Distinguishing Biases from Personal Preferences: An ‘Honest’ Machine Learning Approach

Mahyar Habibi
Bocconi

December, 2023

Abstract

This study proposes a new method for estimating biases at the micro-level in scenarios with multiple bilateral interactions, where the presence of individual preferences and correlated characteristics complicates the analysis. The proposed method comprises two stages. In the first stage, the method introduces a novel approach to extract preferences and characteristics, employing Collaborative Filtering with an 'honest' design. This technique is designed to separate preferences and self-induced outcomes from the constructed embeddings of interacting units. In the second stage, the method utilizes a Double Machine Learning estimator to identify biases at the unit level, based on the embeddings generated in the first stage. The methodology was applied to a dataset of nearly 150,000 film ratings by professional critics, aiming to uncover personal biases among critics towards films directed by women.The results indicate that approximately 5% of critics show a significant bias in favor of films directed by women, once personal preferences and film characteristics are accounted for. However, a ‘naive’ approach that ignores these elements suggests a much higher prevalence of bias among critics.

Keywords: Discrimination, Bias, Collaborative Filtering, Causal Machine Learning.

1 Introduction

Studies in the field of discrimination have predominantly focused on measuring discrimination at an aggregate level. These studies typically examine whether groups, like employers or police officers, demonstrate discriminatory behavior on the whole. Yet, there is a noticeable scarcity of research delving into how widespread and severe discrimination is among individual members or units within these groups. Key questions about the proportion of individuals or units demonstrating discriminatory patterns and the extent of such behavior are largely unaddressed.

Furthermore, the aggregate-level estimates of discrimination, though informative about the average magnitude of such behaviors, obscure the variance of discriminatory practices among micro-level participants within the group under study. This leaves a critical question unanswered: is the observed group-level discrimination a result of a few 'bad apples', or does it indicate a more widespread, albeit subtle, issue? A detailed examination of discriminatory behavior at the individual level is therefore essential, both for a deeper understanding of its causes and for crafting more nuanced and effective policies to counteract these biases.

A major challenge in unit-level discrimination analysis is accounting for unit-specific preferences. Consider a firm's hiring committee evaluating resumes and deciding on interview candidates. The committee selects candidates based on the firm's ideal candidate profile, considering which applicants best match these preferences. However, difficulties arise when these preferred characteristics are associated with attributes like gender, age, or race. For example, firms may differ in how they value learning skills, regardless of an applicant's age. Overlooking this specific preference can lead to incorrect interpretations of age-based discrimination if learning skills and age are correlated. However, in practice, preferences are often complex functions defined over a high-dimensional space of characteristic, many of which remain unobservable to the researcher. This complexity underscores the need for a nuanced approach in discrimination analysis to distinguish between actual discrimination and preference-based selections.

To tackle the complexities arising from unit preferences and characteristics influencing the matching process and outcomes, I introduce a novel approach utilizing causal machine learning (ML) techniques. This method aims to generate unit-level discrimination estimates in situations where each unit within the studied group evaluates multiple individuals or items from a secondary group, and each individual or item in this secondary group is evaluated by multiple units. This framework applies to a broad spectrum of real-world interactions, including but not limited to applications for jobs, housing, credit, as well as online reviews.

In the initial phase, I develop a novel method to extract latent preferences of units and characteristics of items or individuals from the available outcome data. This method builds on the 'honest trees' concept from Athey and Imbens (2016) and incorporates it into collaborative filtering (CF) techniques. This innovative approach is particularly useful in for the effects of preferences and characteristics in cases these factors are not are not explicitly provided. In the second phase, I apply a Double Machine Learning estimator, developed by Chernozhukov et al. (2018), to obtain estimates of unit-level trait-based discrimination, while considering the influence of preferences and characteristics in both the matching and outcome processes.

To test the real-world applicability of the proposed methodology, I collected a dataset of almost 150,000 reviews from Metacritic.com, encompassing over 8,000 films. The aim was to investigate whether film critics demonstrate gender-based discrimination or favoritism in their reviews of films directed by women. Using this methodology, the findings indicated that around 5% of critics showed a preference for films by female directors. Alternatively, a 'naive' model that disregards critics' preferences, the characteristics of the films, and the intrinsic review process suggested that more than 30% of critics gave more favorable reviews to films directed by women. This example emphasizes the importance of considering individual preferences in discrimination studies at the unit level, particularly in contexts where outcomes are significantly influenced by subjective judgments.

The remainder of this paper is organized as follows: Section 2 offers a brief overview of relevant literature. The proposed methodology is described in detail in Section 3. Section 4 discusses the empirical exercise carried out. The paper concludes with Section 5.

2 Literature Review

Becker (1957) taste-based discrimination and the statistical discrimination theory by Arrow (1973) and Phelps (1972) are two principal frameworks in economics for analyzing discriminatory practices. Taste-based discrimination posits that discrimination arises from an individual's personal prejudices or 'tastes' against specific groups. Statistical discrimination, in contrast, is not rooted in personal bias but occurs when decisions are made based on average characteristics of groups, often due to incomplete information about individual capabilities.

Beyond explicit taste-based discrimination, there can be non-discriminatory preferences that incidentally correlate with characteristics of potentially discriminated individuals. Distinguishing between true discrimination and these legitimate preferences is essential. While earlier studies attempted to control for these preferences using observable variables, such approaches often fell short due to unaddressed confounders. Randomized

Controlled Trials (RCTs), like Bertrand and Mullainathan (2004) study on resume callbacks, provide clearer insights. However, RCTs are not always practical or ethical for unit-level discrimination analysis, especially when it involves assigning numerous cases to individual evaluators.

Research on micro-level discrimination has been limited, likely due to data availability issues. Most existing studies in this area have concentrated on law enforcement, where subjective preferences are presumably less influential than in sectors like employment. Ridgeway and MacDonald (2009) identified discriminatory behavior in a small fraction of New York City Police Officers in issuing pedestrain stops. Goncalves and Mello (2021) observed minority drivers receiving fewer ticket discounts in Florida, with more than 40% of officers showing bias. Vomfell and Stewart (2021) further examined police searches in the UK, finding widespread over-searching of ethnic minorities. This paper aims to go a step further by separating biases from broader individual preferences, a distinction that has not been investigated in previous micro-level discrimination studies.

3 Conceptual Framework and Methodology

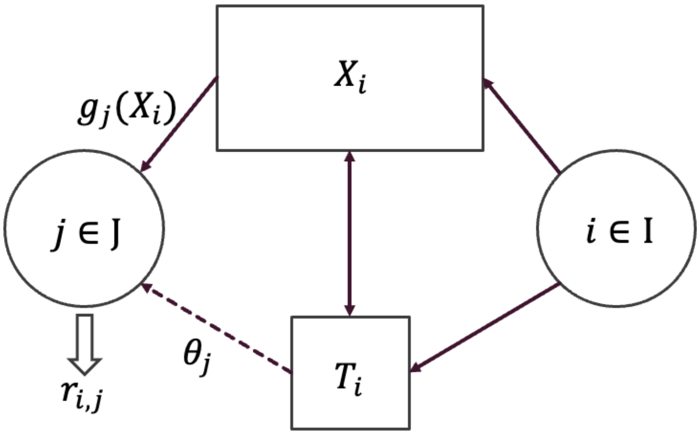
The conceptual framework of this study is based on the interaction between two distinct sets: I, representing items/individuals (e.g., jobseekers), and J, comprising reviewers/judges (e.g., employers). Each individual $i \in I$ is subject to evaluation by multiple reviewers $j \in J$, as determined by a matching process $M(I, J) \rightarrow \{0, 1\}$. As illustrated in Figure 1, during the evaluation, each reviewer j observes not only the general characteristics X_i of the assignee i but also a binary trait $T_i \in \{0, 1\}$, which is the focus of discrimination analysis. A researcher is keen to investigate if the reviewers are influenced by the trait T in their decision-making process, beyond considering the general attributes X of the individuals and the reviewers' non-discriminatory preferences.

Assume the researcher neither observes X nor g, and has no knowledge on the assignment process M. All she observes are traits T, and outcomes $r_{i,j}$. Her objective is to estimate $\theta_j \in \Theta$ for every $j \in J$, as described by the equation:

$$r_{j,i} = \alpha + \theta_j T_i + g_j(X_i) + \varepsilon_{j,i} \tag{1}$$

At first glance, this task seems nearly impossible. How to factor in the unobserved attributes and diverse preferences when no direct evidence are available? The answer lies in a technique widely explored in machine learning: collaborative filtering (CF). This method assumes that if person A has the same opinion as person B on an issue, A is more likely to have B's opinion on a different issue than that of a random person. In essence, collaborative filtering creates a latent database of users' preferences, then uses this data

Figure 1: Conceptual Framewok



Notes: The diagram illustrates the conceptual framework of the analysis. Individual i having characteristics X_i and a binary trait variable T_i is evaluated by individual j. $r_{i,j}$ is the outcome of the evaluation process. θ_j captures the influence of T_i on the outcome.

to predict a user's tastes based on the tastes of similar users. This approach is widely used in various applications, such as recommending books, movies, or music, where the system suggests new items based on the likes and dislikes of similar users rather than analyzing the content of the items themselves.

Collaborative filtering (CF) encompasses a wide range of methods, each suited to various data availability scenarios. In this analysis, I have employed one of the fundamental methods, which requires minimal data input: Regularized Matrix Factorization (RMF), as described by Koren et al. (2009). This method starts with the matrix of reviews $R_{|I|\times|J|}$, where each entry $r_{i,j}$ indicates the assessment of reviewer j for individual i, and missing elements where no review was given. Typically, this matrix is *sparse*, as reviewers usually evaluate only a fraction of the total individuals. RMF effectively transforms the high-dimensional and sparse matrix R into two matrices of lower dimensions, $C_{|I|\times d}$ and $P_{|J|\times d}$, with $d << I, J$ being a hyperparameter of the model. The objective is for the matrix product $W_{|J|\times|I|} = CP'$ to effectively approximate the observed entries in R, while using regularization to avoid over-fitting. For instance, using mean squared error loss and using regularization on the sum of squared parameters in P and C, the loss function would be,

$$L = \sum_{(i,j)\in M} (r_{i,j} - p_j c_i)^2 + \left[\lambda \left(\frac{1}{|I|} \sum C^2 + \frac{1}{|J|} \sum P^2 \right) \right]$$

where M is the set of non missing entries, and λ is a a regularization parameter.

The essence of Regularized Matrix Factorization (RMF) lies in its ability to distill the information in R into two condensed, lower-dimensional matrices: P and C. These matrices act as latent spaces, where P represents the latent preferences of the reviewers, and C captures the characteristics of the individuals being evaluated. In the process of constructing the latent spaces, the model positions similar reviewers and similar individuals/items, close to another in their corresponding matrices. This methodology effectively constructs a refined representation of the characteristics of both reviewers and individuals, based solely on the observed outcomes of their evaluations. RMF's capability to infer rich characteristics from basic outcome data, is what makes it a particularly powerful tool in recommendation systems.

However, applying standard CF methods in examining discrimination can be problematic, as it might incorporate micro-level biases into the resultant embeddings. To mitigate this, I propose a method similar to Athey and Imbens (2016)'s 'honest trees', termed 'honest CF'. This approach begins by ensuring that trait-based biases are not incorporated into the latent preferences of reviewers. It involves factorizing the matrix $R_0 = [r_{i,j}|T_i = 0]$, which includes only the outcomes of individuals from the baseline group with trait $T = 0$. This approach effectively extracts reviewers' latent preferences P^0, uninfluenced by biases associated with the trait.

In the next step, the focus shifts to developing the latent space for individual characteristics, C. The central challenge here lies in separating the outcomes of a particular reviewer, from the representation of items that she has reviewed when estimating the reviewer's trait-based bias to avoid reverse causation. To achieve this, I employ a sample-splitting strategy, dividing the set of reviewers J into K distinct subsets $J_1, J_2, ..., J_K$. Iteratively, I select one subset J_k, use the data from the remaining subsets $J - J_k$ to construct C^k, . This procedure is repeated for each $k \in \{1, ..., K\}$, leading to creation of latent spaces for items that are constructed independently of the outcomes for a particular group of reviewers. Consequently, Equation 1 can be re-formulated in the following form,

$$r_{i,j} = \alpha + \theta_j \, T_i + g(P_j^0 \, , C_i^k) + \varepsilon_{j,i} \tag{2}$$

There are two challenges in estimating the regression specified in Equation 2. First, the matching or assignment process has not been accounted for. Reviewers' preferences and individuals' characteristics are likely to affect the matching process between the two types of players as well as the outcomes of the evaluation. Second, function g needs to be estimated from the data. To overcome these challenges, I use the Double/Debiased ML (DML) method for treatment effect estimation as proposed by Chernozhukov et al. (2018).

To understand the idea behind DML, Consider the following partially linear regression framework as proposed in Robinson (1988):

$$Y = \theta_0 D + g_0(X) + U, \quad \mathbb{E}[U|X, D] = 0$$
$$D = m_0(X) + V, \quad \mathbb{E}[V|X] = 0$$

In this framework, the first equation models the relationship between the treatment variable D and the outcome variable Y. X denotes the vector of control variables influencing both the assignment to treatment and outcomes through unknown functions $m_0(X)$ and $g_0(x)$. The terms V and U represent disturbance variables. This model assumes that conditional on observable features X, the treatment assignment is effectively random, i.e., D is conditionally exogenous. Consequently, θ_0 can be interpreted as the causal effect of treatment on the outcome.

In scenarios where the dimensionality of X is large compared to the sample size N, traditional assumptions facilitating the estimation of $\eta_0 = (m_0, g_0)$ using standard methods become untenable. A simplistic solution might involve employing a machine learning (ML) method to estimate $D\hat{\theta}_0 + \hat{g}_0(X)$ directly. However, such estimator would yield inconsistent estimates of the treatment effect. The inconsistency in this estimation primarily arises from the bias introduced by ML methods while estimating g_0. ML techniques, in an attempt to prevent the estimator's variance from exploding, introduce a bias into the estimator. It is critical to note that even sample splitting does not resolve this issue, as the estimator remains inconsistent even if g_0 is estimated using a separate part of the data. This is because the bias induced by ML methods pertains to the true underlying parameter and is not limited to in-sample bias.

Chernozhukov et al. (2018) propose a method to overcome the regularization biases by using orthogonalization. Their method starts with subtracting the effect of X from D to compute $\hat{V} = D - \hat{m}_0(X)$ where $\hat{m}_0$ is ML estimation of m_0 obtained from the auxiliary sample of observations. After obtaining a preliminary estimate of g_0 from the auxiliary sample, they propose the following double/debiased ML estimator for θ_0 using the main sample of observations,

$$\hat{\theta}_0 = \frac{\sum_{i \in I} \hat{V}_i D_i}{\sum_{i \in I} \hat{V}_i (Y_i - \hat{g}_0(X_i))}.$$

By partialling out the effect of X on D and subtracting $\hat{g}_0$ from the outcome, the regularization induced bias is eliminated in the estimate of $\hat{\theta}_0$.

After minor adjustments, the DML framework can be adapted to estimate micro-level discrimination. This is represented by the model:

$$\begin{aligned} r_{i,j} &= \theta_j T_{i,j} + g(P_j^0, C_i^k) + \varepsilon_{j,i} \\ T_{i,j} &= m(P_j^0, C_i^k) + \epsilon_{j,i} \end{aligned} \tag{3}$$

While the first equation of the model mirrors the interpretation of the earlier framework, the addition of the second equation captures the influence of preferences and characteristics on the matching process between individuals and reviewers.

Algorithm 1 summarizes the proposed methodology to obtain micro-level estimates of discrimination.

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 3.
```

4 Empirical Example

The ensuing section presents an empirical application of the established methodology using a real-world dataset. This approach is instrumental in demonstrating the framework's efficacy to discern micro-level discrimination in practical settings.

4.1 Data Description

A dataset comprising film reviews from professional critics was constructed using data from Metacritic.com, a review aggregator website. Metacritic collects reviews from approximately 100 sources, assigning ratings on a uniform scale of 0-100. These ratings were transformed to a 0-1 scale for this analysis[1]. The objective is to apply the methodology described in Section 3 to explore potential discriminatory patterns in critics' reviews of movies directed by women.

Metacritic provides detailed information, such as the names of film directors. The gender of directors was deduced using their first names and the *gender-guesser* Python library[2], a prevalent tool for name-based gender inference. Entries were removed if *gender-guesser* was unable to make a prediction (name not found in its database) or if the name was

[1] In cases where an explicit rating is absent, Metacritic's evaluators assign a score reflecting their assessment of the article.

[2] https://pypi.org/project/gender-guesser/

Table 1: Summary Statistics of Selected Variables

	Count	Mean	Std.	Min	Med	Max
Year	8,284	2008.7	8.07	1990	2010	2021
Critic Rating	145,522	0.631	0.210	0	67	100
Films' N.o. Critic Reviews	8,284	17.6	9.00	1	16	47

Notes: The table shows summary statistics for the selected variables in the data. The data is limited to the reviews from critics who have evaluated at least 30 movies directed by women.

non-specific to a particular gender. This gender identification procedure was verified for accuracy against a Wikipedia directory of female directors[3], with a misclassification rate under 5%. To maintain simplicity in the analysis, films with more than one director were excluded, as over 95% of movies in the dataset had a single director.

Table I offers a summary of statistics for selected variables in the dataset. Data was collected for films released from 1990 to 2021 and having at least seven critic reviews on Metacritic. This was further narrowed down to critics who had reviewed a minimum of 30 films directed by female directors. The filtered dataset contains over 145,000 reviews from 205 critics, spanning around 8,300 films and 3,900 directors. Films directed by women constitute nearly 14% of the dataset. Each film, on average, garnered reviews from more than 17 critics, with an average rating of 0.63 on a 0-1 scale.

4.2 Estimation

As underscored earlier in this study, the estimation of discrimination or favoritism at the individual level is of considerable significance for several reasons. Chief among them is the potential for aggregate-level bias estimates to be misleading. To illustrate, consider a hypothetical scenario in our context: a seemingly minor bias against movies directed by women could arise either from a general absence of discrimination among critics or from the presence of two distinct groups of reviewers – one disproportionately critical and the other overly favorable towards films by female directors. While both scenarios lead to similar estimates of aggregate-level bias, they depict starkly different realities of micro-level discrimination.

In the assessment of individual-level biases or favoritism, the role of personal preferences among decision-makers is pivotal. For instance, in this analysis, a critic's preference for particular genres or themes – more frequently found in films directed by either gender – might inadvertently color their reviews. This genre or theme preference could manifest as apparent gender bias in reviews, while it truly stems from the critic's own cinematic tastes. Overlooking these personal preferences risks incorrectly categorizing critics as

[3] https://en.wikipedia.org/wiki/List_of_female_film_and_television_directors

biased.

Therefore, I implement the method outlined in Section 3 to obtain individual-level bias/favoritism estimates regarding critics' evaluation of female-directed films. The approach involves estimating the following Double Machine Learning (DML) model

$$\begin{aligned} r_{i,j} &= \theta_j \, FD_{i,j} + g(P^0_j, C^k_i) + \varepsilon_{j,i} \\ FD_{i,j} &= m(P^0_j, C^k_i) + \epsilon_{i,j} \end{aligned} \tag{4}$$

Here, $r_{i,j}$ represents the rating given by critic j to film i, while $FD_{i,j}$ is a binary indicator denoting whether film i was directed by a woman. The parameter θ_j is indicative of the critic-specific bias/favoritism towards films directed by women.

To describe the method in practice, consider the following example of a matrix of ratings,

$$R_{n\times m} = \begin{bmatrix} r_{1,1} & \text{-} & r_{1,3} & \text{-} & \cdots & r_{1,m} \\ r_{2,1} & \text{-} & r_{2,3} & \text{-} & \cdots & \text{-} \\ \text{-} & r_{3,2} & \text{-} & r_{3,4} & \cdots & \text{-} \\ r_{4,1} & \text{-} & \text{-} & \text{-} & \cdots & r_{4,m} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{n,1} & \text{-} & r_{n,3} & \text{-} & \cdots & r_{n,m} \end{bmatrix}$$

In this dataset, typically, critics reviews only a limited selection of films, and correspondingly, each film is assessed by a small group of critics. With a total of over 8,000 films and approximately 200 critics, the dataset comprises less than 150,000 observed ratings, indicating that under 10% of all possible ratings are actually recorded. In matrix R, a '-' signifies a missing rating, denoting a film that a specific critic did not review. This *sparsity* is a common feature in similar contexts where each item or application is evaluated by only a fraction of potential reviewers. Notably, the use of Collaborative Filtering in industrial settings is intended to predict ratings that a user might assign to items they have not yet reviewed (such as books, music, or movies) and to recommend items likely to be highly rated by the user.

As outlined in Algorithm 1, the process begins with applying regularized matrix factorization (RMF) to decompose matrix $R_{I\times J}$ into $C^0_{I\times d}$ and $P^0_{J\times d}$ and . RMF starts by randomly initializing matrices P and C, followed by employing optimization techniques such as gradient descent or its variants to minimize the regularized loss function, detailed in Section 3. This step involves selecting the embedding dimension d and the regularization parameter λ. The value of d was fixed at 100, a commonly adopted figure. For determining λ, a trial and error method was employed, using 10% of the data as a test set and testing several multiples of 10 as potential values for λ. This process led to the selection of $\lambda = 0.01$, which produced a mean squared error (MSE) of 0.074 in the test set.

In the subsequent phase, RMF was implemented following the initial stage outlined in Algorithm 1. Specifically, the RMF algorithm was applied exclusively to the matrix of ratings for films directed by men, omitting the use of a test set. This application aimed to generate P^0, signifying the matrix of critics' latent preferences, deliberately isolated from their evaluations of films directed by women.

Upon deriving P^0, the second step of Algorithm 1 involved randomly dividing critics into $K = 10$ subsets, For each subset k, RMF was then applied to the ratings matrix R_{-k}, comprising ratings data from critics in the remaining subsets, to generate C^k. Here, C^k indicates the film characteristics' embeddings, isolated from the ratings by critics in that particular subset.

In the final step, P^0 and the C^k matrices were utilized to derive DML estimates for the model specified in Equation 4. For these estimates, a binary logistic classifier with L2 regularization was employed as the learner for m in the equation. The regularization parameter was set to the default value of 1, as specified in the *scikit-learn* package. For the learner g, a Random Forest regression was chosen, using the hyperparameters outlined in the *DoubleML* package documentation for the DML estimator in partially linear regression models[4] [5].

To draw a comparison between the outcomes derived by the proposed method and those from a conventional approach, I also estimated the following Ordinary Least Squares (OLS) model:

$$r_{i,j} = \alpha + \beta_j \, FD_{i,j} + \gamma_j + e_{j,i}$$

In this model, β_j represents the OLS estimate of critic j's discrimination/favoritism towards films directed by women. The terms γ_j denotes the critics' fixed effects .

Figure 2 illustrates the distribution of the estimated θ values from Equation 4 using the proposed method, denoted as 'HCF+DML', and compares it with the distribution of the estimated β values from Equation 4.2, labeled 'OLS'. Although there is an overlap between the two distributions, notable differences are evident. The OLS estimates are centered around 0.04, approximately. In contrast, the DML estimates seem to be generally smaller, and centered around, 0.01. Overall, while the mean differences in the two distributions (i.e., the aggregated estimate of discrimination/favoritism) are minor

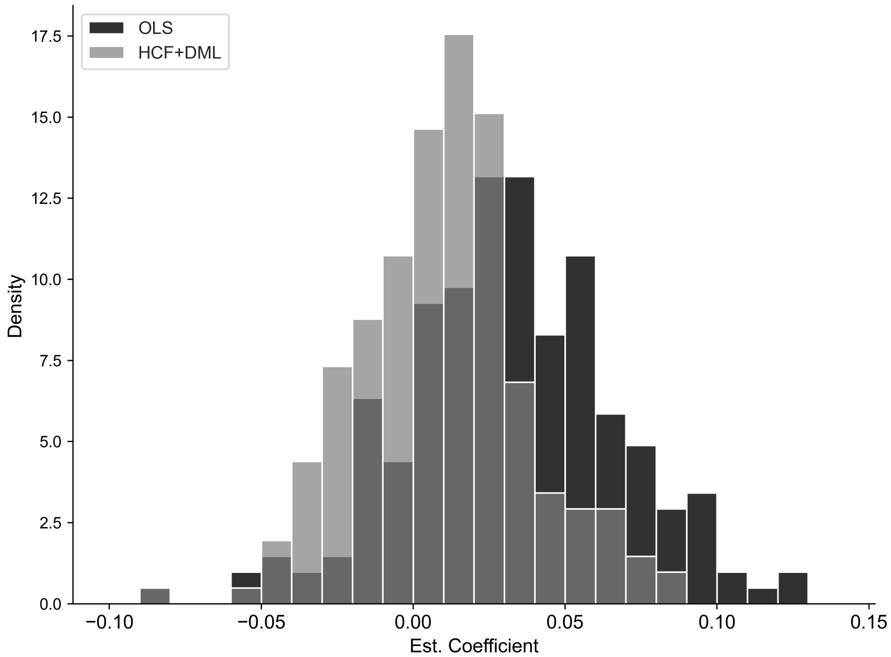
[4] https://docs.doubleml.org/stable/api/generated/doubleml.DoubleMLPLR.html

[5] Since my goal here is to clarify the proposed methodology, I did not engage with hyper-parameter tuning or comparing the results using alternative learning models. However, in practice, trying with different models, and hyper-parameter tuning will be generally helpful to examine the overall robustness of the results.

(around 0.03 scores), there is a marked difference in the distribution tails..

The findings corroborate the hypothesis that aggregate-level estimates might not truly represent the actual distribution of discriminatory behaviors at the individual level. Figure 2 shows that the distribution of OLS estimates, with its heavier right tail on the positive side, implies a greater likelihood of identifying critics who positively discriminate for films directed by women. However, the results from the method employed in this study paint a different picture. Once preferences and characteristics are accounted for, the distribution of estimates indicates far smaller share of critics with high levels of discrimination based on films' directors genders.

Figure 2: Distribution of Estimated Coefficients



Notes: The figure displays the distribution of the micro-level estimates of discrimination obtained via the the proposed methodology (HCF+DML) outlined in Equation 4, and the OLS estimates of the model presented in Equation 4.2.

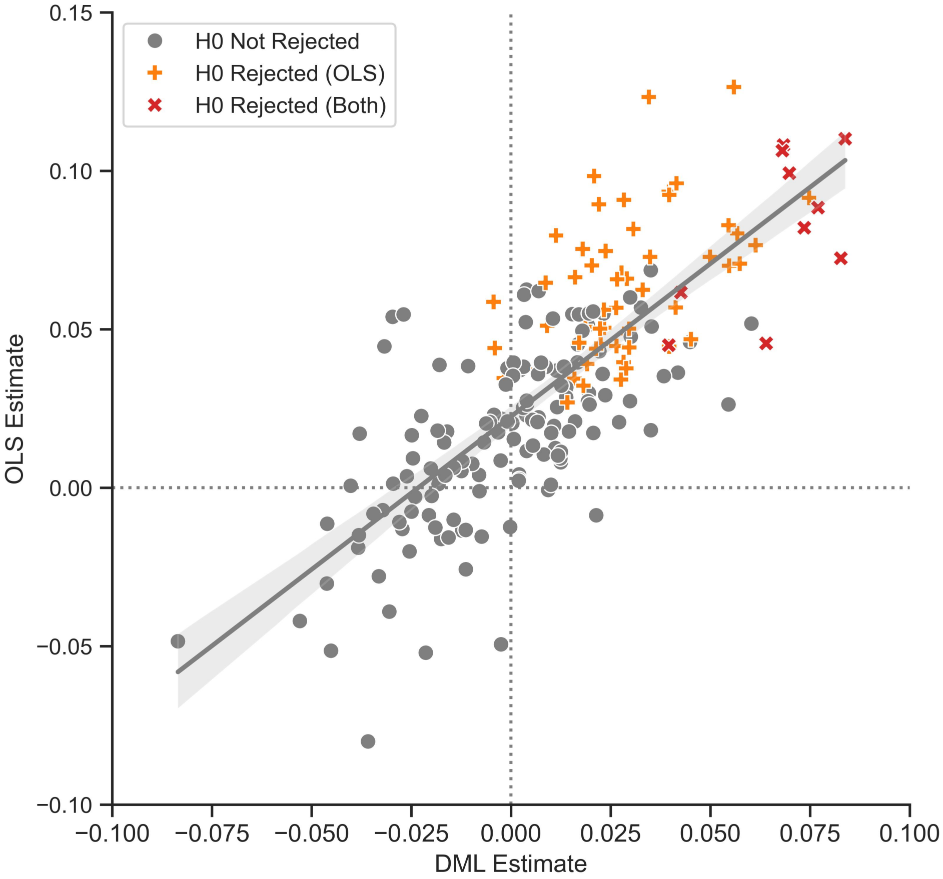
To identify critics whose estimates of discrimination or favoritism are statistically significant from zero, I employed the method by Benjamini and Hochberg (1995) for multiple hypothesis testing[6]. Figure 3 contrasts the estimated coefficients from the OLS model with those obtained via HCF+DML. Although there is a strong correlation between the two sets of estimates, notable differences emerge when assessing the coefficients statistically different from zero using the Benjamini and Hochberg (1995) method with a false

[6] Given the large number of tests, using standard confidence intervals to reject the null hypothesis is not appropriate. Considering a p-value threshold of less than 0.05 for rejecting the null hypothesis of $\beta_c = 0$ in the context of testing 100 estimates would lead to approximately five rejections due to random variation alone. The method by Benjamini and Hochberg (1995) addresses this by controlling the false discovery rate (FDR), i.e., the probability of incorrectly rejecting a true null hypothesis.

discovery rate of 0.10. In the OLS, the null hypothesis is rejected for 64 estimates (all positive). In contrast, the DML estimates reveal 10 statistically significant coefficients.

The findings of this analysis reveal that while factoring in preferences, characteristics, and the matching method as per the proposed methodology in this paper might marginally adjust the overall estimate of discrimination, it significantly alters the micro-level distribution of these estimates. A naive approach, which omits these aspects, suggested that about 30% of critics demonstrate favoritism or discrimination, based on directors' gender . In contrast, the approach used in this study, accounting for these three essential factors, showed that only about 5% of critics show patterns of favoritism toward films directed by women beyond what their preferences and the films' characteristics would justify.

Figure 3: OLS verses DML Estimates



Notes: The figure plots the OLS estimates versus the DML estimates of individual-level discrimination among critics. The statistical significance is tested using Benjamini-Hochberg method using a false discovery rate of 0.10.

5 Conclusion

In this study, I proposed a novel methodology to obtain micro-level estimates of discrimination in an reviewer-applicant setting, that accounts for unobservable preferences, characteristics , and a potentially endogenous matching process . The study proposes 'Honest' Collaborative Filtering, a method to extract latent preferences and characteristics partly isolated from the observed behavior of the individuals. As an empirical example demonstrating the method's use-cases in practice, I analyzed the performance of this method using real-world data from film critic reviews to test for critics' discrimination/favoritism based on gender of the directors. The results suggests that while the aggregate-level estimate of discrimination/favoritism obtained using the proposed method are close to the ones obtained via a naive approach that disregards preferences and characteristics on the two sides, the micro-level estimates provides considerably different pictures on the underlying distribution of discrimination/favoritism.

References

Arrow, K. (1973). The theory of discrimination. *Discrimination in Labor Markets 3*(10), 3–33.

Athey, S. and G. Imbens (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences 113*(27), 7353–7360.

Becker, G. S. (1957). *The Economics of Discrimination.* Chicago: University of Chicago Press.

Benjamini, Y. and Y. Hochberg (1995, January). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological) 57*(1), 289–300.

Bertrand, M. and S. Mullainathan (2004). Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American economic review 94*(4), 991–1013.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018, February). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21*(1), C1–C68.

Goncalves, F. and S. Mello (2021). A few bad apples? racial bias in policing. *American Economic Review 111*(5), 1406–1441.

Koren, Y., R. Bell, and C. Volinsky (2009, August). Matrix Factorization Techniques for Recommender Systems. *Computer 42*(8), 30–37.

Phelps, E. S. (1972). The statistical theory of racism and sexism. *American Economic Review 62*(4), 659–661.

Ridgeway, G. and J. M. MacDonald (2009). Doubly robust internal benchmarking and false discovery rates for detecting racial bias in police stops. *Journal of the American Statistical Association 104*(486), 661–668.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, 931–954.

Vomfell, L. and N. Stewart (2021). Officer bias, over-patrolling and ethnic disparities in stop and search. *Nature Human Behaviour 5*(5), 566–575.