Are We Ranking Fairly? A Bias-Aware Bayesian Rank Aggregation Method for Peer- and Self-Evaluations

Haoyao Ruan[1], Kaiwen (Kevin) Wang[2], Xinlei (Sherry) Wang[1]

1. Department of Mathematics, University of Texas at Arlington, Arlington, Texas
2. Department of Statistical Science, Southern Methodist University, Dallas, Texas

Introduction

- **Rank aggregation (RA)**, which aims to combine multiple ordinal rankings into a single ranking, has broad applications in fields such as elections [1], genomic research [2], and educational evaluation [3].
- In **peer-and self-assessments,** subjective biases can lead to unreliable or unfair outcomes, and existing methods often overlook these unique data mechanisms.
- We propose a novel, bias-aware, Bayesian method **BayeSRank** (**Baye**sian Bias Detection in Peer-and **S**elf-**Rank**ing), validated through simulation studies and real-world data.

Figure: Illustration of Peer- and self-evaluation in a classroom setting. Generated by Adobe illustrator.

- We also invite you to consider these related interesting questions:
  - ❑ Should self-evaluation data be removed to mitigate self-bias?
  - ❑ Is self-bias indicative of a ranker's overall competence?
  - ❑ How will self-bias influence the aggregated results?

Data and Latent Variable Model

	Ranker 1	Ranker 2	Ranker 3	Ranker 4	Ranker 5	⋯
Presenter 1	$\boldsymbol{r_{11}}$	r_{12}	r_{13}	r_{14}	r_{15}	⋯
Presenter 2	r_{21}	$\boldsymbol{r_{22}}$	r_{23}	r_{24}	r_{25}	⋯
Presenter 3	r_{31}	r_{32}	$\boldsymbol{r_{33}}$	r_{34}	r_{35}	⋯
Presenter 4	r_{41}	r_{42}	r_{43}	$\boldsymbol{r_{44}}$	r_{45}	⋯
Presenter 5	r_{51}	r_{52}	r_{53}	r_{54}	$\boldsymbol{r_{55}}$	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋮ ⋱

Table: A sample ranking matrix **R** of dimension $n \times n$, where the entry r_{ij} represents the rank given by Ranker j to Presenter i. The diagonal entries ($\boldsymbol{r_{ii}}$) correspond to self-ranks.

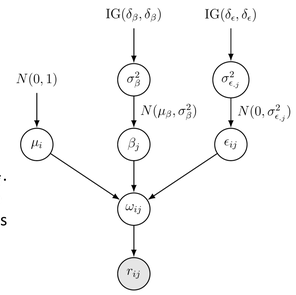
- Assume a **latent importance score**, ω_{ij}, which is associated with the observed ranks r_{ij} of each presenter-ranker pair.
- The unknown true performance, μ_i, is estimated via a probabilistic model that establishes a linear relationship between ω_{ij} and μ_i:

$$\omega_{ij} = \mu_i + I(i=j)\beta_j + \epsilon_{ij}, \text{ where}$$

- $i \in \{1, \dots, n\}$ index presenters being ranked;
- $j \in \{1, \dots, J\}$ index rankers (*note that J = n);
- ω_{ij}: Latent variable of each observed rank;
- μ_i: True academic performance of presenter i;
- β_j: Self-evaluation bias of ranker j with mean μ_β and variance σ_β^2;
- ϵ_{ij}: Random error in the evaluation;
- $\sigma_{\epsilon,j}^2$: The overall variance of each presenter's ranking.

Bayesian Hierarchical Model

Diagram: Bayesian Hierarchical Model. Each observed rank r_{ij} is associated with a latent importance score ω_{ij}, which depends linearly on the true performance μ_i, the self-evaluation bias β_j, and noise ϵ_{ij}. Priors are placed on all variance components and latent variables to capture uncertainty and variation across individuals and rankers.



- **Full probability model** can be expressed as:

$$\boldsymbol{P}(\boldsymbol{r}, \boldsymbol{\Theta}) = \prod_{j=1}^{J}\prod_{i=1}^{n}\{\omega_{(1,j)} < \cdots < \omega_{(n,j)}\} \times \prod_{j=1}^{J}\prod_{i=1}^{n} N(\omega_{ij}|\mu_i + I(i=j)\beta_j, \sigma_{\epsilon_j}^2) \times \prod_{i=1}^{n} N(\mu_i|0,1) \times \prod_{j=1}^{J} N(\beta_j|\mu_\beta, \sigma_\beta^2) \times \mathrm{U}(\mu_\beta|a,b) \times IG(\sigma_\beta^2|\delta_\beta, \delta_\beta) \times \prod_{j=1}^{J} IG(\sigma_{\epsilon,j}^2|\delta_\epsilon, \delta_\epsilon)$$

- **Gibb Sampling [4]** is implemented to generate the posterior samples given the known derived **conditional distributions:**

$$\mu_i \sim N\left(\frac{\sum_{j=1}^{J}\frac{\omega_{ij}-I(i=j)\beta_j}{\sigma_{\epsilon\_j}^2}}{1+\sum_{j=1}^{J}\frac{1}{\sigma_{\epsilon\_j}^2}}, \frac{1}{1+\sum_{j=1}^{J}\frac{1}{\sigma_{\epsilon\_j}^2}}\right), i = 1, \cdots, n;$$

$$\beta_j \sim N\left(\frac{\frac{\omega_{jj}-\mu_j}{\sigma_{\epsilon\_j}^2}+\frac{\mu_\beta}{\sigma_\beta^2}}{\frac{1}{\sigma_{\epsilon\_j}^2}+\frac{1}{\sigma_\beta^2}}, \frac{1}{\frac{1}{\sigma_{\epsilon\_j}^2}+\frac{1}{\sigma_\beta^2}}\right) \times I(-100, 100),\ j = 1, \cdots, J;$$

$$\mu_\beta \sim N\left(\frac{\sum_{j=1}^{J}\beta_j}{J}, \frac{\sigma_\beta^2}{J}\right) \times I(-100, 100);$$

$$\sigma_\beta^2 \sim IG\left(\delta_\beta + \frac{J}{2}, \delta_\beta + \frac{1}{2}\sum_{j=1}^{J}(\beta_j - \mu_\beta)^2\right);$$

$$\sigma_{\epsilon_j}^2 \sim IG\left(\delta_\epsilon + \frac{J}{2}, \delta_\epsilon + \frac{1}{2}\sum_{i=1}^{n}(\omega_{ij} - \mu_i - I(i=j)\beta_j)^2\right), j = 1, \cdots, J;$$

$$\omega_{ij} \sim TN(\mu_i + I(i=j)\beta_j, \sigma_\beta^2; \omega_{(k-1,j)}, \omega_{(k+1,j)}), i = 1, \cdots, n;\ j = 1, \cdots, J.$$

Simulation Study Results

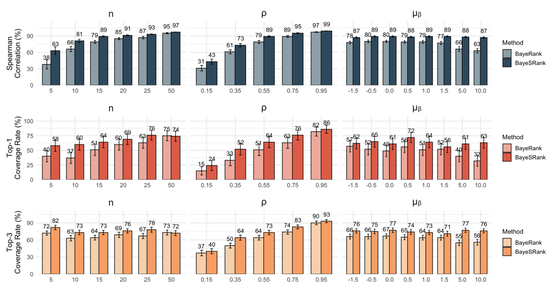


Figure: Comparison of BayeSRank and BayeRank (model without self-evaluation term, β_j) performance across different data-simulation settings. From top to bottom, bar plots display the (1) **Spearman correlation** (in %), (2) **Top-1 coverage rate** (in %), and (3) **Top-3 coverage rate** (in %) for each method across varying values of **n** (number of presenters), **ρ** (correlation between latent variable and true performance, which indicates the ranking quality of rankers), and **μ_β** (mean self-evaluation bias effect). Each bar represents the mean estimate with 95% confidence intervals.

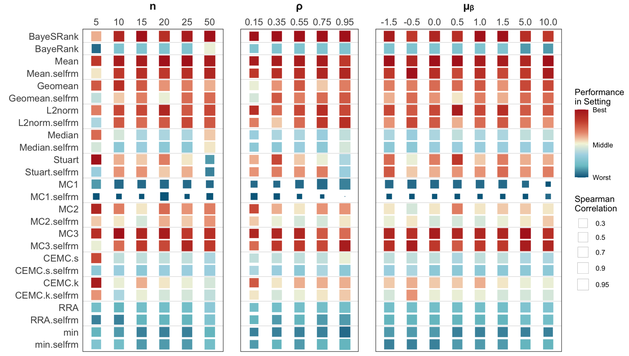


Figure: Spearman correlation results of BayeSRank vs. other methods [5-10] under varying (n, ρ and μ_β) simulation settings. * .selfrm indicates removal of self-valuation data. Within each column, color indicates the relative performance of each method (warmer = better, cooler = worse), while size reflects the magnitude of the average Spearman correlation between estimated and true rankings. BayeSRank consistently ranks among the top-performing methods.

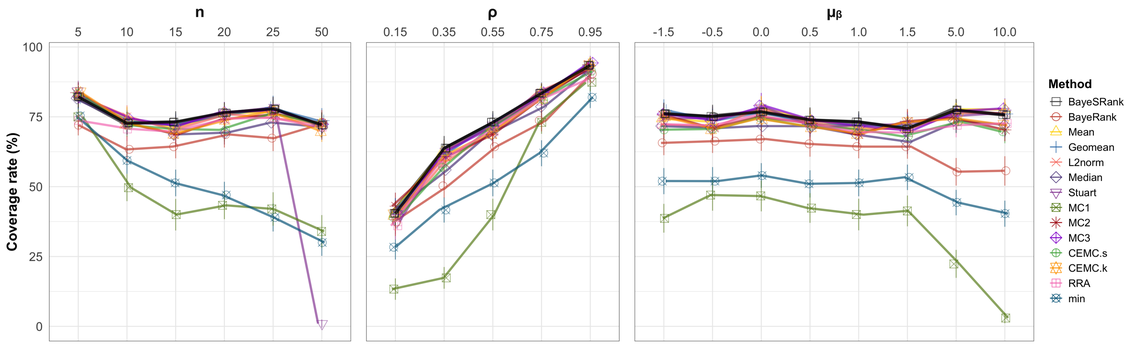
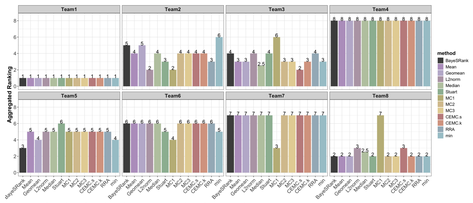


Figure: Top 3 coverage rate (in %) with 95% confidence interval across varying (n, ρ and μ_β) simulation settings. Note that the x-axes are not linearly scaled for n and μ_β, which may visually exaggerate or flatten trends. **BayeSRank (in black)** consistently outperforms or matches other RA methods.

Class Data Example

Aggregated Rankings for 8 Teams by Multiple Methods



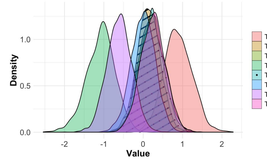


Figure: BayeSRank posterior distribution of true performance ($\mu_i's$).

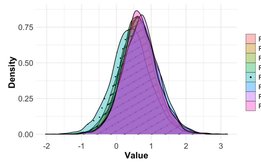


Figure: BayeSRank posterior distribution of self-bias ($\beta_j's$).

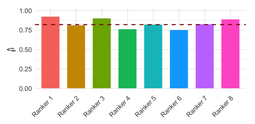


Figure: BayeSRank estimated ranking quality ($\rho_j's$, inversely related to variances).

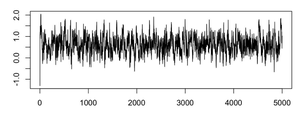


Figure: Trace of BayeSRank posterior samples of mean self-bias (μ_β).

Conclusions

- Study results suggest that removal of self-evaluation data is not the optimal approach, as self-evaluations contribute to ranking accuracy when properly adjusted.
- Through simulations and real-world data examples, we demonstrate BayeSRank's superiority in generating **interpretable, unbiased** evaluations, especially in noisy data settings.
- Our work enhances fairness, transparency, and reliability in peer and self-evaluation systems, offering theoretical and practical implications for bias-aware ranking.

References

1. Borda, JC de. "Mˊemoire sur les' elections au scrutin." Histoire de l'Acad'emie Royale des Sciences (1781).
2. Li, Xue, et al. "A Bayesian latent variable approach to aggregation of partial and top-ranked lists in genomic studies." *Statistics in medicine* 37.28 (2018): 4266-4278.
3. Gramzow, Richard H., et al. "Self-evaluation bias and academic performance: Some ways and some reasons why." Journal of Research in Personality 37.2 (2003): 41-61.
4. Geman, Stuart, and Donald Geman. "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images." IEEE Transactions on pattern analysis and machine intelligence 6 (1984): 721-741.
5. Badgeley, Marcus A., Stuart C. Sealfon, and Maria D. Chikina. "Hybrid Bayesian-rank integration approach improves the predictive power of genomic dataset aggregation." Bioinformatics 31.2 (2015): 209-215.
6. Dwork, Cynthia, et al. "Rank aggregation methods for the web." Proceedings of the 10th international conference on World Wide Web. 2001.
7. Lin, Shili, and Jie Ding. "Integration of ranked lists via cross entropy Monte Carlo with applications to mRNA and microRNA studies." Biometrics 65.1 (2009): 9-18.
8. Kolde, Raivo, et al. "Robust rank aggregation for gene list integration and meta-analysis." Bioinformatics 28.4 (2012): 573-580.
   ❑ R packages:
9. Schimek MG, Budinska E, Ding J, Kugler KG, Svendova V, Lin S, Pfeifer B (2022). _TopKLists: Inference, Aggregation and Visualization for Top-K Ranked Lists_. R package version 1.0.8, <https://CRAN.R-project.org/package=TopKLists>.
10. Kolde R (2022). _RobustRankAggreg: Methods for Robust Rank Aggregation_. R package version 1.2.1, <https://CRAN.R-project.org/package=RobustRankAggreg>.

Contact: haoyao.ruan@uta.edu