Fairness and Discrimination with Multiple Attributes

Agathe Fernandes Machado, Suzie Grondin, François Hu, Philipp Ratz
& **Arthur Charpentier**

Insurance Data Science 2024, Stockholm

Motivation

	CA	HI	GA	NC	NY	MA	PA	FL	TX	AL	ON	NB	NL	QC
Gender	×	×	●	×	●	×	×	●	●	●	●	×	×	●
Age	×	×	●	×*	●	×	●	●	●	●*	●	×	×	●
Driving experience	●	×	●	●	●	●	●	●	●	●	●	●	●	●
Credit history	×	×	●	●	●	×	●*	●	●	×*	×	●*	×	●
Education	×	×	×	×	×	×	●	●	●	●	●	●	●	●
Occupation	×	×	×	●	×	×	●	●	●	●	●	●	●	●
Employment status	×	×	×	●	×	×	●	●	●	●	●	●	●	●
Marital status	●	×	●	●	●	×	●	●	●	●	●	●	●	●
Housing situation	×	×	●	●	●	×	●	●	●	×	×	●	●	●
Address/ZIP code	●	●	●	●	●	●	●	●	●	×	×	●	●	●
Insurance history	●	●	●	●	●	●	●	●	●	●	●	●	●	●

CA: California, HI: Hawaii, GA: Georgia, NC: North Carolina, NY: New York, MA: Massachusetts, PA: Pennsylvania, FL: Florida, TX: Texas, AL: Alberta, ON: Ontario, NB: New-Brunswick, NL: Newfoundland-Labrador, QC: Québec, The Zebra (2022)

```
pip install equipy
```

Notations, risk $\mathcal{R}$ and unfairness $\mathcal{U}$

- $\boldsymbol{X}\in\mathcal{X}$: 'non-sensitive' features,
- $\boldsymbol{A}=(A_1,\cdots A_r)\in\mathcal{A}_1\times\cdots\times\mathcal{A}_r$: r sensitive features,
- $\widehat{Y}$: prediction of response variable (continuous or score for a binary classifier)
- m: predictive model for $(\boldsymbol{x},\boldsymbol{a})$, and $m^\star$ the optimal Bayes estimator,
- ν_f: distribution of $m(\boldsymbol{X},\boldsymbol{A})$
  with cumulative distribution function F_m and quantile function Q_m,
- $\nu_{f|a_i}$: conditional distribution of $m(\boldsymbol{X},\boldsymbol{A})|A_i=a_i$ with $F_{m|a_i}$ and $Q_{m|a_i}$,
- $\mathcal{R}(m)=\mathbb{E}[(Y-m(\boldsymbol{X},\boldsymbol{A}))^2]$: (theoretical) quadratic risk.
- Given a sample $\{(y_i,\boldsymbol{x}_i,\boldsymbol{a}_i)\}$, the empirical risk is $\widehat{\mathcal{R}}_n(m)=\displaystyle\sum_{i=1}^n(y_i-m(\boldsymbol{x}_i,\boldsymbol{a}_i))^2$
- Optimal Bayes estimator $m^\star=\text{argmin}\{\mathcal{R}(m)\}$, is

$$m^\star(\boldsymbol{x},\boldsymbol{a})=\mathbb{E}\big[Y\big|(\boldsymbol{X}=\boldsymbol{x},(\boldsymbol{A}=\boldsymbol{a}))\big]$$

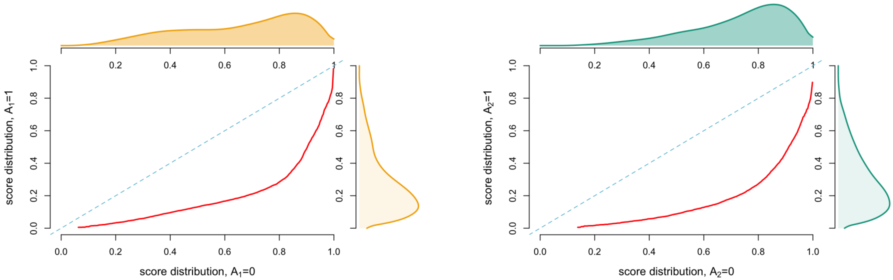
Notations, risk $\mathcal{R}$ and unfairness $\mathcal{U}$

- Theoretical quadratic risk is $\mathcal{R}(m) = \mathbb{E}[(Y - m(\boldsymbol{X}, \boldsymbol{A}))^2]$
- Define Wasserstein (2) distance $\mathcal{W}_2^2(\nu_0, \nu_1) = \min_{Z_0 \sim \nu_0, Z_1 \sim \nu_1} \{\mathbb{E}[(Z_1 - Z_0)^2]\}$

  if Z_0 and Z_1 are univariate (+ technical conditions),

$$\mathcal{W}_2^2(\nu_0, \nu_1) = \int_0^1 (Q_1(u) - Q_0(u))^2 du = \mathbb{E}[(Q_1 \circ F_0(Z_0) - Z_0)^2], \; Z_0 \sim \nu_0$$

  $z \mapsto Q_1 \circ F_0(z)$ is the "optimal Monge" transport, Santambrogio (2015)



Notations, risk $\mathcal{R}$ and unfairness $\mathcal{U}$

- m is strongly fair regarding a single sensitive attribute (SSA) A_i, according to demographic parity – $m(\boldsymbol{X},\boldsymbol{A}) \perp\!\!\!\perp A_i$ – if and only if:

$$\mathcal{U}_i(m) = \max_{a_i\in\mathcal{A}_i} \text{distance}(\nu_m,\nu_{m|a_i}) = 0 \text{ unfairness w.r.t. } A_i$$

- m is strongly fair regarding multiple sensitive attribute (MSA), if and only if:

$$\mathcal{U}(m) = \mathcal{U}_1(m)+\cdots+\mathcal{U}_r(m) = 0 \text{ unfairness w.r.t. } \boldsymbol{A}$$

- m is ϵ-approximately fairregarding A_i if and only if $\mathcal{U}_i(m)\leq\epsilon\cdot\mathcal{U}_i(m^\star)$ (where $m^\star$ is the optimal Bayes estimator)

The Price to Pay for Fairness

- Let $\mathcal{M}$ denote a (general) class of models,
- Let $\mathcal{M}_{\mathsf{F},i}$ denote the subset of fair models with respect to A_i,

$$\mathcal{M}_{\mathsf{F},i} = \{m \in \mathcal{M} : \mathcal{U}_i(m) = 0\}$$

- Fairness is achieved by projection onto the fair subspace

$$\widehat{m} \in \underset{m \in \mathcal{M}}{\text{argmin}} \left\{\widehat{\mathcal{R}}_n(m)\right\} \text{ and } \widehat{m}_{\mathsf{F},i} \in \underset{m \in \mathcal{M}_{\mathsf{F},i}}{\text{argmin}} \left\{\widehat{\mathcal{R}}_n(m)\right\}$$

- The price of fairness of a model class $\mathcal{M}$

$$\mathcal{E}_{\mathsf{F},i}(\mathcal{M}) = \min_{m \in \mathcal{M}_{\mathsf{F},i}} \{\mathcal{R}(m)\} - \min_{m \in \mathcal{M}} \{\mathcal{R}(m)\} \geq 0.$$

- Chzhen et al. (2020) proved that (+mild technical assumptions)

$$\mathcal{E}_{\mathsf{F},i}(\mathcal{M}) = \min_{\mu \in \mathcal{M}} \left\{ \sum_{a_i \in \mathcal{A}_i} \mathbb{P}(A_i = a_i) \cdot \mathcal{W}_2^2(\nu_{m^\star|a_i}, \nu_{\mu|a_i}) \right\}$$

The Price to Pay for Fairness

- Given K distributions $(\nu_1, \ldots, \nu_K)$, and weights $(w_1, \ldots, w_K) \in \mathbb{R}_+^K$, the $\mathcal{W}_2$-Barycenter is the minimizer:

$$\text{Bar}\{(w_k, \nu_k)_{k=1}^K\} = \underset{\nu}{\text{argmin}} \left\{ \sum_{k=1}^{K} w_k \cdot \mathcal{W}_2^2(\nu_k, \nu) \right\}.$$

- SSA ($r = 1$) Chzhen et al. (2020)

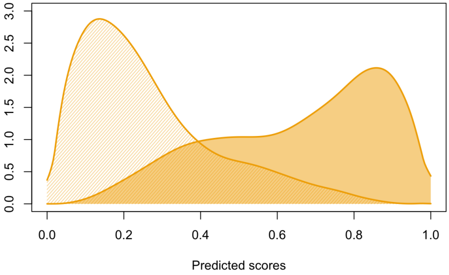
$$\nu_{m_{B_i}} \text{ corresponding to } \min_m \sum_{a_i \in \mathcal{A}_i} \mathbb{P}(A_i = a_i) \cdot \mathcal{W}_2^2\left(\nu_{m^\star|a_i}, \nu_m\right)$$

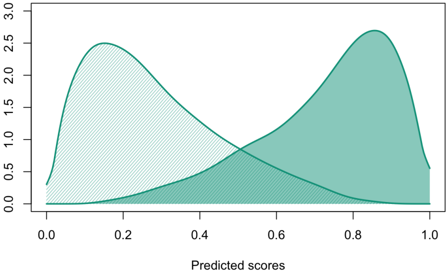
$$m_{B_i}(\boldsymbol{x}, a_i) = \left(\sum_{a' \in \mathcal{A}_i} \mathbb{P}(A_i = a') \cdot Q_{m^*|a'} \right) \circ F_{m^\star|a_i}(m^*(\boldsymbol{x}, a_i)), \ \forall (\boldsymbol{x}, a_i) \in \mathcal{X} \times \mathcal{A}_i.$$

$\rightarrow$ EquiPy: implemented in the function `FairWasserstein` of `fairness` module.

The Case of Multiple Attributes

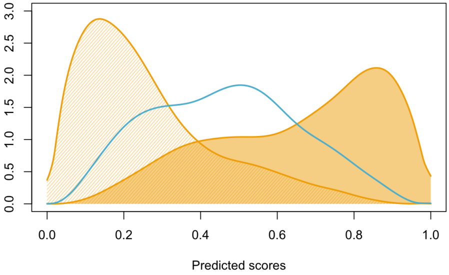
- Consider a machine Learning model m, score predictions and two sensitive attributes, ethnic origin A_1 (White/Black) and gender A_2 (Male/Female).
- Consider densities of $\nu_{m|A_1=0}$, $\nu_{m|A_1=1}$ (left) and $\nu_{m|A_2=0}$, $\nu_{m|A_2=1}$ (right)

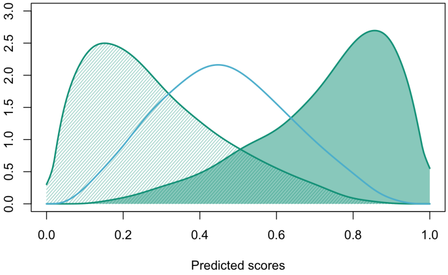




The Case of Multiple Attributes

- Consider a machine Learning model m, score predictions and two sensitive attributes, ethnic origin A_1 (White/Black) and gender A_2 (Male/Female).
- Consider densities of $\nu_{m|A_1=0}$, $\nu_{m|A_1=1}$ (left) and $\nu_{m|A_2=0}$, $\nu_{m|A_2=1}$ (right)
- Plot densities of barycenters, $\nu_{m_{B_1}}$ and $\nu_{m_{B_2}}$





The Case of Multiple Attributes

- Intersectional Fairness, MSA → Single sensitive attribute (SSA), by intersection,

$$\boldsymbol{a} \in \mathcal{A} = \underbrace{\mathcal{A}_1}_{\text{ethnic origin } A_1} \times \underbrace{\mathcal{A}_2}_{\text{gender } A_2} = \{\text{white}, \text{black}\} \times \{\text{male}, \text{female}\}$$

Here $\mathcal{A}$ corresponds to $4 = 2 \times 2$ states,

$$\mathcal{A} = \Big\{(\text{white}, \text{male}), (\text{white}, \text{female}), (\text{black}, \text{male}), (\text{black}, \text{female})\Big\}$$

- Sequential Fairness, MSA, in Hu et al. (2024)

The Case of Multiple Attributes

- MSA ($r \geq 1$) Hu et al. (2024)

$$m_B(\boldsymbol{x}, \boldsymbol{a}) := m_{B_1} \circ m_{B_2} \circ \cdots \circ m_{B_r}(\boldsymbol{x}, \boldsymbol{a})$$

where

$$m_{B_i} \circ m_{B_j}(\boldsymbol{x}, \boldsymbol{a}) = \left(\sum_{a' \in \mathcal{A}_i} \mathbb{P}(A_i = a') Q_{m_{B_j}|a'_i} \right) \circ F_{m_{B_j}|a_i}(m_{B_j}(\boldsymbol{x}, \boldsymbol{a}))$$

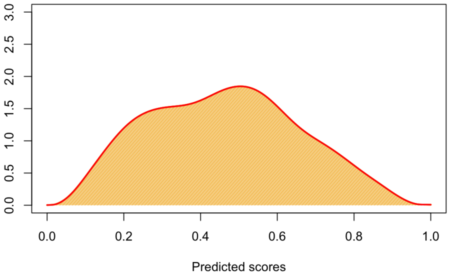
$\forall\ (\boldsymbol{x}, \boldsymbol{a}) \in \mathcal{X} \times \mathcal{A}_{1:r}$, with the i-th component of $\boldsymbol{a}$ denoted a_i.

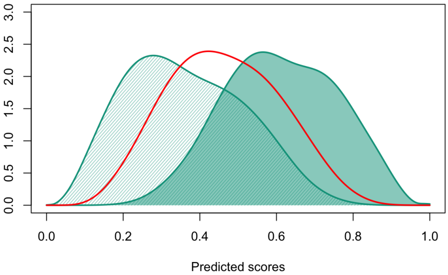
- Hu et al. (2024) proved the associativity of Wasserstein barycenters(fairness mitigation remains unaffected by the order of $A_{1:r}$).

→ `EquiPy`: implemented in the function `MultiWasserstein` of `fairness` module.

The Case of Multiple Attributes

- Given $\nu_{m_{B_1}}$, consider
  - the barycenter $\nu_{m_{B_1}}$ conditional on A_1 (no impact, already fair)
  - the barycenter $\nu_{m_{B_2}}$ conditional on A_2

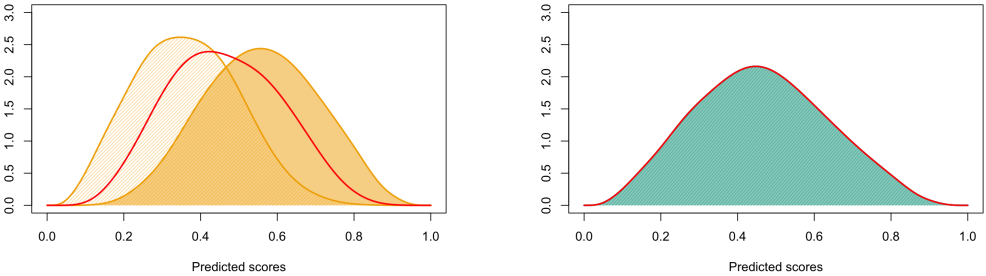




- On the right, distribution of $\nu_{m_{B_2} \circ m_{B_1}}$

The Case of Multiple Attributes

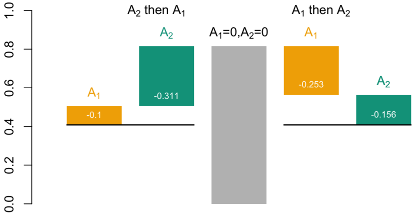
- Given $\nu_{m_{B_2}}$, consider
  - the barycenter $\nu_{m_{B_1}}$ conditional on A_1
  - the barycenter $\nu_{m_{B_2}}$ conditional on A_2 (no impact, already fair)

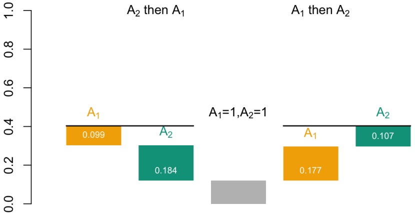


- On the left, distribution of $\nu_{m_{B_1} \circ m_{B_2}}$

The Case of Multiple Attributes

- The order of this sequential approach leads different interpretations,
  - left hand part, A_2 then A_1
  - right hand part, A_1 then A_2





Life insurance dataset

- Public SEER database: `https://seer.cancer.gov`,
- Prediction of one-year mortality of US individuals with melanoma skin cancer,
  $\to$ Use the methodology presented in Sauce et al. (2023), we convert the dataset into survival data, by accounting for exposure over a given time interval.
- Sample size $n = 547,878$ from 2004 to 2018,
- Explanatory variables: 16 features describing patient characteristics (age, gender male/female, ethnic origin) and cancer attributes (tumor size, extent).

$\to$ MSA framework: use of the function `MultiWasserstein`.

Model fitting

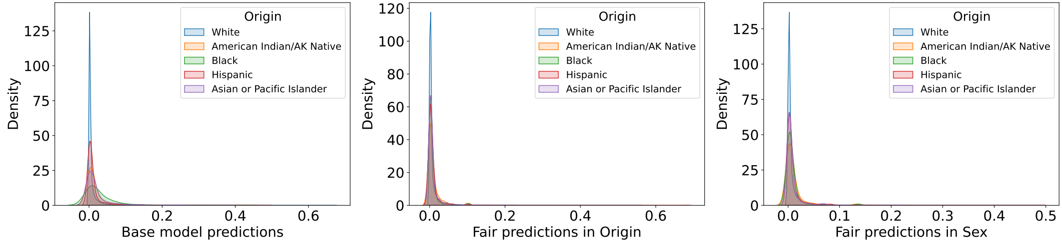
1. Split the data into train and test sets,
2. Fit Logistic Regression m (m can be any ML model, model-agnosticity)
3. Apply m on the test set to obtain $\hat{\boldsymbol{y}}_{\text{test}}$.

We consider different model fitting scenarios, in which we include or exclude sensitive attributes as explanatory variables:

Ethnic origin	Gender	AUC	Unfairness
No	No	0.8652	0.2179
Yes	**Yes**	0.8672	0.2668

Transforming predictions

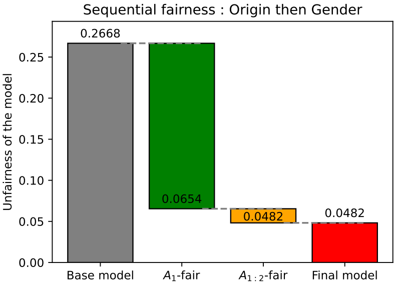
1. Split the test data into calibration and test sets,
2. Specify an order to sequentially correct: A_1 corresponds to ethnic origin and A_2 corresponds to gender,
3. Fit and transform your test predictions using `MultiWasserstein` from `fairness` module.

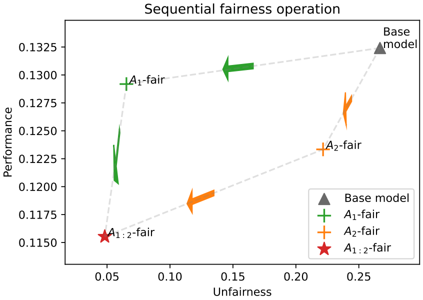


Visualizations

Unfairness and metric calculations with `graphs` module:

- `fair_waterfall_plot`: sequential gain in fairness for the specified order A_1 then A_2,
- `fair_multiple_arrow_plot`: fairness-performance relationship for all potential pathways.





Additional results: Approximate fairness

When correcting biases related to gender, we reduce fairness regarding origin:

Fairness step	Unfairness in origin	Unfairness in gender
Base model	**0.2371**	0.0297
Origin	**0.0345**	0.0309
Origin & Gender	**0.0469**	0.0013

We can prioritize fairness across attributes by specifying $\epsilon = [0, 0.5]$ corresponding to exact fairness in A_1 and 0.5-approximate fairness in A_2.

$$m_B = 0.5 \cdot (m_{B_2} \circ m_{B_1}) + 0.5 \cdot m_{B_1}$$

References

Charpentier, A. (2024). *Insurance: biases, discrimination and fairness*. Springer Verlag.

Chzhen, E., Denis, C., Hebiri, M., Oneto, L., and Pontil, M. (2020). Fair regression with wasserstein barycenters. *Advances in Neural Information Processing Systems*, 33:7321–7331.

Hu, F., Ratz, P., and Charpentier, A. (2024). A sequentially fair mechanism for multiple sensitive attributes. *Annual AAAI Conference on Artificial Intelligence*.

Machado, A. F., Grondin, S., Hu, F., Ratz, P., and Charpentier, A. (2024). Equipy: Sequential fairness using optimal transport in Python. *in progress*.

Santambrogio, F. (2015). Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94.

Sauce, M., Chancel, A., and Ly, A. (2023). AI and Ethics in Insurance: a new solution to mitigate proxy discrimination in risk modeling. *arXiv*, 2307.13616.

The Zebra (2022). Car insurance rating factors by state. *https://www.thezebra.com/*.