1 Introduction

Since 2004, *PS: Political Science and Politics* has published symposia summarizing forecasting models prior to Election Day. Over the years, forecasting teams have updated their models to deliver an accurate prediction of the national two-party vote share by capturing key features of the underlying data generating process.

Our forecast is different. Using ensemble Bayesian model averaging (EBMA), we combine *multiple* forecasting models (most presented in this symposium) into a single prediction. The idea is that each forecasting team develops a parsimonious model given their particular understanding of the true DGP. The theoretical differences of each model are peripheral. Instead, each model receives a weight based how many predictions a given model has made in the past and how well it was able to predict prior elections. Using the weights and each model's forecast for the 2020 election, we then create an out-of-sample forecast that captures the uncertainty and diversity inherent in the wide array of approaches reflected in this symposium. In short, to increase our likelihood of achieving the most accurate prediction, we incorporate the wisdom and knowledge about US elections implicit within each model.

Across subject domains, scholars have shown ensemble predictions to be more accurate than any individual component model (Bates and Granger 1969; Armstrong 2001; Raftery et al. 2005). EBMA builds on the assumption that there is no "best" model and rather collects *all* the wisdom from multiple prediction efforts.

Ensemble Bayesian model averaging (EBMA), originated in the field of weather forecasting (Raftery et al. 2005), where it was created to aggregate across models to improve out-of-sample forecasts (Montgomery et al. 2012a). In the 2012 symposia, Montgomery et al. (2012b) used EBMA to predict a narrow victory for Barack Obama (50.3%), providing a smaller forecasting error than all but three component models (roughly 1.5 percentage points). In 2016, Hollenbach and Montgomery used the same approach to generate a forecast for *Vox.com*, predicting a narrow

popular vote win for Donald Trump (50.9%). This was off by around two percentage points, but did accurately capture a stronger Trump performance than expected by many observers (Matthews 2016).

One important limitation for using EBMA in this exercise is the sparsity of data and paucity of *true* information. The underlying training data is fantastically small – *just the seven elections sinc 1992* – and component models themselves were developed based on a small set of elections. Building a highly complex model on top of such sparse data risks overfitting, i.e., a component model may receive a high weight for being lucky. We therefore wish to make weights less responsive to any individual data point.

To improve EBMA, we we change our approach in two ways. First, to calibrate weights for the ensemble, we use published out-of-sample predictions for presidential elections dating back to 1992.[1] This is more consonant with the core idea of EBMA, where models should be calibrated based on their out-of-sample predictive accuracy (Montgomery et al. 2015). Second, we shift from the expectation maximizion (EM) approach in Montgomery et al. (2012a), and implement a full Markov chain Monte Carlo method that includes a natural approach to regularization. We introduce a concentration tuning parameter (α). When α is set at a high value (e.g., 10), EBMA produces a simple average, i.e., all models are weighted exactly equal. For small α, the "best" models will receive the lions share of the weight. We adopt a leave-one-out cross validation scheme to choose the optimal value for α.

2 Ensemble Bayesian Model Averaging with MCMC

Assume several research teams are interested in predicting some future event y^{t*}, resulting in k forecasting models $(M_1, M_2, ..., M_k)$. Each team has predicted similar events in the past y^t,

[1] Montgomery et al. (2012b) instead rely on "pseudo" out-of-sample predictions generated from each model but not actually generated before the election.

which we term the calibration period. For each model we have a prior probability distribution $M_k \sim \pi(M_K)$ and the probability distribution function (PDF) for y^t is denoted $p(y^t|M_k)$. Applying Bayes' rule, we derive the marginal predictive distribution of y^{t*} given k predictive models as

$$p(y^{t*}) = \sum_{k=1}^{K} p(y^{t*}|M_k)p(M_k|y^t).$$

This PDF can be interpreted as a weighted prediction, with each model weighted by its past predictive performance prior to $t*$. Each forecast is transformed into a normal PDF centered at the individual forecast $\mathcal{N}(f_k^{t*}, \sigma^2)$.

The insight of the EMBA approach is that each component model in the ensemble captures some understanding of the world that is selectively accurate. If we can calculate good weights based on prior performance, the ensemble can provide better out-of-sample forecasts in terms of accuracy and precision than a randomly selected component. The ensemble down-weights poor performers, naturally reducing their influence. The advantages of ensemble weighting is diminished, however, if many of the component models receive zero weight. Unfortunately, with extremely sparse data, this is a common result for the maximum likelihood approach adopted in Montgomery et al. (2012b) (see Graefe et al. 2015). The MCMC estimation scheme, therefore, is designed to regularize the weights to prevent over-concentration on a small number of models.

To do this we rely on a mixture model framework (Bishop 2006; Grimmer et al. 2017). Let $t = [1, \ldots, T]$ be the year of the prediction, $k = [1, \ldots, K]$ indicate the model component, and y_t be the observed outcome for year t. Further, let $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_K]$ be a $T \times K$ matrix of forecasts, where $\mathbf{f}_k = (f_{k1}, f_{k2}, \ldots, f_{kT})$ is the vector of predictions made by model k. We introduce indices $\boldsymbol{\tau} = [\tau_1, \tau_2, \ldots \tau_T]$, which are latent categorical variables indicating which component generated observation t such that $\tau_t \in [1, 2, \ldots, K] \forall t \in [1, 2, \ldots, T]$. With this setup, the likelihood is

$$p(y_t|\boldsymbol{\tau}, \boldsymbol{\sigma^2}, \mathbf{X}) \sim \sum_{k}^{K} N(x_{kt}, \sigma)\mathcal{I}(\tau_t = k), \tag{1}$$

where $\mathcal{I}(\cdot)$ is the standard indicator function. The model is complete by specifying the following priors/hyperpriors:

$$\pi(\boldsymbol{\tau}) \sim \text{Multinomial}(\boldsymbol{\omega}), \pi(\sigma^2) \sim (\sigma^2)^{-1} \tag{2}$$

$$\pi(\boldsymbol{\omega}) \sim \text{Dirichlet}(\boldsymbol{\alpha}) \tag{3}$$

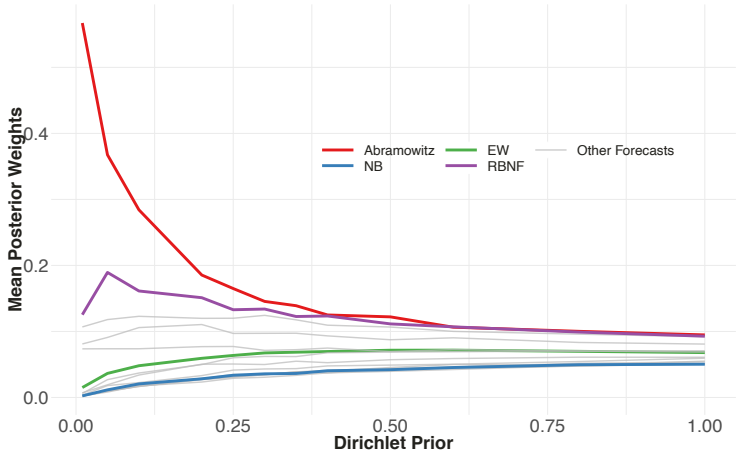
The full MCMC sampling scheme is described in the appendix. In essence, we first estimate $\boldsymbol{\omega}$, which is the overall weight for each model. For each year, we then draw τ_t from a multinomial distribution such that more highly weighted models are more likely to have generated each response. Based on this draw, we can then re-calculate the weights $\boldsymbol{\omega}$ for each model. This process is repeated until we converge to a proper posterior.

Under this framework, researchers must choose the K values for vector $\boldsymbol{\alpha}$, which represents the prior belief that each model best captures the underlying DGP. In theory, researchers could use this prior to give more weight to some models over others. Instead, we choose a single scalar, such that the prior weight is identical across model. High values for α represent a stronger prior, with a value of 1 intuitively indicating one case where each model is known to have been the "best" for one election. Since the data only include seven elections, even a value of $\alpha = 1$ is quite strong. In settings with larger datasets, the prior choice would be less consequential.

Figure 1 shows the model weights for various values of α. The figure shows that the most accurate model (the Abramowitz forecast) receives over half of the final weight in the absence of regularization. However, the weights are much more evenly distributed with higher α values. For our final forecast, we choose the optimal α value using leave-one-out cross validation. This leads us to set $\alpha = 0.225$.[2]

[2]We focus on values of α above 0.1. Below that value, we found the posterior samples to be too unstable.

Figure 1: Regularizing model weights



Model weights for exemplar components as a function of α. Higher values spread out the weights more evenly.

3 The EBMA Forecast for 2020

We apply our ensemble averaging method to the 2020 presidential election by using the historical out-of-sample forecasts of each component model as the calibration-period predictions to estimate the model weights. We use the predictions generously provided by Abramowitz; Campbell; Erikson and Wlezien; Lewis-Beck and Tien; Lockerbie; Fair; DeSart; Graefe; Bruno, Veronique and Nadeau; Lichtman; Rietz, Berg, Nelson, and Forsythe from the models described in this and past symposia. This results in a ensemble of 11 forecasting models and a training period of 7 presidential elections from 1992-2016. Each of the models and their prior predictions can be found in the Appendix along with the RMSE of their combined predictions.[3] The test period will, of course, be the 2020 election.

Table 1 shows the mean of the posterior distribution of weights assigned to each model based on prior forecasts. The model still privileges the Abramowitz forecast and reduces weight for models

[3]NOTE: We are still refining the set of forecasts to include in our ensemble to reflect the composition of the symposium and available forecasts. We are in communication with several authors to get their prediction for the two-party vote share. This forecast reflects the predictions we have at the time of this writing.

Table 1: Mean Posterior Model Weights

Model	Posterior Mean
Abramowitz	0.207
Bruno, Veronique & Nadeau	0.034
Campbell	0.113
DeSart	0.052
Erikson & Wlezien	0.079
Fair	0.065
Graefe	0.033
Lewis-Beck & Tien	0.159
Lichtman	0.032
Lockerbie	0.032
Rietz, Berg, Nelson & Forsythe	0.195

that have historically had higher error rates or fewer true out-of-sample forecasts. For comparison, in a simple average of 11 models, each model's weight would be equal to 0.09. The final mean posterior prediction of the ensemble forecast is 45.80 with a 95% credible interval ranging from 43.58 to 51.28. Figure 2 shows the full posterior density of the ensemble prediction.

4 Conclusion

Ensemble methods are designed to combine insights from a diversity of forecasting methods into a single prediction. In 2016, these forecasts were widely dispersed, predicting both significant wins and significant losses for Trump. In 2020, the wisdom of crowds is offering a much stronger consensus which is reflected in our combined forecast. Nearly every individual indicator – economic growth, performance in office, public approval – points towards a poor performance for President Trump. Still a Biden victory is not a foregone conclusion, and at this point there remains a significant chance of a Trump re-election.

Before concluding, we should recognize that every election is in some way unique. The purpose of this forecasting exercise is in part to uncover the fundamental factors that set the context

Figure 2: Posterior Predictive Density

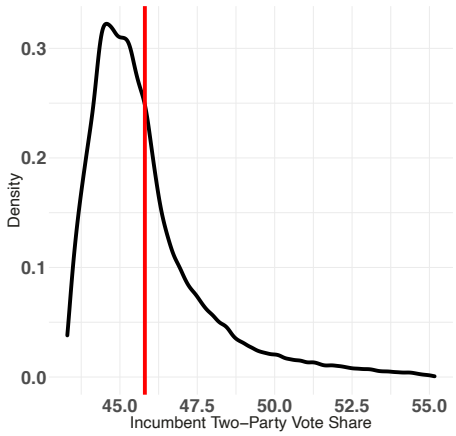


Figure shows the full posterior predictive density of the EBMA forecast. The red vertical line shows the mean prediction.

for elections and restrict the set of plausible outcomes. The year 2020, however, has been unique in so many ways, that there is a higher than normal probability of something unusual happening. There is a risk that models trained on previous elections may not accurately predict 2020. There has never been anything like 2020 in US election history. So far we have experienced a presidential impeachment, a deadly pandemic, a record-setting economic collapse, historically massive street protests for racial justice, and a sitting president attacking the legitimacy of our elections. Unprecedented is not just a word, but a warning. None of the component models include anything like 2020 in their training data. In this setting, all forecasts must necessarily be taken with a grain of salt.

References

Armstrong, J. S. (2001). Combining forecasts. In J. S. Armstrong (Ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Norwell, MA: Kluwer Academic Publishers.

Bates, J. and C. W. Granger (1969). The combination of forecasts. *Operations Research 20*(4), 451–468.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Graefe, A., H. Küchenhoff, V. Stierle, and B. Riedl (2015). Limitations of ensemble bayesian model averaging for forecasting social science problems. *International Journal of Forecasting 31*(3), 943–951.

Grimmer, J., S. Messing, and S. J. Westwood (2017). Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods. *Political Analysis 25*(4), 413–434.

Matthews, D. (2016, Aug). The trump tax: A daily number showing how trump is blowing the election for the gop.

Montgomery, J. M., F. Hollenbach, and M. D. Ward (2012a). Improving predictions using ensemble Bayesian model averaging. *Political Analysis 20*(3), 271–291.

Montgomery, J. M., F. Hollenbach, and M. D. Ward (2012b). Ensemble predictions of the 2012 US presidential election. *PS: Political Science & Politics 45*(4), 651–654.

Montgomery, J. M., F. M. Hollenbach, and M. D. Ward (2015). Calibrating ensemble forecasting models with sparse data in the social sciences. *International Journal of Forecasting 31*(3), 930–942.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review 133*(5), 1155–1174.