RATING CONFERENCE PRESENTATIONS WHEN JUDGES VARY

RICHARD MCELREATH

June 1, 2021

1. Purpose

Academic societies often award prizes for best talk and best poster. The procedure for making these awards depends upon judges, usually conference participants or officers of the society, who may not rate every talk/poster. Furthermore, individual judges almost certainly vary in how they use the rating scale and how much they value different aspects of a talk/poster.

How can consensus about talks/posters be reliably inferred from the available ratings? A fast and common approach is to use the average rating each talk/poster received. However, if a particularly harsh judge rates an excellent talk, it will receive a lower average rating than it deserves. Likewise if a particularly easy-going judge rates a talk, the talk will receive a higher score than it deserves.

There is a large literature of statistical solutions to this problem. What they all have in common is the simultaneous inference of latent features of both the talks and the judges. Here we develop one of these solutions, a solution appropriate when ratings are given in ordered categories like 1 through 5. Of course there is no ground truth in this domain. We just want to reliably estimate the consensus ranking of talks, adjusting evaluatings for different scale effects specific to each judge.

2. Generative Model

Suppose there are N talks (or posters) and M total judges. Each talk is assigned K judges. This implies that each judge will need to rate at least $\lceil KN/M \rceil$ talks. For example, if there are $N = 50$ talks and $M = 10$ judges, with $K = 4$ judges per talk, then each judge should be assigned to at least 20 talks. Ratings are given on an integer scale.

Talks have one or more latent features that influence ratings. In the simplest case, each talk i has exactly one feature, and we assign it a Gaussian prior:

$$z_i \sim \text{Normal}(0, \tau)$$

Department of Human Behavior, Ecology and Culture, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany

E-mail address: richard_mcelreath@eva.mpg.de.

Date: June 1, 2021.

where τ is a scale parameter that determines how differentiated the talks are. The Normal distribution is not required. A thicker tailed distribution could be used instead, so that exceptional talks are not shrunk too much towards the mean. However with an integer rating scale, the ceiling effect of the scale can make thick-tailed priors pointless. Unless the population of judges is very, well, judgmental, it will not be possible to differentiate good talks from excellent talks.

Judges have unique cut-points, which are thresholds on the latent features of talks at which a rating transitions from one value to another. For a scale with x levels (a 1–x rating scale), we require $x - 1$ cut-points. Let L equal the number of required cut-points. Cut-points are specified on the logit scale, such that each is a cumulative probability of a judge assigning a rating or any lower rating. An individual judge j has cut-points distributed as:

$$c_j \sim \text{MVNormal}(\kappa, \mathbf{SRS})$$

with the restriction that the elements of c_j be ordered low to high. In practice, this is achieved by representing all cut-points above the first with increments. The code implementation uses this offset strategy.

The rating given by judge j to talk i is then distributed as:

$$y_{ij} \sim \text{Ordered-Logistic}(z_i, c_j)$$

So in principle this is just an ordinary ordered-logistic IRT model with hierarchical structure in the cut-points.

In the case that more than feature of the talks is rated, the model generalizes by specifying unique cut-points for each judge and feature. For example, feature q of talk i rated by judge j:

$$y_{iqj} \sim \text{Ordered-Logistic}(z_{iq}, c_{jq})$$

This means that each judge requires LQ cut-points, where Q is the number of features rated by each judge. Likewise Q parameters are needed for each talk, and their prior is now:

$$\mathbf{z}_i \sim \text{MVNormal}(0, \mathbf{B})$$

where $\mathbf{B}$ is a suitable covariance matrix. Modeling the covariance among features allows for partial pooling across features, which could improve estimates when there are strong correlations among features.

3. Implementation and Priors

This generative model can be run forwards, to simulate judgments, or backwards to produce posterior distributions for each talk (as well as judge). Inference requires prior distributions for the talk scale τ, the cut-point means κ, and the cut-point covariance matrix $\mathbf{SRS}$, where $\mathbf{S}$ is a vector of scales (standard deviations) and $\mathbf{R}$ is a correlation matrix. Note that all three of these are hyperparameters used to induce shrinkage among talks and judges.

Hyperpriors in these models are commonly chosen using one of three methods: (1) priors can chosen using some set of principles, such as desired shrinkage properties, (2) priors can be decided through prior predictive simulation, and (3) priors can be found through tuning on existing or hold-out data. We don't have data available for (3), at least not yet. But storing conference data will allow us to use this approach in the future. This leaves us with (1) and (2), which can be used together.

A standard set of priors with good shrinkage properties might be:

$$\begin{aligned}\tau &\sim \text{Exponential}(1)\\ \mathbf{S} &\sim \text{Exponential}(1)\\ \mathbf{R} &\sim \text{LKJCorr}(2)\end{aligned}$$

where LKJcorr is the correlation matrix family of distributions described in Lewandowski et al. 2009.

In the case of $Q > 1$, that is more than one feature to be rated, the covariance among cut-points can be done using the convention above. But additional structure can also be specified, producing a covariance matrix with fewer free parameters. For example, suppose $Q = 2$ and $L = 2$ for the simplest case. This requires $QL = 4$ parameters for each judge. These are $[ab]$ for the first feature and $[AB]$ for the second. A complete free correlation matrix then requires 6 correlation parameters:

$$\begin{pmatrix} 1 & \rho_{ab} & \rho_{aA} & \rho_{aB} \\ & 1 & \rho_{bA} & \rho_{bB} \\ & & 1 & \rho_{AB} \\ & & & 1 \end{pmatrix}$$

where the lower triangle is symmetric with the upper triangle. This structure can be reduced by noting that we desire a unique correlation structure within each block of cut-points ab and AB. And then matching positions between blocks should share a correlation ξ, because they belong to the same judge and so share some deviation from the population mean. This gives us:

$$\begin{pmatrix} 1 & \rho_{ab} & \xi & 0 \\ & 1 & 0 & \xi \\ & & 1 & \rho_{AB} \\ & & & 1 \end{pmatrix}$$

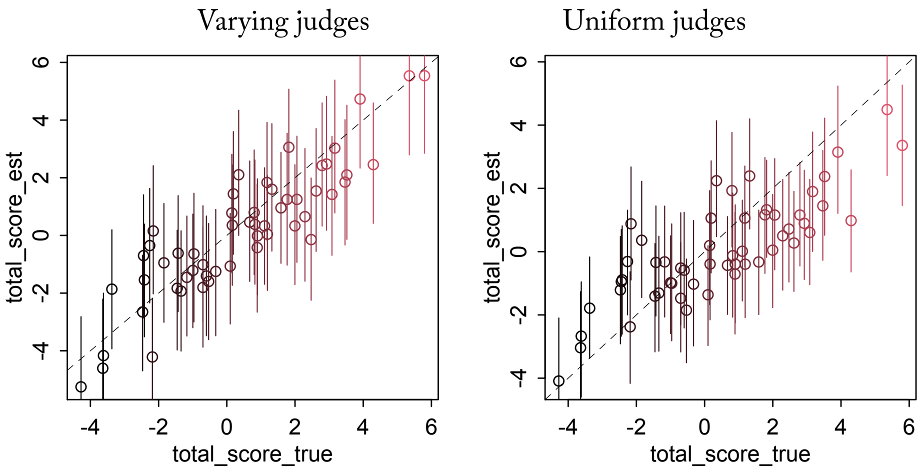
So now there are only 3 parameters to estimate. Very large LQ, the savings can be very great. However it is very difficult to specify priors in this case that ensure a valid correlation matrix. As a result, prior predictive simulations are needed to understand the implications of priors here. In contrast, in the completely free case, the LKJ family of priors can be used, and these are well understood and guarantee valid matrices.

4. Testing

We have evaluated the model's performance on synthetic data, both in order to test that it works (in large samples with complete information) and to understand how it produces inferences. The key assumptions of the generative model are:

(1) Talks vary in multiple dimensions (features) and these features are positively correlated with one another.
(2) Judges vary in how they evaluate each feature as a function of each talk's unobserved latent value for that feature.

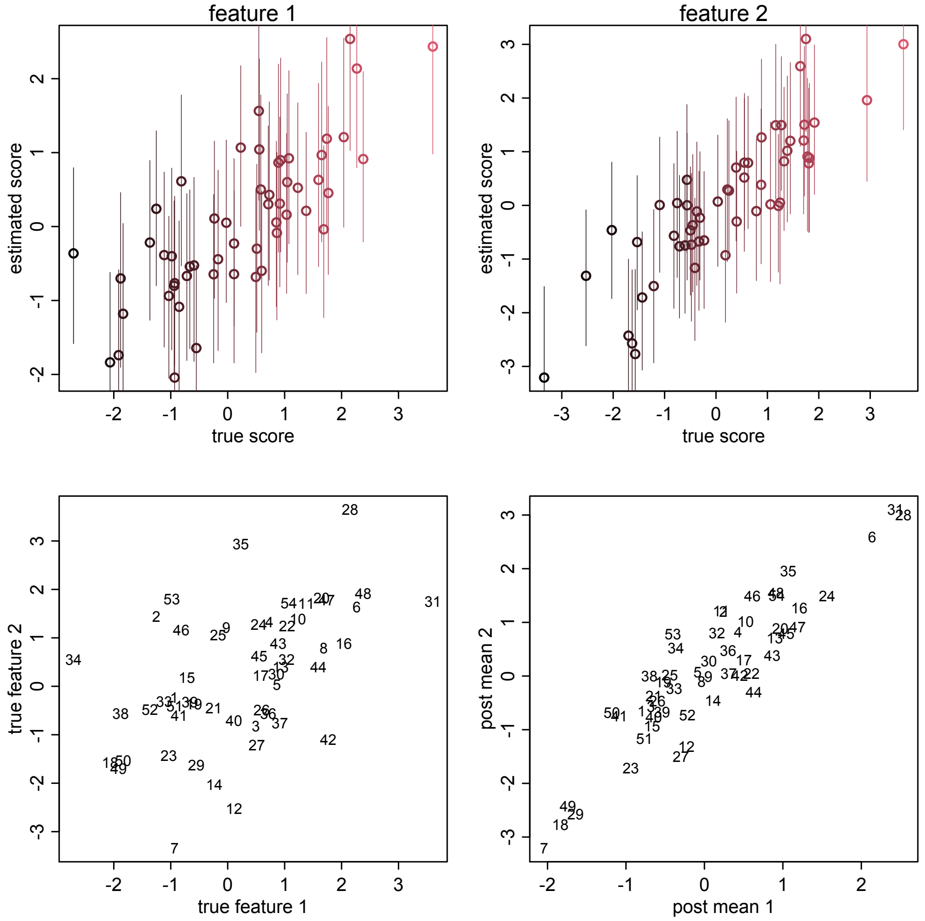
How does the model produce inferences at our defaults of $N = 54$ talks rated by $M = 18$ judges on $Q = 2$ features on a 5-point ($L = 4$) scale, with $K = 4$ judges per talk? The lefthand plot below shows posterior distribution of each talk's total score (by user-defined weights of each feature) against the true total score (using the same weights), for a typical simulation. The points are posterior means and the line segments are 89% compatibility intervals.



These results are quite typical for item-response models. There is a strong posterior correlation between the underlying truth and the inference, but it can be hard to distinguish adjacent items from one another. The inference also shows the expected regression towards the mean, with the oustanding talk on the far right being shrunk along the vertical axis towards the average talk. Converting these posterior means to ranks, the correlation between the true ranks and inferred ranks is, in this example, 0.87.

To appreciate what allowing judges to vary does for inference, consider the same simulated data fit to a model in which all judges share the same feature-specific cut-points. This is the righthand plot above. There is more regression to the mean here, because the variation among judges introduces noise that is uncorrelated with score, reducing differences among talks. The posterior correlation between true rank and inferred rank is now 0.78. Notice that the two best talks benefit a lot from controlling for variation in judges. In a simulation in which judges varied more, including some judges who are extremely harsh or generous in scoring, the benefit of judge-specific cut-points would be greater.

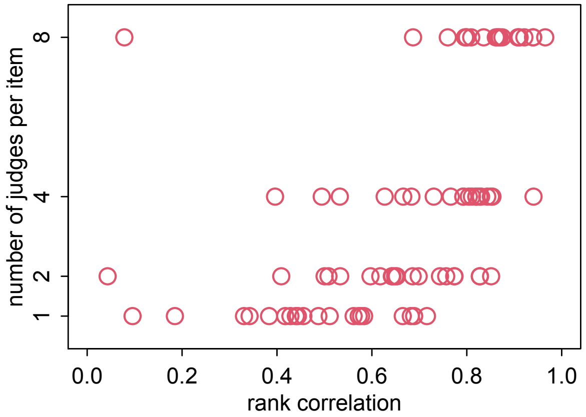
Finally, for completeness, here are the inferred features, for the same simulation with seed 9001. The top row below shows feature 1 (left) and feature 2 (right) against the posterior mean and 89% interval. The bottom row shows the true relationship between the two features (left) and the inferred (right), with each talk's number as its label.



In this simulation, there is a mild correlation between the two features, which lets inference about one feature aid inference about the other. If in fact features are negatively correlated, inference will be harder, and this will be reflected in the width of the posterior total scores.

5. Sensitivity

We might like to know how well talks can be ranked, given different numbers of judges assigned to each talk. In the graph below, we measure the correlation between the true ranks and the inferred posterior mean ranks for $N = 54$ talks rated by $M = 18$ judges on $Q = 2$ features on a 5-point ($L = 4$) scale. We vary K, the number of judges per talk, with 20 simulations at each of 4 values of K. The correlation in ranks increases rapidly with more than $K = 1$, but there is much less improvement from $K = 4$ to $K = 8$. Note that for $K = 8$, each judge must rate 24 different talks. For $K = 4$, it is only 12 talks.



But of course more judges rating each talk provides better inference, both because it provides more information about each talk as well as providing more information about the bias of each judge. The best policy may be to decide the maximum number of talks each judge can assess. Call this number J. Then use the formula $K = JM/N$ to determine the number of judges assigned to each talk. In many cases, the number of judges will not divide evenly into the number of talks. The model still works fine in such cases.

Something else we might to know is which groups of talks can be reliably distinguished from groups of talks below and above. There are many different algorithms for clustering by a matrix of differences. However in this case it is not necessary to cluster the entire sample of talks, as only the top few are interesting. In this particular example, seed 9001, the best talk can be easily distinguished from the others, but there is a large group of talks that are essentially indistinguishable.