Rank Aggregation and Belief Revision Dynamics

Igor Volzhanin (ivolzh01@mail.bbk.ac.uk), Ulrike Hahn (u.hahn@bbk.ac.uk), Dell Zhang (dell.z@ieee.org)
Birkbeck, University of London
London, WC1E 7HX UK

Stephan Hartmann (s.hartmann@lmu.de)
Munich Center for Mathematical Philosophy, LMU München
Geschwister-Scholl-Platz 1, D-80539 München, Germany

Abstract

In this paper, we compare several popular rank aggregation methods by the accuracy of finding the true (correct) ranked list. Our research reveals that under most common circumstances simple methods such as the average or majority actually tend to outperform computationally-intensive distance-based methods. We then conduct a study to compare how actual people aggregate ranks in a group setting. Our finding is that individuals tend to adopt the group mean in a third of all revisions, making it the most popular strategy for belief revision.

Keywords: rank aggregation; distance measure; probabilistic model

Introduction

The problem of *rank aggregation*, where ranked lists from a diverse set of "judges" are combined into a single "consensus" ranked list, is an active research area in computer science. Particularly, rank aggregation has found successful applications in meta-search (Dwork, Kumar, Naor, & Sivakumar, 2001; Renda & Straccia, 2003; Fernández, Vallet, & Castells, 2006), crowd-sourcing (Niu et al., 2015), and recommender systems (Baltrunas, Makcinskas, & Ricci, 2010).

Although extensive studies have already been conducted on this topic by computer scientists, these largely concern only the algorithmic issues, i.e., how to produce the "optimal" ranked list, without questioning the very concept of "optimal". Typically, a distance measure is chosen, and the ranked list with the minimum total distance to all the given ranked lists is presumed to be the best one (Dwork et al., 2001). In this paper, we challenge such a view and address the problem from the perspective of *cognitive science*. Just as importantly, much of the previous research has been theoretical in nature and no empirical work has been conducted to determine how humans actually aggregate ranks. To that end, we went beyond the theoretical models described in section 1 and conducted a group study to better understand real-world rank belief revision. To the best of our knowledge, there has been no similar work to date.

Modeling

In the first instance, we developed a theoretical simulation to test the accuracy of various rank aggregation methods. The simulation can be thought of in terms of the most preferred order in which to display results of a web search.

Given a set of m items (e.g., web pages), we consider n ranked list of them, $\{r_1, \ldots, r_n\}$, each of which is given by a judge (e.g., search engine). One, and only one, of the possible ranking orders (permutations) r_* is deemed to be true (correct).

Each judge is characterised by his "competence" which is defined as the probability of providing the true list.

Our simulation takes the various generated lists and aggregated them into a single list using one of the rules outlined further down in this section.

Unlike in some previous work, such as Fernández et al., for each item in a list we know only its rank position, vis-a-vis other items, and not any other numeric properties. It is often impossible or unrealistic to obtain the scores of individual items and only their relative positioning to each other is available (Dwork et al., 2001; Renda & Straccia, 2003). More importantly, a wealth of psychological research suggests that, in many domains, humans represent faithfully only ranking order information and more detailed information is unhelpful (Stewart, Chater, & Brown, 2006)

For the sake of simplicity our modeling considers that each judge will produce a complete list and no ties are possible. So when ranking items, they will rank all of the choices and will rank them relative to each other in such a way that each item will occupy a unique position. Furthermore, every judge in our model has the same level of competence $c \in [0,1]$. Finally, when a certain rule produces multiple lists that are equally optimal, one of them is selected at random. This work could be generalized straightforwardly in the future by relaxing these constraints.

Following rank aggregation methods have been proposed in previous studies and are widely used in practice, so we will use them in our comparison:

- **majority**: the consensus list is just the ranked list that appears most frequently.
- **average**: the consensus list is generated by ranking the items according to their average rank positions, which is essentially same as the Borda's count (Dwork et al., 2001).
- **Spearman**: the consensus list is the one with the minimum sum of Spearman's footrule to all the given ranked list. Spearman's footrule is defined as the total number of displacements needed to achieve parity between two lists.
- **Kendall**: the consensus list is the one with the minimum sum of Kendall's tau to all the given ranked list. Kendall's

tau is defined as the total number of inversions required to achieve parity between two lists.

- **Kemeny-Snell**: the consensus list is the one with the minimum sum of Kemeny-Snell distance to all the given ranked list. The Kemeny-Snell (KS) distance is similar to Kendall's tau, but more robust when dealing with ties.

While the first two methods are simple and easy to compute, the other three that are based on distance measures and have a high computational complexity. It has been shown that finding the optimally ranked list based on Kendall's tau (known as the Kemeney optimal aggregation) is an NP hard problem with just four full lists(Dwork et al., 2001).

Our research question is then: "which rank aggregation method is most accurate?" Here by accuracy, we mean how often the consensus list returned by a rank aggregation method is indeed the true list.

Computer Simulations

We prepared a simulation in *R*, which samples a number of judges and uses different aggregation methods to determine the list reflective of the group of judges. The generated consensus lists are then compared with the true list to calculate accuracy, which we used as our "performance" measure for the aggregation method. This procedure is repeated across pools of judges of different sizes. In order to smooth out effects of randomness, we performed bootstrapping at each number of judges and took the average value. Therefore each set of judges was simulated several times, before adding additional judges.

Our simulation had a number of parameters that could be altered:

- **list size**: number of unique items in a list
- **competence level**: individual probability of picking the correct list
- **aggregation method**: methods of aggregation described above
- **number of runs**: each run increased the number of judges in a group by one
- **number of simulations**: a number of repeats of the same simulation with the same conditions to smooth out any noise due to randomness

We began with a list size of 4. With no ties there are 24 possible permutations. In the simulation k groups consisting of n number of judges would draw a single list from the full list of permutations. Using one of the aggregation methods, a single list would be selected for each group as the aggregate product, and then compared to the true list. Each group of judges would be re-sampled a number of times to boostrap the results to get a smoother result. Thus, scores reported below are the average results sampled over multiple trials for the same group.

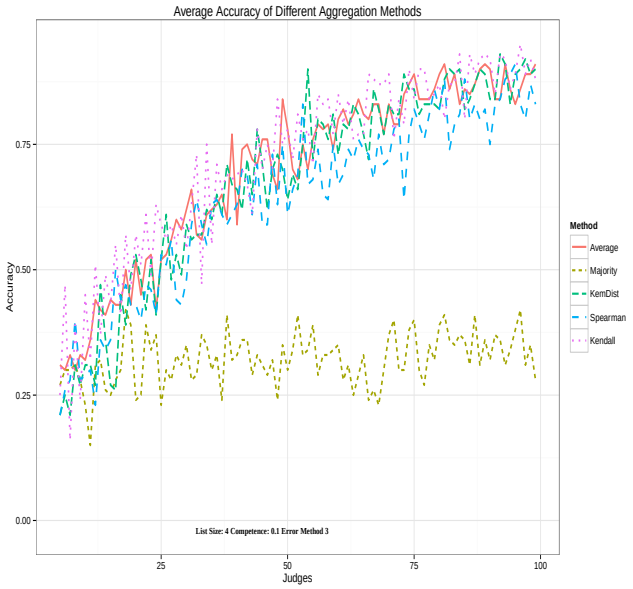

Figure 1: The comparison of aggregation methods in the linear-decay error model.

Error Model One important consideration in the study was the underlying error model that governed a judge's probability of picking the wrong list among all possible permutations. Each judge had a competence measure which reflected the probability of picking the true list. The rest of the probability was distributed among the remaining possible choices. Assuming that judges know anything about the domain in question, the probability of picking a wrong list is likely to be an inverse function of the distance from that list to the true list. Without loss of generality, we used the Kemeny-Snell distance measure $d(\cdot,\cdot)$ to determine the probability of a given list being selected as follows.

$$\Pr[r_i] = \begin{cases} c & \text{if } r_i = r_*, \\ (1-c)\frac{1/d(r_i,r_*)}{\sum_{j\neq *}\left(1/d(r_j,r_*)\right)} & \text{otherwise.} \end{cases} \quad (1)$$

In effect, lists that are closer to the true list, would be more likely to be drawn than the lists further away.

We wanted to see relative performance of the various aggregation methods, as the number of judges increased. For all results, we maintained a constant competence level $c = 0.1$, which meant a 10% chance for a judge to pick the true list r_*. We selected the simulation range from 5 to 100 judges.

Results

After running several different simulations we produced a number of interesting and insightful results. We present our findings in a series of figures that illustrate the relative performance of the different aggregation methods (see figures 1, 2, 3).

The majority rule performs significantly worse than the alternatives and does not increase in accuracy as the number of

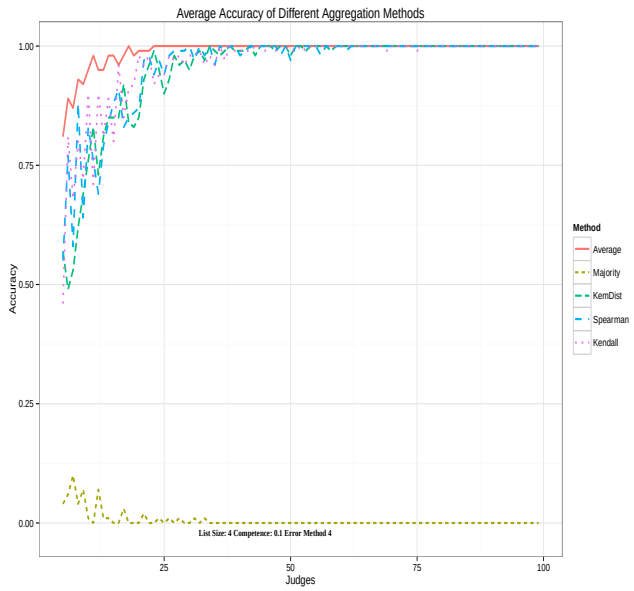

Figure 2: The comparison of aggregation methods in the fastest-decay error model.

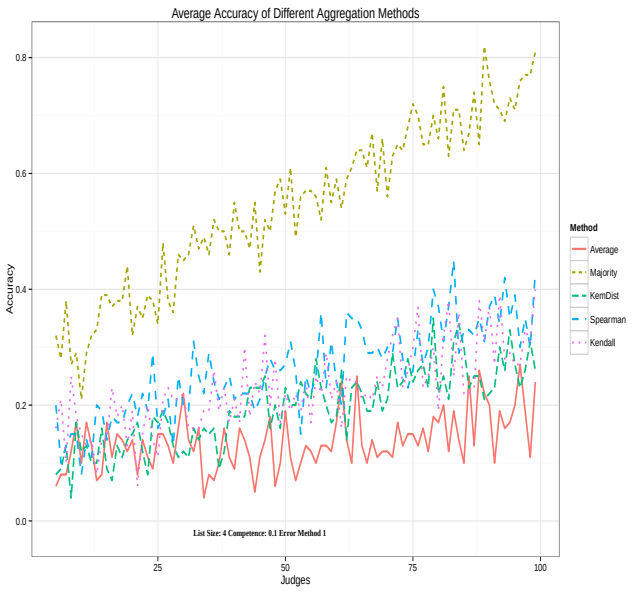

Figure 3: The comparison of aggregation methods in the none-decay error model.

judges increases. On the other hand, the other four methods perform similarly to each other and their accuracy increases as the number of judges goes up, as can be observed in figure 1. It is important to note that the Kemeny-Snell aggregation method does not perform significantly better than the other distance-based methods, despite the fact that the underlying error model is based on the Kemeny-Snell distance! Furthermore, average, which is a very simple method (both computationally and cognitively), performs at least on par with the distance-based methods.

A minor comment regarding high competence is due at this point. When the competence level is above a threshold, e.g., 0.2, we see a quick rise towards perfect accuracy of all methods, which is not particularly interesting, or informative. Therefore, we kept the competence level low and tried to understand how robust different rank aggregation methods would be under the more challenging condition of lower individual competence.

The above *linear-decay* error model as described in Eq. (1) is just one way of converting the underlying KS-distance to the targeted true list into a probability of erroneous list selection. Actually any monotonic decaying transformations – such as an exponential decay – of those distances could be utilised to pick the non-true lists. To generalise our results we considered two extreme cases of monotonic decay functions of distance: at the one end (*fastest-decay*), the selection probability drops so rapidly as a function of distance that only the closest lists stand a chance of being selected; at the other end (*none-decay*), the selection function is flat and the lists of all distances are equally likely to be selected. We have examined both of these extreme cases.

First, let us consider the case where only the ranked lists closest to the true list had a non-zero selection probability (with each list at that distance equally likely to be picked).

From the results of the simulation we see that the majority method plummets almost immediately towards zero accuracy. This is due to the fact that the competence level, i.e., the probability of picking the true list (10%), is significantly lower than the probability of picking any of those closest lists (which is 30% in this example as there are three closest lists in total).

Interestingly, the average method appears to outperform the other methods, and quickly moves towards perfect accuracy as the number of judges increases. There appears to be little difference among the other distance-based methods and they all behave similarly to the average method.

Second, we consider the case where a judge is equally likely to pick any of the wrong lists, regardless of their distance to the true list.

The results of this simulation stand in stark contrast to the other two simulations. The majority method performs significantly better and improves with the number of judges, which is exactly reverse of what was observed in the earlier simulations.

Although the observation was initially quite surprising, the explanation is fairly intuitive. Since the probability of picking the true list is 10%, the remaining probability would be distributed evenly over 23 other possible permutations, which leads to only 3.9% per permutation. Therefore, the ranked list occurred most frequently is almost guaranteed to be the true list, and the majority method would always perform the best.

Just as importantly, the other aggregation methods appear to falter at this stage. Although there is some improvement along with the increase in the number of judges, the

accuracy stays well below 0.5, even for groups with 100 judges. Notably, the average method performed the worst, while the Spearman method performed the best among the three distance-based methods.

Discussion

A few key insights emerge from our modeling efforts. The first and most important one is that there appears to be little benefit of using computationally-expensive distance-based methods to conduct rank aggregation. Secondly, there is clear robustness of adopting the group mean. Accuracy is constantly increases in most scenarios and the method itself is simple enough to calculate and act upon.

The one research question that remains open, however, is what real human subjects would do given a similar task. While it may appear that taking the group mean is advantageous from the accuracy point of view, it is also more difficult to determine than simply adopting the majority opinion for example. To test, this we designed a study that looked at individual rank revision in a group setting.

Experiment - Rank Revision

This experiment was set up to test what rules, if any, individuals use to revise their beliefs in light of new information. Unlike similar studies on the topic which have mostly looked at absolute answers and estimates, we were interested in applying this in the context of rank revision. In other words, our interest was to understand better how participants revise ranked orders when presented with information from their peers. From the modeling exercises above we knew that adopting the group mean is the most beneficial strategy a person can take in most situations, however, we could not locate any research that corroborated this in an empirical study.

Method

Participants Participants for this study were volunteers from the University of London community. Participants were paid 5 for taking part in the study. There were 19 participants who took part, which created three panels of five participants and one panel with four participants (n=19). Each group of participants took part at the same time and were hosted in the same room. No particular exclusion criteria were used and participants were free to self select which of the time slots worked best for them to attend the study. It did not appear that any participants knew each other prior to the study.

Materials & Procedure Participants were seated in a computer lab, spaced apart in a way that prevented them from seeing each others' screens. Each participant had a computer in front of them that contained a NetLogo interface that was connected in a network to other computers in the room. See Figure 4 for a sample interface that each participant saw.

Initially, participants were read basic instructions regarding the task. The task involved each participant to rank four cities from the largest to smallest by population size. Each city was presented in a text box and contained a number along with

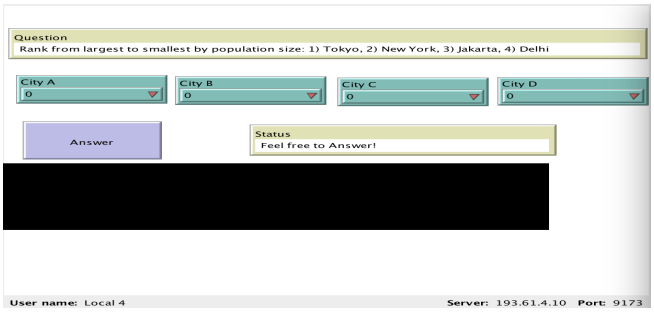

Figure 4: NetLogo participant interface

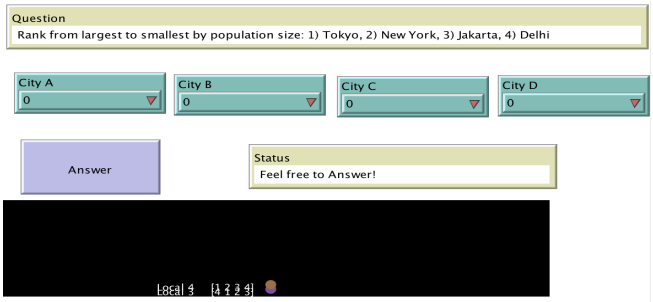

Figure 5: NetLogo participant interface

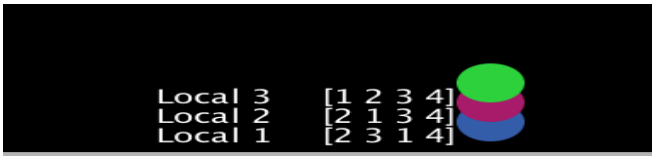

Figure 6: Zoomed in participant view

the name of the city (see example in figure 4). In the drop down box 'City A' they were instructed to put the number of the city they believed to be the largest, 'City B' were to contain the second largest, and so on. After all four boxed were filled, participants had to submit their answers and wait for everyone else in the room to finish. Once, all answers were submitted, participants could see how everyone else had ranked the cities. At this point, everyone had an opportunity to revise their answers in light of additional information (see figure 5 and zoomed in view in figure 6). They repeated this process three times for each question, resulting in four rounds - initial round, plus three revision rounds.

In total, each participant answered 21 questions. There is an initial practice question which participants did in a directed manner, followed by 20 other questions, which were done independently and free from any additional instructions. Each question contained a different set of cities and in different order, but the task was the same. There was only a single experimental condition and all participants were treated the same; they were shown the same set of questions, in the same order.

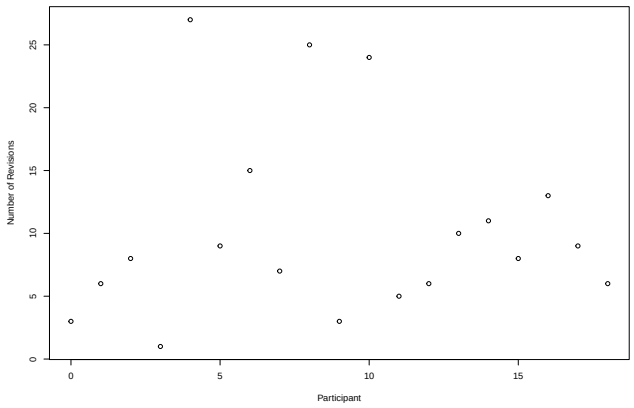

Figure 7: Number of revisions per Participant

Results

In the first instance we were interested in individual belief revision. We analyzed how often individuals changed their answers and what rules they have have used to do so.

Individual Revision Discounting the first question, there were 60 opportunities for revision for each participant (20 questions * 3 revision rounds). On average participants changed their answers 10.3 (SD 7.51) times over the course of the simulation, or about 10% of the time. With some participants changed their answers as little as once, and others changed almost a third of their answers. In total there were 196 revision for all participants. See figure 7 for a visual representation of the number of revisions per participant.

Overall, most revisions occurred in the first round, where almost as many revisions occurred as the subsequent two rounds. Table 1 breaks down revisions by round.

Revisions occurred unevenly between questions. Seven questions had between 13 and 15 revisions, while remaining 13 questions had between five and nine revisions.

The number of revisions made by participants was rather low, but the overall profile of the changes, i.e. mostly in the first round and more for some questions than others, is consistent with some of the other studies in the field of decision-making.

Table 1: Round Revisions

Revision Round 1	Revision Round 2	Revision Round 3
96	58	42

Models of Revision We fitted several models presented in the first part of the paper trying to predict individual belief revision rules that induced the change (such as mean, median and majority models). We decided to restrict our fitting to two models in particular: mean and majority. As these models had very interesting properties and were most likely to be available and calculable to participants. Since ranked lists were presented near each other identifying the majority list, or calculating the mean list was a conceivable task that a participant could engage in prior to revising their beliefs.

In order to test whether participants actually behaved in a way predicted by a model, we generated an answer that a participant would pick if they were guided by a model and then compared the predicted answer with the actual answer in a binary fashion. We used two models: mean - using simple Borda count - and majority lists.

Table 2 demonstrates that there were significantly more revisions that moved towards the mean than majority. In fact, of the 196 total revisions, 62, or 31% were revisions that adopted the group mean, and 44 or 22% that adopted the majority list. On average, the mean model was adopted 3.26 times per participant, while majority model was adopted 2.32 times. The rest of the revisions were not accounted by these two models and were being guided by *unknown* rules.

Naturally, there were instances where both models predicted the same list and the above numbers include revisions where the mean and majority lists coincide. There were 35 revisions where both models predicted the same result.

When removed from the total revision count for each model, there were 27 revisions that adopted the group mean and only 9 revisions that adopted the majority list. This provides strong evidence to suggest that participants in our study adopted the group mean much more readily than the majority list.

Table 2: Model Revision

Model	Total	Model Only	Average	Revision %
Mean	62	27	3.263	31
Majority	44	9	2.316	22

Toward a Model of Rank Belief Revision Our findings suggest that human participants are 3 times more likely to adopt the group mean over the majority list in cases where the two do not coincide. This suggests that computational models that emphasize mean ranks may be closer to the way humans make revisions given additional information in a ranked format.

We did not test other, more complex models on the study dataset. Therefore, it is difficult to say at this point whether adopting the group mean is the most preferred strategy. It should also be noted that revisions represented only 11% of all choices made by participants and most of the time participants did not change their answers and were not influenced by additional information. However, where revisions did occur, in a third of all cases it was towards the group mean, which is a significant finding. Future models that seek to replicate human behaviour should take these findings into account when constructing more human-like models.

Conclusions

Our research outlined a basic error model as well as two limit cases. In all three scenarios, distance-based methods did not produce significantly better results, suggesting that the problem of rank aggregation could be satisfactorily solved by simpler methods such as taking the average or majority.

As the performances of the two simple methods are diametrically opposite, which method should be used depends on the underlying error distribution in a population. Conversely, if one is able to measure accuracy, the performances of various rank aggregation methods can actually inform us the underlying error distribution and allow us to make inferences about the cognitive process of ranking.

In order to expand on our findings, we conducted a lab experiment where we tested actual belief revision in a group setting. Our findings suggest that when revising their answers, participants most often adopted the group mean, suggesting that human cognition gravitates towards this method of revision. This is significant, in light of the fact that adopting group mean is both computationally less strenuous and quite advantageous in most situations. This suggests that human cognition is adaptive in this sense, using a strategy that our modeling shows to be robust in most cases.

References

Baltrunas, L., Makcinskas, T., & Ricci, F. (2010). Group recommendations with rank aggregation and collaborative filtering. In *Proceedings of the 2010 ACM conference on recommender systems (recsys)* (pp. 119–126). Barcelona, Spain.

Dwork, C., Kumar, R., Naor, M., & Sivakumar, D. (2001). Rank aggregation methods for the web. In *Proceedings of the 10th international conference on world wide web (www)* (pp. 613–622). Hong Kong, China.

Fernández, M., Vallet, D., & Castells, P. (2006). Probabilistic score normalization for rank aggregation. In *Proceedings of the 28th european conference on IR research (ecir)* (pp. 553–556). London, UK.

Niu, S., Lan, Y., Guo, J., Cheng, X., Yu, L., & Long, G. (2015). Listwise approach for rank aggregation in crowdsourcing. In *Proceedings of the 8th ACM international conference on web search and data mining (wsdm)* (pp. 253–262). Shanghai, China.

Renda, M. E., & Straccia, U. (2003). Web metasearch: Rank vs. score based rank aggregation methods. In *Proceedings of the 2003 ACM symposium on applied computing (sac)* (pp. 841–846). Melbourne, FL, USA.

Stewart, N., Chater, N., & Brown, G. D. (2006). Decision by sampling. *Cognitive Psychology*, *53*(1), 1–26.