Counterfactual Fairness Through Transforming Data Orthogonal to Bias

Shuyi Chen[1] and Shixiang Zhu[1]

[1]Carnegie Mellon University

September 25, 2024

Abstract

Machine learning models have shown exceptional prowess in solving complex issues across various domains. However, these models can sometimes exhibit biased decision-making, resulting in unequal treatment of different groups. Despite substantial research on counterfactual fairness, methods to reduce the impact of multivariate and continuous sensitive variables on decision-making outcomes are still underdeveloped. We propose a novel data pre-processing algorithm, *Orthogonal to Bias* (`OB`), which is designed to eliminate the influence of a group of continuous sensitive variables, thus promoting counterfactual fairness in machine learning applications. Our approach, based on the assumption of a jointly normal distribution within a structural causal model (SCM), demonstrates that counterfactual fairness can be achieved by ensuring the data is orthogonal to the observed sensitive variables. The `OB` algorithm is model-agnostic, making it applicable to a wide range of machine learning models and tasks. Additionally, it includes a sparse variant to improve numerical stability through regularization. Empirical evaluations on both simulated and real-world datasets, encompassing settings with both discrete and continuous sensitive variables, show that our methodology effectively promotes fairer outcomes without compromising accuracy.

1 Introduction

The fairness concern in machine learning has catalyzed a growing body of research aimed at identifying, understanding, and mitigating the biases present in data and algorithms. Among the various conceptual frameworks developed to address this issue [1, 2, 3, 4, 5, 6], the notion of *counterfactual fairness* [7] emerges as particularly significant. It seeks to ensure that a decision for an individual remains consistent with the decision that would have been made in a counterfactual scenario in which the individual's sensitive variables were different. This concept is particularly powerful as it aligns closely with intuitive notions of individual fairness and justice, offering a rigorous standard by which to measure and rectify bias [8].

Significant progress has been made in achieving counterfactual fairness, with extensive studies exploring various approaches [2, 7, 9, 10, 11]. However, existing methods face two significant challenges when applied in practical scenarios: (i) These methods typically address discrete or binary sensitive variables and struggle to accommodate continuous variables such as age or weight [12]. (ii) While these methods can effectively manage a single sensitive variable, they do not extend well to scenarios involving multiple sensitive variables, thus limiting their practical utility. For instance, socio-economic status in employment hiring processes encompasses diverse factors like education, income, occupation, neighborhood, and parental education, which cannot be neatly categorized into

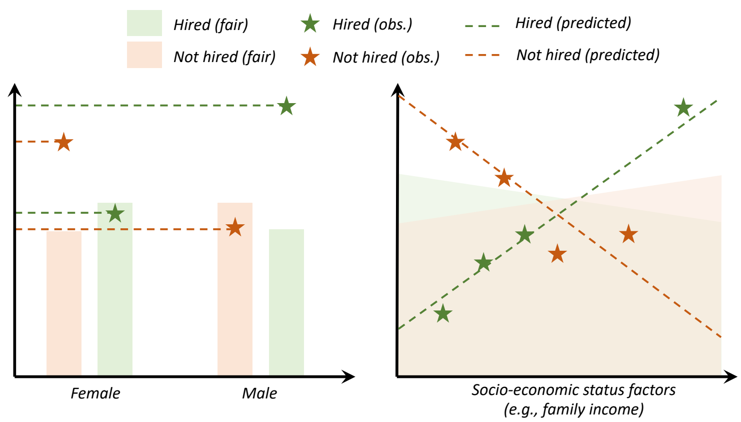


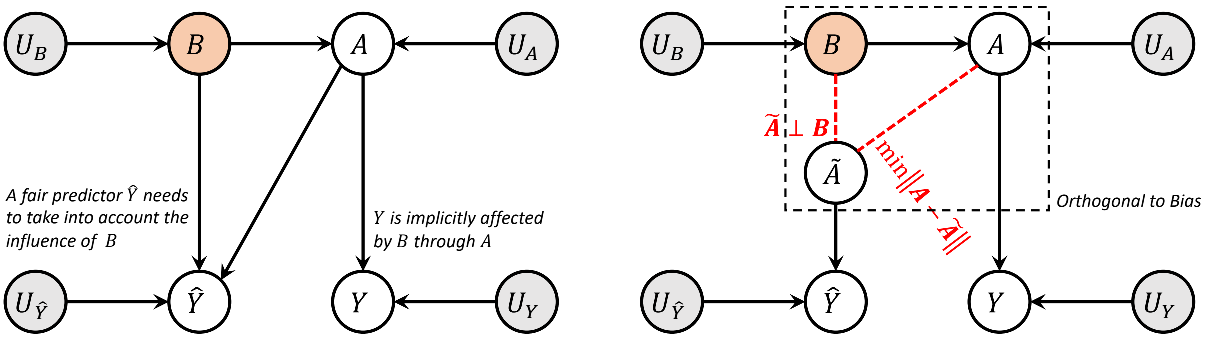
Figure 1: This motivating example illustrates the distinction between binary and continuous sensitive variables in the context of hiring decisions. The dashed lines indicate the predicted hiring decisions, and the shaded indicates the fair decisions. The left panel presents a simpler scenario with a binary sensitive variable, such as gender, where adjustments for fairness are more straightforward due to limited, exhaustively enumerable values. In contrast, the right panel delves into the complexity introduced by a continuous sensitive variable, such as family income. In this case, counterfactual estimation becomes significantly more challenging due to potentially unseen or sparsely distributed values. Additionally, socio-economic status, as a multi-dimensional variable combining various factors, adds complexity to establishing fair decision-making processes.

simple discrete groups, as shown in Figure 1. Training models on historical hiring data directly may inadvertently bias them towards candidates with specific types of experience—factors often correlated with higher socio-economic status [13]. This bias may stem from unrepresentative training data or traditional fairness methods that fail to capture the complex influence of continuous multivariate sensitive variables on decision-making.

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 counterfactual fairness can be attained 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 [14]. 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 to handle high-dimensional features. Then the resulting data readily serves 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 two simulated datasets and three real-world datasets, covering both continuous and discrete sensitive variable settings, demonstrating that our approach enables machine learning models to achieve fairer outcomes with comparable accuracy to current state-of-the-art fair learning methods. The results also demonstrate that our approach is robust beyond the joint normality assumption, 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



(a) The SCM and typical fair learning methods (b) Transforming data with *Orthogonal to Bias* (`OB`)

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.

non-sensitive and sensitive data under some mild conditions;

2. We introduce a model-agnostic data pre-processing algorithm, termed *Orthogonal to Bias* (`OB`), which facilitates counterfactual fairness across a broad spectrum of downstream machine learning applications with continuous multivariate sensitive variables;

3. We validate the enhanced efficacy of our algorithm compared to the existing state-of-the-art methods through evaluations on both synthetic and real datasets.

Related work This section begins by examining the fundamental fairness definitions that form the basis of our modeling framework. Following this, we delve into prominent machine learning techniques designed to ensure fairness, with a particular emphasis on counterfactual fairness.

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 [15, 16, 10, 17]. Metrics such as individual fairness (IF) [3], demographic parity or group fairness [5, 18] and equalized odds [2] 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, equalized odds ensures fairness by requiring that the true positive and false positive rates are equal across different groups [9, 2].

Counterfactual methods, on the other hand, offer a causal perspective on fairness. These approaches determine fairness based on potential changes in outcomes if sensitive variables were altered [2, 7, 9, 10, 11]. Such methods extend the concept of fairness beyond observable measures. For example, the Equal Opportunity (EO) criterion directly compares the actual and counterfactual decisions for the same individual, assuming the individual had the same non-sensitive attributes, providing a more nuanced assessment than equalized odds [9].

While integrating observational fairness into machine learning models is relatively straightforward, achieving counterfactual fairness often requires approximations of causal models or counterfactuals, which presents two major challenges [9, 19]. First, the process of counterfactual estimation often compromises the predictive accuracy of models due to the exclusion of related information [20, 21, 22].

For example, the FairLearning algorithm [7] uses a Markov chain Monte Carlo method to simulate unobserved portions of the causal model, making decisions based solely on variables that are not descendants of sensitive variables. Second, estimating counterfactual distributions for sparse or continuous sensitive variables is difficult and often violates basic causal inference assumptions, namely Positivity [19]. This underscores the ongoing challenge of achieving counterfactual fairness with continuous sensitive variables. Our method addresses these two challenges by ensuring minimal data modification to achieve orthogonality between non-sensitive and sensitive data while maintaining the predictive accuracy of machine learning models.

Counterfactually fair learning. Counterfactual fairness in machine learning is achieved when a decision for an individual would remain unchanged in a counterfactual world where the individual's sensitive variables differ [7]. This pursuit has led to the development of diverse strategies to enhance fairness, categorized into three main types: pre-processing [10], post-processing [19], and in-processing methods [12].

Our approach is most related to [10], which is also a pre-processing technique for counterfactual fairness. In [10], the authors propose two distribution adjustment procedures for making counterfactually fair decisions based on adjusted data. While both procedures remove variables' dependence on sensitive variables under respective conditions, their methods provide no guarantee regarding the scale of modification due to the distribution modification. 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. Furthermore, the method proposed by [10] 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.

Additionally, our work contrasts with [12], the only other effort to the best of our knowledge that specifically tackles counterfactual fairness with continuous sensitive variables. While we include their method as a compared baseline, their approach relies on a generic loss design that lacks explicit guarantees for counterfactual fairness. In contrast, our method offers a theoretical underpinning in the motivation and design of the method, prioritizing minimal data changes to enhance the balance between accuracy and fairness.

Overall, our emphasis on minimal data modification places our proposed algorithm in a unique position within the widely observed fairness-accuracy spectrum [20, 21, 22]. Through our approach, we aim to capture as much information as possible between the target variable Y and the features, including the sensitive features.

2 Methodology

2.1 Problem setup

We jointly define q non-sensitive random variables as $A \in \mathcal{A} \subseteq \mathbb{R}^q$, p sensitive variables as $B \in \mathbb{R}^p$, and the decision variable as $Y \in \mathcal{Y}$. The data generation process in our problem setup is described by a Structural Causal Model (SCM) [14], as shown in Figure 2a. Our setup allows for sensitive variables (B) that are both continuous and multivariate, facilitating the handling of a broader range of real-world applications involving complex sensitive variables.

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 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:

$$B = f_B(U_B), \quad A = f_A(B, U_A), \quad Y = f_Y(A, B, U_Y). \tag{1}$$

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 conditioned 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 counterfactual fairness [7, 10] without being influenced by the bias in B. Formally, the counterfactual fairness in our SCM can be defined as follows:

Definition 2.1 (Counterfactual Fairness)**.** Given a new pair of variables $(\mathbf{b}, \mathbf{a})$, a decision variable Y is considered counterfactually fair if, for any $\mathbf{b}' \in \mathcal{B}$,

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

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 2.1 uses the decision variable Y, it also applies to its predictor $\hat{Y}$ without any loss of generality [10].

2.2 Achieving counterfactual fairness via data decorrelation

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

Consider a dataset $\mathcal{D} = \{(\mathbf{b}_i, \mathbf{a}_i, y_i)\}_{i=1}^n$ 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 dataset $\mathcal{D}$.

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

Assumption 2.2. 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 2.3. *Under Assumption 2.2, $\hat{Y}$ is counterfactually fair when A and B are uncorrelated.*

Proof. We first demonstrate that $\hat{Y}$ is counterfactually fair 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 [14], the conditional expectation of $\hat{Y}_{\mathbf{b}'}$ given $B = \mathbf{b}^*, A = \mathbf{a}^*$ can be written as:

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

where $f_{\hat{Y}}(\cdot;\mathcal{D}) : \mathcal{A} \rightarrow \mathcal{Y}$ denotes the predictor of $\hat{Y}$ trained using data $\mathcal{D}$ and $\mathbb{P}_{U_A|B,A}(u \mid B = \mathbf{b}^*, A = \mathbf{a}^*)$ 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 $f_A(\mathbf{b}, u)$ 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 2.2, 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 $\Sigma_A \in \mathbb{R}^{q\times q}$, $\Sigma_B \in \mathbb{R}^{p\times p}$. Recall that when A and B are uncorrelated, the cross-covariance matrix between A and B is $\Sigma_{[A,B]} \in \mathbb{R}^{q\times p} = \mathbf{0}$. Therefore, the covariance matrix $\Sigma_{[A,B]}$ of joint distribution of $[A, B]$ is

$$\Sigma_{[A,B]} = \begin{bmatrix} \Sigma_A & 0 \\ 0 & \Sigma_B \end{bmatrix} \text{ and } \Sigma_{[A,B]}^{-1} = \begin{bmatrix} \Sigma_A^{-1} & 0 \\ 0 & \Sigma_B^{-1} \end{bmatrix}.$$

Substituting $\Sigma_{[A,B]}$ above into the joint probability density function of A and B, we have

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

As we can observe, 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 implies that, given the SCM in (1), predictor $\hat{Y}$ is counterfactually fair. □

Theorem 2.3 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 minimally adjust the observed data to remove the correlation between non-sensitive and sensitive variables while preserving strong predictive performance.

It is important to emphasize that Assumption 2.2 is applicable across a diverse array of applications that necessitate data standardization. For example, the SCM configuration depicted in Figure 2a, for instance, aligns with Assumption 2.2 when A is linear in terms of U_A, as shown in Appendix A. While the assumption of normality may not always hold perfectly, standardization serves as a pragmatic and effective pre-processing step that aligns real-world data more closely with the required conditions. By utilizing this widely accepted practice, empirical results indicate that our method remains both effective and robust, capable of delivering reliable performance even when applied to real-world datasets that do not fully meet this criterion. In particular, as elaborated in Section 3, our strategy has proven to be promising in experiments that involve both categorical and continuous sensitive variables B.

2.3 Orthogonal to bias

Now we introduce our data pre-processing algorithm, termed as *Orthogonal to Bias* (`OB`). We assume both non-sensitive variables A and sensitive variables B have a zero mean for each column. This allows us to estimate the empirical covariance between A and B by calculating their inner product:

$$\mathrm{cov}(A, B) = \mathbb{E}[(A - \mathbb{E}[A])(B - \mathbb{E}[B])] \approx \langle \mathbf{A}, \mathbf{B} \rangle .$$

Since two variables are uncorrelated if their covariance is zero, achieving orthogonality between the observed data $\mathbf{A}$ and $\mathbf{B}$ ensures their uncorrelation.

Therefore, the key idea of `OB` algorithm is to adjust the observed non-sensitive data $\mathbf{A}$ in such a way that it is orthogonal to the observed sensitive data $\mathbf{B}$, while ensuring minimal changes to non-sensitive data $\mathbf{A}$. Specifically, we follow the idea of Orthogonal to Groups introduced by [23], and define a rank k approximation of $\mathbf{A}$ as $\widetilde{\mathbf{A}} = \mathbf{S}\mathbf{U}^\top$, where $\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_k]$ is a $q \times k$ orthonormal matrix and $\mathbf{S} = \{s_{ij}\}$ is a $n \times k$ matrix with associated scores. The goal is to find a transformed $n \times q$ matrix $\widetilde{\mathbf{A}}$ that is orthogonal to $\mathbf{B}$ 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:

$$\arg\min_{\mathbf{S},\mathbf{U}} \left\| \mathbf{A} - \mathbf{S}\mathbf{U}^\top \right\|_F^2, \quad \text{s.t.} \ \left\langle \mathbf{S}\mathbf{U}^\top, \mathbf{B} \right\rangle = \mathbf{0}, \ \mathbf{U} \in \mathcal{G}_{q,k}, \tag{4}$$

where $\mathcal{G}_{q,k}$ is the Grassman manifold [24]. The last constraint helps prevent degeneracy, such as basis vectors becoming identically zero or encountering solutions with double multiplicity. Here we let $\mathbf{U}$ be orthonormal matrix in practice, *i.e.*, $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_k$.

In the following, we focus on a univariate $\mathbf{b} \in \mathbb{R}^n$ for clarity. We note the procedure can be easily extended for multivariate B [23]. The above constrained optimization (4) can be reformulated in terms of Lagrange multipliers λ_j:

$$\arg\min_{\mathbf{S},\mathbf{U}} \frac{1}{n} \sum_{i=1}^{n} \|\mathbf{a}_i - \sum_{j=1}^{k} s_{ij}\mathbf{u}_j\|^2 + \frac{2}{n} \sum_{j=1}^{k} \lambda_j \sum_{i=1}^{n} s_{ij} b_i, \tag{5}$$

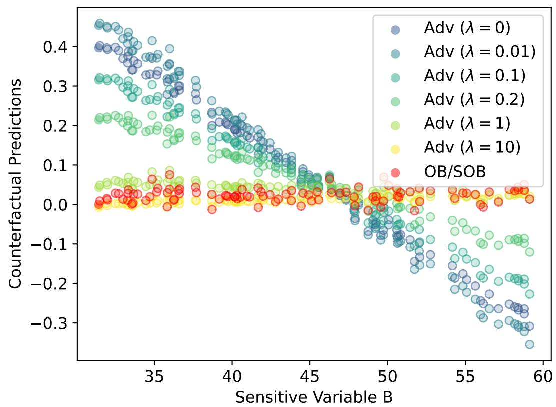
where $\lambda_j \in \mathbb{R}$ denotes the Lagrange multiplier and b_i is the i-th item of $\mathbf{b}$. Here, the factor $2/n$ is introduced to simplify the expression for the optimal solutions. It ensures that the first-order condition for (5) with respect to s_{ij} involves a common factor of $2/n$, which can then be canceled out during computations. Let $\mathbf{S}^*$ and $\mathbf{U}^*$ denote the optimal solutions of $\mathbf{S}$ and $\mathbf{U}$, respectively. The reformulated equation (5) gives a closed-form solution:

$$\lambda_j^* = \frac{\left\langle \mathbf{A}\mathbf{u}_j^{*\top}, \mathbf{b} \right\rangle}{\langle \mathbf{b}, \mathbf{b} \rangle}, \quad \mathbf{U}^* = [\mathbf{u}_1^*, \dots, \mathbf{u}_k^*], \quad \mathbf{S}^* = \{s_{ij}^*\}, \tag{6}$$

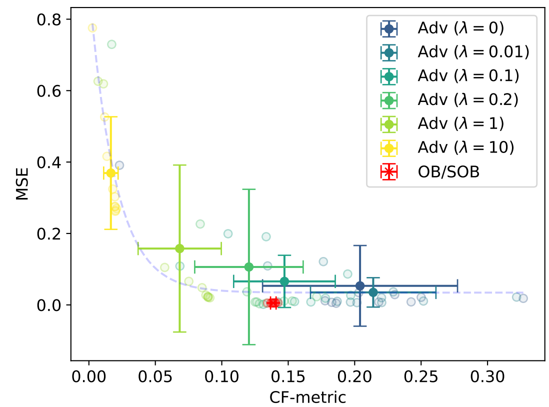
where $\mathbf{U}^*$ consists of the first k right singular vectors of $\mathbf{A}$ and $s_{ij}^* = \mathbf{a}_i\mathbf{u}_j^{*\top} - \lambda_j^* b_i$. The pre-processed non-sensitive data matrix can be obtained by $\widetilde{\mathbf{A}} = \mathbf{S}^*\mathbf{U}^{*\top}$. Detailed derivations of the closed-form solution (6) can be found in Appendix B.

It is noteworthy that the reconstruction error of $\widetilde{\mathbf{A}}$ equals to that of the standard Singular Value Decomposition (SVD) if there is zero correlation between the sensitive and non-sensitive variables. 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 $\mathbf{B}$ and the left singular vectors of the data $\mathbf{A}$ [23]. The comparison of reconstruction error between the `OB` and SVD can be found in Appendix B.

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 [25]. To address this challenge, we introduce a sparse variant of the `OB`



(a) Counterfactual prediction



(b) Predictive error vs counterfactual fairness

Figure 3: Comparison of our approach (OB/SOB) with the primary baseline methods on the synthetic dataset: (a) Counterfactual predictions for 1,000 randomly generated sample points across the sensitive variable, comparing our approach to baseline methods. Our method's predictions remain flat across the sensitive variable; (b) Comparison of predictive error and counterfactual fairness between our approach and baseline methods. The dashed line represents an exponential function fitted to the results of Adv. The error bars indicate standard deviation. Our method is under the curve and achieves a low MSE while maintaining relatively low CF compared to Adv.

algorithm, referred to as SOB. The SOB imposes an ℓ_1-norm penalty for $\mathbf{U}$ to encourage sparsity and improve numerical stability, in addition to the orthogonality constraints in (4). We define h as the ℓ_1 constraint on u. Details of the formulation of SOB can be found in Appendix C. Following derivation by [23, 26], Algorithm 1 illustrates the key steps to implement the SOB algorithm, where η represents the minimum change to terminate the iterative optimization process.

Note that with the additional regularization constraints, the solution favors sparsity while satisfying the orthogonal constraint. Therefore, the SOB also achieves counterfactual fairness under SCM framework with Theorem 2.3.

3 Experiments

In this section, we empirically evaluate the performance of our method using three real-world datasets and two simulated datasets. For the discrete scenario where $Y \in \{0, 1\}$ and $B \in \{0, 1\}$, we utilize two popular datasets: the Adult income dataset [27], with gender as the sensitive variable (male or female) and whether a person earns over \$50K a year as Y, and the COMPAS dataset [28], with race as the sensitive variable (Caucasian or non-Caucasian) and two-year recidivism as Y. In addition, we explore a synthetic loan decision dataset using similar setup as [10]. For scenarios where both Y and A are continuous, we use the Crime dataset [29], with the ratio of an ethnic group per population as the sensitive variable and violent crimes per population as Y. Additionally, we explore a synthetic scenario that allows for further analysis of the relative performances of the approaches. This synthetic scenario involves a pricing algorithm for a fictional car insurance price adapted from [12], following the causal graph in Figure 2a. Details of the experimental setup can be found in Appendix D.

Evaluation metrics We assess the accuracy of decisions using Mean Squared Error (MSE) for continuous value predictions, and Area Under the Curve (AUC) and Accuracy (ACC) for binary

Table 1: Performance on Crime and Synthetic Insurance Price Data

Dataset	Metrics	Baselines		Compared Methods		Ours	
		ML	FTU	Adv ($\lambda = 0.1$)	Adv ($\lambda = 0.2$)	OB	SOB
Crime [29]	MSE	0.4751 (0.0336)	0.4711 (0.0101)	0.4674 (0.0425)	0.4511 (0.0503)	<u>0.4534</u> (0.0110)	**0.4491** (0.0155)
	CF-metric	0.2353 (0.0729)	0.2467 (0.1633)	**0.0095** (0.0057)	<u>0.0066</u> (0.0029)	0.1047 (0.0373)	0.1051 (0.0328)
	Avg. Training Time (s)	68.9089 (7.4695)	62.9821 (6.5349)	136.9746 (4.4359)	131.8830 (4.4037)	**69.2535** (5.1429)	80.7566 (7.4400)
Synthetic	MSE	0.0008 (0.0009)	0.0000 (0.0000)	0.0658 (0.0731)	0.1062 (0.2174)	**0.0054** (0.0001)	**0.0054** (0.0002)
	CF-metric	0.1414 (0.0008)	0.1418 (0.0002)	0.1422 (0.0361)	**0.1189** (0.0409)	0.1309 (0.0023)	<u>0.1296</u> (0.0020)
	Avg. Training Time (s)	36.4220 (13.9463)	32.9465 (7.0428)	55.2336 (2.6324)	63.5752 (0.6989)	**32.8576** (8.6382)	37.5425 (10.8821)

Table 2: Performance comparison on COMPAS data [28]

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	<u>0.5674</u>
AUC	0.7206	0.7225	0.6928	**0.7225**	0.6927	0.6927	0.6928	0.7015	<u>0.7019</u>	0.6764	0.6744
CF-metric	0.2274	0.1406	0.0054	0.1377	0.0060	0.0058	0.0054	**0.0026**	<u>0.0027</u>	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	<u>1.6e-16</u>	<u>1.1e-16</u>

classification problems. In evaluating counterfactual fairness for continuous cases, we utilize the CF-Metric employed by [12], which measures the mean change in predictions for a set of random counterfactual samples for each individual in the test set. These counterfactuals are generated using a variational inference model of the Structural Causal Model (SCM), following the same adversarial training process described by [12]. For classification problems, we adopt CF-Metrics from [10], which measure counterfactual fairness by calculating the average change in predicted scores between two groups. This metric is designed to be zero when decisions are counterfactually fair. For a comprehensive comparison of methods, we also include two additional counterfactual fairness metrics: Affirmative Action (AA) Fairness and Equalized Opportunities (EO) Fairness [9]. Both AA Fairness and EO Fairness are only defined for scenarios with discrete sensitive variables. Definitions of these metrics are provided in Appendix D.

Results with continuous sensitive variables To our knowledge, the approach by [12] is the only other notable attempt to tackle counterfactual fairness with continuous sensitive variables. In our experiments, we included this method (henceforth referred to as `Adv`) along with a standard machine learning approach (`ML`) and the Fairness Through Unawareness (`FTU`) method as baselines. All models were evaluated using a uniform four-layer neural network for prediction tasks.

Figure 3 illustrates the fairness-accuracy trade-off, a well-explored topic in the literature [20, 21, 22]. Moving along the spectrum, `Adv` appears with progressively increasing λ values, reflecting a shift towards greater emphasis on fairness at the expense of accuracy. Our method is located in the lower-left, demonstrating low MSE and CF metrics, with narrow error bars due to its closed-form solution (Appendix B). Furthermore, Table 1 provides detailed numerical results, including mean and standard deviations from 10 repeated experiments, which consistently demonstrate our method's superior performance in achieving lower MSE and competitive CF-metrics across two datasets with continuous variables. Moreover, Table 1 shows that our method is more time-efficient than `Adv`.

Results with discrete sensitive variables We compare our method with the following baselines on datasets that include discrete sensitive variables: Machine Learning (`ML`), a straightforward logistic regression model that uses all variables, regardless of their sensitivity; Fairness through Unawareness (`FTU`), which employs a logistic model with $\mathbf{A}$, specifically excluding sensitive variables; FairLearning

Table 3: Performance comparison on Adult data [30]

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

Table 4: Performance comparison on Synthetic Loan data

Metrics	Baselines		Compared Methods							Ours	
	ML	FTU	FL	EO	AA	$\texttt{FLAP}_1\texttt{(O)}$	$\texttt{FLAP}_2\texttt{(O)}$	$\texttt{FLAP}_1\texttt{(M)}$	$\texttt{FLAP}_2\texttt{(M)}$	$\texttt{OB}_1$	$\texttt{OB}_2$
ACC	0.6618	0.6481	0.6224	0.6237	0.6224	0.6237	0.6224	0.6237	0.6224	**0.6406**	<u>0.6279</u>
AUC	0.9457	0.8986	<u>0.5867</u>	**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**	<u>0.0011</u>	0.0026
EO Fairness	0.5469	0	<u>0.0156</u>	**0**	0.0336	0.0321	<u>0.0156</u>	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**

Algorithm (`FL`), which attains counterfactual fairness by sampling unobserved portions of the causal model using Markov chain Monte Carlo methods [7]; EO-fair and AA-fair Predictors (`EO` and `AA`) by a post-processing algorithm that generate optimal predictors adhering to Equalized Opportunities (EO) and Affirmative Action (AA) [9]; and Fair Learning through Data Pre-processing (`FLAP`) by [10], which modifies data prior to model training to enhance fairness. Logistic regression predictors are adopted for all compared models with classification tasks.

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. For discrete sensitive variables, both `FLAP` and our proposed method (`OB`) can include sensitive variables in the predictor class. Specifically, aside from the predictor class $f_{\hat{Y}}(\mathbf{A}) : \mathcal{A} \to \mathcal{Y}$ discussed in Section 2.2, an standard machine learning predictor $f_{\hat{Y}}(\mathbf{A}, \mathbf{B}) : \mathcal{A} \times \mathcal{B} \to \mathcal{Y}$ that utilizes both sensitive and non-sensitive variables, and an Averaged Machine Learning (AML) predictor $f'_{\hat{Y}}(\mathbf{A}) = \sum f_{\hat{Y}}(\mathbf{A}, \mathbf{B})\mathbb{P}(\mathbf{B})$ can be constructed respectively. We denote scenarios involving predictor $f'_{\hat{Y}}(\mathbf{A})$ or $f_{\hat{Y}}(\mathbf{A})$ as $\texttt{OB}_1$ and $\texttt{OB}_2$, respectively. Similarly, we use $\texttt{FLAP}_1$ and $\texttt{FLAP}_2$ for the respective predictor scenarios. Additionally, [10] introduces two pre-processing methods, Orthogonalization and Marginal Distribution Mapping, denoted as `FLAP(O)` and `FLAP(M)`.

As shown in Table 2, 3, and 4, the accuracy of `OB` is consistently among the highest, often ranking first or second. This supports our claim that the proposed `OB` algorithm effectively preserves most of the information in the non-sensitive data. Additionally, its CF-metric is comparable to those of `FLAP` and `FL`, which are also designed for counterfactual fairness with discrete sensitive variables, indicating that `OB` also maintains counterfactual fairness. Furthermore, despite the general incompatibility between counterfactual fairness and popular group fairness metrics like Equalized Opportunities, EO can also be encouraged with `OB`. Empirically, `OB` achieves relatively low EO and AA Fairness metrics compared to `FLAP` and `FL`. More details and proof can be found in E.

In summary, across all discrete datasets, our approach consistently demonstrates its effectiveness in achieving a better overall balance between accuracy and counterfactual fairness, even when the data does not fully meet our theoretical assumptions.

Table 5: Sensitivity analysis on the COMPAS dataset

Metrics	$\texttt{OB}_1$			$\texttt{OB}_2$		
	$k = 3$	$k = 4$	$k = 5$	$k = 3$	$k = 4$	$k = 5$
ACC	0.5409	0.5507	0.5666	0.5420	0.5516	0.5674
CF-metric	0.0120	0.0157	0.0060	0.0129	0.0162	0.0065
EO Fairness	0	0	0	0	0	0
AA Fairness	0.0006	0.0003	1.6e-16	0.0006	0.0003	1.1e-16

Sensitivity analysis We examine the sensitivity of our OB method to the hyperparameter rank k. Note that rank k plays a key role in information preservation and data approximation. For SOB, additional hyperparameters, such as η—used in Algorithm 1 to set the convergence threshold—are involved in solving the optimization problem presented in (6). However, these do not significantly impact the outcomes.

We utilize the COMPAS dataset for sensitivity analysis, which contains 5 features in addition to the sensitive variable. The limited number of feature set makes it easier to illustrate the different effects of varying k by restricting k to a maximum of 5. As shown in Table 5, there is an empirical trend of improved accuracy (ACC), similar to what is observed in Principal Component Analysis (PCA) [23]. Additionally, the counterfactual fairness metrics (CF-metric and AA Fairness) also improve with a higher k.

4 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 datasets, 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. Despite the advantages of the OB algorithm, it exhibits certain limitations. As an offline, pre-processing approach, it may not integrate as seamlessly with dynamic or real-time systems compared to in-processing methods. Additionally, the theoretical guarantees of OB depend on jointly normal assumption which may not hold in some practical scenarios.

References

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

[2] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3323–3331, Red Hook, NY, USA, 2016. Curran Associates Inc.

[3] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, page 214–226, New York, NY, USA, 2012. Association for Computing Machinery.

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

[5] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 325–333, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[6] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi. Fairness constraints: Mechanisms for fair classification. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 962–970, Sydney, Australia, 20–22 Apr 2017. PMLR.

[7] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, Long Beach, CA, USA, 2017. Curran Associates, Inc.

[8] Zeyu Tang, Jiji Zhang, and Kun Zhang. What-is and how-to for fairness in machine learning: A survey, reflection, and perspective. *ACM Comput. Surv.*, 55(13s), jul 2023.

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

[10] Rui Song Haoyu Chen, Wenbin Lu and Pulak Ghosh. On learning and testing of counterfactual fairness through data preprocessing. *Journal of the American Statistical Association*, 0(0):1–11, 2023.

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

[12] Vincent Grari, Sylvain Lamprier, and Marcin Detyniecki. Adversarial learning for counterfactual fairness. *Machine Learning*, 112(3):741–763, 2023.

[13] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

[14] Judea Pearl. *Causality*. Cambridge University Press, 2 edition, 2009.

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

[16] Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. 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*, pages 502–510, 2011.

[17] Song Wei, Xiangrui Kong, Alinson Santos Xavier, Shixiang Zhu, Yao Xie, and Feng Qiu. Assessing electricity service unfairness with transfer counterfactual learning. *arXiv preprint arXiv:2310.03258*, 2024.

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

[19] Alan Mishler, Edward H Kennedy, and Alexandra Chouldechova. Fairness in risk assessment instruments: Post-processing to achieve counterfactual equalized odds. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 386–400, 2021.

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

[21] Sina Baharlouei, Maher Nouiehed, Ahmad Beirami, and Meisam Razaviyayn. R\'enyi fair inference. *arXiv preprint arXiv:1906.12005*, 2019.

[22] Richard A. Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. A convex framework for fair regression. *ArXiv*, abs/1706.02409, 2017.

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

[24] Alan T James. Normal multivariate analysis and the orthogonal group. *The Annals of Mathematical Statistics*, 25(1):40–75, 1954.

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

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

[27] Ron Kohavi. Census Income. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5GP7S.

[28] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. In *Ethics of data and analytics*, pages 254–264. Auerbach Publications, 2022.

[29] Michael Redmond. Communities and Crime. UCI Machine Learning Repository, 2009. DOI: https://doi.org/10.24432/C53W3X.

[30] Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

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

[32] Carl M. O'Brien. Statistical learning with sparsity: The lasso and generalizations. *International Statistical Review*, 84(1):156–157, 2016.

A Analyzing the Joint Normality of Linear Combinations in SCM

This section demonstrates how the Structural Causal Model (SCM) depicted in Figure 2a aligns with Assumption 2.2 under certain mild conditions. For simplification, we assume both A and B are one-dimensional variables.

Consider a scenario where the underlying exogenous variables U_A and U_B are distributed as follows:

$$
\begin{aligned}
U_A &\sim N(\mu_A, \sigma_A^2), \\
U_B &\sim N(\mu_B, \sigma_B^2),
\end{aligned}
$$

and U_B is independent of U_A.

Given their normal distributions and independence, $[U_A, U_B]$ form a jointly normal distribution. This property of normal distributions ensures that any linear combination of U_A and U_B will also be jointly normal. We define A and B through the following linear combinations:

$$
\begin{aligned}
A &= aU_A + bU_B + c, \\
B &= dU_B + e,
\end{aligned}
$$

where a, b, c, d, and e are constants. This formulation ensures that $[A, B]$ is also jointly normal, as long as A and B are defined as linear combinations of U_A and U_B as described. The synthetic data used for continuous sensitive variables, described in Appendix D.3, follows a similar structure, satisfying Assumption 2.2.

B Closed-form `OB` Solution Derivation

We start by considering (5) 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. (5) can be reformulated as:

$$
\arg\min_{S,U} \left\{ \frac{1}{n} \sum_{i=1}^{n} \| a_i - \sum_{j=1}^{k} s_{ij} u_j^T \|^2 + \frac{2}{n} \lambda_1 \sum_{i=1}^{n} s_{i1} b_i \right\}. \tag{7}
$$

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

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

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

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

Solving it finds a stationary point of

$$
s_{i1} = a_i u_1^T - \lambda_1 b_i. \tag{8}
$$

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 (7) for u_1 is equivalent to the unconstrained scenario. A standard result of linear algebra states that the optimal u_1 for (7) without constraints equivalent to the first right singular vector of $\mathbf{A}$, or equivalently to the first eigenvector of the matrix $\mathbf{A}^\top \mathbf{A}$ [26]. Plugging in the solution for u_1 and setting the derivative with respect to λ_1 equal to 0 leads to

$$\sum_{i=1}^{n} (a_i u_1^T - \lambda_1 a_i)^T b_i = 0. \tag{9}$$

Therefore,

$$\lambda_1 = \frac{\sum_{i=1}^{n} a_i u_1^T b_i}{\sum_{i=1}^{n} b_i^2} = \frac{\langle A u_1^T, b \rangle}{\langle b, b \rangle}, \tag{10}$$

which states λ is is a least squares estimate of $\mathbf{A} u_1^\top$ 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

$$s_{ij} = a_i u_j^T - \lambda_j b_i, \tag{11}$$

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 (11) 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

$$\lambda_k = \frac{\langle A u_k^T, b \rangle}{\langle b, b \rangle}. \tag{12}$$

This results in the closed-form solution in (6). For a more complete proof and discussion of the implications of the solution, we refer to [23]. For example, as noted by [23], an intuitive interpretation of the solution in (11) 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 [31, 32].

We note the following useful Lemma adapted from [23] that quantifies the additional reconstruction error of A due to using `OB` compared to `SVD` is:

Lemma B.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 $\|kPV_k D_k\|_{\mathrm{F}}$.*

C 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:

$$\begin{aligned} &\arg\min_{S,U} \left\|A - SU^T\right\|_F^2 \\ &\text{subject to } \|u_j\|_2 \leq 1, \|u_j\|_1 \leq t, \|s_j\|_2 \leq 1, s_j^T s_l = 0, s_j^T B = 0, \end{aligned} \tag{13}$$

for $j = 1, \ldots, k$, and $l \neq j$. The detailed iterative approach to solving this problem is outlined and explained in [23]. 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.

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 $\mathcal{S}_\theta(x) = \operatorname{sign}(x)(|x| - \theta)\mathbb{1}(|x| \geq \theta)$
    $t \leftarrow t + 1$
  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 $\hat{\mathbf{U}} = [u_1, \ldots u_k]$.
Calculate the attribute matrix $\widetilde{\mathbf{A}} = \hat{\mathbf{S}}\hat{\mathbf{U}}^\top$.
Output: $\widetilde{\mathbf{A}}$

D Experiment Details

The experiments were conducted in a Jupyter Notebook environment with 16 GB RAM and a 16 GB T4 GPU. Each experiment for the continuous case (Crime and Synthetic) was repeated 10 times to calculate the mean and standard deviation. For regression tasks (continuous Y), we utilized a uniform four-layer NN and MSE loss, trained for 2000 epochs with a batch size of 256. `Adv` and `ML` have an input size of $q+1$, while `FTU`, `OB`, and `SOB` have an input size of q. For classification tasks, we employed logistic regression. The hyperparameter K for our method was set to 4 for the Crime dataset and 20 for the Synthetic insurance price dataset.

D.1 Evaluation Metrics

CF-Metric in Continuous Case The Counterfactual Fairness (CF) metric for the continuous case, adopted from [10], quantifies the fairness of decisions made by a machine learning model. This metric assesses the difference in predicted outcomes under hypothetical scenarios where a sensitive attribute $\mathbf{B}$ of an individual might differ, while all other attributes remain constant. It is defined as:

$$CF = \frac{1}{m_{\text{test}}} \sum_{i=1}^{m_{\text{test}}} \mathbb{E}_{(\mathbf{A}'_i, \mathbf{b}') \sim C(i)} \left[\left| \Delta(f_{\hat{Y}}(\theta(\mathbf{A}_i, \mathbf{b}_i)), f_{\hat{Y}}(\theta(\mathbf{A}'_i, \mathbf{b}'_i))) \right| \right],$$

where m_{test} is the number of test instances, and $C(i)$ denotes the set of counterfactual samples for the i-th test instance. These samples are generated based on an adversarial inference process. Δ represents a cost function comparing two predictions, and θ denotes the transformation model applied to the inputs. For `Adv` [12], it's referring to the VAE. And for our method, it's the results of `OB` process. This metric evaluates the average absolute difference between the predicted outcomes under actual and counterfactual attributes, aiming to assess how decisions might vary with changes in the sensitive attribute B, while keeping other variables constant. This measure seeks to ensure decisions remain invariant under hypothetical alterations of sensitive attributes, thus quantifying the model's counterfactual fairness.

CF-Metric in Discrete Case The CF-metric for the discrete case is designed to measure how decisions made by a machine learning model might differ if a sensitive attribute B were altered, assuming all other attributes A remain unchanged. It is defined as:

$$CF = \max_{r,h \in \mathcal{B}} \left(\frac{1}{m_{\text{test}}} \sum_{i=1}^{m_{\text{test}}} \left| f_{\hat{Y}}(\theta(r, \hat{a}^D_M(r, b_i, a_i))) - f_{\hat{Y}}(\theta(b, \hat{a}^D_M(h, b_i, a_i))) \right| \right),$$

where m_{test} is the number of test instances, and $\mathcal{B}$ represents the set of sensitive groups under consideration. $\hat{a}^D_M(h, b^*, a^*)$ is the mapping function to compute non-sensitive attributes assuming that the individual belongs to a different sensitive group h [10]. The functions $f_{\hat{Y}}(\theta(r, \cdot))$ and $f_{\hat{Y}}(\theta(h, \cdot))$ evaluate decision outcomes for real and hypothetical scenarios, respectively. This metric aims for a value of zero, indicating perfect counterfactual fairness across all considered groups.

For AA and EO metric definitions, we refer to [9] for details.

D.2 Real-world Datasets

Adult In the Adult Income dataset [30], 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 variables 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 dataset, we employ `SOB`. As illustrated in Table 3, 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 variables and generates unfair results. As noted by [23, 31], 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 [28] 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.

Crime Crime dataset [29] with the ratio of an ethnic group per population as sensitive variable, and violent crimes per population as Y.

D.3 Simulated Datasets

D.3.1 Synthetic Insurance Price Dataset

We simulate two scenarios for continuous sensitive variable B and decision variable Y following Figure 2a:

$$U \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0.5 \\ 1 \\ 1.5 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 2 \end{bmatrix}\right)$$

The non-sensitive variables A_1, A_2, A_3, A_4 are defined as follows:

$$A_1 \sim \mathcal{N}(7 + 0.1 \cdot B + U_1 + U_2 + U_3, 1),$$

$$A_2 \sim \mathcal{N}(80 + B + U_2, 10),$$
$$A_3 \sim \mathcal{N}(200 + 5 \cdot B + 5 \cdot U_3, 20),$$
$$A_4 \sim \mathcal{N}(10^4 + 5 \cdot B + U_4 + U_5, 1000),$$

where B is defined by the vector:

$$B \sim \mathcal{N}(\big[45, 5\big]).$$

The response variable Y is modeled as:

$$Y \sim \mathcal{N}(2 \cdot (7 \cdot B + 20 \cdot \sum_j A_j), 0.1).$$

D.3.2 Synthetic Loan Decision Dataset

We apply our methods to a synthetic loan dataset example which is a modification from [10]. 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 [10], 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:

$$\begin{aligned} E &= \max\{0, U_E\}, \\ I &= \exp\{0.1U_E + U_I\}. \end{aligned} \tag{14}$$

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

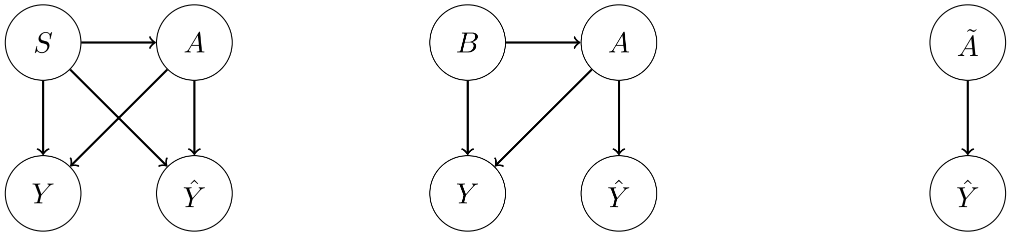
$$Y = \mathbb{1}\{U_Y < \text{expit}(\beta_0 + \beta_1\mathbb{1}\{B = 1\} + \beta_2\mathbb{1}\{B = 2\} + \beta_E E + \beta_I I)\}, \tag{15}$$

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

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 2.2. As evidenced in Table 4, despite the deviation from the assumptions in the tested synthetic dataset, 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, akin to `FL` and `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.

E Explanation for EO Fairness of OB and SOB



(a) The SCM with typical predictor (b) The SCM with FTU preidctor (c) The SCM with `OB` and `SOB`

Figure 4: Causal diagrams showing the relationships between S, A, Y, and $\hat{Y}$.

Here we explain how `FTU` achieves EO-fairness and why predictors based on our proposed methods, `OB` and `SOB`, also exhibit low EO-metric values. We clarify this through the following lemma about EO-fairness, which shows the EO-fairness of a decision $\hat{Y}$ is equivalent to the absence of any causal arrow from B_i to $\hat{Y}$. The proof of the lemma can be found in [9].

Lemma E.1 (EO $\Leftrightarrow$ No $(B \to \hat{Y})$)**.** *Assume the causal graph in Fig. 4a. A decision $\hat{Y}$ satisfies equal opportunities over S if and only if there is no causal arrow between B and $\hat{Y}$.*

With this lemma, it becomes straightforward to establish the EO-fairness of `FTU`, where $\hat{Y}^{FTU}(a, b) = \hat{Y}^{FTU}(a)$ (see Figure 4b). Concurrently, the designs of `OB` and `SOB` leverage a predictor on $\tilde{A}$, which is the closest counterpart to A (in terms of the Frobenius norm) that is orthogonal to B (see Figure 2b and 4c). This configuration endows them with properties similar to `FTU` that promotes EO-fairness, but with the added benefit that modifications are minimal while effectively eliminating the influence of B on A, leading to Theorem 2.3.