Causal Inference and Counterfactuals with Optimal Transport with Applications in Fairness and Discrimination

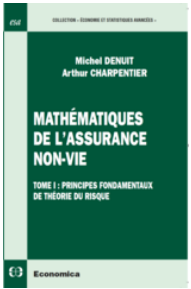
Arthur Charpentier

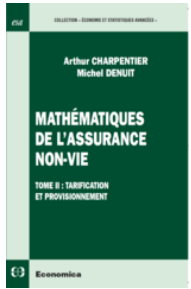
WG Risk / Essec - September 2023

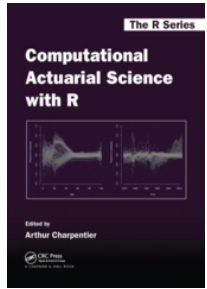
Bio (short)

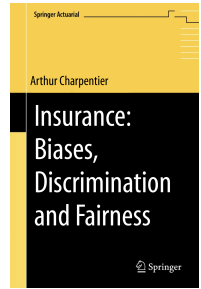
Professor at Université du Québec à Montréal, (https://freakonometrics.github.io/)

- Denuit and Charpentier (2004, 2005) Mathématiques de l'Assurance Non-Vie,
- Charpentier (2014) Computational Actuarial Science with R,
- Bénéplanc et al. (2022) Manuel d'Assurance,
- Charpentier (2023) Insurance: Biases, Discrimination and Fairness.



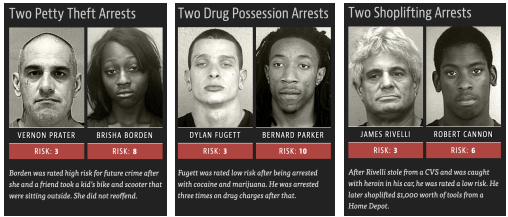


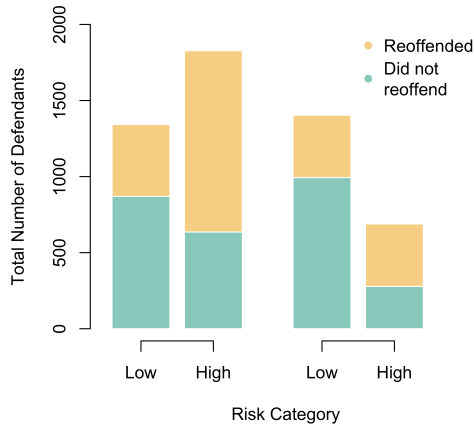




Motivation (1. Propublica, Actuarial Justice)

- Concept of "actuarial justice" as coined in Feeley and Simon (1994)
- **C**orrectional **O**ffender **M**anagement **P**rofiling for **A**lternative **S**anctions (**COMPAS**), Perry (2013)



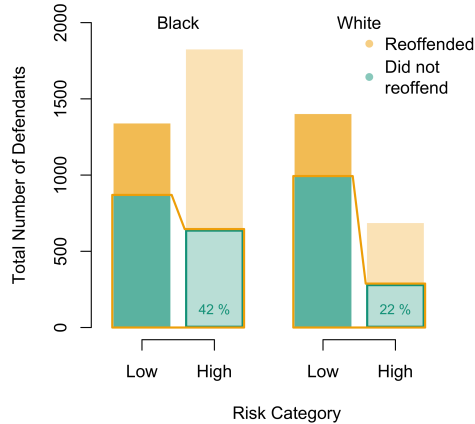


https://github.com/propublica/compas-analysis

- Angwin et al. (2016) Machine Bias
  Dressel and Farid (2018)

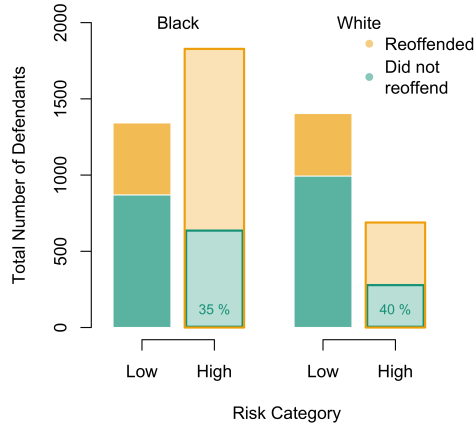
Motivation (1. Propublica, Actuarial Justice)

- From Feller et al. (2016),
  - for White people, among those who did not re-offend, 78% were properly classified, since 22% did re-offend,
  - for Black people, among those who did not re-offend, 58% were properly classified, since 42% did re-offend.

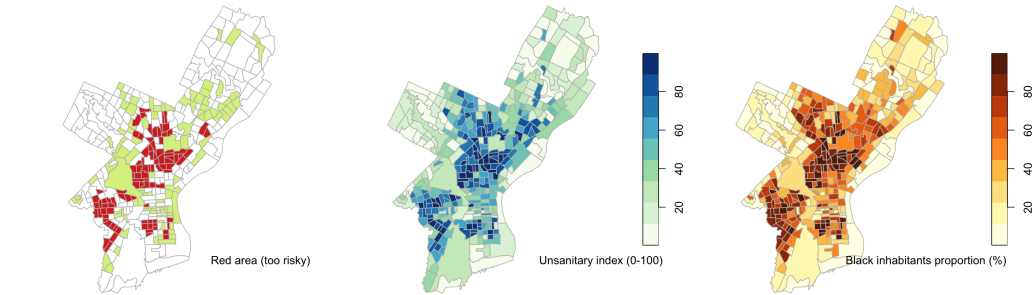


Motivation (1. Propublica, Actuarial Justice)

- From Dieterich et al. (2016),
  - for White people, among those who where classified as high risk, 40% did not re-offend,
  - for Black people, among those who where classified as high risk, 35% did not re-offend.



Motivation (2. Redlining)



(Fictitious maps, inspired by a Home Owners' Loan Corporation map from 1937)

- Federal Home Loan Bank Board (FHLBB) "*residential security maps*" (for real-estate investments), Crossney (2016) and Rhynhart (2020)
- Unsanitary index and proportion of Black inhabitants

Discrimination and Insurance

"*What is unique about insurance is that even statistical discrimination which by definition is absent of any malicious intentions, poses significant moral and legal challenges. Why? Because on the one hand, policy makers would like insurers to treat their insureds equally, without discriminating based on race, gender, age, or other characteristics, even if it makes statistical sense to discriminate* (...) *On the other hand, at the core of insurance business lies discrimination between risky and non-risky insureds. But riskiness often statistically correlates with the same characteristics policy makers would like to prohibit insurers from taking into account.* " Avraham (2017)

"*Technology is neither good nor bad; nor is it neutral* " , Kranzberg (1986)

"*Machine learning won't give you anything like gender neutrality 'for free' that you didn't explicitely ask for* ", Kearns and Roth (2019)

Discrimination and Protected Attributes ?

California

Allowed (with applicable limitations): driving experience, marital status, address/zip code

Prohibited (or effectively prohibited): gender, age, credit history, education, occupation, employment status, residential status, insurance history

Notes & Clarifications: California's insurance commissioner banned gender as of January 2019. Occupation and education are permitted for use in group plans (i.e. for alumni associations and other membership programs).

Georgia

Allowed (with applicable limitations): gender, age, years of driving experience, credit history, marital status, residential status, address/zip code, insurance history

Prohibited (or effectively prohibited): occupation, education, and employment status

Notes & Clarifications: none

Hawaii

Allowed (with applicable limitations): address/zip code, insurance history

Prohibited (or effectively prohibited): gender, age, years of driving experience, credit history, education, occupation, employment status, marital status, residential status

Notes & Clarifications: none

Illinois

Allowed (with applicable limitations): gender, age, years of driving experience, credit history, education, occupation, employment status, marital status, residential status, address/zip code, insurance history

Prohibited (or effectively prohibited): none

Notes & Clarifications: none

Massachusetts

Allowed (with applicable limitations): years of driving experience, address/zip code, insurance history

Prohibited (or effectively prohibited): gender, age, credit history, education, occupation, employment status, marital status, residential status

Notes & Clarifications: none

Michigan

Allowed (with applicable limitations): gender (group-rated policies), age, years of driving experience, credit history, education, occupation, employment status, marital status (group-rated policies), residential status, address/zip code, insurance history

Prohibited (or effectively prohibited): gender (non-group policies), marital status (non-group policies)

Notes & Clarifications: Gender and marital status are permitted only in rate-making for group plans (i.e. for alumni associations and other membership programs). **UPDATE:** Michigan lawmakers approved a major insurance reform bill in May 2019 that will ban insurers in the state from using gender, marital status, address/zipcode, residential status, education and occupation in rate setting. The ban will be enforced starting in July 2020. Insurers will be permitted to use "territory" as approved by the state regulators instead of zip code.

New York

Allowed (with applicable limitations): gender, age, years of driving experience, credit history, marital status, residential status, address/zip code, insurance history

Prohibited (or effectively prohibited): occupation, education, employment status

Notes & Clarifications: none

via The Zebra (2022)

Fairness (Demographic Parity) for Classifiers

- Some notations,

$$\begin{cases} \boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^d : \text{'explanatory' variables} \\ s \in \{\texttt{A}, \texttt{B}\} : \text{"sensitive variable"} \\ y \in \{0, 1\} : \text{classification problem} \\ m(\boldsymbol{x}) : \text{scoring function, classically } m(\boldsymbol{x}, s) = \mathbb{P}[Y = 1 | \boldsymbol{X} = \boldsymbol{x}, S = s] \\ \widehat{y} \in \{0, 1\} : \text{prediction, classically } \widehat{y} = \mathbf{1}(m(\boldsymbol{x}, s) > t) \end{cases}$$

- One could consider multiple sensitive attributes, see Hu et al. (2023b)

Fairness (Demographic Parity) for Classifiers

- Defining "Demographic Parity", Corbett-Davies et al. (2017) or Agarwal (2021)

Weak Demographic Parity,
Decision function $\widehat{y}$ satisfies weak demographic parity if $\widehat{Y} \perp\!\!\!\perp S$, i.e.

$$\mathbb{E}[\widehat{Y}|S = \texttt{A}] = \mathbb{E}[\widehat{Y}|S = \texttt{B}],$$

or

$$\mathbb{E}[\mathbf{1}(m(\boldsymbol{X}, S) > t)|S = \texttt{A}] = \mathbb{E}[\mathbf{1}(m(\boldsymbol{X}, S) > t)|S = \texttt{B}].$$

One can easily obtain weak Demographic Parity using different thresholds

$$\mathbb{E}[\mathbf{1}(m(\boldsymbol{X}, S) > t_{\texttt{A}})|S = \texttt{A}] = \mathbb{E}[\mathbf{1}(m(\boldsymbol{X}, S) > t_{\texttt{B}})|S = \texttt{B}].$$

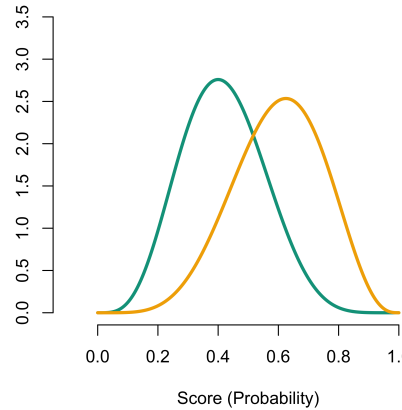
Fairness (Demographic Parity) for Scores

Strong Demographic Parity,

$$\mathbb{E}[\mathbf{1}(m(\boldsymbol{X}, S) \in E)|S = \mathtt{A}] = \mathbb{E}[\mathbf{1}(m(\boldsymbol{X}, S) \in E)|S = \mathtt{B}],$$

for any $E \subset [0, 1]$, or $\mathbb{P}_{\mathtt{A}}[E] = \mathbb{P}_{\mathtt{B}}[E]$,

$$\begin{cases} \mathbb{P}_{\mathtt{A}}[E] = \mathbb{P}[m(\boldsymbol{X}, S) \in E|S = \mathtt{A}] \\ \mathbb{P}_{\mathtt{B}}[E] = \mathbb{P}[m(\boldsymbol{X}, S) \in E|S = \mathtt{B}] \end{cases}$$



Fairness (Demographic Parity) for Scores

- Use some "distance" between $\mathbb{P}_A$ and $\mathbb{P}_B$ (TV, KL, or Wasserstein)

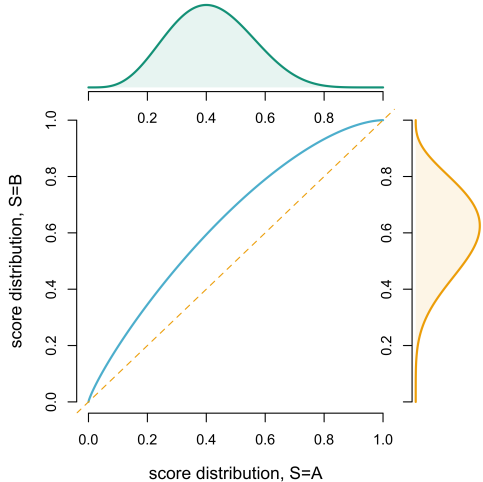
$$\inf_{\pi \in \Pi(\mathbb{P}_A, \mathbb{P}_B)} \{\mathbb{E}[\ell(X, Y)],\ (X, Y) \sim \pi\}$$

or

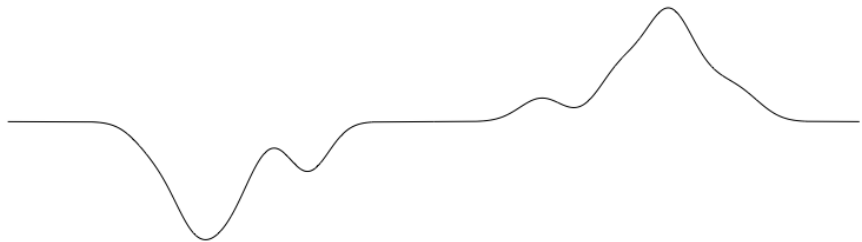
$$\inf_{\pi \in \Pi(\mathbb{P}_A, \mathbb{P}_B)} \left\{ \int \ell(x, y) \pi(dx, dy) \right\}.$$

or using a transport mapping $\mathcal{T}$

$$\inf_{\mathcal{T}: \mathcal{T}_{\#} \mathbb{P}_A = \mathbb{P}_B} \left\{ \int \ell(x, \mathcal{T}(x)) d\mathbb{P}_A(x) \right\}.$$



Fairness and Optimal Transport

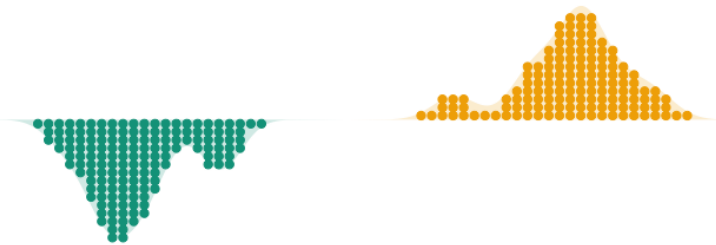


Monge (1781), Mémoire sur la théorie des déblais et des remblais
We want to transport optimally sand from a hole (with shape $-d\mathbb{P}_{\mathtt{A}}$) to a pile (with shape $d\mathbb{P}_{\mathtt{B}}$). "*Rien ne se perd, rien ne se créé, tout se transporte*": $\int d\mathbb{P}_{\mathtt{A}} = \int d\mathbb{P}_{\mathtt{B}}$.

Fairness and Optimal Transport

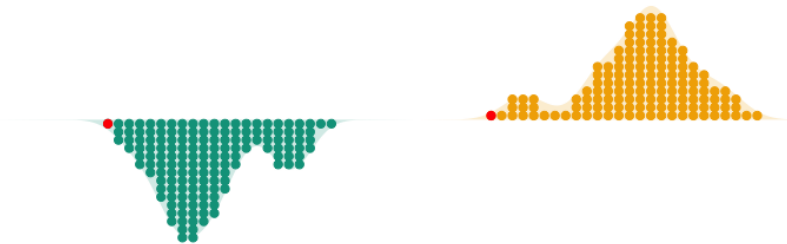
$$\inf_{T:T_{\#}\mathbb{P}_A=\mathbb{P}_B}\left\{\int(x-T(x))^k d\mathbb{P}_A(x)\right\}. = \left(\int_0^1 \left|F_0^{-1}(u)-F_1^{-1}(u)\right|^k \mathrm{d}u\right)^{1/k},$$

Fairness and Optimal Transport



$$\inf_{\mathcal{T}:\mathcal{T}_{\#}\mathbb{P}_{\mathrm{A}}=\mathbb{P}_{\mathrm{B}}} \left\{ \int (x-\mathcal{T}(x))^k d\mathbb{P}_{\mathrm{A}}(x) \right\} . = \left(\int_0^1 \left| F_{\mathrm{A}}^{-1}(u) - F_{\mathrm{B}}^{-1}(u) \right|^k \mathrm{d}u \right)^{1/k},$$

Fairness and Optimal Transport



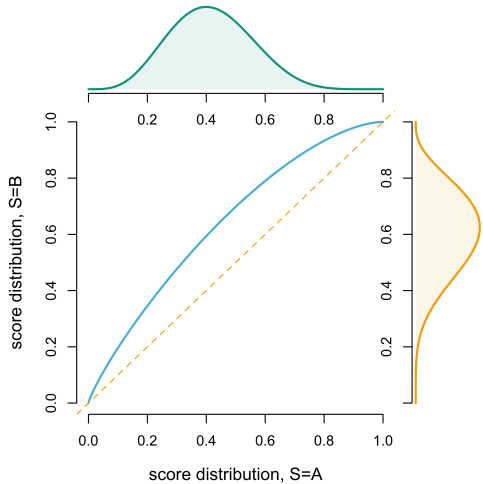
Optimal transport plan is here $\mathcal{T}^\star : x \mapsto y = F_{\mathtt{B}}^{-1} \circ F_{\mathtt{A}}(x)$ (increasing function)

Counterfactual Fairness (and Optimal Transport)

- Used to quantify unfairness,

m satisfies **Strong Demographic Parity** if $W_2 = 0$,

$$W_2 = \left(\int_0^1 \left(F_{\mathrm{A}}^{-1}(u) - F_{\mathrm{B}}^{-1}(u) \right)^2 \mathrm{d}u \right)^{1/2}.$$



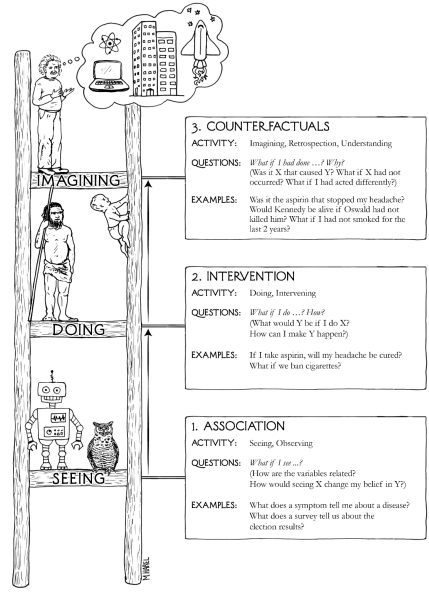
Counterfactual Fairness (and Optimal Transport)

"Ladder of causation" from Pearl et al. (2009)

- 3. **Counterfactuals**
(Imagining, "*what if I had done...*")
- 2. **Intervention**
(Doing, "*what if I do...*")
- 1. **Association**
(Seeing, "*what if I see...*")

Picture source: Pearl and Mackenzie (2018)

What would be the impact of a treatment T on a variable of interest Y ?



Counterfactual Fairness (and Optimal Transport)

- Define individual or counterfactural fairness, Castelnovo et al. (2022)

"*Individual fairness is embodied in the following principle: similar individuals should be given similar decisions. This principle deals with the comparison of single individuals rather than focusing on groups of people sharing some characteristics.* "

- Following Kusner et al. (2017)

A decision is **counterfactually fair** if the prediction in the real world is the same as the prediction in the counterfactual world

$$\mathbb{E}[Y^{\star}_{S\leftarrow \mathtt{A}}|\boldsymbol{X}=\boldsymbol{x}]=\mathbb{E}[Y^{\star}_{S\leftarrow \mathtt{B}}|\boldsymbol{X}=\boldsymbol{x}],\ \forall \boldsymbol{x},$$

where $Y^{\star}_{S\leftarrow \mathtt{A}}$ and $Y^{\star}_{S\leftarrow \mathtt{B}}$ denote "potential outcomes".

- since we use the same $\boldsymbol{x}$ it is a ceteris paribus counterfactual.

(is the counterfactual of a man with height 190 cm a woman with height 190 cm ?)

Counterfactuals and Causal Inference

Ceteris paribus sic stantibus Ceteris paribus is a Latin phrase, meaning "*all other things being equal*" or "*other things held constant*".

Mutatis mutandis Mutatis mutandis is a Latin phrase meaning "*with things changed that should be changed*" or "*once the necessary changes have been made*".

Consider a linear model $\widehat{y} = m(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, where $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$.

What happens if $x_1 \to x_1 + dx_1$?

- Ceteris paribus: $\widehat{y} \to \widehat{y} + \beta_1 dx_1$
- Mutatis mutandis: $\widehat{y} \to \widehat{y} + \beta_1 dx_1 + \beta_2 \frac{r\sigma_2}{\sigma_1} dx_1$

Counterfactuals and Causal Inference

	Gender	Treatment	Outcome (Weight)			Height	⋯
		t_i	y_i	$y^\star_{i,T\leftarrow 0}$	$y^\star_{i,T\leftarrow 1}$	x_i	⋯
1	H	0	75	75	?	172	⋯
2	F	1	52	?	52	161	⋯
3	F	1	57	?	57	163	⋯
4	H	0	78	78	?	183	⋯

(different notations are used $y(1)$ and $y(0)$ in Imbens and Rubin (2015), y^1 and y^0 in Cunningham (2021), or $y_{t=1}$ and $y_{t=0}$ in Pearl and Mackenzie (2018))

ATE & SATE

$$\text{ATE} = \mathbb{E}\left[Y^*_{T\leftarrow 1} - Y^*_{T\leftarrow 0}\right] \text{ and SATE} = \frac{1}{n}\sum_{i=1}^{n} y^\star_{i,T\leftarrow 1} - y^\star_{i,T\leftarrow 0}$$

Counterfactuals and Causal Inference

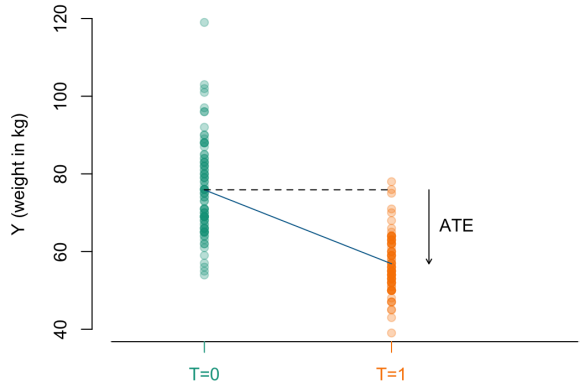
Let $y_i \in \mathcal{D}_0$ and $y_j \in \mathcal{D}_1$

"*what would have been the weight of that person if that person had been a woman, and not a man?* "

(too) simple sample estimate

$$\text{SATE} = \overline{y}_1 - \overline{y}_0$$

$$\text{SATE} = \frac{1}{n_1} \sum_{j \in \mathcal{D}_1} y_j - \frac{1}{n_0} \sum_{i \in \mathcal{D}_0} y_i$$



Counterfactuals and Causal Inference

Consider a third variable x

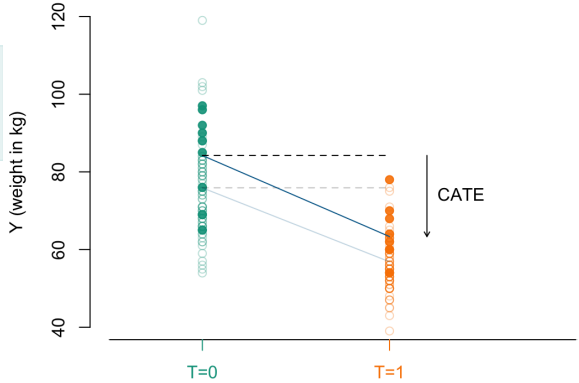
CATE Conditional ATE is CATE(x)

$$\mathbb{E}\left[Y^*_{T\leftarrow 1} - Y^*_{T\leftarrow 0} \middle| X = x\right]$$

A natural estimate would be

$$\text{SCATE} = \frac{1}{k}\sum_{x_j \in \mathcal{V}^k_{x,1}} y_j - \frac{1}{k}\sum_{x_i \in \mathcal{V}^k_{x,0}} y_i$$

as in Abrevaya et al. (2015).



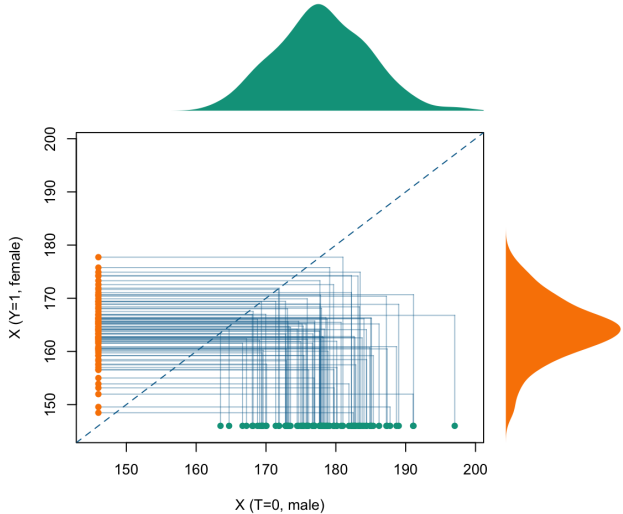
Matching

Between $x_i \in \mathcal{D}_0$ and $x_j \in \mathcal{D}_1$

$$j_i^\star = \underset{j \in \mathcal{D}_1}{\text{argmin}} \{d(x_i, x_j)\},$$

then remove observation from $\mathcal{D}_1$.

Algorithm in Rubin (1973),
described in Stuart (2010) under the name "1:1 nearest neighbor matching",
see Ho et al. (2007) or Dehejia and Wahba (1999)
also called "Greedy Matching"

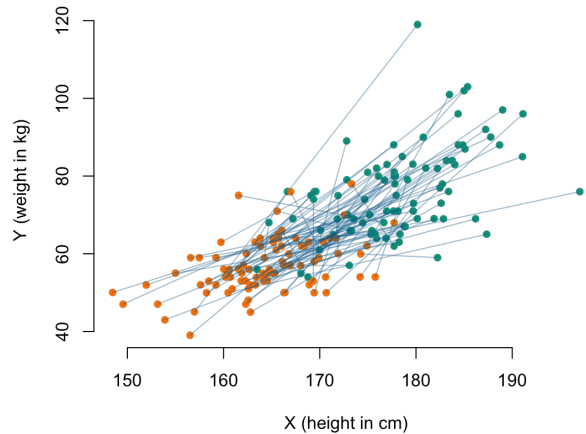


Matching

Each individual (x_i, y_i) in the control group ($\mathcal{D}_0$)
has counterfactual $(x_{j_i^\star}, y_{j_i^\star})$
in the treated group ($\mathcal{D}_1$)...

$$\text{SATE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_{j_i^\star} - y_i\right)$$

i.e. $\text{SATE} = \overline{y}_1 - \overline{y}_0$
(simply permutate observations in $\mathcal{D}_1$)



Matching

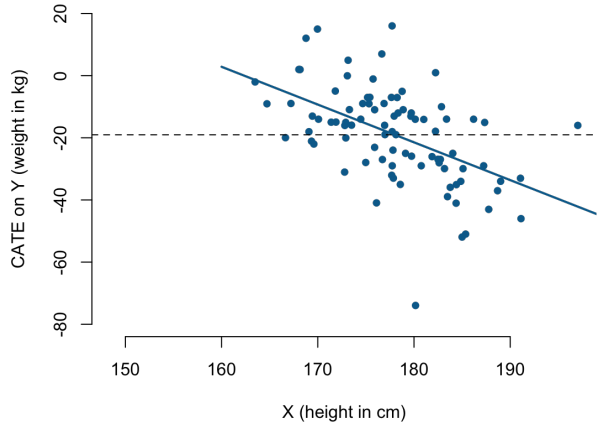
Let V_x^k denote the list of k nearest neighbors of x_i's in $\mathcal{D}_0$ close to x,

$$\text{SCATE}(x) = \frac{1}{k} \sum_{i \in V_x^k} (y_{j_i^\star} - y_i)$$

Here scatter-plot • of

$$\{(x_i, y_{j_i^\star} - y_i)\}_{i=1,\cdots,n}$$

and linear regression ————
Horizontal line – – – is ATE



Optimal Matching

C is the $n \times n$ matrix that quantifies the distance between individuals in the two groups, $C_{i,j} = d(x_i, x_j)^2 = (x_i - x_j)^2$, the optimal matching is solution of

$$\min_{P \in \mathcal{P}} \left\{ \langle P, C \rangle \right\} = \min_{P \in \mathcal{P}} \left\{ \sum_{i,j} P_{i,j} C_{i,j} \right\}, \tag{1}$$

where $\mathcal{P}$ is the set of permutation matrices
$n \times n$ permutation matrix, P, with entries in $\{0, 1\}$, satisfying $P\mathbf{1}_n = \mathbf{1}_n$ and $P^{\star\top}\mathbf{1}_n = \mathbf{1}_n$, see Brualdi (2006).

Optimal Matching

Initial algorithm without the "no-replacement" rule (1:1), total cost 1.06

	7	8	9	10	11	12
1	0.41	0.55	0.22	0.64	0.04	0.25
2	0.28	0.24	0.73	0.22	0.64	0.80
3	0.28	0.47	0.32	0.52	0.16	0.37
4	0.28	0.62	0.81	0.25	0.64	0.85
5	0.41	0.37	0.89	0.25	0.81	0.97
6	0.66	0.76	0.21	0.89	0.22	0.14

	7	8	9	10	11	12	
1	·	·	·	·	1	·	$1 \to 11$
2	·	·	·	1	·	·	$2 \to 10$
3	·	·	·	·	1	·	$3 \to 11$
4	·	·	·	1	·	·	$4 \to 10$
5	·	·	·	1	·	·	$5 \to 10$
6	·	·	·	·	·	1	$6 \to 12$

Optimal Matching

Initial algorithm (not optimal), total cost 2.19

	7	8	9	10	11	12
1	0.41	0.55	0.22	0.64	0.04	0.25
2	0.28	0.24	0.73	0.22	0.64	0.80
3	0.28	0.47	0.32	0.52	0.16	0.37
4	0.28	0.62	0.81	0.25	0.64	0.85
5	0.41	0.37	0.89	0.25	0.81	0.97
6	0.66	0.76	0.21	0.89	0.22	0.14

		7	8	9	10	11	12
1	1	·	·	·	·	1	·
2	2	·	·	·	1	·	·
3	3	1	·	·	·	·	·
4	4	·	1	·	·	·	·
5	5	·	·	1	·	·	·
6	6	·	·	·	·	·	1

$1 \leftrightarrow 11$
$2 \leftrightarrow 10$
$3 \leftrightarrow 7$
$4 \leftrightarrow 8$
$5 \leftrightarrow 9$
$6 \leftrightarrow 12$

Optimal Matching

Initial algorithm (not optimal), another initial shuffle, total cost 1.32

	7	8	9	10	11	12
1	0.41	0.55	0.22	0.64	0.04	0.25
2	0.28	0.24	0.73	0.22	0.64	0.80
3	0.28	0.47	0.32	0.52	0.16	0.37
4	0.28	0.62	0.81	0.25	0.64	0.85
5	0.41	0.37	0.89	0.25	0.81	0.97
6	0.66	0.76	0.21	0.89	0.22	0.14

		7	8	9	10	11	12	
5	1	·	·	1	·	·	·	$1 \leftrightarrow 9$
3	2	1	·	·	·	·	·	$2 \leftrightarrow 7$
4	3	·	·	·	·	1	·	$3 \leftrightarrow 11$
1	4	·	·	·	1	·	·	$4 \leftrightarrow 10$
2	5	·	1	·	·	·	·	$5 \leftrightarrow 8$
6	6	·	·	·	·	·	1	$6 \leftrightarrow 12$

Optimal Matching

Initial algorithm (optimal), total cost 1.27*

	7	8	9	10	11	12
1	0.41	0.55	0.22	0.64	0.04	0.25
2	0.28	0.24	0.73	0.22	0.64	0.80
3	0.28	0.47	0.32	0.52	0.16	0.37
4	0.28	0.62	0.81	0.25	0.64	0.85
5	0.41	0.37	0.89	0.25	0.81	0.97
6	0.66	0.76	0.21	0.89	0.22	0.14

	7	8	9	10	11	12
1	·	·	·	·	1	·
2	·	1	·	·	·	·
3	·	·	1	·	·	·
4	1	·	·	·	·	·
5	·	·	·	1	·	·
6	·	·	·	·	·	1

$1 \leftrightarrow 11$
$2 \leftrightarrow 8$
$3 \leftrightarrow 9$
$4 \leftrightarrow 7$
$5 \leftrightarrow 10$
$6 \leftrightarrow 12$

Optimal Matching

Algorithm 1 SATE, optimal matching case

Require: dataset $\mathcal{D} = \{(y_i, \boldsymbol{x}_i, t_i)\}$

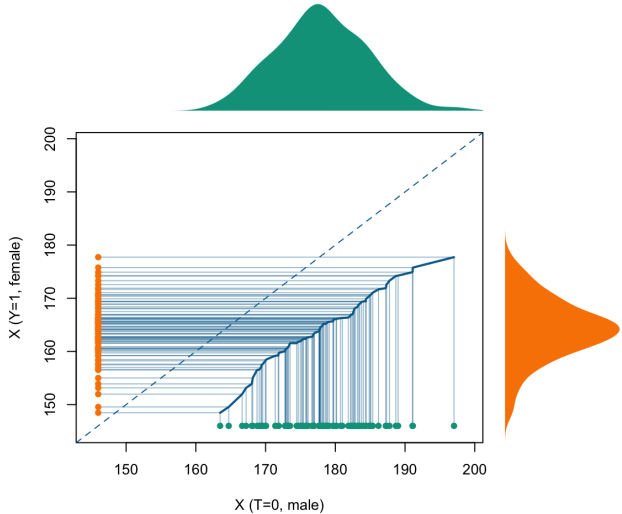
1: $\mathcal{D}_0 \leftarrow$ subset of $\mathcal{D}$ when $t = 0$ (size n) shuffled, with indices i
2: $\mathcal{D}_1 \leftarrow$ subset of $\mathcal{D}$ when $t = 1$ (size n), with indices j
3: $C \leftarrow$ matrix $n \times n$, $C_{i,j} = d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ between points in $\mathcal{D}_0$ and $\mathcal{D}_1$
4: $P^* \leftarrow$ solution of Problem (1)
5: $\text{SATE} \leftarrow \dfrac{1}{n_0} \displaystyle\sum_{i=1}^{n} y_i^0 - P_i^{*\top} \boldsymbol{y}^1$

Actually Problem (1) has a simple explicit solution (based on ranks) and a nice geometric interpretation

Optimal Matching

$y_i \in \mathcal{D}_0$ and $y_j \in \mathcal{D}_1$
Let r_i denote the rank of $y_i \in \mathcal{D}_0$ and s_j denote the rank of $y_j \in \mathcal{D}_1$

Match $y_i \in \mathcal{D}_0$ with $y_{j_i^\star} \in \mathcal{D}_1$ such that $r_i = s_j$.

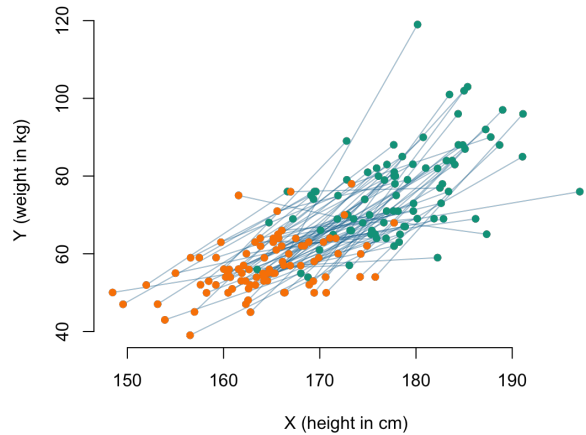


Optimal Matching

Each individual (x_i, y_i)
in the control group
has counterfactual $(x_{j_i^\star}, y_{j_i^\star})$
in the treated group...

$$\text{SATE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_{j_i^\star} - y_i\right)$$

i.e. $= \overline{y}_1 - \overline{y}_0$
(again, simple permutation)



Optimal Matching

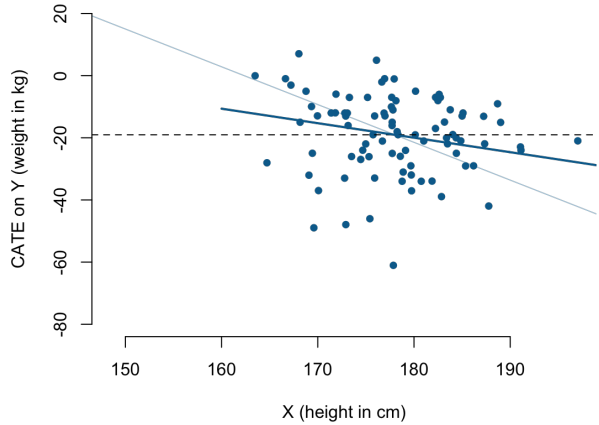
Let $V_{\boldsymbol{x}}^k$ denote the list of k nearest neighbors of x_i's in $\mathcal{D}_0$ close to x,

$$\text{SCATE}(\boldsymbol{x}) = \frac{1}{k} \sum_{i \in V_{\boldsymbol{x}}^k} (y_{j_i^\star} - y_i)$$

Here scatter-plot • of

$$\left\{ (x_i, y_{j_i^\star} - y_i) \right\}_{i=1,\cdots,n}$$

and linear regression ————
Horizontal line — — — is ATE
(same as the previous one)



Optimal Coupling

$r_i^{(1)}$ denote the rank of $x_i^{(1)}$ in the treated dataset $\{x_1^{(1)}, \cdots, x_n^{(1)}\}$. The procedure then becomes simply a matching based on ranks, in the sense that $j_i^\star$ satisfies $r_{j_i^\star}^{(1)} = r_i^{(0)}$, as discussed in Chapter 2 of Santambrogio (2015).

In a very general setting, if $\boldsymbol{a}_0 \in \mathbb{R}_+^{n_0}$ and $\boldsymbol{a}_1 \in \mathbb{R}_+^{n_1}$ satisfy $\boldsymbol{a}_0^\top 1_{n_0} = \boldsymbol{a}_1^\top 1_{n_1}$ (identical sums), define

$$U(\boldsymbol{a}_0, \boldsymbol{a}_1) = \left\{ M \in \mathbb{R}_+^{n_0 \times n_1} : M\mathbf{1}_{n_1} = \boldsymbol{a}_0 \text{ and } M^\top \mathbf{1}_{n_0} = \boldsymbol{a}_1 \right\}.$$

This set of matrices is a convex polytope (see Brualdi (2006)).

In our case, let U_{n_0,n_1} denote $U\left(\mathbf{1}_0, \frac{n_0}{n_1}\mathbf{1}_1\right)$

$$P^* \in \underset{P \in U_{n_0,n_1}}{\operatorname{argmin}} \left\{ \langle P, C \rangle \right\} \text{ or } \underset{P \in U_{n_0,n_1}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} P_{i,j} C_{i,j} \right\}. \tag{2}$$

Optimal Coupling

	7	8	9	10	11	12	13	14	15	16
1	0.41	0.55	0.22	0.64	0.04	0.25	0.24	0.77	0.74	0.55
2	0.28	0.24	0.73	0.22	0.64	0.80	0.76	0.76	0.12	0.10
3	0.28	0.47	0.32	0.52	0.16	0.37	0.27	0.68	0.63	0.45
4	0.28	0.62	0.81	0.25	0.64	0.85	0.58	0.32	0.51	0.48
5	0.41	0.37	0.89	0.25	0.81	0.97	0.91	0.81	0.05	0.25
6	0.66	0.76	0.21	0.89	0.22	0.14	0.33	0.96	0.99	0.79

	7	8	9	10	11	12	13	14	15	16
1	·	·	1/5	·	3/5	·	1/5	·	·	·
2	·	2/5	·	·	·	·	·	·	·	3/5
3	3/5	·	·	·	·	·	2/5	·	·	·
4	·	·	·	2/5	·	·	·	3/5	·	·
5	·	1/5	·	1/5	·	·	·	·	3/5	·
6	·	·	2/5	·	·	3/5	·	·	·	·

Quantile CATE

If $X_0 \sim F_0$, then $X_1 = \mathcal{T}(X_0) \sim F_1$, where $\mathcal{T} : x_0 \mapsto x_1 = F_1^{-1} \circ F_0(x_0)$.

Mutatis mutandis Quantile CATE

$$\text{QCATE}(u) = \mathbb{E}\big[Y^*_{T\leftarrow 1} | X = F_1^{-1}(u)\big] - \mathbb{E}\big[Y^*_{T\leftarrow 0} | X = F_0^{-1}(u)\big],\ u \in (0,1),$$

where F_t is the cumulative distribution function of X, conditional on $T = t$

Mutatis mutandis CATE

$$\mathbb{E}\big[Y^*_{T\leftarrow 1} | X = \mathcal{T}(x)\big] - \mathbb{E}\big[Y^*_{T\leftarrow 0} | X = x\big],\ \mathcal{T} = F_1^{-1} \circ F_0$$

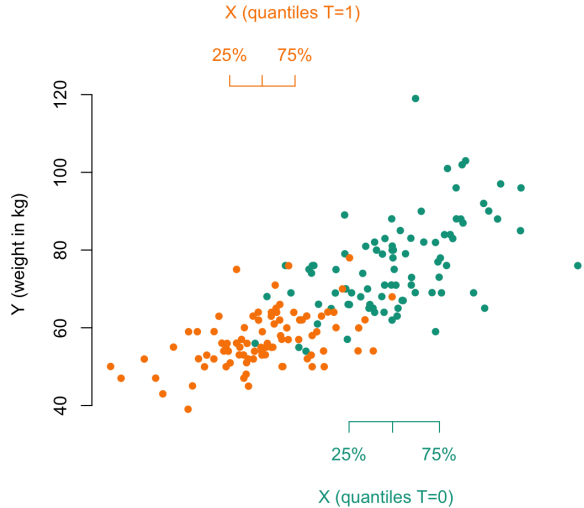
where x is considered with respect to the control group.

Quantile CATE

$(x_i, y_i) \in \mathcal{D}_0$ and $(x_j, y_j) \in \mathcal{D}_1$

Instead of x scale, visualize

$$\begin{cases} \text{top :} & \text{probability}, x_i \in \mathcal{D}_0 \\ \text{bottom :} & \text{probability}, x_j \in \mathcal{D}_1 \end{cases}$$

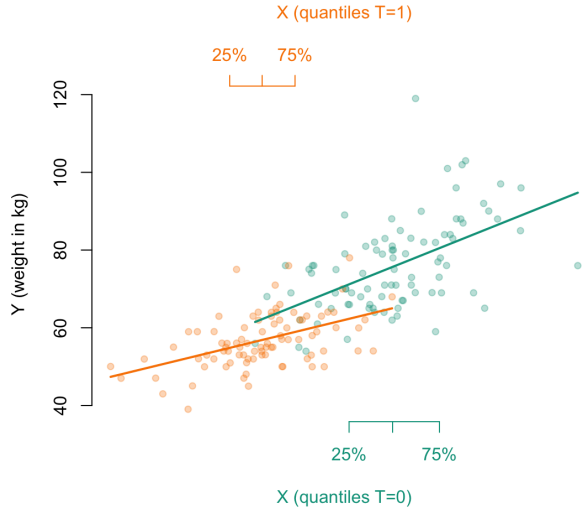


Quantile CATE

$$\begin{cases} \text{top :} & \text{probability, } x_i \in \mathcal{D}_0 \\ \text{bottom :} & \text{probability, } x_j \in \mathcal{D}_1 \end{cases}$$

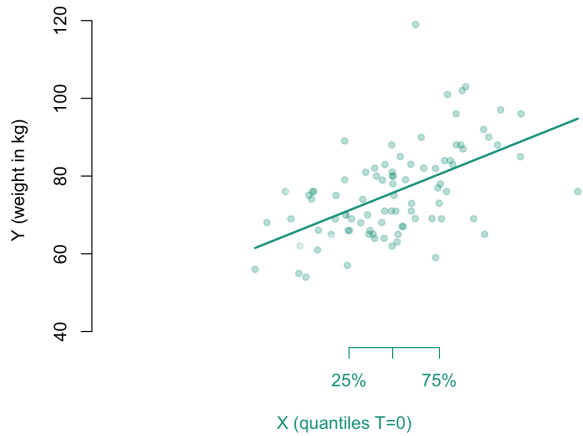
Fit two models m_0 and m_1

$$\begin{cases} m_0(x) = \mathbb{E}[Y|X = x, T = 0] \\ m_1(x) = \mathbb{E}[Y|X = x], T = 1 \end{cases}$$



Quantile CATE

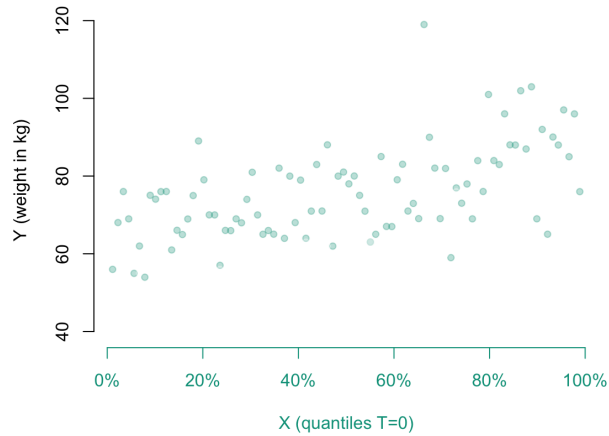
Instead of (x_i, y_i) (in $\mathcal{D}_0$)



Quantile CATE

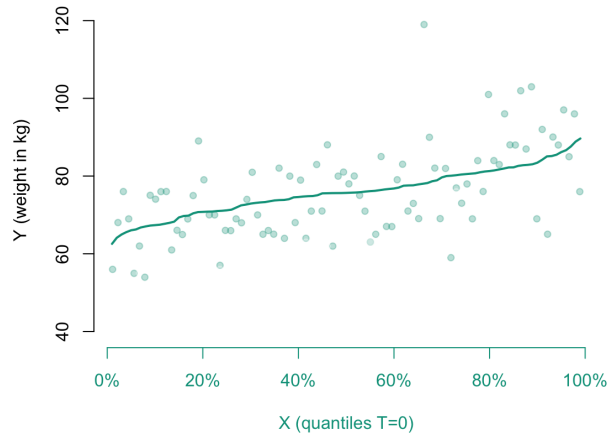
Plot $(F_0(x_i), y_i)$ (in $\mathcal{D}_0$)

Note: $F_0(x_i) \propto r_i$



Quantile CATE

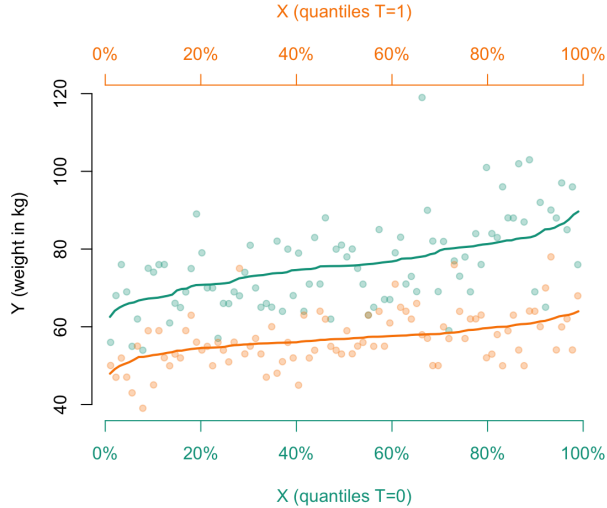
$\mu_0(u) = \mathbb{E}[Y|X = F_0^{-1}(u), T = 0]$
i.e. $\mu_0(u) = m_0(F_0^{-1}(u))$



Quantile CATE

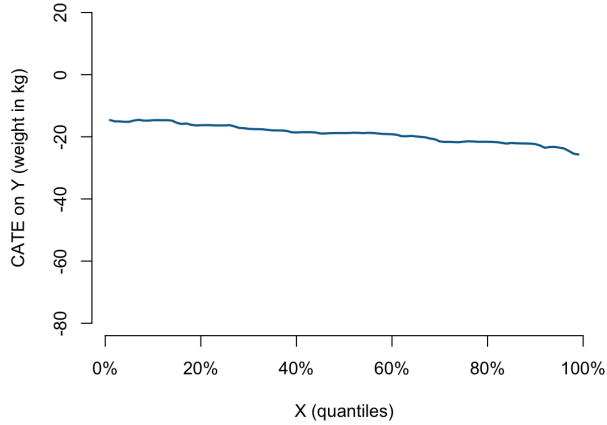
$\mu_0(u) = \mathbb{E}[Y|X = F_0^{-1}(u), T = 0]$
i.e. $\mu_0(u) = m_0(F_0^{-1}(u))$

$\mu_1(u) = \mathbb{E}[Y|X = F_1^{-1}(u), T = 0]$
i.e. $\mu_1(u) = m_1(F_1^{-1}(u))$



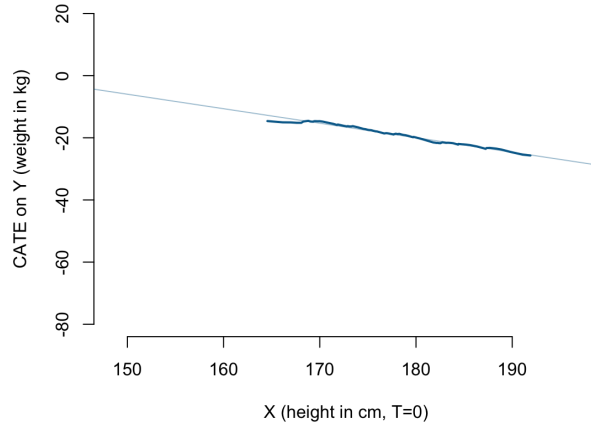
Quantile CATE

Thus QCATE(u)
$= \mathbb{E}[Y^*_{T\leftarrow 1}|X = F_1^{-1}(u)]$
$-\mathbb{E}[Y^*_{T\leftarrow 0}|X = F_0^{-1}(u)]$
for $u \in (0, 1)$.



Quantile CATE

Mutatis mutandis CATE
$= \mathbb{E}\big[Y^*_{T\leftarrow 1}|X = \mathcal{T}(x)\big]$
$\quad - \mathbb{E}\big[Y^*_{T\leftarrow 0}|X = x\big]$,
where $\mathcal{T} = F_1^{-1} \circ F_0$
and x is considered with respect to the control group.



Quantile CATE

Mutatis mutandis Sample CATE Consider two models, $\widehat{m}_0(x)$ and $\widehat{m}_1(x)$, that estimate, respectively, $\mathbb{E}[Y|X=x, T=0]$ and $\mathbb{E}[Y|X=x, T=1]$,

$$\text{SCATE}(x) = \widehat{m}_1(\widehat{\mathcal{T}}(x)) - \widehat{m}_0(x)$$

where $\widehat{\mathcal{T}}(x) = \widehat{F}_1^{-1} \circ \widehat{F}_0(x)$, with $\widehat{F}_0$ and $\widehat{F}_1$ denoting the empirical distribution functions of x conditional on $t=0$ and $t=1$, respectively.

Mutatis mutandis Gaussian CATE Consider two models two models, $\widehat{m}_0(x)$ and $\widehat{m}_1(x)$, that estimate, respectively, $\mathbb{E}[Y|X=x, T=0]$ and $\mathbb{E}[Y|X=x, T=1]$,

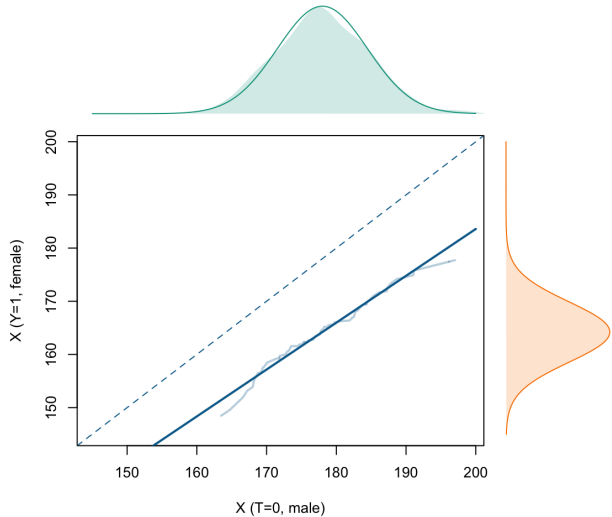
$$\text{SCATE}(x) = \widehat{m}_1(\widehat{\mathcal{T}}_{\mathcal{N}}(x)) - \widehat{m}_0(x)$$

where $\widehat{\mathcal{T}}_{\mathcal{N}}(x) = \overline{x}_1 + s_1 s_0^{-1}(x - \overline{x}_0)$, $\overline{x}_0$ and $\overline{x}_1$ being respectively the averages of x in the two sub-populations, and s_0 and s_1 the sample standard deviations.

Quantile CATE

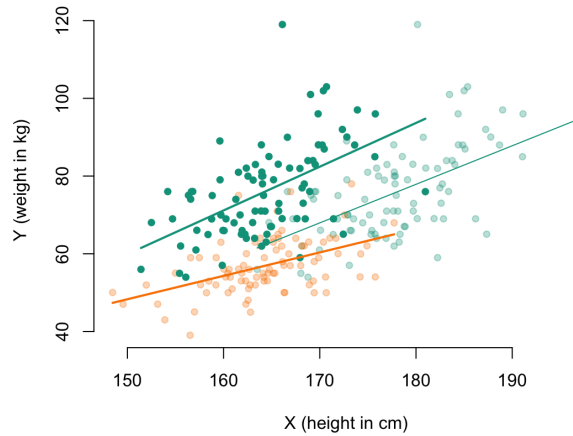
As previously,
consider $x_i \in \mathcal{D}_0$ and $x_j \in \mathcal{D}_1$

Suppose $X|T=0 \sim \mathcal{N}(\mu_0, \sigma_0^2)$
and $X|T=1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$



Quantile CATE

Previously, we had a matching,
$i \in \mathcal{D}_0 \leftrightarrow j \in \mathcal{D}_1$
We have here an explicite mapping $\mathcal{T}$

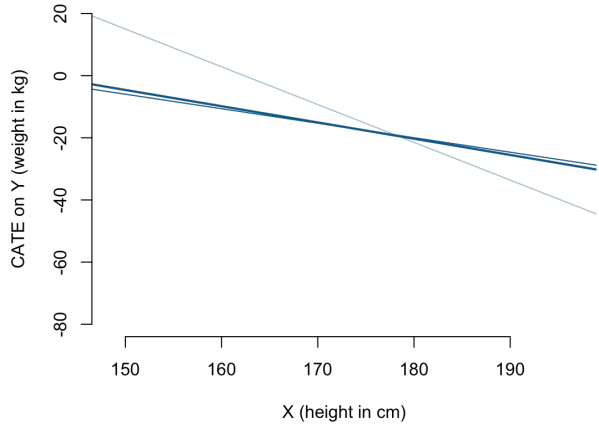


Quantile CATE

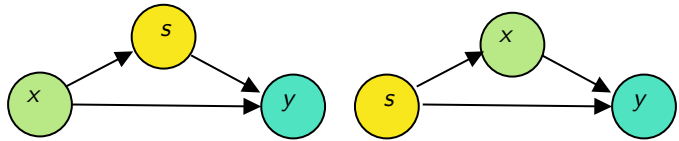
We can actually plot
$x \mapsto \widehat{m}_1(\widehat{\mathcal{T}}_{\mathcal{N}}(x)) - \widehat{m}_0(x)$

(and not only estimate it from matched samples)

Here $\widehat{m}_0$ and $\widehat{m}_1$ are linear



Mutatis mutandis counterfacturals



(a) propensity score (b) our approach

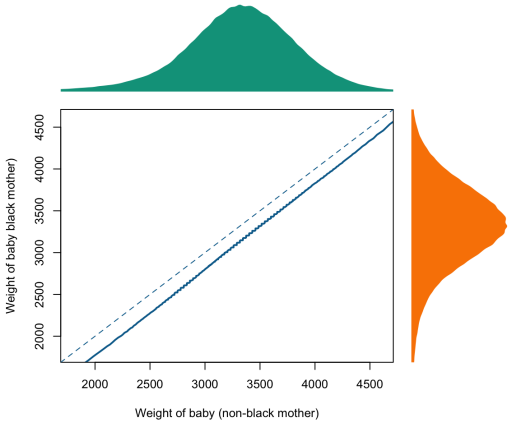
- Charpentier et al. (2023a) defined mutatis mutandis counterfactural fairness,

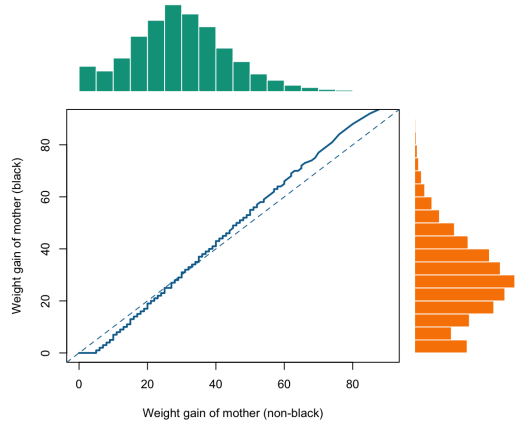
$$\mathbb{E}[Y^{\star}_{S\leftarrow \mathtt{A}}|\boldsymbol{X}=\boldsymbol{x}]=\mathbb{E}[Y^{\star}_{S\leftarrow \mathtt{B}}|\boldsymbol{X}=\mathcal{T}^{\star}(\boldsymbol{x})],\ \forall \boldsymbol{x}.$$

(probability to get surgery when delivering a baby for Black / non-Black mother)

Mutatis mutandis counterfacturals

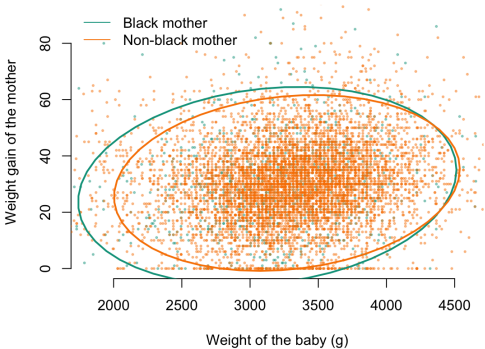
$x_1 \leftrightarrow x_1$ (newborn weight) and $x_2 \leftrightarrow x_2$ (weight gain of the mother)

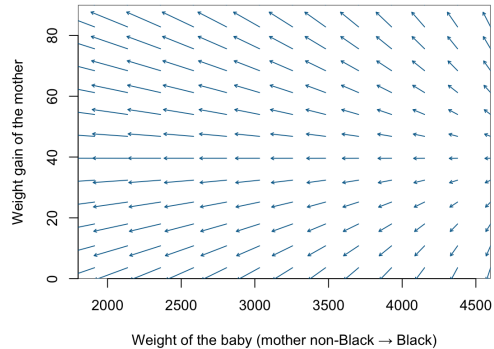




Mutatis mutandis counterfacturals

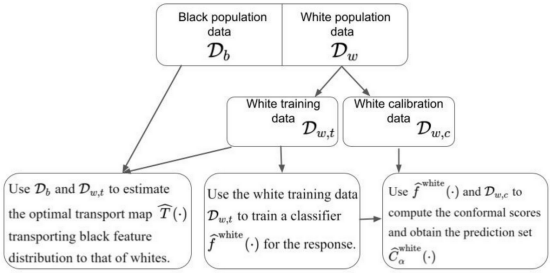
$(x_1, x_2) \leftrightarrow (x_1, x_2)$ (newborn weight, weight gain of the mother)





Back to Actuarial Justice

- See Berk et al. (2021)



Improving Fairness in Criminal Justice Algorithmic Risk Assessments Using Optimal Transport and Conformal Prediction Sets*

Richard A. Berk
University of Pennsylvania
Arun Kumar Kuchibhotla
Carnegie Mellon University
Eric Tchetgen Tchetgen
University of Pennsylvania

Abstract

In the United States and elsewhere, risk assessment algorithms are being used to help inform criminal justice decision-makers. A common intent is to forecast an offender's "future dangerousness." Such algorithms have been correctly criticized for potential unfairness, and there is an active cottage industry trying to make repairs. In this paper, we use counterfactual reasoning to consider the prospects for improved fairness when members of a disadvantaged class are treated by a risk algorithm as if they are members of an advantaged class. We combine a machine learning classifier trained in a novel manner with an optimal transport adjustment for the relevant joint probability distributions, which together provide a constructive response to claims of bias-in-bias-out. A key distinction is made between fairness claims that are empirically testable and fairness claims that are not. We then use confusion tables and conformal prediction sets to evaluate achieved fairness for estimated risk. Our data are a random sample of 300,000 offenders at their arraignments for a large metropolitan area in the United States during which decisions to release or detain are made. We show that substantial improvement in fairness can be achieved consistent with a Pareto improvement for legally protected classes.

*Cary Coglianese and Sandra Mayson provided many insightful suggestions for legal conceptions of fairness and the prospect for criminal justice reform. Emanual Candès offered several very instructive insights when commenting in this work at the Stanford/Berkeley Online Causal Inference Seminar. We also received very helpful feedback from a group of researchers at MIT and Harvard who work on causal inference. In that regard, a special thanks go to Devavrat Shah. Thanks also go to three thoughtful reviewers.

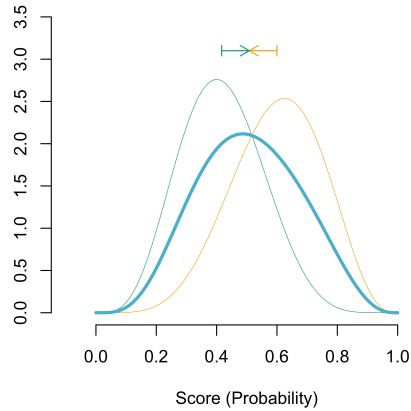
Mitigation with Wasserstein Barycenter

- If $W_2 \neq 0$ can we mitigate discrimination ?
- Use of Wasserstein Barycenter
  see Charpentier et al. (2023b)
- In Euclidean spaces

$$\boldsymbol{z}^\star = \underset{\boldsymbol{z}}{\text{argmin}} \left\{ \sum_{i=1}^{n} \omega_i d(\boldsymbol{z}, \boldsymbol{z}_i)^2 \right\},$$

- For probability measures

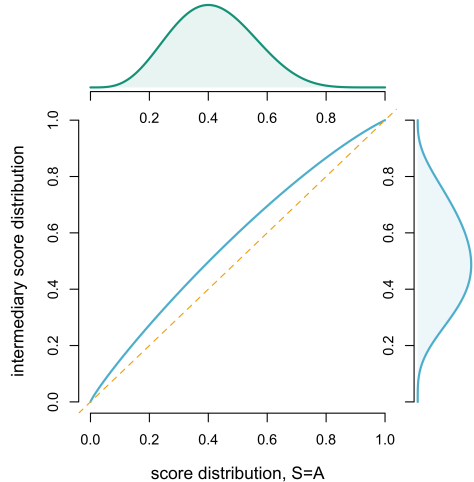
$$\mathbb{P}^\star = \underset{\mathbb{Q}}{\text{argmin}} \left\{ \sum_{i=1}^{n} \omega_i d(\mathbb{Q}, \mathbb{P}_i)^2 \right\},$$



Mitigation with Wasserstein Barycenter

Given scores $m(\boldsymbol{x}, s = \mathtt{A})$ and $m(\boldsymbol{x}, s = \mathtt{B})$, the "fair barycenter score" is

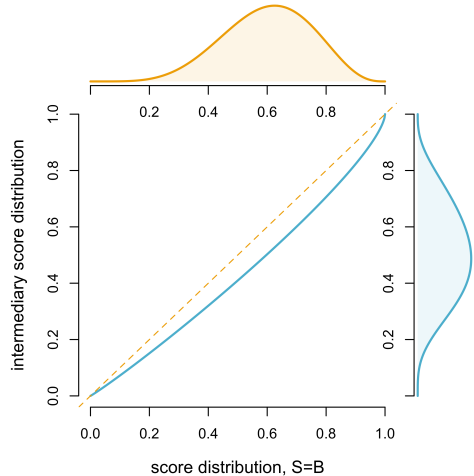
$$
\begin{aligned}
& m^\star(\boldsymbol{x}, s = \mathtt{A}) \\
= \; & \mathbb{P}[S = \mathtt{A}] \cdot m(\boldsymbol{x}, s = \mathtt{A}) \\
+ \; & \mathbb{P}[S = \mathtt{B}] \cdot F_{\mathtt{B}}^{-1} \circ F_{\mathtt{A}}(m(\boldsymbol{x}, s = \mathtt{A}))
\end{aligned}
$$



Mitigation with Wasserstein Barycenter

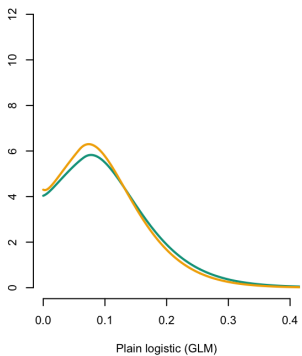
Given scores $m(\boldsymbol{x}, s = \texttt{A})$ and $m(\boldsymbol{x}, s = \texttt{B})$, the "fair barycenter score" is

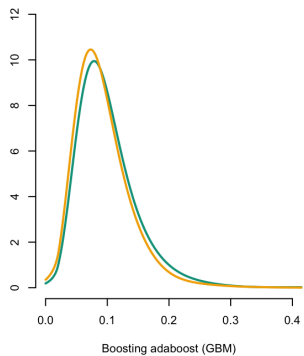
$$\begin{aligned} & m^\star(\boldsymbol{x}, s = \texttt{B}) \\ = \; & \mathbb{P}[S = \texttt{A}] \cdot F_{\texttt{A}}^{-1} \circ F_{\texttt{B}}\big(m(\boldsymbol{x}, s = \texttt{B})\big) \\ + \; & \mathbb{P}[S = \texttt{B}] \cdot m(\boldsymbol{x}, s = \texttt{B}) \end{aligned}$$

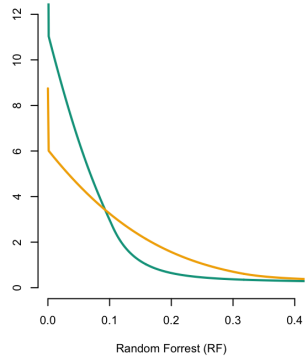


Mitigation with Wasserstein Barycenter

- If the two models are balanced, $m^\star$ is also balanced.
- Annual claim occurrence (motor insurance, Charpentier et al. (2023b))
- Three models (plain GLM, GBM, Random Forest)

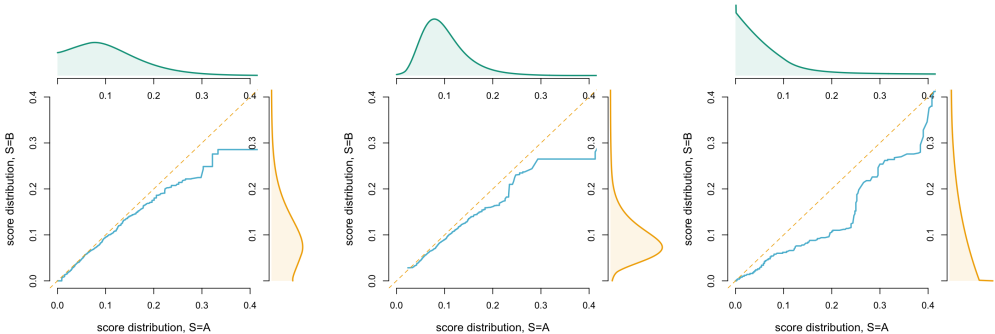






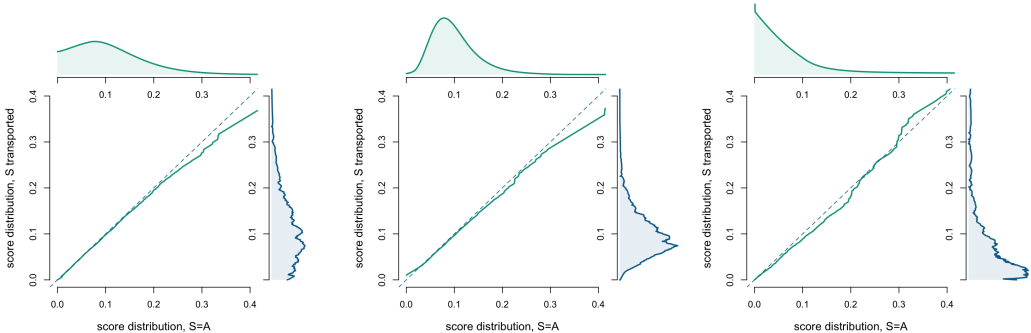
Mitigation with Wasserstein Barycenter

- Predictions are different for men ($= \texttt{A}$) and women ($S = \texttt{B}$)



- since $W_2 \neq 0$ consider post processing mitigation

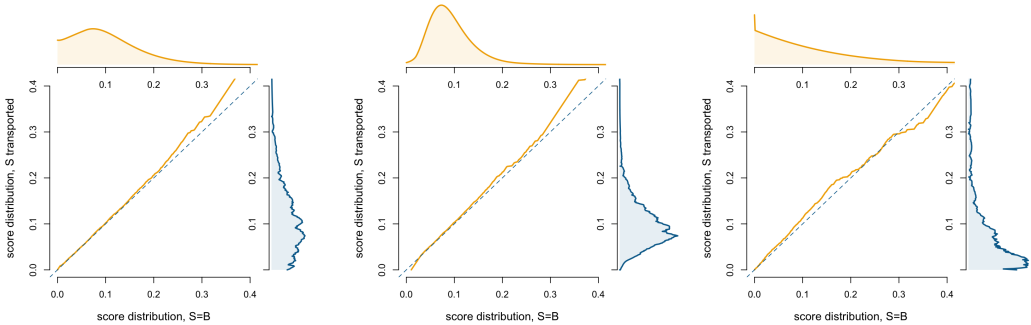
Mitigation with Wasserstein Barycenter



- Given scores $m(\boldsymbol{x}, s = \mathtt{A})$ and $m(\boldsymbol{x}, s = \mathtt{B})$, the "fair barycenter score" is

$$m^\star(\boldsymbol{x}, s = \mathtt{A}) = \mathbb{P}[S = \mathtt{A}] \cdot m(\boldsymbol{x}, s = \mathtt{A}) + \mathbb{P}[S = \mathtt{B}] \cdot F_{\mathtt{B}}^{-1} \circ F_{\mathtt{A}}(m(\boldsymbol{x}, s = \mathtt{A}))$$

Mitigation with Wasserstein Barycenter

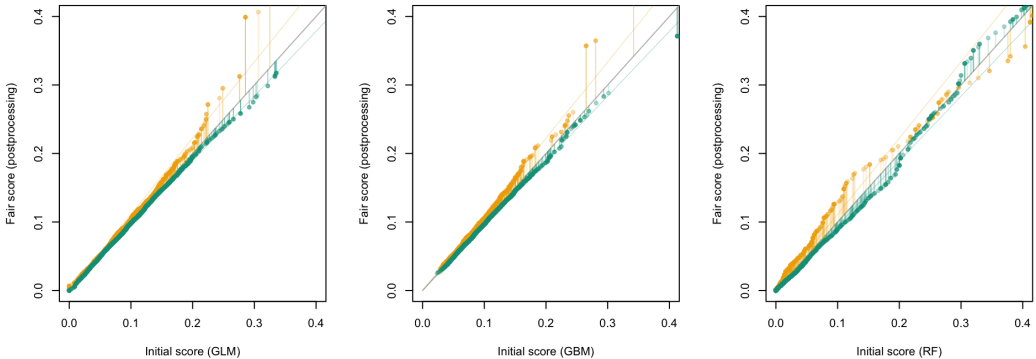


- Given scores $m(\boldsymbol{x}, s = \mathtt{A})$ and $m(\boldsymbol{x}, s = \mathtt{B})$, the "fair barycenter score" is

$$m^\star(\boldsymbol{x}, s = \mathtt{B}) = \mathbb{P}[S = \mathtt{A}] \cdot F_{\mathtt{A}}^{-1} \circ F_{\mathtt{B}}(m(\boldsymbol{x}, s = \mathtt{B})) + \mathbb{P}[S = \mathtt{B}] \cdot m(\boldsymbol{x}, s = \mathtt{B})$$

Mitigation with Wasserstein Barycenter

- We can plot $\{(m(\boldsymbol{x}_i, \texttt{A}), m^\star(\boldsymbol{x}_i, \texttt{A})\}$ and $\{(m(\boldsymbol{x}_i, \texttt{B}), m^\star(\boldsymbol{x}_i, \texttt{B})\}$



Mitigation with Wasserstein Barycenter

- Numerical values, for initial occurence probability of 5%, 10% and 20%, we have

	A (men)				B (women)			
	$\times 0.94$	GLM	GBM	RF	$\times 1.11$	GLM	GBM	RF
$m(\boldsymbol{x}) = 5\%$	4.73%	4.94%	4.80%	4.42%	5.56%	5.16%	5.25%	6.15%
$m(\boldsymbol{x}) = 10\%$	9.46%	9.83%	9.66%	8.92%	11.12%	10.38%	10.49%	12.80%
$m(\boldsymbol{x}) = 20\%$	18.91%	19.50%	18.68%	18.26%	22.25%	20.77%	21.63%	21.12%

On Multiple Sensitive Attributes

- In Euclidean spaces (see Ungar (2010)), global barycenter coincides with the barycenter of barycenters ("associativity" property).

- In Hu et al. (2023b) we consider the "folktables" dataset, and $\boldsymbol{S} = (S_1, S_2)$, race and gender.

Mitigation ? (brief conclusion)

- If it is mandatory to mitigate, there are robust techniques that can guarantee fairness
- Supreme Court Justice Harry Blackmun stated, in 1978,

"*In order to get beyond racism, we must first take account of race. There is no other way. And in order to treat some persons equally, we must treat them differently*."

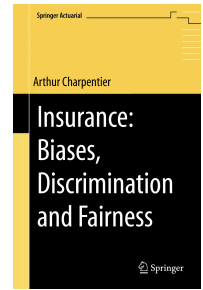
Knowlton (1978), cited in Lippert-Rasmussen (2020)

- In 2007, John G. Roberts of the U.S. Supreme Court submits

"*The way to stop discrimination on the basis of race is to stop discriminating on the basis of race*"

Sabbagh (2007) and Turner (2015)

- To go further,
  Charpentier (2023) Insurance: Biases, Discrimination and Fairness.



References

Abrevaya, J., Hsu, Y.-C., and Lieli, R. P. (2015). Estimating conditional average treatment effects. *Journal of Business & Economic Statistics*, 33(4):485–505.

Agarwal, S. (2021). Trade-offs between fairness and interpretability in machine learning. In *IJCAI 2021 Workshop on AI for Social Good*.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals and it's biased against blacks. *ProPublica*, May 23.

Avraham, R. (2017). Discrimination and insurance. In Lippert-Rasmussen, K., editor, *Handbook of the Ethics of Discrimination*, pages 335–347. Routledge.

Berk, R. A., Kuchibhotla, A. K., and Tchetgen, E. T. (2021). Improving fairness in criminal justice algorithmic risk assessments using optimal transport and conformal prediction sets. *arXiv*, 2111.09211.

Brualdi, R. A. (2006). *Combinatorial matrix classes*, volume 13. Cambridge University Press.

Bénéplanc, G., Charpentier, A., and Thourot, P. (2022). *Manuel d'assurance*. Presses Universitaires de France.

Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I. G., and Cosentini, A. C. (2022). A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 12(1):1–21.

References

Charpentier, A. (2014). *Computational actuarial science with R*. CRC press.

Charpentier, A. (2023). *Insurance: biases, discrimination and fairness*. Springer Verlag.

Charpentier, A., Flachaire, E., and Gallic, E. (2023a). Causal inference with optimal transport. In Thach, N. N., Kreinovich, V., Ha, D. T., and Trung, N. D., editors, *Optimal Transport Statistics for Economics and Related Topics*. Springer Verlag.

Charpentier, A., Hu, F., and Ratz, P. (2023b). Mitigating discrimination in insurance with wasserstein barycenters. *BIAS, 3rd Workshop on Bias and Fairness in AI, International Workshop of ECML PKDD*.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. *arXiv*, 1701.08230.

Crossney, K. B. (2016). Redlining. *https://philadelphiaencyclopedia.org/essays/redlining/*.

Cunningham, S. (2021). *Causal inference*. Yale University Press.

Dehejia, R. H. and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448):1053–1062.

Denuit, M. and Charpentier, A. (2004). *Mathématiques de l'assurance non-vie: Tome I Principes fondamentaux de théorie du risque*. Economica.

References

Denuit, M. and Charpentier, A. (2005). *Mathématiques de l'assurance non-vie: Tome II Tarification et provisionnement*. Economica.

Dieterich, W., Mendoza, C., and Brennan, T. (2016). Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc*, 7(7.4):1.

Dressel, J. and Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580.

Feeley, M. and Simon, J. (1994). Actuarial justice: The emerging new criminal law. *The futures of criminology*, 173:174.

Feller, A., Pierson, E., Corbett-Davies, S., and Goel, S. (2016). A computer program used for bail and sentencing decisions was labeled biased against blacks. it's actually not that clear. *The Washington Post*, October 17.

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis*, 15(3):199–236.

Hu, F., Ratz, P., and Charpentier, A. (2023a). Fairness in multi-task learning via wasserstein barycenters. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases – ECML PKDD*.

References

Hu, F., Ratz, P., and Charpentier, A. (2023b). A sequentially fair mechanism for multiple sensitive attributes. *ArXiv*, 2309.06627.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Kearns, M. and Roth, A. (2019). *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press.

Knowlton, R. E. (1978). Regents of the university of california v. bakke. *Arkansas Law Review*, 32:499.

Kranzberg, M. (1986). Technology and history:" kranzberg's laws". *Technology and culture*, 27(3):544–560.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 4066–4076. NIPS.

Lippert-Rasmussen, K. (2020). *Making sense of affirmative action*. Oxford University Press.

Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*.

Pearl, J. et al. (2009). Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146.

References

Pearl, J. and Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic books.

Perry, W. L. (2013). *Predictive policing: The role of crime forecasting in law enforcement operations*. Rand Corporation.

Rhynhart, R. (2020). Mapping the legacy of structural racism in philadelphia. *Philadelphia, Office pf the Controller*.

Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics*, pages 159–183.

Sabbagh, D. (2007). *Equality and transparency: A strategic perspective on affirmative action in American law*. Springer.

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

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1):1.

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

Turner, R. (2015). The way to stop discrimination on the basis of race. *Stanford Journal of Civil Rights & Civil Liberties*, 11:45.

Ungar, A. A. (2010). *Barycentric calculus in Euclidean and hyperbolic geometry: A comparative introduction*. World Scientific.