An augmented confidence-weighting algorithm

Motivation

A simple way to aggregate judgments on the occurrence of an event is confidence-weighting. Putting a higher weight on more confident forecasters' judgments could produce a more accurate prediction when confidence correlates with the informativeness of forecasters' signals. Confidence-weighting would not perform well when there is systematic under or over confidence. To illustrate, consider a True/False question. Suppose forecasters who believe the correct answer is 'false' are on average less confident than the forecasters who pick 'true'. Then, confidence-weighting would often predict 'true' even when the correct answer is 'false'. If the aggregator knew about the relative under-confidence on the one side (with respect to the average confidence on the other side), it could be possible to adjust the aggregate prediction accordingly. This project explores if meta-beliefs can be used to infer and correct for a potential relative under-confidence in the correct answer to a binary question.

The Method

Consider the framework in Wilkening et al. (2021). Each forecaster in a crowd of N receives a signal s_i from $\{s_1, \ldots, sm\} \cup \{s_\emptyset\}$, normalized so that $s_i := p(T|s_i)$ where $p(T|s_i)$ is forecaster i's posterior belief on the probability of event being true. Forecasters are asked to report i) a probabilistic prediction on probability of T and ii) a probabilistic meta-prediction on the average of others' predictions. Forecaster i's truthful prediction and meta-prediction are given by s_i and $M^p(Q|s_i) = s_i EP(\mathbf{Q}|T) + (1 - s_i) EP(\mathbf{Q}|F)$ where $\mathbf{Q}$ is the information matrix defined as in Wilkening et al. (2021). To simplify the notation a bit, let $\mu^F = EP(\mathbf{Q}|F)$ and $\mu^T = EP(\mathbf{Q}|T)$ be the expected average forecast in states F and T respectively. Then we can write $M^p(Q|s_i) = \mu^F + (\mu^T - \mu^F)s_i$. With a slight abuse of notation, let P_i and M_i^p also denote forecaster i's reported prediction and meta-prediction on the event. Suppose forecasters' prediction reflects their true signal but meta-predicitions are reported with a random error. Then, we can formulate:

$$M_i^p = \mu^F + (\mu^T - \mu^F)P_i + \epsilon_i \tag{1}$$

where ϵ_i is the random error. Estimating this linear model (using data from this event only) produces $\hat{\mu}^F$ and $\hat{\mu}^T$, which represent the estimated average forecast in states F and T. Intuitively, these estimates show the expected confidence when the associated state is the actual state. To illustrate, suppose $\{\hat{\mu}^F, \hat{\mu}^T\} = \{0.3, 0.9\}$. The average confidence in a given question is predicted to be 90% (70%) if the correct answer is T (F). So, average forecast is *relatively under-confident* in state F (w.r.t the average confidence in state T). Suppose the correct answer is F. Based on $\{\hat{\mu}^F, \hat{\mu}^T\}$, we may infer that the realized average forecast is less confident than it would be if the actual state were T instead. Then, it is more likely that the realized average forecast could be on the wrong side of 0.5, leading to an inaccurate aggregate judgment. It could be possible to improve confidence-weighted estimate by selective extremization and anti-extremization to compensate for the under-confidence. Here's a description of the aggregation procedure:

```
Algorithm 1: Augmented confidence-weighting
Collect $\{P_i, M_i^p\}$ from each forecaster $i$;
Calculate $\bar{P}^F$ and $\bar{P}^T$ (average among forecasters with $P_i < 0.5$ and $P_i > 0.5$ respectively);
Estimate $\{\hat{\mu}^F, \hat{\mu}^T\}$ from the regression in Equation 1;
if $\hat{\mu}^F < \hat{\mu}^T$ (under-condidence at F) then
    Extremize $\bar{P}^F$ (towards 0) and anti-extremize $\bar{P}^T$;
else if $\hat{\mu}^F < \hat{\mu}^T$ (under-condidence at T) then
    Extremize $\bar{P}^T$ (towards 1) and anti-extremize $\bar{P}^F$;
Calculate average forecast $\bar{P}$ using extremized (or anti-extremized) $\bar{P}^F$ and $\bar{P}^T$;
Pick 'false' if $\bar{P} \leq 0.5$, 'true' else;
```

For this algorithm to improve accuracy, the under-confidence in the actual state ($\hat{\mu}^F$ vs. $\hat{\mu}^T$) should relate to average forecast $\bar{P}$ being on the wrong side of 0.5 due to lower confidence in the actual state. Also, the extent of (anti-)extremization could be important. The section below presents preliminary evidence from the General Knowledge and States datasets in Wilkening et al. (2021).

Preliminary evidence

Figure 1 depicts the average predictions in all questions (General Knowledge and States data pooled). The left (right) panels include the questions in which the correct answer is ‘false’. In each panel, the x-axis represents the state at which the estimated confidence is lower. In 529 of the 550 tasks, the estimated average confidence is lower in ‘false’.

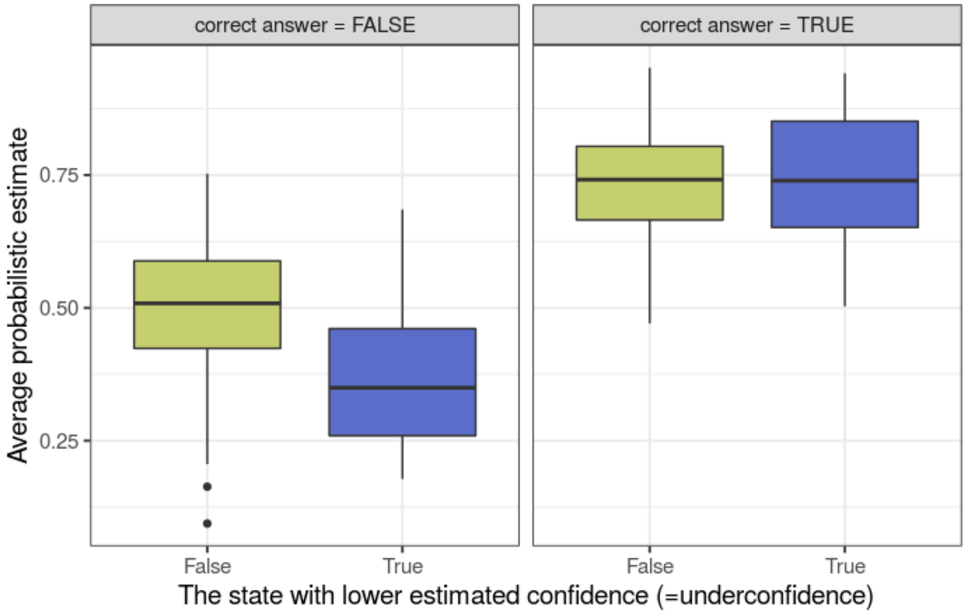

Figure 1

The left panel shows that, in questions where ‘false’ is predicted to be the under-confident state, the realized average prediction falls above 0.5 in more than half of the cases. In these questions, the augmented confidence-weighting will adjust the average prediction towards 0, improving the accuracy. In the same panel, for the questions with under-confidence in ’true’ (blue box plot), the adjustment will be in the incorrect direction.

What should be the extent of extremization? Consider the extremization function used by Baron et al. (2014) and Martinie et al. (2020), given by $t(p) = p^a/[p^a + (1-p)^a]$. If the recalibration parameter a is greater (less) than 1, the average prediction is extremized (anti-extremized). Figure

2 reproduces Figure 6 in Wilkening et al. (2021). Panels differ in terms of extremization and anti-extremization parameters used in augmented confidence weighting. For example in the top-left panel, the extremization and anti-exremization are done with $a = 2.2$ and $a = 0.9$ respectively.

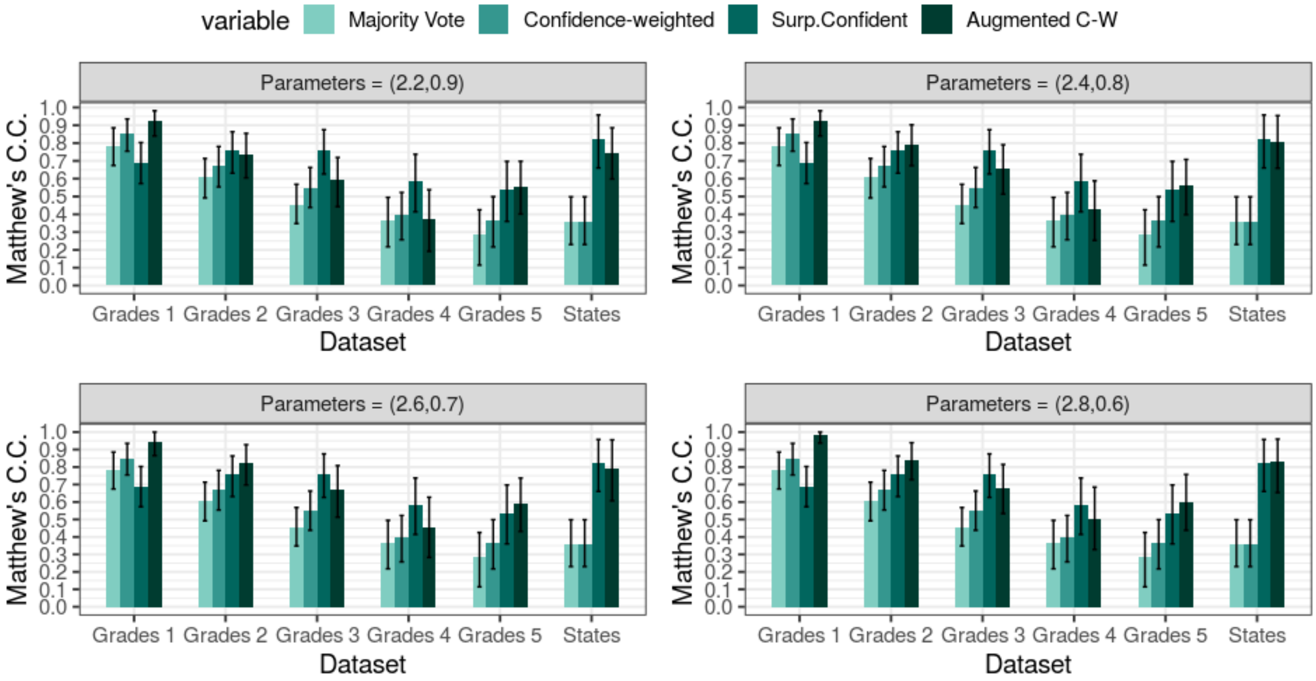

Figure 2

The augmented confidence-weighting performs better than majority vote an standard confidence-weighting in almost all data sets and parameter values. Furthermore, the performance is comparable to the SC algorithm in Grade 5 questions and in the States data. However, the augmented confidence-weighting performs better than the SC algorithm in Grade 1 questions. To understand why, consider a question where the correct answer is 'true' and the average forecast is 0.8. Suppose the extremization happens to be in the wrong direction. Even under the strongest extremization considered in Figure 2, the resulting average forecast is very likely to be above 0.5. Augmented confidence-weighting could flip the prediction on the outcome in more tricky questions where the average forecast is close to 0.5. As shown in Figure 1, average confidence on 'false' is lower than 50% in many questions where 'false' is the correct answer (i.e. average forecast is higher than 0.5). Augmented confidence-weighting improves accuracy in these questions while not distorting the properly confident aggregate predictions in other questions.

What next?

In the datasets considered, the procedure predicts under-confidence in 'false' in most questions (529 out of 550). In almost all cases, the average forecast is adjusted downwards. However, in 267 out of 550 questions the correct answer is 'true'. So the adjustment is in the opposite direction of the correct answer in many cases. Nevertheless, it was effective because most average forecasts around 0.5 needed an adjustment towards 0 (leftmost box plot in Figure 1). The algorithm should be tested with new datasets to see if it correctly predicts under-confidence in 'true' in questions where the correct answer is 'true' but the average forecast is less than 0.5 (i.e., most average forecasts around 0.5 need an adjustment towards 1).

Figure 2 suggests that augmented confidence-weighting could improve over simple confidence-weighting for various levels of extremization parameters. More work could be done on how the parameters should be set.

References

Baron, J., Mellers, B. A., Tetlock, P. E., Stone, E., and Ungar, L. H. (2014). Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis*, 11(2):133–145.

Martinie, M., Wilkening, T., and Howe, P. D. (2020). Using meta-predictions to identify experts in the crowd when past performance is unknown. *Plos one*, 15(4):e0232058.

Wilkening, T., Martinie, M., and Howe, P. D. (2021). Hidden experts in the crowd: Using meta-predictions to leverage expertise in single-question prediction problems. *Management Science*.