Using Rater Confidence/Certainty

Melissa Wong

January 17, 2020

In MAADCAP, raters assign a confidence to each score entered. The rater can choose between three confidence levels: Low, Medium or High. However, MAADCAP analytics do not currently make use of this information. Note that this "rater confidence" information is distinctly different from the statistical meaning of confidence. Therefore to avoid confusion, through this document the Low/Medium/High scores will be referred to as "rater certainty". This document proposes an approach for incorporating the rater certainty information into the analytics.

1 Combining Rater Certainty

First, we are interested in combining individual Rater Certainty levels to get an overall certainty level. Note that this overall certainty level is distinct from consensus. Consider the examples in the table below. Case 1 there is high consensus among the scores, but the overall certainty is low. Case 2 there is low consensus among the scores, and the overall certainty is what? A naive approach might be to "average" the certainty levels and conclude the overall certainty is Medium. However, as we will see such a conclusion is misleading.

Case 1		Case 2	
Score	certainty	Score	certainty
1	Low	1	High
1	Low	3	Med
1	Low	5	Low

The certainty levels are categorical which can be represented by a multinomial distribution:

$$p(x_{Low}, x_{Med}, x_{High}) \propto \prod_{i=Low}^{High} \theta_i^{x_i}$$

where the $\theta_i^{x_i}$ are the true, but unknown, probability for each category and the x_i are the observed counts per certainty category. The Dirichlet distribution is the conjugate prior for the multinomial distribution:

$$p(\theta_{Low}, \theta_{Med}, \theta_{High}; \alpha_{Low}, \alpha_{Med}, \alpha_{High}) \propto \prod_{i=Low}^{High} \theta_i^{\alpha_i - 1}$$

If we choose a non-informative prior ($\alpha_{Low} = \alpha_{Med} = \alpha_{High} = 1$), then the posterior distribution of the θ_i is $Dirichlet(1 + x_{Low}, 1 + x_{Med}, 1 + x_{High})$.

From the posterior distribution, we can then estimate the probability and a credible interval for each certainty category. This approach can be applied to estimate certainty on an individual question and/or over all questions. Finally, this approach can easily be extended if the number of certainty categories is increased.

Surface plots of the non-informative prior and overall certainty for Case 1 are shown in Figure 1.

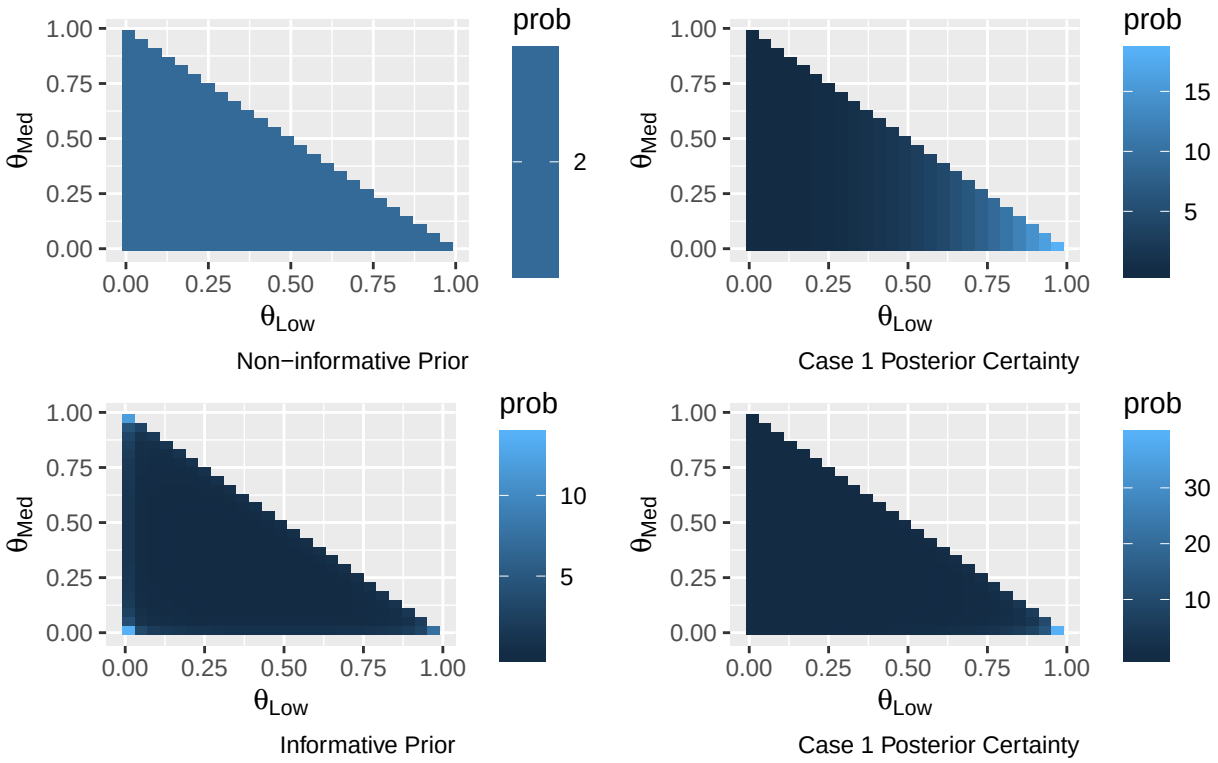
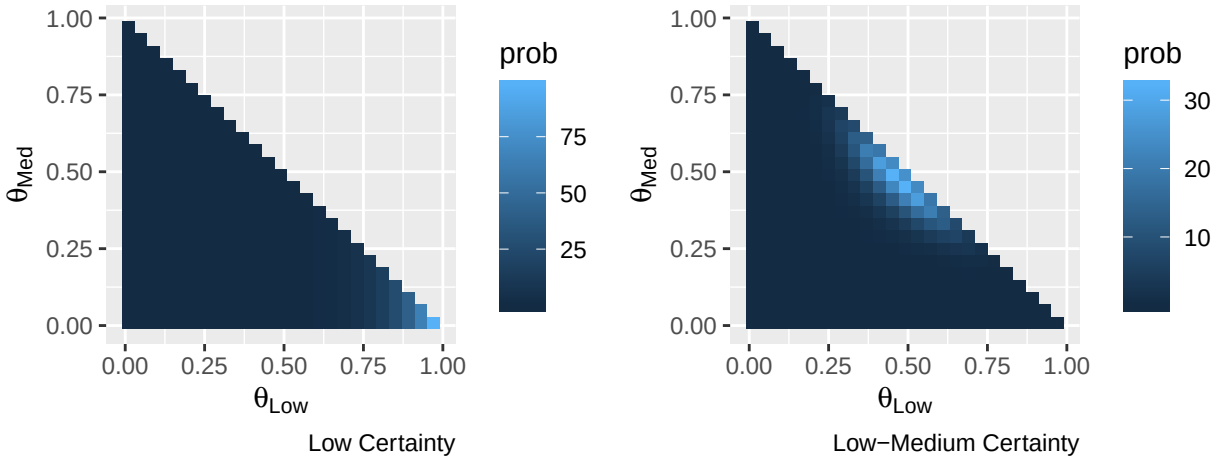


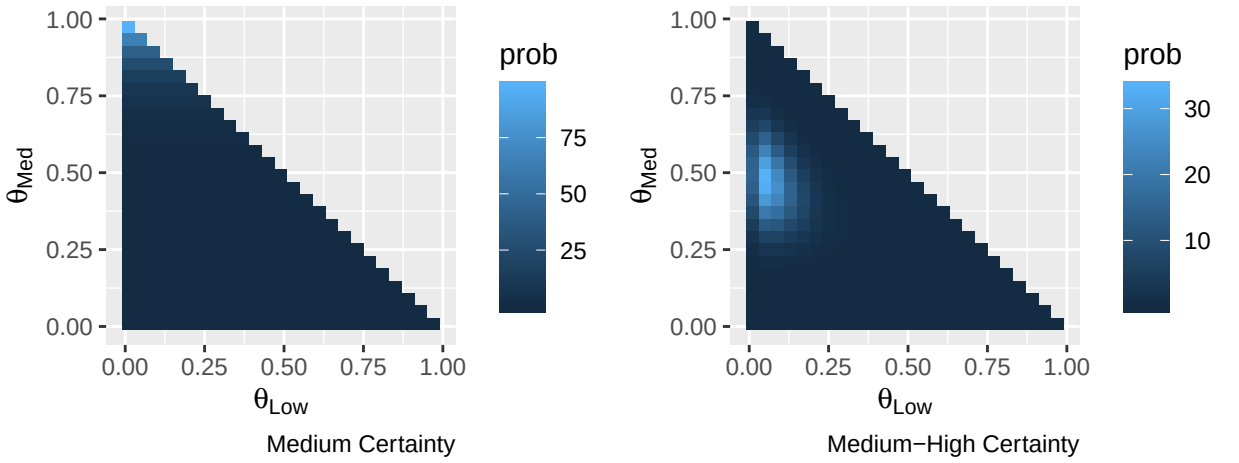
Figure 1: Case 1 Prior and Posterior Certainty

1.1 Visualizing Rater Certainty

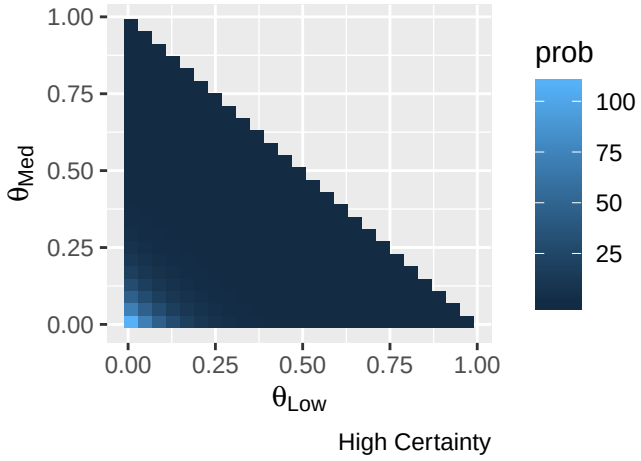
If the majority of the observations are Low certainty, the overall certainty level concentrates around the bottom right corner as the number of observations increases ($\theta_{Low} \to 1, \theta_{Med} = \theta_{High} \to 0$). On the other hand, an equal number of Low and Medium scores, with comparatively few High certainty scores, will be concentrated along the upper boundary on the plot.



Similary, a majority of Medium observations concentrates around the top left ($\theta_{Med} \to 1, \theta_{Low} = \theta_{High} \to 0$). An equal number of High and Medium scores, with comparatively few Low certainty scores, will be concentrated along the Y-axis.



And finally, a majority of High observations concentrates the distribution around the origin ($\theta_{High} \to 1, \theta_{Low} = \theta_{Med} \to 0$).



1.2 Ambiguous Confidence

Returning to Case 2, where there were an equal number of Low, Medium and High certainty scores, we can see from the plot below that the distribution is quite different from a Medium certainty distribution. In this case, one cannot draw a conclusion about the most likely overall certainty level.

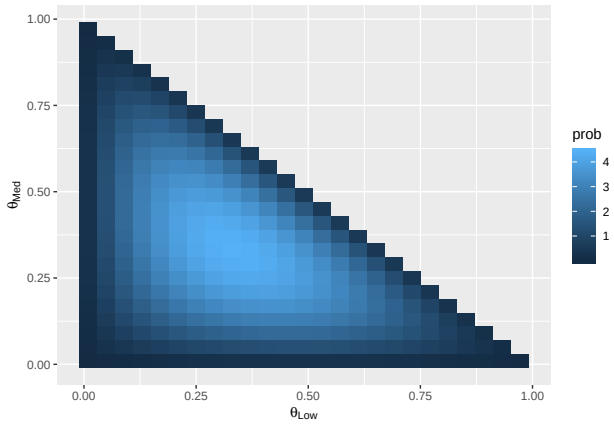


Figure 2: Case 2 Confidence

Similarly, the overall certainty is ambiguous when there are an approximately equal number of Low and High observatons and comparatively few Medium observations. The distribution will be concentrated around the center of the X-axis. Again this situation is not equivalent to a naive "averaging" of Low and High observations to conclude overall certainty is Medium.

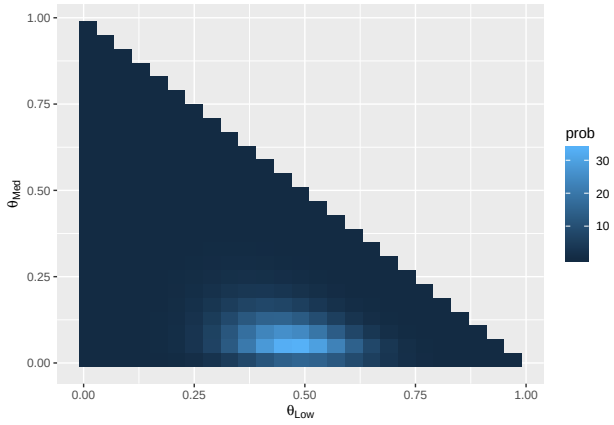


Figure 3: Ambiguous Confidence

2 Using Certainty in Consensus/Agreement

Currently, rater certainty is not incorported when calculating Tastle's Consensus or Agreement. In other words, a Low certainty score is treated the same as a High certainty score when calculating those statistics. DeGroot (DeGroot 1974) proposed using a stochastic model for reaching a consensus among subjective ratings. DeGroot's method can be adapted for MAADCAP to incorporate the certainty into the Consensus/Agreement statistics.

To use DeGroot's method, we need to assign numerical weights to the categories. For illustration, let Low = 0.2, Med = 0.5 and High = 0.7. Further, assume remaining weights are equally distributed among other raters (i.e., modeling a Delphi process as described by DeGroot).

2.1 Case 1

Consider the following scores and certainty levels.

Score	Confidence
1	Med
2	Med
3	Med
4	Med
5	Med

The one-step transtion matrix, P, is

0.50	0.12	0.12	0.12	0.12
0.12	0.50	0.12	0.12	0.12
0.12	0.12	0.50	0.12	0.12
0.12	0.12	0.12	0.50	0.12
0.12	0.12	0.12	0.12	0.50

We can then solve for the steady state weights using the following system of equations

$$\boldsymbol{\pi} = \boldsymbol{\pi} P$$
$$\sum_{i=1}^{n} \pi_i = 1$$

and the resulting steady state weights are

$$\boldsymbol{\pi} = [0.2, 0.2, 0.2, 0.2, 0.2]$$

The steady state weights are identical which is expected since all the raters had the same certainty in their scores. Tastle's Consensus (0.43) and scaled Tastle's Agreement around median (0.51) are thus calculated as usual (see "Summary of MAADCAP Statistics" for details).

Note: If all raters choose the same certainty level *and* we follow the rule of equally distributing remaining weight, the one-step transition matrix will always be doubly stochastic and ergodic. Under these conditions, the steady state probabilities will be $\pi_1 = \pi_2 = ... = \pi_n = \frac{1}{n}$ where n is the number of raters regardless of the numerical values chosen for Low/Med/High.

2.2 Case 2

Now consider the following:

Score	Confidence
1	Low
2	High
3	Med
4	Low
5	Med

The one-step transition matrix, P, is

0.20	0.20	0.20	0.20	0.20
0.08	0.70	0.08	0.08	0.08
0.12	0.12	0.50	0.12	0.12
0.20	0.20	0.20	0.20	0.20
0.12	0.12	0.12	0.12	0.50

and the resulting steady state weights are

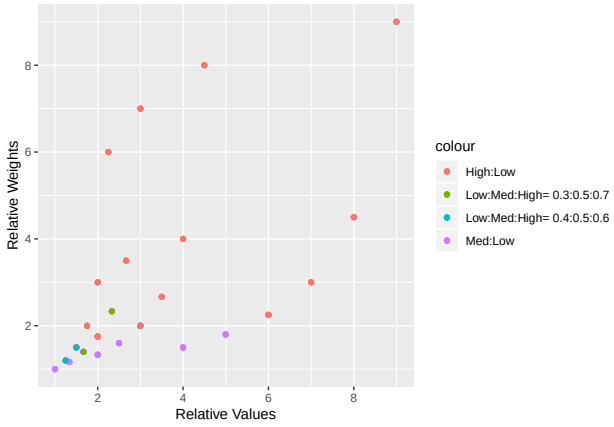
$$\boldsymbol{\pi} = [0.13, 0.34, 0.2, 0.13, 0.2]$$

Now the scores are not equally weighted; instead there is more weight placed on the scores with Medium and High certainty. Further, we expect the Consensus and Agreement to be higher since they are calculated about the mean and median respectively (3 in both cases for this example). Using these new weights results in both Tastle's Consensus (0.47) and scaled Tastle's Agreement (0.55) consistent with expectations.

It should be noted that this approach implicitly assumes the each certainty level means the same thing to every rater. In reality, that is probably not true. However, in absence of any information about consistency of raters' interpretation of the certainty levels, DeGroot's method gives a quantitative and repeatable way to make use of the certainty information versus ignoring it altogether. In the future, MAADCAP might allow each user to specify his/her individual numerical interpretation of the categories or even do away with the certainty categories altogether and instead select a certainty level on a scale from 0 to 1 for each question. Either option can still be used with DeGroot's method for reaching consensus. However, one would need to evaluate whether the additional burder on the rater is worth the potential improvement.

2.3 Sensitivity Analysis

Since DeGroot's method requires assigning a numerical value to the Low/Med/High categories, a natural question then is how to choose the values if they are not specified by the rater. The plot below shows the effect on relative weights depending on the relative values chosen for each category. Unsurprisingly, choosing extreme values for Low (0.1) and High (0.9) results in an extremely large High:Low relative weight (~9). Choosing either Low:Med:High = 0.4:0.5:0.6 or Low:Med:High = 0.3:0.5:0.7 yields relative weights between 1.2 and 2.33. This is a reasonable choice since it permits differentiation between the certainty levels without extreme relative weight differences.



References

DeGroot, Morris H. 1974. "Reaching a Consensus." *Journal of the American Statistical Association* 69 (345): 118–21.