Ranking and Rating Rankings and Ratings*

Jingyan Wang, Nihar B. Shah
Carnegie Mellon University
{jingyanw, nihars}@cs.cmu.edu

Abstract

Cardinal scores collected from people are well known to suffer from miscalibrations. A popular approach to address this issue is to assume simplistic models of miscalibration (such as linear biases) to de-bias the scores. This approach, however, often fares poorly because people's miscalibrations are typically far more complex and not well understood. *It is widely believed that in the absence of simplifying assumptions on the miscalibration, the only useful information in practice from the cardinal scores is the induced ranking.* In this paper we address the fundamental question of whether this widespread folklore belief is actually true. We consider cardinal scores with arbitrary (or even adversarially chosen) miscalibrations that is only required to be consistent with the induced ranking. We design rating-based estimators and prove that despite making no assumptions on the ratings, they strictly and uniformly outperform all possible estimators that rely on only the ranking. These estimators can be used as a plug-in to show the superiority of cardinal scores over ordinal rankings for a variety of applications, including A/B testing and ranking. This work thus provides novel fundamental insights in the eternal debate between cardinal and ordinal data: It *ranks* the approach of using ratings higher than that of using rankings, and *rates* both approaches in terms of their estimation errors.

Introduction

"A raw rating of 7 out of 10 in the absence of any other information is potentially useless." (Mitliagkas et al. 2011)

"The rating scale as well as the individual ratings are often arbitrary and may not be consistent from one user to another." (Ammar and Shah 2012)

Consider two items that need to be evaluated (for example, papers submitted to a conference) and two reviewers. Suppose each reviewer is assigned one distinct item for evaluation, and this assignment is done uniformly at random. The two reviewers provide their evaluations (say, in the range $[0, 1]$) for the respective item they evaluate, from which the better item must be chosen. However, the reviewers' rating scales may be miscalibrated. It might be the case that the first reviewer is lenient and always provides scores in $[0.6, 1]$ whereas the second reviewer is more stringent and provides scores in the range $[0, 0.4]$. Or it might be the case that one reviewer is moderate whereas the other is extreme – the first reviewer's 0.2 is equivalent to the second reviewer's 0.1 whereas the first reviewer's 0.3 is equivalent to the second reviewer's 0.8. More generally, the miscalibration of the reviewers may be arbitrary and unknown. Then is there any hope of identifying the better of the two items with any non-trivial degree of certainty?

A variety of applications involve collection and aggregation of human preferences or judgments in terms of cardinal scores (numeric ratings). A perennial problem with eliciting cardinal scores is that of miscalibration – the systematic errors introduced due to incomparability of cardinal scores provided by different people (see (Poston 2008; Griffin and Brenner 2008) and references therein).

This issue of miscalibration is sometimes addressed by making simplifying assumptions about the form of miscalibration, such as linear bias models (Paul 1981; Roos, Rothe, and Scheuermann 2011; Baba and Kashima 2013; Ge, Welling, and Ghahramani 2013; MacKay et al. 2017). However, the calibration issues with human-provided scores are often significantly more complex causing significant violations to these simplified assumptions (see (Griffin and Brenner 2008) and references therein). Moreover, the algorithms for post-hoc correction often try to estimate the individual parameters which may not be feasible due to low sample sizes. For instance, John Langford notes from his experience as the program chair of the ICML 2012 conference:

"We experimented with reviewer normalization and generally found it significantly harmful." (Langford 2012)

It is commonly believed that when unable or unwilling to make simplifying assumptions on the bias in cardinal scores, the only useful information is the ranking of the scores (Rokeach 1968; Freund et al. 2003; Harzing et

*This work was originally published at the International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019) (Wang and Shah 2019). The interested reader is referred to the full version of this paper on arXiv (Wang and Shah 2018) for proofs and additional material.

al. 2009; Mitliagkas et al. 2011; Ammar and Shah 2012; Negahban, Oh, and Shah 2012). This perception gives rise to a second approach towards handling miscalibrations – that of using only the induced ranking or otherwise directly eliciting a ranking and not scores from the use. As noted by Freund et al.:

> *"[Using rankings instead of ratings] becomes very important when we combine the rankings of many viewers who often use completely different ranges of scores to express identical preferences."* (Freund et al. 2003)

These motivations have spurred a long line of literature on analyzing data that takes the form of partial or total rankings of items (Cook et al. 2007; Baskin and Krishnamurthi 2009; Ammar and Shah 2012; Negahban, Oh, and Shah 2012; Rajkumar et al. 2015; Shah et al. 2016; Shah and Wainwright 2018).

In this paper, we contest this widely held belief by addressing the following two **fundamental** questions:

- In the absence of simplifying modeling assumptions on the miscalibration, is there any estimator (based on the scores) that can outperform estimators based on the induced rankings?
- If only one evaluation per reviewer is available, and if each reviewer may have an arbitrary miscalibration, is there hope of estimation better than random guessing?

Our theory shows that the answer to both questions is "Yes". One need not make simplifying assumptions about the miscalibration and yet guarantee a performance superior to that of any estimator that uses only the induced rankings.

In more detail, we consider settings where a number of people provide cardinal scores for one or more from a collection of items. The calibration of each reviewer is represented by an unknown monotonic function that maps the space of true values to the scores given by this reviewer. These functions are arbitrary and may even be chosen adversarially. We present a class of estimators based on cardinal scores given by the reviewers which *uniformly* outperforms any estimator that uses only the induced rankings. A compelling feature of our estimators is that they can be used as a plug-in to improve ranking-based algorithms in a variety of applications, such as A/B testing and ranking.

The techniques used in our analyses draw inspiration from Stein's shrinkage (Stein 1956; James and Stein 1961) and empirical Bayes (Robbins 1956). Our setting with 2 reviewers and 2 papers presented subsequently in the paper carries a close connection to the classic two-envelope problem (for a survey on the two-envelope problem, see (Gnedin 2016)), and our estimator in this setting is similar in spirit to the randomized strategy (Cover 1987) proposed by Thomas Cover.

Our work provides a new perspective on the eternal debate between cardinal scores and ordinal rankings. It is often believed that ordinal rankings are a panacea for the miscalibration issues with cardinal scores. Here we show that *ordinal estimators are not only statistically inadmissible (that is, Pareto-inefficient), they are also strictly and uniformly beaten by our cardinal estimators*. Our results thus uncover a new point on the tradeoff between cardinal and ordinal data collection. The theoretical results and insights established in this paper are envisaged to serve as a crucial building block towards the design of rating-based estimators under more benign assumptions on miscalibrations, and for more complex settings of data collection, in the future.

Preliminaries

Consider a set of n items denoted as $\{1, \ldots, n\}$ or $[n]$ in short. Each item $i \in [n]$ has an unknown value $x_i \in \mathbb{R}$. For ease of exposition, we assume that all items have distinct values. There are m reviewers $\{1, \ldots, m\}$ and each reviewer evaluates a subset of the items. The calibration of any reviewer $j \in [m]$ is given by an unknown, strictly-increasing function $f_j : \mathbb{R} \to \mathbb{R}$. When reviewer j evaluates item i, the reported score is $f_j(x_i)$. We make no other assumptions on the calibration functions $f_1, \ldots, f_m$. We use the notation $\succ$ to represent a relative order of any items, for instance, we use "$1 \succ 2$" to say that item 1 has a greater value (ranked higher) than item 2. We assume that m and n are finite.

Every reviewer is assigned one or more items to evaluate. We denote the assignment of items to reviewers as $A = (S_1, \ldots, S_m)$, where $S_j \subseteq [n]$ is the set of items assigned to reviewer $j \in [m]$. We use the notation Π to represent the set of all permutations of n items. We let $\pi^* \in \Pi$ denote the ranking of the n items induced by their respective values $(x_1, \ldots, x_n)$, such that $x_{\pi^*(1)} > x_{\pi^*(2)} > \cdots > x_{\pi^*(n)}$. The goal is to estimate this underlying "true" ranking π^* from the evaluations of the reviewers. We consider two types of settings: an ordinal setting where estimation is performed using the rankings induced by each reviewer's reported scores, and a cardinal setting where the estimation is performed using the reviewers' scores (which can have an arbitrary miscalibration and only need to be consistent with the rankings). Formally:

- **Ordinal:** Each reviewer j reports a total ranking among the items in S_j, that is, the ranking of the items induced by the values $\{f_j(x_i)\}_{i \in S_j}$. An ordinal estimator observes the assignment A and the rankings reported by all reviewers.
- **Cardinal:** Each reviewer j reports the scores for the items in S_j, that is, the values of $\{f_j(x_i)\}_{i \in S_j}$. A cardinal estimator observes the assignment A and the scores reported by all reviewers.

In order to compare the performance of different estimators, we use the notion of *strict uniform dominance*. Informally, we say that one estimator strictly uniformly dominates another if it incurs a strictly lower risk for all possible choices of the miscalibration functions and the item values.

In more detail, suppose that you wish to show that an estimator $\widehat{\pi}_1$ is superior to estimator $\widehat{\pi}_2$ with respect to some metric for estimating π^*. However, there is a clever adversary who intends to thwart your attempts. The adversary can choose the miscalibration functions of all reviewers $\{f_1, \ldots, f_m\}$, the true ranking π^* of the items, and the values of all items $\{x_1, \ldots, x_n\}$. The only constraints in this choice are that the miscalibration functions $f_1, \ldots, f_m$ must be strictly monotonic and that the item values $x_1, \ldots, x_n$ should induce the ranking π^* (such that $x_{\pi^*(1)} > x_{\pi^*(2)} >$

$\cdots > x_{\pi^*(n)})$. The items are then assigned to reviewers according to the (possibly random) assignment A. The reviewers now provide their ordinal or cardinal evaluations as described earlier, and the two estimators $\widehat{\pi}_1$ and $\widehat{\pi}_2$ use these evaluations to compute their estimates. We say that estimator $\widehat{\pi}_1$ strictly uniformly dominates $\widehat{\pi}_2$, if $\widehat{\pi}_1$ always incurs a strictly smaller (expected) error than $\widehat{\pi}_2$. Formally:

Definition 1 (Strict uniform dominance)**.** *Let $\widehat{\pi}_1$ and $\widehat{\pi}_2$ be two estimators for the true ranking π^*. Estimator $\widehat{\pi}_1$ is said to strictly uniformly dominate estimator $\widehat{\pi}_2$ with respect to a given loss function $L : \Pi \times \Pi \to \mathbb{R}$ if*

$$\mathbb{E}[L(\pi^*, \widehat{\pi}_1)] < \mathbb{E}[L(\pi^*, \widehat{\pi}_2)],$$

for all π^* and all permissible $\{f_1, \ldots, f_m, x_1, \ldots, x_n\}$. The expectation is taken over any randomness in the assignment A and the estimators.

Note that strict uniform dominance is a stronger notion than comparing estimators in terms of their minimax (worst-case) or average-case risks. Moreover, if an estimator $\widehat{\pi}_2$ is strictly uniformly dominated by some estimator $\widehat{\pi}_1$, then the estimator $\widehat{\pi}_2$ is statistically inadmissible (see (Wasserman 2010, Definition 12.17) for the definition of statistical inadmissibility). Finally, for ease of exposition, we focus on the 0-1 loss, $L(\pi^*, \pi) = \mathbb{1}\{\pi^* \neq \pi\}$.

A canonical setting

Consider a canonical setting that involves two items and two reviewers (that is, $n = 2$, $m = 2$), where each reviewer evaluates one of the two items. The ideas in this setting are directly applicable towards designing uniformly superior estimators for other applications.

In this canonical setting, each of the two reviewers evaluates one of the two items chosen uniformly at random without replacement, that is, the assignment A is chosen uniformly at random from the two possibilities $(S_1 = 1, S_2 = 2)$ and $(S_1 = 2, S_2 = 1)$. Since each reviewer is assigned only one item, the ordinal data is vacuous. Then the natural ordinal baseline is an estimator which makes a guess uniformly at random:

$$\widehat{\pi}_{\text{can}}(A, \{\}) = \begin{cases} 1 \succ 2 & \text{with probability } 0.5 \\ 2 \succ 1 & \text{with probability } 0.5. \end{cases}$$

In the cardinal setting, let y_1 denote the score reported for item 1 by its respective reviewer, and let y_2 denote the score for item 2 reported by its respective reviewer. Since the calibration functions are arbitrary (and may be adversarial), it appears hopeless to obtain information about the relative ordering of x_1 and x_2 from just this data. Indeed, as we show below, standard estimators such as the sign test — ranking the items in terms of their reviewer-provided scores — provably fail to achieve this goal. More generally, the following theorem holds for all deterministic estimators, that is, estimators given by deterministic mappings from $\{A, y_1, y_2\}$ to the set $\{1 \succ 2, 2 \succ 1\}$.

Theorem 2. *No deterministic (cardinal or ordinal) estimator can strictly uniformly dominate the random-guessing estimator $\widehat{\pi}_{can}$.*

This theorem demonstrates the difficulty of this problem by ruling out all deterministic estimators. Our original question still remains: is there any estimator that can strictly uniformly outperform the random-guessing ordinal baseline?

We show that the answer is yes, with the construction of a randomized estimator denoted as $\widetilde{\pi}^{\text{our}}_{\text{can}}$. This estimator is based on a function $w : [0, \infty) \to [0, 1)$ which may be chosen as any arbitrary strictly-increasing function. For instance, one could choose $w(x) = \frac{x}{1+x}$ or w as the sigmoid function. Given the scores y_1, y_2 reported for the two items, let $\widehat{i}^{(1)} \in \operatorname{argmax}_{i \in \{1,2\}} y_i$ denote the item which receives the higher score, and let $\widehat{i}^{(2)}$ denote the remaining item (with ties broken uniformly). Then our randomized estimator outputs:

$$\widetilde{\pi}^{\text{our}}_{\text{can}}(A, y_1, y_2) = \begin{cases} \widehat{i}^{(1)} \succ \widehat{i}^{(2)} & \text{with probability } \frac{1+w(|y_1 - y_2|)}{2} \\ \widehat{i}^{(2)} \succ \widehat{i}^{(1)} & \text{otherwise.} \end{cases} \quad (1)$$

Note that the the output of this estimator is independent of the assignment A.

As an example, suppose that the values of the two items are $(x_1 = 4, x_2 = 7)$. Suppose the calibration function f_1 of reviewer 1 maps the values of these two items to $(f_1(x_1) = 1, f_1(x_2) = 5)$, and the calibration function f_2 of reviewer 2 maps them to $(f_2(x_1) = 6, f_2(x_2) = 8)$. Now we observe the ratings $(y_1 = 1, y_2 = 8)$ with probability 0.5, in which case the estimator reports item 2 as greater with probability $\frac{1+w(7)}{2}$. With probability 0.5, we observe $(y_1 = 6, y_2 = 5)$, in which case the estimator reports item 2 as greater with probability $1 - \frac{1+w(1)}{2} = \frac{1-w(1)}{2}$. Since the function w is strictly-increasing, we have $w(7) > w(1)$. Using this fact and averaging the outcomes over these two cases yields a probability of success strictly greater than 0.5. The following theorem now proves this result formally.

Theorem 3. *The randomized estimator $\widetilde{\pi}^{our}_{can}$ strictly uniformly dominates the random-guessing baseline $\widehat{\pi}_{can}$.*

The contrast between deterministic estimators and randomized estimators arises from the fact that a deterministic estimator "commits" to an action (deciding which item has a greater value). It performs well if the situation is aligned with this action (when the scores under miscalibration are consistent with the true ordering of the two items). However, due to its prior commitment it may fail if the situation is not aligned. In contrast, a randomized estimator balances out good and bad cases. The probability of the good case (correct estimation) is greater than the probability of the bad case (incorrect estimation) for the randomized estimator (1), because it exploits the monotonic structure of the calibration functions, whereas this structure is lost in ordinal data.

Additional results in the arXiv version

The analysis for the canonincal setting conveys the key ideas underlying more general results. We now outline additional material that is included in the arXiv version of this paper (Wang and Shah 2018):

- **Noisy setting:** a setting with noisy observations ($y = f(x) + \text{noise}$).

- **A/B testing and ranking:** results on the superiority of cardinal data over ordinal data by using the proposed estimator (1) in the canonical setting as a plug-in component in two applications, A/B testing and ranking.
- **Simulation:** simulation on A/B testing and ranking, and simulation on the tradeoff between estimation under perfect calibration vs. miscalibration.
- **Related work:** connections between our results and prior work, in particular the connection to Cover's estimator for the two-envelope problem (Cover 1987).
- **Proofs:** the proofs of all theoretical results.

Acknowledgments

This work was supported in part by NSF grants CRII: CIF: 1755656 and CCF: 1763734. The authors thank Bryan Parno for very useful discussions on biases in conference peer review, and Pieter Abbeel for pointing out the related work on the two-envelope problem.

References

Ammar, A., and Shah, D. 2012. Efficient rank aggregation using partial data. In *SIGMETRICS/PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems*.

Baba, Y., and Kashima, H. 2013. Statistical quality estimation for general crowdsourcing tasks. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Baskin, J. P., and Krishnamurthi, S. 2009. Preference aggregation in group recommender systems for committee decision-making. In *ACM Conference on Recommender Systems*.

Cook, W. D.; Golany, B.; Penn, M.; and Raviv, T. 2007. Creating a consensus ranking of proposals from reviewers' partial ordinal rankings. *Computers & Operations Research*.

Cover, T. M. 1987. *Pick the Largest Number*. Springer New York. 152–152.

Freund, Y.; Iyer, R. D.; Schapire, R. E.; and Singer, Y. 2003. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*.

Ge, H.; Welling, M.; and Ghahramani, Z. 2013. A Bayesian model for calibrating conference review scores. http://mlg.eng.cam.ac.uk/hong/unpublished/nips-review-model.pdf [Online; accessed 03/01/2019].

Gnedin, A. 2016. Guess the larger number. *preprint arXiv:1608.01899*.

Griffin, D., and Brenner, L. 2008. *Perspectives on Probability Judgment Calibration*. Wiley-Blackwell. chapter 9.

Harzing, A.-W.; Baldueza, J.; Barner-Rasmussen, W.; Barzantny, C.; Canabal, A.; Davila, A.; Espejo, A.; Ferreira, R.; Giroud, A.; Koester, K.; et al. 2009. Rating versus ranking: What is the best way to reduce response and language bias in cross-national research? *International Business Review*.

James, W., and Stein, C. 1961. Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1.

Langford, J. 2012. ICML acceptance statistics. http://hunch.net/?p=2517 [Online; accessed 05/14/2018].

MacKay, R. S.; Kenna, R.; Low, R. J.; and Parker, S. 2017. Calibration with confidence: a principled method for panel assessment. *Royal Society Open Science*.

Mitliagkas, I.; Gopalan, A.; Caramanis, C.; and Vishwanath, S. 2011. User rankings from comparisons: Learning permutations in high dimensions. In *Allerton Conference on Communication, Control, and Computing*.

Negahban, S.; Oh, S.; and Shah, D. 2012. Iterative ranking from pair-wise comparisons. In *Advances in Neural Information Processing Systems*.

Paul, S. R. 1981. Bayesian methods for calibration of examiners. *British Journal of Mathematical and Statistical Psychology*.

Poston, R. S. 2008. Using and fixing biased rating schemes. *Commun. ACM*.

Rajkumar, A.; Ghoshal, S.; Lim, L.-H.; and Agarwal, S. 2015. Ranking from stochastic pairwise preferences: Recovering Condorcet winners and tournament solution sets at the top. In *International Conference on Machine Learning*.

Robbins, H. 1956. An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1.

Rokeach, M. 1968. The role of values in public opinion research. *Public Opinion Quarterly*.

Roos, M.; Rothe, J.; and Scheuermann, B. 2011. How to calibrate the scores of biased reviewers by quadratic programming. In *AAAI Conference on Artificial Intelligence*.

Shah, N. B., and Wainwright, M. J. 2018. Simple, robust and optimal ranking from pairwise comparisons. *Journal of Machine Learning Research*.

Shah, N. B.; Balakrishnan, S.; Bradley, J.; Parekh, A.; Ramchandran, K.; and Wainwright, M. J. 2016. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. *Journal of Machine Learning Research*.

Stein, C. 1956. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1.

Wang, J., and Shah, N. B. 2018. Your 2 is my 1, your 3 is my 9: Handling arbitrary miscalibrations in ratings. *preprint arXiv:1806.05085*.

Wang, J., and Shah, N. B. 2019. Your 2 is my 1, your 3 is my 9: Handling arbitrary miscalibrations in ratings. In *International Conference on Autonomous Agents and Multiagent Systems*.

Wasserman, L. 2010. *All of Statistics: A Concise Course in Statistical Inference*. Springer Publishing Company, Incorporated.