Walk a Mile in Their Shoes: a New Fairness Criterion for Machine Learning

Norman Matloff
matloff@cs.ucdavis.edu
Department of Computer Science
University of California, Davis
Davis, CA 95616, USA

October 14, 2022

Abstract

The old empathetic adage, “Walk a mile in their shoes,” asks that one imagine the difficulties others may face. This suggests a new ML counterfactual fairness criterion, based on a group level: How would members of a nonprotected group fare if their group were subject to conditions in some protected group? Instead of asking what sentence would a particular Caucasian convict receive if he were Black, take that notion to entire groups; e.g. how would the average sentence for all White convicts change if they were Black, but with their same White characteristics, e.g. same number of prior convictions? We frame the problem and study it empirically, for different datasets. Our approach also is a solution to the problem of covariate correlation with sensitive attributes.

1 Introduction

The issue of fairness in machine learning (ML) is the subject of increasing concern, with many interesting algorithms proposed, both for fair prediction and for assessing fairness. As always, though, “The devil is in the details.” Exactly what do we mean by “fair”? Many criteria for fairness have been proposed. See for instance [3] [35] [4] [26] [6] for overviews, some of them written from a critical point of view.
Much work has been done on fair ML from counterfactual points of view. What sentence would this Caucasian convict receive if he had been Black but otherwise with the same relevant characteristics? See [21] for a comprehensive review. In this paper, a novel approach to the counterfactual is proposed.

1.1 Relation to the Legal Realm

Having served as an expert witness in a number of discrimination cases in litigation (e.g. [7]), I view an important point in choosing fairness criteria to be consideration of legal issues, including legal standards for evidence but even more important, development of methodology that is easily understood by judges and juries.

[39], [17] and [2] give detailed analyses of how fairness criteria developed in the ML literature may or may not be consistent with US federal statutes and case law. The legal situation in the European Union is apparently less precisely formulated at this point [36], but will certainly evolve in the near future; this may produce constraints. See [32] and [33] for interesting South and East Asian perspectives.

A key point regarding statistical analyses presented in litigation is the Daubert standard [24]) [12] [16]. The vast majority of fair ML methods do not account for sampling variability, i.e. do not include mechanisms for significance testing and confidence intervals. Accordingly, most published methods may be vulnerable to challenge in court.

Another key point in litigation is that the statistical analysis presented be intuitively clear and reasonable to judges and juries ([34]). Indeed, in the case of judges, the US Federal Judicial Center, a federal agency chaired by the Chief Justice of the Supreme Court, found the issue of statistical literacy pressing enough to commission a guide to statistical methods for judges [11]. This was viewed as essential even though presumably most judges had some exposure to the subject in school; that will not be the case for some, likely most, jurors. It is thus imperative that fair ML methodology be easily grasped on an intuitive level, a major goal of this paper.

There is also concern among many regarding intersectionality, the pernicious interaction between the separate discriminatory actions against different protected groups [9], [10]. As explained in [9],

Intersectionality is a lens for examining societal unfairness which orig-

---

1 The authors in [14] speculate that they are the first to have developed such mechanisms; a literature search seems to confirm this.
inally arose from the observation that sexism and racism have inter-twined effects, in that the harm done to Black women by these two phenomena is more than the sum of the parts.

Methodology aimed at fairness must take intersectionality into account. Linear models of effects of race, gender and age, for instance, may require inclusion of interaction terms.

1.2 Contributions of This Paper

Although much of the presentation here will be motivated by legal aspects, the need for fair ML goes far beyond the legal realm. A firm or government agency, for instance, may be concerned that its AI tool may be effective at addressing some objective function, but at the same time may be unfair to some important groups.

Indeed, no single fairness criterion is best for all possible applications. Thus it is desirable for ML analysts to have available a variety of fairness criteria to choose from.

This paper proposes a new fairness criterion, motivated specifically by these desiderata:

(a) A criterion should be easily relatable to not only ML analysts, but also to non-STEM academics, the press, participants in litigation, and the general public.

(b) As such, a criterion should be kept simple, describable easily in plain language.

(c) Conditional distributions can be confusing. Fairness criteria with lesser explicit involvement with conditional quantities may be preferred.

(d) Again in the interest of simplicity, a criterion should not consist of many numbers; it should consist of a single number for each protected (base group, counterfactual group) pair.

(e) A criterion should include a mechanism for generating associated standard errors, thus enabling statistical inference.

(f) A criterion should be easily adaptable to intersectionality analyses.
A criterion should properly account for covariates that may be related both to the outcome of interest and to status as a member of a protected group.

Concerning that last point, let $X$ and $S$ denote our feature vector and indicator of protected status, respectively. Let $Y$ be the outcome of interest, and let $\hat{Y}$ denote the predicted value of $Y$, given $X$. Ideally we wish that $\hat{Y}$ is independent of $S$, or nearly so. This goal is obstructed by the fact that there may be features in $X$ that are correlated with $S$, so that the model makes some use of $S$ after all. Let’s call this the $X,S$ correlation problem.

### 1.3 Organization of This Paper

As seen in the paper’s title and abstract, the proposed method is inspired by the old empathetic adage, “Walk a mile in their shoes,” which asks that one imagine the difficulties others may face. Accordingly the proposed fairness criterion will be called Walk a Mile (WaM).

WaM will be motivated in Section 2 then described in detail in Section 5. In Section 6 we illustrate the criterion on various real datasets. Section 7 will be devoted to mathematical derivations and empirical methods to assess the statistical accuracy of WaM. Extensions of WaM, such as for false-positive rates, will be discussed in Section 8. Finally, Section 10 consists of conclusions and discussion of future work.

### 2 Introducing WaM

WaM takes the counterfactual view of a protected group as a whole. As a quick example, consider some data from the 2000 US census (similar to the Adult dataset, another common example in fair ML studies). WaM would yield statements like,

Hey, you men out there! Your mean income in this study was $63554.04. But if you were women of the same occupation, age and education, your mean would be only $53829.78.

Or, in COMPAS, an algorithm for assisting judges in sentencing of defendants:
Take note, Caucasian defendants! Your mean decile risk score is 3.67. (Smaller values are better.) But if you were Black while having the same characteristics as yours, your mean would be 4.19.

This directly addresses the X,S correlation problem, in a novel way. Unlike methods that attempt to remove the influence of S in X, we adjust for that influence.

In the employment example above, in which S is gender, the occupation variable in X is strongly related to S. This variable codes for one of six computer-professional occupations in the study, and is strongly gendered:

|       | A  | B  | C  | D  | E  | F  |
|-------|----|----|----|----|----|----|
| men   | 0.20 | 0.22 | 0.34 | 0.02 | 0.04 | 0.17 |
| women | 0.31 | 0.23 | 0.33 | 0.04 | 0.03 | 0.06 |

Men are only 2/3 as likely as women to be in occupation A, while the male rate for occupation F is nearly triple that of women.

And of course there is considerable variation in pay across occupations in the study:

|       | A  | B  | C  | D  | E  | F  |
|-------|----|----|----|----|----|----|
| mean $| 50396 | 51374 | 53640 | 67019 | 68798 | 69494 |

WaM adjusts for this by equalizing the distribution of X: How would men fare with their distribution of X if they were women?

3 Why WaM?

Other than the desideratum regarding standard errors, most of the multitude of existing fairness criteria satisfy at least several of the desiderata listed above. Yet, not all the desiderata are important in all application settings. With that in mind, let us discuss some of the desiderata,

3.1 Use of Conditional Quantities

Desideratum (c) states,

Conditional distributions can be confusing. To the degree possible, fairness criteria that do not explicitly involve conditional quantities are preferred.
This one may be surprising, as a number of fairness criteria are based on explicit conditional quantities, such as the true positive rate (TPR, conditional probability) and regression functions (conditional mean) [3].

Many such criteria are used successfully in practice, such as in [19]. There a credit-equity measure is proposed,

\[
SP = \frac{1}{2} \left| (FPR_{x_α=1} - FPR_{x_α=0}) + (FNR_{x_α=1} - FNR_{x_α=0}) \right|
\]

This may be effective in a corporate internal introspective document, but difficult to comprehend by most jurors.

It is well known that conditional quantities are subject to frequent misunderstanding [30]. Such confusion is not limited to classroom settings, and indeed arises often in ordinary discourse. The authors in [20] cite a number of misleading newspaper headlines, as well as court cases. After the Covid-19 pandemic emerged, for instance, it was common to see news article with titles such as “Coronavirus Testing: What Is a False Positive rate?” [25], and there was much confusion over the efficacy of policy [27]

Careful presentation, say in describing data to a jury, may succeed in ameliorating such problems, but it is desirable to have available criteria that less explicitly cite conditional quantities. Though WaM technically does involve conditional quantities, it does so in an implicit manner that is easier for non-STEM people to grasp.

It should be noted in passing that if a fairness analysis involves ranks, other major problems arise [22], as two individuals may have roughly equal traits yet be far apart in rankings.

3.2 Dimensionality of the Fairness Criterion

Desideratum (d) says,

Again in the interest of simplicity, a criterion should not consist of many numbers; it should consist of a single number for each protected (base group, counterfactual group) pair.

Many methods involve setting up multiple scenarios of real or hypothetical individual cases, with a typical scenario being say, “Consider a 36-year-old single
woman applying for credit, with a current bank balance of $1200, a yearly income of $63,250” and so on. Again, for an internal, introspective corporate analysis, this finely detailed approach may be informative, indeed necessary. But in many settings, again such as litigation, it is best to keep to a single summary number. WaM satisfies this criterion that our fairness measure consists of a single number, as opposed to measures involving pointwise counterfactual fairness, such as [21].

Computing a single summary number is especially useful in situations in which multiple policies, multiple ML prediction methods and so on are being compared over multiple protected groups. If possible, having just a single number for each case in the comparison is desirable.

On the other hand, however, note Einstein’s famous quote, “Everything should be made as simple as possible, but not simpler.” Some settings are fundamentally complex, and it will be seen below that WaM can help uncover more complex fairness problems.

### 3.3 Ability of the Consumer (e.g. Juror) to Empathize

It is well known in legal circles that empathy is crucial in judges and jurors. As pointed out in [23], this is especially important when, for instance, jurors and a defendant are of a different race. Presumably this holds for gender as well.

The proposed method’s name, Walk a Mile, is motivated by empathetic aims, arising from the saying, “Walk a mile in their shoes.” [28] recounts the history of this phrase, which goes back at least as early as the Cherokee tribe of Native Americans.

WaM is aimed at capturing this spirit, helping the intended recipient of the presentation, say a jury, relate emotionally. The hypothetical example given earlier in this paper,

```plaintext
Hey, you men out there! Your mean income in this study was $63554.04. But if you were women of the same occupation, age and education, your mean would be only $53829.78.
```

should be something that strikes a chord with either men or women. Men, for instance, may be somewhat detached when presented with evidence of discrimination against women, but may better relate to comparisons such as the above, which bring in their own gender on a groupwide level.
3.4 Amenability to Intersectionality Issues

The analyst may find that accounting for intersectional issues may complicate implementation for some types of fairness criteria. But intersectionality flows naturally from the definition of WaM, with no special extensions or special cases required. An example will be presented shortly.

3.5 Further Comments

It is not claimed here that WaM uniquely satisfies the desiderata (again, ignoring the standard errors issue). Yet, different application settings may have different goals, and the novel method presented here should be very effective in a number of settings. The WaM approach to measuring fairness should have simple, immediate relevance in many discussion venues, such as policymaking, litigation, social science research and so on. It thus should be part of any analyst’s toolkit.

4 Notation and Background

As above, let $X$ be a vector of features, such as credit history and jobs, and let $S$ indicate membership in a sensitive group, such as a racial minority.

Let $Y$ denote the score value of interest—say a credit rating or a sentence given to a convicted criminal. Or, the “score” may be a direct outcome, such as income in our example above. Note that though one could distinguish between an outcome $Y$ and a decision $D$ as in [26], this is not done here, in the interests of simplicity; indeed, for convenience let us refer to $Y$ as a “policy,” whether a formal one such as the decile risk score, or a virtual one imposed by society such as via discriminatory wages. For binary $Y$, we assume coding as a dummy/one-hot variable, coded 0,1.

Note that while $S$ will typically be a scalar, such as gender or race, it could be an intersectional vector, consisting of gender and race. In the vector case, assume for notational convenience that $S$ has been recoded to scalar form, in which $S$ is categorical, with $s$ values. If $S$ is continuous, such as age, assume that it has been discretized to $s$ values.

Though WaM may remind some of covariate shift methods, such as [29], [40], its goals and operation are actually opposite to those methods. They are concerned with adapting a fitted ML model in the face of future distributional shifts in data to
be predicted. In WaM, we utilize the fact that the distribution of $X|S$ varies with $S$, as a means of measuring unfairness.

One more point to get started: Let $U$ and $V$ be any random variables for which the quantities below exist. Here and below we will be using the Law of Total Expectation,

$$EV = E[ E(V|U)]$$

and the Law of Total Variance ("Pythagorean Theorem" for variance) [31],

$$Var(V) = E[Var(V|U)] + Var[E(V|U)]$$

5 Details of the Method

Now, in the context of the probabilistic generative process of our data, let $X_i$, $i = 1, ..., s$ denote a random vector of length $p$ having the distribution of $X|S = i$. Then define

$$\mu_i(t) = E(Y|X_i = t), \ i = 1, ..., s$$

Here are the quantities that WaM estimates,

$$\nu_{ij} = E[\mu_j(X_i)], \ i, j = 1, ..., s$$

Then $\nu_{ij}$ is the mean outcome that members of group $i$ would have if they were subjected to the policies used for group $j$.

In the case $i = j$, we simply have

$$\nu_{ii} = E[\mu_i(X_i)] = E[E(Y|X)|S_i] = E(Y|S = i)$$

using the Tower Theorem for conditional expectations ([18]). That is, $\nu_{ii}$ is the mean (non-counterfactual) outcome in group $i$.

These quantities $\nu_{ij}$ must be estimated from our data. Let $X_{ik}$ denote the $k^{th}$ realization of $X_i$ in group $i$, $k = 1, ..., n_i$, and let $Y_{ik}$ denote the associated $Y$
value. In other words, considering only the rows $R_i$ of our training data for which $S = i$, then $X_{ik}$ is the value of $X$ in the $k^{th}$ such row, and $Y_{ik}$ is the value of $Y$ in that row. Let $N_i$ denote the number of such rows, with $n$ rows in all.

We fit one’s favorite predictive algorithm, say linear regression or random forests, taking the training set to be $R_i$, yielding the estimated regression function for group $i$, $\hat{\mu}_i(t)$. We can then form the counterfactual outcomes at the individual level,

$$\hat{Y}_{jk} = \hat{\mu}_i(X_{jk})$$

(7)

Our WaM estimates are then

$$\hat{\nu}_{ij} = \frac{N_i}{\sum_{k=1}^{N_i} \hat{\mu}_j(X_{ik})}$$

(8)

For the purpose of forming $\hat{\mu}_i(t)$, linear or logistic regression, k-nearest neighbors, gradient boosting, and random forests all work in a straightforward manner. Neural networks are also fine for continuous $Y$; in the case of categorical $Y$, the final-stage activation function must be chosen appropriately to produce class probabilities.

SVM also is fine for the categorical $Y$ case, provided calibration methods are used on the output scores [13]. For continuous $Y$, various forms of support vector regression are available, even for advanced forms such as quantile regression [15].

It is important to note that unlike many ML fairness criteria, WaM has no problem with unbalanced data, since $\hat{\mu}_i(t)$ is estimated separately for each $i$. However, one does of course need the generative process underlying $(X_{ik}, Y_{ik})$ to be that of $E(Y|X_i)$.

Other than of course requiring that the expectations exist, there are really no statistical assumptions underlying WaM.

### 6 Empirical Evaluation

Here we illustrate WaM via various datasets, using both parametric regression methods, either linear or logistic, and random forests or k-Nearest Neighbors. The software is available at https://github.com/matloff/WAMfair.
The examples are presented solely for illustrating the workings of the method. No feature engineering is performed, and default values are used for hyperparameters.

6.1 Census Data

This is a dataset from the 2000 US Census, focused on Silicon Valley. Variables are age, education, occupation (among six programmer and engineer job categories), gender, wage income, and number of weeks worked.

First, a look at gender (example from above), with a linear model:

|            | men cf. | women cf. |
|------------|---------|-----------|
| men, act.  | 63554.04| 53829.78  |
| women, act.| 58646.87| 50403.48  |

We use “act.” and “cf.” to denote the actual and counterfactual cases. Women as a group had a mean wage of about $50,403, but if they had been men with the same characteristics, their mean would have been $58,647.

The random forest analysis is similar:

|            | men cf. | women cf. |
|------------|---------|-----------|
| men act.   | 63554.04| 53728.65  |
| women act. | 59464.30| 50403.48  |

(The diagonal elements do not change in this second analysis, due to \(A\).)

6.2 Employment Race Test Data

This data is from [5]. Re’sume’s were sent to random employers, with the same qualifications but some with “Black-sounding” given names, such as Lakisha. \(Y\) is an indicator variable recording whether the employer called the applicant after receiving the re’sume’. A number of variables were used, such as education, years of experience, military background and so on.

Only the variables ”education”, ”ofjobs”, ”yearsexp”, ”honors”, ”volunteer”, ”military”, ”empholes”, ”workinschool”, ”email”, ”computerskills”, ”race”, ”call” and ”adid” were used.

Here is the random forests analysis:
We see that “Caucasian” applicants had a 10% callback rate, but would have had only a 3% rate if they had been Black with the same qualifications. Interestingly, though “Black” applicants did not seem to benefit from having “White” names.

As noted, WaM automatically includes intersectional cases, no extension or special cases involved. Again with random forests:

| act | bcf | wcf |
|-----|-----|-----|
| b   | 0.06| 0.05|
| w   | 0.03| 0.10|

Here ‘fb’ means female Black, ’mb’ means male Black, and so on. So, Black women would get a boost from being White men, with their actual 7% callback rate increasing to 10%, but not, for instance, from being Black men. The latter, on the other hand, would benefit by being White men, but not much by being White women.

The latter result, of course, may depend on the job type, and a more segmented analysis is needed, together with domain expertise, but WaM does seem to have uncovered some interesting, complex relationships.

### 6.3 COMPAS

We used the variables 'age', 'juvenile_felony_count', 'decile_score', 'juvenile_misdemeanor_count', 'juvenile_other_count', 'prior_records', 'sex' and 'race'. Due to insufficient data, the Asian and Native American instances were changed to Other.

Using a linear regression model to predict decile score, we have

|       | African-American.cf | Caucasian.cf | Hispanic.cf | Other.cf |
|-------|---------------------|--------------|-------------|----------|
| African-American.act | 5.28               | 4.78         | 4.41        | 4.13     |
| Caucasian.act         | 4.19               | 3.67         | 3.20        | 2.94     |
| Hispanic.act          | 4.36               | 3.77         | 3.32        | 3.07     |
| Other.act             | 4.36               | 3.76         | 3.31        | 3.05     |
We showed two of these numbers in Section 1 above, with possible implications that the COMPAS system rates Black defendants unfairly highly. Various other numbers in the table suggest that as well.

The random forest analysis is similar:

|                | African-American.cf | Caucasian.cf | Hispanic.cf | Other.cf |
|----------------|----------------------|--------------|-------------|----------|
| African-American.act | 5.28                 | 4.73         | 4.28        | 4.16     |
| Caucasian.act     | 4.15                 | 3.67         | 3.20        | 3.00     |
| Hispanic.act      | 4.27                 | 3.74         | 3.32        | 3.07     |
| Other.act         | 4.26                 | 3.70         | 3.22        | 3.05     |

What about gender?

|         | Female | Male |
|---------|--------|------|
| Female  | 4.05   | 3.60 |
| Male    | 5.04   | 4.50 |

The COMPAS algorithm seems to have a substantial bias against women. Their mean decile score was 4.05, but would have been only 3.60 had they been men with the same characteristics.

### 6.4 German Credit Data

This is another commonly used example in ML fairness research, though one of the aspects considered below is less common. Let’s start with the standard aspect, investigating possible gender bias. Using a logistic model for a ‘Good’ credit rating, we have

|         | female.cf | male.cf |
|---------|-----------|---------|
| female.act | 0.72      | 0.64    |
| male.act  | 0.70      | 0.65    |

Women in general had a 72% rate of being assigned a Good rating, but if they were men with the same characteristics, only 64% would get this score. There may be bias against men here. The results using random forests are similar:

|        | female | male |
|--------|--------|------|
| female | 0.72   | 0.66 |
| male   | 0.72   | 0.65 |

Let’s take a look at (discretized) age, using random forests:
Possibly a slight bias in favor of the middle-aged consumers, less favorable to the young and old.

Is there discrimination against foreign workers in the credit realm? Again, running an analysis with random forests, we have

|               | not for.cf | for.cf |
|---------------|------------|--------|
| not for.act   | 0.89       | 0.74   |
| for.act       | 0.84       | 0.69   |

About 69% of foreign workers have Good credit, but if they were citizens with the same characteristics, 84% would be judged good credit risks. So, there may indeed be some discrimination here.

7 Assessment of Sampling Variance

Claims of discrimination are sensitive. As such, there must be some indication of the accuracy of the fairness criterion used, in the form of statistical standard errors ("error bars"). This is vital in academia, the corporate world and especially in litigation, where, as previously noted, case law requires it.

Let’s recall the notation for our data: Define $Y$, $X$ and $X_i$ as before. Denote the $k$th data point in group $i$ by $X_{ik}$ and $Y_{ik}$, respectively. Data points are i.i.d. across $k$ for fixed $i$, and independent across $i$. Let $n_i$ denote the number of data points in group $i$. Vectors are written as columns.

7.1 Exact Expression for Variance in the Linear Case

For notational convenience, it is assumed here that all data has been centered, thus having mean 0. This eliminates the intercept term for the coefficient vector, and the need to have a 1s column in the design matrix.

Here we have

$$E(Y|X_i) = \beta_i^T X_i$$  \hspace{1cm} (9)
where \( \beta_i = (\beta_{i1}, \beta_{i2}, \ldots, \beta_{ip})^T \) and \( T \) denotes matrix transpose.

In this setting,

\[
\nu_{ij} = E(\beta_j^T X_i) = \beta_j^T E(X_i)
\] (10)

Using OLS, we obtain an unbiased linear estimate \( \hat{\beta}_j \) of \( \beta_j \). Denote its covariance matrix by \( \text{Cov}(\hat{\beta}_j) \) (computed via the standard formula and available for instance in R via the \text{vcov()} \ function). We are not assuming normal \( Y_{ij} \), but under our i.i.d. conditions \( \hat{\beta}_j \) will be asymptotically normal.

Then our estimated \( \nu_{ij} \) is

\[
\hat{\nu}_{ij} = \hat{\beta}_j^T A_i
\] (11)

where our sample estimate of \( EX_i \) is

\[
A_i = \frac{1}{n_i} \sum_{k=1}^{n_i} X_{ik}
\] (12)

Then conditioning on \( A_i \) and using (3), we have

\[
\text{Var}(\hat{\beta}_j^T A_i) = \text{Var}(A_i^T \hat{\beta}_j)
\] (13)

\[
= E \left[ \text{Var}(A_i^T \hat{\beta}_j | A_i) \right] + \text{Var} \left[ E(A_i^T \hat{\beta}_j | A_i) \right]
\] (14)

\[
= E \left[ A_i^T \text{Cov}(\hat{\beta}_j) A_i \right] + \text{Var} \left[ A_i^T \beta_j \right]
\] (15)

\[
= E \left[ A_i^T \text{Cov}(\hat{\beta}_j) A_i \right] + \beta_j^T \text{Cov}(A_i) \beta_j
\] (16)

\[
= E \left[ A_i^T \text{Cov}(\hat{\beta}_j) A_i \right] + \frac{1}{n_i} \beta_j^T \text{Cov}(X_i) \beta_j
\] (17)

where \( \text{Cov}(A_i) \) and \( \text{Cov}(X_i) \) are the covariance matrices of \( A_i \) and \( X_i \), respectively. Hence we have the exact expression for variance of WaM.

These quantities can then easily be estimated directly from their sample analogs. In fact, our estimate for (17) is simply
\[ A_i^T \hat{\text{Cov}}(\hat{\beta}_j) A_j + \frac{1}{n_i} \hat{\beta}_i^T \hat{\text{Cov}}(X_i) \hat{\beta}_j \]  

(18)

where \( \hat{\text{Cov}} \) denotes covariance matrices estimated from the data.

Note that in (11) we have the product of a quantity \( \hat{\beta}_i \) that converges in distribution (to a multivariate Gaussian distribution with mean \( \beta_i \)) and a quantity \( A_j \) that converges in probability (to \( EX_i \)). In other words, \( \hat{\nu}_{ij} \) too is asymptotically normal (18), with standard error given by the square root of (18), and confidence intervals can be formed.

### 7.2 The Bootstrap

For nonparametric models, say random forests, one may apply bootstrap methods (137). One does resampling of the data many time, generating many values of \( \hat{\nu}_{ij} \). Our new point estimate is the average of these values, and the standard error is their standard deviation.\(^2\)

### 7.3 Examples

We earlier broke down the COMPAS data by gender, finding an apparent bias against women. Here are the standard errors, based on 100 resamplings:

| act | cf | myGrpEst | theirGrpEst | bias | biasSE |
|-----|----|----------|-------------|------|--------|
| Female | Male   | 4.11 | 3.65 | 0.46 | 0.05 |
| Male | Female | 4.60 | 5.16 | -0.56 | 0.08 |

So for instance an approximate 95% confidence interval for that 0.46 figure would be \( 0.46 \pm 1.96 \times 0.05 \), or about 0.36 to 0.56. The bias is real.

As another example, consider the Employment Race Test Data, Section 6.2. Again with 100 resamples, the bootstrap results were

| act | cf | myGrpEst | myGrpSE | theirGrpEst | theirGrpSE | bias | biasSE |
|-----|----|----------|---------|-------------|-------------|------|--------|
| b   | w  | 0.06 | 0.06 | 0.01 | 0.00 | 0.01 |
| w   | b  | 0.10 | 0.04 | 0.01 | 0.06 | 0.01 |

\(^2\)The software computes only bootstrap results, not the formula derived above for the linear case.
So there is a genuine difference in callback ratings. If the White applicants were
to apply under “Black” names with the same employment-relevant characteristics,
their chances of being called back by employers would only be about 40% as high.

8 Extensions

The WaM approach can be extended, applying it to various common fairness crite-
ria. For instance, we could address questions such as:

Consider loan applications, predicting default, with the criterion false
positive rate. If Caucasian applicants were to have their present qualifi-
cations but were Black, what would their false positive rate be?

Assume and the classification rule (if distributions were fully known)

\[
\hat{Y} = \begin{cases} 
1, & \mu_j(X_i) \geq 0.5 \\
0, & \mu_j(X_i) < 0.5 
\end{cases}
\] (19)

Let’s use \(\gamma_{ij}\) to denote the false positive rate that those in group \(i\) would experience
if they were subjected to the treatment of those in group \(j\). The quantities being
estimated are

\[
\gamma_{ij} = \frac{E\{1_{[0.5,1]}[\mu_j(X_i) \cdot (1 - \mu_j(X_i))]\}}{E\{1_{[0.5,1]}[\mu_j(X_i)]\}}
\] (20)

These are then estimated by their sample analogs.

For instance, consider a Dutch census dataset ([38]). The outcome of interest is
whether a respondent’s occupation is considered prestigious, using various demo-
graphic, economic and household characteristics. Here are the results using a k-
Nearest Neighbors analysis:

|          | male c.f. | female c.f. | myGrpEst | theirGrpEst | bias | biasSE |
|----------|-----------|-------------|----------|-------------|------|--------|
| male     | 1         | 2           | 0.19     | 0.22        | -0.03| 0.01   |
| female   | 2         | 1           | 0.21     | 0.22        | -0.01| 0.01   |

For example, the FP rate for men was about 19%, but would be 22% if they were
women with the same characteristics.
9 Accounting for Model Edge Bias

Fit of one’s regression model is important for WaM. Standard fit assessment techniques may be used, but here we discuss some issues concerning bias at the edges of one’s dataset.

To motivate this discussion, think of a simple setting in which we use k-Nearest Neighbors to predict human weight from height. Consider a data point in which, say, the person is unusually tall. Then his/her neighbors are likely to be shorter than this person, thus lighter. In other words, the prediction on this data point may be biased downward.

The situation with random forests is similar. The leaf in a decision tree into which this hypothetical tall person falls is also a neighborhood, and again, the other points in this leaf will likely be shorter/lighter.

But remedies have been developed for both k-NN [8] and random forests [1], in the form of calculating a linear fit within a neighborhood or leaf. These will soon be incorporated in our WaM software.

10 Conclusions and Future Work

We have introduced a new approach to fairness in machine learning, a group-counterfactual measure we call WaM. It serves as a simple measure of fairness that can be easily grasped by nontechnical people, and it is a solution to the X,S correlation problem.

We believe WaM opens the door to extensive further research. We mention two possible areas:

- Extension to other fairness and predictive power measures: In Section 8, we extended WaM to False Positive Rates. There are various other rates that may be of interest in WaM’s group-counterfactual approach. AUC may also be extended to WaM.

- As explained earlier, any report of a fairness measure in ML should be accompanied by a standard error. This is especially important in litigation, but also vital in any context in which the report will be used in decision making. The bootstrap solution we presented here is convenient, but is time consuming. Moreover, it cannot easily be used in pre-planning, e.g. of sample size,
needed in a study. For such purposes, explicit formulas for estimator variance such as (17) can be helpful, and worthy of further study. Extension of (17) to the generalized linear model, notably the logistic and Poisson regression cases, should be pursued.

References

[1] Athey, S., Tibshirani, J., and Wager, S. Generalized random forests. *The Annals of Statistics* 47, 2 (2019), 1148 – 1178.

[2] Bao, M., Zhou, A., Zottola, S., Brubach, B., Desmarais, S., Horowitz, A., Lum, K., and Venkatasubramanian, S. It’s complicated: The messy relationship between rai datasets and algorithmic fairness benchmarks, 2021.

[3] Barocas, S., Hardt, M., and Narayanan, A. *Fairness and Machine Learning: Limitations and Opportunities*. 2021.

[4] Berk, R., Heidari, H., Jahbari, S., Kearns, M., and Roth, A. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods and Research* 50 (2018).

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

[6] Corbett-Davies, S., and Goel, S. The measure and mismeasure of fairness: A critical review of fair machine learning, 2018.

[7] Court of Appeal, Sixth District, C. Reid v. google inc, 2007.

[8] Elizabeth Yancey, R., Xin, B., and Matloff, N. Modernizing k-nearest neighbors. *Stat* 10, 1 (2021), e335. e335 sta4.335.

[9] Foulds, J. R., and Pan, S. An intersectional definition of fairness. *CoRR* abs/1807.08362 (2018).

[10] Freedman, S. Intersectional discrimination in eu gender equality and non-discrimination.

[11] Freedman, D., and Kaye, D. Reference guide on statistics. In *Reference Manual on Scientific Evidence*. 2011.
[12] GASTWIRTH, J. *Statistical Science in the Courtroom*. Statistics for Social and Behavioral Sciences. Springer New York, 2012.

[13] HUANG, Y., LI, W., MACHERET, F., GABRIEL, R. A., AND OHNOMACHADO, L. A tutorial on calibration measurements and calibration models for clinical prediction models. *Journal of the American Medical Informatics Association* 27, 4 (2020), 621–633.

[14] HURLIN, C., PERIGNON, C., AND SAURIN, S. The fairness of credit scoring models.

[15] HWANG, C., AND SHIM, J. A simple quantile regression via support vector machine. In *Proceedings of the First International Conference on Advances in Natural Computation - Volume Part I* (Berlin, Heidelberg, 2005), ICNC’05, Springer-Verlag, p. 512–520.

[16] KAYE, D. The dynamics of daubert: Methodology, conclusions, and fit in statistical and econometric studies. *Virginia Law Review* 87 (2001).

[17] KLEINBERG, J., LUDWIG, J., MULLAINATHAN, S., AND SUNSTEIN, C. R. Discrimination in the Age of Algorithms. *Journal of Legal Analysis* 10 (04 2019), 113–174.

[18] KLENKE, A. *Probability Theory: A Comprehensive Course*. World Publishing Corporation, 2012.

[19] KOZODOI, N., JACOB, J., AND LESSMANN, S. Fairness in credit scoring: Assessment, implementation and profit implications. *European Journal of Operational Research* 297, 3 (mar 2022), 1083–1094.

[20] KRAMER, W., AND GIGERENZER, G. How to confuse with statistics or: The use and misuse of conditional probabilities.

[21] KUSNER, M. J., LOFTUS, J. R., RUSSELL, C., AND SILVA, R. Counterfactual fairness, 2018.

[22] LAHOTI, P., GUMMADI, K., AND WEIKUM, G. ifair: Learning individually fair data representations for algorithmic decision making.

[23] LINDE, D. Juror empathy and race.

[24] MAHLE, S. The impact of daubert v. merrell dow pharmaceuticals, inc., on expert testimony: with applications to securities litigation. *Florida Bar Journal* 73, 3 (1999).
[25] Maybin, S., and Casserly, J. Coronavirus testing: What is a false positive?, 2020.

[26] Mitchell, S., Potash, E., Barocas, S., D’Amour, A., and Lum, K. Algorithmic fairness: Choices, assumptions, and definitions. Annual Review of Statistics and Its Application 8, 1 (2021), 141–163.

[27] Morris, J. Israeli data: How can efficacy vs. severe disease be strong when 60 of hospitalized are vaccinated?, 2021.

[28] Mueller, S. Developing empathy: Walk a mile in someone’s shoes, 2020.

[29] Polo, F. M., and Vicente, R. Effective sample size, dimensionality, and generalization in covariate shift adaptation, 2021.

[30] Prodromou, T. Secondary school students’ reasoning about conditional probability, samples, and sampling procedures.

[31] Ross, S. A First Course in Probability. Pearson Prentice Hall, 2010.

[32] Sambasivan, N., Arnesen, E., Hutchinson, B., and Prabhakaran, V. Non-portability of algorithmic fairness in india, 2020.

[33] Shin, D. D. Toward fair, accountable, and transparent algorithms: Case studies on algorithm initiatives in korea and china. Javnost - The Public 26, 3 (2019), 274–290.

[34] Thompson, W. Are juries competent to evaluate statistical evidence? Law and Contemporary Problems 52 (1989).

[35] Verma, S., and Rubin, J. Fairness definitions explained. In 2018 IEEE/ACM International Workshop on Software Fairness (FairWare) (2018), pp. 1–7.

[36] Wachter, S., Mittelstadt, B., and Russell, C. Bias preservation in machine learning: The legality of fairness metrics under eu non-discrimination law. West Virginia Law Review 123, 3 (2021).

[37] Wager, S., Hastie, T., and Efron, B. Confidence intervals for random forests: The jackknife and the infinitesimal jackknife, 2014.

[38] World Bank. ’the dutch virtual census of 2001, ipums subset’, 2001.

[39] Xiang, A., and Raji, I. D. On the legal compatibility of fairness definitions, 2019.
[40] ZHANG, T., YAMANE, I., LU, N., AND SUGIYAMA, M. A one-step approach to covariate shift adaptation, 2021.