The Inconsistent Judge

Dan Scholz
Jade Vinson
Derek Oliver Zaba
Dept. of Systems Science and Mathematics
Washington University
St. Louis, MO 63130

Advisor: Hiro Mukai

Summary

We provide a judging process that is robust enough to ensure that the intrinsically best papers are chosen in spite of randomness and subjectivity in the judging process. We increase the scope of the problem by introducing inconsistency into the judging process. We model this inconsistency by expressing the actual paper score as the sum of the intrinsic numerical score, the overall bias of the judge, and an error term:

$$S_{jp} = S_p + B_j + \epsilon_{jp}.$$

We use an iterative computer-guided process to determine the judging procedure. After each round of judging, the computer program uses bias estimates to calculate confidence intervals for the intrinsic score of each paper. These confidence intervals are used to reject as many papers as possible while guaranteeing, within a specified level of confidence, that the top W papers advance to the next round. Less cautious rejection criteria in each round adapt the method to select the winning papers from among the top $2W$ papers.

We did a computer simulation over a range of values for the parameters. Intrinsic scores were normally distributed with mean 50 and standard deviation 20; bias and consistency parameters were varied. We compare the method results to the intrinsic scores of the papers. For $P = 100$, $J = 8$, $W = 3$, the method proved correct 95% of the time with an average of 175 papers read.

Assumptions

Given that papers have an absolute intrinsic ranking, we assume that papers also have an associated intrinsic numerical score. The score that a judge gives a paper reflects not only the intrinsic score of the paper and the overall bias of the judge but also the inconsistency of the judge. This assumption is more realistic than assuming that all judges would agree to an absolute ranking and will produce a more robust judging procedure. We assume:

- Papers have intrinsic scores which follow a normal distribution.
- Judges have constant numerical bias.
- The range of biases for all judges follows a normal distribution.
- Judges' inconsistency follows a normal distribution.

The normal distribution is used for analytical convenience and is justified by historical precedent [DeGroot 1986, 263–264].

The Model

We express our assumptions mathematically by equating the score S_{jp} that judge j assigns paper P to the sum of the intrinsic score S_p of the paper, a bias term B_j for the judge, and an error term ϵ_{jp} for the score:

$$S_{jp} = S_p + B_j + \epsilon_{jp}.$$

Our model has parameters μ, σ, B, and Δ. The distribution of intrinsic scores is parameterized by μ and σ. That is, S_p is a random variable with distribution $N(\mu, \sigma^2)$. Parameter B is a measure for the bias of all the judges; the bias B_j for a particular judge comes from the distribution $N(0, \sigma^2)$. The parameter Δ is measure for the overall consistency of the judges. The error for an individual grading, ϵ_{jp}, comes from the distribution $N(0, \Delta^2)$. The terms B_j and ϵ_{jp} account for the subjective nature of the judging process.

The Method

Our model estimates the intrinsic scores of papers by producing estimates for the bias of the judges and adjusting their scores accordingly. Our confidence in the estimated intrinsic scores is used to reject as many papers as possible while maintaining that the probability of rejecting one of the top W papers is less than a predetermined α. The method proceeds as follows:

Distribution of Papers

Our model distributes papers according to the following prioritized criteria:

- No judge reads the same paper more than once.
- The numbers of papers read by each judge for a given round do not differ by more than one. This minimizes time spent reading for a round.
- Workload is distributed equally among the judges.

Estimation of the Intrinsic Score Distribution

Since the distribution and values of the actual intrinsic scores are not known, we attempt to estimate them. We estimate the mean and variance of the intrinsic scores after the first round. The mean and variance are estimated by the following (see **Appendix A**):

$$\hat{\mu} = \overline{S}_{jp}, \qquad \hat{\sigma^2} = \frac{1}{J}\sum_j \frac{1}{P(j)-1}\sum \left(X_j - \mu_j\right)^2,$$

where the $P(j)$ denotes the number of papers judge j has read.

Calculation of Bias

After round one, each judge will have scored approximately P/J papers. If the average of the scores for a given judge is significantly greater than the mean of the scores, either the judge is positively biased or the judge happened to receive a sample of unusually good papers, or both. If $X_1, \ldots, X_n$ are the scores of papers reas by a particular judge, then the conditional distribution for B_j after round one (see **Appendix B**) is normal with mean and variance given by

$$B_j^1 = \frac{\sum(X_i - \mu)}{n + \frac{\sigma^2+\Delta^2}{B^2}}, \qquad V_j^1 = \frac{1}{\frac{1}{B^2} + \frac{n}{\sigma^2+\Delta^2}}.$$

Note that in the special case when $B = 0$, the estimate for B_j is also zero, but if B is large, the distribution has mean approximately $\overline{X}$ and variance σ^2/n.

Recalculation of the Bias

If the judges are unusually consistent, i.e., if Δ is very small, we would like our judging procedure to recognize and take advantage of this fact. In the extreme case, when $\Delta = 0$, we can precisely rank all P papers with only $P + J - 1$ readings: Start by dividing the papers evenly in the first round; in the second round, judge A retires while each of the other judges reads one of judge A's papers; by subtracting out each judge's bias relative to judge A, we learn the precise ranking of the papers.

The method optimized for the trivial case above is successful because it uses the fact that $\Delta = 0$ to calculate the biases exactly after the second round. We could adapt this simple example to improve our judging procedure. In the first round, the biases are estimated according to the preceding section. These are used for the first cut. In subsequent rounds, first re-estimate Δ. Using the new value, re-estimate the biases B_j and their variances V_j of our estimates; if the new value of Δ is small, so is the uncertainty of our bias estimate. The combination of a small inconsistency Δ and accurate knowledge of the biases would allow us to calculate an estimated intrinsic score more

accurately. With sharpened values of the estimated intrinsic score, more papers could confidently be eliminated after each round. We derive and present the formulas for this bias re-estimation in **Appendix B**.

Estimation of Intrinsic Scores

The estimated bias for each judge is used to calculate for each paper p a net score that estimates the intrinsic score after taking into account the bias of the judges and the number of readings. The mean and variance for the net score are (see **Appendix B**)

$$\text{mean} = \frac{\frac{\mu}{\sigma^2} + \sum_{j \text{ judges } p} \frac{S_{jp} - B_j}{V + \Delta^2}}{\frac{1}{\sigma^2} + (\#\text{ readings}) \frac{1}{V+\Delta^2}}, \qquad \text{variance} = \frac{1}{\frac{1}{\sigma^2} + (\#\text{ readings}) \frac{1}{V+\Delta^2}}.$$

Here $V = \max V_j$ is used instead of V_j to simplify forthcoming calculations.

Rejection of Papers

At the end of each round, we seek to eliminate as many papers as possible while still ensuring that the best W papers are selected within a specified degree of confidence. **Appendix D** derives the inequality

$$\Pr(\text{mistaken rejection}) < W \cdot \sum_{p=1}^{R} \Phi\left(\frac{Sj, p - S_{j,p-w+1}}{\sqrt{2}\sigma_T}\right).$$

The variable S_{jp} reflects the computed score of paper p in ascending order, $S_{j,p-w+1}$ is the score of the paper whose score is ranked w, and σ_T^2 is the total variance of the score distributions. This inequality lends itself to an iterative process in which the lowest papers are rejected one by one until the inequality reaches a desired level of confidence. If there still remain more than W papers after the confidence level is reached, a new round is initiated. This iterative process involves repeating the earlier steps of this section.

Model Implementation

We simulated the model with a C++ program to demonstrate its validity and scope. The simulation compares the actual $2W$ intrinsic score winners to the W model-determined winners. Due to time constraints, the re-estimation of the bias was omitted from the simulation.

Initialization

The simulation assumes that there are 100 papers, 8 judges, and 3 winning papers. It generates absolute intrinsic scores from a normal distribution with mean 50 and standard deviation 20. The parameters B and Δ, which determine the generation of judges' parameters, are varied over a realistic range, empirically determined to be between 5 and 10. For all simulation calculations, the judging process assumes $B = 8$ and $\Delta = 8$, in order to demonstrate the validity of the model with no knowledge of the distributions of the biases and inconsistencies.

Simulation

We ran the program for 1,000 competitions with various levels of confidence per round and distributions of scores. The model was successful approximately 93–97% of the time with 160–190 readings. Due to the slack in the confidence inequalities, a strict lower bound of 0.7 confidence per round produced these encouraging results while significantly reducing the total number of readings.

Real-World Implementation

A fully implemented computer program would allow a judging team to input the number of papers to be evaluated, the number of winning papers, and the number of judges. The program would have the judges input the scores for each paper that they judged in round one. The program would then ask for a degree of confidence that the winning papers will be drawn from the top $2W$ papers. Output is the designation of the papers that are to be advanced to the next round. The judges now enter their scores for round two, and the process is repeated until W papers remain.

Stability

The formulas used thus far have relied upon exact values for parameters μ and Δ for the distribution of intrinsic scores to greatly simplify calculations. This information, however, would not be available in actual implementation of our judging process. Fortunately, small inaccuracies in the calculated values of $\hat{\mu}$ and $\hat{\sigma}$ do not undermine the validity of our judging process.

Strengths and Weaknesses

Our model provides a great deal of flexibility for variations in the judging procedure. We do not assume that the judges will agree in absolute ranking for a given competition. We are able to do this with a confidence of 95% for 3

winning papers with an average of 175 total readings. These numbers reflect simulation without re-estimating the biases. If re-estimation calculations are implemented, the numbers would improve.

Shortcomings of the model include its assumption that papers have an intrinsic score. This confines the validity of the model to more technical papers. Additionally, the model does not account for a situation in which the inconstancy of judges may vary in the distribution of the scores they report. This would happen if one judge used the full range of scores from 0 to 100 and another judge had a tighter range of reported scores. In this sense, the model does not fully reflect reality. The assumptions of a normal distribution and simulation over normally distributed data are also approximations of reality.

Appendix A: Estimation of Intrinsic Score Distribution

We seek to estimate the mean and variance of the intrinsic scores by observing the scores of the first round. The most reasonable estimate for the mean of the intrinsic scores is simply the mean of the scores observed in the first round, $\hat{\mu} = \overline{X}$. The scores assigned to various papers in the first round by a particular judge are of the form $S_{jp} = S_p + B_j + \epsilon_{jp}$ for the various values of p. The variance of these numbers for fixed j is the sum of the variance of the intrinsic score and the inconsistency Δ^2. Since the Δ^2 is insignificant compared to σ^2, we may approximate σ^2 by the variance Var S_{jp} for a fixed value of j. Averaging this variance over all judges yields a reasonable estimate for the variance σ^2:

$$\hat{\sigma}^2 = \frac{1}{J} \sum \text{Var } S_{jp}.$$

Appendix B: Estimation of Biases and Net Scores

Theorem. *Suppose that A is a random variable with distribution $N(\mu, \sigma^2)$ and A is hidden from observation, but the independent random variables $X_i = N(A, \sigma_i^2)$ are observed. Given observations X_i, the conditional distribution for A is normal with*

$$\textit{mean} = \frac{\frac{\mu}{\sigma^2} + \sum \frac{X_i}{\sigma_i^2}}{\frac{1}{\sigma^2} + \sum \frac{1}{\sigma_i^2}}, \qquad \textit{variance} = \frac{1}{\frac{1}{\sigma^2} + \sum \frac{1}{\sigma_i^2}}.$$

Proof: [EDITOR'S NOTE: This theorem was formulated by the authors, who could not find a reference for it. For reasons of space, we omit their proof, which is based on results in DeGroot [1986].]

Corollary 1. *If a judge's bias comes from the distribution $N(0, B^2)$ and the scores $X_1, \ldots, X_n$ reflect the bias, the variation σ^2 of intrinsic scores of papers with mean μ, and the inconsistency Δ of the judge, then our estimate for the bias of this judge and the variance of this estimate are*

$$B_j = \frac{\sum(X_i - \mu)}{n + \frac{\sigma^2+\Delta^2}{B^2}}, \qquad V_j = \frac{1}{\frac{1}{B^2} + \frac{n}{\sigma^2+\Delta^2}}.$$

Corollary 2. *Suppose that the intrinsic score of a paper comes from $N(\mu, \sigma^2)$. The scores S_{jp} reflect the intrinsic score of the paper, the biases of the judges, and the inconsistency Δ of the judging process. Our estimates B_j of the biases each have variance V. Then our estimate of the intrinsic score for paper p has mean and variance:*

$$\begin{aligned} \textit{mean} &= \frac{\frac{\mu}{\sigma^2} + \sum_{j \text{ judged } p} \frac{S_{jp}-B_j}{V+\Delta^2}}{\frac{1}{\sigma^2} + (\#\ \textit{readings})\frac{1}{V+\Delta^2}} \\ \textit{variance} &= \frac{1}{\frac{1}{\sigma^2} + (\#\ \textit{readings})\frac{1}{V+\Delta^2}}. \end{aligned}$$

Each paper has a normal score distribution with mean $\hat{S}_j$ and variance σ_T^2. The variances are the same for each paper. Then $S_{jp} - B_j$ (based on our best estimate of B_j, which may change from round to round), which is our best estimate of a judge's score, has variance $V_j + \Delta^2\sigma^2$. If we just up all bias variances to $V = \max V_j$, this becomes $V + \Delta^2\sigma^2$. So overall for this paper,

$$\text{mean} = \frac{\frac{\mu}{\sigma^2} + \sum \frac{S_{jp}-B_j}{V+\Delta^2\sigma^2}}{\frac{1}{\sigma^2} + n\left(\frac{1}{V+\Delta^2\sigma^2}\right)}, \qquad \text{variance} = \frac{1}{\frac{1}{\sigma^2} + n\left(\frac{1}{V+\Delta^2\sigma^2}\right)}.$$

Note that $\sigma_T = V + \Delta^2\sigma^2$.

Appendix C: Re-estimation of Parameters

First we seek to re-estimate the parameter Δ. If we consider all papers (at least two) read by both judge j and judge k, the differences are distributed according to

$$S_{jp} - S_{kp} = B_j - B_k + (\epsilon_{jp} - \epsilon_{kp}) = B_j - B_k + N\big(0, 2\Delta^2\big).$$

By computing the variance of the differences $S_{jp} - S_{kp}$ for a fixed pair of independent judges, we obtain an estimate of the variance $2\Delta^2$. The more papers the pair of judges has read in common, the more precise this estimate will be. We obtain a still more precise estimate of $2\Delta^2$ by averaging these variances

over each pair of judges, weighting each average according to the number of papers read by both judges:

$$\hat{\Delta}^2 = \frac{\frac{1}{2}\sum(P(j,k)-1)\text{Var}\left[S_{jp}-S_{kp}\right]}{\sum(P(j,k)-1)},$$

where $P(j,k)$ is the number of papers read by both judges j and k. Using the updated estimate of Δ, we may now re-estimate the biases B_j as well as their variances V_j. We use an iterative procedure and demonstrate that for $\Delta \neq 0$ the successive calculations for B_j and V_j converge. We cannot rigorously demonstrate the validity of this iterative procedure. Instead, we justify this procedure by intuitively motivating each step. [EDITOR'S NOTE: For reasons of space, we omit the details.]

Appendix D: The Confidence Inequality

Theorem. *Let $S_{j1},\ldots,S_{jP}$ denote the computed scores of the papers sorted in ascending order and let σ_T denote the standard deviation of the score estimates. If $R \leq P-W$, then the probability of accidentally rejecting one of the best W papers by rejecting the R lowest-ranked papers is bounded by*

$$\text{Pr(mistaken rejection)} < W \cdot \sum_{p=1}^{R} \Phi\left(\frac{Sj,p-S_{j,p-w+1}}{\sqrt{2}\sigma_T}\right).$$

Proof: To mistakenly eliminate one of the best W papers, it is necessary that one of the rejected papers have an intrinsic score greater than that of one of the top W papers. Thus

$$\begin{aligned}
\text{Pr(mistaken rejection)} &< \sum_{p=1}^{R}\sum_{q=P-W+1}^{P} \text{Pr}(S_p > S_q) \\
&< \sum_{p=1}^{R}\sum_{q=P-W+1}^{P} \text{Pr}(S_p > S_{P-W+1}) \\
&= W \cdot \sum_{p=1}^{R} \text{Pr}(S_p - S_{P-W+1}) \\
&= W \cdot \sum_{p=1}^{R} \Phi\left(\frac{S_{j,p}-S_{j,P-W+1}}{\sqrt{2}\sigma_T}\right).
\end{aligned}$$

References

DeGroot, Morris H. 1986. *Probability and Statistics.* Reading, MA: Addison-Wesley.