another latent variable with a prior distribution that needed to be updated along with every other distribution. In a dataset with a total size of over 28000, in which a majority of individuals were not asked every question of interest, this meant adding well over 30000 random variables to the calculation. Given that PyMC3 does not benefit from hardware accelerators such as GPUs or TPUs, this computation cost was simply insurmountable for our group.

4 Conclusions

In our study, we found concrete evidence that U.S citizens are very likely to perceive themselves as more liberal than they truly are, as inferred based on their opinions on political issues. Furthermore, we also have evidence to suggest that an individual's perceived political leaning is not an unbiased estimator of their true political leaning. This finding has major implications both for individuals who hold political positions, as well citizens. For politicians, this evidence may be useful in shifting campaign strategy. Given that individuals see themselves as more liberal-leaning than they actually are, politicians may increasingly turn to performative activism as a way to pull in more voters. For citizens, it calls into question whether our sources of entertainment and information are distracting or pulling us away from the issues and matter most to us. Still, more needs to be done to understand why this perception gap exists, and what effects it has on our already fragile democracy.