Project Proposal

Manik Goyal (mg4106) | Mark Dijkstra (md3507)

October 28, 2020

1 Description

Goal The goal of this project is to translate the open source library for the presidential election forecasting project done by Andrew Gelman and Merlin Heidemanns[2, 3] into Pyro and then empirically summarize the various methodologies for evaluating the model. This summary of evaluation metrics is worthwhile because being able to accurately predict the outcome of the elections in this upcoming 2020 election given 2016's blunder[4] (when Hillary seemed to be the winner via the predictions) will help us understand where we went wrong and that the model can actually predict new out of sample data.

Thus, we want to be able to criticize our model based on the predictions our model makes on the upcoming 2020 election and then be able to describe why our results follow the true unforeseen outcome or not. As a result, this will allow us to determine at a high level whether the assumptions taken by the model were justified, or whether other assumptions needed to be taken in place. For example, there is the decision to lessen the level of the recession caused by Covid and we would like to see if this was a justifiable assumption.[5]

In addition, another goal of this project to grasp the capabilities of Pyro and then be able to contribute to Pyro's open source library if any drawbacks or shortcuts are to be found during our code translation endeavor from Stan and R.

Data To solve this task of forecasting presidential elections, we observe the state and national polls that are conducted by different pollster/survey houses till the election day. These polls are used to represent an estimate of that day's support for Democrats/Republicans. To do this only the number of respondents who report a preference in favor of either party are considered. [6] Further to form the Abramowitz's Time-for-Change fundamentals model[1] we use the annualized growth rate of GDP in the second quarter of the election year and the incumbent president's net approval rating. The Economist paper[3] states the usage of real disposable income, non-farm payrolls and the stock market to better grasp the voter behavior; which although appears to not be a part of the GitHub repository publicly shared.

2 Proposal

Modeling The model used is a Dynamic Bayesian model which is a bayesian network that relates variables to each other over adjacent time steps. The model has a backward component to derive the latent variables (both at the state level and national level) and then a forward component to predict how these values will change with time till election.[7]

The backward portion is categorized by: $y_i^d \sim Binom(n_i, \pi_{i,j}^d)$ where y_i^d is the number of respondents in favor of voting for democratic party candidate, n_i is the total number of respondents reporting a political party preference (dem / rep), and $\pi_{i,j}^d$ is the share of voters who would vote for democratic candidate on day j. Thus the prior is $\pi^d_{i,j} = \text{logit}^{-1}(\beta_{i,j} + \delta_j)$ where, $\beta_{i,j}$ is the state-level effect and δ_j is the nation-level effect with hierarchical priors given on those as: $\beta_{i,j} \sim N(\beta_{i,j+1}, \sigma^2_\beta)$, and $\delta_j \sim N(\delta_{j+1}, \sigma^2_\delta)$[8]

The model shares polling information across states and across time using an interstate correlation matrix by taking state level election results (& other state-level predictors) and then setting any negative correlations to 0 and keeping the positive correlations as is. Furthermore, with time, the weights of the estimate get updated by $\beta_{s,t}|\beta_{s,t+1} \sim \mathcal{N}\big(\beta_{b,t+1}, \Sigma\big)$ where the priors on Σ try to estimate how much public opinion can change day to day on the state level.

Then the model uses this polling data and combines it with fundamental / economic factors model based on the Abramovitz "Time-for-Change" model[1]. The Time for Change Models is a general linear model that takes into account: net approval, annualized growth of real GDP in second quarter of the election year, and whether the candidate was an incumbent. This results in a model that predicts the vote share as such: $V = c_0 + c_1 NetApproval + c_2 GDPGrowth + c_3 IsIncumbent$. This model is used as the prior of β from above. However, the model proposed by Gelman uses a variant of the fundamentals model with data that is not readily available to us currently. We will call the model we described previously model A and the model that includes all of the additional data Gelman[3] uses Model B. Model B includes real disposable income, non-farm payrolls and the stock market to better showcase voter behavior as a part of their fundamental model, which as we said gets fed into our vote estimate as the prior.

In the end the model uses elastic net regularization to do a general linear model to predict the percent of the popular vote that a candidate will have in the election.

Inference The inference aspect of this project is by the means of Markov Chain Monte Carlo to estimate the posterior distribution determined by our model above. Moreover, the MCMC is ran everyday up until the date of the election for a total of 20,000 simulations to allow state polling averages to drift randomly allowing for subsequent prior distribution updates.[5]

Criticism The current evaluation techniques used in the model involves leave-one-out cross-validation to reduce the issue of overfitting to the data set and then predicting a single point estimate which consequently tells us the probability of the winner of the upcoming presidential election. What the model evaluation metrics fail to tell us is how accurately the inference describes other crucial details. Such crucial details include the average win, the extreme wins, and how varied our data may be. In order to see if the model adequately considers these aspects, we want to do posterior predicative checks to simulate data from our model and then compare the mins, maxs, average, variance, and skewness of our simulated data with the real historic predictions. By doing this we can see if the model wholly describes our data. Moreover, we will also do population posterior checks which involves the bootstrap method of resampling data from our sample and then we can check the same above test statistics (mean, max, min, etc ...) and then see if our model or inference accurately describes the data. Moreover, the model that the Economist describes makes strong assumptions about the weight that covid should play with respect to the ongoing recession and places the overall effect at a 40% hit. We would like to manipulate this hyper-parameter and then see how much the predictions above change with respect to a change in this hyper-parameter.

References

[1] Abramowitz A. *Time-for-Change Fundamentals Model*. https://pollyvote.com/en/components/models/retrospective/fundamentals-plus-models/time-for-change-model/. [Online; accessed 07-Oct-2020]. 2020.

[2] Gelman A. *Github Code*. https://github.com/TheEconomist/us-potus-model. [Online; accessed 23-Oct-2020]. 2020.

[3] Economist. *Forecasting the US elections*. https://projects.economist.com/us-2020-forecast/president/how-this-works. [Online; accessed 23-Oct-2020]. 2020.

[4] Azari J. Gelman A. "19 things we learned from the 2016 election". In: (). URL: http://www.stat.columbia.edu/~gelman/research/published/what_learned_in_2016_5.pdf.

[5] Morris G. Heidemanns M. Gelman A. "An Updated Dynamic Bayesian Forecasting Model for the U.S. Presidential Election". In: (). URL: http://www.stat.columbia.edu/~gelman/research/published/Harvard_Data_Science_Review.pdf.

[6] Huffington. *2016 General Election: Trump vs. Clinton Data*. https://elections.huffingtonpost.com/pollster/2016-general-election-trump-vs-clinton. [Online; accessed 23-Oct-2020]. 2020.

[7] Kremp. *Stan Implementation of Drew Linzer's dynamic Bayesian election forecasting mode*. http://www.slate.com/features/pkremp_forecast/report.html. [Online; accessed 07-Oct-2020]. 2020.

[8] Drew A. Linzer. "Dynamic Bayesian Forecasting of Presidential Elections in the States". In: *Journal of the American Statistical Association* 108.501 (2013), pp. 124–134. eprint: https://doi.org/10.1080/01621459.2012.737735. URL: https://doi.org/10.1080/01621459.2012.737735.