Turner, B., & Steyvers, M. (2011). A Wisdom of the Crowd Approach to Forecasting. 2nd NIPS workshop on Computational Social Science and the Wisdom of Crowds.

A Wisdom of the Crowd Approach to Forecasting

Brandon M. Turner and Mark Steyvers
Department of Cognitive Sciences
University of California, Irvine
Irvine, CA 92697-5100
turner.826@gmail.com

Abstract

The "wisdom of the crowd" effect refers to the phenomenon that the mean of estimates provided by a group of individuals is more accurate than most of the individual estimates. This effect has mostly been investigated in general knowledge or almanac types of problems that have pre-existing solutions. Can the wisdom of the crowd effect be harnessed to predict the future? We present two probabilistic models for aggregating subjective probabilities for the occurrence of future outcomes. The models allow for individual differences in skill and expertise of participants and correct for systematic distortions in probability judgments. We demonstrate the approach on preliminary results from the Aggregative Contingent Estimation System (ACES), a large-scale project for collecting and combining forecasts of many widely-dispersed individuals.

1 Introduction

When a group of people make an estimate about a quantity, often the mean of these estimates is better than the majority of the group [1]. The so called wisdom of the crowd (WoC) phenomenon is typically studied in the context of a single magnitude estimate (e.g., the weight of an oxen [2]). However, the WoC effect has recently been studied in a variety of other tasks involving more complicated sources of information such as rank-ordering tasks [3-5]. The aggregation models for these tasks have demonstrated that it is possible to infer people's expertise or skill directly from the answers they provide. These latent measures of expertise outperformed self-report measures such as confidence ratings, in terms of correlation with the actual accuracy of the answers.

In this paper, we develop probabilistic models for the aggregation of human judgments in a forecasting situation, in which individual differences such as latent expertise play a key role. We build on previous statistical models [6, 7] that assume that there is a "shared truth" among all participants. However, it is well-known that subjects are generally biased in their responses, such as responding with confidence that is higher than their accuracy [9]. As such, many models of confidence treat these responses as a "distortion" from the true probability.

The models presented in this article include a generative process that explains how individuals produce their forecasts after some distortion of the aggregate knowledge or shared truth. Because there are a variety of ways this distortion could occur within the models, we refer to bias as systematic distortion and random variability as random distortion. The models allow for differences in latent expertise of the participants as well as question difficulty. In addition, the aggregation models also allow for systematic deviations from the shared truth such that the distribution of individual judgments is systematically distorted from the latent group knowledge. Model parameters related to latent expertise, question difficulty, and systematic distortion are estimated on the basis of judgments from a group of participants that provide forecasts on a number of questions. In order to reliably estimate the systematic distortion parameters, we also require that some subset of forecasting questions have a known outcome (i.e., they have already resolved).

We compare the probabilistic aggregation models to a standard linear unweighted average of the forecasting judgments. Previously, it has been suggested that the unweighted linear average is difficult to outperform in many forecasting situations [8].

2 Experiment

The forecasting judgments were collected by the Aggregative Contingent Estimation System (ACES), a large-scale project for collecting and combining forecasts of many widely-dispersed individuals (`http://www.forecastingace.com/aces`). Participants involving members of general public were asked to estimate the probability of future events occurring, such as the outcome of presidential elections in Taiwan. At first, there were no known answers to any of these questions. However, the questions were designed such that an answer would be determined at some fixed date. The data set included the judgments from a one-month period involving 817 participants and 51 questions (18 of which resolved by the time this dataset was put together).

3 Models

We let $y_{i,j}$ represent the probability estimate provided by Subject i on Question j and let x_j be the result of the resolved Question j. For this particular data set, x_j is only partially observed. That is, some questions have not been resolved and thus x_j on unresolved questions is treated as a missing observation.

3.1 Unweighted Linear Average

The base model we used simply calculates the unweighted linear average of the estimates provided by each of the subjects. Thus, predictions $\widehat{\mu_j}$ for this model are obtained by evaluating

$$\widehat{\mu_j} = \frac{1}{n}\left(\sum_i y_{i,j}\right),$$

where n is the number of responses obtained on Question j.

Despite its simplicity, the unweighted linear average is a formidable estimate for forecasting data. Some authors have argued that it is impossible to beat the unweighted linear average by more than 20% (e.g., [8]). Thus, our goal in modeling this data is to provide estimates that are better than, or as good as, the unweighted linear average.

We now present two Bayesian models that each assume a distortion occurs prohibiting subjects from accurately forecasting the probability of events. We examine this distortion in two ways. The first model assumes that the probabilities provided by each subject are distorted versions of the true latent probability on a question-specific basis and the second assumes that the distortion occurs on a subject-specific basis.

For both Bayesian models, predictions for Question j are obtained by taking the mean of the posterior distribution for μ_j, or

$$\widehat{\mu_j} = \frac{1}{K}\left(\sum_{k=1}^{K} \mu_{j,k}\right), \tag{1}$$

where K is the number of samples obtained using Markov chain Monte Carlo (MCMC), and $\mu_{j,k}$ is the kth sample of the posterior corresponding to μ_j.

3.2 Question Distortion Model

The first model we examine treats the observed confidence ratings $y_{i,j}$ as perturbations of the latent mean μ_j for Question j. The Question Distortion Model (QDM) captures distortion in two ways. First, the model allows for systematic distortion τ_j for the jth question. Second, the model assumes a random distortion process, governed by the dispersion parameter ω_j. Modeling the data in this way allows for a "staggering" effect, which can accommodate any question-specific biases that may occur.

The left panel of Figure 1 shows a graphical diagram for this model. Because the subjective probabilities are bounded by zero and one, we used the flexible beta distribution reparameterized to have mean $y'_{i,j}$ and dispersion parameter η, which was assumed to be fixed across all subjects. The distortion of the true latent probability occurs at the node $y'_{i,j}$. We assume that $y'_{i,j}$ is a beta random variable with mean $\tau_j + \mu_j$ and dispersion parameter ω_j. Each of these question-specific parameters assume a common hyper-distribution.

The data x_j influences the mean μ_j only if Question j has a solution. Otherwise, the mean μ_j was estimated on the basis of the individual confidence ratings. Thus, if a question is resolved, then the posterior distribution for μ_j will balance the two sources of information by assuming that the probability of an event occurring is a Bernoulli random variable with mean μ_j and that the estimates provided by the subjects are beta random variables with mean equal to $\mu_j + \tau_j$.

3.3 Individual Distortion Model

The second model we examine is very similar to the QDM. The Individual Distortion Model (IDM) also captures distortion in two ways. As in the QDM, the IDM assumes a random distortion process controlled by ω_j. Second, the IDM assumes that the systematic distortion is not due to the question-specific biases, but is due to subject-specific biases.

The right panel of Figure 1 shows a graphical diagram of the IDM. All model specifications were equal to the QDM, with the exception of the subject-specific distortion parameter. Here, we assume that the $y'_{i,j}$s are beta random variables with mean $\mu_j + \tau_i$ and dispersion parameter ω_j. Both question-specific and subject-specific parameters are assumed to arise from common hyper-distributions.

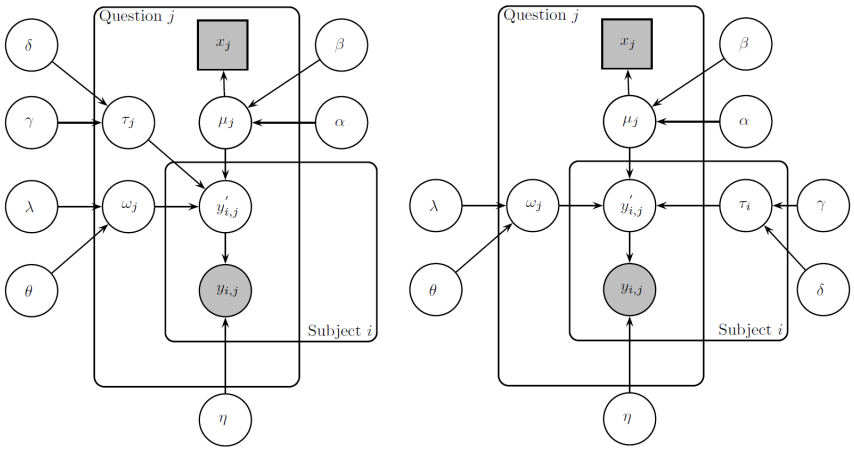

Figure 1: Graphical diagrams of the Question Distortion Model (left panel), and the Individual Distortion Model (right panel).

4 Results

We used JAGS to estimate the joint posterior distribution. For each model, we obtained 1,000 samples from the joint posterior after a burn-in period of 1,000 samples, and collapsed across two chains. Once estimation was complete, we compared the predictive power of the models relative to the unweighted linear average. To evaluate the models, we computed the Brier score by calculating

$$B_j = \sum_{i=1}^{2} \left(X_{i,j} - \phi_{i,j} \right),$$

where $\phi_{1:2,j} = \{\widehat{\mu_j}, 1 - \widehat{\mu_j}\}$ and $X_{1:2,j}$ is the outcome of the jth question with one minus the outcome of the jth question. For example, if the jth question resolved and the event did occur, then

$X_{1:2,j} = \{1, 0\}$. Thus, the best possible Brier score is zero, and the worst possible score is two. If an event did not resolve in the one-month duration, the predictions were not scored and they had no effect on the model's performance.

Once a Brier score was obtained for each of the three models, we computed the percent improvement of the two Bayesian models over the unweighted linear average. The QDM was better than the unweighted linear average at forecasting future events by 4.7%. This is a sizable difference because the Brier score was obtained on only 18 resolved questions. The IDM performed much better, defeating the unweighted linear average by 9.6% percent. This is a dramatic difference, especially on so few data.

As a final examination of the model, we plotted the posterior predictive distributions against each observed probability estimate. Figure 2 shows these results. The predictions of the QDM (left panel) are markedly different from the predictions of the IDM (right panel). In particular, the predictions of the QDM seem to cover lower areas of the prediction space than the IDM, and tend to show little variation as a function of the observed probability estimate. By contrast, the predictions of the IDM tend to be more variable, possibly due to the increased number of parameters accounting for individual differences rather than question differences.

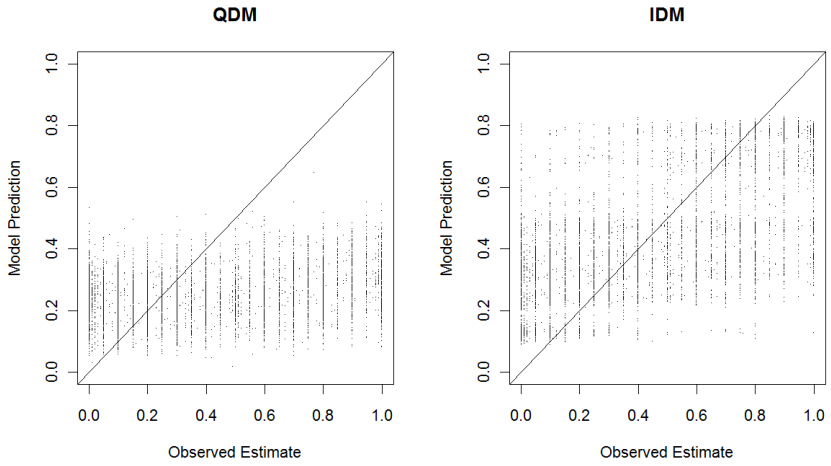

Figure 2: The posterior predictive distribution for the QDM (left panel) and the IDM (right panel).

5 Conclusions

In this article, we have illustrated two Bayesian models for improving forecasting accuracy. The first of these models, the QDM, assumed that any distortion that may be present was due entirely to the questions that subjects were responding to. This model performed favorably to the unweighted linear average, increasing estimation accuracy by 4.7%.

The second model we examined was very similar, but altered the assumption about where the distortion from the true latent probability occurred. The IDM model assumed that the distortion was a subject-specific process. Modifying this one assumption resulted in a major improvement in the model performance. The IDM surpassed the unweighted linear average by 9.6%, more than twice the improvement obtained by the QDM.

Acknowledgments

MS acknowledges support from the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20059. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

References

[1] Surowiecki, J. (2004). *The Wisdom of Crowds.* New York: Random House.

[2] Galton, F. (1907). Vox Populi. *Nature*, 75, 450-451.

[3] Steyvers, M., Lee, M.D., Miller, B., & Hemmer, P. (2009). The Wisdom of Crowds in the Recollection of Order Information. In Y. Bengio and D. Schuurmans and J. Lafferty and C. K. I. Williams and A. Culotta (Eds.) Advances in Neural Information Processing Systems, 22, pp. 1785-1793. MIT Press.

[4] Lee, M. D., Steyvers, M., de Young, Mindy, & Miller, B. (in press). Inferring Expertise in Knowledge and Prediction Ranking Tasks. *TopiCS*.

[5] Yi, S.K.M., Steyvers, M., & Lee, M.D. (in press). The Wisdom of Crowds in Combinatorial Problems. *Cognitive Science*.

[6] Batchelder, W.H. & Romney, A.K. (1988). Test theory without an answer key. *Psychometrika*, 53, 71-92.

[7] Merkle, E.C., & Steyvers, M. (2011). A Psychological Model for Aggregating Judgments of Magnitude. Conference on Social Computing, Behavioral Modeling, and Prediction, 11.

[8] Armstrong, J. S. (2001) *Principles of Forecasting*. Norwell, MA: Kluwer Academic.

[9] Kahneman, D. & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47, 263-291.