Merkle, E.C., & Steyvers, M. (2011). A Psychological Model for Aggregating Judgments of Magnitude. In J. Salerno et al. (Eds.), Social Computing, Behavioral Modeling, and Prediction, LNCS6589, 11, pp. 236-243. Berlin: Springer-Verlag.

A Psychological Model for Aggregating Judgments of Magnitude

Edgar C. Merkle[1] and Mark Steyvers[2]

[1] Department of Psychology
Wichita State University
edgar.merkle@wichita.edu
[2] Department of Cognitive Sciences
University of California, Irvine
mark.steyvers@uci.edu

Abstract. In this paper, we develop and illustrate a psychologically-motivated model for aggregating judgments of magnitude across experts. The model assumes that experts' judgments are perturbed from the truth by both systematic biases and random error, and it provides aggregated estimates that are implicitly based on the application of nonlinear weights to individual judgments. The model is also easily extended to situations where experts report multiple quantile judgments. We apply the model to expert judgments concerning flange leaks in a chemical plant, illustrating its use and comparing it to baseline measures.

Keywords: Aggregation, magnitude judgment, expert judgment.

1 Introduction

Magnitude judgments and forecasts are often utilized to make decisions on matters such as national security, medical treatment, and the economy. Unlike probability judgments or Likert judgments, magnitude judgments are often unbounded and simply require the expert to report a number reflecting his or her "best guess." This type of judgment is elicited, for example, when an expert is asked to forecast the year-end value of a stock or to estimate the number of individuals afflicted with H1N1 within a specific geographic area.

Much work has focused on aggregating the judgments of multiple experts, which is advantageous because the aggregated estimate often outperforms the individuals [1]. A simple aggregation method, the *linear opinion pool*, involves a weighted average of individual expert judgments. If experts are equally weighted, the linear opinion pool neglects variability in: (1) the extent to which experts are knowledgeable, and (2) the extent to which experts can translate their knowledge into estimates. As a result, a variety of other methods for assigning weights and aggregating across judges have been proposed. These methods have most often been applied to probability judgments [2,3].

In the domain of probability judgments, the *supra-Bayesian* approach to aggregation [4,5] has received attention. This approach involves initial specification

J. Salerno et al. (Eds.): SBP 2011, LNCS 6589, pp. 236–243, 2011.

of one's prior belief about the occurrence of some focal event. Experts' judged probabilities are then used to update this prior via Bayes' theorem. This method proves difficult to implement because one must also specify a distribution for the experts' judgments conditional on the true probability of the focal event's occurrence. As a result, others have focused on algorithms for weighting expert judgments. Kahn [6] derived weights for a logarithmic opinion pool (utilizing the geometric mean for aggregation), with weights being determined by prior beliefs, expert bias, and correlations between experts. This method requires the ground truth to be known for some items. Further, Cooke [7] developed a method to assign weights to experts based on their interval (i.e., quantile) judgments. Weights are determined both by the extent to which the intervals are close to the ground truth (for items where the truth is known) and by the width of the intervals. This method can only be fully utilized when the ground truth is known for some items, but it can also be extended to judgments of magnitude.

Instead of explicitly weighting experts in an opinion pool, Batchelder, Romney, and colleagues (e.g., [8,9]) have developed a series of Cultural Consensus Theory models that aggregate categorical judgments. These models account for individual differences in expert knowledge and biases, and they implicitly weight the judgments to reflect these individual differences. The aggregated judgments are not obtained by averaging over expert judgments; instead, the aggregated judgments are estimated as parameters within the model. Additionally, these models can be used (and are intended for) situations where the ground truth is unknown for all items.

In this paper, we develop a model for aggregating judgments of magnitude across experts. The model bears some relationships to Cultural Consensus Theory, but it is unique in that it is applied to magnitude judgments and can handle multiple judgments from a single expert for a given item. The model is also psychologically motivated, building off of the heuristic methods for psychological aggregation described by Yaniv [10] and others. Finally, the model can be estimated both when the ground truth is unknown for all items and when the ground truth is known for some items.

In the following pages, we first describe the model in detail. We then illustrate its ability to aggregate expert judgments concerning leaks at a chemical plant. Finally, we describe general use of the model in practice and areas for future extensions.

2 Model

For a given expert, let X_{jk} represent a magnitude prediction of quantile q_k for event j (in typical applications that we consider, $q_1 = .05; q_2 = .5; q_3 = .95$). All predictions, then, may be represented as a three-dimensional array $\boldsymbol{X}$, with the entry X_{ijk} being expert i's prediction of quantile q_k for event j. The model is intended to aggregate across the i subscript, so that we have aggregated quantile estimates for each event.

We assume an underlying, "true" uncertainty distribution for each event j, with individual experts' uncertainty distributions differing systematically from the true distribution. Formally, let V_j be the true uncertainty distribution associated with event j. Then

$$V_j \sim \mathrm{N}(\mu_j^*, \sigma_j^{2*}), \tag{1}$$

where normality is assumed for convenience but can be relaxed. For events with known outcomes, we can fix μ_j^* at the outcome. Alternatively, when no outcomes are known, we can estimate all μ_j^*.

We assume that each expert's uncertainty distribution differs systematically from the true distribution. The mean (μ_{ij}) and variance (σ_{ij}^2) of expert i's distribution for event j is given by:

$$\begin{aligned} \mu_{ij} &= \alpha_i + \mu_j^*, \\ \sigma_{ij}^2 &= \sigma_j^{2*}/D_i, \end{aligned} \tag{2}$$

where the parameter α_i reflects systematic biases of expert i and receives the hyperdistribution:

$$\alpha_i \sim \mathrm{N}(\mu_\alpha, \phi_\alpha),\ i = 1, \ldots, N. \tag{3}$$

The parameter D_i reflects expertise: larger values reflect less uncertainty. D_i is restricted to lie in $(0, 1)$ $\forall$ i because, by definition, expert uncertainty can never be less than the "true" uncertainty.

We assume that expert i's quantile estimates arise from his/her unique uncertainty distribution and are perturbed by response error:

$$X_{ijk} \sim \mathrm{N}(\Phi^{-1}(q_k)\sigma_{ij} + \mu_{ij}, \gamma_i^2), \tag{4}$$

where $\Phi^{-1}(\cdot)$ is the inverse cumulative standard normal distribution function. The response error reflects the fact that the expert's quantile estimates are likely to stray from the quantiles of his/her underlying normal distribution.

We are estimating a Bayesian version of the above model via MCMC, so we need prior distributions for the parameters. These are given by

$$\begin{array}{lll} D_i \sim \mathrm{Unif}(0,1) & \gamma_i^2 \sim \mathrm{Unif}(0,1000) & \mu_j^* \sim \mathrm{N}(0,1.5) \\ \mu_\alpha \sim \mathrm{N}(0,4) & \phi_\alpha \sim \mathrm{Unif}(0,1000) & \sigma_j^{2*} \sim \mathrm{Gamma}(.001,.001) \end{array} . \tag{5}$$

The above priors are minimally informative, though this is not obvious for the priors on μ_j^* and μ_α. We further describe the rationale for these priors in the application. An additional advantage of the Bayesian approach involves the fact that we can easily gauge uncertainty in the estimated ground truth.

In summary, the model developed here is unique in that: (1) the ground truth for each item is estimated via a model parameter, and (2) specific distributional assumptions are very flexible. Psychological aspects of the model include the facts that: (1) judges are allowed to have systematic response biases; and (2) judges are allowed to vary in expertise.

3 Application: Flange Leaks

To illustrate the model, we use data on the causes of flange leaks at a chemical plant in the Netherlands [11,12]. Ten experts estimated the number of times in the past ten years that specific issues (called "items" below) caused a failure of flange connections. Issues included "changes in temperature caused by process cycles," "aging of the gasket," and "bolts improperly tightened." For each issue, the experts reported 5%, 50%, and 95% quantiles of their uncertainty distributions. There were fourteen issues assessed, yielding a total of $10 \times 14 \times 3 = 420$ observations. The ground truth (i.e., true number of times the issue led to a failure in the past ten years) was known for eight of the fourteen issues.

3.1 Implementation Details

The judgments elicited from experts varied greatly from item to item, reflecting the fact that some issues were much more frequent than others. Figure 1 illustrates this phenomenon in more detail. The figure contains 8 panels, each of which displays individual expert judgments in grey (labeled `E1` through `E10`). The dots with horizontal lines reflect experts' 5%, 50%, and 95% quantile judgments, and the horizontal line in each panel reflects the truth. The vast differences in magnitude can be observed by, e.g., comparing judgments for item 5 with those for item 6 (keeping in mind that the y-axis changes for each item). Further, focusing on item 5, we can observe extreme intervals that are far from the truth (for item 5, close to zero).

Both the vast differences in judged magnitudes and the large outliers lead to poor model performance. To minimize these issues, we fit the model to standardized, log-transformed judgments for all items. These transformations make model estimation more stable, and the resulting model estimates can be "untransformed" to obtain aggregated estimates on the original scale. The transformed data also lead to default prior distributions on the μ_j^*: because we are dealing with standardized data, a reasonable prior would be $\text{N}(0, 1)$. Because experts tend to under- or overestimate some quantities, we have found it helpful to increase the prior variance to, say, 1.5. This also leads to the prior variance on μ_α: because we are dealing with standardized data, it would be quite surprising to observe a mean bias outside of $(-2, 2)$.

We used three implementations of the model to aggregate estimates in different ways. In the first implementation (`M1`), we fit the model to expert judgments for all 14 items, excluding the ground truth. This implementation reflects a situation where the truth is unknown and estimates must be aggregated. In the second implementation (`M2`), we used the same data but included the ground truth for the first four items. This reflects a situation where the truth unfolds over time or was already known to the analyst, so that experts can be "graded" on some items but not others. The third model implementation (`M3`) held out the ground truth for one item at a time. Thus, the estimated interval for the

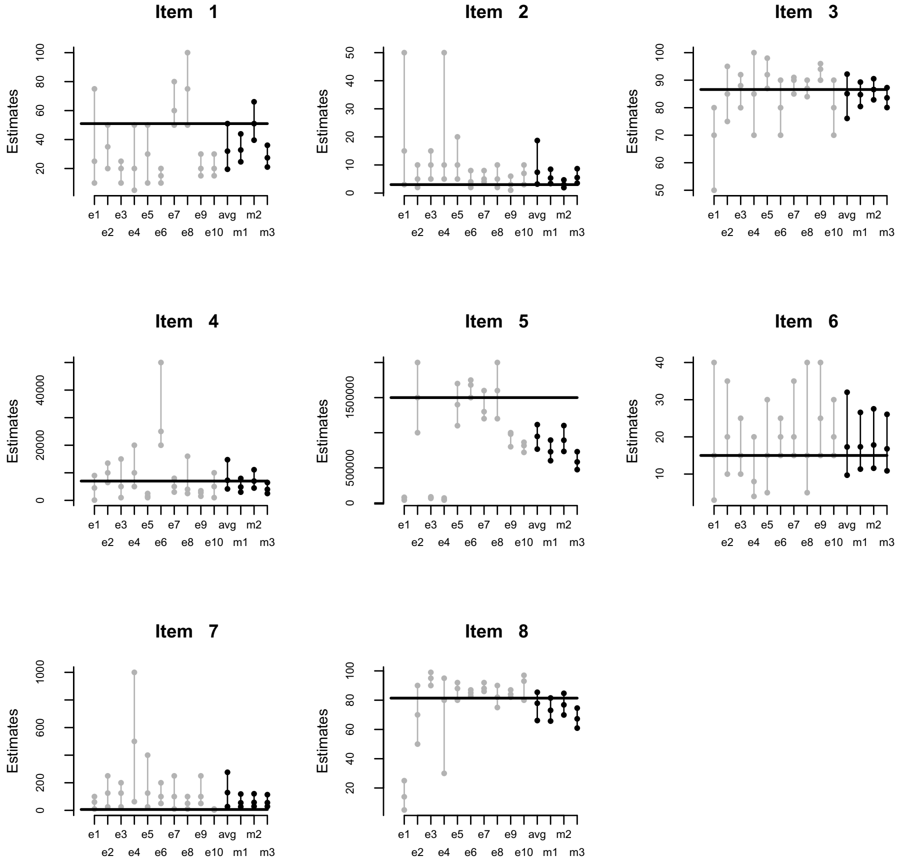

Fig. 1. Magnitude estimates of ten experts and four aggregation methods for eight items. Vertical lines with points reflects 5%, 50%, and 95% quantile judgments; "e" labels represent experts, and the "avg" label represents the unweighted average aggregation. "M1" represents model-based aggregations with no ground truth, "M2" represents model-based aggregations with ground truth for the first four items, and "M3" represents model-based aggregations for all items except the one being estimated. The horizontal line in each panel represents the ground truth.

first item utilized the ground truth for items 2–8. The estimated interval for the second item utilized the ground truth for items 1 and 3–8, and so on. Finally, for comparison, we also aggregated the estimates via an unweighted average.

3.2 Main Results

Figure 1 contains the four aggregated intervals (black lines), with `avg` being unweighted averages and `M1, M2, M3` being the model implementations. It can be seen that, as compared to the experts' reported intervals, the aggregated intervals are consistently narrower and closer to the truth. The model-based aggregations also result in narrower intervals, as compared to the simple average across experts. Comparing the three aggregated intervals with one another, it can be seen that the (M2) intervals are centered at the true value for Items 1–4. However, for items where the ground truth was not known (items 5–8), the intervals are very similar for all three methods. This implies that the model did not benefit from "knowing" the ground truth on some items.

We now focus on statistical summaries of Figure 1. Table 1 displays four statistics for each expert and for the aggregated intervals. The first is the number of intervals (out of eight) that covered the truth. The other four are ratio measures, comparing each expert/model to the model-based aggregations with no ground truth (M1). These measures include the average ratio of interval widths, defined as:

$$\text{Width} = \frac{1}{8}\sum_{j=1}^{8} \frac{(X_{ij3} - X_{ij1})}{(\widehat{X}_{j3} - \widehat{X}_{j1})}, \tag{6}$$

where i is fixed and the $\widehat{X}$ are aggregated quantiles from M1. The next measure is a standardized bias ratio, defined as:

$$\text{Bias} = \frac{\sum_j |X_{ij2} - \mu_j^*|/\mu_j^*}{\sum_j |\widehat{X}_{j2} - \mu_j^*|/\mu_j^*}, \tag{7}$$

where μ_j^* is the ground truth. The last measure is the average likelihood of the ground truth under each expert's interval (assuming normality), relative to the sum of likelihoods under the expert's interval and under the aggregated interval from M1:

$$\text{LR} = \frac{1}{8}\sum_{j=1}^{8} \frac{\phi((\mu_j^* - X_{ij2})/\sigma_{ij})}{\phi((\mu_j^* - X_{ij2})/\sigma_{ij}) + \phi((\mu_j^* - \widehat{X}_{j2})/\widehat{\sigma}_j)}, \tag{8}$$

where $\phi()$ is the standard normal density function. This is a measure of the relative likelihood of the truth for the expert's intervals, as compared to the M1 intervals.

For the width and bias measures described above, values less than one imply that the expert (or other aggregation) did better than M1 and values greater than one imply that the expert (or aggregation) did worse than M1. The likelihood measure, on the other hand, must lie in (0,1). Values less than .5 imply that the expert did worse than M1, and values greater than .5 imply the converse.

Table 1 contains the statistics described above for each expert and for the aggregations. Statistics for M2 are not included because its estimated intervals are centered at the ground truth for the first four items (leading to inflated performance measures). Examining the individual expert judgments, the table shows that no expert beats M1 on all four measures. Expert 8 is closest: her intervals are wider than the model, but she covers more items, exhibits less bias, and exhibits greater truth likelihoods.

Examining the aggregations, M1 beats the unweighted average on both interval width and bias. The two aggregation methods are similar on coverage and

Table 1. Statistics for individual experts' judgments and model-based, aggregated judgments. "Coverage" is number of intervals that cover the truth (out of 8), "Width" is ratio of of mean interval width to that of M1, "Bias" is ratio of mean standardized, absolute bias to that resulting from M1, and "LR" is the likelihood of the truth for each interval relative to M1.

Expert	Coverage	Width	Bias	LR
e1	3	2.8	1.5	0.28
e2	6	2.1	2.1	0.53
e3	3	1.3	2.3	0.19
e4	4	4.1	8.4	0.29
e5	3	1.9	2.3	0.45
e6	2	1.7	1.9	0.34
e7	4	1.2	1.7	0.53
e8	7	1.8	0.9	0.61
e9	1	1.0	1.7	0.39
e10	4	1.1	0.4	0.36
avg	4	1.9	2.2	0.55
M1	4	–	–	–
M3	2	0.9	1.1	0.26

likelihood. Comparing M1 with M3, M3 is worse on coverage and likelihood ratio. These findings, which generally match the visual results in Figure 1, imply that having access to the ground truth for some items did not improve model estimates.

4 Conclusions

The psychological model for aggregating magnitude judgments described in this paper exhibits good statistical properties, including an ability to consistently outperform expert judgments and judgments from simple aggregation methods. Unexpectedly, the model was unable to benefit from the ground truth (though see [13] for a similar result in a different context). We view this to be related to the model assumption that experts consistently over- or underestimate the ground truth across items. This assumption is violated if, say, experts are very knowledgeable about some items but not others. If the assumption is violated, then the model cannot obtain good estimates of expert biases because the biases are changing across items. This leads to the disutility of the ground truth. A potential solution would involve the addition of "item saliency" parameters into the model, though we may quickly find that the model is overparameterized.

While these and other extensions may be considered, the model developed here exhibits reasonable performance and has not been tailored to the content area in any way. It may generally be applied to domains where magnitude judgments are utilized, and it is very flexible on the number of quantile judgments elicited, the presence of ground truth, and the extent to which all experts judge all items. The model also utilizes interpretable parameters, allowing the analyst to maintain some control over the model's behavior. This feature may be especially useful in situations where the costs of incorrect judgments are high. More generally, the model possesses a unique set of features that give it good potential for future application.

References

1. Surowiecki, J.: The wisdom of crowds. Anchor, New York (2005)
2. Clemen, R.T., Winkler, R.L.: Combining probability distributions from experts in risk analysis. Risk Analysis 19, 187–203 (1999)
3. O'Hagan, A., Buck, C.E., Daneshkhah, A., Eiser, J.R., Garthwaite, P.H., Jenkinson, D.J., Oakley, J.E., Rakow, T.: Uncertain judgements: Eliciting experts' probabilities. Wiley, Hoboken (2006)
4. Morris, P.A.: Combining expert judgments: A Bayesian approach. Management Science 23, 679–693 (1977)
5. Genest, C., Zidek, J.V.: Combining probability distributions: A critique and annotated bibliography. Statistical Science 1, 114–148 (1986)
6. Kahn, J.M.: A generative Bayesian model for aggregating experts' probabilities. In: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, pp. 301–308 (2004)
7. Cooke, R.M.: Experts in uncertainty: Opinion and subjective probability in science. Oxford University Press, New York (1991)
8. Batchelder, W.H., Romney, A.K.: Test theory without an answer key. Psychometrika 53, 71–92 (1988)
9. Karabatsos, G., Batchelder, W.H.: Markov chain estimation for test theory without an answer key. Psychometrika 68, 373–389 (2003)
10. Yaniv, I.: Weighting and trimming: Heuristics for aggregating judgments under uncertainty. Organizational Behavior and Human Decision Processes 69, 237–249 (1997)
11. Cooke, R.M.: Experts in uncertainty: Opinion and subjective probability in science. Oxford, New York (1991)
12. Cooke, R.M., Goossens, L.L.H.J.: TU Delft expert judgment data base. Reliability engineering and system safety 93, 657–674 (2008)
13. Dani, V., Madani, O., Pennock, D., Sanghai, S., Galebach, B.: An empirical comparison of algorithms for aggregating expert predictions. In: Proceedings of the Twenty-Second Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI 2006), Arlington, Virginia, pp. 106–113. AUAI Press (2006)