Columbia Law School
Scholarship Archive

Faculty Scholarship | Faculty Publications

3-2024

Orthogonalizing Inputs

Talia B. Gillis
Columbia Law School, gillis@law.columbia.edu

Follow this and additional works at: https://scholarship.law.columbia.edu/faculty_scholarship

 Part of the Computer Law Commons, and the Science and Technology Law Commons

Recommended Citation

Talia B. Gillis, *Orthogonalizing Inputs*, SYMPOSIUM ON COMPUTER SCIENCE AND LAW (CSLAW '24), MARCH 12-13 (2024).
Available at: https://scholarship.law.columbia.edu/faculty_scholarship/4438

This Article is brought to you for free and open access by the Faculty Publications at Scholarship Archive. It has been accepted for inclusion in Faculty Scholarship by an authorized administrator of Scholarship Archive. For more information, please contact scholarshiparchive@law.columbia.edu.

Orthogonalizing Inputs

Talia B. Gillis
tbg2117@columbia.edu
Columbia University
New York, NY, USA

ABSTRACT

This paper examines an approach to algorithmic discrimination that seeks to blind predictions to protected characteristics by orthogonalizing inputs. The approach uses protected characteristics (such as race or sex) during the training phase of a model but masks these during deployment. The approach posits that including these characteristics in training prevents correlated features from acting as proxies, while assigning uniform values to them at deployment ensures decisions do not vary by group status.

Using a prediction exercise of loan defaults based on mortgage HMDA data and German credit data, the paper highlights the limitations of this orthogonalization strategy. Applying a lasso model, it demonstrates that the selection and weights on protected characteristics are inconsistent. At the deployment stage, where uniform values for race or sex are given to the model, the variations between models lead to meaningful differences in outcomes and resultant disparities.

The core challenge is that orthogonalization assumes an accurate model estimation of the relationship between protected characteristics and outcomes, which can be isolated and neutralized during deployment. In reality, when correlations are pervasive and predictions are constrained by regularization, feature selection can be unstable and driven by the efficiency of the prediction. This analysis casts doubt on the continued reliance on input scrutiny as a strategy in discrimination law and cautions against the myth of algorithmic colorblindness.

ACM Reference Format:
Talia B. Gillis. 2024. Orthogonalizing Inputs. In *Symposium on Computer Science and Law (CSLAW '24), March 12–13, 2024, Boston, MA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3614407.3643698

1 INTRODUCTION

As algorithms are increasingly used in critical decision-making domains, there is growing academic and policy attention to how to capture the benefits of increased prediction accuracy while guaranteeing that outcomes are fair and non-discriminatory. The challenges of how to define fairness and how these notions align with legal definitions of discrimination has been the focus of an extensive algorithmic fairness literature in recent years [Caton and Haas 2020; Mitchell et al. 2021].

This paper focuses on one approach to algorithmic fairness of attempting to blind an algorithm to sensitive or protected characteristics, such as race or sex, as well as their proxies. The approach, which I refer to as 'orthogonalizing inputs' seeks to eliminate the direct and indirect effects of a protected characteristic on a decision by using the predictive power from input variables that is orthogonal to protected characteristics. The conceptual foundation of this strategy was notably developed by Yang and Dobbie [2020][1] who build on a framework developed by Pope and Sydnor [2011]. At the heart of this approach is the use of a protected characteristic, like race, in the initial stages of training the prediction model, with a focus on estimating and isolating the impact of race on the predictive outcome. According to the approach, this prevents other inputs that correlate with the protected characteristic from acting as proxies for the protected characteristic. Then, at the deployment stage when the prediction is used for decision-making the protected characteristic of individuals can be substituted for a uniform value across all groups. This ensures, according to the approach, that the algorithm is devoid of both direct and indirect proxy influences of race.

This paper argues that the orthogonalization approach fails to neutralize the direct and indirect impact of protected characteristics in predictions primarily due to its inability to estimate and isolate the impact of a protected characteristic on predictions in many settings. Although Yang and Dobbie [2020] focus on a linear regression setting, I consider the application of orthogonalization in the machine-learning context. I argue that orthogonalization heavily relies on accurately estimating the decision weights (or coefficients) of a protected characteristic in prediction models, ensuring the impact of a protected characteristic is fully accounted for at deployment during decision-making. Yet, in many settings, especially with high-dimensional data featuring ubiquitous correlations, the weight assigned to a protected characteristic might reflect its tenuous relevance to the prediction and not a reflection of the true estimate of the underlying relationship between a feature and the predicted outcome.

Through a simulation exercise, based on the analysis in Mullainathan and Spiess [2017], I underscore the complexities associated with the orthogonalization technique rooted in the sensitivity of feature selection to noise in the training dataset. Focusing on a prediction of a borrower's default risk using a lasso model, I show that the model does not consistently select the feature 'race' or 'sex' when provided this feature at training and that even when selected for the prediction function, the weight on the feature is inconsistent. The first exercise uses data reported under the Home Mortgage Disclosure Act (HMDA) on mortgage applicants, to which I add simulated default rates. The second exercise uses the German Credit data [Dua and Graff 2017], frequently in computer science research,

[1] Yang and Dobbie refer to this as the "colorblinding-inputs" algorithm, at page 346.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CSLAW '24, March 12–13, 2024, Boston, MA, USA
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0333-1/24/03…$15.00
https://doi.org/10.1145/3614407.3643698

containing loan records of clients at a German bank. The result of these exercises is that the implementation of the orthogonalization approach may yield different results that depend on small and random differences in the training dataset. While my focus is on the variability of weight on protected characteristics arising from small and random differences in the training dataset, variability could potentially be much greater in cases where there is a difference between the training and deployment populations, such as with demographic shift [Giguere et al. 2022], covariate shift [Rezaei et al. 2021] and label shift [Lipton et al. 2018b].

The algorithmic blinding approach of orthogonalization merits special attention. In the algorithmic fairness literature the shortcomings of algorithmic blinding were documented early on and some scholars have suggested that blindness should not be treated as an algorithmic fairness notion at all [Makhlouf et al. 2021, see e.g.,]. However, the dominance of scrutinizing inputs as a way to approach algorithmic discrimination persists in legal and policy debates [Gillis 2022]. The recent Supreme Court decision in *Students for Fair Admissions* [SFF 2023], signifies the ascendancy of race-neutral interpretations in discrimination law, potentially influencing AI decision-making and domains that increasingly rely on algorithms like employment and lending decisions.[2] Given these prevailing perspectives, a thorough examination of orthogonalization, seen as an advanced blinding method, is imperative. This paper aims to spotlight the inherent challenges and inefficacies in pursuing fairness through neutralizing inputs.

2 RELATED WORK

This paper discusses a particular approach for verifying that algorithmic predictions are fair and non-discriminatory and therefore sits at the intersection of the technical algorithmic fairness literature and discussions of algorithmic discrimination in the law and policy literature. The broader literature on defining, measuring, and ensuring fair algorithms has been surveyed recently by Mitchell et al. [2021], Caton and Haas [2020] and Pessach and Shmueli [2023], among others. These methodologies are typically categorized by the phase at which they intervene in the machine learning pipeline. Notably, orthogonalization, the focal point of this paper, is a pre-processing method, emphasizing fairness by blinding the deployment algorithm to protected characteristics and their proxies [Kusner et al. 2017].

Drawing inspiration from the field of econometrics, the cornerstone of orthogonalization, as articulated in Yang and Dobbie [2020] and Pope and Sydnor [2011], is countering potential biases stemming from correlated traits acting as proxies for the protected attributes. This builds upon the econometrics literature addressing 'omitted variable bias' and the broader repercussions of variable exclusion in regression analyses[Angrist and Pischke 2009; Jung et al. 2018].

Distinguishing the use of protected characteristics at the training and deployment stages, a key feature of the orthogonalization approach, has also been discussed in the algorithmic fairness literature. The division between training and deployment with respect to the use of protected characteristics is partly driven by the requirement, on the one hand, that decisions not vary on the basis of protected characteristics, with the recognition, on the hand, that many techniques to mitigate an the use of proxies requires awareness of protected characteristics [Harned and Wallach 2019; Kim 2022]. Comparable strategies, termed Disparate Learning Processes (DLPs), leverage protected traits in training but abstain from them during deployment [Kamiran and Calders 2012; Kamishima et al. 2011]. These approaches have encountered criticism regarding potential revelations of protected characteristics and their influence on prediction accuracy [Lipton et al. 2018a]. Blindness to protected characteristics has been criticized on other grounds, relating primarily to the impact of blindness techniques on fairness of outcomes. Kleinberg et al. [2018], for example, argue that genuine fairness efforts, aimed at outcome equalization, should be race-aware even during decision-making.

The paper also relates to legal debates surrounding the algorithmic omission of protected characteristics and their proxies, as well as the challenges of adapting traditional discrimination doctrines to algorithmic settings. While Yang and Dobbie [2020] advocate for algorithmic blindness mandated by the Equal Protection Clause, others suggest doctrines like 'disparate impact' may necessitate a race-conscious approach [Bent 2019; Gillis 2022; Gillis and Spiess 2019; Kim 2022].

There has been some work to directly engage with the approach in Yang and Dobbie [2020] and the foundational framework in Pope and Sydnor [2011]. Key contributions include Bartlett et al. [2021], who demonstrate the challenges of debiasing proxies, and Altenburger and Ho [2019], who suggest that under certain circumstances, simple exclusion of protected traits might be preferable to orthogonalization when applying the method to a random forest algorithm.

In this paper, I focus on the implementation of the orthogonalization approach in the machine learning setting and show the practical implications of the instability of feature selection. Although the Yang and Dobbie [2020] method originated in the linear regression context, its adaptation to machine learning realms, where correlations may be ubiquitous, remains under-explored.[3] The need to consider the application of the method to machine learning is particularly important given its emergence as a policy solution to AI discrimination [Prince and Schwarcz 2019]. My focus here is on the disconnect between strategies targeting model estimation and prediction technologies purely aimed at optimizing predictions, illuminating the ensuing policy ambiguity.

3 ORTHOGONALIZING INPUTS

The starting point for the orthogonalization approach is that fair and non-discriminatory decision-making requires that decisions be blind to protected characteristics. While this position remains

[2]The case centered on college admissions and Equal Protection Clause, however, it potentially speaks more generally to the consideration of race. In particular, Justice Gorsuch's concurring opinion argued that Title VI of the Civil Rights Act bars affirmative action and that Title VII, employment discrimination, contains similar language. Because of the close relationship between employment discrimination and fair lending, such as Title VIII Fair Housing Act, this gestures towards Gorsuch's understanding of a race-blind requirement in other domains.

[3]Pope and Sydnor [2011] apply their analysis to logit regression analysis but do not consider machine learning models. Altenburger and Ho [2019] analyze the framework using a random forest but do not consider the instability properties of the prediction. Importantly, their use of 'contentious' and 'correlated' variables departs from the original approach in Yang and Dobbie [2020].

somewhat contentious [Kim 2022], may overlook the systemic inequalities and historical biases that shape features and labels [Gillis 2022] and could depend on the particular legal domain of discrimination law, the notion of some formal requirement of blindness to protected characteristics, meaning that decisions do not vary on the basis of these characteristics, is ubiquitous in computer science literature and legal discussions.

The orthogonalization approach can be seen as a response to the challenges of simply omitting a protected characteristics from an algorithm's inputs. The approach demonstrates how when a protected characteristic is excluded, the coefficients or inputs that correlate with the protected characteristic can partially reflect the omitted protected characteristic, a statistical phenomena known more generally as 'omitted variable bias.' This arises because variables that correlate with a protected characteristic can play a dual role–a features's predictive power might arise from its ability to predict an outcome independent of a protected characteristic but also from its role as a proxy for a protected characteristic. An example might be the prediction of credit risk using employment history, where employment history is correlated with race. A consumer's employment history may provide important information on credit risk, however it may also act a proxy for race. Avery et al. [2012] document, for example, how a credit file variable 'average age of accounts on credit report' serves as a proxy for borrower age in a prediction of loan performance in this way.[4]

The key distinction under orthogonalization is separating the model training stage from the decision-making or deployment stage with respect to the use of protected characteristics. Focusing on the example of race, in the training stage the algorithm is race-aware in the sense that the algorithm uses 'race' as one of its inputs. This produces an estimate of the weight given to race in forming the prediction. However, in the screening stage, meaning the stage at which the prediction is applied to a particular person, the algorithm is given an uninformative value of an individual's race. This means that even though 'race' was used to train the algorithm, there is no differential treatment on the basis of race at deployment.

Formally, Yang and Dobbie [2020] consider a case in which we are trying to predict outcome Y_i for individual i, where there are three types of inputs. There is the protected characteristic, such as race, $X_i^{protected}$. Then there are inputs that correlate with race, $X_i^{correlated}$, and inputs that do not correlate with race, $X_i^{noncorrelated}$. Consider, for example, a lender looking to predict default risk of loan applicants where some applicants are non-White minority applicants, $X_i^{noncorrelated}$ might be characteristics like 'age'[5] that are not correlated with racial minority status, and $X_i^{correlated}$ might be applicant income and credit history that correlate with race. The lender first uses past loans and their outcome (whether they were defaulted on or not) to estimate a prediction model and then uses this model to predict default risk for new applicants.

One possibility is to estimate a race-blind linear model:

$$Y_i = \beta_0 + \beta_1 X_i^{noncorrelated} + \beta_2 X_i^{correlated} + \epsilon_i \quad (1)$$

Yang and Dobbie [2020] show that simply estimating a linear model without race would mean that β_2 would not only reflect the direct impact of those features on the outcome but also some of the weight of 'race' which has been omitted. They therefore argue that the following linear regression should be estimated at the training stage:

$$Y_i = \beta_0 + \beta_1 X_i^{noncorrelated} + \beta_2 X_i^{correlated} + \beta_3 X_i^{protected} + \epsilon_i \quad (2)$$

Estimating equation 2 produces coefficients β_1, β_2 and β_3, where β_2 no longer partially reflects the role of $X_i^{correlated}$ as a proxy for 'race.' Applying the estimated function to predict default for future borrowers would arguably trigger 'disparate treatment' prohibitions as it treats borrowers differently on the basis of race. Therefore, when applying this model to future borrowers, 'race' is set to the same value for all borrowers, for example mean race ($\bar{X}^{protected}$), so that the model does not distinguish between different racial groups when implemented. Arguably, this satisfies the requirement to not discriminate on the basis of race while preventing correlates from serving as proxies for race.

4 ORTHOGONALIZATION APPROACH IN PRACTICE

How would this framework apply in the context of machine learning? In their paper, Yang and Dobbie [2020] apply this method to an ordinary least squares (OLS) or linear regression and do not demonstrate their method in the machine learning context. Prince and Schwarcz [2019] provide a general discussion of the method in the context of artificial intelligence but do not discuss the implementation details.[6] To assess the method's efficacy in machine learning, I will use the lasso algorithm owing to its analogous output characteristics with linear regression. Future work should consider the implications of the approach in other model classes.

I use two simulation exercises to consider how the orthogonalization approach might play out in practice. The first exercise considers a hypothetical mortgage lender and relies on data reported under the Home Mortgage Disclosure Act (HMDA).[7] The second exercise is based on the publicly available the German credit data used in computer science research.[8]

4.1 HMDA Data

To demonstrate how this approach would play out in practice I consider a hypothetical lender. This lender takes data on past loans and loan performance to predict the default risk of new borrowers. My hypothetical lender uses loan information reported by mortgage lenders under HMDA to predict creditworthiness.

Specifically, I use the Boston Fed HMDA dataset with more than 40 variables (many of which are categorical, taking on a fixed number of possible values) to which I add simulated default rates. Default rates are simulated because HMDA contains only mortgage

[4] Avery et al. [2012] do not document the same proxy effect for race and gender. Interestingly, Avery et al. [2012] refer to a situation where a variable's predictive power arises from acting as a proxy for a protected characteristic as 'disparate impact' and not 'disparate treatment.' This is surprising given that the proxy effect is arguably leading to a direct conditioning on a protected characteristic.

[5] Note that in some settings, age itself could be a protected characteristic.

[6] Pope and Sydnor [2011] also focus on the application to an OLS regression but expand the analysis to probit and logit regressions.

[7] Home Mortgage Disclosure Act (HMDA) (12 U.S.C. § 2801)).

[8] The dataset can be found at https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data.

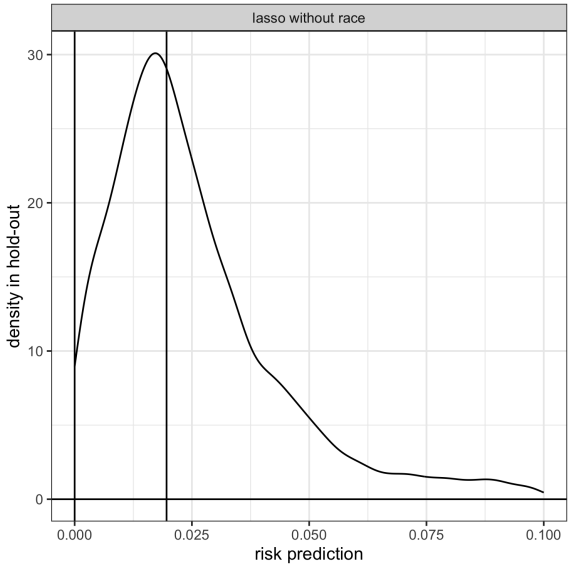

Figure 1: Distribution of predicted risk

The graph shows the distribution of predicted default probabilities for all borrowers in the holdout set of 2,000 borrowers. The graph is cutoff at 10%*, meaning that only borrowers with a default risk of less than* 10% *are plotted. The vertical line is the median predicted default probability (of the full sample, not just the borrowers with a risk below* 10%*).*

application information and not mortgage performance information. Simulated default rates are produced by a model that relates application outcomes to borrower and loan characteristics which are then calibrated to known default rates from the literature.[9] Details on the Boston Fed HMDA dataset [Munnell et al. 1996] and the model I use to simulate default rates can be found in previous work [Gillis 2022; Gillis and Spiess 2019].

One important feature of HMDA reporting, is that mortgage originators are required to report the race of a loan applicant so that the HMDA dataset includes information on whether a borrower is a non-Hispanic White borrower or a minority borrower [Munnell et al. 1996].

4.1.1 Simulation. The prediction of loan default as a function of individual characteristics of the loan and applicant is made using a lasso regression for its interpretability and similarity to a standard linear regression. The algorithm is trained on a sample of 2,000 past borrowers. This function can then be applied to new borrowers, which is a subset of borrowers from the HMDA dataset.

In Figure 1 the model's prediction function is applied to a holdout set, meaning a subset of 2,000 borrowers that is drawn from the same distribution but was not used to train the prediction function. In the real world, this is likely to be a group of new applicants for which the lender is deciding whether to extend a loan and at what price. Borrowers who are to the left of the distribution have a lower probability of default. The default probabilities can either be used as thresholds for binary lending decisions or as a way to price loans for credit risk.

To demonstrate the practical challenges in applying the orthoganlization method to machine learning I repeat the exercise of fitting a model to the training data set 10 times, each time with a slightly different training set. In the repeated training of the models, I include the 'race' feature so that the function can select 'race' as a predictor and determine the weight on the feature.[10] To create 10 comparable datasets with slightly different noise, I randomly draw 2,000 observations from my full dataset 10 times, in a procedure similar to Mullainathan and Spiess [2017]. Because these 10 datasets are randomly drawn from the same full sample, they should be similar, although they are unlikely to be identical. I then fit a lasso regression to each of the 10 training datasets and let the algorithm choose which of the many characteristics to include in the model.

4.1.2 Results. When comparing the function trained on the 10 iterations of the training set the feature selection and weights vary. Importantly, whether the model selects the feature 'race' in its prediction function and the weight of 'race' varies by iteration. Figure 2 plots the weights on the variable 'race' where each column represents a different random draw from the data set, which resulted in different prediction functions. We can see that for 8 of the draws, the feature 'race' is not selected at all. In draw 5 the 'race' feature receives the largest weight of around -0.02, which can be interpreted as meaning that non-Hispanic White borrower have a predicted default probability that is 2 percentage points lower that a racial minority borrower. Considering that the average default probability is only 1.7%, this is a meaningful difference. Iteration 10 has a smaller weight of around -0.012, which can be interpreted as a decrease of predicted default probability of 1.2 percentage points for non-Hispanic White borrowers. The weights on 'race' in the rest of the iterations is 0.

The results so far reflect the different weights on the 'race' variable at the training stage. Because the orthogonalization approach distinguishes between the use of a protected characteristic during the training and deployment stages, we now consider how the prediction functions are applied when the individual's race is hidden at the deployment stage. As discussed above, this would be achieved by substituting the race of all new loan applicants with a uniform value.

Applying the orthogonalization method would lead to different results depending on the iteration. To see this, consider the resulting predictions if after the training stage of the model we provide uniform 'race' information at the screening stage. This would mean that all applicants would receive the same 'race' value regardless of their race such as mean race, $\bar{X}_{protected}$,[11] and that the weight on this variable would be determined by the weight of the function that was trained on the specific iteration of the training dataset. Once we have the prediction for the 2,000 applicants we can examine the disparities in default prediction for White and minority applicants. These disparities would not be the result of a direct consideration of an applicant's race, as this information is not provided at the

[9] Given this limitation of the data, the analysis should not be interpreted as an empirical analysis of default rates but limited to the conceptual argument. One way to consider the exercise is that the prediction is closer to a prediction of application outcomes than of mortgage default. While the labels needed for this prediction are contained in HMDA this type of prediction exercise lacks real world realism as lenders are unlikely to build prediction models for their own loan decisions.

[10] The 'race' feature in my simulation is a binary variable that equals 1 if a borrower is non-Hispanic White and 0 if the borrower belongs to a racial minority, such as Black, Asian or Hispanic.

[11] In this case $\bar{X}_{race} = 0.8$, because there are many more White borrowers in the dataset.

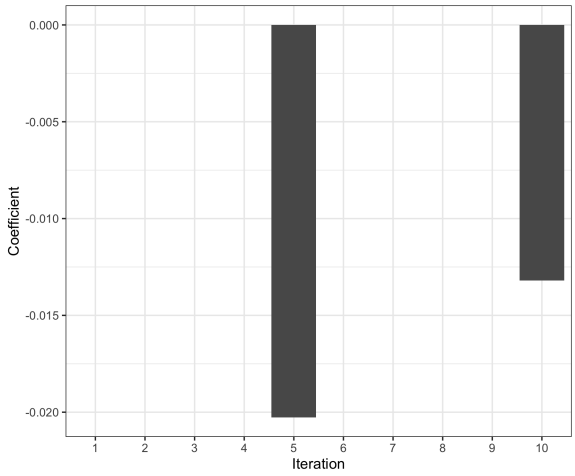

Figure 2: Weight on race feature

Each of the 10 columns represents a different random draw of 2,000 observations from the full data, on which a lasso regression was fitted. The columns plot the weight the lasso regression placed on the race variable, which is one of the 40 model inputs.

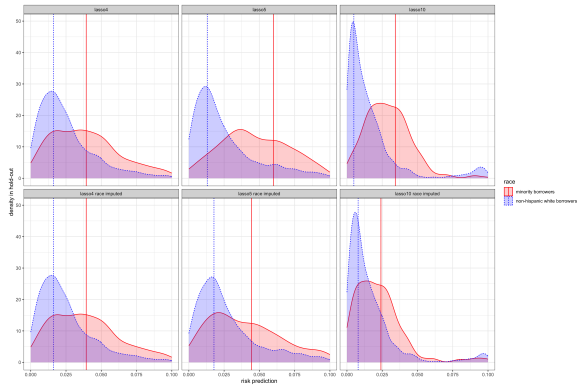

Figure 3: Disparities after orthogonalizing race

The top row shows the distribution of the risk predictions for non-Hispanic White borrowers and racial minority borrowers, from a lasso regression using all inputs. The graph on the top left corner is the prediction from random draw 4 of the dataset, the graph at the top in the middle is the prediction from random draw 5 and the graph on the top right corner is the prediction from random draw 10 of the dataset for training purposes. The bottom row shows the prediction when applicant race is substituted for a uniform value of 'mean race' (0.8). The vertical lines represent the median predicted default probability for non-Hispanic White borrowers and racial minority borrowers separately.

screening stage, but rather a result of the different distributions of the other features (credit history, income, etc.) for White and minority applicants.

Figure 3, plots the default probably separately for White and minority applicants for 3 of the 10 iterations of training sets using a holdout set not used for training the model, as if they were a new group of loan applicants. The top row shows the disparities produced by iteration 4 (top left figure), iteration 5 (top middle figure) and iteration 10 (top right figure) if we were to apply the prediction function to new applicants including their race. Despite the lasso being trained on very similar training datasets, the disparities for White and minority borrowers are different in the three iterations.

In iteration 5, where 'race' is selected and receives a large negative weight, the disparities between White and minority borrowers are greater than in iteration 4 where 'race' is not selection. The disparities in iteration 10, where the weight is negative but of a smaller magnitude than 5, disparities are larger than 4 but smaller than 5. Note that even when 'race' is not selected, as in iteration 4, there are still significant disparities between White and minority borrowers.

The orthogonalization method would not allow the deployment of the prediction function as reflected in the top row of Figure 3 and would instead require that for the new applicants being considered 'race' be substituted with a uniform value for all applicants. The lower figures reflect the predicted default disparities for the new applicants when applying the orthogonalization method using the value of 'mean race' instead of the individual's true race.

For iteration 4, the left column of Figure 3, the top and bottom row are identical, and the median for White and minority borrowers does not change (the vertical lines represents the median for White and minority borrowers). This is because in iteration 4, the variable 'race' has no weight so that substituting the characteristic does not change the prediction function. Comparing iterations 5 and 10 reveals that reduction in disparities is more meaningful with iteration 5 when substituting the 'race' variable for a uniform value, which is what we would expect with a larger model weight. The comparison of all three iterations demonstrates the inconsistency of the method in reducing disparities and the remaining disparities after orthogonalization.

The conclusion of this exercise is that even though the training datasets of iterations 4, 5 and 10 are very similar, the lasso regression made different choices with respect to the weight on 'race.' The orthogonalization method, which uses the coefficient or weight on 'race' for the screening and deployment stage, will therefore yield different results based on the random draw.[12]

4.2 German Credit Data

The second simulation also considers a hypothetical lender that takes data on past loans and their performance to predict the default risk of new borrowers. The experiment is based on the German credit data [Dua and Graff 2017] used in computer science research and frequently in machine learning credit scoring research [see e.g., Dastile et al. 2020; Lin et al. 2012]. The dataset contains 1,000 observations with 700 borrowers classified as a 'good' credit risk (creditworthy) and 300 as a 'bad' credit risk (not creditworthy). There are 30 features for each borrower, including borrower 'sex' which will be the protected characteristic for this example. The other features cover characteristics like income, age, employment, marital status and information on assets and past credit.

4.2.1 Simulation. Similar to the HMDA simulation, a lasso will be used to fit a prediction of creditworthiness (probability of 'bad' risk) using features from the training data set. The algorithm is trained on a sample of 700 past borrowers and then applied to 300 new borrowers, which is a subset of borrowers from the German credit dataset.

[12]For a general discussion of the instability with respect to the variables selected by the lasso regression see Mullainathan and Spiess [2017] and for the application to fair lending see Gillis and Spiess [2019].

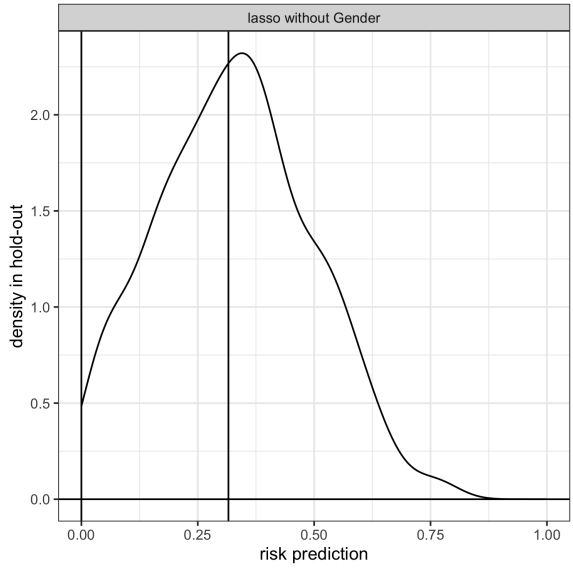

Figure 4: Distribution of predicted risk

The graph shows the distribution of 'bad' credit risk probability for all borrowers in the holdout set of 300 *borrowers. The vertical line is the median predicted default probability,*

In Figure 4 the model's prediction function is applied to a holdout set, meaning a subset of 300 borrowers that is drawn from the same distribution but was not used to train the prediction function. Similar to the HMDA example, the exercise of fitting a model to the training dataset is repeated 10 times, each time with a randomly drawn training set. In the repeated training of the models, 'sex' is included as an input so that the function can select 'sex' as a predictor and determine the weight on the feature.[13]

4.2.2 Results. Similar to the HMDA example, whether the model selects the feature 'sex' in its prediction function and the weight of 'sex' varies by iteration. Figure 5 plots the weights on the variable 'sex' where each column represents a different random draw from the data set and therefore a different prediction function. For 3 of the draws, the feature 'sex' is not selected at all. There are 5 draws for which there is a negative weight on 'sex' (with varying weights) and 2 draws for which the sign of the weight flips to positive. To understand the meaning of the weights on sex, iteration 4, for example, puts a weight of roughly −0.05 on the feature 'sex' so that being male is associated with a 5 percentage point decrease in predicted default probability.

Here too, applying the orthogonalization method would lead to different results depending on the iteration. At deployment, instead of providing the applicant's sex, the function is provided with the mean sex value, $\bar{X}_{protected}$,[14] and the weight on this variable is determined by the weight of the function that was trained on the specific iteration of the training dataset.

[13]The 'sex' feature in my simulation is a binary variable that equals 1 if a borrower is male and 0 if the borrower is female.

[14]In the German credit data, roughly 70% of the dataset is male so that $\bar{X}_{protected}$ = 0.7.

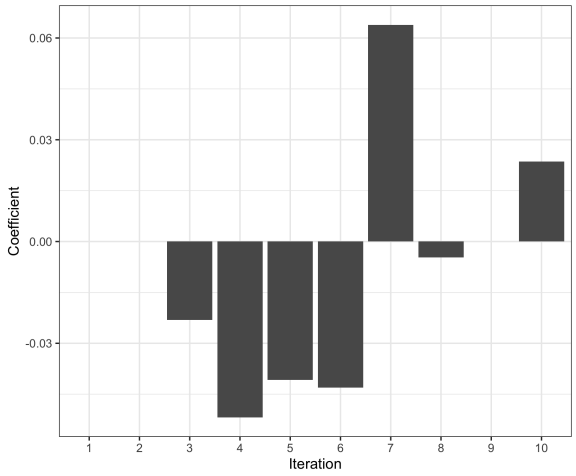

Figure 5: Weight on sex feature

Each of the 10 *columns represents a different random draw of 700 observations from the full data, on which a lasso regression was fitted. The columns plot the weight the lasso regression placed on the sex variable, which is one of the* 30 *model inputs.*

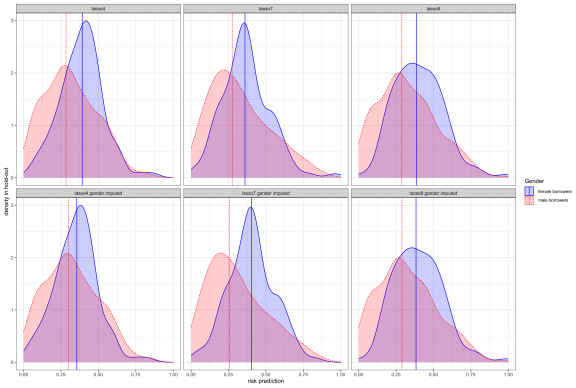

Figure 6: Disparities after orthogonalizing sex

The top row shows the distribution of the risk predictions for female and male borrowers, from a lasso regression using all inputs. The graph on the top left corner is the prediction from random draw 4 *of the dataset, the graph at the top middle is the prediction from random draw* 7*, and the graph on the top right corner is the prediction from random draw* 8 *of the dataset. The bottom row shows the prediction when applicant sex is substituted for a uniform value of 'mean sex'* (0.7)*. The vertical lines represent the median predicted default probability for female and male applicants.*

Figure 6, plots the default probably separately for female and male applicants for 3 of the 10 iterations of training sets using a holdout set not used for training the model, as if they were a new group of loan applicants. The top row shows the disparities produced by iteration 4 (top left figure), iteration 7 (top middle figure) and iteration 8 (top right figure) if we were to apply the prediction function to new applicants including their sex. Despite the lasso being trained on very similar training datasets, the disparities for female and male borrowers are different in the three iterations. In iteration 8, where 'sex' receives a small negative weight, the disparities between female and male does not change much at deployment. In iteration 7, where the weight on 'sex' is positive, the disparity

increases after deployment.[15] In iteration is 3, where 'sex' receives a negative weight, disparities decrease with substitution of sex for a uniform value at deployment. Overall, this simulation exhibits even greater variability than with the HMDA data, which may partially be due to the smaller dataset.

5 DISCUSSION

The simulation exercises in the previous sections reveal a limitation in how to interpret machine learning algorithms. In a standard regression analysis, the coefficients represent some estimation of the impact of the independent variables on the predicted dependent variables. This is not necessarily the case with a lasso model. Although the lasso regression output function looks similar to the output of an OLS regression, it should be interpreted differently. Machine learning is constructed to optimize the prediction accuracy, particularly in high dimensional data when there may be many correlated features. Therefore, the fact that even small amounts of noise in the data can change the variables that are selected by the algorithm in forming the prediction may not matter as long as the prediction accuracy is somewhat stable. When there are many possible features that predictions can depend on, and algorithms choose from a large, expressive class of potential prediction functions, then many rules that look very different have qualitatively similar prediction properties [Black et al. 2022]. Which of these rules is chosen in a given draw of the data then may come down to a flip of a coin.

One of the reasons the orthogonalization method goes wrong in the machine learning context is because it essentially provides a model different information at the deployment stage than what it was trained on. The method asks the algorithm to optimize the prediction when it has access to race or sex, only to restrict this access when applying the prediction function. This may not be a problem when the prediction was based on estimating the model, thereby arguably isolating the effect of race or sex on the prediction. However, when using a machine learning algorithm, the use of race or sex is instrumental in optimizing the prediction accuracy and is not a substantive evaluation of its contribution to the prediction.

In addition to the general instability of feature selection, it is important to highlight that the selection of a protected characteristic in a predictive model does not necessarily indicate a genuine relationship between that characteristic and the outcome being predicted. To illustrate this, consider a simplified example in the context of predicting default probability, denoted as Y, for new borrowers. Imagine a lender who can potentially assess borrowers based on three characteristics: race R, income X_1, and credit score X_2. In the true, but unknown to the lender, model of default risk, suppose the risk is a function of a linear combination of income and credit score: $Y = f(aX_1 + bX_2)$. Assume also that race R has no direct relationship to default probability. However, income X_1 and credit score X_2 might be correlated with race. For instance, certain groups might, on average, have different income levels or credit scores due to a variety of socio-economic factors.

[15] One way to intuitively understand this result is that the positive weight for male borrowers can bring down the disparities where overall female borrowers are predicted to default at higher rates. When this equalizing effect of 'sex' is not longer possible because all applicants receive the same 'sex' value, disparities increase.

In an unconstrained scenario, a predictive model would ideally use X_1 and X_2 to estimate default probability, without needing to consider R. However, the situation becomes more complex when we introduce a constraint, such as the one imposed by lasso regressions, that can help prevent overfitting by penalizing the absolute size of the regression coefficients. Consider a scenario where the tuning parameter in the lasso regression is set in such a way that only one variable can have a non-zero weight in the model. The model might select either X_1 or X_2 for prediction. However, note that R may potentially reflect information on both X_1 and X_2 in a manner that is efficient for the regularized prediction function. Given this possible advantage of the aggregated information reflected in R over each of the other features separately, a prediction function may choose the only feature without a true relationship to the outcome.

This thought experiment highlights two key points. First, the instability in feature selection, particularly under constraints like regularization, can lead to the inclusion of variables that do not have a causal relationship with the outcome. Second, it demonstrates that protected characteristics might be chosen not for their direct relevance but because they inadvertently encapsulate other relevant information in a condensed form. This should caution us against using a method that relies heavily on estimating the true relationship of a protected characteristic with the outcome at the training stage as a way to account for the potential impact of protected characteristics on decisions.

Limitations and Future Work. The simulation exercises in this paper are limited in several ways. The HMDA data contains simulated default rates, limiting their interpretation to reflect real-world default predictions. The German credit data, although it is used frequently in computer science research, is a relatively small dataset (1,000 observations) reported from a single bank. Future work should consider the implications of the orthogonalization approach on additional and potentially larger datasets. Moreover, the exercises in this paper are limited to lasso predictions, that are known for their instability. Future work could consider the implications of the orthogonalization approach in additional model classes.

6 CONCLUDING REMARKS

Applying the orthogonalization method to the machine learning context creates practical and conceptual difficulties. Practically, the variable selection of the lasso is unstable, and even small amounts of noise lead to different variable selection. This could lead to differences in disparities between groups based on small differences in training datasets. Conceptually, a lasso algorithm is not meant to estimate a model, as with an OLS regression, so that it is problematic to interpret the weights of different variables as reflecting some underlying model, as the orthogonalization method does.

The implications of this exercise go beyond the specific example of the orthogonalization approach and suggest deeper skepticism around blindness approaches to fairness and the scrutiny of inputs as a viable path for algorithmic discrimination law. Achieving the underlying goals of discrimination law is likely to require both a race-aware approach at deployment [Meursault et al. 2022] and a direct measurement of decision outcomes [Gillis 2022]. Current law, that risks adopting a myth of algorithmic colorblindness, should

instead confront the practical and theoretical shortcomings of attempting to sterilize inputs from protected characteristics.

ACKNOWLEDGMENTS

This paper benefited from the comments and suggestions of Jann Spiess and James Hicks, and the superb research assistance of Zara Hall.

REFERENCES

2023. Students for Fair Admissions, Inc. v. President and Fellows of Harvard College. 600 U.S. 181. Supreme Court of the United States.

Kristen M Altenburger and Daniel E Ho. 2019. When algorithms import private bias into public enforcement: the promise and limitations of statistical debiasing solutions. *Journal of Institutional and Theoretical Economics* 175, 1 (2019), 98–122.

Joshua D Angrist and Jörn-Steffen Pischke. 2009. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Robert B Avery, Kenneth P Brevoort, and Glenn Canner. 2012. Does Credit Scoring Produce a Disparate Impact? *Real Estate Economics* 40 (2012), S65–S114.

Robert Bartlett, Adair Morse, Nancy Wallace, and Richard Stanton. 2021. Algorithmic discrimination and input accountability under the civil rights acts. *Berkeley Tech. LJ* 36 (2021), 675.

Jason R Bent. 2019. Is algorithmic affirmative action legal. *Geo. LJ* 108 (2019), 803.

Emily Black, Manish Raghavan, and Solon Barocas. 2022. Model multiplicity: Opportunities, concerns, and solutions. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 850–863.

Simon Caton and Christian Haas. 2020. Fairness in machine learning: A survey. *Comput. Surveys* (2020).

Xolani Dastile, Turgay Celik, and Moshe Potsane. 2020. Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing* 91 (2020), 106263.

Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

Stephen Giguere, Blossom Metevier, Yuriy Brun, Bruno Castro da Silva, Philip S Thomas, and Scott Niekum. 2022. Fairness guarantees under demographic shift. In *Proceedings of the 10th International Conference on Learning Representations (ICLR)*.

Talia B Gillis. 2022. The input fallacy. *Minn. L. Rev.* 106 (2022), 1175.

Talia B Gillis and Jann L Spiess. 2019. Big data and discrimination. *The University of Chicago Law Review* 86, 2 (2019), 459–488.

Zach Harned and Hanna Wallach. 2019. Stretching Human Laws to Apply to Machines: The Dangers of a" Colorblind" Computer. *Fla. St. UL Rev.* 47 (2019), 617.

Jongbin Jung, Sam Corbett-Davies, Ravi Shroff, and Sharad Goel. 2018. Omitted and included variable bias in tests for disparate impact. *arXiv preprint arXiv:1809.05651* (2018).

Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems* 33, 1 (2012), 1–33.

Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. 2011. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*. IEEE, 643–650.

Pauline T Kim. 2022. Race-aware algorithms: Fairness, nondiscrimination and affirmative action. *Cal. L. Rev.* 110 (2022), 1539.

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. 2018. Algorithmic fairness. In *Aea papers and proceedings*, Vol. 108. American Economic Association, 22–27.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in neural information processing systems* 30 (2017).

Wei-Yang Lin, Ya-Han Hu, and Chih-Fong Tsai. 2012. Machine Learning in Financial Crisis Prediction: A Survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42, 4 (2012), 421–436. https://doi.org/10.1109/TSMCC.2011.2170420

Zachary Lipton, Julian McAuley, and Alexandra Chouldechova. 2018a. Does mitigating ML's impact disparity require treatment disparity? *Advances in neural information processing systems* 31 (2018).

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. 2018b. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*. PMLR, 3122–3130.

Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. 2021. On the applicability of machine learning fairness notions. *ACM SIGKDD Explorations Newsletter* 23, 1 (2021), 14–23.

Vitaly Meursault, Daniel Moulton, Larry Santucci, and Nathan Schor. 2022. One Threshold Doesn't Fit All: Tailoring Machine Learning Predictions of Consumer Default for Lower-Income Areas. (2022).

Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application* 8 (2021), 141–163.

Sendhil Mullainathan and Jann Spiess. 2017. Machine learning: an applied econometric approach. *Journal of Economic Perspectives* 31, 2 (2017), 87–106.

Alicia H Munnell, Geoffrey MB Tootell, Lynn E Browne, and James McEneaney. 1996. Mortgage lending in Boston: Interpreting HMDA data. *The American Economic Review* (1996), 25–53.

Dana Pessach and Erez Shmueli. 2023. Algorithmic fairness. In *Machine Learning for Data Science Handbook: Data Mining and Knowledge Discovery Handbook*. Springer, 867–886.

Devin G Pope and Justin R Sydnor. 2011. Implementing anti-discrimination policies in statistical profiling models. *American Economic Journal: Economic Policy* 3, 3 (2011), 206–231.

Anya ER Prince and Daniel Schwarcz. 2019. Proxy discrimination in the age of artificial intelligence and big data. *Iowa L. Rev.* 105 (2019), 1257.

Ashkan Rezaei, Anqi Liu, Omid Memarrast, and Brian D Ziebart. 2021. Robust fairness under covariate shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 9419–9427.

Crystal S Yang and Will Dobbie. 2020. Equal protection under algorithms: A new statistical and legal framework. *Mich. L. Rev.* 119 (2020), 291.