Distinguishing Biases from Personal Preferences: An Honest Machine Learning Approach

Mahyar Habibi
Lyft

Zahra Khanalizadeh
University of Washington

Negar Ziaeian
University of Warwick

1 Introduction

This study develops a novel method for detecting micro-level biases in settings characterized by repeated bilateral interactions—such as firms reviewing applicants, judges evaluating defendants, or critics rating films—where personal preferences and correlated characteristics can obscure discriminatory behavior. Traditional empirical strategies often conflate taste-based variation with true bias or rely on assumptions that limit identification to average effects. In contrast, our approach uncovers individual-level discrimination by separating evaluators' discriminatory biases from preferences over true underlying skills or fit that may correlate with the trait. This framework is well-suited for high-dimensional, sparse outcome data where traditional covariate adjustment is infeasible.

2 Conceptual Framework and Methodology

The conceptual framework of this study is based on the interaction between two distinct sets: I, representing items/individuals (e.g., job seekers), and J, comprising reviewers/judges (e.g., employers). Each individual $i \in I$ is subject to evaluation by multiple reviewers $j \in J$, as determined by a matching process $M(I, J) \to \{0, 1\}$. Each applicant is associated with a trait T (e.g., gender) that could be subject to bias from employers in the evaluation process.

As illustrated in Figure 1, there are two routes through which trait T can influence the outcome Y: directly through employers' biases, or indirectly through the *legitimate* path capturing employers' preferences over applicants' characteristics. The goal is to separate trait-related biases from personal preferences over characteristics correlated with the trait.

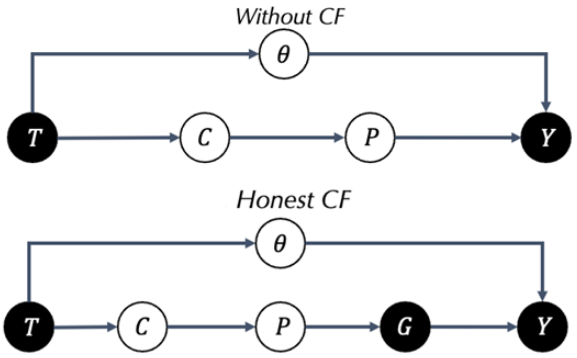

Figure 1: Conceptual Framework. Black nodes are observable; circles are unobservable for the researcher.

Algorithm 1 summarizes the proposed methodology for obtaining micro-level estimates of discrimination. The procedure consists of two main stages.

Algorithm 1 Estimating Micro-Level Coefficients of Discrimination

1. Factorize the matrix of outcomes R into P^0 and C^0.
2. Split the set of reviewers J into K mutually exclusive subsets $J = \{J_1, ..., J_K\}$.
3. *for* J_k in $\{J_1, ..., J_K\}$:
   - 3.1 Construct the critic-movie ratings matrix R_{-k} for all critics not in J_k;
   - 3.2 Factorize R_{-l} to obtain C^k;
4. Obtain DML estimates of the model specified in Equation 1.

In the first stage, we apply *honest collaborative filtering*, a matrix factorization technique with an "honest" design that isolates evaluator preferences and item characteristics without contaminating these embeddings with traits subject to bias (e.g., gender or race). Specifically:

- We extract bias-free latent preferences P^0, estimated using outcomes from the baseline group ($T = 0$),
- We construct item characteristic embeddings C^k via *sample splitting*, ensuring that each evaluator's ratings do not influence the item representations used in their bias estimation.

In the second stage, with the bias-free latent preferences and independently constructed item embeddings in hand, we estimate micro-level discrimination using a *double machine learning* (DML) estimator. This approach regresses outcomes on the trait of interest, flexibly controlling for the learned latent characteristics and preferences from the first stage to obtain unbiased unit-level estimates of bias.

$$\begin{aligned} r_{i,j} &= \theta_j T_{i,j} + g(P_j^0, C_i^k) + \varepsilon_{j,i} \\ T_{i,j} &= m(P_j^0, C_i^k) + \epsilon_{j,i} \end{aligned} \tag{1}$$

3 Simulation: Gender-Based Discrimination in Job Applications

We simulate a gender-differentiated job application and hiring process, where only application submissions and interview outcomes are observed. Despite the limited information, our method accurately recovers employer-specific gender bias coefficients, even when the researcher mis-specifies the underlying dimensionality and functional form of preferences and characteristics.

Figure 2 compares the employer bias estimates from our method to the true bias parameters in the simulation. Despite misspecification in the latent preference space and the match value function, the estimates remain closely centered around the 45-degree line. To assess statistical significance while controlling for multiple testing, we apply the Benjamini-Hochberg procedure. The null hypothesis of $\theta = 0$ is falsely rejected in only four out of 100 instances where the estimated coefficient has the wrong sign.

4 Empirical Application: Gender-Based Favoritism in Film Reviews

We apply the same methodology to a dataset of professional film reviews to examine potential gender bias in evaluations of female-directed films. Figure 3 compares critics' bias estimates from a standard OLS model with those obtained using our HCF+DML method. Although the estimates are strongly correlated, the OLS model identifies 64 significant coefficients, while the DML method finds only 10 after adjusting for multiple testing. This highlights how failing to account for latent preferences and item characteristics can substantially distort estimates of discrimination.

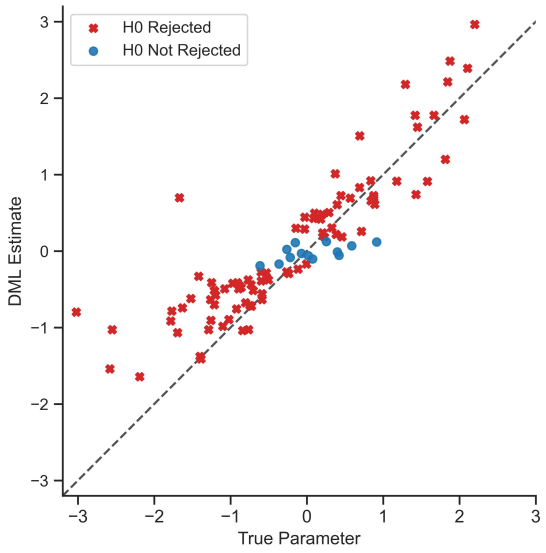

Figure 2: Estimates of Employers' Bias in the Simulated Labor Market.

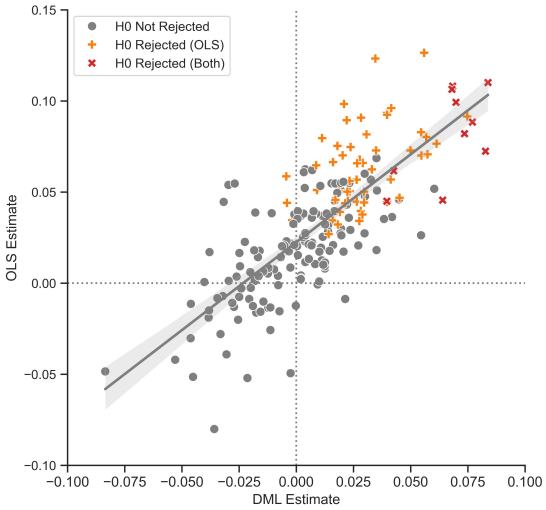

Figure 3: OLS versus DML Estimates of Critic Biases.