LinkedIn

An Outcome Test of Discrimination for Ranked List

Jonathan Roth, Assistant Professor, Brown University

Guillaume Saint-Jacques, Sr Data Science Manager, Apple

YinYin Yu, Applied Research Manager – Experimentation & Causal Inference, LinkedIn

Introduction

- Large literature in economics and computer science focused on detecting bias on the part of a decision maker (humans, algorithms)
- Canonical setting involves a binary decision: grant pre-trial release, give a loan, admit to college
- In many practical settings, the decision maker produces a ranked list
    - National Residency Match Program (Roth, 1984) or lab experiments (Castillo and Petrie, 2010)
    - Online platforms: Google and LinkedIn produce search results, FB, Twitter, LinkedIn produce Feeds
- Detecting and quantifying bias for different types of algorithms:
    - Pointwise classification algorithms
    - Pointwise ranking algorithms
    - Listwise ranking algorithms

Yu and Saint-Jacques (2022)

This paper: **how to test for bias in ranked lists?**

Overview of the paper

- Develop tests of bias in ranked lists, in the spirit of Becker (1957)
- Show that a sharp testable implication of unbiased behavior is a set of moment inequalities
- Intuitively, these inequalities say that regardless of protected categories, you shouldn't be able to flip the positions of adjacent candidates and systematically improve the objective
- Illustrates how these inequalities can be tested in data generated by listwise ranking algorithms
- Validation exercise using data from LinkedIn

Model

- There is a **Ranker** and an **Auditor**
- For each query q, Ranker produces a ranking of J candidates, $j_q(1), \dots, j_q(J),$ with
  - characteristics $X_{1q}, \dots, X_{Jq}$ and
  - group status $G_{1q}, \dots, G_{Jq}$ (e.g., gender, race)
- The Auditor observes the ranked list of candidates and their
  - realized outcomes (labels) : $Y_{j_q(1)q}, \dots., Y_{j_q(J)q}$ and
  - group status: $G_{1q}, \dots, G_{Jq}$
- This captures the fact that the Auditor may not observe all the information used by the Ranker. Results easily extends to the case where X is observed

Definition of unbiasedness

- Intuition: the Ranker is unbiased if it ranks candidates to maximizes its given objective, regardless of group status → ***equal opportunity for equally qualified candidates*** (Yu and Saint-Jacques, 2022)
- Auditor wants to test if the Ranker is unbiased in the sense that it maximizes

Information set of the ranker

$$E[\sum_r w_r Y_{j_q(r)q} | I_q]$$

Decreasing sequence of weights

Realized outcome (label) at rank r by candidates $j_q(r)$ in query q

- Example: Net Discounted Cumulative Gain (NDCG), a common objective for ranking algorithms, can be put in this form

Main results

- A sharp testable implication of unbiasedness is that

$$E\left[Y_{j_q(r)q} - Y_{j_q(r+1)q} \middle| G_q = g\right] \geq 0 \;\; for\ all\ r$$

- Intuition: we should not be able to improve the objective (on average) by swapping candidates in positions r and $r+1$ whenever we see group orientation G_q
    - E.g., flip candidates ranked 1 and 2 in all queries where 1 is M and 2 is F
- Show that statistical tests can be implemented using results from moment inequalities literature (Canay and Shaikh, 2017)
- Discuss several important implementation details:
    - adjust for position bias ($Y_{j_q(r)q}$ is affected by rank)
    - dealing with high dimensional moments
    - controlling for observed covariates X

Empirical implementation on LinkedIn data

- We apply our methodology to data from **InstaJobs** at LinkedIn
- **InstaJobs:** pointwise classification algorithm that scores candidates' propensity to apply for a job and receive recruiter interaction, s_{j_q}
  - candidate receives a notification for a job if $s_{j_q} \geq \bar{S}$
- Reframe as pointwise ranking algorithm:
  - query = job,
  - candidates are ranked in decreasing order of their scores (up till the cutoff),
  - outcome, Y_{j_q}, is observed for all candidates who received a notification
- Validate listwise outcome test results against that from pointwise outcome test (Yu and Saint-Jacques, 2022)

Moment estimates for adjacent ranks

Figure 1: Moments Comparing Men and Women in Adjacent Ranks

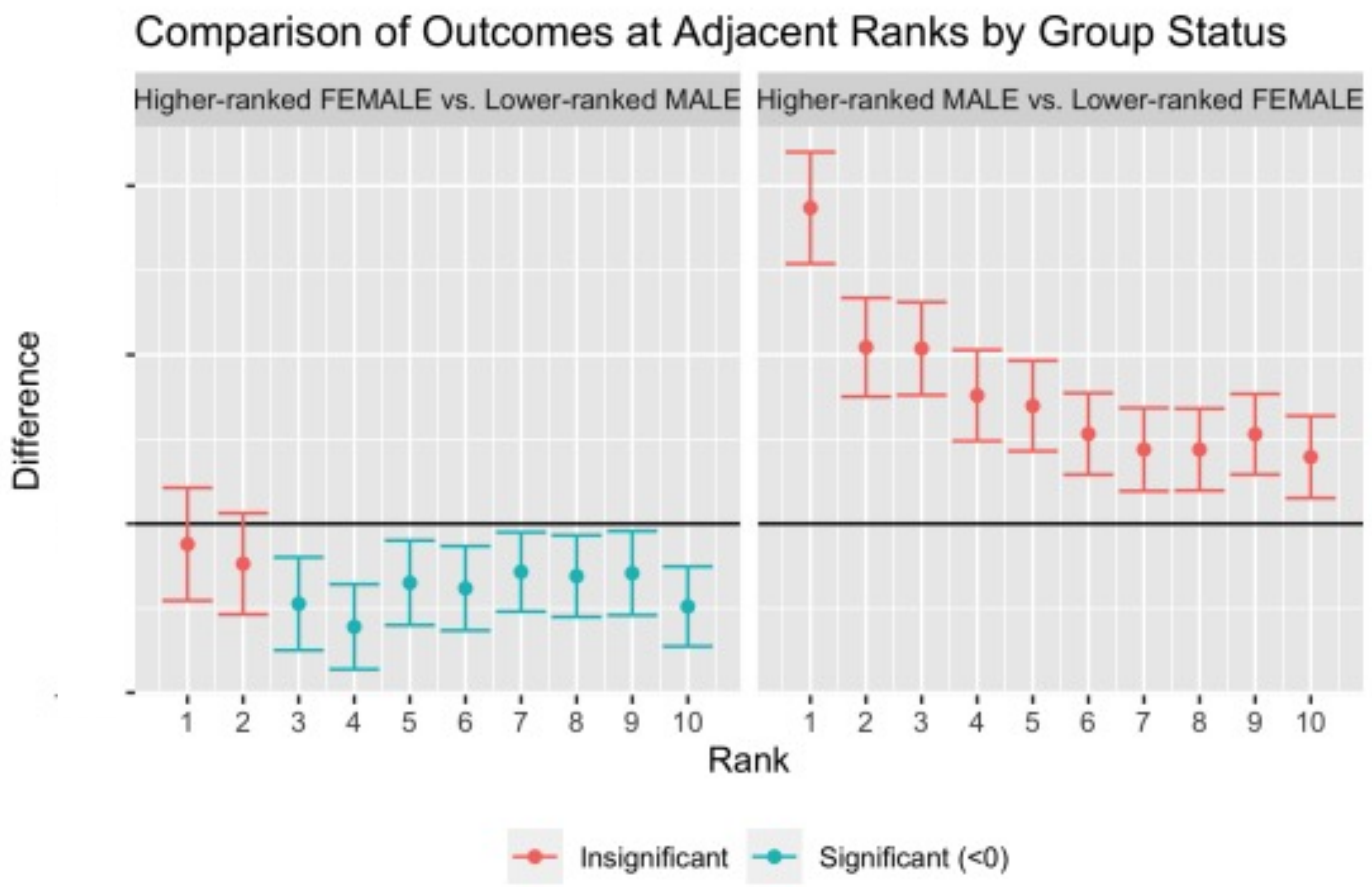

Figure 2: Pointwise Comparison of Outcome by Score

Evidence of bias: lower-ranked M systematically realize better outcomes than higher ranked F – consistent with pointwise implementation of the outcome test, where the scores do not appear to be calibrated by gender

Joint hypothesis test

H_0: the algorithm is unbiased in ranking pairs of candidates in adjacent ranks given their genders

Higher Rank	Lower Rank	p-val
All	All	0.02803
FEMALE	FEMALE	1.00000
MALE	FEMALE	1.00000
FEMALE	MALE	0.00042
MALE	MALE	1.00000

Table 1: p-values for Joint Hypotheses

Reject joint H_0 that the algorithm is unbiased in ranking candidates between adjacent ranks

Algorithm systematically ranked M below F despite the former having better outcomes

Takeaways

- This paper proposes a test for discrimination for ranked lists, extending the framework of Becker (1957) for detecting taste-based discrimination in binary decisions
- This paper is a companion to Yu and Saint-Jacques (2022), which discusses how to detect and quantify algorithmic bias in pointwise classification and pointwise ranking algorithms
- Our approach is complementary to other approaches based on different definitions of bias, such as **demographic parity** (Zehlike et al., 2017); **disparity treatment/impact** (Singh and Joachims, 2018); or **equal opportunity** (Hardt et al, 2016)
- LinkedIn is actively measuring and developing mitigation strategies for identified bias

Jonathan Roth, Guillaume Saint-Jacques, and YinYin Yu. 2022. An Outcome Test of Discrimination for Ranked Lists. In 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). Association for Computing Machinery, New York, NY, USA, 350–356. https://doi.org/10.1145/3531146.3533102

Yu, YinYin, and Guillaume Saint-Jacques. "Choosing an algorithmic fairness metric for an online marketplace: Detecting and quantifying algorithmic bias on LinkedIn." *arXiv preprint* *arXiv:2202.07300* (2022).

Thank you

LinkedIn