Causal Fair Machine Learning via Rank-Preserving Interventional Distributions

Ludwig Bothmann[1,2,*], Susanne Dandl[1,2] and Michael Schomaker[1,3]

[1]Department of Statistics, LMU Munich, Germany

[2]Munich Center for Machine Learning (MCML)

[3]Center for Infectious Disease Epidemiology, School of Public Health, University of Cape Town, South Africa

Abstract

A decision can be defined as fair if equal individuals are treated equally and unequals unequally. Adopting this definition, the task of designing machine learning models that mitigate unfairness in automated decision-making systems must include causal thinking when introducing protected attributes. Following a recent proposal, we define individuals as being normatively equal if they are equal in a fictitious, normatively desired (FiND) world, where the protected attribute has no (direct or indirect) causal effect on the target. We propose *rank-preserving interventional distributions* to define an estimand of this FiND world and a *warping method* for estimation. Evaluation criteria for both the method and resulting model are presented and validated through simulations and empirical data. With this, we show that our warping approach effectively identifies the most discriminated individuals and mitigates unfairness.

Keywords

fairness in ML, causal thinking, interventional distributions, stochastic interventions, rank-preserving interventions, quasi-individual fairness

1. Introduction

Automated decision-making (ADM) systems can support human decision-makers by predicting some variable of interest via a machine learning (ML) model. The data used for learning such ML models can have historical bias, i.e., show normatively undesirable discrimination against certain groups of protected attributes (PAs). When left unaddressed, this historical bias leads to biased ML models, generating fairness problems in ADM systems. The research field of fair machine learning (fairML) has quickly grown around this problem in recent years, giving birth to various "fairness metrics" (such as, e.g., demographic parity, see [1] for an overview).

A critique raised by [2] is that the question of *what fairness is* – as a philosophical concept – is rarely discussed. Hence, the proposed fairness metrics lack a philosophical justification, making it unclear which concept of fairness is measured by the respective metrics. They propose a consistent concept of fairness and outline how this should be integrated into the design of

✉ ludwig.bothmann@stat.uni-muenchen.de (L. Bothmann); susanne.dandl@stat.uni-muenchen.de (S. Dandl); michael.schomaker@stat.uni-muenchen.de (M. Schomaker)

🌐 https://www.slds.stat.uni-muenchen.de/people/bothmann/ (L. Bothmann)

🆔 0000-0002-1471-6582 (L. Bothmann); 0000-0003-4324-4163 (S. Dandl); 0000-0002-8475-0591 (M. Schomaker)

CEUR Workshop Proceedings (CEUR-WS.org)

ML models in ADM systems. Following Aristotle [3], they define a treatment as being fair "if equals are treated equally and if unequals are treated unequally". Furthermore, they distinguish between descriptively unfair treatment (which can occur without PAs) and normatively unfair treatment (which is a causal notion). For this, they conceive a *fictitious, normatively desired (FiND) world*, where the PA has no (direct or indirect) causal effect on the target variable. Individuals are normatively considered equal if they are equal in the FiND world.

We build upon this work by proposing concrete estimands and estimation procedures. As a starting point, we define a directed acyclic graph (DAG) that describes the causal relations in the real world. The DAG in the FiND world is then created by deleting all arrows that constitute paths from the PA to the target. We achieve this through specific stochastic interventions leading to *rank-preserving interventional distributions*. This intervention is rank-preserving in the following sense: Individuals of the disadvantaged group maintain the rank they have in the real world (compared with other individuals of the disadvantaged group) as population-wide rank in the FiND world (compared with all individuals), see Section 3.1.

After identifying the estimand, we propose a *warping method* for estimation that maps real-world data to a *warped world* which in turn approximates the FiND world, see Section 3.2. We call this a *quasi-individual* approach because individual "merits" are pulled through to the warped world. Finally, an ML model is trained on the warped data that can be used at prediction time after warping the new observation. Since final prediction models are trained and evaluated in the warped world, our approach can be considered to be a *pre-processing approach* [see 4, for a categorization of different approaches in fairML]. We propose evaluation metrics both for evaluating the warping method in a simulation study and for evaluating an ML model using warped data in an applied use case in Section 4. In a simulation study, we show that our warping method is able to approximate the FiND world, identify the most discriminated individuals, and eliminate the effects of the PA in the warped world (Section 5). Finally, we apply the proposed methodology to German Credit data, showing how to use our framework in practice (Appendix A).

2. Related Work

In addition to *group fairness* concepts (see, e.g., [5] for an overview), approaches of (non-causal) *individual fairness* have been proposed, starting with [6], who require that similar individuals should be treated similarly (see also [7, 8, 9]). An early notion of *causal fairness* was made by [10], who conceive a fictitious world where an individual belongs to a different subgroup of the PA, defining a decision as fair if it is equal in the real and fictitious world. For a thorough explanation of how this differs from our FiND world, see [2]. Including causality in the fairness debate and conceiving a fictitious world was also proposed by, e.g., [11, 12, 13, 14, 15, 16], where different ideas underlie those fictitious worlds. With the fairness concept introduced by [2], we distinguish the real world and the FiND world by the idea that in the FiND world, there must be no causal effects from the PA on the target – neither indirectly, nor directly. This means that we delete all arrows starting in the PA and eventually leading to the target variable (dashed arrows in Figure 1). This idea differs from what the literature on path-specific effects [e.g., 17, 18, 19] conceives. However, we believe that this more adequately captures the legal requirements of

many laws that demand that individuals must not be discriminated against based on the PA[1] – rendering it irrelevant which path in the DAG this effect follows.

3. Methods

As derived in [2], in order to derive the decision basis for fair decisions, we must conceive a "fictitious, normatively desired (FiND) world in which the PA has no causal effect" on the target variable, "neither directly nor indirectly". In the following, we adopt this idea, elaborate it further by concretely specifying causal and statistical estimands, and derive an estimation method, thereby building concrete and actionable algorithms for approximating the FiND world by what we call a "warped world" and for using Causal Fair ML (cfML) in applied use-cases.

Our method consists of four fundamental steps: (i) We first define the estimand as the joint distribution in the FiND world, described by stochastic interventions, leading to *rank-preserving interventional distributions* (see Section 3.1); (ii) We then estimate the joint/conditional distributions of interest in the FiND world, based on a specific g-formula type-of factorization that follows from specifying the respective identification assumptions and allows us to "warp" the real-world data into the warped world (see Section 3.2.1); With this, we can (iii) train an ML model (for predicting the target) in this warped world (see Section 3.2.2), and (iv) predict on a new observation in the warped world using the above warping models and ML model (see Section 3.3).

3.1. Estimand

3.1.1. DAGs in the Real and the FiND World

Deriving a DAG falls into the realm of *Causal Discovery* (see, e.g., [20] for a review of current methods). Since this is a notoriously hard challenge in practice, the alternative is to define the DAG with expert knowledge, as is typically done in epidemiology and medicine, where knowledge from human decision-makers is readily available [21]. In the remainder, we are agnostic to the question of how the DAG was constructed and will assume that all DAGs are correct; note that this may be an optimistic (and untestable) assumption and can hamper success in practice.

Two DAGs must be defined: the DAG in the real world and the DAG in the FiND world – where the PAs have no causal effect on the target. Figure 1 shows the two DAGs that we assume for the example of the German Credit data. Note that these DAGs are chosen for illustrative purposes and not because there is empirical evidence or expert knowledge that justifies exactly those DAGs. We reduced the feature set for a clearer presentation: Age (a confounder C) is the numerical age of an individual; Gender (the PA G) is assumed to be binary (classes *female* and *male*) in the remainder, but note that an extension on multi-categorical gender is methodologically straightforward; Savings (feature X_S) is a binary variable, indicating if the person has small savings (1) or not (0); Amount (feature X_A) is the amount of credit applied for; and Risk (target Y) is the binary risk category with values good (1) and bad (0).

---

[1]E.g., Charter of Fundamental Rights of the European Union: https://www.citizensinformation.ie/en/government-in-ireland/european-government/eu-law/charter-of-fundamental-rights/
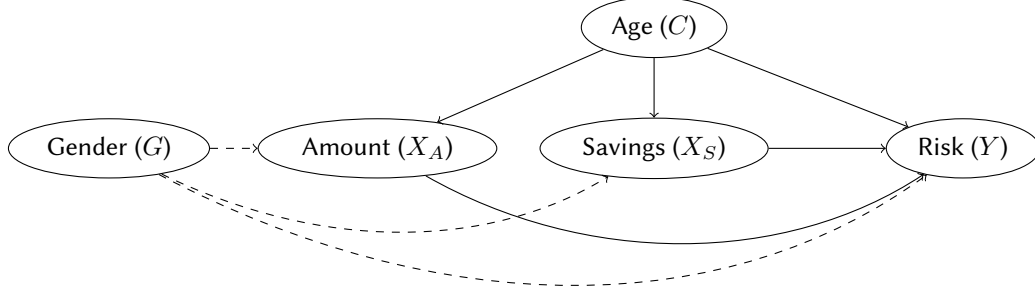
Figure 1: Assumed DAGs for credit risk example. In the FiND world, only solid arrows exist, and in the real world, all arrows exist.

3.1.2. Rank-Preserving Interventional Distributions

There are several possible interventions that can delete the dashed arrows in Figure 1 and, hence, lead to the FiND world DAG. We propose the following idea of "rank-preserving interventional distributions", which we believe to be the best way of defining those interventions when aiming to mitigate unfairness. We assume that the given DAGs (as shown in Figure 1) correctly mirror the causal relationships in both the real world and the FiND world. Slightly adapting the notation and terminology of [22], a general structural causal model (SCM) is given by

$$X_j := f_j(pa(X_j), U_j), \quad j \in \{1, \ldots, p\},$$

where $U_1, \ldots, U_p$ denote exogeneous independent random variables, and $pa(X)$ are parent nodes of X. In our example, the SCM in the real world (i.e., pre-intervention) is given by

$$G := f(U_G)$$
$$C := f(U_C)$$
$$X_A := f_A(G, C, U_A)$$
$$X_S := f_S(G, C, U_S)$$
$$Y := f_Y(G, C, X_A, X_S, U_Y),$$

which entails a joint distribution that can be factorized according to our working order:

$$P(Y, X_S, X_A, C, G) = P_Y(Y|X_S, X_A, C, G)P_S(X_S|C, G)P_A(X_A|C, G)P_C(C)P_G(G). \quad (1)$$

For the FiND world, we must make all descendants from the PA neutral w.r.t. the PA. We achieve this by a fictitious intervention rule d_p on the mediators and outcome only, i.e., no "modification" of the potentially sensitive PA is required (Eq. 2). This intervention leads to a joint post-intervention distribution $P_p(G, C, X_A^{d_p} X_S^{d_p}, Y^{d_p})$ in which the dashed arrows have been removed; thus, no effect of Gender on the mediators and the outcome exists – but the distributions of males and females are comparable and still in line with the data-generating process on which we want to train our ML model. Additionally, our suggested intervention is "rank-preserving" in the sense that the quantile of female customers within their strata is

transported into the FiND world (see Figure 2a). Thereby, all PA-dependent quantities are transformed into their FiND-world counterparts. Note that we can factorize the joint post-intervention in line with the pre-intervention distribution (Eq. 1), where the mediators and outcomes are replaced by their post-intervention counterparts. This leads to a g-formula type of factorization, which we can use for plug-in estimation of the relevant counterfactual distributions. A similar, quantile-based approach, can be found earlier in [23] which uses quantile regression forests for estimation.

$$
d_p = \begin{cases}
X_A^{(i)} = \tilde{x}_A^{(i)} & \text{where } \tilde{x}_A^{(i)} \text{ is the } (p_A^{(i)} \times 100)\% \text{ quantile of the conditional} \\
& \text{mediator distribution among the reference PA value, i.e.,} \\
& P(X_A \leq \tilde{x}_A^{(i)} | C = c^{(i)}, G = m) = p_A^{(i)}, \text{ and } p_A^{(i)} \text{ is determined} \\
& \text{by the pre-intervention quantile of unit } i, \text{ i.e.,} \\
& p_A^{(i)} = P(X_A \leq x_A^{(i)} \mid C = c^{(i)}, G = g^{(i)}). \\[2mm]
X_S^{(i)} = \tilde{x}_S^{(i)} & \text{where } \tilde{x}_S^{(i)} \text{ is the } (p_S^{(i)} \times 100)\% \text{ quantile of the conditional} \\
& \text{mediator distribution among the reference PA value, i.e.,} \\
& P(X_S \leq \tilde{x}_S^{(i)} | C = c^{(i)}, G = m) = p_S^{(i)}, \text{ and } p_S^{(i)} \text{ is determined} \\
& \text{by the pre-intervention quantile of unit } i, \text{ i.e.,} \\
& p_S^{(i)} = P(X_S \leq x_S^{(i)} | C = c^{(i)}, G = g^{(i)}). \\[2mm]
Y^{(i)} = \tilde{y}^{(i)} & \text{where } \tilde{y}^{(i)} \text{ is the } (p_Y^{(i)} \times 100)\% \text{ quantile of the counterfactual} \\
& \text{outcome distribution for the reference PA value, i.e.,} \\
& P(Y \leq \tilde{y}^{(i)} \mid \tilde{x}_A^{(i)}, \tilde{x}_S^{(i)}, C = c^{(i)}, G = m) = p_Y^{(i)}, \text{ and } p_Y^{(i)} \text{ is} \\
& \text{based on the pre-intervention quantile of unit } i, \text{ i.e.,} \\
& p_Y^{(i)} = P(Y \leq y^{(i)} \mid X_A = x_A^{(i)}, X_S = x_S^{(i)}, C = c^{(i)}, G = g^{(i)}).
\end{cases} \tag{2}
$$

3.2. Estimation

We base our estimation algorithm on the factorization derived above, i.e., we use the empirical distributions of both G and C to implement the intervention to sequentially obtain the respective post-intervention distributions of X_A, X_S, and Y. To determine the distributions and quantiles needed to facilitate the intervention implementation, our proposed algorithm uses the empirical distributions for the PA reference group (i.e., male customers) and a residual-based approach for the non-reference group (i.e., female customers). Alternatively, we could data-adaptively estimate the quantiles from the conditional distributions [24], but we do not pursue this more complicated approach further in this manuscript.

More generally, we approximate the FiND world by "warping" the target and the features affected by the PA (see Figure 2b). Once a preprocessed dataset containing "cleaned" features and target variables is available, standard ML techniques can be applied, prioritizing high predictive performance. This approach does not necessitate the incorporation of "fairness metrics" in the training process.[2] The three key steps of our proposed algorithm are:

---

[2]For a more in-depth exploration of the philosophical rationale and the potential implications regarding the intro-
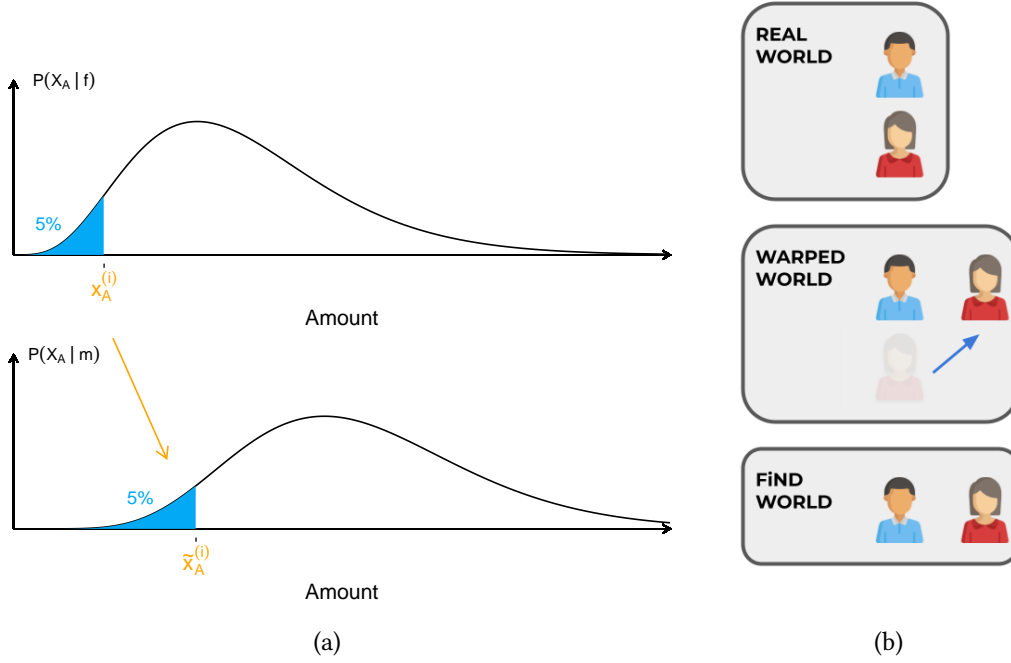
Figure 2: (a) Rank-preserving interventional distribution: Transform female observation to corresponding quantile in male distribution. (b) Real world discriminates between males and females, FiND world does not discriminate, warped world estimates FiND world by warping females to male level.

(1) Derive a warping from the real world to the warped world (see Section 3.2.1)
(2) Train and test an ML model using the warped data (see Section 3.2.2).
(3) At the time of prediction, use warping and trained ML model (see Section 3.3)

3.2.1. Warping for Approximating the FiND World

We propose to implement the interventions defined above (see Eq. 2) by the following residual-based estimation method. For determining the female intervention values, we must estimate – for each variable to be warped – (i) the individual probability rank of female i (e.g., $p_A^{(i)}$ for variable X_A) and (ii) the corresponding quantile of the male distribution (e.g., $\tilde{x}_A^{(i)}$). This means that we must estimate full distributions (not just location parameters) of $X_A | C = c, G = g$ for all values of c and g (analogously for X_S and Y), which becomes prohibitively complex in situations with finite data and numeric confounders C or features X_*. In our algorithm proposed below, we reduce estimation complexity by only estimating models for the location parameters of these distributions and derive individual probability ranks by using residuals of those models. The five steps of this warping algorithm are explained for feature Amount (X_A), warping for other variables works analogously:

(1) Estimate prediction models $\pi_A^f(C)$ for the female and $\pi_A^m(C)$ for the male population, where we are agnostic on the model class and can choose any ML model, since we only rely on point

---

duction of unfairness to the ADM system utilizing the trained ML model, please refer to [2].

predictions and model residuals on training data.

(2) Compute residuals as

$$r_f^{(i)} = \pi_A^f(c^{(i)}) - x_A^{(i)} \; \forall i \in I_f,$$
$$r_m^{(i)} = \pi_A^m(c^{(i)}) - x_A^{(i)} \; \forall i \in I_m,$$

where I_f and I_m are the female and male index set, respectively.

(3) Compute the individual probability rank of female i as ranked within the female residuals, i.e., telling us how "exceptionally high or low" her value is in comparison to other females, by

$$p_f^{(i)} = \frac{|\{j \in I_f : r_f^{(j)} \leq r_f^{(i)}|\}}{|I_f|}.$$

(4) Set $q_m^{(i)}$ to the empirical $p_f^{(i)}$-quantile of the residuals of the male model π_A^m, i.e.,

$$q_m^{(i)} = \min\{r \in R_m : \frac{|\{j \in R_m : j \leq r\}|}{|R_m|} \geq p_A^{(i)}\},$$

where $R_m = \{r_m^{(i)} : i \in I_m\}$ is the set of male residuals.

(5) Warp $x_A^{(i)}$ to the sum of male prediction and warped residual, i.e.,

$$\hat{x}_A^{(i)} = \pi_A^m(c^{(i)}) + q_m^{(i)}.$$

Analogously, we can warp X_S and Y (where in the latter case, warped values of Amount and Savings must be plugged into the male prediction in step (5)). However, note that for warping of non-continuous variables (such as Savings and Risk), we define the models to predict the probability scores, not the hard labels. That way, the warped values, e.g., $\hat{x}_S^{(i)}$, are no longer binary, but may be $\in [-1, 2]$. If we need hard labels – e.g., for learning a binary prediction model, such as for the target variable Y in Section 3.2.2 – we can simply threshold these scores. On the other hand, for use in further warping steps (such as for warping of $y^{(i)}$), we can directly use the raw values by plugging them into the prediction function, thereby pulling through finer information than if we would threshold earlier in the process.

Now, we have warped all Gender-dependent quantities $(x_S^{(i)}, x_A^{(i)}, y^{(i)})$ of female individuals to their warped world counterparts $(\hat{x}_S^{(i)}, \hat{x}_A^{(i)}, \hat{y}^{(i)})$, approximating their FiND world counterparts $(\tilde{x}_S^{(i)}, \tilde{x}_A^{(i)}, \tilde{y}^{(i)}) \; \forall i \in I_f$. To have a complete warped world data set $\mathcal{D}_w = \left(\left(\hat{\mathbf{x}}^{(1)}, \hat{y}^{(1)}\right), \ldots, \left(\hat{\mathbf{x}}^{(n)}, \hat{y}^{(n)}\right)\right)$, we set warped male values and values of non-warped features (e.g., Age) to their real-world value. Additionally to having warped the training data, we have also estimated warping functions that can be applied for new test data at the time of prediction.

3.2.2. Training ML Models in the Warped World

We can now use the warped world data $\mathcal{D}_w$ to train a prediction model for the warped target $\hat{Y}$. Assuming that the warping cleaned the data from any PA discrimination, we do not have to

account for any fairness metrics in this training step but can just focus on training a model that has high predictive performance. Since we assume that all Gender-related discrimination was eliminated through the warping, we do not use Gender G as a feature in this model (see Section 5 for an investigation of what happens if this assumption is wrong, e.g., due to a misspecified DAG). As a result, we obtain a trained model $f(\hat{\mathbf{x}})$ which can be used for prediction.

3.3. Prediction

Warp New Data. Consider a new observation $\mathbf{x}^* = (g^*, x_A^*, x_S^*, c^*)$. If this is a male observation, no warping must be done; if this is a female observation, we use the estimated warping functions of Section 3.2.1 as follows for X_A and analogously for X_S (but not for Y):
  (1) Compute individual residual r_f^* w.r.t. female model $\pi_A^f(c^*)$ as $r_f^* = \pi_A^f(c^*) - x_A^*$.
  (2) Compute individual probability rank p_f^* w.r.t. female population I_f as above.
  (3) Set q_m^* to the empirical p_f^*-quantile of training data residuals of male model π_A^m as above.
  (4) Warp x_A^* to the sum of male prediction and warped residual, i.e., $\hat{x}_A^* = \pi_A^m(c^*) + q_m^*$.
After carrying out the same steps for warping X_S, we finally obtain the warped observation $\hat{\mathbf{x}}^* = (\hat{x}_A^*, \hat{x}_S^*, c^*)$ (recall that we do not use Gender as a feature in the prediction model).

Predict New Data. For predicting the target in the warped world $\hat{y}^*$, we plug the warped observation $\hat{\mathbf{x}}^*$ into the prediction model trained on the warped world data, i.e., $\hat{y}^* = f(\hat{\mathbf{x}}^*)$.

4. Evaluation

We propose evaluation criteria that can be used for two purposes: Section 4.1 describes how to evaluate our proposed warping method for rank-preserving interventional distributions in a simulation study. Section 4.2 describes how the warped data and resulting ML models can be evaluated in an applied use-case. We denote with $\hat{\pi}^{(i)}$, $\hat{\phi}^{(i)}$, and $\hat{\psi}^{(i)}$ the predicted target of individual i in the real, warped, and FiND world, respectively.

4.1. Evaluation of Warping Method

We can evaluate our warping method w.r.t. (i) the warped data, asking, e.g., if the FiND world is recovered by the warping and w.r.t (ii) the final ML model – using the warped data.

(W1) Recovering FiND world. In a simulation study we can compare the warped and the FiND world distributions for investigating if the warping procedure recovers the FiND world. For numerical features, we compare warped world and FiND world empirical distributions by Kolmogorov-Smirnov (KS) tests, and for binary features, we use binomial tests. Additionally, we use a t-test to test the null hypothesis that there is no discrimination in the warped world between male and female subgroups w.r.t. risk predictions. If the method works, p-values of these tests should be consistently high, indicating that the null hypotheses cannot be rejected.

(W2) Identifying strongest discriminated individuals. In addition to the population-wide perspective of (W1), we are interested in the individual perspective, i.e., if the warping method also recovers the individual ranks of the FiND world w.r.t. the target variable prediction. If this would be the case, we could identify individuals who are most strongly affected by discrimination in the real world by comparing real world and warped world predictions in an applied use case. For the warped class of the PA, we compute individual risk prediction differences between the real world and the warped world, $d_1^{(i)} = \hat{\pi}^{(i)} - \hat{\phi}^{(i)}$ and between the real world and the FiND world, $d_2^{(i)} = \hat{\pi}^{(i)} - \hat{\psi}^{(i)}$, respectively. We use a t-test to test the null hypothesis that the means of these differences are equal. If the method works, p-values of these tests should be consistently high, and differences $d_1^{(i)} - d_2^{(i)}$ should be small. Correlation between ranks of $d_1^{(i)}$ and $d_2^{(i)}$ should be high, too.

4.2. Evaluation in an Applied Use-Case

How can the model be evaluated in an applied use case, i.e., how can we know if the warping method worked and if it removed unfairness? In our opinion, this cannot be answered by evaluating the final ML model w.r.t some "classical" fairML metrics.[3] Once we have successfully warped the data from the real to the warped world (approximating the FiND world), we reduced the problem to finding a model with good predictive performance. However, we can train models in the real and the warped world and then compare their behavior:

(UC1) Comparing performance. Test performance of the ML models in the real world $\hat{\pi}(\cdot)$ and the warped world $\hat{\phi}(\cdot)$ can be compared, assuming that both models fit "their" world equally well. However, this must not be misinterpreted as either of those models being better than the other one, as the models are merely modeling different worlds.

(UC2) Comparing predictions and identifying strongest discriminated individuals. For each individual i, the predictions in the real and the warped worlds can be compared by computing the difference $d_1^{(i)}$. As for (W2), this analysis can reveal individuals that are discriminated most in the real world (either positively or negatively). Additionally, these differences can be aggregated on the subgroup level, and tests can be computed to test the null hypothesis that predictions do not change between the two worlds for the respective subgroup.

(UC3) Identifying strongest warped individuals. We can also ask which individuals are affected most by the warping. These individuals' feature vectors have the largest distance between the real and the warped world, i.e., we can compare $\mathbf{x}^{(i)}$ and $\tilde{\mathbf{x}}^{(i)}$ by a suitable distance metric for each individual $i \in \{1, \ldots, n\}$.

---

[3]As elaborated on thoroughly in [2] (main points were summarized above), these kinds of metrics (such as demographic parity, equalized odds, etc.) do not reflect a clearly defined concept of fairness and, hence, are not suitable for deciding if an ML model entails unfairness. However, as readers might still be interested in the respective values – in the sense of a descriptive or explorative analysis –, we provide the resulting metrics in Appendix B.

(UC4) Identifying important features. For each feature, we can compare the empirical distributions in the real and the warped world, i.e., of X_j and $\tilde{X}_j$ for each $j \in \{1, \ldots, p\}$. We compute distances for each (normalized) feature and, thereby, can identify features that vary most between the two worlds, indicating that these features carry most of the real-world discrimination w.r.t. the PA.

5. Simulation Study

For investigating the behavior of the proposed method, we first conduct a simulation study where we know the true DAG in both the real and the FiND world. Subsequently, we apply the methods to the German Credit dataset in Appendix A. We seek to answer the following **research questions**:

(RQ1) Does our warping method work as expected? In other words, does this method recover the distributions in the FiND world (W1), and is it able to correctly identify the individual ranks of the target in the FiND world (W2)?

(RQ2) How does misspecification of the DAG affect the results?

(RQ3) What effects does the direction of warping have on performance (e.g., if subgroup A of the PA is warped to subgroup B, versus the other way around)?

5.1. Simulation Setup

Data simulation setup. We simulate data from the DAGs depicted in Figure 1. Here, the real-world data simulation contains all arrows, while the FiND world data simulation only contains solid arrows by setting Gender to *male* for all observations, thereby eliminating the Gender effect. The distributions utilized here are (left: real-world, right: FiND world):[4]

$$
\begin{array}{ll}
G \sim \mathrm{B}(\pi_G) & G \sim \mathrm{B}(\pi_G) \\
C \sim \mathrm{Ga}(\alpha_C, \beta_C) & C \sim \mathrm{Ga}(\alpha_C, \beta_C) \\
X_A | C, G \sim \mathrm{Ga}(\alpha_A(C, G), \beta_A(C, G)) & \tilde{X}_A | C \sim \mathrm{Ga}(\alpha_A(C, m), \beta_A(C, m)) \\
X_S | C, G \sim \mathrm{B}(\pi_S(C, G)) & \tilde{X}_S | C \sim \mathrm{B}(\pi_S(C, m)) \\
Y | X_A, X_S, C, G \sim \mathrm{B}(\pi_Y(X_A, X_S, C, G)) & \tilde{Y} | \tilde{X}_A, \tilde{X}_S, C \sim \mathrm{B}(\pi_Y(\tilde{X}_A, \tilde{X}_S, C, m)),
\end{array}
$$

where we use linear combinations of the features combined with a log- and logit-link for the Gamma and Binomial models, respectively, and mirror the Gender distribution of the German Credit data with $\pi_f = 31\%$ females. We perform $M = 1,000$ simulations on data sets of size $N_{tr} = 10,000$ for training and of size $N_{te} = 1,000$ for test, for each world, using the same seed for the two worlds to ensure comparability. Note that Gender and Age are then identical in both worlds, and only the descendants of Gender have differing values. We refer to this setup as (SIM1). To answer the misspecification behavior question (RQ2), we modify the simulation slightly by sampling Age from $C \sim \mathrm{Ga}(\alpha_C(G), \beta_C(G))$ but ignoring this effect for warping. We refer to this setup as (SIM2).

---

[4]Concrete values can be found in `simulation_study.R` in https://github.com/slds-lmu/paper_2023_cfml

Warping and prediction models. For warping models, we estimate models following the same distributional assumptions as in the simulation, i.e., by estimating the parameter vectors of the Gamma and logistic regressions separately for male and female observations of the training data. With these models, we apply the above warping strategy. As prediction models for the target variable, we train logistic regression models on the training data, warped training data, and FiND world training data, separately.

5.2. Results

With these models, we can now answer the above research questions (using a significance threshold of $\alpha = 5\%$ for all tests):

(RQ1) Figure 3a shows the distribution of Amount X_A in the different worlds for male and female observations, aggregated over all iterations of the simulation study. The null hypothesis of equal distributions in the warped and the FiND world is only rejected in $0\%, 0.4\%, 0\%$ of the iterations for Amount, Savings, and Risk, respectively. The mean difference between male and female risk predictions in the real world is 0.1122 (95% CI: $(0.1117, 0.1127)$). In the warped world, this is reduced to -0.0016 $(-0.0021, -0.0012)$, meaning that even if the difference between subgroups is still significantly non-zero, it is smaller by a factor of 70, i.e., we effectively reduced PA discrimination (and the direction switched from positive to negative).

Investigating individual predictions (W2), we see that correlations between ranks in the warped and the FiND world are high (0.892). Figure 3b shows individual risk prediction differences between the real world and the warped world as well as between the real world and the FiND world for females in one iteration (males are identical in the real world and the FiND world). The most strongly negatively affected individuals are at the upper end of the distribution. As shown, the most discriminated individuals (large difference between the FiND world and the real world) are correctly identified (large difference between the warped world and the real world). In 81% of iterations, the null hypothesis of equal differences d_1 and d_2 cannot be rejected. In cases with $p < 0.05$, the mean difference is -0.0023 – meaning that the deviation between the warped world and the FiND world is also minimal in these cases. We conclude that warping (i) recovers the marginal distributions in the FiND world, (ii) diminishes discrimination to a very small value, and (iii) correctly identifies the most discriminated individuals.

(RQ2) The null hypothesis of equal distribution in the warped world and the FiND world is rejected in $17\%, 4\%, 0\%$ of the iterations for Amount, Savings, and Risk, respectively. In the real world, the mean difference between risk predictions is 0.1723 $(0.1718, 0.1728)$, which is higher than in (SIM1). In the warped world, this is reduced to 0.0355 $(0.0350, 0.0360)$ (reducing by a factor of 4.9) – far less than above. The correlation of ranks (0.9518) is higher than above, since discrimination in the FiND world is higher in (SIM2). In 6.9% of iterations, the null hypothesis of equal differences d_1 and d_2 cannot be rejected. In cases with $p < 0.05$, the mean difference is 0.026 – far higher than above. We conclude that misspecification of the DAG is a relevant factor for degrading the performance of our approach.

(RQ3) By switching the warping direction and warping male to female values, we observe the following: Recovering of marginal FiND world distributions is equally successful as when warping female to male values. The null hypothesis is rejected in $0\%, 0.3\%, 0\%$ of the iterations for Amount, Savings, and Risk, respectively (see also Figure 3c). The mean difference between

Real, warped, and FiND world

(a)

Risk prediction differences

(b)

Real, warped, and FiND world

(c)

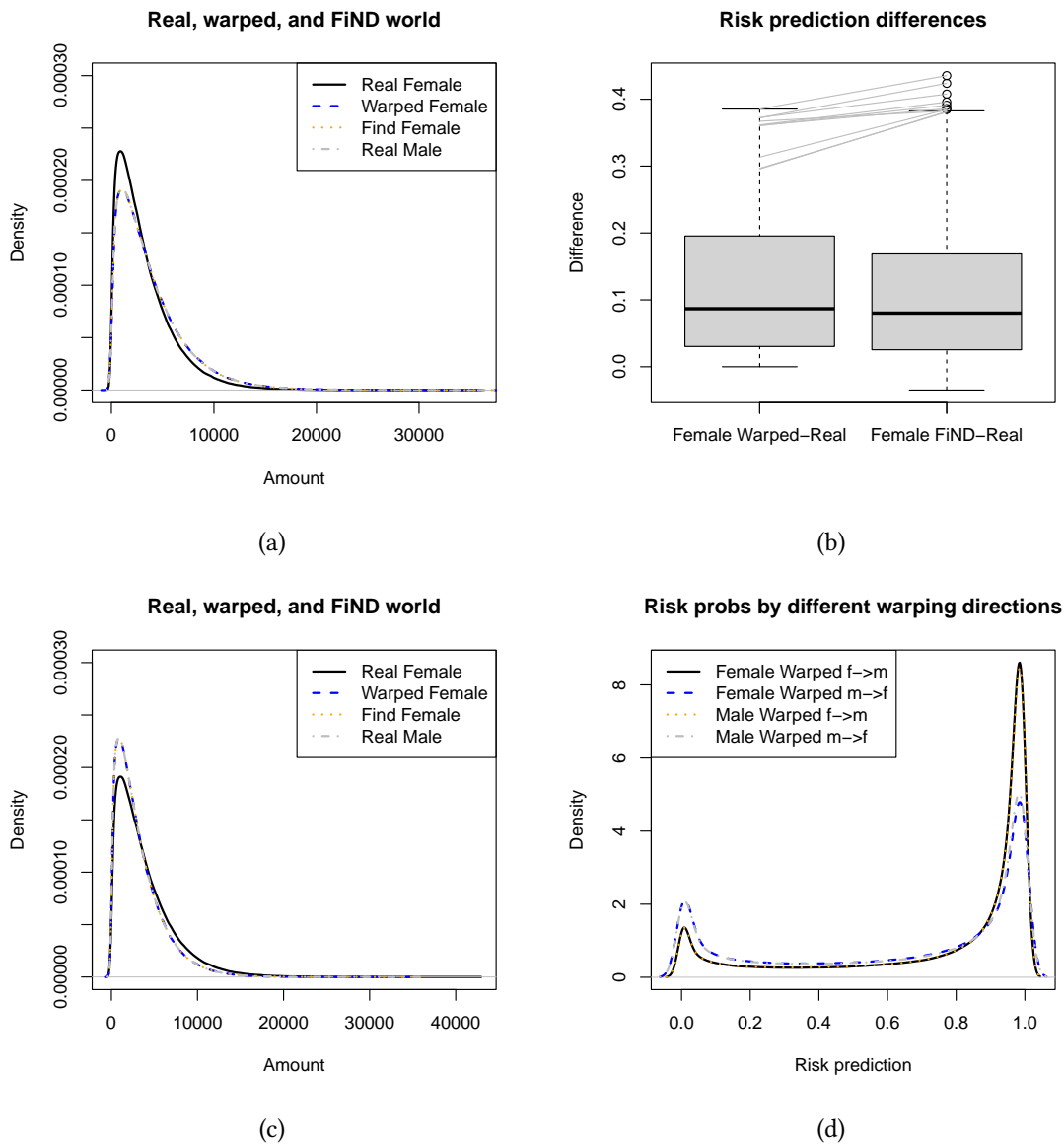## Risk probs by different warping directions

(d)

Figure 3: (a) Distribution of Amount, (b) most discriminated individuals, (c) distribution of Amount for reverse warping, i.e., warping males to females, (d) risk predictions for different warping directions.

risk predictions in the warped world is reduced to 0.0065 $(0.0060, 0.0071)$ – slightly worse than in the analysis of RQ1, which is due to the imbalance of the data. The mean correlation of ranks compared with ranks of RQ1 is high (0.9595), meaning that individual ranks are comparable for both warping directions. In 34% of iterations, the null hypothesis of equal differences d_1 and d_2 cannot be rejected. In cases with $p < 0.05$, the mean difference is 0.0073, meaning that the deviation between the warped world and the FiND world is also minimal in these cases (although a bit higher than in the analysis of RQ1, due to data imbalance). This means we can

also mitigate discrimination and preserve individual ranks with changing the warping direction. Most interestingly, the general level of the risk predictions changes, as shown in Figure 3d. This also makes sense, since we are now warping male to female values.

5.3. Discussion

We have shown that for the simulation setup above, our proposed method works as expected, recovering the marginal distributions in the FiND world and individual ranks; direction of warping does not make a relevant difference. However, as this is just an initial study, these investigations should be extended by follow-up work. As subsequent investigations, we would propose to (at least): (i) consider other, diverse DAGs, (ii) compare different ML models for warping and target prediction, and (iii) investigate behavior on other empirical data sets.

A general limitation of our method is that it depends on knowing the true DAG. As shown in RQ2, misspecifying the DAG degrades the performance of the method. Hence, special care should be given to identifying the true DAG in an applied use case by strongly connecting expert knowledge on the subject matter and rigorous application of causal discovery methods.

Practical feasibility: The computational costs of our method are rather small for the presented analysis. The computations for the German credit data (learning warping models, training models in both worlds, warping, and predicting in the warped world) took 0.17 seconds on a 3,4 GHz Intel Core i5, using one core of the CPU. This increases with, e.g., (i) the data size, (ii) the complexity of the DAG (since more models have to be trained), and (iii) the ML models used (e.g., training a multilayered neural networks takes more time than training the logit models used here). We do not expect the computational time to be a relevant constraint – also because learning the different models for warping can be parallelized. From a practical viewpoint, the much more important challenge is finding the DAG, since this involves carefully interweaving expert knowledge and application of methods of causal discovery [see, e.g., 20].

6. Conclusion and Outlook

We have presented rank-preserving interventional distributions as a framework to identify a FiND world where no causal effects of a PA exist. Additionally, we have proposed a warping method for estimating FiND world distributions with real-world data. A simulation study showed that the method works for the investigated simulation setup (see Section 5.3 for limitations), and we demonstrated in the Appendix how the method can be applied to empirical data (Appendix A). Analyses can be reproduced via a public GitHub repository, which also contains code for applied use cases.[5] Apart from extending the study as outlined in Section 5.3, further work should compare our method to other methods that conceive a fictitious world for tackling fairness issues of ML models (see references in Section 2).

Acknowledgments

References

[1] S. Barocas, M. Hardt, A. Narayanan, Fairness and Machine Learning, 2019. URL: http://www.fairmlbook.org.

[2] L. Bothmann, K. Peters, B. Bischl, What Is Fairness? Philosophical Considerations and Implications For FairML, 2023. doi:10.48550/arXiv.2205.09622.

[3] Aristoteles, Aristotelis Opera, volume 2 of *ex rec. Immanuelis Bekkeri ed. Acad. Regia Borussica*, de Gruyter, 1831. Book V.

[4] S. Caton, C. Haas, Fairness in Machine Learning: A Survey, ACM Computing Surveys (2023) 3616865. doi:10.1145/3616865.

[5] S. Verma, J. Rubin, Fairness definitions explained, in: Proceedings of the International Workshop on Software Fairness, ACM, Gothenburg Sweden, 2018. doi:10.1145/3194770.3194776.

[6] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel, Fairness through awareness, in: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, Association for Computing Machinery, New York, NY, USA, 2012, pp. 214–226. doi:10.1145/2090236.2090255.

[7] Y. Bechavod, C. Jung, S. Z. Wu, Metric-Free Individual Fairness in Online Learning, in: H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 11214–11225. URL: https://proceedings.neurips.cc/paper/2020/file/80b618ebcac7aa97a6dac2ba65cb7e36-Paper.pdf.

[8] A. Chouldechova, A. Roth, The Frontiers of Fairness in Machine Learning, arXiv:1810.08810 (2018). doi:10.48550/arXiv.1810.08810.

[9] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, On the (im)possibility of fairness, arXiv:1609.07236 (2016). doi:10.48550/arXiv.1609.07236.

[10] M. J. Kusner, J. Loftus, C. Russell, R. Silva, Counterfactual Fairness, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

[11] J. Zhang, E. Bareinboim, Equality of Opportunity in Classification: A Causal Approach, in: Advances in Neural Information Processing Systems, volume 31, Curran Associates, Inc., 2018. URL: https://proceedings.neurips.cc/paper_files/paper/2018/hash/ff1418e8cc993fe8abcfe3ce2003e5c5-Abstract.html.

[12] J. Zhang, E. Bareinboim, Fairness in Decision-Making — The Causal Explanation Formula, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2018. doi:10.1609/aaai.v32i1.11564.

[13] R. Nabi, I. Shpitser, Fair inference on outcomes, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI'18/IAAI'18/EAAI'18, AAAI Press, New Orleans, Louisiana, USA, 2018, pp. 1931–1940. doi:10.5555/3504035.3504270.

[14] R. Nabi, D. Malinsky, I. Shpitser, Learning Optimal Fair Policies, in: Proceedings of the

36th International Conference on Machine Learning, PMLR, 2019, pp. 4674–4682. URL: https://proceedings.mlr.press/v97/nabi19a.html.

[15] R. Nabi, D. Malinsky, I. Shpitser, Optimal Training of Fair Predictive Models, in: Proceedings of the First Conference on Causal Learning and Reasoning, PMLR, 2022, pp. 594–617. URL: https://proceedings.mlr.press/v177/nabi22a.html.

[16] S. R. Pfohl, T. Duan, D. Y. Ding, N. H. Shah, Counterfactual Reasoning for Fair Clinical Risk Prediction, in: Proceedings of the 4th Machine Learning for Healthcare Conference, PMLR, 2019, pp. 325–358. URL: https://proceedings.mlr.press/v106/pfohl19a.html.

[17] S. Chiappa, Path-Specific Counterfactual Fairness, Proceedings of the AAAI Conference on Artificial Intelligence 33 (2019) 7801–7808. doi:10.1609/aaai.v33i01.33017801.

[18] Y. Chikahara, S. Sakaue, A. Fujino, H. Kashima, Learning Individually Fair Classifier with Path-Specific Causal-Effect Constraint, in: Proceedings of The 24th International Conference on Artificial Intelligence and Statistics, PMLR, 2021, pp. 145–153. URL: https://proceedings.mlr.press/v130/chikahara21a.html.

[19] N. Weinberger, Path-Specific Effects, The British Journal for the Philosophy of Science 70 (2019) 53–76. doi:10.1093/bjps/axx040.

[20] A. R. Nogueira, A. Pugnana, S. Ruggieri, D. Pedreschi, J. Gama, Methods and tools for causal discovery and causal inference, WIREs Data Mining and Knowledge Discovery (2022). doi:10.1002/widm.1449.

[21] M. A. Hernán, J. M. Robins, Causal Inference: What If, Boca Raton: Chapman & Hall/CRC, 2020.

[22] J. Pearl, Causality: Models, Reasoning and Inference, 2nd ed., Cambridge University Press, 2009. URL: https://yzhu.io/courses/core/reading/04.causality.pdf.

[23] D. Plečko, N. Meinshausen, Fair Data Adaptation with Quantile Preservation, Journal of Machine Learning Research 21 (2020) 1–44. URL: http://jmlr.org/papers/v21/19-966.html.

[24] N. S. Hejazi, D. Benkeser, I. Díaz, M. J. van der Laan, Efficient estimation of modified treatment policy effects based on the generalized propensity score, 2022. doi:10.48550/arXiv.2205.05777.

[25] J. Wiśniewski, P. Biecek, fairmodels: A Flexible Tool For Bias Detection, Visualization, And Mitigation, 2022. URL: http://arxiv.org/abs/2104.00507, arXiv:2104.00507 [cs, stat].

A. German Credit Data

DAG, warping, and prediction models. We assume the same DAG as in the simulation study, depicted in Figure 1. For warping and prediction models, we use the same models as in the simulation study.

Evaluation. For the evaluation of the behavior of our method for this applied use case, we use the evaluation strategies defined in Section 4.2. Models are trained on randomly sampled 80% of the training data (i.e., 800 from 1,000 observations) and tested on the remaining 20%.

(UC1) Test accuracy in the real world is 71% for both the male and the female subgroup. In the warped world, male accuracy is comparable, with test accuracy of 70%. However, female accuracy increases to 75%, showing increasing performance for the discriminated subgroup.

(UC2) Table 1 shows individuals whose predictions differ most in the two worlds, either positively or negatively. Regressing this difference on features reveals that the risk prediction of young women grows strongly through warping, indicating that this subgroup was discriminated against most strongly in the real world (see Figure 4a). Figure 4b compares female predictions in both worlds and shows the most strongly affected individuals. Figure 4c shows prediction differences for female and male subgroups. While mean differences for females are significantly positive ($p < 10^{-12}$), the mean differences for males do not change significantly ($p = 0.26$). However, individual predictions and ranks of males do change: Figure 4d shows partial effects of Age and Amount on the prediction difference.

Table 1
Most discriminated individuals for German Credit data. Last column shows the difference in risk prediction between the two worlds (warped world - real world).

Gender	Age	Amount	Saving	Pred-diff
female	22	1567	1	0.20
female	20	1282	1	0.20
...	...	...	...	...
male	57	2225	1	-0.03
male	66	766	0	-0.03

(UC3) Investigating the effect of warping on the individuals reveals similar results as investigating the prediction differences in (UC2) and are omitted for the sake of concise presentation.
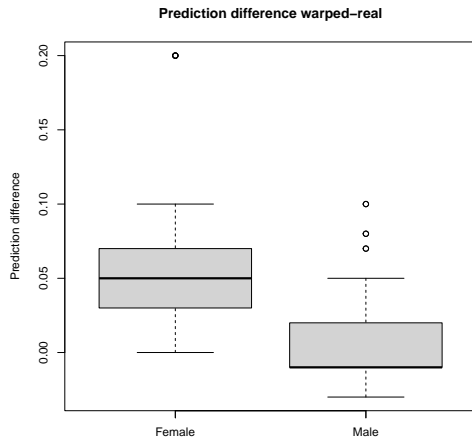
(UC4) The normalized feature differences between the real world and the warped world for Age, Amount, and Savings are $0.00, 0.01, 0.24$, respectively. This reveals that Savings is affected most by the warping and, hence, carries the strongest discrimination effect in the real world.
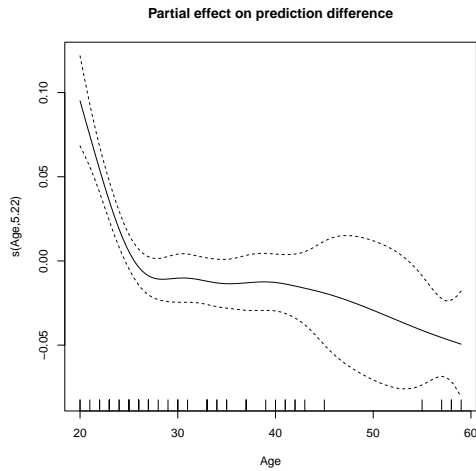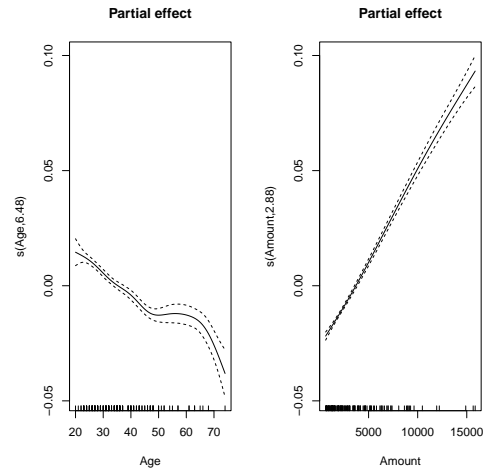
(a) Female Age effect on predictions.
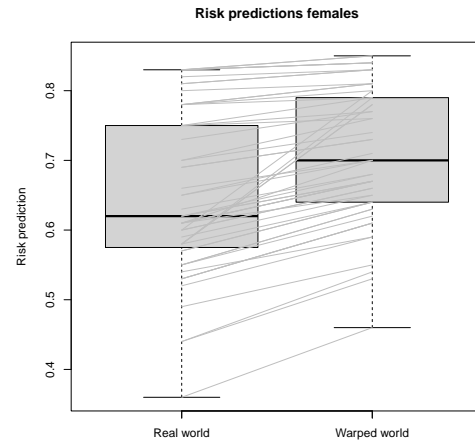
(b) Risk predictions for females in two worlds.

(c) Prediction diffs between worlds.

(d) Male Age and Amount effect on predictions.

Figure 4: Evaluation of results for German Credit data.

B. Classical FairML Metrics

As elaborated on thoroughly in [2], "classical" fairML metrics such as demographic parity, equalized odds, etc., do not reflect a clearly defined concept of fairness and, hence, are not suitable for deciding if an ML model entails unfairness. Hence, they can also not be used as quality criteria for evaluating our warping method. However, since some of these metrics are still popular, one might be interested – from a descriptive or explorative point of view – how these metrics change after applying the above proposed warping approach. For this reason, we provide respective results in the following, strongly emphasizing that such results are neither suitable for proving nor for disproving that our method works.

Simulation study.   For the simulation study described above, Tables 2 – 4 summarize some group fairness metrics (see, e.g., [5, 4] for an overview). We display ratios of different metrics, comparing the *male* and *female* subgroup, where a value smaller than 1 indicates that the respective metric in the *male* subgroup is larger than in the *female* subgroup (i.e., *female* value divided by *male* value). The tables show the following mean values (averaged over simulation iterations)[6]:

- ACC: Ratio of subgroup-specific accuracies, a.k.a. *Overall accuracy equality*

- PPV: Ratio of subgroup-specific positive predictive values (precisions), a.k.a. *Predictive parity*

- FPR: Ratio of subgroup-specific false positive rates, a.k.a. *Predictive equality*

- TPR: Ratio of subgroup-specific true positive rates, a.k.a. *Equal opportunity*

- STP: Ratio of subgroup-specific positively predicted rates, a.k.a. *Statistical parity* or *Demographic parity*

- No. checks passed: In each iteration, we check for each of the values of ACC, PPV, FPR, TPR, STP if it is inside the interval $(\epsilon, \frac{1}{\epsilon})$, where we use $\epsilon = 0.95$ as tolerance value. This number reports the total number of checks passed, which is $\in \{0, 1, 2, 3, 4, 5\}$ for each iteration, i.e., the mean (as reported in the tables) is $\in (0, 5)$.

Table 2 shows that for the scenario of correctly specified DAG and using the larger subgroup (*male*) as a reference group, the metrics are considerably closer to 1 for the warped and FiND world, compared with the real world. Warped and FiND world values are very close, only for FPR there seems to be a (small) difference between warped and FiND world values.

Table 2
Group fairness metrics w.r.t. RQ1, i.e., no misspecification and no reverse warping.

World	ACC	PPV	FPR	TPR	STP	No. checks passed
Real	0.9391	0.9337	1.0409	0.9563	0.8718	1.4440
Warped	1.0041	1.0023	0.9760	1.0028	1.0004	4.7040
FiND	0.9998	0.9999	1.0019	0.9999	0.9997	4.6040

Table 3 shows that for the scenario of a misspecified DAG, warped-world values are slightly more different than FiND world values but still considerably closer to 1 than real-world values. Only for FPR, real-world values are closer to 1 than warped-world values.

Table 4 shows that for the scenario of reverse warping, the picture is somewhat inconclusive: ACC and TPR of the real and warped world are very similar, for FPR the real-world values are closer to 1 where for PPV and STP the warped-world values are closer to 1. In the warped world, 76% more checks are passed compared to the real world.

---

[6]We used the R Package `fairmodels` [25].

Table 3

Group fairness metrics w.r.t. RQ2, i.e., misspecification and no reverse warping.

World	ACC	PPV	FPR	TPR	STP	No. checks passed
Real	0.9101	0.9022	1.0751	0.9401	0.8178	0.4070
Warped	1.0350	1.0207	0.7103	1.0208	1.0000	4.0000
FiND	1.0001	1.0001	0.9999	1.0000	0.9997	4.5180

Table 4

Group fairness metrics w.r.t. RQ3, i.e., no misspecification and reverse warping.

World	ACC	PPV	FPR	TPR	STP	No. checks passed
Real	0.9391	0.9337	1.0409	0.9563	0.8718	1.4440
Warped	0.9393	0.9544	1.4364	0.9556	1.0012	2.5320
FiND	0.9999	1.0000	0.9992	0.9997	0.9992	4.6390

German Credit Data. Figure 5 shows the same metrics for the analysis of the German credit data. For the real world, 1 check is passed (equal opportunity ratio), where for the warped world, all 5 checks are passed.
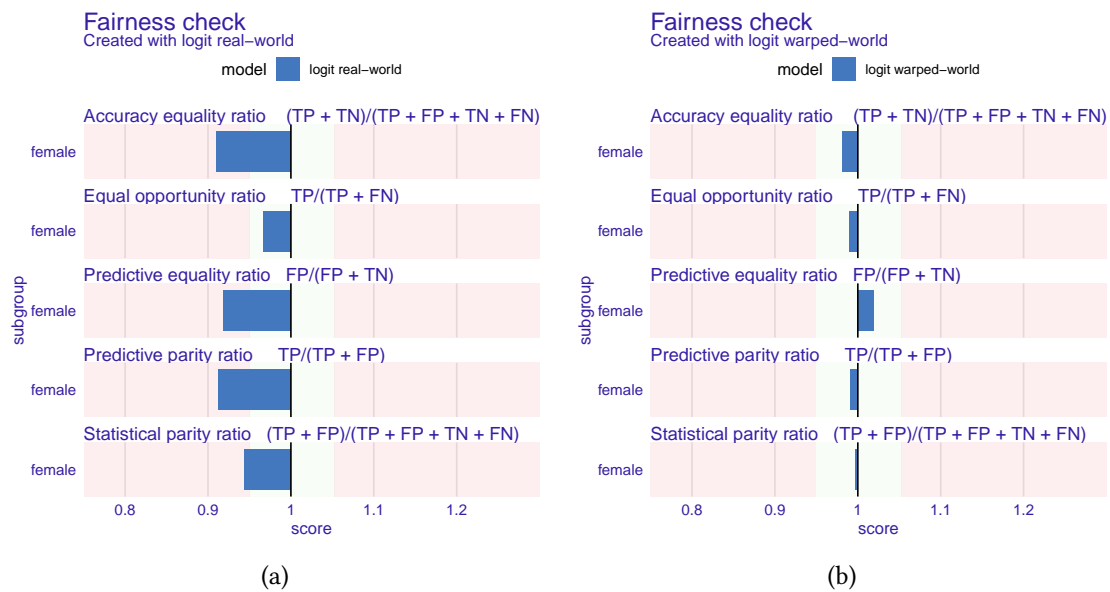
Figure 5: Group fairness metrics for German credit data in (a) real world and (b) warped world, produced by R-Package `fairmodels` by [25].