Data Mining and Knowledge Discovery (2024) 38:2062–2086
https://doi.org/10.1007/s10618-024-01024-z

Bias-aware ranking from pairwise comparisons

Antonio Ferrara[1,2,3] · Francesco Bonchi[1,4] · Francesco Fabbri[5] · Fariba Karimi[3,6] · Claudia Wagner[2,7]

Received: 5 December 2023 / Accepted: 14 April 2024 / Published online: 31 May 2024

Abstract
Human feedback is often used, either directly or indirectly, as input to algorithmic decision making. However, humans are biased: if the algorithm that takes as input the human feedback does not control for potential biases, this might result in biased algorithmic decision making, which can have a tangible impact on people's lives. In this paper, we study how to detect and correct for evaluators' bias in the task of *ranking people (or items) from pairwise comparisons*. Specifically, we assume we are given pairwise comparisons of the items to be ranked produced by a set of evaluators. While the pairwise assessments of the evaluators should reflect to a certain extent the latent (unobservable) true quality scores of the items, they might be affected by each evaluator's own bias against, or in favor, of some groups of items. By detecting and amending evaluators' biases, we aim to produce a ranking of the items that is, as much as possible, in accordance with the ranking one would produce by having access to the latent quality scores. Our proposal is a novel method that extends the classic Bradley-Terry model by having a bias parameter for each evaluator which distorts the true quality score of each item, depending on the group the item belongs to. Thanks to the simplicity of the model, we are able to write explicitly its log-likelihood w.r.t. the parameters (i.e., items' latent scores and evaluators' bias) and optimize by means of the alternating approach. Our experiments on synthetic and real-world data confirm that our method is able to reconstruct the bias of each single evaluator extremely well and thus to outperform several non-trivial competitors in the task of producing a ranking which is as much as possible close to the unbiased ranking.

Keywords Pairwise comparisons · Evaluators' bias · Rankings · Bradley-Terry · Fairness

Responsible editor: Rita P. Ribeiro.

Extended author information available on the last page of the article

1 Introduction

Human decision bias is well-studied in the field of human computation (Chen et al. 2013; Kamar et al. 2015; Hube et al. 2019; Almaatouq et al. 2020; Liu et al. 2022): human characteristics, opinions, cognitive and social biases, as well as the way the human computation task is formulated, can result in biased human feedback. In turn, if the algorithm that takes as input the human feedback does not take into account and control for such potential biases, this might result in biased algorithmic decision making. In this paper, we focus on the task of ranking humans (or items) starting from pairwise comparisons, i.e., a collection of triples $\langle i, j, e \rangle$ indicating that, according to the evaluator e, the item i is superior to j. Pairwise comparisons are widely used to collect relevance feedback from humans since comparative assessments of two items are easy and fast to obtain: this type of human feedback is typically collected by means of crowdsourcing experiments in which crowdworkers are presented with a pair of items/people from which they should select the more relevant one for a given task (e.g., searching for information about a certain topic, or selecting people for a given position). Alternatively, indirect relevance feedback can be inferred from human behavior (e.g., their click behavior, or the time they spend exploring certain options). Pairwise comparisons have been used in various contexts including hiring at scale (Kotturi et al. 2020; Sarma et al. 2016; Kuttal et al. 2021), investigating aspects of the labor activity (Koshkalda et al. 2020), assessing political biases (Kuo et al. 2020), and comparing political texts (Carlson and Montgomery 2017). Furthermore, various methods have been proposed in the literature to rank from pairwise comparisons (Bradley and Terry 1952; Negahban et al. 2012; Chen et al. 2013). However, pairwise comparisons, based on human judgements, can be affected by the quality of the evaluators (Chen et al. 2013; Bugakova et al. 2019) and by various types of biases, including ordering effects and presentation biases (Bugakova et al. 2019; Beaver and Gokhale 1975; Davidson and Beaver 1977).

Of special interest for this paper, *in-group favoritism* and *out-group prejudice* are two prominent bias mechanisms that spring into action when humans judge people or items that belong to certain identity groups such as nationality, gender, or even political parties. These bias mechanisms produce systematic deviations from rational decision making that puts one group in an unfavorable position by systematically skewing the decisions against items/people of this group even if they objectively should be preferred. Our goal is to develop a method to accurately measure and account for group biases in pairwise comparisons.

More formally, we assume to have a collection of items (or people) to be ranked, and we assume that we are given pairwise comparisons of the items produced by a set of evaluators. While the pairwise assessments of the evaluators should reflect to a certain extent the latent (unobservable) true quality scores of the items, they might be affected by each evaluator's own bias against, or in favor, of some groups of items. By detecting and amending evaluators' biases, we aim to produce a ranking of the items that is, as much as possible, in accordance with the unbiased ranking, i.e., the one we would produce by having access to the latent quality scores.

Our proposed method, dubbed BARP (Bias-Aware Ranker from Pairwise comparisons), is based on a novel model that enhances the Bradley-Terry (BT) model (Bradley and Terry 1952) by having a bias parameter for each evaluator which distorts the true quality score of each item, depending on the group the item belongs to. Thanks to the simplicity of the model, we are able to write explicitly its log-likelihood w.r.t. the parameters (i.e., items' latent scores and evaluators' bias). Then, the disentanglement of the true latent scores from the evaluators' bias can be achieved by maximum likelihood estimation.

The main principles and concepts of our problem and our proposed method are depicted in Fig. 1.

Interestingly and contrarily to most of the fair ranking literature, our method does not require to indicate any group as protected, instead, all groups are treated equivalently and the method is able to detect and fix bias in favor or against any group and without any prior information of the evaluators preferences.

We assess empirically, on both synthetic and real-world datasets, the performance of BARP in detecting evaluators' bias, thus producing a ranking that is, as much as possible, close to the unbiased ranking. We compare BARP with several non-trivial baselines, including the BT model (Bradley and Terry 1952) (which does not take into account the possibility of bias), some recent variants of the BT model, which take into consideration some forms of evaluators' quality and bias (Chen et al. 2013; Bugakova et al. 2019), spectral methods for ranking from pairwise comparison (Negahban et al. 2012), and methods for fair ranking (Zehlike et al. 2017) used as post-processing in combination with the BT model.

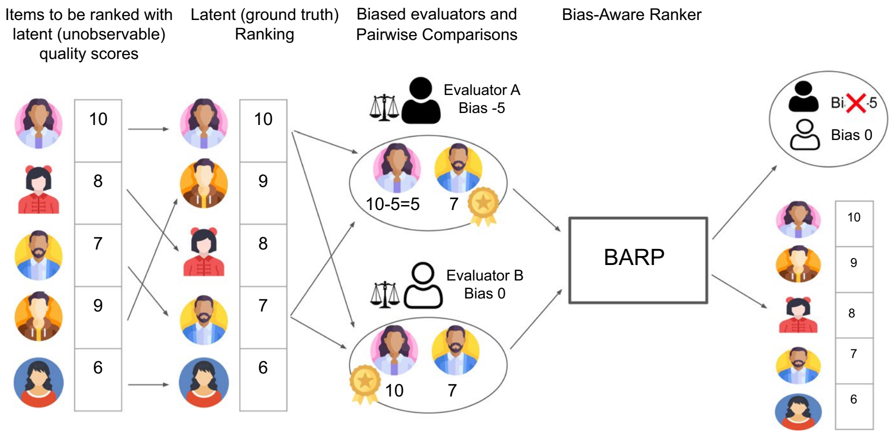
Items to be ranked with latent (unobservable) quality scores
Latent (ground truth) Ranking
Biased evaluators and Pairwise Comparisons
Bias-Aware Ranker
10
8
7
9
6
10
9
8
7
6
Evaluator A
Bias -5
10-5=5
7
Evaluator B
Bias 0
10
7
BARP
Bi 5
Bias 0
10
9
8
7
6

Fig. 1 Depiction of the main principles and concepts of the problem and the proposed method, BARP. We are given a set of items to be ranked, in this case people, that belong to two groups: in this example, we consider binary gender as the attribute defining the two groups. Each items has a latent (unobservable) score which implicitly defines a ground truth (correct) ranking that we would like to produce, but we can not directly because we can not observe the latent scores. Instead, what we are given as input is a set of pairwise comparisons performed by some evaluators. The pairwise comparisons may be affected by the evaluators' own biases in favour or against a specific group. For instance, in the figure, Evaluator A has a bias against female candidates, while Evaluator B has no bias. As an effect of the implicit bias of Evaluator A, their pairwise comparisons are likely not to correctly reflect the true value of the candidates. BARP, by analyzing many input pairwise comparisons by many different evaluators, is able to detect and estimate evaluators' bias and to take this into account while producing a final ranking that is as close as possible to the ground-truth one

Our experiments on synthetic data (with ground-truth evaluators' bias) confirm that BARP is able to reconstruct the bias of each single evaluator extremely well ($MSE < 0.3$, w.r.t. evaluators' bias uniformly distributed in $[-5, 5]$). Thanks to this, the ranking produced by BARP is much closer to the unbiased ranking than those produced by all the baselines: the stronger is the evaluators' bias, the larger the performance gap between our method and the baselines.

Our experiments on real-world data demonstrate the utility of BARP in identifying otherwise unknown biased evaluators. Utilizing the IMDB-WIKI-SbS dataset, which comprises pairwise comparisons of face snapshots, BARP effectively identify evaluators who frequently misperceive the ages of males compared to females, and vice versa. Furthermore, we showcase the applicability of our method in the context of admissions at law schools. We illustrate how BARP can mitigate differences in the rankings of individuals from different groups.

The rest of the paper is organized as follows. The next section presents a survey of the related literature. Section 3 introduces the formal problem statement, while Sect. 4 describes the BARP model, the learning method and its derivation. Section 5 contains the experiments on synthetic data. Results on real-word datasets are presented in Sect. 6. Finally, in Sect. 7, we conclude with a discussion on the limitations of our work and future research pathways.

2 Related work

Ranking from pairwise comparisons dates back to Kendall and Smith (1940) and, in the course of time, different methods have been proposed. Early works include counting and heuristic methods, such as David's score (David 1987). Seminal works, grounded in statistical and probability methods, like Bradley-Terry (1952) and Thurstone (1927) models, make distributional assumptions on the relationship between the comparisons and the ranking. The item's scores and ranking can, then, be recovered with maximum likelihood optimizations. Other methods exploit the interpretation of pairwise comparison as a directed graph, where nodes represent items and directed edges represent pairwise comparisons, leading to the use of random walk and spectral-based methods for ranking items. Examples in this category are RankCentrality (Negahban et al. 2012), SerialRank (Fogel et al. 2014), and GNNRank (He et al. 2022). Furthermore, Ranking from pairwise comparisons is substantially different from learning to rank, even when a pairwise learning to rank approach is employed. The goal of learning to rank is to construct a ranking model to rank new, unseen items by exploiting the information about features learned in the training data. Ranking from pairwise comparisons, instead, aims at constructing a rank of the available items, having at disposal only limited information about pairwise comparisons among them.

Several biases can affect pairwise comparisons. One of these is the order of presentation of the items in the pair. For example, in an experiment investigating a subject's sensitivity to small electrical shocks, the response to the second shock could be strongly modified by the first shock (Beaver and Gokhale 1975). In the literature, variations of the Bradley-Terry model have been proposed to model and account for such ordering effects (Beaver and Gokhale 1975; Davidson and Beaver 1977). An additional series of effects relevant to our study are related to the evaluators and their behaviors. Evaluators's biases have received a lot of interest in other fields, for example in Natural Language Processing. How evaluators' demographics and beliefs can bias toxic language detection is pointed out in Sap et al. (2021), Liu et al. (2022), while Sap et al. (2019) highlights how insensitivity of evaluators to differences in dialect can lead to racial bias in automatic hate speech detection models. Finally, Geva et al. (2019) shows how the use of evaluator identifiers as features and the use of models that are able to recognize the most productive evaluators improves performances of Natural Language Processing models in various language understanding tasks.

Instead, the role of evaluators' biases in crowdsourced pairwise comparisons didn't receive an adequate attention. The main method that considers evaluators' quality is CrowdBT (Chen et al. 2013), where a parameter models evaluators' quality by quantifying the probability that an evaluator answers sincerely or not to the comparison tasks. Furthermore, evaluators' bias is considered in FactorBT (Bugakova et al. 2019). Similarly to FactorBT, our methods deal with evaluators' group biases, but with various significant differences. FactorBT models the probability that an evaluator, instead of basing their choices on the scores of the items of a pair, chose an item because of certain characteristics of the item, for example, being placed at the top of a screen. In FactorBT, how the perception of the scores is affected by evaluators' bias and how the perception of the scores relates to the probability that an item is selected are not considered. Instead, we model evaluators' behaviour with a parameter that directly relates to the perceived scores of the items and we establish a connection between the variation of the perceived score due to group biases and the probability of an item being selected. Furthermore, FactorBT analyzes the problem mainly from an accuracy perspective, while we investigate how our method affects the exposure and ranking of the different groups and we show how our method can be used to estimate the group bias of each individual evaluator.

Lastly, we want to point out the connection between the literature on fair ranking and our work. In a ranking, the desired good for an individual is to be ranked as higher as possible, and, in broad terms, fair ranking tries to avoid members of certain groups being systematically ranked lower than those of privileged groups (Zehlike et al. 2017). In the last years, various methods have been developed to achieve fair ranking, such as Singh and Joachims (2018), Zehlike and Castillo (2020), Celis et al. (2017), García-Soriano and Bonchi (2021). However, many of these methods are focused on learning to rank and are not well tailored to the task of ranking from pairwise comparisons considered in this paper.

3 Problem statement

We consider a collection of n items $I = \{1, \dots, n\}$ where each item $i \in I$ belongs to one group $g_i \in G$ (where G is a discrete set), and has a latent (*unobservable*) quality score $s_i \in \mathbb{R}$. We denote $\mathbf{g}$ and $\mathbf{s}$ the vectors of groups membership and quality scores for all items.

We are given a set of pairwise comparisons among the items in I, produced by a set E of m evaluators. Each evaluator receives pairs of items to evaluate: we denote the set of labeled pairs by the k-th evaluator as $Q_k = \{(i,j) \in I \times I : i \succ_k j\}$, where $\succ_k$ is a relation, with $i \succ_k j$ meaning that the k-th evaluator preferred the object i to the object j. Let Q be the multi-set of all the pairwise comparisons. In this paper, we assume that the multi-set Q is noisy and potentially inconsistent. Although the pairwise comparisons should reflect to a certain extent the latent quality scores of the items, the input comparisons might be affected by each evaluator's own bias against, or in favor, some groups in G. Moreover, inconsistent information might exist in Q, such as, for instance, $i \succ_{k_1} j$ and $j \succ_{k_2} i$ for two different evaluators $k_1, k_2 \in E$. Another example of inconsistent information is $i \succ_k j$, $j \succ_k l$, and $l \succ_k i$.

The problem we tackle in this paper is to produce a ranking r^* of the items in I, that corrects for the evaluators bias, i.e., it is as close as possible to the ranking induced by the latent quality scores, denoted $r(\mathbf{s})$. We will hence evaluate the quality of the ranking r^* in terms of Kendall's Tau correlation (Kendall 1938, 1945) with $r(\mathbf{s})$.

4 Method

In order to solve our problem we need to relate the scores $\mathbf{s}$ and the pairwise comparisons Q. In devising our method, we build on top of the classic Bradley-Terry (BT) model (Bradley and Terry 1952). The BT model assumes that the probability that item i is preferred over j is defined as:

$$P(i \succ j) = \frac{e^{s_i}}{e^{s_i} + e^{s_j}}, \tag{1}$$

where e^x is the exponential function.

This probability has the following meaning: when evaluators are facing the decision of preferring one item over another, given the fact that they observe the item but they don't know the true score of the item, they prefer the correct item (i.e., the item with unknown higher score) with a certain probability that depends on the unknown scores. For example, when two items are similar (i.e. have similar unknown scores), it is hard for an evaluator to chose the correct one, while when one items is significantly better than the other, the evaluator is able to chose the best one correctly with a high probability.

Given the observed pairwise comparisons and the BT relational assumption, it is possible to estimate the underlying unknown scores according to the maximum likelihood principle. Indeed, the log-likelihood l of the scores can then be written as:

$$l(\mathbf{s}) = \sum_{(i,j)\in Q} \log\left(\frac{e^{s_i}}{e^{s_i} + e^{s_j}}\right), \tag{2}$$

and $\hat{s} = \arg\max_{\mathbf{s}} l(\mathbf{s})$ represents the vector of scores that better approximates the latent, unobserved vector of scores $\mathbf{s}$. The maximum of the log-likelihood can be found with standard numerical methods.

We observe that in the Bradley-Terry model, all evaluators are treated equally, and the model does not account for any differences in the quality of their contributions.

Hence, with some further assumptions, different behaviors of the evaluators can be modeled. For instance, in CrowdBT (Chen et al. 2013) it is assumed that evaluators might respond sincerely, randomly, or that they could be malicious or poorly informed. In particular, in CrowdBT, there is an additional parameter that models the probability that each evaluator agrees with the true pairwise preference, and allows for an interpolation between the possible behaviors above.

In our work, we focus on evaluators' behavior but from a different perspective. We assume that the perception of each evaluator is affected by the group membership of an item. For example, if evaluators, in a hiring scenario, exhibit a *gender bias* they might consider more frequently men as more fit for the position than women, when confronted with mixed-gender pairs. Essentially, we assume that evaluators can have an implicit preference for items belonging to a group which may lead to biased relevance feedback.

4.1 Single binary attribute

For sake of simplicity of exposition, we start presenting the case of one single binary attribute (for example a binary version of the attribute gender), which induces two groups (males and females). Later we will extend our model to deal with multiple attributes and with non-binary attributes. It is worth stressing that *our method does not require to indicate any group as protected*: all groups are treated equivalently and the method is able to detect and fix bias in favor or against any group and without any prior information of the evaluators preferences.

In the binary setting, we model the bias of each evaluator $k \in E$ with a bias parameter $\theta_k \in \mathbb{R}$. Taken one of the two groups as reference (it is not relevant which one), the interpretation is that an evaluator perceives the scores of the item i as $s_i + \theta_k$, if the item i belongs to the reference group, or equivalently, $s_i - \theta_k$ if the item belongs to the other group. In the gender example, suppose we take the group females as reference: the evaluator $k \in E$ with a bias parameter $\theta_k \in \mathbb{R}$, in a direct comparison between a male and a female, will be biased in favor of the female if $\theta_k > 0$, or in favor of the male if $\theta_k < 0$.

Formally, let $\gamma_k = e^{\theta_k}$, the probability that item i is preferred over j by the k-th evaluator is as follows:

$$P(i \succ_k j) = \frac{\gamma_k^{\delta_{g_i g}} e^{s_i}}{\gamma_k^{\delta_{g_i g}} e^{s_i} + \gamma_k^{\delta_{g_j g}} e^{s_j}}, \tag{3}$$

where δ is the Kronecker delta, that is $\delta_{g_i g} = 1$, if $g_i = g$ and $\delta_{g_i g} = 0$, if $g_i \neq g$. Such a multiplicative factor on the probability (or equivalently an additive factor on the scores) is grounded in the statistical literature of pairwise comparisons, as Davidson and Beaver (1977) used an analog parameter to model the effect of the order of presentation within the pairs, sometimes also referred to as the home advantage effect. Here, we take a similar functional form and, instead of using it to account for order of presentation effects, we extend it to account for group bias in the evaluations.

Furthermore, one may notice that when $g_i = g_j$ the factor γ_k cancels out, correctly reflecting the assumption that the evaluators show only a preferential behavior when faced with the evaluation of items belonging to two different classes, and not affecting within classes evaluations.

Equation 3, can also be written in terms of θ_k as:

$$P(i \succ_k j) = \frac{e^{s_i+\theta_k\delta_{g_ig}}}{e^{s_i+\theta_k\delta_{g_ig}} + e^{s_j+\theta_k\delta_{g_jg}}}, \tag{4}$$

This corresponds to Eq. 1, where, instead of the original scores s_i, we have biased scores $s_i + \theta_k\delta_{g_ig}$ resulting from biased evaluations. This probability has the following meaning: when evaluators are facing the decision of preferring one item over another, they prefer the correct item with a certain probability that depends, not directly on the unknown scores as in BT model, but rather on their biased perception of the unknown scores.

Disentangling the contribution of the scores and the group bias of the evaluators, we can obtain an unbiased estimation of the unknown scores. Indeed, given the observed pairwise comparisons Q_k for each evaluator k, one can compute $\hat{\mathbf{s}}$ (i.e. the estimate of the scores $\mathbf{s}$) and the estimate of θ_k (we will call them $\hat{\theta}_k$) via a maximum likelihood estimation. In particular, the log-likelihood for the parameters $\mathbf{s}$ and $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)$ can be written as:

$$l(\mathbf{s}, \boldsymbol{\theta}) = -\sum_{k=1}^{m} \sum_{(i,j)\in Q_k} \log\left(1 + e^{-\left(s_i+\theta_k\delta_{g_ig}-s_j-\theta_k\delta_{g_jg}\right)}\right). \tag{5}$$

One way to optimize the objective function $l(\mathbf{s}, \boldsymbol{\theta})$ is to use the alternating variables approach, also referred to as the coordinate descend or coordinate search method, Wright (2015), Nocedal and Wright (1999), which involves the iterative repetition of the following two steps. Firstly, keep $\boldsymbol{\theta}$ constant and optimize over $\mathbf{s}$. Then, keep $\mathbf{s}$ constant and optimize over $\boldsymbol{\theta}$. In particular, we used the equation of the first iteration of BFGS (Nocedal and Wright 1999) to determine, at each step, the parameters updates for both $\mathbf{s}$ and $\boldsymbol{\theta}$, with the gradients explicitly expressed as:

$$\begin{aligned}\frac{dl}{ds_i}(\mathbf{s}, \boldsymbol{\theta}) = \sum_{k=1}^{m}\Bigg(&\sum_{i:(i,j)\in Q_k} \left(1 + e^{\left(s_i+\theta_k\delta_{g_ig}-s_j-\theta_k\delta_{g_jg}\right)}\right)^{-1} \\ &- \sum_{i:(j,i)\in Q_k} \left(1 + e^{\left(s_i+\theta_k\delta_{g_ig}-s_j-\theta_k\delta_{g_jg}\right)}\right)^{-1}\Bigg),\end{aligned} \tag{6}$$

where the sum on all $i : (i,j) \in Q_k$ means the sum on all i such that $i \succ_k j$ (by the definition of Q_k), and

$$\frac{dl}{d\theta_k}(\mathbf{s}, \boldsymbol{\theta}) = \sum_{(i,j)\in Q_k} \frac{\delta_{g_ig} - \delta_{g_jg}}{\left(1 + e^{\left(s_i+\theta_k\delta_{g_ig}-s_j-\theta_k\delta_{g_jg}\right)}\right)}. \tag{7}$$

The pseudocode of our heuristic method, dubbed BARP (Bias-Aware Ranker from Pairwise comparisons), is summarized in Algorithm 1. BARP takes as input, the collection of items I, together with the classes that they belong $\mathbf{g}$, the set of pairwise comparison Q, the number of evaluators m. As stopping criteria, it also takes thresholds for the gradients and a maximum number of iterations. It outputs the maximum likelihood estimation of the scores $\hat{\boldsymbol{s}}$, from which the ranking of the object r^* is inferred, and the maximum likelihood estimation of the bias parameters of the evaluators $\hat{\boldsymbol{\theta}}$.

The algorithm starts by initializing $\mathbf{s}^{(0)}$ and $\boldsymbol{\theta}^{(0)}$, until the maximum number of iterations or the stopping criteria are met, the algorithm updates the estimate for $\mathbf{s}$ and $\boldsymbol{\theta}$ with the following procedure. The algorithm computes the gradient with respect to $\mathbf{s}$ and uses it to compute a one step update for $\mathbf{s}$. Then, the gradient of the log-likelihood with respect to the bias parameter vector $\boldsymbol{\theta}$ is computed and used to update $\boldsymbol{\theta}$, in a similar way of what done for $\mathbf{s}$. The algorithm is flexible on how to exactly compute the updates of the parameters.

In details, we initialize $\mathbf{s}^{(0)}$ and $\boldsymbol{\theta}^{(0)}$ to the 0 vectors, alternatively the vectors could take any random initialization, we set the maximum number of iterations to 1000, and, as further stopping criteria, we stop if the norms of the gradient vectors for $\mathbf{s}$ and $\boldsymbol{\theta}$ are both smaller than 10^{-5}.

Algorithm 1 BARP

```
Input  : $I$,$\mathbf{g}$,$Q$, $m$, stopping criteria, max_iters
Output: $\hat{\boldsymbol{s}}$,$\hat{\boldsymbol{\theta}}$
$\mathbf{s}^{(0)}, \boldsymbol{\theta}^{(0)} \leftarrow$ initialization;
for $t \leftarrow 0$ to $(max\_iters - 1)$ do
    for $i \leftarrow 0$ to $(n - 1)$ do
        compute $\frac{dl}{ds_i}(\mathbf{s}^{(t)}, \boldsymbol{\theta}^{(t)})$ from eq. 6;
    end
    compute $\mathbf{s}^{(t+1)}$ using $\mathbf{s}^{(t)}$ and the gradient $\frac{dl}{d\mathbf{s}}$ ;
    for $k \leftarrow 0$ to $(m - 1)$ do
        compute $\frac{dl}{d\theta_k}(\mathbf{s}^{(t+1)}, \boldsymbol{\theta}^{(t)})$ from eq. 7;
    end
    compute $\boldsymbol{\theta}^{(t+1)}$ using $\boldsymbol{\theta}^{(t)}$ and the gradient $\frac{dl}{d\boldsymbol{\theta}}$ ;
    if stopping criteria are met then
        Break
    end
end
return $\hat{\boldsymbol{s}} = \mathbf{s}^{(t+1)}$ and $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(t+1)}$
```

4.2 Multiple and non-binary attributes

We next extend BARP to account for the presence of multiple binary attributes and non-binary attributes.

Consider q binary attributes and for each item i, let $\vec{g}_i \in \{0,1\}^q$ be the vector of length q, where the t-th component is 1 if i belongs to the reference group for the t-th attribute and 0 otherwise. Similar to the single binary case, we model the behavior of the k-th evaluator by a bias parameter vector of length q, $\vec{\theta}_k \in \mathbb{R}^q$. Equation 4, can hence be rewritten to account for the multiple binary attributes as:

$$P(i \succ_k j) = \frac{e^{s_i + \vec{\theta}_k \cdot \vec{g}_i}}{e^{s_i + \vec{\theta}_k \cdot \vec{g}_i} + e^{s_j + \vec{\theta}_k \cdot \vec{g}_i}}, \tag{8}$$

where $\vec{x} \cdot \vec{y}$ indicates the scalar product between the vectors $\vec{x}$ and $\vec{y}$. The log-likelihood for the parameters $\mathbf{s}$ and the parameter matrix $\vec{\boldsymbol{\theta}} = (\vec{\theta}_1, \dots, \vec{\theta}_m)$ can be written as:

$$l(\mathbf{s}, \vec{\boldsymbol{\theta}}) = -\sum_{k=1}^{m} \sum_{(i,j) \in Q_k} \log \left(1 + e^{-\left(s_i + \vec{\theta}_k \cdot \vec{g}_i - s_j - \vec{\theta}_k \cdot \vec{g}_i \right)} \right), \tag{9}$$

while the gradients become:

$$\begin{aligned} \frac{dl}{ds_i}(\mathbf{s}, \vec{\boldsymbol{\theta}}) = \sum_{k=1}^{m} \Bigg(& \sum_{i:(i,j) \in Q_k} \left(1 + e^{\left(s_i + \vec{\theta}_k \cdot \vec{g}_i - s_j - \vec{\theta}_k \cdot \vec{g}_i \right)} \right)^{-1} \\ & - \sum_{i:(j,i) \in Q_k} \left(1 + e^{\left(s_i + \vec{\theta}_k \cdot \vec{g}_i - s_j - \vec{\theta}_k \cdot \vec{g}_i \right)} \right)^{-1} \Bigg), \end{aligned} \tag{10}$$

and

$$\frac{dl}{d(\vec{\theta}_k)_t}(\mathbf{s}, \vec{\boldsymbol{\theta}}) = \sum_{(i,j) \in Q_k} \frac{(\vec{g}_i)_t - (\vec{g}_j)_t}{\left(1 + e^{\left(s_i + \vec{\theta}_k \cdot \vec{g}_i - s_j - \vec{\theta}_k \cdot \vec{g}_j \right)} \right)}, \tag{11}$$

where $(\vec{x})_t$ indicates the t-th component of the vector $\vec{x}$. In this way, we are able to estimate the bias of the evaluators with respect to multiple binary attributes.

The case of a single attribute with more than two groups is dealt with in an analogous way. Assume that the attribute of interest induces q groups. Then, we will need a vector $\vec{\theta}_k \in \mathbb{R}^{q-1}$ of $q-1$ parameters to model the bias of the evaluator k. We only need $q-1$ parameters as the effect on the q-th group is implicitly determined as $-\sum_{t=1}^{q-1} (\theta_k)_t$, where $(\theta_k)_t$ indicates the t-th component of θ_k (This extends the particular case of a single binary attribute, where there are two groups and one parameter and θ_k for, e.g., male is equal to $-\theta_k$ for, e.g., female). Furthermore, for an item i, the vector $\vec{g}_i \in \{0,1\}^{q-1}$ will have all components as 0 if i belongs to the q-th group and all components as 0 except the t-th equal to 1 if i belongs to the t-th of the other $q-1$ groups. Note that any of the groups can equivalently be the q-th group.

It is further possible to combine multiple (non binary) attributes. Let's assume to have $\vec{g}_i \in \{0,1\}^q$ and $\vec{g}'_i \in \{0,1\}^{q'}$ (representing two different attributes with $q+1$ and $q'+1$ groups respectively) we can simply stack $\vec{g}_i$ on the top of $\vec{g}'_i$, obtaining a new vector of $q+q'$ components. Analogously, it will be necessary a parameter vector $\theta_k \in \mathbb{R}^{q+q'}$ of length $q+q'$ for each evaluator k.

4.3 Dealing with intersectionality

Based on Crenshaw's theory (Crenshaw 2013), "Intersectionality states that interaction along multiple dimensions of identity produces unique and differing levels of discrimination for various possible sub-groups" (Gohar and Cheng 2023). Intersectionality focuses on the fact that sub-groups, defined as the intersection of multiple attributes (e.g., black women), might be particularly penalised or favored. BARP is able to deal with intersectionality by adding the following components. For each intersectional group that needs to be modeled, a new component, with value 1 if item i belong to that intersectional group and 0 otherwise, needs to be added to each vector $\vec{g}_i$. Furthermore, for each intersectional group and for each evaluators' bias vector $\vec{\theta}_k$, a new bias parameter, that will be estimated by BARP, needs to be added. In general, we recommend incorporating an intersectional component when experts' knowledge indicates potential discrimination issues related to such intersectionality. Furthermore, one can firstly use BARP without intersectional parameters and then compare the distributions of the scores of the items of intersectional groups with the distribution of all items' scores. If there are significative differences for an intersectional group, it suggests that such group has been discriminated/favoured and that explicit evaluators' bias parameters for that intersectional group should be added.

5 Experiments on synthetic data

In this section, we assess the effectiveness of our method using synthetic data. The advantage of using synthetic data is to have the ground truth for the evaluators' biases and the items' latent quality scores. In this way, we can assess how well our method is able to estimate evaluators' bias and if it can effectively reconstruct an unbiased ranking. Furthermore, it allows us to explore how certain factors, such as the total number of comparisons, affect the performance of our method. More in details, our empirical analysis is aimed at answering the following research questions:

- To what extent does evaluators' bias affect ranking from pairwise comparisons methods?
- To what extent can BARP reconstruct the ground truth (unobservable) ranking, thus mending evaluators' bias?
- Can BARP estimate the bias of each evaluator correctly?
- Which conditions affect the performance of our method?

In Sect. 6, we also present an empirical analysis on real-world datasets.

5.1 Synthetic dataset generation

Our experimental setup consists of 100 items to be ranked, each assigned a latent quality score (denoted as $\mathbf{s}$), drawn from a normal distribution $N(0, 5)$. Among these items, 70 belong to group $g_a \in G$, while the remaining 30 belonged to group $g_b \in G$. Additionally, we consider 50 evaluators, each characterized by a group bias parameter sampled from a normal distribution. The bias parameter represented an additive value, either positive or negative, applied uniformly to all elements within one of the two groups. Subsequently, taking into account their respective biases, each evaluator assesses 100 randomly generated pairs. More specifically, consider two items $i, j \in I$ with latent quality score s_i and s_j respectively, that are to be compared by an evaluator $k \in E$ with bias parameter θ_k: following the literature, e.g., Chen et al. (2013), Negahban et al. (2012), Chen and Suh (2015), the pairwise comparison is produced stochastically according the probability in Eq. 4. This process allows us to accumulate a total of 5000 labeled pairwise comparisons.

5.2 Baselines and measures

As baselines for comparison we consider the Bradley-Terry model (Bradley and Terry 1952) (BT), which consists in the maximization of the likelihood of Eq. 2, as a baseline for parametric statistical methods. We also considered CrowdBT (Chen et al. 2013), a variation of the Bradly-Terry, that includes evaluators and estimation of their reliability, and FactorBT (Bugakova et al. 2019) a variation of CrowdBT, that assumes that evaluators, with a certain probability, instead of basing their choices on the scores of the items, can chose an item because of its belonging to a certain group. Furthermore, we include RankCentrality (Negahban et al. 2012) as a representative of graph and spectral methods, that exploit the representation of the items as nodes in a graph and of the pairwise comparisons as edges. With this graph representation, the ranking of the items can be seen as the stationary distribution of a random walk. Finally, we consider the FA*IR post-processing method (Zehlike et al. 2017), that consists in re-ranking an output ranking (in this case the output of BT) to ensure that each group have enough members at the top of the ranking so to satisfy certain statistical constraints.

The main measure we consider is the reconstruction accuracy of the true latent ranking, computed in terms of Kendall's Tau correlation coefficient, adjusted for ties, Tau-b (Kendall 1945), between the true latent scores $\mathbf{s}$ and the scores $\hat{\mathbf{s}}$ estimated by the different methods. The measure ranges between -1 (complete inversion) and 1 (perfect agreement). Any pair of pairs $(s_i, \hat{s}_i)$ and $(s_j, \hat{s}_j)$, where $i < j$, is said to be concordant if either both $s_i > s_j$ and $\hat{s}_i > \hat{s}_j$ holds or both $s_i < s_j$ and $\hat{s}_i < \hat{s}_j$; otherwise, they are said to be discordant. The pair $\{(s_i, \hat{s}_i), (s_j, \hat{s}_j)\}$ is said to be tied if and only if $s_i = s_j$ or $\hat{s}_i = \hat{s}_j$; a tied pair is neither concordant nor discordant. Then, the Kendall's Tau-b coefficient is defined as follows:

$$\tau_b = \frac{n_c - n_d}{\sqrt{(n_c + n_d + t_s)(n_c + n_d + t_{\hat{s}})}}$$

where n_c,n_d are respectively the the number of concordant and discordant pairs, t_s is the number of tied values in $\mathbf{s}$ only and $t_{\hat{s}}$ is the number of tied values in $\hat{\mathbf{s}}$ only.

Further, we measure group specific differences in exposure in a ranking following Singh and Joachims (2018). We define the exposure for a group g as:

$$\text{Exposure}(g) := \frac{1}{|g|} \sum_{i \in g} \frac{1}{\log_2 \left(r_i^* + 1 \right) + 1}.$$

Then, the difference in exposure received on average by the items in two groups g_a and g_b is simply computed as $\text{Exposure}(g_a) - \text{Exposure}(g_b)$. Values, positive or negative, far from 0 are undesirable.

5.3 Results

Figures 2 (Left) and (Right) show that our method is able to correctly reconstruct the unbiased rank, despite the existence of group biases in the evaluators' judgments. In the two plots, we progressively increase the average group bias of the evaluators and its standard deviation, reporting on the horizontal axis the mean (or std) of the normal distribution from which the group bias of each evaluator is sampled. Our results show that the baseline methods become increasingly inaccurate when rising the average evaluators' group bias and also suffer from increases in the standard deviation, while the accuracy of our method is hardly affected.

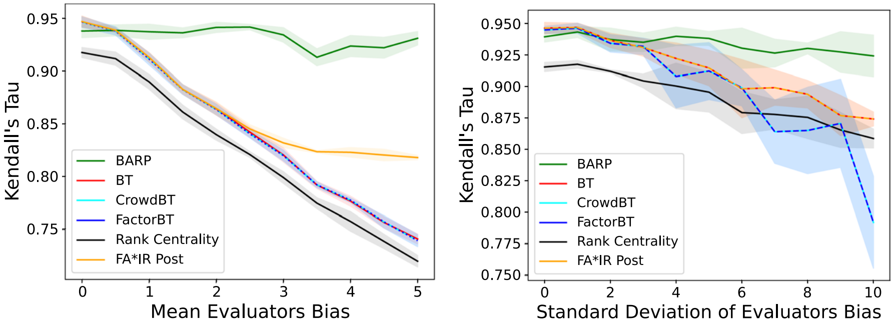

Fig. 2 (Left) Comparison of different methods with increasing evaluators' group bias. As the group bias of the evaluators increases, the Kendall's Tau for all the methods, except BARP, decreases. This indicates that baseline methods, including FactorBT and FA*IR that have access to the group information, are not able to correctly reconstruct the ranking of the items if the evaluators exhibit a systematic group bias. (Right) Comparison of different methods with increasing variance in evaluators' group bias. The mean of the evaluators is 0, hence, some evaluators favor one group and some others favor the other group, while on average the comparisons result equally biased in both directions. However, accounting for the group bias of the single evaluators ensures that the accuracy of our method is not decreasing

Increasing the mean of evaluators bias, BT and Rank Centrality are not able to deal with the fact that items of the penalized group are erroneously being relegated to lower positions in the ranking. CrowdBT and, interestingly, also FactorBT try to find evaluators that sometimes answer randomly, but, since the comparisons are driven by systematic bias, both fails to detect such evaluators' bias and they both behave as the simple BT model. Finally, the FA*IR post-processing, when the group bias in the ranking become significative, by partially addressing this bias it is able to mitigate part of the accuracy reduction.

Instead, when the standard deviation is increased and some evaluators might show extreme group bias, CrowdBT and FactorBT, mistakenly, assume that some evaluators answer adversarially, that is always answers the opposite of what the true scores would suggest. We noticed this behaviour by looking at the estimated parameters of the evaluators, and this explains their worst performance. Furthermore, since on average evaluators are equally biased in both directions, FA*IR does not detect significant violations and does not modify the ranking output of BT.

In the case in which the evaluators' mean bias is different from 0, the reduction in the accuracy of the methods, except BARP, is connected with the fact that the items of the group penalized are erroneously being relegated to lower positions in the ranking. In substance, methods that do not correctly model bias tend to penalize one group if the judgements of the evaluators are biased. We show this in Fig. 3 (Left) by computing the differences in the exposure received on average by the items in different groups. Our results indicate that with increasing average group bias of evaluators most methods show increasing differences in the exposure of different groups in a ranking. Unlike most methods, BARP based rankings are not affected by increasing evaluator bias and lead to relatively equal group exposure.

Besides being able to correctly rank items even in the presence of evaluators' group bias, our method further allows to identify the amount of group bias of each single evaluator. We investigate this in Fig. 3 (Right). In particular, we create

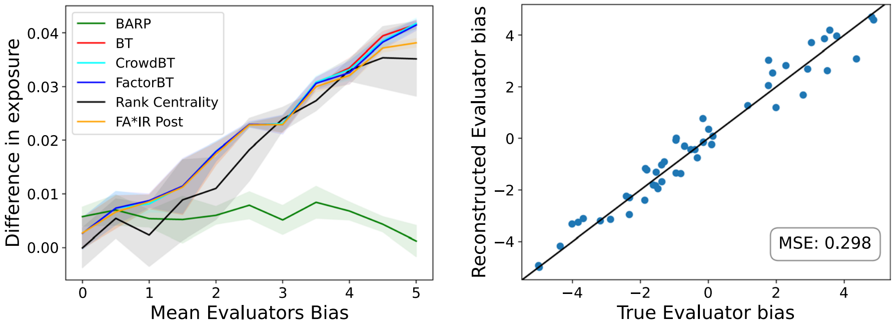

Fig. 3 (Left) Difference in the exposure of the groups. Mending the bias of the evaluators BARP avoids that the items belonging to different groups receive different exposure in the ranking. (Right) Reconstruction of the individual bias of each evaluator. Each point represents an evaluator. The horizontal axis depicts the simulated ground-truth bias of each evaluator and the vertical axis represents the BARP based estimation of the individual biases. For each evaluator, we can individuate their bias quite accurately

comparisons where the single evaluators have different biases uniformly sampled in $[-5, 5]$. A negative bias means that the evaluator favors group g_a and a positive one means that the evaluator instead has a bias in favor of the items belonging to group g_b. Figure 3 (Right) reports a scatter-plot comparing the ground-truth evaluators' bias with the estimated one. Our results show that BARP allows to accurately reconstruct this bias starting from the pairwise comparisons of items only.

Next, we investigate multiple factors that might affect the performance of our method. Firstly, we are interested in how the total number of pairwise comparisons available, the amount of pairwise comparisons evaluated by each evaluator, and the number of evaluators affect the accuracy of our ranking recovery method. We analyze this in Fig. 4 (Left). We notice two main effects. The first is that when there are more total pairs, the method is able to better reconstruct the ground truth ranking (i.e. Kendall's Tau closer to 1). This is as expected, as having more pairwise comparisons implies having more information, allowing for better estimations. The second effect refers to the number of evaluators that produce a fixed total number of pairwise comparisons. Moving along the anti-diagonal in the heatmap, one can see that it is slightly better to have less evaluators evaluating more pairs rather than having more evaluators evaluating less pairs. This is because, with more information per evaluator, we are better able to estimate the group bias of each evaluator, and hence to better reconstruct the true ranking. However, this effect is less pronounced compared to the effect of the total amount of pairwise comparisons which is the most important factor that effects the accuracy of our method.

Another factor that might affect the results is the dispersion of the ground truth scores. If all items are similar to each other with respect to their ground truth scores, reconstructing the ranking is more difficult, even in the absence of bias. We study this factor in Fig. 4 (Right) by reporting the Kendall's Tau score achieved by our method while varying the relative distance between the scores. For example, when the scores are normally distributed with a small variance it is

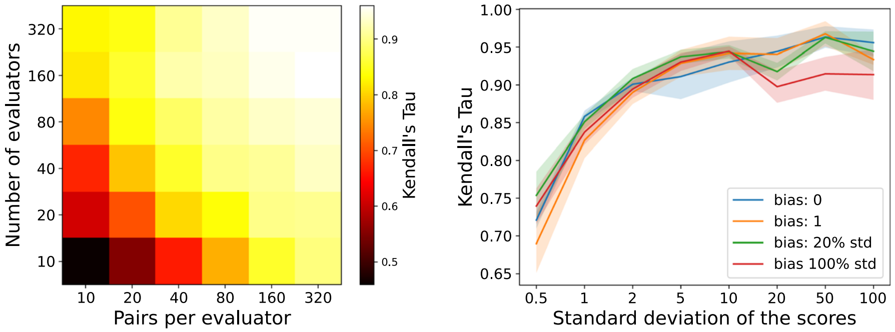

Fig. 4 (Left) Number of pairwise comparisons. The heatmap contains the Kendall's Tau ranking reconstruction scores (1 is best), varying the number of evaluators and the number of pairs annotated by each evaluator. Increasing the total number of pairs, i.e. moving up or to the right in the heatmap, the ranking becomes more accurate. (Right) The effect of the values of the scores. The higher the distance between the scores, the easier it is to reconstruct the original ranking

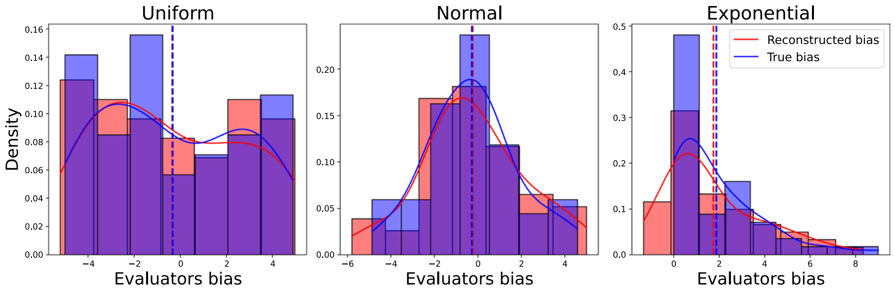

Fig. 5 Reconstruction of evaluators' bias under different distributions of the bias: uniform (Left), normal (Center), and exponential (Right)

harder to correctly recover the original ranking for the above-mentioned reasons (as the comparisons depend on the relative differences shifting the mean of the scores doesn't have any effect).

Furthermore, one could argue that the distribution of the group bias of the evaluators could depend on different situations and scenarios. Figure 5 illustrates the capability of BARP to reconstruct the bias under different distributions: uniform (Left), normal (Center), and exponential (Right). Our results show that independently of the shape of the group bias distribution that we use to sample the evaluator bias, BARP accurately estimates the bias of evaluators.

Finally, in Fig. 6, we examined evaluators with a bias against groups having attribute values of 1 for two binary attributes. This bias specifically penalizes items with both attribute values set to 1 (depicted as red items in the figure). On the left, we can see how the ranking obtained with Bradly-Terry (similarly happens for the other methods, exept BARP), ranks systematically lower items of the disadvataged groups. On the other hand, we illustrate how BARP can reconstruct the ranking of

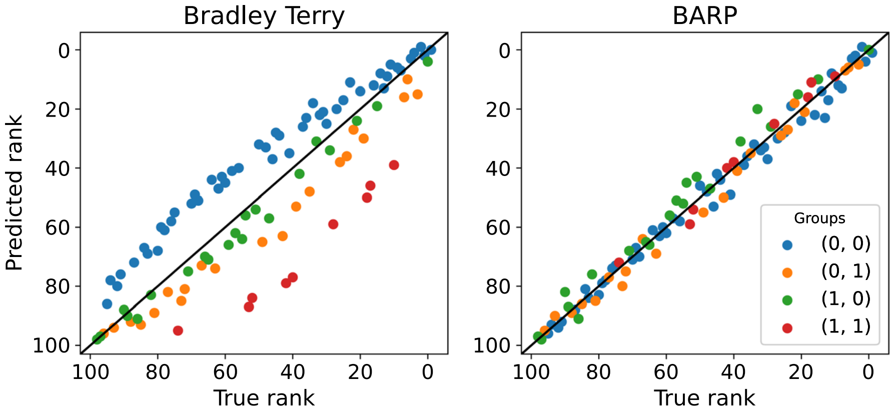

Fig. 6 Ranking with multiple groups. (Left) The ranking reconstruct by BT (as an example) penalizes certain groups over others due to its inability to account for evaluators group bias. (Right) BARP avoids inequalities between the groups and better reconstructs the true rank of the items

the items accurately and contrast the disadvantage of certain group of items, even in the presence of combined multiple groups effects.

6 Experiments on real-world data

In this section we report applications of our method to two real-world datasets. In these cases, contrarily to the case of synthetic data discussed in the previous section, we do not have access to the evaluators' ground truth bias.

6.1 People's age evaluation dataset

For the empirical analysis, we consider the IMDB-WIKI-SbS (Pavlichenko and Ustalov 2021), a recently released large-scale dataset for evaluating pairwise comparisons. The dataset contains 9150 images appearing in 250,249 pairs annotated via crowdsourcing. The dataset uses a subset of images and information about age and gender from the IMBD-WIKI dataset (Rothe et al. 2018), with a balanced distributions of age and gender. In the pairwise comparison task crowdworker were asked to judge which of the two presented people is older. The pairs of face-shot images were presented to 4,091 workers, with each pair annotated by only one unique worker. In this way, on average, each worker annotated around 60 pairs (excluding additional control tasks).

In order to investigate the gender group bias of the evaluators, we apply BARP to the set of pairwise preferences. Figure 7 reports the distribution of the gender group bias of the annotators according to the estimation of our model. Choosing as convention the group g to be the male group, the interpretation of the bias of the evaluators is the following: an evaluator k with a bias of $\theta_k > 0$, perceives males as if they were θ_k years older than what they are, when they are compared to females. Instead, a bias of $\theta_k < 0$ implies that the males are perceived as θ_k years younger than what

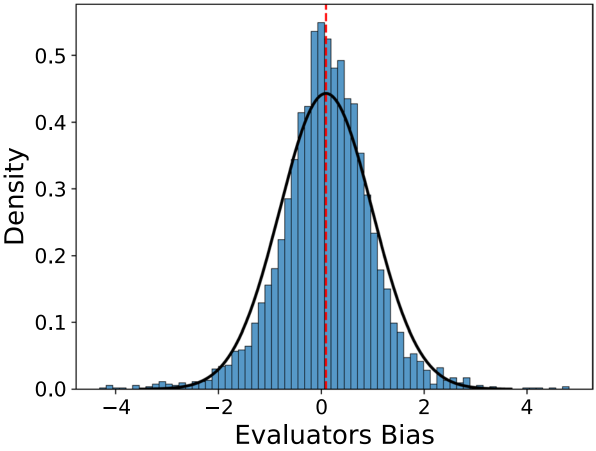

Fig. 7 IMDB-WIKI-SbS dataset: Distribution of the group bias of the evaluators. Positive values indicate a tendency to perceive males as older, and negative values indicate a tendency to perceive females as older, when compared with the opposite gender

they are, when compared to the opposite gender. The distribution of the bias of the evaluators resembles the distribution of a Gaussian with a mean close to 0. This means that, averaging on all evaluators, we would not see a clear effect against a group, but single evaluators might still exhibit a gender bias in their age evaluations.

We further investigate, in Fig. 8, the estimated gender group bias of the individual evaluators by relating them to the amount and types of errors that evaluators are committing in the evaluation of gender-mixed and single-gender pairs. Note that the ground-truth scores for age are only used to calculate the errors while the BARP

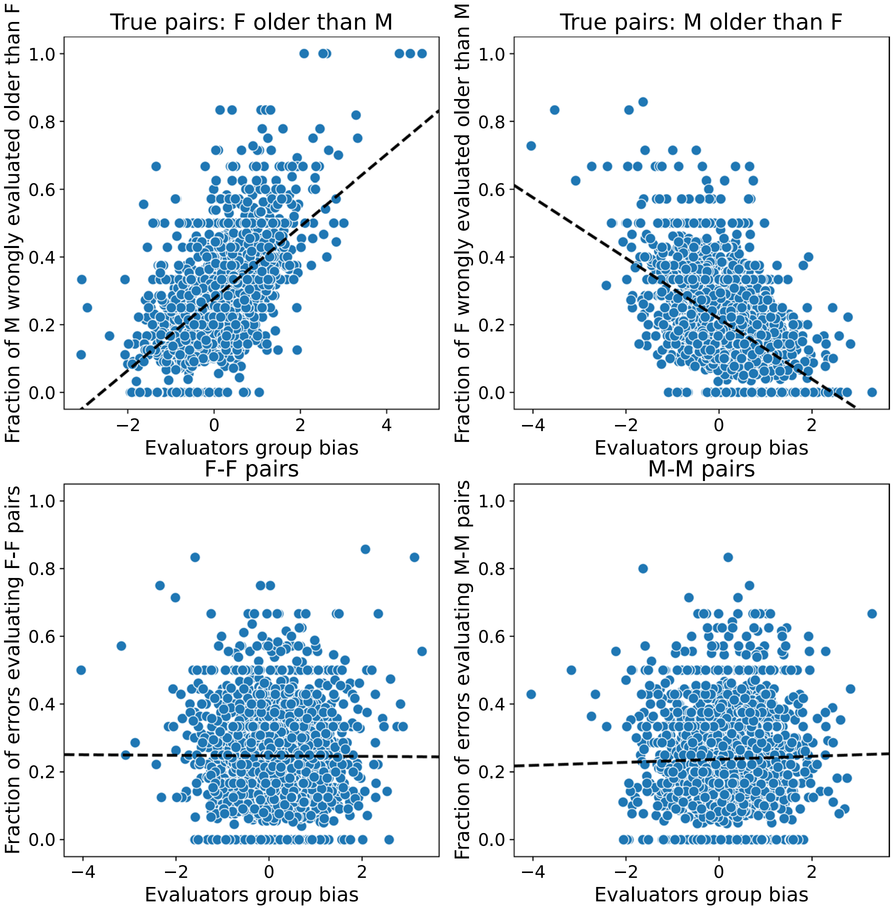

Fig. 8 Relationship between evaluators' gender group bias and errors in the evaluation of pairwise comparisons. The plots in the first row depict pairs in which a male (M) face is compared with a female (F) face. The horizontal axis depicts the BARP estimate of the evaluator bias, while the vertical axis depicts the error rate. In the left plot only the pairs in which the woman is older than the man are depicted, and vice versa on the right. One can see that the BARP based estimate for the gender bias of the evaluator is correlated with the fraction of errors that evaluators conduct. The plots on the second row represent within group comparisons (i.e. male-male or female–female). While evaluators also commit errors here, the error seems to be unrelated from the group bias of evaluators

estimates do not use this information. In the plots of the first row, there are the pairs in which a male (M) is compared with a female (F). On the left plot, the pairs in which the female is older than the man, and vice versa on the right. The plots on the second row represent within groups comparisons (i.e. male-male or female–female). On the horizontal axis is displayed the bias of each evaluator as computed with our method, while on the vertical axis is represented the error rate that an evaluator has when evaluating pairs of specific groups. The results are as expected. In the first plot, a high positive bias (i.e. perceiving man as older) is associated with more often evaluating a man as older than a woman, when this is not true. The opposite happens in the second plot, evaluators with negative bias, tend to commit more often the error of wrongly evaluating a woman as older than a man. In the bottom row, instead, we can see how the evaluators' group bias doesn't affect error when evaluating pairs of the same gender.

6.2 Law schools scores

As an additional example of application of BARP, we consider the US data from the Law School Admission Council survey (Wightman 1998; Alvarez and Ruggieri 2023). The dataset contains information about students' law school admissions test scores (LSAT), undergraduate grade-point average (UGPA), and z-scores of the first-year average grades (ZFYA). It also contains demographic information about the students. Following Alvarez and Ruggieri (2023), we consider as attributes a student's binary gender (male/female) and a binary race variable (white/non-white). The dataset comes in the form of tabular data and, in order to apply ranking from pairwise comparison methods, we randomly select $n = 1000$ students and simulate pairwise comparisons considering their scores and grades in the following way. For each of the three LSAT, UGPA, and ZFYA, we sample 10,000 pairs, creating a pairwise preference relationship based on which student has a higher score/grade.

While ranking students based on their scores and grades might be considered as an objective criterion, in reality, racial disparities in access to education may result in systematic variations in scores (Reeves and Halikias 2017). We aim to investigate whether our proposed method can detect this bias and control for it. First, we use BARP to estimate the bias of the different scores/grades, considering in one case the gender and in the other the race. The results are reported in table 1.

We can notice that, for gender, the bias of the columns used to evaluate are nearly symmetric and close to zero, while for race the biases are higher and of the same sign. When we are not expecting bias, i.e. the case of gender, algorithms non group-bias-aware reconstruct scores and rank similarly as BARP. This is shown in the first

Table 1 Estimated group bias with the BARP model

	LSAT	UGPA	ZFYA
Gender	0.334	– 0.358	0.074
Race	– 1.532	– 0.728	– 1.786

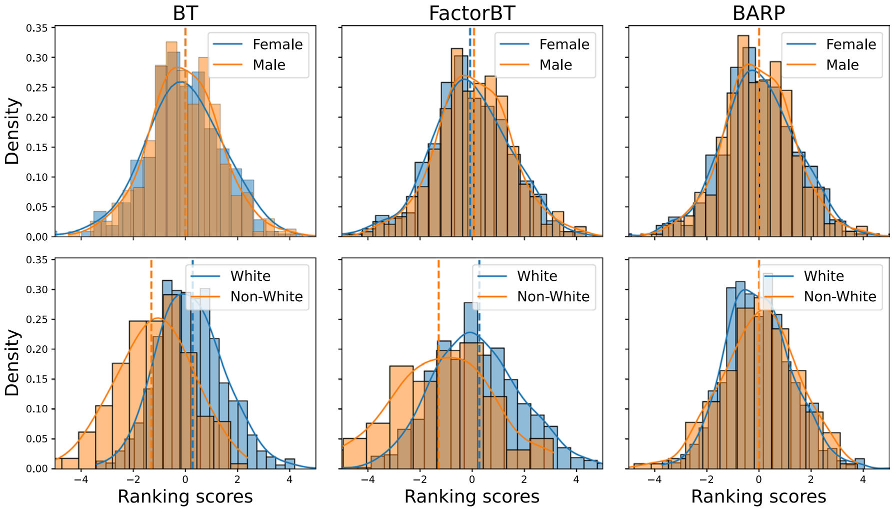

Fig. 9 Distribution of the ranking scores. Even in the presence of group bias, the case of the attribute race, BARP tackles the differences in the distributions of the ranking scores across the different groups

row of Fig. 9. Instead, in the presence of group biases, i.e. the case of race, in the second row, we expect an algorithm that doesn't account for group bias to present inequalities in the distribution with respect to race. The second row on the left of Fig. 9 shows how Bradely-Terry, as an example of non group bias aware ranker, ranks a group, in this case white students, systematically higher than the other group, non-white students. Furthermore, also FactorBT is not able to correctly address the discrepancy in the scores in the presence of group biases. The reason can be traced in the fact that FactorBT does not models how the perception of the scores is affected by evaluators' bias. BARP, on the other hand, on the second row on the right of Fig. 9, by accounting for the group bias relative to the race, ranks the two groups more similarly.

7 Conclusions, limitations and future work

In this paper, we study the problem of correctly ranking items from pairwise comparisons in presence of evaluators' bias against some groups of items. We propose an algorithm that estimates the group bias of the individual evaluators and accounts for this bias while recovering the unknown ranking of the items. Our experiments confirm that our method is very effective at detecting and fixing the biases intrinsic in the pairwise comparisons, thus producing rankings which are closer to the latent one than methods that do not consider the possibility of biased evaluators. Furthermore, our experiments show how mending the bias of the evaluators avoids items from different groups receiving different exposure in the ranking.

We next discuss some limitations of this work which hint potential pathways for future research. First, it is often difficult to distinguish if the bias comes from the evaluators or if it is the problem of different skill sets between different groups. Here, we assume that there should not be systematic differences in the true skills/ scores for different groups. In reality, due to certain societal discrimination, the true scores may be related to the group identity. In such cases, our method can serve as a corrective method to reduce those systematic group-level biases. Second, we assume that the bias of each evaluator can be modeled with a constant parameter for each group, but there might be instances in which the bias might depend also on other factors, such as, for example, on the scores of the items themselves. We leave this as a future work. Furthermore, similar to all ranking methods from pairwise comparisons, BARP requires a minimum amount of pairwise comparison that scales with $n \log n$. Therefore the accuracy of the method decreases with the increasing the number of objects if this is not followed by an adequate increase in the number of comparisons.

Despite these limitations, we believe that our model can be useful in detecting and addressing biases against groups of people in ranking from pairwise comparisons and we hope that our work can motivate further research in this topic.

Appendix: Additional experiments

Number of pairs evaluated by each evaluator

In the analysis with simulated data, we assumed that each evaluator assesses the same number of comparisons. However, in real-word scenarios, as for example happens in the IMDB-WIKI-SbS dataset, the number of pairs evaluated is different across evaluators. In Fig. 10, we show that this assumption does not significatively affects the recovery of the true ranking of the items. We consider four different distributions for the number of pairs evaluated by the evaluators. We keep fixed the number of evaluators at 50 and the total number of comparisons at 5000. As sake of

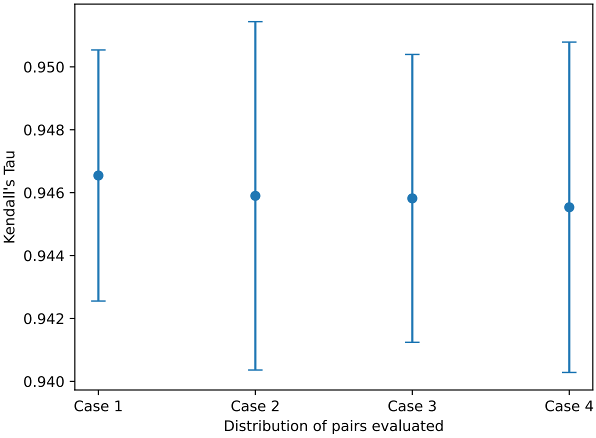

Fig. 10 The distribution of pairs evaluated by the evaluators does not significantly affect the recovery of the true ranking of the items. On the horizontal axis are displayed four different distributions, while on the vertical axis, are reported the mean and standard deviation, across ten repetitions, of the Kendall's Tau correlation between the reconstructed ranking and the true ranking

example, we consider four different alternatives. In the first case, all evaluators evaluates 100 comparisons each, in the second case we assume that the number of pairs evaluated by each evaluator is Poisson distributed with expected rate 100 (hence the total number of pairs evaluated is 5000 on average), in the third case, each of 10, 3 0, 50, 70, 90, 110, 130, 150, 170 and 190 pairs are evaluated by 5 evaluators, and lastly we consider an extreme case in which half evaluators evaluate only 5 pairs each while the other half evaluate 195 pairs each. In Fig. 10 we can see that despite considering different distributions for the pairs evaluated by the evaluators the Kendall's Tau between the reconstructed ranking and the true ranking does not differs significatively. One reason lies in the fact that even if the bias of a single evaluator could be less precise if they estimated only few pairs, it is also true that the impact on the reconstructed ranking of that evaluator is low since they only evaluated few pairs.

Approximation and stopping criteria

On the vertical axis of Fig. 11 we report the approximation obtained in terms of the Kendall's Tau, while on the horizontal axis we report the number of iterations of BARP. Each line correspond to a typical run with the tolerance for the gradient specified in the legend. Each run is performed on a different set of observed pairwise comparisons. Since multiple runs, even with different initial parameter initializations, on the same set of pairwise comparisons are almost equivalent, this allows us to compare how the approximation due to the choice of a stopping criteria compare with the overall noise in the observations. We can notice some facts. The first is that after the first around hundred of iterations, the Kendall's Tau is almost stable, meaning that at the ranking level the order of only very few pairs is swapped. The method does not rely very much on the choice of the values of stopping criteria on the gradient since even setting the values as big as 10^{-1} leads to meaningful results. Lastly, the noise of stopping at a certain point (as long as we passed the initial

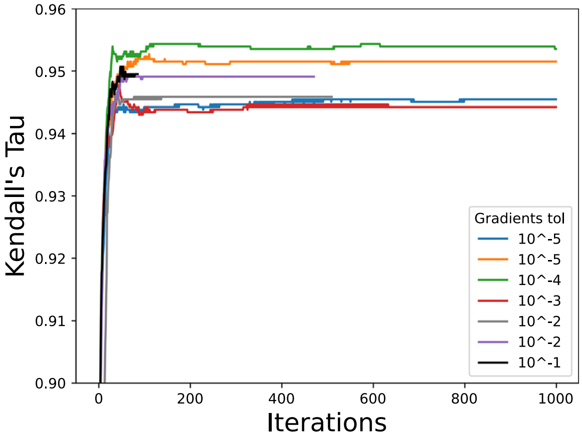

Fig. 11 Approximation of BARP in reconstructing the original ranking in terms of the Kendall's Tau correlation with respect to the number of iterations and the stopping criteria on the gradients

around hundred iterations) is small compared to the overall noise that the actually observed comparisons induce.

Acknowledgements This work has received funding from the European Union's Horizon 2020 research and innovation programme under Marie Sklodowska-Curie Actions (grant agreement number 860630) for the project "NoBIAS - Artificial Intelligence without Bias". This work reflects only the authors' views and the European Research Executive Agency (REA) is not responsible for any use that may be made of the information it contains.

Funding Open access funding provided by the project "NoBIAS - Artificial Intelligence without Bias".

Availability of data and materials The code and datasets to reproduce our experiments can be found at: https://github.com/Ambress92/Bias-Aware-Ranker-from-Pairwise-comparisons.

Declarations

Ethics approval and consent to participate This paper focuses on bias and fairness issues in ranking from pairwise comparisons. Although the goal is to mitigate such issues, there are several ethical aspects to be considered. First, in order to tackle bias issues we need access to the sensitive attributes (for example gender or race) for which we want to mitigate biases. Furthermore, ranking from pairwise comparison methods, and hence BARP as such, can be used in decision making contexts, such as hiring or university admissions, that can affect human lives. Moreover, we considered biases against groups of people, however discriminatory behaviours against single individuals might still exist and special attention is required in such cases.

References

Almaatouq A, Krafft P, Dunham Y, Rand DG, Pentland A (2020) Turkers of the world unite: multilevel in-group bias among crowdworkers on amazon mechanical Turk. Soc Psychol Personal Scince 11(2):151–159

Alvarez JM, Ruggieri S (2023) Counterfactual situation testing: Uncovering discrimination under fairness given the difference. Preprint arXiv:2302.11944

Beaver RJ, Gokhale D (1975) A model to incorporat within-pair order effects in paired comparisons. Commun Stat Theory Methods 4(10):923–939

Bradley RA, Terry ME (1952) Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika 39(3/4):324–345

Bugakova N, Fedorova V, Gusev G, Drutsa A (2019) Aggregation of pairwise comparisons with reduction of biases. Preprint arXiv:1906.03711

Carlson D, Montgomery JM (2017) A pairwise comparison framework for fast, flexible, and reliable human coding of political texts. Am Polit Sci Rev 111(4):835–843

Celis LE, Straszak D, Vishnoi NK (2017) Ranking with fairness constraints. Preprint arXiv:1704.06840

Chen X, Bennett PN, Collins-Thompson K, Horvitz E (2013) Pairwise ranking aggregation in a crowdsourced setting. In: Proceedings of the sixth ACM international conference on web search and data mining, pp 193–202

Chen Y, Suh C (2015) Spectral MLE: Top-K rank aggregation from pairwise comparisons. In: International conference on machine learning, pp 371–380

Crenshaw K (2013) Demarginalizing the intersection of race and sex: a black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. In: Feminist legal theories. Routledge, pp 23–51

David HA (1987) Ranking from unbalanced paired-comparison data. Biometrika 74(2):432–436

Davidson RR, Beaver RJ (1977) On extending the Bradley-Terry model to incorporate within-pair order effects. Biometrics 693–702

Fogel F, d'Aspremont A, Vojnovic M (2014) Serialrank: spectral ranking using seriation. Adv Neural Inf Process Syst 27

García-Soriano D, Bonchi F (2021) Maxmin-fair ranking: individual fairness under group-fairness constraints. In: KDD '21: The 27th ACM SIGKDD conference on knowledge discovery and data mining, pp 436–446

Geva M, Goldberg Y, Berant J (2019) Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In: Inui K, Jiang J, Ng V, Wan X (eds) Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing, EMNLP-IJCNLP, pp 1161–1166

Gohar U, Cheng L (2023) A survey on intersectional fairness in machine learning: Notions, mitigation, and challenges. Preprint arXiv:2305.06969

He Y, Gan Q, Wipf D, Reinert GD, Yan J, Cucuringu M (2022) GNNRank: learning global rankings from pairwise comparisons via directed graph neural networks. In: International conference on machine learning, pp 8581–8612

Hube C, Fetahu B, Gadiraju U (2019) Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In: Proceedings of the 2019 CHI conference on human factors in computing systems

Kamar E, Kapoor A, Horvitz E (2015) Identifying and accounting for task-dependent bias in crowdsourcing. In: Proceedings of the AAAI conference on human computation and crowdsourcing, vol 3, no 1, pp 92–101

Kendall MG (1938) A new measure of rank correlation. Biometrika 30(1/2):81–93

Kendall MG (1945) The treatment of ties in ranking problems. Biometrika 33(3):239–251

Kendall MG, Smith BB (1940) On the method of paired comparisons. Biometrika 31(3/4):324–345

Koshkalda I, Kniaz O, Ryasnyanska A, Velieva V (2020) Motivation mechanism for stimulating the labor potential. Res World Econ 11(4):53–61

Kotturi Y, Kahng A, Procaccia A, Kulkarni C (2020) Hirepeer: impartial peer-assessed hiring at scale in expert crowdsourcing markets. In: Proceedings of the AAAI conference on artificial intelligence, vol 34, pp 2577–2584

Kuo T-S, Hankin M, Miranda L, Ying A, Wang C (2020) Assessing political bias using crowdsourced pairwise comparisons. In: Proceedings of the AAAI conference on human computation and crowdsourcing

Kuttal SK, Chen X, Wang Z, Balali S, Sarma A (2021) Visual resume: exploring developers's online contributions for hiring. Inf Softw Technol 138:106633

Liu H, Thekinen J, Mollaoglu S, Tang D, Yang J, Cheng Y, Liu H, Tang J (2022) Toward annotator group bias in crowdsourcing. In: Proceedings of the 60th annual meeting of the association for computational linguistics

Negahban S, Oh S, Shah D (2012) Iterative ranking from pair-wise comparisons. Adv Neural Inf Process Syst 25

Nocedal J, Wright SJ (1999) Numerical optimization

Pavlichenko N, Ustalov D (2021) IMDB-WIKI-SbS: an evaluation dataset for crowdsourced pairwise comparisons. Preprint arXiv:2110.14990

Reeves RV, Halikias D (2017) Race gaps in sat scores highlight inequality and hinder upward mobility. Brookings Institute, Washington

Rothe R, Timofte R, Gool LV (2018) Deep expectation of real and apparent age from a single image without facial landmarks. Int J Comput Vis 126(2–4):144–157

Sap M, Card D, Gabriel S, Choi Y, Smith NA (2019) The risk of racial bias in hate speech detection. In: Proceedings of the 57th annual meeting of the association for computational linguistics, pp 1668–1678

Sap M, Swayamdipta S, Vianna L, Zhou X, Choi Y, Smith NA (2021) Annotators with attitudes: how annotator beliefs and identities bias toxic language detection. Preprint arXiv:2111.07997

Sarma A, Chen X, Kuttal S, Dabbish L, Wang Z (2016) Hiring in the global stage: profiles of online contributions. In: 2016 IEEE 11th international conference on global software engineering (ICGSE), pp 1–10

Singh A, Joachims T (2018) Fairness of exposure in rankings. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, pp 2219–2228

Thurstone LL (1927) The method of paired comparisons for social values. J Abnorm Soc Psychol 21(4):384

Wightman LF (1998) LSAC national longitudinal bar passage study. LSAC research report series

Wright SJ (2015) Coordinate descent algorithms. Math Program 151(1):3–34

Zehlike M, Bonchi F, Castillo C, Hajian S, Megahed M, Baeza-Yates R (2017) FA*IR: a fair top-k ranking algorithm. In: Proceedings of the 2017 ACM on conference on information and knowledge management, pp 1569–1578

Zehlike M, Castillo C (2020) Reducing disparate exposure in ranking: a learning to rank approach. In: Proceedings of the web conference, pp 2849–2855

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Antonio Ferrara[1,2,3] · **Francesco Bonchi**[1,4] · **Francesco Fabbri**[5] · **Fariba Karimi**[3,6] · **Claudia Wagner**[2,7]

✉ Antonio Ferrara
antonio.ferrara@centai.eu

[1] CENTAI, Turin, Italy

[2] GESIS, Cologne, Germany

[3] TU Graz, Graz, Austria

[4] EURECAT, Barcelona, Spain

[5] Spotify, Barcelona, Spain

[6] Complexity Science Hub, Vienna, Austria

[7] RWTH Aachen University, Aachen, Germany