Counterfactual Fairness through Transforming Data Orthogonal to Bias

Shuyi Chen
shuyic@alumni.cmu.edu
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Shixiang Zhu
shixianz@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

ABSTRACT

Machine learning models have shown exceptional prowess in solving complex issues across various domains. Nonetheless, these models can sometimes exhibit biased decision-making, leading to disparities in treatment across different groups. Despite the extensive research on fairness, the nuanced effects of multivariate and continuous sensitive variables on decision-making outcomes remain insufficiently studied. We introduce a novel data pre-processing algorithm, *Orthogonal to Bias* (OB), designed to remove the influence of a group of continuous sensitive variables, thereby facilitating counterfactual fairness in machine learning applications. Our approach is grounded in the assumption of a jointly normal distribution within a structural causal model (SCM), proving that counterfactual fairness can be achieved by ensuring the data is uncorrelated with sensitive variables. The OB algorithm is model-agnostic, catering to a wide array of machine learning models and tasks, and includes a sparse variant to enhance numerical stability through regularization. Through empirical evaluation on simulated and real-world datasets—including the adult income and the COMPAS recidivism datasets—our methodology demonstrates its capacity to enable fairer outcomes without compromising accuracy.

KEYWORDS

Counterfactual fairness, Data pre-processing, Algorithmic Fairness

1 INTRODUCTION

In recent years, machine learning has emerged as a pivotal technology, driving advancements across a broad spectrum of real-world applications, from healthcare diagnostics [16], hiring decision-making systems [12], to loan assessments [26]. Its ability to learn from and make predictions or decisions based on large data sets has been transformative in addressing complex problems. However, the broader application prospects of some machine learning techniques is hampered by the implicit biases ingrained in the data it learns from, leading to outcomes that systematically and unfairly disadvantage certain groups [4, 9, 19, 23]. This issue of bias in machine learning models is not merely a technical challenge but a fundamental concern that threatens to reinforce existing social inequalities.

Such fairness concern in machine learning has catalyzed a growing body of research aimed at identifying, understanding, and mitigating biases present in data and algorithms. Among the various conceptual frameworks developed to address this issue [9, 15, 18, 20, 34, 35], the notion of *counterfactual fairness* [24] stands out. Counterfactual fairness seeks to ensure that a decision made by a machine learning model would remain unchanged if a sensitive (or protected) variable of an individual were different, all else being equal. This concept is particularly powerful as it aligns closely with intuitive notions of individual fairness and justice, offering a rigorous standard against which to measure and rectify bias.

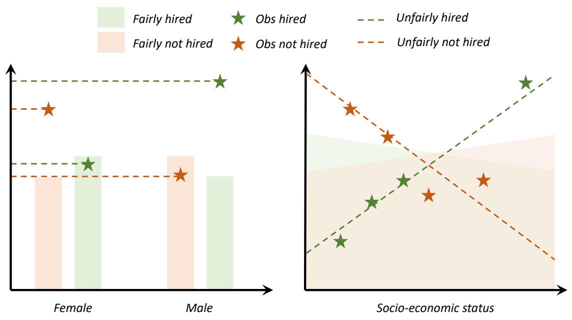


Figure 1: This motivating example illustrates the distinction between binary and continuous sensitive variables in the context of hiring decisions. The dashed lines indicate the predicted hiring decisions and the shaded area indicate the unbiased true decisions. The left panel simplifies the scenario with a binary sensitive variable, such as Gender, where adjustments for fairness are more straightforward due to the clear dichotomy in data. The right panel, however, delves into the complexity introduced by a continuous sensitive variable, like socio-economic status, demonstrating the intricate task of achieving fairness across a spectrum.

Numerous previous research efforts have focused on attaining counterfactual fairness, making significant strides in this area [11, 19, 20, 24, 30]. A key example is the study by [30], which targets non-discrimination through adjusting the predictions based on the empirical joint distribution of the data, aiming to achieve equal bias and accuracy across different groups. However, these approaches typically presuppose a complete understanding of the causal relationships among all variables without unmeasured confounders and consider sensitive variables to be binary (or categorical), so that they can be easily isolated or adjusted. As a result, traditional methods of fair learning face challenges in addressing situations involving multivariate and continuous sensitive variable with intricate inter-dependency.

Take the employment hiring process as an instance shown in Figure 1, socioeconomic status can't be easily classified into a small number of discrete categories. It is a complex, multidimensional variable influenced by education, income, occupation, and even subtler factors like neighborhood and parental education. A candidate's profile is a complex amalgamation of their experiences, skills, educational background, and personal characteristics. A model trained on historical hiring data might inadvertently learn to favor candidates from prestigious universities or with certain types of work experiences—criteria that are often correlated with socioeconomic status. This form of implicit bias emerges either because the data used for training lacks comprehensive representation of all possible candidate profiles or because traditional approaches to fairness do not adequately address the subtle interplay and impact of various sensitive variables on decision-making outcomes.

To tackle these challenges, we develop a novel data pre-processing approach that aims to remove the influence of a group of continuous sensitive variables from the data, thereby ensuring counterfactual fairness in subsequent machine learning tasks. We first prove that the counterfactual fairness can be attainable easily by making the data uncorrelated with the group of sensitive variables, based on the assumption of a jointly normal distribution within a structural causal model (SCM) framework [29]. This assumption is applicable across a broad spectrum of applications where standardization of data is required, ensuring that all variables, both sensitive and non-sensitive, are normalized to have a mean of zero and a variance of 1. Motivated by this understanding, we consider all sensitive variables collectively and propose a data pre-processing algorithm, referred to as *Orthogonal to Bias* (`OB`). This algorithm is designed for minimal data adjustments to achieve orthogonality between non-sensitive and sensitive data. To facilitate numerical stability, we also present a sparse variant of this algorithm which incorporates a regularization term. Then the resulting data is ready to serve as input for machine learning models in downstream tasks without being influenced by the undesirable bias associated with the complexities of sensitive variables. It is also important to note that our proposed algorithm is model-agnostic, making it suitable for a variety of machine learning models and tasks. Lastly, we evaluate our algorithm's performance on a simulated data set and two real-world data sets, including the adult income data and the COMPAS recidivism data, demonstrating that our approach enables machine learning models to achieve fairer outcomes with comparable accuracy to current state-of-the-art fair learning methods. In the numerical results, we found that our approach is not limited by the assumptions about data distribution, indicating its applicability to a wider range of scenarios.

Our contributions in this work can be summarized as follows:

(1) We show that achieving counterfactual fairness is feasible by ensuring orthogonality between non-sensitive and sensitive data when they are jointly normal.
(2) We introduce a model-agnostic data-pre-processing algorithm, termed as *Orthogonal to Bias* (`OB`), which facilitates counterfactual fairness across a broad spectrum of downstream machine learning applications.
(3) We validate the enhanced efficacy of our algorithm compared to the existing state-of-the-arts through evaluations on both synthetic and real-world data sets.

2 LITERATURE REVIEW

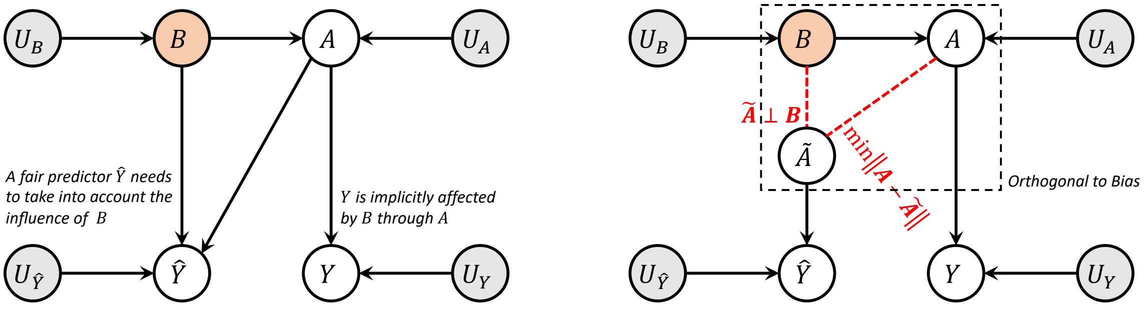
This work is related to several streams of algorithmic fairness literature which we review in this section.

Fairness in Machine Learning. The pursuit of fair decision-making in machine learning has led to diverse approaches for defining and quantifying fairness. Researchers commonly adopt either observational or counterfactual approaches to formalize fairness. Observational methods typically characterize fairness through metrics derived from observed data and predicted outcomes [19, 21, 25, 31]. Metrics such as individual fairness (IF) [15], demographic parity or Group Fairness [22, 35] and equalized odds [20] fall under this category. The key idea for the observational fairness metric is viewing fairness as treating similar individuals or individuals belonging to the same groups similarly. For example, IF defines fairness as treating any two individuals who are similar with respect to a particular task similarly [35].

In contrast, counterfactual approaches proposes a causal approach to defining fairness. These definitions assess fairness based on how predictions would change if sensitive attributes were altered [11, 19, 20, 24, 30]. With the help of the potential outcome concept, the measuring of fairness is no longer restricted to the observable quantities. For instance, the Equal Opportunity (EO) definition, akin to individual fairness, directly compares the actual and counterfactual decisions of the same individual, rather than relying on comparisons between the observation of two similar individuals [30].

While observational definitions of fairness can be incorporated into optimization problems, either by treating the fairness condition as a constraint [15] or directly optimizing the fairness metric as an objective function [35], achieving counterfactual definitions of fairness often require an approximation of the causal model or the counterfactuals since the counterfactuals are unobservable. For example, in the FairLearning algorithm proposed by [24], the unobserved parts of the graphical causal model are sampled through the Markov chain Monte Carlo method. Then they use only the non-descendants of sensitive variable to make the decision. However, this approach may sacrifice a significant amount of information in the data and lead to lower prediction accuracy. Moreover, when estimating the counterfactual distribution is not directly feasible due to sparse or continuous sensitive variables, unexpected relationships between sensitive and non-sensitive attributes may persist. This can either compromise fairness by considering variables related with sensitive variables, or compromise the accuracy by removing all related information regarding sensitive variables from the training data. As we discuss later, our aim is to minimize data changes while ensuring counterfactual fairness under certain assumptions, thereby better addressing the trade-off between accuracy and fairness.

Fair learning approaches. Fairness learning in machine learning aims to prevent discrimination and can be generally categorized into three stages. Firstly, pre-processing approaches [10, 13, 19], which are most closely related to our work, involve modifying the data to eliminate or neutralize any preexisting bias, followed by the application of standard ML techniques. Secondly, in-processing approaches [2, 28] either by a fairness regularizer to the loss function objective, which penalizes discrimination, or imposing a constraint, thereby mitigating disparate treatment. Such methods are normally model-spefiic. The third type of approach is post-processing approaches adjust predictors learned using standard ML techniques after the fact to enhance their fairness properties [7, 14, 17, 20].



(a) The SCM and the typical fair learning

(b) Fair learning by transforming data *orthogonal to bias*

Figure 2: Illustration of (a) the structural causal model (SCM) and a common fair learning strategies, as well as (b) the proposed data pre-processing algorithm *Orthogonal to Bias* (OB). The white nodes A, Y, and $\hat{Y}$ are the non-sensitive variables, the decision variable, and its prediction, respectively. The red nodes B represent the sensitive variable. The $\hat{A}$ is the transformed data that is orthogonal to bias in B. The gray nodes represent exogenous variables.

Our approach is most related to the work by [19], where the authors propose two distribution adjustment procedures, Orthogonalization and Marginal Distribution Mapping, for making counterfactually fair decisions based on adjusted data. While both procedures remove attributes' dependence on sensitive variables under respective conditions, their methods provide no guarantee regarding the scale of modification to the data distribution. In contrast, our work introduces an exact approach to solving an optimization problem that guarantees minimal modification to the data while ensuring counterfactual fairness under specific assumptions. We believe that our emphasis on minimal data modification places our proposed algorithm in a unique position in the widely observed fairness-accuracy spectrum [1, 6, 7]. Through our approach, we aim to capture as much information as possible between the target variable Y and features, including the sensitive features. Additionally, the method proposed by [19] presuppose sensitive variables to be binary (or categorical) so that they can be easily isolated or adjusted based on empirical probability mass function. It does not address situations involving multivariate or continuous sensitive variable with intricate inter-dependency, whereas our proposed OB algorithm aims to remove the influence of a group of continuous sensitive variables from the data, thereby ensuring counterfactual fairness in subsequent machine learning tasks.

3 METHODOLOGY

3.1 Problem setup

We jointly define q non-sensitive variables as $A \in \mathcal{A} \subseteq \mathbb{R}^q$, p sensitive variables as $B \in \mathbb{R}^p$, and decision variable as $Y \in \mathcal{Y}$. The data generation process in our problem setup can be described by a Structural Causal Model (SCM) [29] as shown in Fig. 2 (a). To be specific, we consider the set of endogenous variables $V = \{B, A, Y, \hat{Y}\}$, where $\{B, A, Y\}$ are the observed variables and $\hat{Y}$ is the prediction of Y we made based on B and A. We assume that U_B, U_A, and U_Y, which are the exogenous variables that affect B, A, and Y respectively, are independent of each other. The structural equations are described with the functions $F = \{f_Y, f_A, f_B\}$, one for each component in V, detailed as follows:

$$\begin{aligned} B &= f_B\left(U_B\right), \\ A &= f_A\left(B, U_A\right), \\ Y &= f_Y\left(A, U_Y\right). \end{aligned} \tag{1}$$

According to the above SCM, the bias present in the sensitive variables B can transmit to the predictor $\hat{Y}$ via the non-sensitive variables A. This means that, if there are any differences in the distribution of A conditioning on B, the decision variable $\hat{Y}$ based on A might be unfair.

In this paper, we aim to design a predictor $\hat{Y}$ that achieves the counterfactual fairness [19, 24] without being influenced by the bias in B. Formally, the counterfactual fairness in our SCM can be defined as follows:

Definition 3.1 (Counterfactual Fairness)**.** Given a new pair of attributes $(\mathbf{b}, \mathbf{a})$, a decision variable Y is considered counterfactually fair if, for any $\mathbf{b}' \in \mathcal{B}$,

$$Y_{\mathbf{b}'}(U) | \left\{B = \mathbf{b}^*, A = \mathbf{a}^*\right\} \overset{d}{=} Y_{\mathbf{b}^*}(U) | \left\{B = \mathbf{b}^*, A = \mathbf{a}^*\right\}, \tag{2}$$

where $P \overset{d}{=} Q$ indicates that random variables P and Q are equal in distribution, and $Y_{\mathbf{b}}(U)$ represents the counterfactual outcome of Y when $B = \mathbf{b}$.

The above definition implies that the distribution of the counterfactual result should not depend on the sensitive variables conditional on the observed data. Note that although Definition 3.1 uses the decision variable Y, it also applies to its predictor $\hat{Y}$ without any loss of generality [19].

3.2 Achieving counterfactual fairness via data decorrelation

To clarify and streamline the presentation of our findings, we begin by illustrating that counterfactual fairness can be attained under conditions where sensitive and non-sensitive variables exhibit no correlation and are together normally distributed.

Consider a data set $\mathcal{D} = \{(\mathbf{b}_i, \mathbf{a}_i, y_i)\}_{i=1}^n$ with n observed data tuples, where $\mathbf{b}_i$, $\mathbf{a}_i$, and y_i represent the i-th observation of the sensitive, non-sensitive, and decision variables, respectively. We use $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_n]^\top \in \mathbb{R}^{n \times q}$ to denote the data matrix of non-sensitive variables A, and use $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_n]^\top \in \mathbb{R}^{n \times p}$ to denote the data matrix of sensitive variables B in data set $\mathcal{D}$.

To establish the connection between counterfactual fairness and data uncorrelation, we introduce the following assumption:

Assumption 3.2. Given the structural model defined in (1), the sensitive variable A and non-sensitive variables B are joint normal.

Building on this assumption, we present the following theorem:

Theorem 3.3. *Under Assumption 3.2, $\hat{Y}$ is counterfactually fair when A and B are uncorrelated.*

Proof. We first demonstrate that $\hat{Y}$ is counterfactually fair, which is achieved when the model's predictions for Y are not influenced by the sensitive variable B. In the following proof, we focus on the case of a binary predictor $\hat{Y}$ for simplicity, noting that our findings can be seamlessly applied to predictors that yield continuous outcomes. This simplification allows us to establish fairness by showing that the expected outcomes are equivalent, which for a Bernoulli random variable, also indicates a distributional equivalence. Following the well-established "Abduction-Action-Prediction" method from [29], the conditional expectation of $\hat{Y}_{\mathbf{b}'}$ given $B = \mathbf{b}^*, A = \mathbf{a}^*$ can be written as:

$$\begin{aligned}&\mathbb{E}(\hat{Y}_{\mathbf{b}'} | B = \mathbf{b}^*, A = \mathbf{a}^*)\\ =& \int f_{\hat{Y}}\left(f_A\left(\mathbf{b}', u\right); \mathcal{D}\right) \mathbb{P}_{U_A|B,A}\left(u \mid B = \mathbf{b}^*, A = \mathbf{a}^*\right) du,\end{aligned} \tag{3}$$

where $f_{\hat{Y}}(\cdot; \mathcal{D}) : \mathcal{A} \to \mathcal{Y}$ denotes the predictor of $\hat{Y}$ trained using data $\mathcal{D}$ and $\mathbb{P}_{U_A|B,A}\left(u \mid B = \mathbf{b}^*, A = \mathbf{a}^*\right)$ denotes the conditional density of U_A given $B = \mathbf{b}^*$ and $A = \mathbf{a}^*$. To argue for counterfactual fairness, it suffices to show

$$\mathbb{E}(\hat{Y}_{\mathbf{b}'} | B = \mathbf{b}^*, A = \mathbf{a}^*) = \mathbb{E}(\hat{Y}_{\mathbf{b}^*} | B = \mathbf{b}^*, A = \mathbf{a}^*),$$

if the data generating process for the observed data $f_A\left(\mathbf{b}, u\right)$ does not depend on the value of $\mathbf{b}$, indicating A's independence from B.

Next, we prove that A is independent of B when they are uncorrelated under Assumption 3.2, which is a commonly accepted statistical result. Consider the mean vectors for A and B as $\boldsymbol{\mu}_A$ and $\boldsymbol{\mu}_B$, respectively, with covariance matrices Σ_A, Σ_B. Recall that when A and B uncorrelated, the covariance matrix Σ of A and B is

$$\Sigma = \begin{bmatrix} \Sigma_A & 0 \\ 0 & \Sigma_B \end{bmatrix} \text{ and } \Sigma^{-1} = \begin{bmatrix} \Sigma_A^{-1} & 0 \\ 0 & \Sigma_B^{-1} \end{bmatrix}.$$

Substituting Σ above into the joint probability density function of A and B, we have

$$\begin{aligned}\mathbb{P}(\mathbf{a}, \mathbf{b}) =& \frac{1}{\sqrt{(2\pi)^{q+p}|\Sigma|}} \cdot\\ & \exp\left(-\frac{1}{2}\begin{bmatrix} \mathbf{a} - \boldsymbol{\mu}_A \\ \mathbf{b} - \boldsymbol{\mu}_B \end{bmatrix}^\top \Sigma^{-1} \begin{bmatrix} \mathbf{a} - \boldsymbol{\mu}_A \\ \mathbf{b} - \boldsymbol{\mu}_B \end{bmatrix}\right)\\ =& \frac{1}{\sqrt{(2\pi)^{q+p}|\Sigma_A||\Sigma_B|}} \cdot\\ & \exp\left(-\frac{1}{2}(\mathbf{a} - \boldsymbol{\mu}_A)^\top \Sigma_A^{-1} (\mathbf{a} - \boldsymbol{\mu}_A)\right) \cdot\\ & \exp\left(-\frac{1}{2}(\mathbf{b} - \boldsymbol{\mu}_B)^\top \Sigma_B^{-1} (\mathbf{b} - \boldsymbol{\mu}_B)\right)\\ =& \mathbb{P}_A(\mathbf{a})\mathbb{P}_B(\mathbf{b}).\end{aligned} \tag{4}$$

As we can observe that the joint distribution decomposes into the product of their marginal distributions, hence demonstrating their statistical independence when A and B are jointly normal and uncorrelated.

Finally, establishing A and B as jointly normal and uncorrelated leads to the inference that, given the SCM in (1), predictor $\hat{Y}$ is counterfactually fair.

□

Theorem 3.3 suggests that achieving counterfactual fairness in the predictor $\hat{Y}$ is possible by decorrelating non-sensitive variables A from sensitive variables B. This insight motivates us to develop a data pre-processing algorithm aimed at adjusting the observed data with minimal changes to achieve uncorrelation between non-sensitive and sensitive variables.

It is important to emphasize that Assumption 3.2 is applicable across a broad spectrum of applications where standardization of data is required, ensuring that all variables, both sensitive and non-sensitive, are normalized to have a mean of zero and a variance of 1. Moreover, our investigations have revealed that this assumption is not always critical for the success of our data pre-processing algorithm. Our approach has demonstrated effectiveness even when applied to data sets that do not meet this criterion. In particular, as elaborated in Section 4, our strategy has proven to be promising in experiments that involve categorical sensitive variables B, rather than continuous ones.

3.3 Orthogonal to bias

In this section, we develop a data pre-processing algorithm, termed as *Orthogonal to Bias* (OB). We first standardize both non-sensitive variables A and sensitive variables B, to achieve a normal distribution for each. Then the empirical covariance between A and B can be estimated by

$$\operatorname{cov}(A, B) = \mathbb{E}[(A - \mathbb{E}[A])(B - \mathbb{E}[B])] \approx \frac{1}{n}\sum_{i=1}^{n} \mathbf{a}_i^\top \mathbf{b}_i = \langle \mathbf{A}, \mathbf{B} \rangle .$$

Given that two variables are uncorrelated if their covariance is zero, orthogonality between observed data $\mathbf{A}$ and $\mathbf{B}$ guarantees uncorrelation. Therefore, the OB algorithm aims to adjust the observed non-sensitive data $\mathbf{A}$ in such a way that it is orthogonal to the observed sensitive data **B**, while ensuring minimal changes to non-sensitive data **A**.

Specifically, we follow the idea of Orthogonal to Groups introduced by [3], and define a rank k approximation of **A** as $\widetilde{\mathbf{A}} = \mathbf{S}\mathbf{U}^\top$, where $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_k]$ is a $q \times k$ orthonormal matrix and $\mathbf{S} = \{s_{ij}\}$ is a $n \times k$ matrix. The goal is to find a transformed $n \times q$ matrix $\widetilde{\mathbf{A}}$ that is orthogonal to **B** with minimal change to matrix, as measured by the Frobenius norm $\|X\|_\mathrm{F} = \sqrt{\sum_i \sum_j x_{ij}^2}$. Formally, we aim to solve the following constrained optimization problem:

$$\begin{aligned} &\arg\min_{\mathbf{S},\mathbf{U}} \left\|\mathbf{A} - \mathbf{S}\mathbf{U}^\top\right\|_F^2, \\ &\text{s.t. } \left\langle \mathbf{S}\mathbf{U}^\top, \mathbf{B} \right\rangle = \mathbf{0}, \\ &\mathbf{U}^\top\mathbf{U} = \mathbf{I}_k, \end{aligned} \tag{5}$$

where the last constraint requires **U** to be orthonormal matrix, which helps prevent degeneracy, such as basis vectors becoming identically zero or encountering solutions with double multiplicity.

As shown by [3], the original problem (5) can be reformulated using Lagrange multipliers as follows:

$$\arg\min_{\mathbf{S},\mathbf{U}} \frac{1}{n} \sum_{i=1}^{n} \left\| \mathbf{a}_i - \sum_{j=1}^{k} s_{ij}\mathbf{u}_j^\top \right\|^2 + \frac{2}{n} \sum_{j=1}^{k} \lambda_j \sum_{i=1}^{n} s_{ij} B_i. \tag{6}$$

The factor $2/n$ is introduced to simplify the expression for the optimal solutions. It ensures that the first-order condition for (6) with respect to s_{ij} involves a common factor of $2/n$, which can then be canceled out during computations. Let $\mathbf{S}^*$ and $\mathbf{U}^*$ denote the optimal solutions of **S** and **U**, respectively. As demonstrated in Appendix A, the reformulated equation (6) yields a closed-form solution:

$$\begin{aligned} \lambda_j^* &= \frac{\left\langle \mathbf{A}\mathbf{u}_j^{*\top}, \mathbf{B} \right\rangle}{\langle \mathbf{B}, \mathbf{B} \rangle} \\ \mathbf{U}^* &= \left[\mathbf{u}_1^*, \ldots, \mathbf{u}_k^*\right] \\ \mathbf{S}^* &= \{s_{ij}^*\}, \end{aligned} \tag{7}$$

where $\mathbf{U}^*$ consists of the first k right singular vectors of **A** and $s_{ij}^* = \mathbf{a}_i\mathbf{u}_j^{*\top} - \lambda_j^* B_i$. Detailed derivations of the closed-form solution (7) can be found in Appendix A. The processed non-sensitive data matrix is therefore $\widetilde{\mathbf{A}} = \mathbf{S}^*\mathbf{U}^{*\top}$.

It is noteworthy that when the correlation between the sensitive variable and the observed data is minimal, $\widetilde{\mathbf{A}}$ yields a reconstruction error similar to that of the standard Singular Value Decomposition (SVD). It is because the additional reconstruction error of $\widetilde{\mathbf{A}}$ relative to SVD with the same rank is proportional to the collinearity between the subspace spanned by **B** and the left singular vectors of the data **A** [3]. The exact expression for OB's additional reconstruction error compared to SVD can be found in Appendix A.

*Sparse Orthogonal to Bias (*SOB*).* When the number of features p exceeds the number of observations n, estimating a low-dimensional structure from high-dimensional data can become numerically unstable [36]. To address this challenge, we introduce a sparse variant of the OB algorithm, referred to as SOB. The SOB imposes an ℓ_1-norm penalty for **U** to encourage sparsity and improve numerical stability, in addition to the orthogonality constraints in (5). We define h as the ℓ_1 constraint on u. Details of the formulation of SOB can be found in Appendix B. Following derivation by [3, 33], Algorithm 1 illustrates the key steps to implement the SOB algorithm, where η represents the minimum change to terminate the iterative optimization process.

Algorithm 1 Sparse Orthogonal to Bias (SOB)

```
Input: Non-sensitive and sensitive data $\mathbf{A}$ and $\mathbf{B}$, rank $k$;
Standardize $\mathbf{A}$ and $\mathbf{B}$;
for $i = 1, \ldots, k$ do
    Set $t = 1$, $\theta = 1$, and $s_i^{(0)} = 0$;
    Randomly initialize $u_i^{(0)}$;
    while $\left\|u_i^{(t)} - u_i^{(t-1)}\right\|_F > \eta$ and $\left\|s_i^{(t)} - s_i^{(t-1)}\right\|_F > \eta$ do
        Compute $\beta_i^{(t)}$ with
            $\beta_i \leftarrow \left(\mathbf{B}^\top\mathbf{B}\right)^{-1} \mathbf{B}^\top P_{i-1}\mathbf{A}u_i$
        with $P_{i-1} = I_{n\times n} - \sum_{l=1}^{i-1} s_l s_l^\top$
        Update $s_i$ as
            $s_i^{(t)} \leftarrow \frac{P_{i-1}\mathbf{A}u_i - \beta_i\mathbf{B}}{\|P_{i-1}\mathbf{A}u_i - \beta_i\mathbf{B}\|_2}$
        Update $u_i$ as
            $u_i^{(t)} \leftarrow \frac{\mathcal{S}_\theta\left(\mathbf{A}^\top s_i\right)}{\|\mathcal{S}_\theta\left(\mathbf{A}^\top s_i\right)\|_2}$
        where $\mathcal{S}_\theta(x) = \mathrm{sign}(x)(|x| - \theta)\mathbb{1}(|x| \geq \theta)$
        $t \leftarrow t + 1$
    end while
    if $\left\|\mathbf{A}^\top s_i\right\|_1 \leq h$ then
        Set $\theta = 0$
    else
        Set $\theta > 0$ such that $\left\|u_i^{(t)}\right\|_1 = h$
    end if
end for
Set $\hat{\mathbf{S}} = [d_1 s_1, \ldots, d_k s_k]$ where $d_i = s_i^\top \mathbf{A}u_i$, and $\hat{\mathbf{U}} = [u_1, \ldots u_k]$.
Finally calculate the attribute matrix $\widetilde{\mathbf{A}} = \hat{\mathbf{S}}\hat{\mathbf{U}}^\top$.
Output: $\widetilde{\mathbf{A}}$
```

Note that with the additional regularization constraints in this case, the solution favors sparsity while satisfying the orthogonal constraint. Therefore, SOB also achieves at counterfactual fairness under SCM framework with Theorem 3.3.

4 EXPERIMENTS

In this section, we present the results of our experiments conducted on both synthetic and real data sets. In Section 4.2 and 4.3, we compare the proposed OB with several existing methods: Machine Learning (ML), a logistic regression baseline which uses all attributes whether sensitive or not; Fairness through Unawareness (FTU), which simply fits a logistic model with **A**, excluding the sensitive variables from the model; Equalized Odds (EO), a post-processing algorithm chosen to balance false positives and false negatives while minimizing the expected loss proposed by [30]; FairLearning Algorithm (FL), which achieves counterfactual fairness by sampling unobserved parts of the graphical causal model using Markov chain Monte Carlo methods [24]; Affirmative Action

(AA), an post processing algorithm that produces fair equalized odds (EO) and affirmative action (AA) predictors by positing a causal model and considering counterfactual decisions [30]; Fair Learning through Data pre-processing (FLAP) by [19].

Among the compared methods, FTU and FLAP are pre-processing methods, FL is an in-processing approach, and AA and EO are post-processing approaches. Note that for both FLAP and OB, the predictor class used for prediction can include sensitive variables. Specifically, in addition to the predictor class $f_{\hat{Y}}(A) : \mathcal{A} \rightarrow \mathcal{Y}$ discussed in Section 3.2, for a machine learning predictor $f_{\hat{Y}}(A, B) : \mathcal{A} \times \mathcal{B} \rightarrow \mathcal{Y}$ that utilizes both sensitive and non-sensitive variables, an Averaged Machine Learning (AML) predictor $f^{'}_{\hat{Y}}(A) = \int f_{\hat{Y}}(A, B)\mathbb{P}(B)dB$ or $f^{'}_{\hat{Y}}(A) = \sum f_{\hat{Y}}(A, B)\mathbb{P}(B)$, depending on whether B is continuous or binary, can be constructed respectively. Therefore, we denote $\mathtt{OB}_1$ and $\mathtt{OB}_2$ for scenarios involving the training of $f_{\hat{Y}}(A, B)$ or $f_{\hat{Y}}$ with OB-processed data, respectively. Similar designations are used for $\mathtt{FLAP}_1$ and $\mathtt{FLAP}_2$. Additionally, since [19] introduces two pre-processing methods, Orthogonalization and Marginal Distribution Mapping, we denote them as FLAP(O) and FLAP(M) respectively. Logistic regression predictors are utilized for all models, denoted as $\hat{Y}$. The experiments were conducted in a Jupyter Notebook environment with 16 GB RAM.

4.1 Evaluation metrics

We assess the accuracy of the decisions concerning the ground truth with Area Under the Curve (AUC) and Accuracy (ACC). As for the evaluation of counterfactual fairness, we adopt two metrics introduced by [19]: CF-metrics measures counterfactual fairness by calculating the average change in predicted scores between groups with the most significant difference. Additionally, we incorporate CF Bound to evaluate counterfactual fairness in the absence of non-sensitive conditions. This metric computes the maximum absolute value of the bounds' average of predicted scores for a sample randomly selected from the set. Due to computational challenges, we only evaluate CF-Bound for the simulated data set. For a comprehensive comparison of the methods, we additionally incorporate two observational fairness metrics: EO Fairness, as defined by [30], and AA Fairness, proposed in [30]. In Tables 1 to 3, we highlight the best-performing method in bold and underline the second-best for each metric used.

4.2 Synthetic data

We first apply our methods to a synthetic loan data set example which is a modification from [19]. Using synthetic data allows us to repeat the random data generation process and provide the average results. It also enables us to observe how fair learning models respond to changing effects resulting from different levels of unfair treatment among different groups. The presented example illustrates a scenario in which a bank evaluates loan applications based on the applicant's education level (E) and annual income (I), determining approval ($Y = 1$) or rejection ($Y = 0$). The population comprises three possible race groups: $B = \{0, 1, 2\}$. Similar to [19], we generate B according to $B = \mathbb{1}\{U_B < 0.76\} + \mathbb{1}\{U_B > 0.92\}$, where $U_B \sim \text{Uniform}(0, 1)$. Let U_E and U_I be two standard normal random variables with mean $\mu_E = \lambda_{E0} + \mathbb{1}\{B = 1\}\lambda_{E1} + \mathbb{1}\{B = 2\}\lambda_{E2}$ and $\mu_I = \log(\lambda_{A0} + \mathbb{1}\{B = 1\}\lambda_{A1} + \mathbb{1}\{B = 2\}\lambda_{A2})$, respectively. Then the education year and annual income for each race group follows the following distribution:

$$E = \max\{0, U_E\},$$
$$I = \exp\{0.1U_E + U_I\}. \tag{8}$$

The bank's decision is simulated using a logistic model:

$$Y = \mathbb{1}\{U_Y < \text{expit}(\beta_0 + \beta_1\mathbb{1}\{B = 1\} + \beta_2\mathbb{1}\{B = 2\} + \beta_E E + \beta_I I)\}, \tag{9}$$

where $U_Y \sim \text{Uniform}(0, 1)$ and $\text{expit}(u) = (1 + e^{-u})^{-1}$.

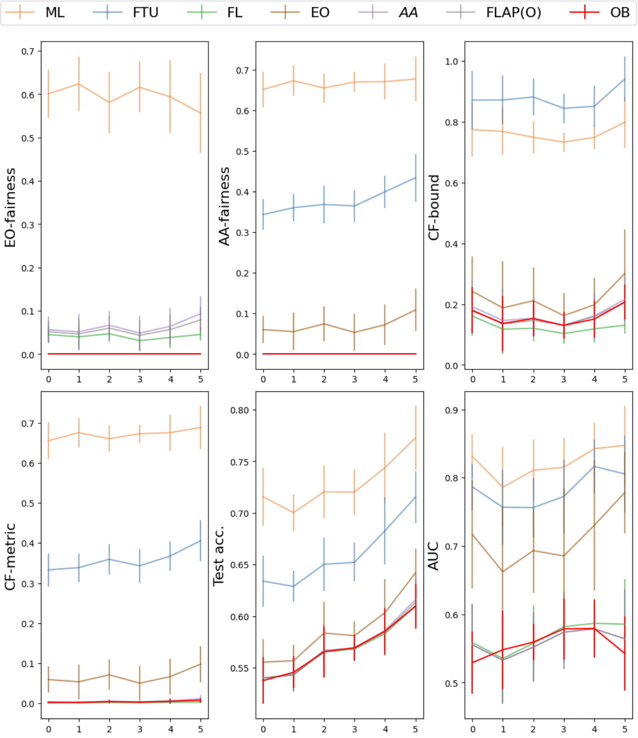


Figure 3: Synthetic loan data. The x-axis shows different levels of the effect of the sensitive variable on education level, β_E, as defined in (9). With 10 repeated experiments, the lines represent the averaged results for each method, and the bars represent their standard deviation. The red line represents the results of the proposed method OB.

In this example, the parameters λ_{E1} and λ_{E2} determine the extent of the mean difference in education years across the three race groups, while the parameters λ_{I1} and λ_{I2} dictate the magnitude of the mean difference in log income among these three race groups. β_1 and β_2 characterize the direct effect of the race information on the loan approval rate.

It is important to note that the sensitive variable B is categorical, and the data generating process does not exactly conform to Assumption 3.2. As evidenced in Table 1 and Figure 3, despite the deviation from the assumptions in the tested synthetic data set, our method consistently showcases comparatively high AUC and ACC compared to most methods. Notably, its accuracy outperforms FL and FLAP, two other counterfactually fair methods. Moreover, our method achieves low CF-metric and CF Bound, akin to FL and

Table 1: Performance comparison of our method and other existing methods with synthetic loan data

Metrics	Baselines		Compared Methods							Ours	
	ML	FTU	FL	EO	AA	FLAP$_1$(O)	FLAP$_2$(O)	FLAP$_1$(M)	FLAP$_2$(M)	OB$_1$	OB$_2$
ACC	0.6618	0.6481	0.6224	0.6237	0.6224	0.6237	0.6224	0.6237	0.6224	**0.6406**	0.6279
AUC	0.9457	0.8986	0.5867	**0.6682**	0.5714	0.5668	0.5837	0.5875	0.5863	0.5704	0.5856
CF-metrics	0.6291	0.3906	0.0031	0.0355	0.0034	0.0016	0.0032	**0.0002**	**0.0002**	0.0011	0.0026
CF Bound	0.8690	0.9464	0.1836	0.1071	0.0918	0.0937	0.1847	0.0690	**0.0670**	0.0830	0.2340
EO Fairness	0.5469	0	0.0156	**0**	0.0336	0.0321	0.0156	0.0301	0.0180	**0**	**0**
AA Fairness	0.6235	0.4559	**5.6e-18**	0.0370	**1.1e-18**	**3.3e-18**	**6.7e-18**	0.0012	0.0038	**4.6e-17**	**4.3e-17**

Table 2: Performance comparison of our method and other existing methods on Adult data

Metrics	Baselines		Compared Methods							Ours	
	ML	FTU	FL	EO	AA	FLAP$_1$(O)	FLAP$_2$(O)	FLAP$_1$(M)	FLAP$_2$(M)	SOB$_1$	SOB$_2$
ACC	0.7612	0.7604	0.7594	**0.7680**	0.7644	0.7357	0.7151	0.7548	0.7594	0.7655	0.7597
AUC	0.8128	0.8036	0.7680	**0.7991**	0.7682	0.7682	0.7680	0.7651	0.7649	0.7806	0.7809
CF-metric	0.2779	0.2338	**0.0228**	0.2047	0.0268	0.0280	**0.0228**	0.0280	**0.0228**	0.0529	0.0600
EO Fairness	0.1536	0	0.2853	**0**	0.2811	0.2780	0.2853	0.2780	0.2853	0.0002	0.0005
AA Fairness	0.3034	0.2574	**0**	0.2259	**0**	2.2e-17	2.2e-17	2.8e-17	2.8e-17	0.0001	0.0004

FLAP, indicating a high degree of counterfactual fairness. This desirable characteristic can be attributed to the property of OB, which minimally modifies in the non-sensitive data while ensuring counterfactual fairness. Furthermore, in terms of observational fairness metrics, OB exhibits an overall better performance compared to FL and FLAP with lower EO and AA Fairness metrics.

4.3 Real data

We also apply our methods to two real data sets: the Adult Income data set from the UCI Machine Learning Repository[1] and the COMPAS recidivism data from ProPublica [5].

Adult. In the Adult Income data set, we aim to predict whether an individual's income exceeds $50,000, considering features such as sex, race, age, work class, education, occupation, marital status, capital gain, and loss. The sensitive attributes are sex and race. The training set consists of 32,561 samples, and the test set comprises 16,281 samples.

Due to the large sample size of Adult data set, we employ SOB. As illustrated in Table 2, similar to OB's performance in synthetic data, the accuracy is comparatively high compared to all other tested fair learning approaches. Notably, the accuracy is even higher than the vanilla ML model, which utilizes both sensitive and non-sensitive attributes and generates unfair results. As noted by [3, 8], the additional regulation with SOB may contribute to high out-of-sample prediction performances. Moreover, its CF-metric is comparable to that of FLAP and FL and is much lower than baselines, implying counterfactual fairness attributed to OB. Additionally, it achieves both low EO and AA Fairness metrics.

COMPAS. The COMPAS recidivism data includes demographic information such as sex, age, race, and record data (prior counts, juvenile felonies counts, and juvenile misdemeanors counts) for over 10,000 criminal defendants in Broward County, Florida. The goal is to predict whether they will re-offend in the next two years.

As depicted in Table 3, similar to our method's performance in the two previous data sets, the accuracy with OB is comparatively high. Moreover, its CF-metric is similar compared to that of FLAP and FL, implying counterfactual fairness attributed to OB. Additionally, it achieves both relatively low EO and AA Fairness metrics compared to FLAP and FL, suggesting better observational fairness.

In summary, across three datasets, our approach consistently exhibits its effectiveness in maintaining an overall better balance between accuracy, observational fairness, and counterfactual fairness. Additionally, we observed that using SOB for the synthetic loan data and COMPAS results in slightly lower ACC and AUC, while achieving similar fairness metrics.

4.4 Case Study with Continuous Decision Variables

To showcase the empirical efficacy of OB in decorrelating **A** and **B** and achieving counterfactually fair predictions, we present an additional case study of two synthetic data sets featuring continuous decision variable Y. We apply FTU, FL, and OB to the two simulated data sets. We evaluate the accuracy of $\hat{Y}$ generated by different methods using Root Mean Square Error (RMSE) compared to ground truth Y. We examine the counterfactual fairness by examining the KL-divergence between predictions using the observed (actual) data and counterfactual data with a different sensitive variable following

[1] https://archive.ics.uci.edu/dataset/2/adult

Table 3: Performance comparison of our method and other existing methods on COMPAS data

Metrics	Baselines		Compared Methods							Ours	
	ML	FTU	FL	EO	AA	FLAP$_1$(O)	FLAP$_2$(O)	FLAP$_1$(M)	FLAP$_2$(M)	OB$_1$	OB$_2$
ACC	0.5744	0.5726	0.5598	**0.5710**	0.5609	0.5605	0.5599	0.5607	0.5607	0.5666	0.5674
AUC	0.7206	0.7225	0.6928	0.7225	0.6927	0.6927	0.6928	0.7015	0.7019	**0.6764**	**0.6744**
CF-metric	0.2274	0.1406	0.0054	0.1377	0.0060	0.0058	0.0054	**0.0026**	0.0027	0.0060	0.0065
EO Fairness	0.1046	0	0.1374	**0**	0.1405	1.7e-06	3.3e-06	6.7e-07	1.2e-06	**0**	**0**
AA Fairness	0.2258	0.1460	**0**	0.1424	**0**	2.9e-07	5.6e-07	8.2e-07	3.0e-07	1.6e-16	1.1e-16

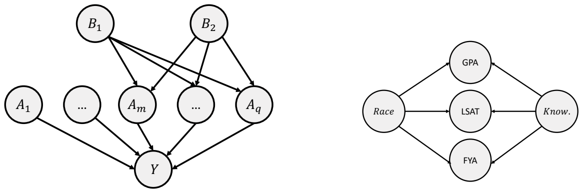
Table 4: Performance comparison on synthetic data sets with continuous decision variable

Metrics	Syn (Cont. Y)			LSAT		
	FTU	FL	OB	FTU	FL	OB
RMSE	0.50	0.51	0.53	0.86	0.89	0.18
KL Div.	0.35	0.00	0.00	1.42	0.28	0.01
$Corr(\widetilde{\mathbf{A}}, \mathbf{B})^*$	0.52	0.52	0.00	0.30	0.30	0.00

* The average pairwise correlation between $\widetilde{\mathbf{A}}$ and $\mathbf{B}$.

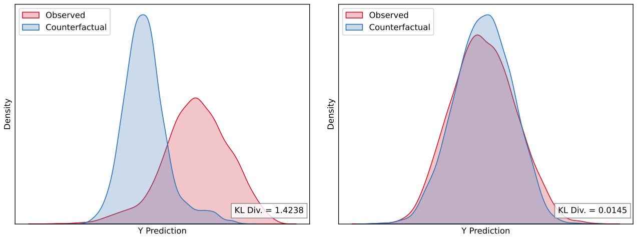
[30]. We additionally include the average pair-wise correlations between $\widetilde{\mathbf{A}}$ and $\mathbf{B}$ to verify their uncorrelation.

While the detailed setup of the two synthetic cases is provided in Appendix C, we present a brief overview here. Causal models for the two synthetic cases with continuous decision variables are outlined in Figure 4. The first synthetic dataset, Syn (Cont. Y), features a continuous Y following a normal distribution with its mean linearly dependent on non-sensitive variables A, which in turn are affected by two categorical sensitive variables. The other synthetic dataset, LSAT, is derived from a survey conducted by the Law School Admission Council across 163 law schools in the United States [32]. Here, the decision variable Y represents the first-year average grade (FYA), while the sensitive variable is the students' race. The results summarized in Table 4 indicate that OB effectively decorrelates **A** and **B** in both cases. It achieves a similar Root Mean Square Error (RMSE) in synthetic data and lower RMSE in the LSAT dataset compared to FTU and FL. Additionally, the predicted distributions of FYA under observed and counterfactual sensitive variables for the LSAT case are nearly identical for OB, as evidenced by the KL-divergence measure in Table 4 and highly overlapped curves in Figure 5. Thus, the predictions of OB align with the definition of counterfactual fairness outlined in Definition 3.1.



(a) Synthetic Data **(b) LSAT**

Figure 4: (a) represents the causal model for a synthetic data set used in Section 4.4, with B being the binary sensitive variable and Y being a continuous outcome of interest, along with the intermediate variable A that is influenced by B. (b) represents the causal model for LSAT, a semi-synthetic example used by [24]. We extract a latent variable, student's knowledge (K), and assume such a variable affects GPA, LSAT, and FYA scores to apply the FL method in this case.



(a) FTU **(b) OB**

Figure 5: The red line represents the $\hat{Y}$ distribution utilizing the observed variables, while the blue line represents the prediction distribution utilizing the counterfactual data. (a) shows the distributions of $\hat{Y}$ using FTU method, while (b) presents the distributions of $\hat{Y}$ using OB processed data.

5 CONCLUSION

In conclusion, this paper demonstrates that achieving counterfactual fairness is feasible by ensuring the uncorrelation between non-sensitive and sensitive variables under certain conditions. Building on this insight, we present the Orthogonal to Bias (OB) algorithm, a novel approach to addressing fairness challenges in machine learning models. OB achieves counterfactual fairness by decorrelating data from sensitive variables under mild conditions. The resulting data pre-processing algorithm effectively removes bias in predictions while making minimal changes to the original data. Importantly, OB is model-agnostic, ensuring its adaptability to a variety of machine learning models. Through comprehensive evaluations on simulated and real-world data sets, we demonstrate that OB strikes a great balance between fairness and accuracy, outperforming compared methods and offering a promising solution to the complex issue of bias in machine learning.

REFERENCES

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. In *International conference on machine learning*. PMLR, 60–69.

[2] Sina Aghaei, Mohammad Javad Azizi, and Phebe Vayanos. 2019. Learning Optimal and Fair Decision Trees for Non-Discriminative Decision-Making. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence* (Honolulu, Hawaii, USA) *(AAAI'19/IAAI'19/EAAI'19)*. AAAI Press, Article 175, 9 pages. https://doi.org/10.1609/aaai.v33i01.33011418

[3] Emanuele Aliverti, Kristian Lum, James E Johndrow, and David B Dunson. 2021. Removing the influence of group variables in high-dimensional predictive modelling. *Journal of the Royal Statistical Society. Series A,(Statistics in Society)* 184, 3 (2021), 791.

[4] McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. 2021. What We Can't Measure, We Can't Understand: Challenges to Demographic Data Procurement in the Pursuit of Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 249–260. https://doi.org/10.1145/3442188.3445888

[5] Julia Angwin and Jeff Larson. 2023. Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[6] Sina Baharlouei, Maher Nouiehed, Ahmad Beirami, and Meisam Razaviyayn. 2019. R\’enyi Fair Inference. *arXiv preprint arXiv:1906.12005* (2019).

[7] Richard A. Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. A Convex Framework for Fair Regression. *ArXiv* abs/1706.02409 (2017). https://api.semanticscholar.org/CorpusID:12641090

[8] Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.

[9] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems* 29 (2016).

[10] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building classifiers with independency constraints. In *2009 IEEE international conference on data mining workshops*. IEEE, 13–18.

[11] Silvia Chiappa. 2019. Path-specific counterfactual fairness. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 7801–7808.

[12] Lee Cohen, Zachary C. Lipton, and Yishay Mansour. 2020. Efficient Candidate Screening Under Multiple Tests and Implications for Fairness. In *1st Symposium on Foundations of Responsible Computing, FORC 2020, June 1-3, 2020, Harvard University, Cambridge, MA, USA (virtual conference) (LIPIcs, Vol. 156)*, Aaron Roth (Ed.). Schloss Dagstuhl - Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 1:1–1:20. https://doi.org/10.4230/LIPIcs.FORC.2020.1

[13] Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. 2019. Flexibly fair representation learning by disentanglement. In *International conference on machine learning*. PMLR, 1436–1445.

[14] Brian d'Alessandro, Cathy O'Neil, and Tom LaGatta. 2017. Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big data* 5, 2 (2017), 120–134.

[15] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (Cambridge, Massachusetts) *(ITCS '12)*. Association for Computing Machinery, New York, NY, USA, 214–226. https://doi.org/10.1145/2090236.2090255

[16] Qizhang Feng, Mengnan Du, Na Zou, and Xia Hu. 2022. Fair machine learning in healthcare: A review. *arXiv preprint arXiv:2206.14397* (2022).

[17] Benjamin Fish, Jeremy Kun, and Ádám D Lelkes. 2016. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM international conference on data mining*. SIAM, SIAM, Miami, Florida, USA, 144–152.

[18] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. 2016. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS symposium on machine learning and the law*, Vol. 1. Barcelona, Spain, Curran Associates, Inc., Barcelona, Spain, 11.

[19] Rui Song Haoyu Chen, Wenbin Lu and Pulak Ghosh. 2023. On Learning and Testing of Counterfactual Fairness through Data Preprocessing. *J. Amer. Statist. Assoc.* 0, 0 (2023), 1–11. https://doi.org/10.1080/01621459.2023.2186885 arXiv:https://doi.org/10.1080/01621459.2023.2186885

[20] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (Barcelona, Spain) *(NIPS'16)*. Curran Associates Inc., Red Hook, NY, USA, 3323–3331.

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

[22] Aria Khademi, Sanghack Lee, David Foley, and Vasant Honavar. 2019. Fairness in algorithmic decision making: An excursion through the lens of causality. In *The World Wide Web Conference*. Association for Computing Machinery, New York; NY; United States, 2907–2914.

[23] Niki Kilbertus, Manuel Gomez Rodriguez, Bernhard Schölkopf, Krikamol Muandet, and Isabel Valera. 2020. Fair Decisions Despite Imperfect Predictions. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 108)*, Silvia Chiappa and Roberto Calandra (Eds.). PMLR, 277–287. https://proceedings.mlr.press/v108/kilbertus20a.html

[24] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc., Long Beach, CA, USA. https://proceedings.neurips.cc/paper_files/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf

[25] Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. 2011. k-NN as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. 502–510.

[26] Amitabha Mukerjee, Rita Biswas, Kalyanmoy Deb, and Amrit P Mathur. 2002. Multi–objective evolutionary algorithms for the risk–return trade–off in bank loan management. *International Transactions in operational research* 9, 5 (2002), 583–597.

[27] Carl M. O'Brien. 2016. Statistical Learning with Sparsity: The Lasso and Generalizations. *International Statistical Review* 84, 1 (2016), 156–157. https://doi.org/10.1111/insr.12167 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/insr.12167

[28] Pranita Patil and Kevin Purcell. 2022. Decorrelation-Based Deep Learning for Bias Mitigation. *Future Internet* 14, 4 (2022). https://doi.org/10.3390/fi14040110

[29] Judea Pearl. 2009. *Causality* (2 ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511803161

[30] Yixin Wang, Dhanya Sridhar, and David M Blei. 2019. Equal opportunity and affirmative action via counterfactual predictions. *arXiv preprint arXiv:1905.10870* (2019).

[31] Song Wei, Xiangrui Kong, Alinson Santos Xavier, Shixiang Zhu, Yao Xie, and Feng Qiu. 2024. Assessing Electricity Service Unfairness with Transfer Counterfactual Learning. *arXiv preprint arXiv:2310.03258* (2024).

[32] Linda F. Wightman. 1998. LSAC National Longitudinal Bar Passage Study. LSAC Research Report Series. https://api.semanticscholar.org/CorpusID:151073942

[33] Daniela M Witten, Robert Tibshirani, and Trevor Hastie. 2009. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* 10, 3 (2009), 515–534.

[34] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 54)*, Aarti Singh and Jerry Zhu (Eds.). PMLR, Sydney, Australia, 962–970. https://proceedings.mlr.press/v54/zafar17a.html

[35] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. In *Proceedings of the 30th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 28)*, Sanjoy Dasgupta and David McAllester (Eds.). PMLR, Atlanta, Georgia, USA, 325–333. https://proceedings.mlr.press/v28/zemel13.html

[36] Hui Zou, Trevor Hastie, and Robert Tibshirani. 2006. Sparse principal component analysis. *Journal of computational and graphical statistics* 15, 2 (2006), 265–286.

A CLOSED-FORM OB SOLUTION DERIVATION

We start by considering (6) when $k = 1$. The OB algorithm aims to find the closest rank-1 matrix (vector) approximation to the original set of data that satisfies the orthogonal condition. (6) can be reformulated as:

$$\arg\min_{S,U}\left\{\frac{1}{n}\sum_{i=1}^{n}\left\|a_i - \sum_{j=1}^{k} s_{ij}u_j^T\right\|^2 + \frac{2}{n}\lambda_1\sum_{i=1}^{n} s_{i1}b_i\right\}. \tag{10}$$

Some algebra and the orthonormal condition on u_1 allow us to express the loss function to be minimized as:

$$\begin{aligned} L(s_1, u_1) &= \frac{1}{n}\sum_{i=1}^{n}(a_i - s_{i1}u_1^T)^T(a_i - s_{i1}u_1^T) + \frac{2}{n}\lambda_1\sum_{i=1}^{n} s_{i1}b_i \\ &= \frac{1}{n}\sum_{i=1}^{n}(a_i^T a_i - 2s_{i1}a_i u_1^T + s_{i1}^2) + \frac{2}{n}\lambda_1\sum_{i=1}^{n} s_{i1}b_i. \end{aligned}$$

The function is quadratic, and its partial derivative with respect to s_{i1} is

$$\frac{\partial}{\partial s_{i1}}L(s_1, u_1) = \frac{1}{n}(-2a_i u_1^T + 2s_{i1}) + \frac{2}{n}\lambda_1 b_i.$$

Solving it finds a stationary point of

$$s_{i1} = a_i u_1^T - \lambda_1 b_i. \tag{11}$$

So the optimal score for the i-th subject is obtained by projecting the observed data onto the first basis and then subtracting $\lambda_1 b$. The constraint does not involve the orthonormal basis u_1, hence the solution of (10) for u_1 is equivalent to the unconstrained scenario. A standard result of linear algebra states that the optimal u_1 for (10) without constraints equivalent to the first right singular vector of A, or equivalently to the first eigenvector of the matrix $A^T A$ [33]. Plugging in the solution for u_1 and setting the derivative with respect to λ_1 equal to 0 leads to

$$\sum_{i=1}^{n}(a_i u_1^T - \lambda_1 a_i)^T b_i = 0. \tag{12}$$

Therefore,

$$\lambda_1 = \frac{\sum_{i=1}^{n} a_i u_1^T b_i}{\sum_{i=1}^{n} b_i^2} = \frac{\langle A u_1^T, b\rangle}{\langle b, b\rangle}, \tag{13}$$

which states λ is is a least squares estimate of Au_1^T over b.

Now consider the more general case when $k > 1$. The derivatives with respect to the generic element s_{ij} can be calculated easily due to the constraint on U, which simplifies the computation. The optimal solution for the generic score s_{ij} is given by

$$s_{ij} = a_i u_j^T - \lambda_j b_i, \tag{14}$$

since $u_i^T u_j = 0$ for all $i \neq j$ and $u_j^T u_j = 1$ for $j = 1, ..., k$.

The global solution for $\lambda = (\lambda_1, ...\lambda_k)$ can be derived from least squares projection since we can interpret (14) as a multivariate linear regression where the k columns of the projected matrix AU^T are response variables and a a covariant. Therefore, the optimal value for general k is then equal to the multiple least squares solution

$$\lambda_k = \frac{\langle A u_k^T, b\rangle}{\langle b, b\rangle}. \tag{15}$$

This results in the closed-form solution in (7). For a more complete proof and discussion of the implications of the solution, we refer to [3]. For example, as noted by [3], an intuitive interpretation of the solution in (14) is that the optimal scores for the j-th dimension are obtained by projecting the original data over the j-th basis and then subtracting j-times the observed value of b. Moreover, as the constraints of OB do not involve any vector u_j, the optimization with respect to the basis can be derived from known results in linear algebra. The optimal value for the vector u_j, with $j = 1, ..., k$, is equal to the first k right singular values of A, sorted accordingly to the associated singular values [8, 27].

We note the following useful Lemma adapted from [3] that quantifies the additional reconstruction error of A due to using OB compared to SVD is:

Lemma A.1. *Let $\hat{\mathbf{A}} = V_k D_k U_k^T$ denote the best rank-k approximation of the matrix $\mathbf{A}$ obtained from the truncated SVD of rank k. Let $[P]_{ij} = \frac{1}{n} + \frac{b_i b_j}{\sum_{i=1}^{n} b_i^2}$. The additional reconstruction error of the* OB *algorithm compared to SVD is $\|kPV_kD_k\|_{\mathrm{F}}$.*

B FORMULATION OF SOB

To enhance the applicability of the OB algorithm, particularly in scenarios with a large number of features, we incorporate an ℓ_1-norm penalty for the matrix U. This addition aims to promote sparsity in U and enhance the numerical stability of the approximation. The modified algorithm, denoted as SOB, is formulated as follows:

$$\begin{aligned} &\arg\min_{S,U}\left\|A - SU^T\right\|_F^2 \\ &\text{subject to } \left\|u_j\right\|_2 \leq 1, \left\|u_j\right\|_1 \leq t, \left\|s_j\right\|_2 \leq 1, s_j^T s_l = 0, s_j^T B = 0, \end{aligned} \tag{16}$$

for $j = 1, \ldots, k$, and $l \neq j$. The detailed iterative approach to solving this problem is outlined and explained in [3]. The main idea is that although the minimization problem is not jointly convex in s and u, it can be addressed iteratively. When s is fixed, the minimization step is equivalent to a sparse matrix decomposition with constraints on the right singular vectors of A. On the other hand, when u is fixed, the solution for s is obtained by rearranging the constraints and solving a univariate optimization problem. This iterative process ensures orthogonality among the vectors s_j.

C ADDITIONAL EXPERIMENT RESULTS

We include an additional experiments with continuous Y to demonstrate some additional properties of OB.

C.1 Evaluation metrics

In this section, we evaluate the model performance and fairness using three key metrics: Root Mean Square Error (RMSE) and the KL-divergence between observed (actual) data predictions and counterfactual data predictions. We also include Variable Correlation and Frobenius norm of $\widetilde{\mathbf{A}} - A$ to validate the effect of OB.

RMSE. RMSE is a widely adopted metric for evaluating prediction performance. It effectively quantifies the overall accuracy of our model's predictions.

KL-Divergence. We use KL-divergence to measure the distance between the distribution of observed data predictions and counterfactual data predictions. Additionally, visualization of these distributions serves as an intuitive indicator of counterfactual fairness. Ideally, if our model satisfies counterfactual fairness, these distributions should perfectly overlap, resulting in a KL-divergence of 0.

Variable Correlations. We calculate the average pairwise correlations between variables $A\&B$ over the given variables. Correlation close to zero indicates that our framework successfully mitigates the impact of B.

C.2 Data description

We conduct additional experiments using synthetic data sets and a real-world data set. Here, we provide an overview of these data sets:

Synthetic data sets. The casual graph used to generate the synthetic data is shown in Figure 4a. It is crafted to simulate high-dimensional data, incorporating a larger number of variables ($n = 10,000, q = 3, p = 40$, with additional 8 features that are independent of sensitive variables). It consist of 10,000 samples, ensuring a substantial sample size for analysis.

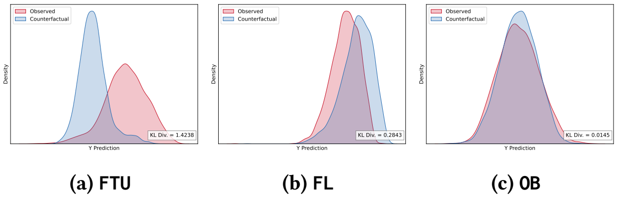
Let n denote the number of samples and p_a represent the number of features of A. Let p_b denote the number of features of B and p_x represent the number of features of X, which is unrelated to A and B. Let B follow the Bernoulli distribution with a probability equal to 0.7. Then $A_j = (\sum_{i=1}^{p_b} B_i + \varepsilon) * (i * j)$. X is unrelated with B and $X \sim \mathcal{N}(0, p_a * p_b * 0.05)$. Let $Y = \sum_{i=1}^{p_a} A_i + \sum_{i=1}^{p_x} X_i + \varepsilon$, ε is the noise and $\varepsilon \sim \mathcal{N}(0, 0.5)$. We split the data set 75/25 into a train/test set. For the counterfactual data set, we only generate 80% counterfactual data for all sensitive variables B_i.

To generate the data, we follow the casual graph as Figure 4a and set $n = 10000, p_b = 3, p_a = 40, p_x = 8$. The casual graph is similar with Figure 4a but with different number of variables. By comparing (c) and (d) to rest of the figures, we observe that OB effectively increases overlap of the observed and counterfactual distributions, indicating improved counterfactual fairness compared to FTU or FL.

Law School data set. This real-world data set is derived from a survey conducted by the Law School Admission Council across 163 law schools in the United States [32]. The casual graph is shown in Figure 4b. It contains comprehensive information on 21,790 law students, including their entrance exam scores (LSAT), grade-point averages (GPA) collected prior to law school, and their first-year average grade (FYA). The data set serves the purpose of predicting whether an applicant will achieve a high FYA, while ensuring that these predictions remain unbiased by an individual's race and sex. However, the LSAT, GPA, and FYA scores may exhibit bias due to underlying social factors. We split the data set into training (80%) and testing (20%) subsets to fit our model. Our framework is compared against the fair learning (FL) method proposed by [24].

We postulate that a latent variable, a student's knowledge, affects GPA, LSAT, and FYA scores. The casual graph corresponding to this model is shown in Figure 4b.This is a short-hand for the distributions:

$$
\begin{aligned}
GPA &\sim \mathcal{N}(b_g + w_G^K K + w_G^R R, \sigma_G),\\
FYA &\sim \mathcal{N}(w_F^K K + w_F^R R, 1),\\
LSAT &\sim \text{Poisson}(exp(b_L + w_L^K K + w_L^R R)),\\
K &\sim \mathcal{N}(0, 1).
\end{aligned}
$$



(a) FTU

(b) FL

(c) OB

Figure 6: The red line represents the $\hat{Y}$ distribution utilizing the observed variables, while the blue line represents the prediction distribution utilizing the counterfactual data. (a) and (b) shows the distributions of $\hat{Y}$ using FTU and FL methods respectively, while (c) presents the distributions of $\hat{Y}$ using OB processed data.

We perform inference on this model using an observed training set to estimate the posterior distribution of K. We use the probabilistic programming language Stan to learn K. We utilize K to predict FYA.

C.3 Synthetic Results

Table 4 summarizes the results for three models on both synthetic data sets and real data set: unaware model, which ignores the sensitive variable, fairness learning [24], and the proposed OB. According to Table 4, our framework effectively reduces the impact of sensitive variables B on Y and A with minimal information loss while achieving desirable prediction performance. More detailed discussions are listed as follows:

Prediction Performance. In order to achieve improved counterfactual fairness, our framework makes a trade-off by slightly sacrificing prediction performance. According to Table 4, the resulting degradation in prediction performance is low in the both the high-dimensional synthetic dataset and LSAT. Compared to fair learning, the OB technique gains greater flexibility in adjusting the data with minimal modifications in high-dimensional data matrices, resulting in less dent to performance in these cases.

Counterfactual Fairness. KL-divergence metric in Table 4 and Figure 6 clearly illustrate that OB achieves improved counterfactual fairness, as evidenced by the decreased KL-divergence and increased overlap between the blue and red distributions. In particular, in the high-dimensional experiment, our framework exhibits a minor degradation of 3.43% in prediction performance, while significantly enhancing counterfactual fairness. Therefore, our framework effectively isolates the impact of sensitive variables on non-sensitive variables and outcomes, while maintaining a high level of prediction performance.

Variable Correlations. To verify how OB affects the raw data sets, we note the correlation results demonstrate that our framework significantly decreases the correlation between $A\&B$ across all data sets. This reduction indicates that our framework successfully mitigates the impact of B.

C.4 LSAT result

In particular, here we discuss the impact of `OB` on real data set. In Table 4, we observe that our model achieves lower variable correlations between the sensitive variables B and A, as well as between B and Y. This indicates that the OG technique successfully removes the impact of the sensitive variable, race, on GPA, LSAT, and FYA while introducing minimal changes to the original data set. Furthermore, our model outperforms fair learning (`FL`) in terms of prediction performance, as evidenced by smaller RMSE and MAPE values. Figure 6 further demonstrates our model's effectiveness in achieving better counterfactual fairness, as indicated by the increased overlap between the red and blue distributions.