Is There a Partial Consensus Ordering Between Rankings?

Srinath Sampath* Joseph S. Verducci†

Abstract

We propose an innovative approach to the problem recently posed by Hall and Schimek (2012): determining at what point the agreement between two rankings of a long list of items degenerates into noise. We modify the method of estimation in Fligner and Verducci's (1988) multistage model for rankings, from maximum likelihood of conditional agreement over a sample of rankings to a locally smooth estimator of agreement. Through simulations we show that this innovation performs very well under several conditions. Some ramifications are discussed as planned extensions.

Key Words: partial rankings, top-K rank list, rank aggregation, multistage model, maximum likelihood estimation, consensus

1. Introduction

Ranking a group of objects is a fundamental activity in practically every field of inquiry. Judges – both human and machine – show their preferences among a group of objects by assigning them ranks. Two judges who rank the same group of objects may exhibit similar or dissimilar preference patterns. For example, voters of candidates running for political office may express their preferences with a rank of 1 for their favorite candidate, a rank of 2 for their second favorite candidate, and so on. The earliest papers on ranking objects to reflect one's preferences date back to at least the late 18th century, when Borda (1781) formally provided a contradiction of the 'ordinary method' used in the electoral process of the period of having each voter pick only one best candidate.

Rank aggregation is a vital area of research with applications in practically every field of inquiry. Two resources that explore the rich variety of questions being researched in this area are Fligner and Verducci (1993), and Marden (1995). In many situations, ranks are the only data available from a survey or experiment. The area has also garnered considerable attention in the last decade from fields as diverse as bioinformatics and search engine algorithms.

Closely related to the topic of rank aggregation is the question of *Top-K* specification, which seeks to rank the best K items from a long list. If two assessors independently rank the same long list of objects, the point K in the list where the second assessor becomes uninformative about the first is of considerable interest. As data sets grow in size and number, practitioners are increasingly pressured to focus on that subset of the data where the two assessors show the greatest concordance, and potentially discard or underweight the signal elicited from the remainder of the data.

This is a recent question in ranking literature and has witnessed an elegant approach by Hall and Schimek (2012). The algorithm anchors one assessor's ranks for a group of objects, and digitizes the second assessor's concordance or discordance with the first assessor's ranks respectively through a sequence of ones and zeros. Here concordance occurs when the second assessor's rank for a given object falls within a distance d of the first assessor's rank for the same object. The paper estimates K based on this Bernoulli sequence.

*The Ohio State University, 1958 Neil Avenue, 404 Cockins Hall, Columbus, OH 43210-1247
†The Ohio State University, 1958 Neil Avenue, 404 Cockins Hall, Columbus, OH 43210-1247

We tackle the question of Top-K specification using the forward-looking multistage ranking model framework developed by Fligner and Verducci (1988). At every stage, each assessor is assumed to select the most preferred object from the remaining objects. Hence one of the great strengths of this model is that the stages become independent, and so the probability calculations at every stage avoid the need to condition based on the outcomes in prior stages, which in turn leads to mathematical tractability and closed-form solutions. Another advantage of this model is that the higher the deviation between the two assessors' rankings of a given object, the greater the penalty assigned to the mismatch. In this way, the model provides a nuanced approach to capturing the discordances between the two assessors.

Section 2 describes the model framework and the role of the truncated geometric distribution in the assignment of probabilities to the discordances between the assessors' ranks. The determination of the maximum likelihood estimator and the explicit stopping rule are given in section 3. We use simulations to analyze the behavior and accuracy of the multistage model in Section 4. A discussion of our work and planned extensions are provided in Sections 5 and 6 respectively.

2. The Model

A *ranking* or *permutation* of n distinct objects is a vector of length n, with each component corresponding to an object, and the value of each component being the rank of that object, namely the quantity $1+$ the number of other objects that are considered superior, in either a qualitative or quantitative sense. We use $\pi = [\pi(1), \ldots, \pi(n)]$ to denote this ranking or permutation.

An *ordering* or *inverse permutation* of n objects, labeled 1 to n, is a vector of length n, with each component i giving the label of the object that has rank $i, i = 1, \ldots, n$. The *ordering* or *inverse permutation* associated with π is specified by the mapping $\pi^{-1}(j) = i$ if $\pi(i) = j$, $i = 1, \ldots, n$, $j = 1, \ldots, n$.

We now consider the situation where the n objects are ordered sequentially according to two independent processes. The processes may be the qualitative ranking schemes of two judges who evaluate a shortlist of books that are finalists for a prestigious prize. Another example is two technologies that independently measure the severity of a disease in a group of patients, and rank the patients based on their risk level. It is possible that the two ranking processes are initially governed by common parameters, and, as one moves further down the list, the ranks start to diverge from each other, and ultimately, at rank K, become completely uninformative about each other. We are interested in determining the value of K where the two ranking schemes become uninformative about each other.

Our approach is to fix one of the two assessors' rankings as the *reference ranking* or *ground truth*, and evaluate the deviation of the other assessor's ranks – the *generated ranking* – from the reference ranking, using the Fligner and Verducci (1988) multistage ranking approach as follows. For illustrative purposes, we show the first stage and then generalize to other stages.

Stage 1: Here all n objects are available. The second assessor selects the ith best available object, as specified by π^{-1}, and incurs the penalty $V_1 = \nu = i - 1$, with truncated geometric probability

$$P(V_1 = \nu) = \left(\frac{1 - r_1}{1 - r_1^n}\right) r_1^{\nu},\ \nu = 0, \ldots, n-1,\ 0 < r_1 < 1. \tag{1}$$

Stage j ($j = 2, \ldots, n-1$): In stage j, $n - j + 1$ objects are available. The second assessor picks the ith best object available, as specified by π^{-1}, and incurs a penalty $V_j =$

$\nu = i - 1$, with truncated geometric probability

$$P(V_j = \nu) = \left(\frac{1 - r_j}{1 - r_j^{n-j+1}} \right) r_j^{\nu},\ \nu = 0, \ldots, n - j,\ 0 < r_j < 1. \tag{2}$$

We assume independent choices at each stage of the ranking process, and so the $\{V_j \mid j = 1, \ldots, n-1\}$ are independent. $\{V_1, \ldots, V_{n-1}\}$ is therefore the *discordance* or *penalty vector* between the reference ranking and the generated ranking. Since the probabilities in (1) and (2) are decreasing functions of ν, the model does indeed penalize the second assessor appropriately for larger departures from the reference ranking at a given stage.

The limiting distribution of the V_j as $r_j \to 1$ is discrete uniform on the set $\{0, \ldots, n - j\}$, $j = 1, \ldots, n - 1$, which removes all skill from the second assessor with respect to the reference ranking. Here the second assessor randomly selects from the remaining objects.

For mathematical convenience, we make the substitution $\theta_j = -\log r_j$, $j = 1, \ldots, n - 1$. The condition $r_j \to 1$, which leads to the limiting uniform distribution, is now equivalent to the condition $\theta_j \to 0$.

The original problem statement of determining the value of K where the two ranking schemes are uninformative about each other, is equivalent to determining the value of K for which $\theta_K > 0$, and $\theta_j = 0$ for all $j > K$. Informally, we seek the final stage where the second assessor exhibits some level of concordance with the reference ranking, before devolving into uniformly random selections of the remaining unranked objects. Once past this stage K, the two assessors become uninformative about each other.

3. Parameter Estimation and the Stopping Rule

Recall that the $r_j,\ j = 1, \ldots, n - 1$, are stage-wise measures of the second assessor's concordance with the reference ranking. Since the probability mass function of the penalties in the jth stage is inversely proportional to r_j, a lower r_j leads to higher concordance between the two assessors. With the transformation

$$\theta_j = -\log r_j,\ j = 1, \ldots, n - 1 \tag{3}$$

for the determination of the MLEs of the θ_j's, it can be reasonably assumed that these θ_j's vary gradually from one stage to the next. Specifically, we assume a common value θ for all θ_j in a window of the form $s < j < s + w + 1,\ s = 0, \ldots, n - w - 1$, and with width w. In this manner we calculate a set of local MLEs, $\hat{\theta}_j$, and note that the successive local MLEs use overlapping rank data as the window moves forward through the stages.

3.1 Maximum Likelihood Estimation

Consider a sequence of successive stages $s + 1$ through $s + w$ (both inclusive). The likelihood function of the fixed r, given by $e^{-\theta}$, is given by

$$\begin{aligned} L(r) &= P(V_{s+1} = \nu_{s+1}) \times \ldots \times P(V_{s+w} = \nu_{s+w}) \\ &= \left(\frac{1 - r}{1 - r^{n-s-1}} \right) r^{\nu_{s+1}} \times \ldots \times \left(\frac{1 - r}{1 - r^{n-s-w-1}} \right) r^{\nu_{s+w}}. \end{aligned}$$

The log likelihood is therefore

$$\log L(r) = w \log(1 - r) + (\log r) \sum_{j=s+1}^{s+w} \nu_j - \sum_{i=0}^{w-1} \log(1 - r^{(n-s)-i}). \tag{4}$$

Differentiating (4) with respect to r gives the maximum likelihood estimator as the solution to the equation $\bar{V}_s = g_s(r)$, where

$$\bar{V}_s = \frac{1}{w} \sum_{j=s+1}^{s+w} V_j,$$

and

$$g_s(r) = \frac{r}{1-r} - \frac{1}{w} \sum_{\ell=(n-s)-(w-1)}^{n-s} \ell(\frac{r^\ell}{1-r^\ell})$$

is an increasing function of r.

3.2 The Stopping Rule

The calculations in subsection 3.1 result in MLEs for the stage-wise r_j's, which in turn can be used to compute the stage-wise $\hat{\theta}_j$'s by applying the transformation (3). We now generate a large number of simulations from the multistage model, compute the stage-wise $\hat{\theta}_j$'s from each simulation, and, for each stage j, plot $q(j)$, the $(1-\alpha)$th quantile of $\hat{\theta}_j$. Then the stopping stage K is estimated by K^*, the earliest stage at which $\hat{\theta}_{K^*} > q(K^*)$, and $\hat{\theta}_j > q(j)$ for at most α percent of the remaining $j > K^*$.

4. Analysis of the Behavior and Accuracy of the Multistage Model using Simulations

We simulated data to represent the ranking schemes of two assessors on 200 objects. The black line in the two exhibits of Figure 1 represents the true θ used in the simulation. The linear descent over the first 100 stages represents the decreasing concordance between the two assessors. The cliff event at stage 101 represents the abrupt discordance between the second assessor and the preference ranking, after which the second assessor ranks the remaining objects at random for the remainder of the stages, which is represented by the horizontal line. The left and right graphs show the computed forward moving average MLEs with window widths of 20 and 40 stages respectively. While the blue line represents the computed $\hat{\theta}$'s, the red line represents the 'reverse' $\hat{\theta}$'s, where the MLEs are calculated as before, but the roles of the first and second assessors are reversed, with the second assessor providing the reference ranking, and the first assessor providing the generated ranking.

It is clear that while the reversal of roles between the two assessors leads to a slightly different picture of the degradation of information, and an average of the two estimators could be used for Top-K detection, in practice this is not necessary. It is also evident that both window width assumptions capture the cliff event very well, with the narrower window MLEs providing the more accurate picture of the cliff event than the wider window MLEs. Furthermore, all four MLE curves show a positive bias, which can be reduced by underweighting the earlier stages in each MLE computation. Once past the cliff event, all four MLEs succeed in capturing the flatness of θ in the later stages, with the wider-window MLEs showing better results. It is also clear that the wider-window MLEs show lower variability in the earlier stages.

Figure 2 displays the results of a simulation involving two assessor rankings and 400 stages. The left graph calculates the MLEs using rolling 40-stage windows, while the right graph uses 80-stage windows. Here again we see that the $\hat{\theta}$'s and the reverse $\hat{\theta}$'s are very similar, and that all MLE curves get close to the more modest cliff event at stage 182. All four MLEs again show a positive bias, with the bias decreasing closer to the cliff event. As with the 200-stage runs, all four MLEs are successful in capturing the flatness of θ in

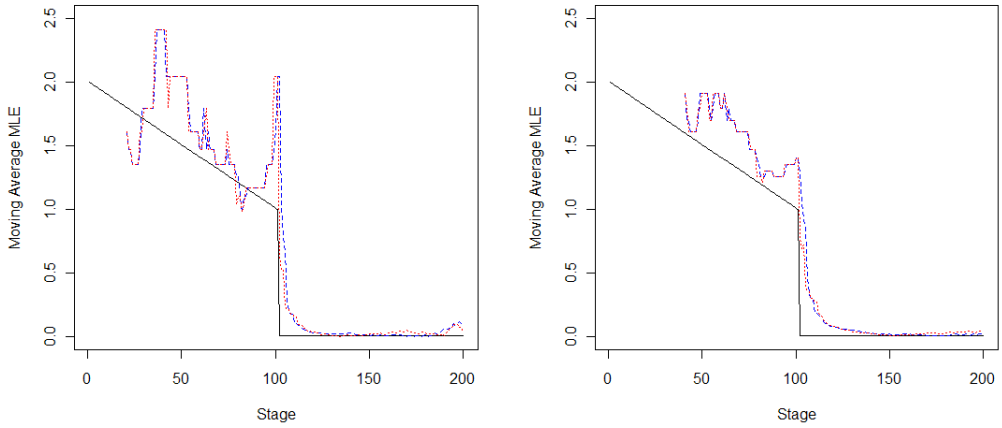

Figure 1: Forward Moving Average MLE for a 200-stage simulation. L: window width 20, R: window width 40. Black line: true θ, Blue line: $\hat{\theta}$, Red line: reverse $\hat{\theta}$.

the later stages. As expected, the wider-window MLEs show lower variability in the earlier stages.

Figures (3) and (4) show, respectively, the form of the MLEs computed for windows of widths 10, 20, 30 and 40 stages for a list of 200 objects, and MLEs computed for windows of widths 20, 40, 60 and 80 stages for a list of 400 objects. We note the following:

1. The presence of the positive bias of $\hat{\theta}$ that increases as the window width increases
2. Volatility decreases as the window width increases
3. The modest cliff event is best picked up by the MLE created using the narrowest window width, and the MLEs created using larger window widths miss the cliff event by increasing margins, and
4. The flat region is best picked up by the MLE created using the narrowest window width, and the MLEs created using larger window widths miss the cliff miss the flat region by increasing margins.

5. Discussion

We have proposed an innovative approach to determine the point K of discordance between the ranks of two assessors who evaluate a list of objects. Our approach uses the multi-stage ranking model framework developed by Fligner and Verducci (1988), and exploits its forward-looking and graded assignment of penalties to the mismatched ranks provided by the assessors at every stage. Simulations show that our approach is very successful in recognizing the overall shape of the parameter curve, in particular, cliff events and flat regions. The positive bias of $\hat{\theta}$, caused by equal-weighting the earlier stages, can be reduced by underweighting them. The underlying mathematics is tractable and the algorithm is quick to program and execute, providing an elegant tool to evaluate the concordance between two assessors' preferences among a list of objects, and a technique to estimate the point of degeneration between the rankings.

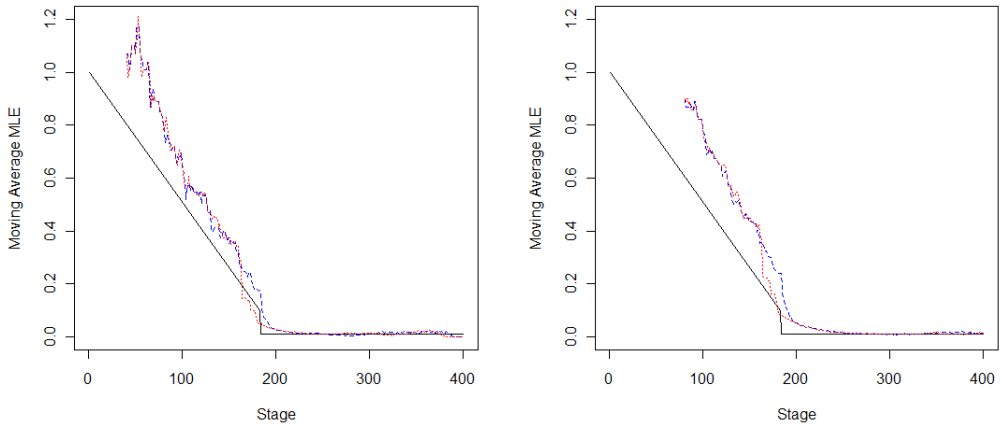

Figure 2: Forward Moving Average MLE for a 400-stage simulation. L: window width 40, R: window width 80. Black line: true θ, Blue line: $\hat{\theta}$, Red line: reverse $\hat{\theta}$.

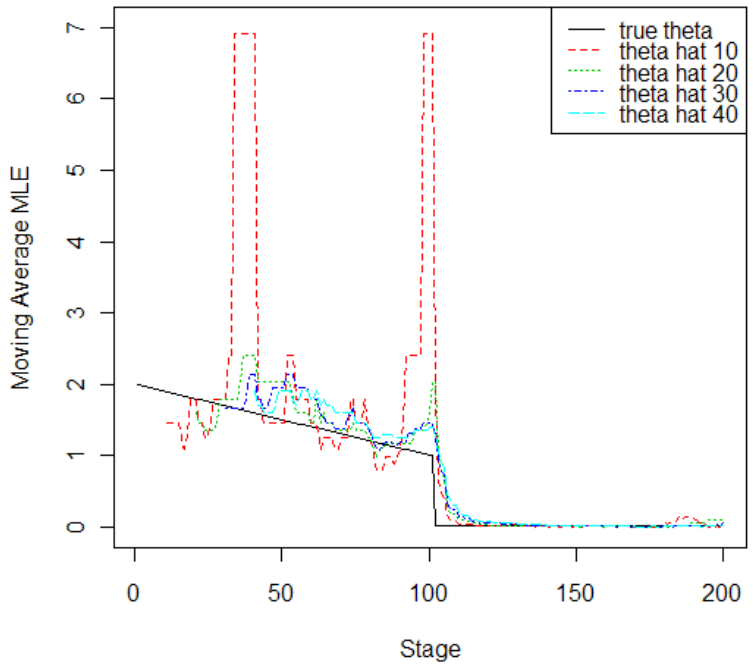

Figure 3: Forward Moving Average MLE: 200 Stages and Widths 10, 20, 30, 40

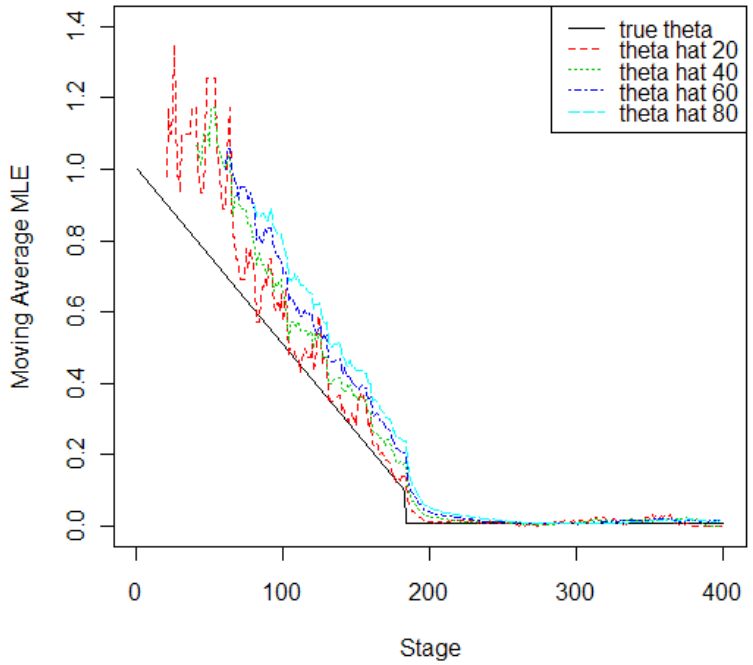

Figure 4: Forward Moving Average MLE: 400 Stages and Widths 20, 40, 60, 80

6. Planned Extensions

A valuable extension to our approach is to the process of aggregating ranks provided by more than two assessors. An example from the world of search algorithms is the consolidation of ranks provided by three or more search engines in response to the same input phrase.

We are also interested in studying the behavior of our approach on data generated from copulas, especially those with asymmetric tails. An evaluation of our technique on customer retention data from a financial services company is also under consideration.

Finally we are interested in exploring the possibility of developing an algorithm to self-tune the window width based on the underlying data. The window widths in our approach are currently input items, and typically multiples of ten. We hope to develop an algorithm which determines, at each stage, how wide the window should be, and represented either as a fixed number or as a percentage of the total number of stages available for analysis.

REFERENCES

Borda, J.-C. (1781), "Mémoire sur les Élections au Scrutin," *Histoire de l'Académie Royale des Sciences*, 31–34.

Fligner, M. A., and Verducci, J. S. (1988), "Multistage Ranking Models," *Journal of the American Statistical Association*, 83:403, 892–901.

Fligner, M. A., and Verducci, J. S. (1993), *Probability Models and Statistical Analyses for Ranking Data*, Lecture Notes in Statistics, 80, Springer-Verlag. Editors.

Hall, P., and Schimek, M.G. (2012), "Moderate-Deviation-Based Inference for Random Degeneration in Paired Rank Lists," *Journal of the American Statistical Association*, 107:498, 661–672.

Marden, J. I. (1995), *Analyzing and Modeling Rank Data*, Monographs on Statistics and Applied Probability, 64, Chapman & Hall.