Using Cognitive Models to Improve the Wisdom of the Crowd

Michael D. Lee (he/him)
mdlee@uci.edu

University of California Irvine
The wisdom of the crowd is the idea that aggregating individual behavior leads to good group decisions (Surowiecki, 2004)

- **Signal and noise**: averaging estimates amplifies common signal and cancels (independent) noise
- **Jigsaw puzzle**: combining diverse individual knowledge can piece together a more complete group answer
Cognitive Modeling and the Wisdom of the Crowd

- The wisdom of the crowd is usually treated as a statistical problem.
  - We treat it as a **cognitive modeling problem** because the data being aggregated are the outputs of cognitive processes.
- Cognitive models can improve the wisdom of the crowd by:
  - Making inferences about **expertise** to boost the signal.
  - **Debiasing** of inherently biased cognitive processes.
  - Providing a **representational substrate** to combine diverse knowledge.
  - Generating **predictions** about how people would have responded, in the absence of their behavioral data, to preserve individual differences.
Amplifying Expertise in Rankings

(Lee, Steyvers, de Young, & Miller, 2012; Lee, Steyvers, & Miller, 2014; Selker, Lee, & Iyer, 2017)
Lee et al. (2014) collected 78 people’s rankings of the 10 largest European cities in order of population. A standard statistical aggregation is the Borda Count - sums the rank of each item across people, and orders the items according to those sums. A cognitive model of ranking is provided by Thurstone models (Thurstone, 1927; Böckenholt, 2006).

<table>
<thead>
<tr>
<th>Rank</th>
<th>Person 1</th>
<th>Person 2</th>
<th>Person 3</th>
<th>...</th>
<th>Sum</th>
<th>Borda Aggregate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>London</td>
<td>London</td>
<td>Budapest</td>
<td></td>
<td>1 + 1 + 3 + ... = 289</td>
<td>London</td>
</tr>
<tr>
<td>2</td>
<td>Berlin</td>
<td>Berlin</td>
<td>Paris</td>
<td></td>
<td>4 + 8 + 2 + ... = 335</td>
<td>Paris</td>
</tr>
<tr>
<td>3</td>
<td>Madrid</td>
<td>Madrid</td>
<td>London</td>
<td></td>
<td>2 + 2 + 9 + ... = 339</td>
<td>Berlin</td>
</tr>
<tr>
<td>4</td>
<td>Rome</td>
<td>Rome</td>
<td>Warsaw</td>
<td></td>
<td>4 + 4 + 6 + ... = 366</td>
<td>Rome</td>
</tr>
<tr>
<td>5</td>
<td>Paris</td>
<td>Warsaw</td>
<td>Hamburg</td>
<td></td>
<td>3 + 3 + 8 + ... = 371</td>
<td>Madrid</td>
</tr>
<tr>
<td>6</td>
<td>Warsaw</td>
<td>Budapest</td>
<td>Rome</td>
<td></td>
<td>8 + 9 + 5 + ... = 461</td>
<td>Hamburg</td>
</tr>
<tr>
<td>7</td>
<td>Bucharest</td>
<td>Bucharest</td>
<td>Bucharest</td>
<td></td>
<td>9 + 10 + 10 + ... = 514</td>
<td>Vienna</td>
</tr>
<tr>
<td>8</td>
<td>Hamburg</td>
<td>Paris</td>
<td>Madrid</td>
<td></td>
<td>10 + 6 + 1 + ... = 519</td>
<td>Budapest</td>
</tr>
<tr>
<td>9</td>
<td>Vienna</td>
<td>Hamburg</td>
<td>Berlin</td>
<td></td>
<td>6 + 5 + 4 + ... = 538</td>
<td>Warsaw</td>
</tr>
<tr>
<td>10</td>
<td>Budapest</td>
<td>Vienna</td>
<td>Vienna</td>
<td></td>
<td>7 + 7 + 7 + ... = 558</td>
<td>Bucharest</td>
</tr>
</tbody>
</table>
Thurstone Model

Latent Ground Truth
Thurstone Model

Latent Ground Truth

Expert

Observed Ranking \( y_1 = (1, 2, 3, 4, 5) \)
Thurstone Model

Latent Ground Truth

Expert

Observed Ranking $y_1 = (1, 2, 3, 4, 5)$

Intermediate

Observed Ranking $y_2 = (1, 2, 3, 5, 4)$
Thurstone Model

Latent Ground Truth

Expert

Intermediate

Novice

Observed Ranking $y_1 = (1, 2, 3, 4, 5)$

Observed Ranking $y_2 = (1, 2, 3, 5, 4)$

Observed Ranking $y_3 = (2, 1, 3, 5, 4)$
Thurstone Model Inference

- Implemented the model in JAGS to allow for fully Bayesian inference using computational sampling (Plummer, 2003)
- Joint posterior distribution for $\mu$ provides a wisdom of the crowd aggregate, summarized by marginal posteriors graphically.
Tau Distance Measure

- Tau distance measures the number of pair-wise swaps needed to convert one ranked list into another (Kendall, 1938)
  - Thurstone aggregate is $\tau = 9$ from the truth
  - Borda aggregate is $\tau = 11$ from the truth

<table>
<thead>
<tr>
<th>Borda</th>
<th>Thurstone</th>
<th>Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>London</td>
<td>London</td>
<td>London</td>
</tr>
<tr>
<td>Berlin</td>
<td>Berlin</td>
<td>Berlin</td>
</tr>
<tr>
<td>Madrid</td>
<td>Rome</td>
<td>Madrid</td>
</tr>
<tr>
<td>Rome</td>
<td>Paris</td>
<td>Rome</td>
</tr>
<tr>
<td>Paris</td>
<td>Madrid</td>
<td>Paris</td>
</tr>
<tr>
<td>Warsaw</td>
<td>Vienna</td>
<td>Bucharest</td>
</tr>
<tr>
<td>Budapest</td>
<td>Hamburg</td>
<td>Hamburg</td>
</tr>
<tr>
<td>Bucharest</td>
<td>Budapest</td>
<td>Warsaw</td>
</tr>
<tr>
<td>Hamburg</td>
<td>Bucharest</td>
<td>Budapest</td>
</tr>
<tr>
<td>Vienna</td>
<td>Warsaw</td>
<td>Vienna</td>
</tr>
</tbody>
</table>
• The main panel show the tau distances for
  - the individual participants in yellow
  - the distribution of Thurstone model answers in blue
  - the Thurstone and Borda count answers as dotted lines

European City Populations
  78 people
  10 items
The inset panel shows the relationship between the posterior mean of the inferred individual expertise $\sigma_j$ and the $\tau_j$ accuracy measure for each participant. - there is a strong predictive correlation between the two.
Diversifying Expertise in Ranking

(Lee, Bradford, & Tejeda, 2022; Montgomery & Lee, in preparation)
Recent Events

- Participants chose a subset of items to rank
  - instructions were to “choose as many amendments as you can, while remaining confident you have put them in the correct order”
Thurstone Model of Subset Ranking

Latent Ground Truth

Person 1

Observed Ranking
$y_1 = (1, 2, 5)$

Person 2

Observed Ranking
$y_2 = (2, 5, 4)$

Person 3

Observed Ranking
$y_3 = (1, 3, 5)$
The ten amendments question was answered by 203 participants:
- modal number of items ranked was four
- about 60% of participants ranked five or fewer items
Spatial Estimation

(Montgomery & Lee, 2022)
• Montgomery, Baldini, Vanderkerckhove, and Lee (2023) asked people to estimate the location of US cities
  - they first provided a point estimate
  - then provided a radius centered on the point estimate large enough to be confident the circle contained the city
Model-based methods improved the wisdom of the crowd performance of statistical aggregation
- model-based performance was best when expertise was allowed to vary both by person and by city
• Asked participants to select all the countries that
  - could be Ethiopia
  - could not be Ethiopia
• Often evidence of framing effects, with greater uncertainty in the absence framing
• Statistical aggregation can piece together diversity of geographic knowledge to identify all of the countries
• Model-based aggregation promises to improve on this by
  - identifying and upweighting region-specific expertise
  - accounting for individual differences in uncertainty management
  - incorporating *within-individual* information provided by the different framings as well as between-individual information
Probability Estimation

(Lee & Danileiko, 2014)
A total of 145 people answered 40 probability/percentage questions about football (soccer)
- “what is the probability that a team that is ahead at the 5th minute will win?” [59%]
- answers calculated from about 6000 first-division professional games
Expertise and De-Biasing

- The model assumes that people experience the true probability with an accuracy that depends on their expertise
  - the experienced probability is $\pi'_{ij} \sim \text{Gaussian}(\pi_i, 1/\sigma_j^2)$
    - where $\pi_i$ is the truth and $\sigma_j$ is the expertise
- People then estimate their knowledge with some level of miscalibration (e.g., Lichtenstein et al., 1978)
  - the miscalibrated estimate is the linear-in-log-odds $\psi_{ij} = \delta_j \log \left( \frac{\pi'_{ij}}{1-\pi'_{ij}} \right)$ for a calibration $\delta_j \sim \text{beta}(5, 1)$
Model Results

- The model’s inferences form a group aggregate that is more accurate than the mean or median
  - the relative expertise of individuals is inferred without access to the ground truth
Inferring Expertise

- Inferred expertise is a much better predictor of accuracy than self-reported expertise or a trivia test
The model helps debias the underweighting of large probabilities and (especially) the overweighting of small probabilities.
Competitive Bidding

(Lee, Zhang, & Shi, 2011)
Combining the “Wisdom” on the Price is Right

- Four players bid in order to win a prize
  - the closest to the true price of the prize \textit{without going over} wins
  - based on data from 200+ games, players often make strategic bids of $1, or $1 more than another bid
Infer a group estimate using a decision model of how people convert their knowledge into bids, assuming people bid to maximize their probability of winning.

This aggregate outperforms:
- single bids, and statistical aggregates of bids
- model-based aggregates that do not account for strategic bidding
Category Learning

(Danileiko & Lee, 2017)
• In category learning, simple statistical aggregation based on the majority response for each stimulus provides a good crowd performance
  - but does not generalize beyond the stimuli for which behavioral data are available
• Models of individual decision making can be learned from available behavioral data and then make predictions about how the individual would categorize unseen stimuli
  - the majority of these predicted decisions provides a model-based wisdom of the crowd
Zeithamova and Maddox (2006) studied a task in which people learned two categories of Gabor patch stimuli from feedback. Use General Recognition Theory to infer a decision bound for each participant (Ashby & Gott, 1988), both quantitative and qualitative individual differences in the decision strategies.
• The individual strategies can then classify new stimuli generated from the categories.
The same approach applies to a more difficult information-integration category learning task presented by Zeithamova and Maddox (2006).
The model-based crowd generalizes the good performance of the mode to new stimuli for which behavioral data do not exist.
Sequential Tasks

(Thomas, Coon, Westfall, & Lee, 2021)
In bandit problems people repeatedly choose between options that have fixed but known rates of reward:
- goal is to maximize total reward
- requires an early exploration strategy that matures to a late exploitation strategy
Thomas et al. (2021) consider 42 people completing an 8-trial bandit problem
- rewarded between 50–64% of trials
The behavior-based majority performs extremely poorly
- about as bad as the worst participant, and 10% below optimal
Maladaptive Crowd Behavior

733 choices vs 319 choices

337 vs 141

235 vs 74

129 vs 28

474 vs 128

221 vs 49

158 vs 40

167 vs 15
Inferring Decision Models

- To maintain the contribution of the entire crowd, infer individual-level models from observed behavior
- Modeling inference incorporates
  - **qualitative** differences in strategy over a set of psychological and reinforcement learning models (Sutton & Barto, 1998)
  - **quantitative** inferences about the parameterization of the specific strategy
The individual models make predictions about what each person would have chosen in the crowd’s current game state - the crowd decision is the majority of these predicted choices.

The model-based crowd matches optimal performance.
Summary

amplify the expertise signal

piece together the jigsaw

debias the signal

maintain crowd diversity

ranking probability estimation

subset spatial estimation

probability estimation competitive bidding

category sequential problems


The Balloon Analogue Risk Task (BART: Lejuez et al., 2002) is a sequential task where on each trial people can:
- pump a balloon to increase its value at the risk of it bursting and losing all value
- bank the current value

A statistical wisdom the crowd approach would be to follow the majority pump-versus-bank decision, but this is fragile:
- conservative “bank early” people drop out of the crowd
- the crowd progressively contains only risk takers, potentially leading to many burst balloons
In this example, the crowd banks at 7 pumps, but only just
Debiasing Anchored Estimates

- Lee, Villarreal, and Montgomery (2023) debiased anchored estimates
  - using the same approach as for debiasing probability estimates
  - assumes an anchoring effect of around 40%
Debiasing Anchored Estimates

- The model-based approach improves statistical aggregation when the anchoring effect occurs.
• The parameters of the Thurstone model are
  - stimulus location parameters $\mu_i$ for the $i$th item, indicating their true position on an assumed underlying single dimension corresponding to the ranking criterion
  - individual expertise parameters $\sigma_j$ for the $j$th person, indicating how accurately they represent information about each stimulus
• The cognitive process is that when the $j$th person wants to rank the stimuli, they take a mental sample $x_{ij}$ for each of the $i = 1, 2, \ldots$ stimuli
• The mental sample is drawn from a Gaussian distribution centered on the true location of the item, with a standard deviation determined by the expertise of the person
  \[
  x_{ij} \sim \text{Gaussian}(\mu_i, \sigma_j^2)
  \]
• The order of the mental samples then determines the ranking $y_j$ the person produces
  \[
  y_j \sim \text{rank}(x_j)
  \]
model{
    for (j in 1:nItems)
        muTmp[j] ~ dunif(-100,100)
        mu[j] = muTmp[j] - mean(muTmp)
    }
    for (i in 1:nPeople)
        sigmaTmp[i] ~ dunif(0,1)
        sigma[i] = sigmaTmp[i] / mean(sigmaTmp)
    }
    for (i in 1:nData)
        for (j in 1:nItems)
            x[i,j] ~ dnorm(mu[j],1/sigma[person[i]]^2)
    }
    for (j in 1:setN[i])
        y[i,set[i,j]] ~ dinterval(x[i,set[i,j]], ...
            sort(x[i,set[i,1:setN[i]]]))