BaBE: Enhancing Fairness via Estimation of Explaining Variables

Ruta Binkyte
CISPA Helmholtz Center for
Information Security
Germany
ruta.binkyte-sadauskiene@cispa.de

Daniele Gorla
Università di Roma "La Sapienza"
Italy

Catuscia Palamidessi
Inria Saclay and École Polytechnique
France

ABSTRACT

We consider the problem of unfair discrimination between two groups and propose a pre-processing method to achieve fairness. Corrective methods like statistical parity usually lead to bad accuracy and do not really achieve fairness in situations where there is a correlation between the sensitive attribute S and the legitimate attribute E (explanatory variable) that should determine the decision. To overcome these drawbacks, other notions of fairness have been proposed, in particular, conditional statistical parity and equal opportunity. However, E is often not directly observable in the data. We may observe some other variable Z representing E, but the problem is that Z may also be affected by S, hence Z itself can be biased. To deal with this problem, we propose BaBE (Bayesian Bias Elimination), an approach based on a combination of Bayes inference and the Expectation-Maximization method, to estimate the most likely value of E for a given Z for each group. The decision can then be based directly on the estimated E. We show, by experiments on synthetic and real data sets, that our approach provides a good level of fairness as well as high accuracy.

KEYWORDS

Fairness, Explainability

1 INTRODUCTION

One of the first group of fairness notions proposed in literature was *statistical parity* (SP) [11], which enforces the probability of a positive prediction to be equal across different groups. Let the prediction and the group be represented, respectively, by the random variables $\hat{Y}$ and S, both of which are assumed to be binary for simplicity, and let $\hat{Y} = 1$ stand for the positive prediction. Then SP is formally described by $\mathbb{P}[\hat{Y} = 1|S = 1] = \mathbb{P}[\hat{Y} = 1|S = 0]$, where $\mathbb{P}[\cdot|\cdot]$ represents conditional probability.

However, SP has been criticized for causing loss of accuracy and for ignoring circumstances that could justify disparity. A more refined notion is *conditional statistical parity* (CSP) [21], which allows some disparity as long as it is legitimated by explaining factors. For example, a hiring decision positively biased towards Group 1 could be justified if Group 1 has a higher education level than Group 0 in average. CSP is formally defined by $\mathbb{P}[\hat{Y} = 1|S = 1, E = e] = \mathbb{P}[\hat{Y} = 1|S = 0, E = e]$, for all e, where E is a random variable representing the ensemble of explaining features.

The most common pre-processing approach to achieve CSP (or an approximation of it) consists in editing the label Y (decision) in the training data, according to some heuristic, so to ensure that the number of samples with $Y = 1$, $S = 1$, and $E = e$ are approximately the same number as those with $Y = 1$, $S = 0$, and $E = e$. One problem, however, is that often E is not directly observable in the data. Usually, we can observe some other variable Z that is representative of E, but the problem is that Z may be also influenced by the sensitive attribute S, hence Z itself can be biased. We illustrate this scenario with the following examples.

EXAMPLE 1. *The SAT (Scholastic Assessment Test) is a standardized test widely used for college admissions in the United States aiming at indicating the skill level of the applicant, and therefore her potential to succeed in college. However, the performance at the test can be affected by other socio-economic, psychological, and cultural factors. For instance, a recent study [16] points out that, on average, black students are less likely to undergo the financial burden of retaking the test than white students. This causes a racial gap in the scores, since retaking the test usually improves the result. Another study [17] reports that, on average, girls score approximately 30 points less on SAT than boys, despite the fact that girls routinely achieve higher grades in school. According to [17], the cause is the higher sensitivity to stress and test anxiety among females.*

EXAMPLE 2. *Many healthcare systems in the United States rely on prediction algorithms to identify patients in need of assistance. One of the most used indicators is the individual healthcare expenses, as they are easily available in the insurance claim data. However, healthcare spending is influenced not only by the health condition, but also by the socio-economic status. A recent study [31] shows that typical algorithms used by these healthcare systems are negatively biased against black patients, in the sense that, for the same prediction score, black patients are in average sicker than the white ones. According to [31], this is due to the bias in the healthcare spending data, since black patients spend less on healthcare due to lower financial capabilities and lower level of trust towards the white-dominated medical system and practitioners.*

In the above examples, the "true skills" and the "true health status", respectively, are the legitimate features E (explanation) on
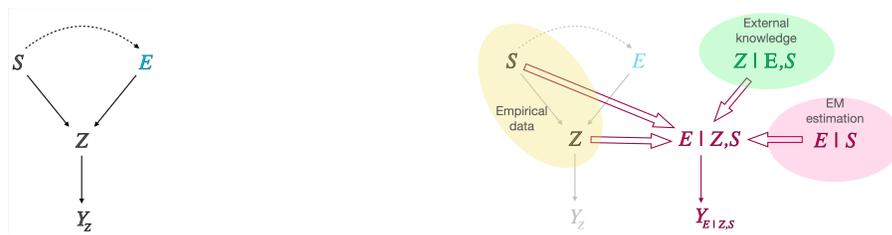
Figure 1: Left: illustration of the causal relation between the data. Right: illustration of our pre-processing method.

which we should base the decision. Unfortunately E is not directly observable. What we can observe, instead, is the result of the SAT test and the healthcare-related spending, respectively. These are represented by the variable Z. These indicators, however, do not faithfully represent E, because they are influenced also by other factors, namely the economical status (or the gender), and the race, respectively. These are the sensitive attribute S.

The line of research that advocates the use of statistical parity [6, 8, 20, 26, 27] adheres to the "we are all equal" principle [14], and makes the basic assumption that *E and S are independent*. However, in many cases, like for instance in decisions regarding the medical treatment of genetic illnesses, race or gender could have a direct effect on the likeliness of the medical condition. For example, in our second running example, the real health status is on average lower in the black population because of socio-economic factors. Hence, we allow the possibility of a link between the sensitive attribute S and the explaining value E, and aim to remove the discrimination introduced by the link between S and Z. The method we propose to remove the discrimination works equally well whether or not there is a link between S and E, and it does not modify this relation.

To summarize, in the original (unfair) scenario the decision Y is based on Z, which is influenced by both E and S. The situation is represented in Figure 1 (left). The arrow from S to Z represents that there is a causal relation between S and Z, and similarly for the other solid arrows[1], while the dashed arrow between S and E represents a relation that may or may not be present. In order to take a fair decision, we would like to base the decision Y only on E, but, as explained before, E may not be directly available. Therefore, we need to determine what is the most likely value of E for the given values of S and Z. To this purpose, we will derive the conditional distribution of E given Z and S, i.e. $\mathbb{P}[E|Z,S]$. The objective is illustrated in Figure 1 (right).

Note that E can be multi-dimensional, and that we represent the effect of other possible latent variables by the randomness in the distribution of the data.

The method we propose uses a combination of the *Bayes theorem* and the *Expectation-Maximization method* (EM) [10], a powerful statistical technique to estimate unobservable variables as the maximum likelihood parameters of empirical data observations. We call our method BaBE, for *Bayesian Bias Elimination.*

BaBE relies on some additional knowledge, namely an estimation of the conditional distribution of Z given S and E, i.e., $\mathbb{P}[Z|E,S]$. This estimation can be obtained by collecting additional data. For instance, for Example 2, we could use the richer set of biomarkers,

like in [31]. Alternatively, it can be produced by studies or experiments in a controlled environment. For instance, for Example 1, we could assess skills in some subjects by in-depth examinations, and derive statistics about their SAT performance both at the first attempt and after several retakes. Another possibility is to collect data on the subsequent performance of the students that have been accepted, and of those who have not been accepted in the school in question but have been accepted in another school.

One obvious question that may arise is: what are the advantages of deriving $\mathbb{P}[Z|E,S]$, rather than directly $\mathbb{P}[E|Z,S]$, from the additional data? (The derivation of the latter from the former is the essence of our proposal.) We argue that, while in general there may not be any advantage, there are real-life situations in which $\mathbb{P}[Z|E,S]$ is more "universal" than $\mathbb{P}[E|Z,S]$, in the sense that the first does not depend on the distribution of $E|S$ (E given S), while the latter does. As a consequence, the knowledge of the first can be re-used in different contexts, while the latter cannot. One typical example is the study of symptoms (Z) induced by certain diseases (E), which may also depend on the gender or other characteristics such as ethnicity, age, etc. (S): $\mathbb{P}[Z|E,S]$ can be statistically estimated from medical data D collected by some hospitals, and it is reasonable to assume that it does not depend on the distribution of $E|S$, which, in contrast, can vary greatly depending on the geographical area, on the social context, etc. Also $\mathbb{P}[E|Z,S]$ could be estimated from D, but it may depend on $E|S$. For example, in towns that are very polluted (area A_1), the risk that coughing (symptom, Z) indicates lung cancer rather than a simple cold (diseases, E) may be much higher than in the (less polluted) area A_2 where the data D were collected. The idea of BaBE to predict diseases in A_1 is to estimate $\mathbb{P}[Z|E,S]$ in the area where complete data (including E) are available, in this case, area A_2, assuming that the same $\mathbb{P}[Z|E,S]$ is valid also in A_1. Then, we estimate the empirical probability (frequency) $\mathbb{P}[Z|S]$ in A_1. Subsequently, using the above $\mathbb{P}[Z|E,S]$ and $\mathbb{P}[Z|S]$, the BaBE method allows us to derive $\mathbb{P}[E|S]$ in A_1. Finally, by applying the Bayes theorem to the above probabilities, we derive $\mathbb{P}[E|Z,S]$ in A_1.

We note that the scenario we are considering is the same as that of machine learning (ML). Indeed, in machine learning, we assume the existence of a dataset (for instance, historical data), i.e., a representation of the joint distribution $\mathbb{P}[E,Z,S]$. In the case of ML, we typically derive directly the prediction of E for a given value of Z and S. However, it may happen that $\mathbb{P}[E|Z,S]$ depends on the distribution of $E|S$, which can vary greatly depending on the context. The effect of the distribution shift is a well known problem

---

[1]Note that E is what in causality is called *a mediator*.

in ML, impeding the deployment of the model in populations that are different from the one in the training data [32, 35].

In contrast, $\mathbb{P}[Z|E, S]$ may be more "universal", and this is exactly the case in which our BaBE method is applicable, also in case of a distribution shift (on E). In this case, it is convenient to invest in the estimation of $\mathbb{P}[Z|E, S]$, which can be done once and then transferred to different contexts. Indeed, one advantage of our approach is that it allows the transfer of causal knowledge. Namely, once we learn the relation $\mathbb{P}[Z|E, S]$, the method can be applied to a population with different proportions, i.e. different $\mathbb{P}[E|S]$ (but the same $\mathbb{P}[Z|E, S]$). For more discussion about this point, we refer to [1–3, 33, 36]. Another case in which our method presents an advantage over ML is when it is possible to estimate causal prior knowledge from experimental data, which is typically small. Machine learning algorithms need large data sets to achieve a good performance, whereas Bayesian statistics can be suitable also for small sample sizes [19, 29].

Once $\mathbb{P}[E|Z, S]$ is estimated, we pre-process the training data by assigning a decision Y based on the most likely value e of E, for given values of S and Z. If e does not have enough probability mass, however, we may not achieve CSP, or even a good approximation of it. In such case, we can base the decision on a threshold for the estimated E, aiming at achieving equal opportunity (EO) [18] instead, that we regard as a relaxation of CSP. Formally, EO is descibed as follows: $\mathbb{P}[\hat{Y} = 1|Y = 1, S = 1] = \mathbb{P}[\hat{Y} = 1|Y = 1, S = 0]$, where Y represents the "true decision", i.e., the decision based on a threshold for the real value of E.

We validate our method by performing experiments,[2] both on synthetic datasets and on the real 'The National Health and Nutrition Examination Survey' (*NHANES*) data set [13], featuring biological and chronological age of the patients. In both cases, we obtain a very good estimation of $\mathbb{P}[E|S]$, and we achieve a good level of both accuracy and fairness.

Summarizing, our contributions are as follows:

- We propose an approach to estimate the distribution of an explaining variable E, using the Expectation-Maximization method (EM). To the best of our knowledge, this is the first time that EM is used to achieve fairness without assuming the independence between E and the sensitive attribute S. From the above, we then derive an estimation of $\mathbb{P}[E|Z, S]$.
- Using the estimation of $\mathbb{P}[E|Z, S]$, we show how to to estimate the values of E and Y for each value of Z and S. These estimations are then used to *pre-process the data in order to achieve CSP and/or EO.*
- We show experimentally that our proposal outperforms other approaches for fairness, in terms of CSP, EO, accuracy, and other metrics for fairness and precision of the estimations.

Related Work. The notion of fairness that we consider in this work was introduced in [21] and it is known nowadays as *conditional statistical parity* (CSP) [9]. In [21], CSP is achieved through data pre-processing, by applying *local massaging* or *local preferential sampling* techniques. However, the authors consider only an explanatory variable E which is part of the data at the time of deployment of their method.

Note that our Z, although observable, cannot be considered as an explanatory variable, because we are assuming it is influenced by the sensitive attribute in a way that would make it unfair to base the decision on Z. To better understand the difference, consider one of the main examples used in [21] to illustrate the idea, which is a kind of *Berkeley admission anomaly*, an instance of the *Simpson paradox* [15]. In this example, the admittance in a certain university looks biased against females, but the disparity can actually be explained by the fact that female students tend to choose a more selective program. In this case, the explanatory variable is a mediator (the choice of the program), and it is assumed to be legitimate as a cause for disparity. By contrast, in our example the observed score is considered to be influenced by social discrimination, hence it cannot be directly used as an explanatory variable.

The work closest to ours is [6], where there is a model containing a latent variable whose distribution is discovered through the Expectation Maximization method. However, in [6] the notion of fairness considered is *statistical parity* (SP). Using SP as a constraint (thus applying a sort of *self-fulfilling prophecy* approach) and other constraints such as the preservation of the total ratio of positive decisions, the authors determine what the distribution $\mathbb{P}[Z|E, S]$ should be, they distribute the probability mass uniformly on all attributes, and they finally apply the EM method to determine the fair labels. In contrast, we are aiming at discovering what is the most probable value of E for each combination of values of the other attributes (S and Z), so as to take a fair decision based on E, considered as the explanatory variable. We do not require statistical parity, nor do we assume a uniform distribution on all attributes. Instead, we use external knowledge as prior knowledge for applying the EM method. Another difference is that they optimize accuracy with respect to the observed biased labels, whereas we consider accuracy towards the true fair label dependent on E, considered as the actual attribute on which the decision should be made.

Similar in spirit to [6], [26] tries to discover the latent variable which is maximally informative about the decision, while minimizing the correlation with the sensitive attribute (statistical disparity); this is done via a deep learning technique. Also [22, 25, 27] use deep learning latent variable models: [22, 25] consider latent confounders and [27] considers the sensitive attribute as a confounder. The situations in which these assumptions apply are quite different from the problem we study, since they aim at eliminating the effect of the confounder, while for us the unobservable variable is a mediator, and we want to use it as the basis for a fair decision. As a consequence, the notion of fairness those works aim at achieving is not suitable for our case. [7] introduces path-specific counterfactual fairness, where (among other cases) they consider the latent cause of a mediator between the sensitive attribute and the decision. This is more similar to our notion of fairness. However, [7] assumes that the latent variable is independent from the sensitive attribute; as such, their method is not directly applicable to our problem. [8] uses probabilistic circuits to impose statistical parity and to learn a relationship between the latent fair decision and other variables. Finally, [12] uses a notion of fairness called *disparate impact*, which is similar to statistical disparity, except that it is defined as a ratio (instead of a difference) between the probabilities of positive decisions for each group. Similarly to our work, [12] applies a corrective factor to the outcome of the observed variable Z, but their goal is to

---

[2]The software used for implementing our approach and for performing the experiments is available at https://github.com/BaBE-Algorithm/BaBE.

minimize the disparate impact (within a certain allowed threshold α), which is again in the spirit of minimizing statistical disparity. Also their technique is very different: they consider the distributions on the observed variable Z for each group, and they compute new distributions that minimize the earth movers' distance and achieve the threshold α. Then, they map each value of Z (for each group) on the new distribution so to maintain the percentile.

2 PRELIMINARIES AND NOTATION

$\hat{E}$, $\hat{Y}$ and Y notations. In this paper, $\hat{E}$ (with generic value $\hat{e}$) represents the estimation of the explanatory variable E. Similarly, $\hat{Y}_{\hat{E}}$ (with generic value $\hat{y}$) represents the estimation of the decision, based on $\hat{E}$, rather than the prediction of the model. To put it in context, recall that we are proposing a pre-processing method: $\hat{y}$ represents the value that we assign as decision in a sample of the training data during the pre-processing phase. The fairness and precision notions are defined with respect to these estimations. We use Y_Z to indicate the biased decision based on Z, and Y_E for the "true" decision based on E. When clear from the context, we may use Y instead of Y_E.

The Expectation-Maximization Framework. Let O be a random variable depending on an unknown parameter θ. Given that we observe $O = o$, the aim is to find the value of θ that maximizes the probability of this observation, and that therefore is *its best explanation*. To this purpose, we use the *log-likelihood function* $L(\theta) = \log \mathbb{P}[O = o|\theta]$. A Maximum-Likelihood Estimation (MLE) of the parameter is then defined as $\text{argmax}_\theta L(\theta)$ (which is the θ that maximizes $\mathbb{P}[O = o|\theta]$, since log is monotone). The Expectation-Maximization (EM) framework [10, 28, 37] is a powerful method for computing $\text{argmax}_\theta L(\theta)$.

2.1 Metrics for the quality of estimations

The Wasserstein distance. This distance is defined between probability distributions on a metric space. Let $\mathcal{X}$ be a set provided with a distance d, and μ, ν be two discrete probability distributions on $\mathcal{X}$. The Wasserstein distance between μ and ν is defined as

$$\mathcal{W}(\mu, \nu) = \min_\alpha \sum_{x,y \in \mathcal{X}} \alpha(x, y) \, d(x, y), \qquad (1)$$

where α represents a *coupling*, i.e., a joint distributions with marginals μ and ν satisfying the properties $\sum_{y \in \mathcal{X}} \alpha(x, y) = \mu(x)$ and $\sum_{x \in \mathcal{X}} \alpha(x, y) = \nu(y)$.

Accuracy. Let X, Y be two random variables with support $\mathcal{X}$ and $\mathcal{Y}$ respectively, and joint distribution $\mathbb{P}[X, Y]$. Let $f : \mathcal{X} \rightarrow \mathcal{Y}$ be a function that, given $x \in \mathcal{X}$, *estimates* the corresponding y, and let $\hat{y}$ be the result, i.e., $\hat{y} = f(x)$. The accuracy of f is defined as the expected value of $\mathbb{1}_{\hat{y}=y}$, that is the function that gives 1 if $\hat{y} = y$, and 0 otherwise. When the distribution is unknown, the accuracy is estimated empirically via a set of pairs $\{(x_i, y_i) \mid i \in \mathcal{I}\}$ independently sampled from $\mathbb{P}[X, Y]$ (*testing set*), and is defined as

$$Acc(\hat{Y}, Y) = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \mathbb{1}_{\hat{y}_i = y_i} \text{ where } \hat{y}_i = f(x_i). \qquad (2)$$

Distortion. If the variable to be predicted ranges over a metric space, and the metric is important for decision-making (like the case of E in our examples), accuracy is not always the best way to
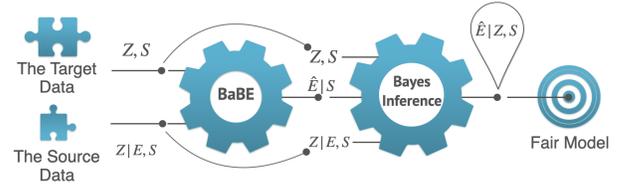
Figure 2: The pipeline of BaBE application. The variable E is observable in the source data and $\mathbb{P}[Z|E, S]$ can be derived. The target data is the one where E is not observable and we want to recover it using $\mathbb{P}[Z|E, S]$ derived from the source data. We input $\mathbb{P}[Z|E, S]$ and statistics from the observable variables in the target data to BaBE and estimate $\hat{E}$ consistent with the target distribution (possibly different than in the source data). We then again use $\mathbb{P}[Z|E, S]$ (from source data), observable variables (from the target data) and $\hat{E}$ (BaBE estimation) to inference $\hat{E}|Z, S$ for each sample in the target data.

measure the quality of the estimation. Arguably, it is more suitable to use the *distortion*, i.e., the expected distance between the true value and its estimation. Using the testing set $\{((z_i, s_i), e_i) \mid i \in \mathcal{I}\}$, the distortion in the estimation of E is defined as

$$Dist(\hat{E}, E) = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} |\hat{e}_i - e_i|, \text{ where } \hat{e}_i = f(z_i, s_i). \qquad (3)$$

2.2 Metrics for fairness

SP, CSP, and EO are rarely achieved, since they require a perfect match. It is therefore useful to quantify the level of (un)fairness, i.e., the difference between the two groups. We will use the following metrics:

Statistical parity difference (SPD)

$$\mathbb{P}[\hat{Y}_{\hat{E}} = 1|S = 1] - \mathbb{P}[\hat{Y}_{\hat{E}} = 1|S = 0]. \qquad (4)$$

Conditional statistical parity difference (CSPD)

$$\mathbb{P}[\hat{Y}_{\hat{E}} = 1|E, S = 1] - \mathbb{P}[\hat{Y}_{\hat{E}} = 1|E, S = 0]. \qquad (5)$$

Equal opportunity difference (EOD)

$$\mathbb{P}[\hat{Y}_{\hat{E}} = 1|Y_E = 1, S = 1] - \mathbb{P}[\hat{Y}_{\hat{E}} = 1|Y_E = 1, S = 0]. \qquad (6)$$

3 THE BABE METHOD

In this section we describe the BaBE approach. We briefly recall the problem: we have a data model represented in Figure 1, where S is the sensitive attribute, E is the explanatory variable on which a fair decision should be based, and Z is an observed but biased version of E. We need to estimate the distribution $\mathbb{P}[E|Z, S]$. The first step is to estimate the distribution of E for each group, $\mathbb{P}[E|S]$. We accomplish this task by adapting the Expectation-Maximization method to our particular setting. Then, from $\mathbb{P}[E|S]$ we derive, using the Bayes theorem, the estimation of $\hat{\mathbb{P}}[E|S, Z]$, from which we finally derive $\hat{E}$ and $\hat{Y}_{\hat{E}}$. The pipeline of the process is provided in Figure 2.

3.1 Deriving $\hat{\mathbb{P}}[E|S]$

We estimate the unknown parameter $\mathbb{P}[E|S]$ as the MLE of a sequence of samples $(\bar{z}, \bar{s}) = \{(z_i, s_i) \mid i \in [1, N]\}$,[3] assuming that we know the effect of the bias, i.e., $\mathbb{P}[Z|E, S]$. We denote by $\varphi_s[z, \bar{z}]$ the empirical probability of $Z = z$ given $S = s$, i.e., the frequency of z in the samples with $S = s$. Algorithm 1 estimates $\mathbb{P}[E|S]$ by starting with the uniform distribution and by iteratively computing at step t a new estimation $\hat{\mathbb{P}}[E|S]^{(t)}$ from the previous one, getting closer and closer to the MLE. The proof of correctness of Algorithm 1 is provided in [5] (the archival version of this paper).

We have experimentally verified that our method is quite efficient: The running time of Algorithm 1 on the data of Section 4 is a few seconds. Details are reported in the additional material.

Algorithm 1 BaBE: Bayesian Bias Elimination

Data: $\{(z_i, \mathbb{P}[Z = z_i|E = e, S = s])\}_{i \in \{1..N\}}$ and γ (an allowed error in estimating $\mathbb{P}[E = e|S = s]$)
Result: An approximation (up to γ) $\hat{\mathbb{P}}[E|S]$ of the MLE

Compute $\varphi_s[z, \bar{z}]$, for every $z \in \mathcal{Z}$ and $s \in \mathcal{S}$
$\hat{\mathbb{P}}[E = e|S = s]^{(0)} = \frac{1}{|\mathcal{E}|}$, for every $e \in \mathcal{E}$
$t = 0$
repeat
$\quad t = t + 1$
$\quad \hat{\mathbb{P}}[E = e|S = s]^{(t)} = \sum_{z \in \mathcal{Z}} \varphi_s[z, \bar{z}] \frac{\mathbb{P}[Z=z|E=e,S=s]\hat{\mathbb{P}}[E=e|S=s]^{(t-1)}}{\sum_{e' \in \mathcal{E}} \mathbb{P}[Z=z|E=e',S=s]\hat{\mathbb{P}}[E=e'|S=s]^{(t-1)}}$, for every $e \in \mathcal{E}$ and $s \in \mathcal{S}$
until $\forall e \in \mathcal{E} \, \forall s \in \mathcal{S}. \left| \hat{\mathbb{P}}[E = e|S = s]^{(t)} - \hat{\mathbb{P}}[E = e|S = s]^{(t-1)} \right| < \gamma$
return $\hat{\mathbb{P}}[E|S] = \hat{\mathbb{P}}[E|S]^{(t)}$

3.2 Deriving $\hat{\mathbb{P}}[E|Z, S]$ from $\hat{\mathbb{P}}[E|S]$

Given the data $\{(z_i, s_i) \mid i \in [1, N]\}$, the conditional distributions $\mathbb{P}[Z|E, S]$, and the estimation $\hat{\mathbb{P}}[E|S]$, we use the Bayes formula to estimate $\hat{\mathbb{P}}[E = e|Z = z, S = s]$ as

$$\hat{\mathbb{P}}[E = e|Z = z, S = s] = \frac{\mathbb{P}[Z=z|E=e,S=s]\hat{\mathbb{P}}[E=e|S=s]}{\mathbb{P}[Z=z|S=s]}$$

3.3 Deriving $\hat{E}$ and $\hat{Y}_{\hat{E}}$ from $\hat{\mathbb{P}}[E|Z, S]$

We propose two ways to derive $\hat{Y}_{\hat{E}}$ for pre-processing the samples in the training data, depending on how much probability mass is concentrated on the mode of $\hat{\mathbb{P}}[E|Z, S]$. We denote by τ the threshold for the values of E that qualify for the positive decision.

Method 1. Given z and s, if $\hat{\mathbb{P}}[E|Z = z, S = s]$ is unimodal and has a large probability mass (say, 50% or more) on its mode, then we can safely set $\hat{E}$ to be that mode. Namely, if $\max_e \hat{\mathbb{P}}[E = e|Z = z, S = s] \geq 0.5$ then we set $\hat{e} = \arg\max_e \hat{\mathbb{P}}[E = e|Z = z, S = s]$, and we can then use $\hat{e}$ directly to set $\hat{Y}_{\hat{E}} = 1$ or $\hat{Y}_{\hat{E}} = 0$ in those samples with $Z = z$ and $S = s$, depending on whether $\hat{e} \geq \tau$ or not, respectively. Our experimental results show that this method gives a good accuracy.

Method 2. If $\hat{\mathbb{P}}[E|Z = z, S = s]$ is dispersed on several values, so that no value is strongly predominant, then it is impossible to estimate individual values for E with high accuracy. However, we can still accurately estimate Y_E as follows: Let $\sigma_0 = \sum_{e<\tau} \hat{\mathbb{P}}[E = e|Z = z, S = s]$ and $\sigma_1 = \sum_{e \geq \tau} \hat{\mathbb{P}}[E = e|Z = z, S = s]$. If $\sigma_0 < \sigma_1$, then we set $\hat{Y}_{\hat{E}} = 1$; otherwise, $\hat{Y}_{\hat{E}} = 0$.

4 EXPERIMENTS

In this section, we test BaBE on scenarios corresponding to Examples 1 and 2, using synthetic data sets and a real data set respectively. We compare our results with those achieved by the following well-known pre-processing approaches that aim to satisfy statistical parity, as well as machine learning algorithms trained on the data set where E is observable.

4.1 Metrics

We will use the following metrics to measure fairness: Statistical parity difference (*SPD*, Equation 4), Conditional statistical parity difference (*CSPD*, Equation 5), Equal opportunity difference (*EOD*, Equation 6). The performance is measured by accuracy ($Acc(\hat{Y}, Y)$, Equation 2), distortion ($Dist(\hat{E}, E)$, Equation 3), and the Wasserstein distance between the true and estimated distributions ($\mathcal{W}(\mu, \nu)$, Equation 1).

4.2 Other Algorithms for Comparison

The first approach we compare with is the *disparate impact remover* (DI) [4, 12].[4] DI has a parameter λ, which represents the minimum allowed ratio between the probability of success ($\hat{Y} = 1$) of each group (hence $\lambda = 1$ corresponds to statistical parity). For the experiments, we use $\lambda = 0.8$.

The second algorithm we compare with ours is the *naive Bayes* (NB) [6].[5] NB also applies the EM method; however, in contrast to our work, NB assumes that E and S are independent, and uses EM to take decisions that optimize the trade-off between SPD and accuracy.

Finally, we compare the performance of BaBE with the ML methods that are trained on the data where E is observed (the source data). The model is then used to predict E (from S and Z) in the data sets where the distribution of $E|S$ is different from the source data, but the learned mechanism ($Z|E, S$) is the same. We used linear regression (LG) and decision tree regression (DT) for the experiments.[6]

4.3 Synthetic data sets with distributions shifts

This group of experiments is aimed at testing how BaBE copes with the transfer of knowledge to populations with different distributions. For this purpose, we generate a synthetic set, that we call "source data", where the mean of E for group 0, *mean0*, is 40 and the mean of E for group 1, *mean1*, is 80. The groups in this data set are about even in size. We use this set of "source data" to estimate the distributions $\mathbb{P}[Z|E, S]$.

Then, we generate three different data sets 1, 2 and 3 where *mean1* is still 80, while *mean0* varies from 40 to 80, representing a

---

[3]We use the notation $[a, b]$ to represent the integers from a to b.

[4]We use the implementation by [4].
[5]Implementation kindly provided by the authors of [6].
[6]We use scikit-learn implementations of the machine learning algorithms [34].
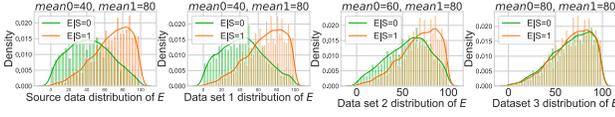
Figure 3: The distribution of $E|S$ in the source data and in the new populations.

distribution shift, w.r.t. the source data, on the E for group 0. Varying *mean0* will also allow us to validate the claim that our method works well regardless of E being independent of S or not. The percentage of the two groups in these new data sets also changes: we have set the group 1 to be 60% of the population, and, consequently, the group 0 to be 40%.

4.3.1 Generation of the synthetic data sets. In this section we explain how to generate various data sets containing tuples of the form (s_i, e_i, z_i, y_i). First, we generate a data set of 30K elements $\{s_i\}_{i \in [1,30K]}$ representing values for the sensitive variable (group) S, where each s_i is sampled from the Bernoulli distribution $\mathcal{B}(0.5)$. This means that the two groups are about even. Then, we set the domain of E to be equal to $[0, 99]$, and to each of the elements s_i in the sequence we associate a value e_i for the variable E, sampled from the normal distribution $\mathcal{N}(mean0, sd)$, if $s_i = 0$, and from $\mathcal{N}(mean1, sd)$, if $s_i = 1$,[7] where the mean *mean1* is set to 80, and the standard deviation sd is set to 30. On the other hand, the value of *mean0*, is 40 in the source data, and varies from 40 to 80 in the data sets 1, 2 and 3. Finally, to each pair (s_i, e_i) we associate a value z_i with Z by applying a bias to e_i. More precisely, $z_i = 100 \times sigmoid(e_i/10-5) - 100 \times sigmoid(e_i/10-5) \times 0.2 + \varepsilon$ for $S = 0$ and $z_i = 100 \times sigmoid(e_i/10-5) + 100 \times sigmoid(e_i/10-5) \times 0.02 + \varepsilon$ for $S = 1$, where ε is a noise term sampled from $\mathcal{N}(1, 0.05)$. The threshold for the decision is $E = 80$, namely: $Y = 1$ if $E > 80$ and $Y = 0$ otherwise. This is used to associate a decision y_i to each tuple (s_i, e_i, z_i). The distribution of $E|S$ for the source data and the data sets 1-3 are shown in Figure 3.

4.3.2 Application of BaBE. We use the source data to estimate $\mathbb{P}[Z|E, S]$. This conditional probability is then used to estimate $\mathbb{P}[E|S]$ in the other data sets, where it is different from the "source data", in the following way: We take a random subset of the data from the data sets 1-3 (80%), remove the E values from them, and use them to compute the empirical distribution $\mathbb{P}[Z|S]$ and to produce the estimate $\hat{\mathbb{P}}[E|S]$ by applying our BaBE method.

We verify that these data sets satisfy the conditions for Method 1 (cf. Section 3.3), and we apply this method to the remaining (20%) of the data (testing data sets) to infer the values of $\hat{E}$ and $\hat{Y}_{\hat{E}}$ for each sample. We compare our estimates to the true values of E and Y in the testing data.

Figure 4 shows the Wasserstein distances between the true distributions and the estimated ones. As we can see, BaBE manages to estimate E quite well: the distance w.r.t. E is very small.

Figure 5 shows the accuracy with respect to the true E (discretized values). BaBE is able to achieve a much better accuracy than other methods. DT algorithm is the second best performer;

---

[7]To keep the samples in the range of E, we re-sample the values that are lower than 0 or higher than 99. We also discretize them by rounding to the nearest integer.
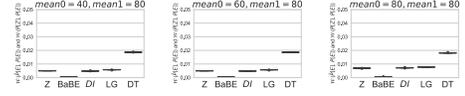
Figure 4: Experiments on the synthetic data sets: The Wasserstein distance between $\hat{\mathbb{P}}[Z]$ and $\mathbb{P}[E]$ and between $\hat{\mathbb{P}}[E]$ and $\mathbb{P}[E]$.

however DT is still performing worse than BaBE, despite being trained on the data set where E is observable.
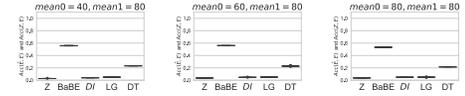
Figure 5: Experiments on the synthetic data sets: The accuracy between Z and E, and between $\hat{E}$ and E.

Figure 6 shows the accuracy with respect to Y for the two groups. Once again the performance of BaBE is better than other pre-processing methods. The overall performance of all methods is better than measuring the accuracy with respect to E. This is not surprising, as achieving good accuracy in a binary setting is an easier task. DT achieves almost the same accuracy as BaBE on the dataset 1 ($mean0 = 40, mean1 = 80$) which has the same distribution of $E|S$ as the training data. However, the performance of DT decreases on the data sets where the distribution of $E|S$ is different from the training data ($mean0 = 60$ and $mean0 = 80$).
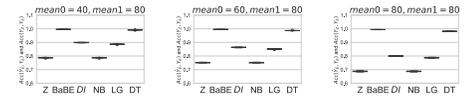
Figure 6: Experiments on the synthetic data sets: The accuracy between $\hat{Y}_Z$ and Y_E, and between $\hat{Y}_{\hat{E}}$ and Y_E.

Figure 7 shows the distortion (Equation 3). BaBE again produces results that are closer to the true values of E than the ones produced by other methods.
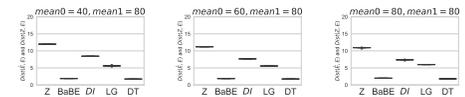
Figure 7: Experiments on the synthetic data sets: The distortion between Z and E and between $\hat{E}$ and E.

Figure 8 shows the conditional statistical parity difference on admission for each group, conditioned on E. The values for BaBE are close to zero, indicating the absence of discrimination. The DI method has decreased the discrimination with respect to Z. NB results are worse than the initial discrimination: it is possible that

the non linear bias function together with the accuracy constraints inbuilt in the algorithm has impeded its performance. DT once again is a second best performer after BaBE.
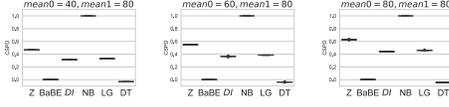
Figure 8: Experiments on the synthetic data sets: Conditional Statistical Parity Difference (CSPD). We recall that, for BaBE, DI, NB, LG and DT, the CSPD is defined as $\mathbb{P}[\hat{Y}_{\hat{E}} = 1|E, S = 1] - \mathbb{P}[\hat{Y}_{\hat{E}} = 1|E, S = 0]$. For Z, the definition is similar, with $\hat{Y}_{\hat{E}}$ replaced by Y_Z.

Figure 9 shows the probabilities of positive prediction when the true decision is positive, and the corresponding difference in equal opportunity. We note that the prediction based on Z has a high probability to be positive for the group 1, but not for the group 0, therefore EOD for Z is close to 1. On the other hand, BaBE's prediction is based on the estimation of E, and hence tends to be equal to the true decision yielding EOD close to zero. Quite surprisingly, DI gives bad results, even though it is supposed to equalize the distributions for $S = 0$ and $S = 1$. However, DI decreases the mean for $S = 1$ instead of increasing the mean for $S = 0$, leaving the values for $S = 0$ below the positive decision threshold (80). Similar considerations apply to NB and LG.
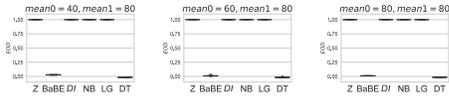
Figure 9: Experiments on the synthetic data sets: Equal Opportunity Difference (EOD). We recall that, for BaBE, DI, NB, LG and DT, the EOD is defined as $\mathbb{P}[\hat{Y}_{\hat{E}} = 1|Y_E = 1, S = 1] - \mathbb{P}[\hat{Y}_{\hat{E}} = 1|Y_E = 1, S = 0]$. For Z, the definition is similar, with $\hat{Y}_{\hat{E}}$ replaced by Y_Z.

Finally, Figure 10 compares the statistical difference (SPD) of the prediction $\hat{Y}_{\hat{E}}$ obtained with the various methods and the SPD of Y_Z. The SPD for $\hat{Y}_{\hat{E}}$ is defined in (4), for Y_Z is defined as $\mathbb{P}[Y_Z = 1|S = 1] - \mathbb{P}[Y_Z = 1|S = 0]$. When $mean0 = 80$, that is, the same as $mean1$, BaBE achieves, correctly, $SPD = 0$. In contrast, DI and NB do not achieve equal distribution for S=1 and S=0, which is surprising since the algorithms are geared towards equality. We hypothesize that the performance of the algorithms is impeded by the non linear bias function and in-built accuracy constraints. The DT algorithm is closest to the performance of BaBE, as it is more suitable to handle non-linearity in the data set than other ML model (LG).

4.4 The real-world data set

The National Health and Nutrition Examination Survey (NHANES) [13] is a series of studies that are intended to evaluate the health and nutritional status of adults and children in the United States.
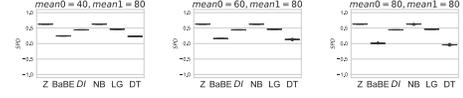
Figure 10: Experiments on the synthetic data sets: Statistical Parity Difference (SPD). We recall that, for BaBE, DI, NB, LG and DT, the SPD is defined as $\mathbb{P}[\hat{Y}_{\hat{E}} = 1|S = 1] - \mathbb{P}[\hat{Y}_{\hat{E}} = 1|S = 0]$. For Z, the definition is similar, with $\hat{Y}_{\hat{E}}$ replaced by Y_Z.

The survey is unique in that it incorporates in-depth interviews and detailed physical examinations. Health-related questions and demographics are included in the NHANES interview. For the survey, the sample was selected to represent the US population of all ages. To produce reliable statistics, NHANES oversamples individuals aged 60 and over, African Americans, and Hispanics. NHANES is a popular source for studying biological aging [23, 24, 30, 38]. The data set consists of 8243 samples. For our experiments, we use three variables from the data set, race (black or white), which is out S, chronological age (20-90), which is our Z, and an estimate of the biological age of the original KDM [8] biological age (variable 'kdm0') which is our E. We choose chronological or biological age 75 or more as the threshold to set $Y_Z = 1$ and $Y_E = 1$. This age group shows the most racial disparity in biological aging in the NHANES data set. Additionally, it is a reasonable age to check for age-related diseases or consider retirement.

Experiments on the NHANES data are carried out using Method 2 (cf. Section 3.3) of the BaBE method. This is because the conditional distribution of $Z|E, S$ does not allow the accurate estimation of every individual E. However, it still allows us to recover the aggregated distribution and estimate $\hat{Y}_{\hat{E}}$. In the experiments we consider only the fairness metric EODS, because the statistical disparity in the NHANES data is very small (owing to the oversampling of the minority population), so SPD is not interesting, and CSPD is not relevant because we apply Method 2.

The boxplots are obtained by repeating the experiments ten times with the same parameters. We report the results for the values of $mean0$ equal to 40, 60 and 80.

Figure 11 shows the accuracy resulting from the application of BaBE, DI, and NB to the NHANES data set. As we can see, BaBE achieves better overall accuracy and significantly better accuracy for $S = 1$.

Figure 12 shows the equal opportunity from the application of BaBE, DI, and NB to the NHANES data set. BaBE achieves EOD close to zero. DI and NB preprocessing methods do not differ significantly from the estimated EOD considering the original Z.

4.5 Discussion

Our experiments show that BaBE performs well for the fairness notions for which BaBE is designed, i.e., CSPD and EOD, while maintaining good accuracy.

BaBE performs well also when $\mathbb{P}[E|S]$ is different from that of the data in which $\mathbb{P}[Z|E, S]$ has been computed (Figure 4), which shows that BaBE is compatible with the transfer of causal knowledge to

---

[8]Klemera and Doubal's method for calculating the biological age from the set of biomarkers.
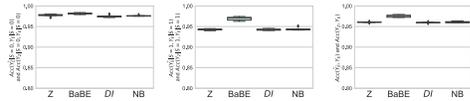
Figure 11: Experiments on the NHANES data. The Accuracy for the two groups separately, and overall.
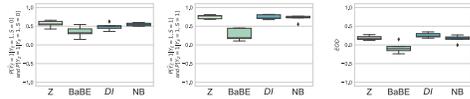
Figure 12: Experiments on the NHANES data. Equal opportunity difference (EOD), for the two groups separately, and overall.

populations with different distributions. On the contrary, DI and NB highly depend on the distribution as they always aim to minimize SPD. Note that minimizing SPD in the NHANES data set would still result in discrimination against black people, who on average have higher biological age than white people of the same chronological age.

The results of machine learning algorithms LG and DT show the sensitivity to the change in distribution of $E|S$. For example, the accuracy with respect to Y_E of LG and DT is highest on the data set 1, where $E|S$ is the same as in the training data ($mean0 = 40, mean1 = 80$) and degrades in the data sets 2 and 3, where it is different. In addition, LG performs worse the DT in all experiments. This is expected, because the relationship between Z and E is non linear. We note that BaBE is able to recover $\hat{E}$ without restrictions on the functional relationship in the data set. We acknowledge, that it is possible that a more complex machine learning algorithm could perform better on the proposed data set than linear regression, however it would imply higher computational cost. Moreover, it might still be affected by the distribution shift [32].

It is important to mention that the performance of BaBE is dependent on the invertibility of $\mathbb{P}[Z|E, S = s]$ (seen as stochastic matrix, aka bias matrix), because invertibility is necessary for the uniqueness of the MLE. However, even when the matrix is not invertible, we are able to obtain favorable results. Indeed, in all our experiments the bias matrices we produce from the synthetic data are not invertible, to mimic the more realistic scenarios. Preliminary experiments show that the diagonal deterministic matrix produces the highest precision for the estimation of distributions $\mathbb{P}[E|S]$, and highest accuracy of the prediction $\hat{Y}_{\hat{E}}$. We leave a more systematic study on how precision and accuracy depend on $\mathbb{P}[Z|E, S = s]$ as a topic for future work.

5   CONCLUSIONS AND FUTURE WORK

We have proposed BaBE, a framework to use knowledge of a biasing mechanism from domain-specific studies to perform data pre-processing, aiming at achieving Conditional Statistical Parity and Equal Opportunity when the explaining variable, and, consequently, the true decision, are not contained in the data. The BaBE

algorithm uses the bias mechanism to estimate the probability distributions of the explaining variable, and it performs equally well even when the population distributions are different from the ones in which the study of the bias was conducted. A distinguishing feature of our approach is that *we do not need to assume that the explaining variable is independent of the sensitive attribute*. One challenging direction for future work is to explore how the precision of the estimation, the accuracy of the prediction, and the fairness level depend on the form of the matrices $\hat{\mathbb{P}}[E|Z, S]$, and how the latter depends on the matrices representing the external knowledge (i.e., the bias mechanism) $\hat{\mathbb{P}}[Z|E, S]$.

We trust our method to serve as a tool to enhance interdisciplinary collaboration between domain experts and ML Fairness practitioners.

REFERENCES

[1] E. Bareinboim and J. Pearl. 2013. Causal Transportability with Limited Experiments. In *AAAI*.

[2] Elias Bareinboim and Judea Pearl. 2013. A General Algorithm for Deciding Transportability of Experimental Results. *Journal of Causal Inference* 1, 1 (may 2013), 107–134. https://doi.org/10.1515/jci-2012-0004

[3] Elias Bareinboim and Judea Pearl. 2014. Transportability from multiple environments with limited experiments: Completeness results. *Advances in neural information processing systems* 27 (2014).

[4] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63, 4/5 (2019), 4–1.

[5] Ruta Binkyte, Daniele Gorla, and Catuscia Palamidessi. 2023. BaBE: Enhancing Fairness via Estimation of Latent Explaining Variables. *arXiv preprint arXiv:2307.02891* (2023).

[6] Toon Calders and Sicco Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. *Data Min. Knowl. Discov.* 21 (09 2010), 277–292. https://doi.org/10.1007/s10618-010-0190-x

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

[8] YooJung Choi, Meihua Dang, and Guy Van den Broeck. 2020. Group Fairness by Probabilistic Modeling with Latent Fair Decisions. *CoRR* abs/2009.09031 (2020).

[9] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*. 797–806.

[10] A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Proceedings of the Royal Statistical Society* B-39 (1977), 1–38.

[11] 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*. 214–226.

[12] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Sydney, NSW, Australia) *(KDD '15)*. Association for Computing Machinery, New York, NY, USA, 259–268. https://doi.org/10.1145/2783258.2783311

[13] National Center for Health Statistics. NHANES. *https://www.cdc.gov/nchs/nhanes/about_nhanes.htm*.

[14] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2021. The (Im)possibility of fairness: different value systems require different mechanisms for fair decision making. *Commun. ACM* 64, 4 (2021), 136–143.

[15] Madelyn Glymour, Judea Pearl, and Nicholas P Jewell. 2016. *Causal inference in statistics: A primer.* John Wiley & Sons.

[16] Joshua Goodman, Oded Gurantz, and Jonathan Smith. 2020. Take Two! SAT Retaking and College Enrollment Gaps. *American Economic Journal: Economic Policy* 12, 2 (May 2020), 115–58. https://doi.org/10.1257/pol.20170503

[17] Brenda Hannon. 2012. Test Anxiety and Performance-Avoidance Goals Explain Gender Differences in SAT-V, SAT-M, and Overall SAT Scores. *Personality and individual differences* 53 (11 2012), 816–820. https://doi.org/10.1016/j.paid.2012.06.003

[18] Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Eds.), Vol. 29. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf

[19] Dirk Heerwegh. 2014. Small sample Bayesian factor analysis. *Phuse. Retrieved from http://www. lexjansen. com/phuse/2014/sp/SP03. pdf* (2014).

[20] Rashidul Islam, Shimei Pan, and James R Foulds. 2022. Fair Inference for Discrete Latent Variable Models. *arXiv preprint arXiv:2209.07044* (2022).

[21] Faisal Kamiran, Indrė Žliobaitė, and Toon Calders. 2013. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and Information Systems* 35, 3 (June 2013), 613–644. https://doi.org/10.1007/s10115-012-0584-8

[22] Matt J. Kusner, Joshua R. Loftus, Chris Russell 0001, and Ricardo Silva. 2017. Counterfactual Fairness. *CoRR* abs/1703.06856 (2017). http://arxiv.org/abs/1703.06856

[23] Dayoon Kwon and Daniel W Belsky. 2021. A toolkit for quantification of biological age from blood chemistry and organ function test data: BioAge. *GeroScience* 43 (2021), 2795–2808.

[24] Wen Liu, Jia Wang, Miao Wang, Huimin Hou, Xin Ding, Lingzhi Ma, and Ming Liu. 2023. Oxidative Stress Factors Mediate the Association Between Life's Essential 8 and Accelerated Phenotypic Aging: NHANES 2005-2018. *The Journals of Gerontology: Series A* (2023), glad240.

[25] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. 2017. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems* 30 (2017).

[26] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard S. Zemel. 2016. The Variational Fair Autoencoder. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1511.00830

[27] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2019. Fairness through causal awareness: Learning causal latent-variable models for biased data. In *Proceedings of the conference on fairness, accountability, and transparency*. 349–358.

[28] Geoffrey J McLachlan and Thriyambakam Krishnan. 2007. *The EM algorithm and extensions*. John Wiley & Sons.

[29] Daniel McNeish. 2016. On using Bayesian methods to address small sample problems. *Structural Equation Modeling: A Multidisciplinary Journal* 23, 5 (2016), 750–773.

[30] LM Nguyen, JJ Chon, EE Kim, JC Cheng, and JL Ebersole. 2022. Biological aging and periodontal disease: analysis of NHANES (2001–2002). *JDR Clinical & Translational Research* 7, 2 (2022), 145–153.

[31] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.

[32] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems* 32 (2019).

[33] Judea Pearl and Elias Bareinboim. 2014. External Validity: From Do-Calculus to Transportability Across Populations. *Statist. Sci.* 29, 4 (nov 2014). https://doi.org/10.1214/14-sts486

[34] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12 (2011), 2825–2830.

[35] Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. 2008. *Dataset shift in machine learning*. Mit Press.

[36] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. 2021. Towards Causal Representation Learning. *CoRR* abs/2102.11107 (2021).

[37] C. F. Jeff Wu. 1983. On the Convergence Properties of the EM Algorithm. *The Annals of Statistics* 11, 1 (1983), 95–103.

[38] Yanyan Xu, Xiaoling Wang, Daniel W Belsky, William V McCall, Yutao Liu, and Shaoyong Su. 2023. Blunted rest–activity circadian rhythm is associated with increased rate of biological aging: an analysis of NHANES 2011–2014. *The Journals of Gerontology: Series A* 78, 3 (2023), 407–413.