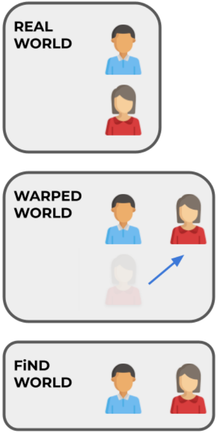
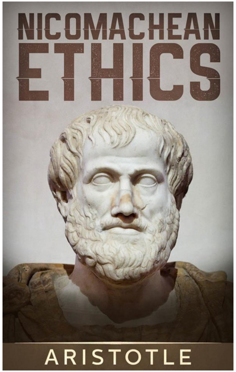
Causal Fair Machine Learning via Rank-Preserving Interventional Distributions

Ludwig Bothmann
Susanne Dandl
Michael Schomaker

AEQUITAS, Kraków, Poland – 01.10.2023



- FairML aspires to mitigate ML-related unfairness in ADM systems.
- Widely overlooked question: **“What is fairness?”**, i.e., what is the basic philosophical concept of fairness which the metrics shall measure? [Bothmann et al., 2023b]



Source: https://www.bol.com/nl/nl/p/nicomachean-ethics/9200000077435159/

Fairness – Basic Concept

Munich Center for Machine Learning

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Fairness since Aristotle [Aristotle, 2009]:

> Equals have to be treated equally,
> unequals have to be treated unequally.

$\Rightarrow$ Treatment / action aspect

$\Rightarrow$ Two normative definitions have to be made (specific to a task):

1. Measure of (task-specific) "equality" (ἡ ἀξία / "worthiness" $\rightarrow w^{(i)}$)
2. Concrete (un-)equal treatment, based on worthiness $\rightarrow s(w^{(i)})$

Role of ML in the ADM process

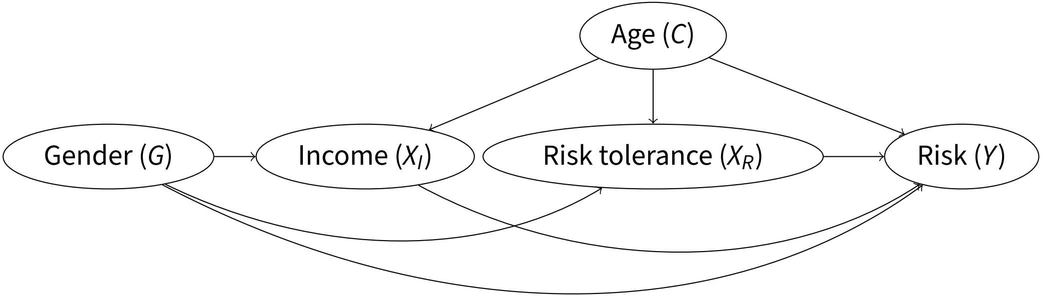
Definition (Fair treatment). A treatment $t^{(i)}$ of an individual i is called **fair** iff it is determined by a normative function of the individual's worthiness $w^{(i)}$, i.e., $t^{(i)} = s(w^{(i)})$.[1]

- Task of ML: Estimate worthiness $w^{(i)}$, e.g., $\pi^{(i)}$ (classification)
  - Fairness problems if $\pi^{(i)} \neq \hat{\pi}(\mathbf{x}^{(i)})$, i.e., if not **individually well-calibrated**.
- Protected attributes (PAs) change worthiness normatively, discrimination must not be based on PA, example:
  - i and j differ only in PA = Gender, i.e., $w^{(i)} = w^{(j)}$, even if $\pi^{(i)} \neq \pi^{(j)}$
  - decision based on $\hat{\pi}^{(i)} \neq \hat{\pi}^{(j)}$ is unfair
  - conceive **fictitious, normatively desired (FiND) world** where true probability $\phi^{(i)} = \phi^{(j)}$
  - estimate $\phi^{(i)}$ and base decision on $\hat{\phi}^{(i)}$

[1] See Bothmann et al. [2023b] for details.

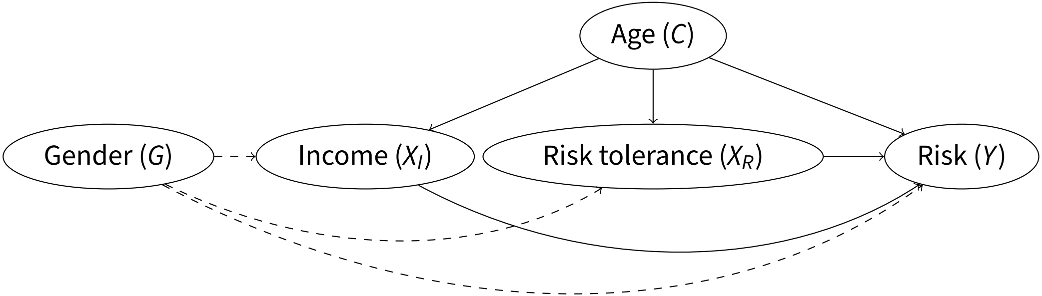
FiND world

⇒ Fictitious, normatively desired (FiND) world: PAs have no causal effect on the target.



FiND world

$\Rightarrow$ Fictitious, normatively desired (FiND) world: PAs have no causal effect on the target.



Structural Causal Model (Real World)

$$
\begin{aligned}
G &:= f(U_G) \\
C &:= f(U_C) \\
X_I &:= f_I(G, C, U_I) \\
X_R &:= f_R(G, C, U_R) \\
Y &:= f_Y(G, C, X_I, X_R, U_Y),
\end{aligned}
$$

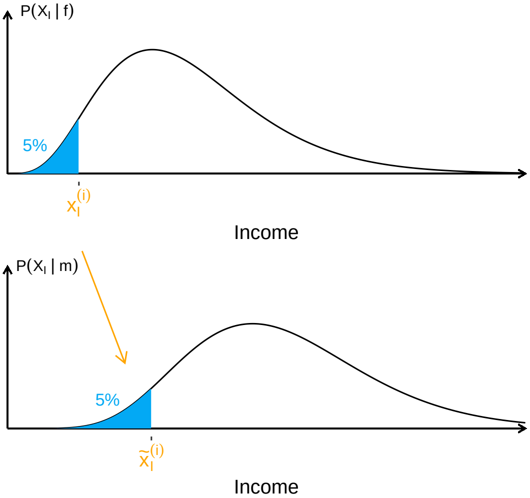
$\Rightarrow$ Joint distribution can be factorized:

$$
P(Y, X_R, X_I, C, G) = P_Y(Y|X_R, X_I, C, G)P_R(X_R|C, G)P_I(X_I|C, G)P_C(C)P_G(G). \tag{1}
$$

Rank-Preserving Interventional Distributions[2]

How to maintain individual “merits” in FiND world?

$\Rightarrow$ Group-specific individual ranks shall be preserved



[2] See Bothmann et al. [2023a] for details, and similar idea by Plečko and Meinshausen [2020]

Rank-Preserving Interventional Distributions

- Make all descendants from the PA neutral w.r.t. the PA, i.e., all PA-dependent quantities are transformed into their FiND-world counterparts.
- Fictitious intervention rule d_p leads to a joint post-intervention distribution $P_p(G, C, X_I^{d_p} X_R^{d_p}, Y^{d_p})$, which can be factorized in line with the pre-intervention distribution.

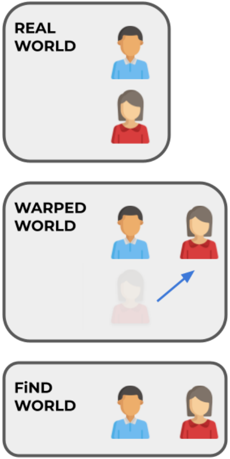
Estimation – Warping

Munich Center for Machine Learning

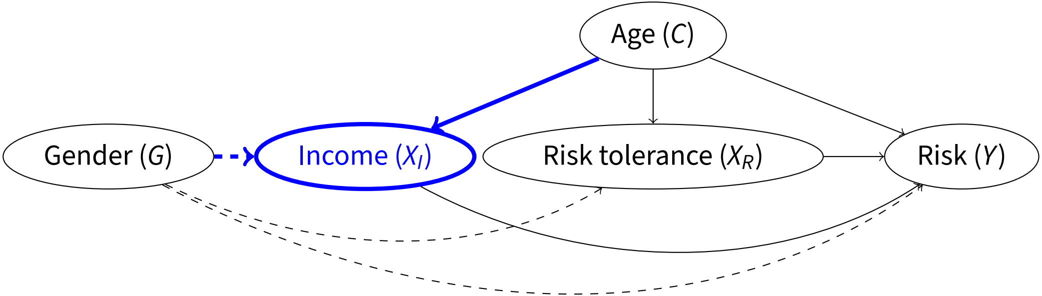
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Faced with real-world data, we propose a warping approach to approximate the FiND world:

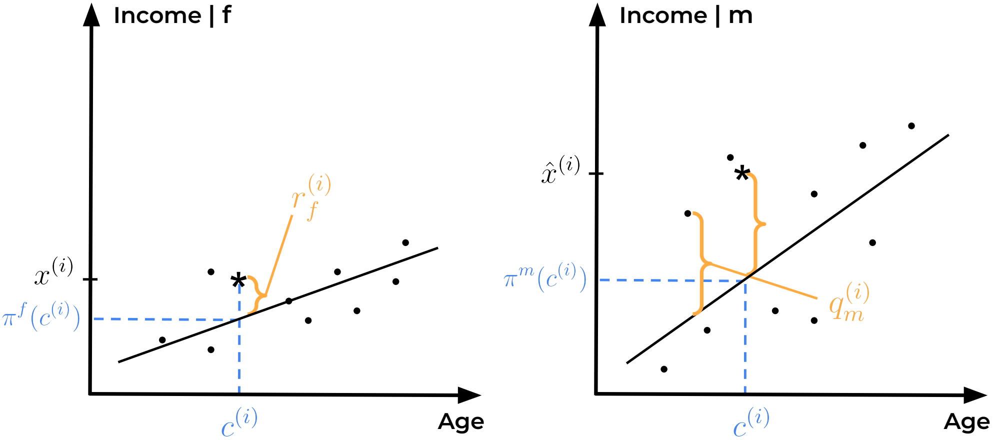
- Derive a warping from real world to warped world
- Train and test an ML model using the warped data
- At prediction: Use warping and trained ML model



Example: Warping for Income



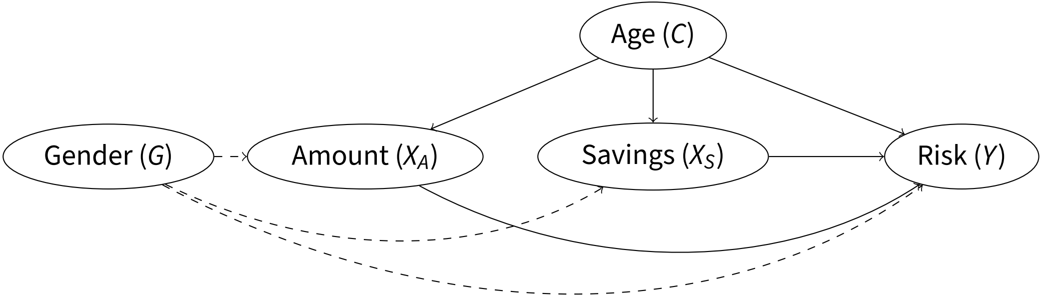
Residual-based Warping



Simulation Study – Research Questions

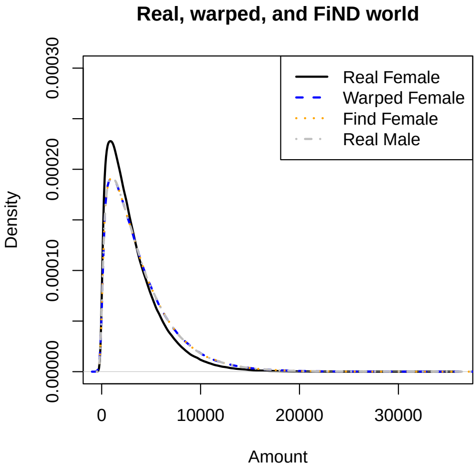
1. Does the warping method work? I.e., does it recover the distributions in the FiND world, and can it correctly identify the individual ranks of the target in the FiND world?
2. What effects does the warping direction have on performance (e.g., if subgroup A of the PA is warped to subgroup B, versus the other way around)?
3. How does misspecification of the DAG affect the results?
4. How does the warping method affect "classical" fairML metrics (e.g., statistical parity)?

Simulation Study – DAG



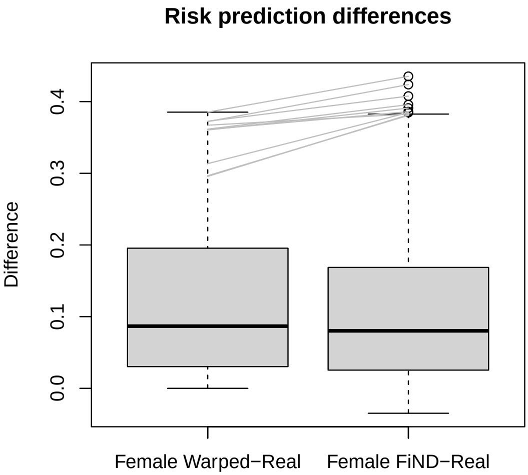
Simulation Study – Results – RQ1

Marginal distribution in FiND world is recovered:



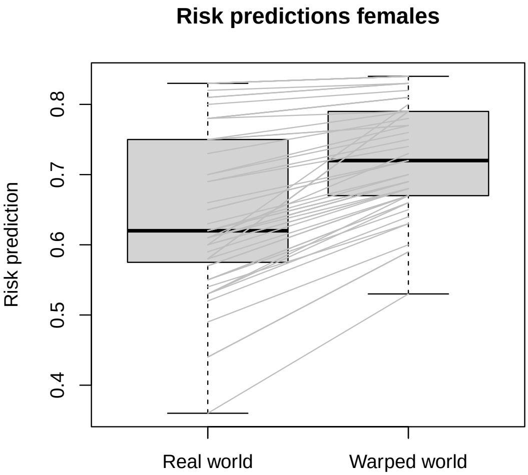
Simulation Study – Results – RQ1

The strongest discriminated individuals are found:



German Credit Data

All females have higher values in warped world, but to a different degree:



German Credit Data

Table: Most discriminated individuals for German Credit data.

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

Conclusion

- We **define** a treatment as **fair** if equals are treated equally and unequals unequally.
- In **FiND world**, normatively equal individuals are numerically equal, **PA have no effect**.
- Rank-preserving interventional distributions **identify** the FiND world.
- Warping method **estimates** the FiND world distributions.
- **Warping works** for the investigated simulation setup and empirical data.

Discussion

Extensions will be necessary:

- Rank-preserving interventional distributions:
  - Formulation for general SCMs
  - Solidify quantile approach for non-numeric variables
- Warping: Investigate other approaches, e.g., Plečko and Meinshausen [2020]
- Experiments:
  - Consider other, diverse DAGs
  - Compare different ML models for warping and target prediction
  - Investigate behavior under misspecification
  - Investigate behavior on other empirical data sets
- Compare our method to other methods that conceive a fictitious world for tackling fairness issues of ML models such as Zhang and Bareinboim [2018a,b], Nabi and Shpitser [2018], Nabi et al. [2019, 2022], Pfohl et al. [2019].

References I

Aristotle. *The Nicomachean ethics (book V)*. Oxford World's Classics. Oxford University Press, 2009. ISBN 978-0-19-921361-0. doi: 10.1093/actrade/9780199213610.book.1. URL https://doi.org/10.1093/actrade/9780199213610.book.1.

L. Bothmann, S. Dandl, and M. Schomaker. Causal Fair Machine Learning via Rank-Preserving Interventional Distributions. arXiv, 2023a. doi: 10.48550/arXiv.2307.12797.

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

R. Nabi and 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, pages 1931–1940, New Orleans, Louisiana, USA, Feb. 2018. AAAI Press. ISBN 978-1-57735-800-8. doi: 10.5555/3504035.3504270.

R. Nabi, D. Malinsky, and I. Shpitser. Learning Optimal Fair Policies. In *Proceedings of the 36th International Conference on Machine Learning*, pages 4674–4682. PMLR, May 2019. URL https://proceedings.mlr.press/v97/nabi19a.html.

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

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

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

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

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

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

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Causal Fair Machine Learning via Rank-Preserving Interventional Distributions

Ludwig Bothmann
Susanne Dandl
Michael Schomaker

01.10.2023

Role of ML in the ADM process

Definition (Fair treatment). A treatment $t^{(i)}$ of an individual i is called **fair** iff it is determined by a normative function of the individual's worthiness $w^{(i)}$, i.e., $t^{(i)} = s(w^{(i)})$, where $s(\cdot)$ is a (strictly) monotonic function.

Task of ML: Estimate worthiness $w^{(i)}$, e.g., $\pi^{(i)}$ (classification) or $\mu^{(i)}$ (regression)
$\Rightarrow$ ML model cannot be unfair per se but might induce unfairness of ADM.

Two sources of imprecision in estimating $w^{(i)} = \pi^{(i)}$ (for classification):

- We don't know $\pi^{(i)}$: Coarse information via $\pi(\mathbf{x}^{(i)})$
- We don't know $\pi(\cdot)$: Estimation via $\hat{\pi}(\cdot)$

$\Rightarrow$ Fairness problems arise if ML model is not **individually well-calibrated**, i.e., if not $\pi^{(i)} = \hat{\pi}(\mathbf{x}^{(i)})$.

Unfair treatment

Definition (Descriptively unfair treatment). Assume a pair of individuals i and j who differ only with respect to feature X. Assume that feature $\boldsymbol{X}$ **is not a causal reason** for a difference in the true probabilities, i.e., $\pi^{(i)} = \pi^{(j)}$. A treatment is called **descriptively unfair w.r.t. feature** X if these individuals are treated differently, i.e., $t^{(i)}(= s(\hat{\pi}^{(i)})) \neq t^{(j)}(= s(\hat{\pi}^{(j)}))$, in a process due to differing estimated individual probabilities $\hat{\pi}^{(i)} \neq \hat{\pi}^{(j)}$.

$\rightarrow$ Example credit risk ($\pi^{(i)}$ is payback probability):

- i and j differ only in $X =$ Gender
- (a) true recidivism probability $\pi^{(i)} = \pi^{(j)}$ (if Gender is not causal)
  - decision based on $\hat{\pi}^{(i)} \neq \hat{\pi}^{(j)}$ is descriptively unfair
- (b) true recidivism probability $\pi^{(i)} \neq \pi^{(j)}$ (if Gender is causal)
  - decision based on $\hat{\pi}^{(i)} \neq \hat{\pi}^{(j)}$ is **not** descriptively unfair

Protected attributes (PA)

Definition (Normatively unfair treatment). Assume a pair of individuals i and j who differ only with respect to feature A. Assume that feature ***A* is a causal reason** for a difference in the true probabilities, i.e., $\pi^{(i)} \neq \pi^{(j)}$. Assume that feature A is a PA. A treatment is called *normatively unfair w.r.t. feature A* if these individuals are treated differently, i.e., $t^{(i)}(= s(\hat{\pi}^{(i)})) \neq t^{(j)}(= s(\hat{\pi}^{(j)}))$, in a process due to differing estimated individual probabilities $\hat{\pi}^{(i)} \neq \hat{\pi}^{(j)}$, as feature A must not be invoked for the determination of equality, i.e., the decision basis for the treatment.

$\rightarrow$ Example credit risk ($\pi^{(i)}$ is payback probability in real world, $\phi^{(i)}$ in fictitious world):

- i and j differ only in $A =$ Gender
- decision based on $\hat{\pi}^{(i)} \neq \hat{\pi}^{(j)}$ is unfair
- true **corrected** recidivism probability $\phi^{(i)} = \phi^{(j)}$ (even if Gender is causal in real world)

$\Rightarrow$ Estimate $\phi^{(i)}$ and base decision on $\hat{\phi}^{(i)} \stackrel{?}{=} \hat{\phi}^{(j)}$.

Rank-Preserving Interventional Distributions

$$d_p = \begin{cases} X_I^{(i)} = \tilde{x}_I^{(i)} & \text{where } \tilde{x}_I^{(i)} \text{ is the } (p_I^{(i)} \times 100)\% \text{ quantile of the conditional} \\ & \text{mediator distribution among the reference PA value, i.e.,} \\ & P(X_I \leq \tilde{x}_I^{(i)} | C = c^{(i)}, G = m) = p_I^{(i)}, \text{ and } p_I^{(i)} \text{ is determined} \\ & \text{by the pre-intervention quantile of unit } i\text{, i.e.,} \\ & p_I^{(i)} = P(X_I \leq X_I^{(i)} \mid C = c^{(i)}, G = g^{(i)}). \\ X_R^{(i)} = \tilde{x}_R^{(i)} & \ldots \\ Y^{(i)} = \tilde{y}^{(i)} & \text{where } \tilde{y}^{(i)} \text{ is the } (p_Y^{(i)} \times 100)\% \text{ quantile of the counterfactual} \\ & \text{outcome distribution for the reference PA value, i.e.,} \\ & P(Y \leq \tilde{y}^{(i)} \mid X_I = \tilde{x}_I^{(i)}, X_R = \tilde{x}_R^{(i)}, C = c^{(i)}, G = m) = p_Y^{(i)}, \text{ and } p_Y^{(i)} \text{ is} \\ & \text{based on the pre-intervention quantile of unit } i\text{, i.e.,} \\ & p_Y^{(i)} = P(Y \leq y^{(i)} \mid X_I = x_I^{(i)}, X_R = x_R^{(i)}, C = c^{(i)}, G = g^{(i)}). \end{cases} \tag{2}$$

Residual-based Warping

Munich Center for Machine Learning

LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

1. Estimate prediction models for female $\pi_l^f(C)$ and male $\pi_l^m(C)$ population.
2. Compute residuals $r_f^{(i)} = \pi_l^f(c^{(i)}) - x_l^{(i)} \; \forall i \in I_f$, and $r_m^{(i)} = \pi_l^m(c^{(i)}) - x_l^{(i)} \; \forall i \in I_m$.
3. Compute individual probability rank of female i as $p_f^{(i)} = \frac{|\{j \in I_f : r_f^{(j)} \leq r_f^{(i)}|\}}{|I_f|}$.
4. Set $q_m^{(i)}$ to the empirical $p_f^{(i)}$-quantile of the residuals of the male model π_l^m, i.e., $q_m^{(i)} = \min\{r \in R_m : \frac{|\{j \in R_m : j \leq r\}|}{|R_m|} \geq p_l^{(i)}\}$, with $R_m = \{r_m^{(i)} : i \in I_m\}$ the set of male residuals.
5. Warp $x_l^{(i)}$ to the sum of male prediction and warped residual, i.e., $\hat{x}_l^{(i)} = \pi_l^m(c^{(i)}) + q_m^{(i)}$.

Simulation Study – Setup

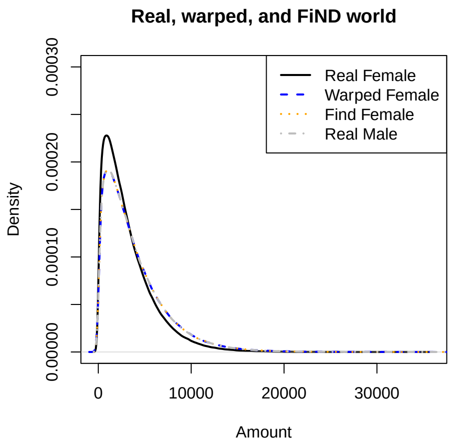
Confounder: Age – Features: Amount (numeric) and Savings (binary)

$$
\begin{aligned}
G &\sim \mathrm{B}(\pi_G) \\
C &\sim \mathrm{Ga}(\alpha_C, \beta_C) \\
X_A | C, G &\sim \mathrm{Ga}(\alpha_A(C, G), \beta_A(C, G)) \\
X_S | C, G &\sim \mathrm{B}(\pi_S(C, G)) \\
Y | X_A, X_S, C, G &\sim \mathrm{B}(\pi_Y(X_A, X_S, C, G))
\end{aligned}
$$

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

Simulation Study – Results – RQ1

Marginal distribution of Amount:

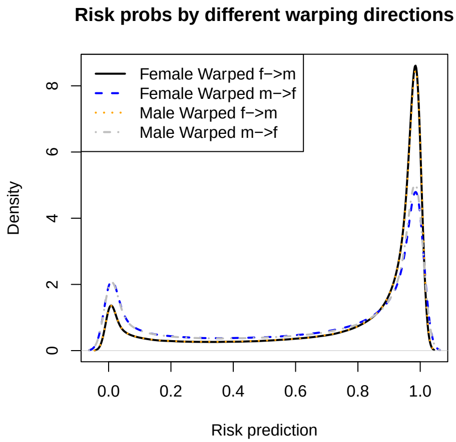


Mean difference between male and female risk predictions (95% CI):

- Real world: 0.1122 (0.1117, 0.1127)
- Warped world: -0.0016 (-0.0021, -0.0012)

Simulation Study – Results – RQ2

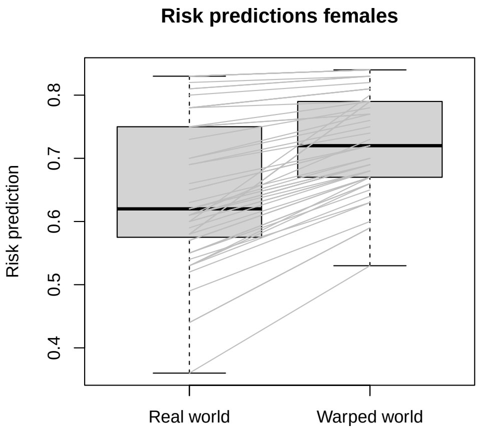
General level shifts:

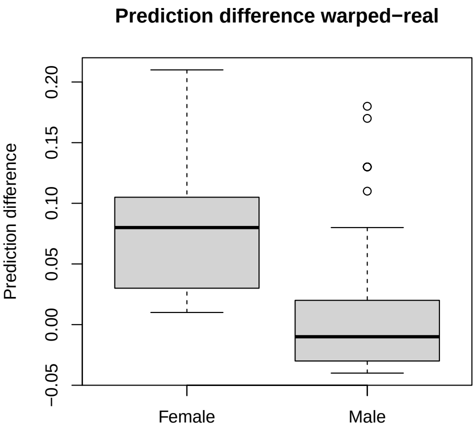


Mean difference between male and female risk predictions (95% CI)

- Real world: 0.1122 (0.1117, 0.1127)
- Warped world: 0.0065 (0.0060, 0.0071)

German Credit Data





Classical FairML Metrics – Simulation Study

Munich Center for Machine Learning

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Table: Group fairness metrics w.r.t. RQ1 – Average over simulation runs

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

Classical FairML Metrics – German Credit Data