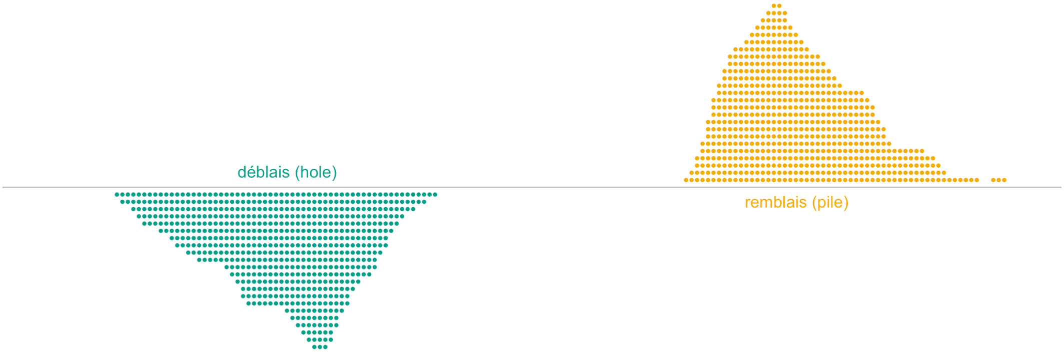
Causal Inference and Counterfactuals with Optimal Transport With Applications in Fairness and Discrimination

Arthur Charpentier[1], Emmanuel Flachaire[2], Ewen Gallic[2]
François Hu[1] & Philipp Ratz[1]

SINCLAIR (Saclay INdustrial Collaborative Laboratory for Artificial Intelligence Research), July 2023

[1] UQAM, Canada, [2] Aix-Marseille University (AMSE), France

Optimal Transport

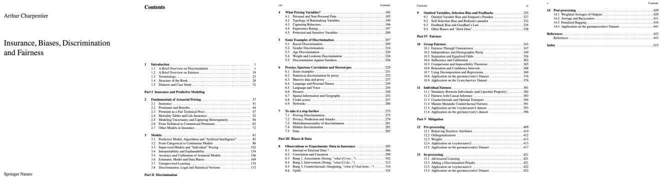


Monge (1781), *Mémoire sur la Théorie des Déblais et des Remblais*

Motivation

- "*at the core of insurance business lies discrimination between risky and non-risky insureds*", 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)

See Charpentier (2022, 2023a) for more details.


Arthur Charpentier

Insurance, Biases, Discrimination and Fairness

Springer Nature

Contents

Motivation

Classical **group fairness** concepts, such as demographic parity (independence, $\widehat{Y} \perp\!\!\!\perp S$), equalized odds (separation, $\widehat{Y} \perp\!\!\!\perp S \mid Y$) or predictive parity (sufficiency, $Y \perp\!\!\!\perp S \mid \widehat{Y}$), see Hardt et al. (2016), Berk et al. (2017) or Corbett-Davies et al. (2017)

But one could also consider **individual fairness** concepts
"*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.* ", Castelnovo et al. (2022)

"what would have been the decision if that individual had a different gender/race ?"

Bonus if some discrimination is observed, is it possible to **mitigate** it...? Grari et al. (2022) and Hu et al. (2023a,b)

Introduction

Duivesteijn and Feelders (2008), Luong et al. (2011)
A decision $\widehat{Y}$ is fair if

$$d_y(\widehat{y}_i, \widehat{y}_j) \leq d_x(\boldsymbol{x}_i, \boldsymbol{x}_j),\ \forall i, j = 1, \cdots, n,$$

whatever the sensitive/protected attribute S (0 or 1).

Kusner et al. (2017) A decision is fair if the prediction in the real world is the same as the prediction in the counterfactual world

$$\mathbb{E}[Y^\star_{S\leftarrow 0}|\boldsymbol{X}=\boldsymbol{x}] = \mathbb{E}[Y^\star_{S\leftarrow 1}|\boldsymbol{X}=\boldsymbol{x}],\ \forall \boldsymbol{x},$$

where $Y^\star_{S\leftarrow 0}$ and $Y^\star_{S\leftarrow 1}$ denote "potential outcomes".

Why would $\boldsymbol{x}$ and $\boldsymbol{x}$ remain the same if S changes ?

(Gordaliza et al. (2019) and Black et al. (2020))

Preambule: Ceteris Paribus & Mutatis Mutandis

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 $\boldsymbol{X} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$.

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

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

Causal inference

- ## Angrist and Pischke (2009, 2014)

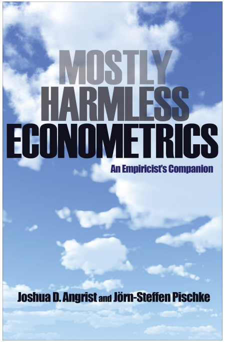
A naive comparison of averages by hospitalization status tells us something about potential outcomes, though not necessarily what we want to know. The comparison of average health conditional on hospitalization status is formally linked to the average causal effect by the equation below:

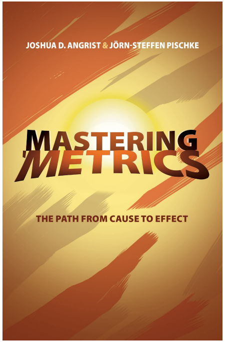
$$\underbrace{E[\mathrm{Y}_i|\mathrm{D}_i=1]-E[\mathrm{Y}_i|\mathrm{D}_i=0]}_{\text{Observed difference in average health}} = \underbrace{E[\mathrm{Y}_{1i}|\mathrm{D}_i=1]-E[\mathrm{Y}_{0i}|\mathrm{D}_i=1]}_{\text{average treatment effect on the treated}} + \underbrace{E[\mathrm{Y}_{0i}|\mathrm{D}_i=1]-E[\mathrm{Y}_{0i}|\mathrm{D}_i=0]}_{\text{selection bias}}$$

The term

$$E[\mathrm{Y}_{1i}|\mathrm{D}_i=1]-E[\mathrm{Y}_{0i}|\mathrm{D}_i=1]=E[\mathrm{Y}_{1i}-\mathrm{Y}_{0i}|\mathrm{D}_i=1]$$

is the *average causal effect of hospitalization on those who were hospitalized.* This term captures the averages difference between the health of the hospitalized, $E[\mathrm{Y}_{1i}|\mathrm{D}_i=1]$, and what would have happened to *them* had they not been hospitalized, $E[\mathrm{Y}_{0i}|\mathrm{D}_i=1]$. The observed difference in health status however, adds to this causal effect a term called *selection bias.* This term is the difference in average Y_{0i} between those who





Causal inference

"Ladder of causation" from Pearl (2009)

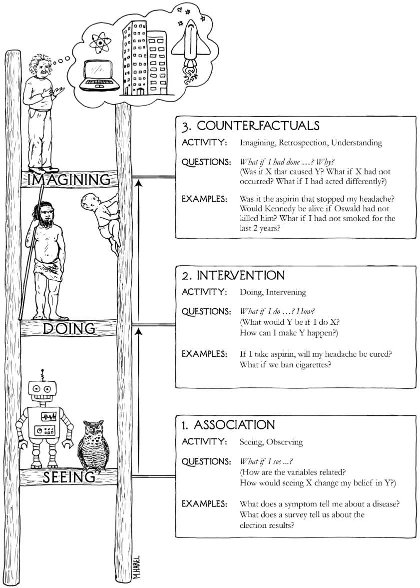
3. **Counterfactuals**
(Imagining, "*what if I had done...*")

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 ?



Causal inference

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

(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

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

Causal inference

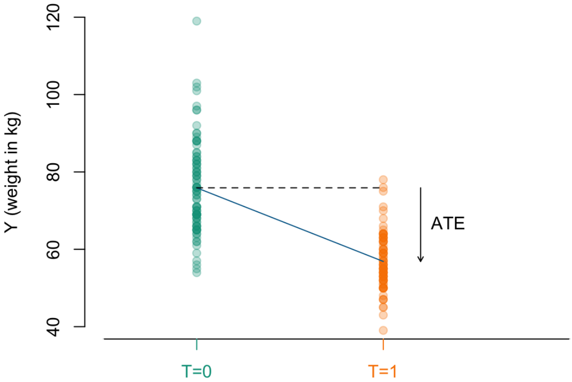
Let $y_i \in \mathcal{D}_0$ and $y_j \in \mathcal{D}_1$

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

(too) simple sample estimate

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

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



Causal inference

Consider a third variable x

CATE

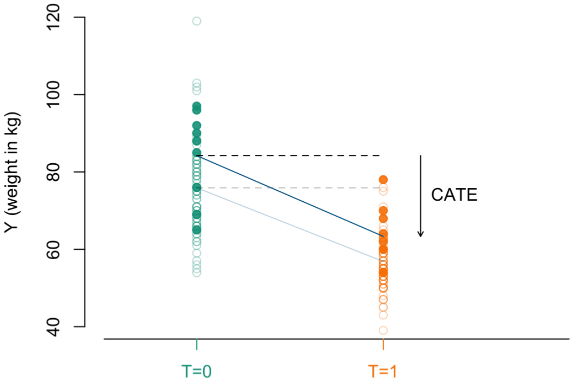
Conditional ATE is CATE(x)

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

A natural estimate would be

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

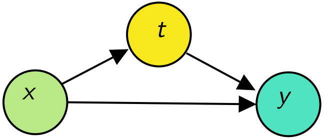
as in Abrevaya et al. (2015).



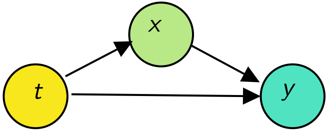
Causal inference

This CATE(x) is a **ceteris paribus CATE** that does not take into account possible correlation between x, y and t.

Classical approach is based on the use of the propensity score, implying that x might influence the treatment t



(a) propensity score



(b) our approach

Mutatis mutandis CATE

Given $\boldsymbol{x}$ "in the reference group" (0)

$$\text{CATE}(\boldsymbol{x}) = \mathbb{E}\left[Y^*_{T\leftarrow 1}\middle|\boldsymbol{x}_{T\leftarrow 1}\right] - \mathbb{E}\left[Y^*_{T\leftarrow 0}\middle|\boldsymbol{x}\right]$$

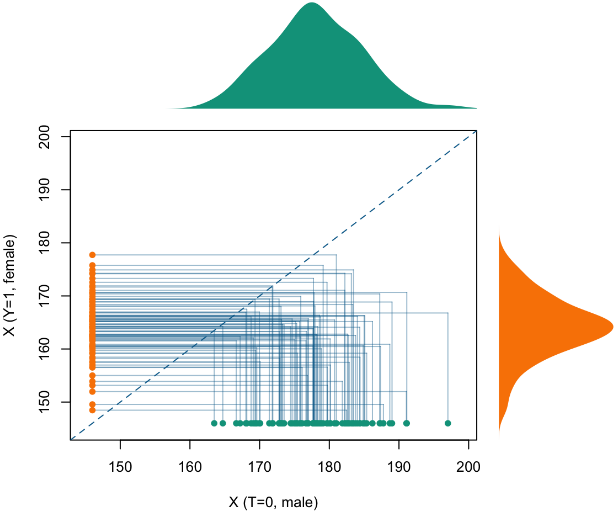
Matching

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"



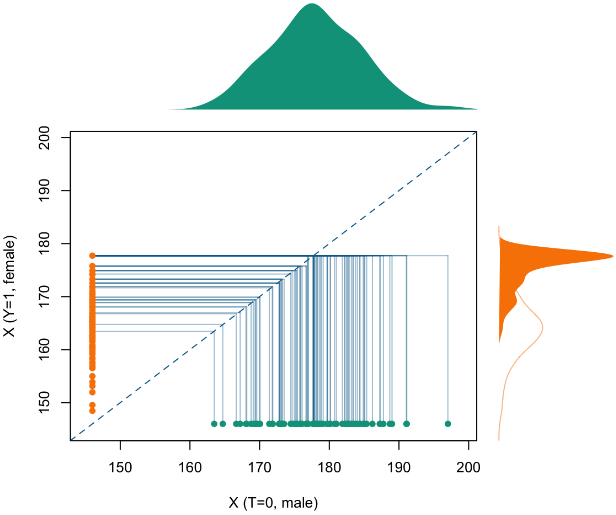
Matching

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

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

and keeping observation from $\mathcal{D}_1$ will distort distribution of x,

$$\{x_j, j \in \mathcal{D}_1\} \overset{\mathcal{L}}{\neq} \{x_{j_i^\star}, i \in \mathcal{D}_0\}$$

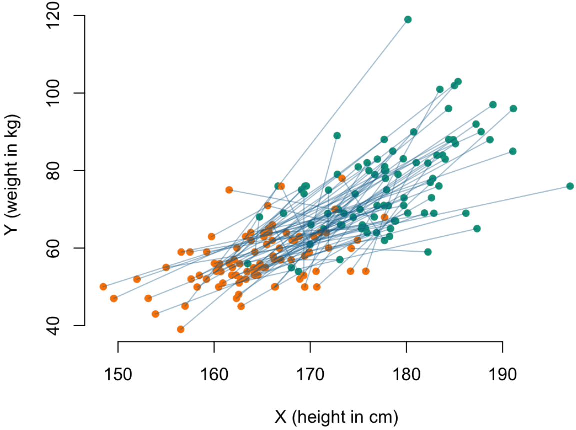


Matching

Done properly
(removing observation),
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 $(\mathcal{D}_1)$...

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

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



Matching

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

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

Here scatter-plot • of

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

and linear regression ———

Horizontal line – – – is ATE

Matching

Algorithm 1 SATE, matching case (classical)

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

$\mathcal{D}_0 \leftarrow$ subset of $\mathcal{D}$ when $t = 0$ (size n) shuffled, with indices i

$\mathcal{D}_1 \leftarrow$ subset of $\mathcal{D}$ when $t = 1$ (size n), with indices j

for $i = 1$ to n **do**

$\quad j_i^\star = \underset{j:t_j=1}{\text{argmin}}\{d(\boldsymbol{x}_i, \boldsymbol{x}_j)\}$ in $\mathcal{D}_1$,

$\quad d_i \leftarrow y_{j_i^\star}^{(1)} - y_i^{(0)}$

$\quad$ remove observation $j_i^\star$ from $\mathcal{D}_1$

end for

$\text{SATE} \leftarrow \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} d_i$

Matching

Algorithm 2 SCATE, matching case (classical)

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

$\mathcal{D}_0 \leftarrow$ subset of $\mathcal{D}$ when $t=0$ (size n) shuffled, with indices i

$\mathcal{D}_1 \leftarrow$ subset of $\mathcal{D}$ when $t=1$ (size n), with indices j

$V_{\boldsymbol{x}}^k \leftarrow$ list of k nearest neighbors of $\boldsymbol{x}_i$'s in $\mathcal{D}_0$ close to $\boldsymbol{x}$

for $i=1,2,\cdots,n$ **do**

  $j_i^\star = \underset{j:t_j=1}{\text{argmin}}\{d(\boldsymbol{x}_i,\boldsymbol{x}_j)\}$ in $\mathcal{D}_1$,

  $d_i \leftarrow y_{j_i^\star}^{(1)} - y_i^{(0)}$

  remove observation $j_i^\star$ from $\mathcal{D}_1$

end for

$\text{SCATE}(\boldsymbol{x}) \leftarrow \dfrac{1}{k}\sum_{i\in V_{\boldsymbol{x}}^k} d_i$

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	·	$1 \rightarrow 11$
2	·	·	·	1	·	·	$2 \rightarrow 10$
3	·	·	·	·	1	·	$3 \rightarrow 11$
4	·	·	·	1	·	·	$4 \rightarrow 10$
5	·	·	·	1	·	·	$5 \rightarrow 10$
6	·	·	·	·	·	1	$6 \rightarrow 12$

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 3 SATE, optimal matching case

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

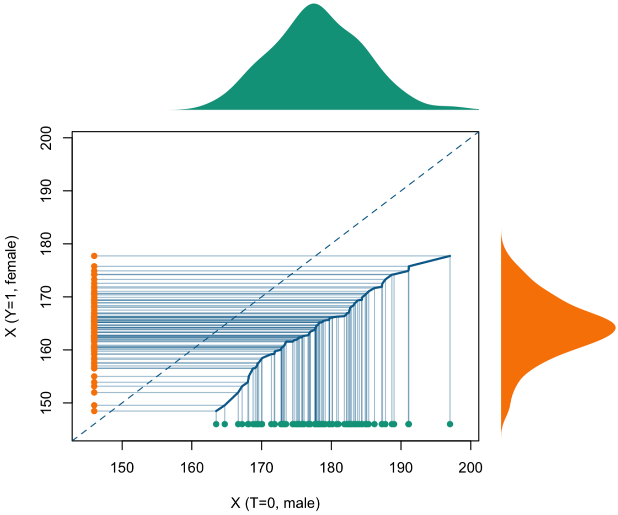
1: $\mathcal{D}_0 \leftarrow$ subset of $\mathcal{D}$ when $t = 0$ (size n) shuffled, with indices i
2: $\mathcal{D}_1 \leftarrow$ subset of $\mathcal{D}$ when $t = 1$ (size n), with indices j
3: $C \leftarrow$ matrix $n \times n$, $C_{i,j} = d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ between points in $\mathcal{D}_0$ and $\mathcal{D}_1$
4: $P^* \leftarrow$ solution of Problem (1)
5: 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 $r_i = s_j$.

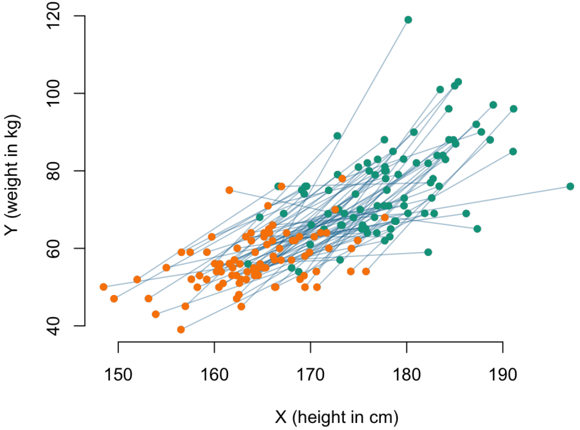


Optimal Matching

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

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

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,

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

Here scatter-plot ● of

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

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 $U\left(\mathbf{1}_0, \dfrac{n_0}{n_1}\mathbf{1}_1\right)$

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

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	·	·	·	·

Optimal Coupling

Algorithm 4 SATE, optimal coupling 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_0) shuffled, with indices i

2: $\mathcal{D}_1 \leftarrow$ subset of $\mathcal{D}$ when $t = 1$ (size n_1), with indices j

3: $C \leftarrow$ matrix $n_0 \times n_1$, $C_{i,j} = d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ between points in $\mathcal{D}_0$ and $\mathcal{D}_1$

4: $P^* \leftarrow$ solution of Problem (4)

5: $\text{SATE} \leftarrow \dfrac{1}{n_0} \displaystyle\sum_{i=1}^{n} y_i^0 - P_i^{*\top} \boldsymbol{y}^1$

Optimal Coupling

Algorithm 5 SCATE, optimal coupling 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_0) shuffled, with indices i

2: $\mathcal{D}_1 \leftarrow$ subset of $\mathcal{D}$ when $t = 1$ (size n_1), with indices j

3: $V_{\boldsymbol{x}}^k \leftarrow$ list of k nearest neighbors of $\boldsymbol{x}_i$'s in $\mathcal{D}_0$ close to $\boldsymbol{x}$

4: $C \leftarrow$ matrix $n_0 \times n_1$, $C_{i,j} = d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ between points in $\mathcal{D}_0$ and $\mathcal{D}_1$

5: $P^* \leftarrow$ solution of Problem (4)

6: $\text{SCATE}(\boldsymbol{x}) \leftarrow \dfrac{1}{k} \displaystyle\sum_{i \in V_{\boldsymbol{x}}^k} y_i^0 - P_i^{*\top} \boldsymbol{y}^1$

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

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

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

Mutatis mutandis CATE

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

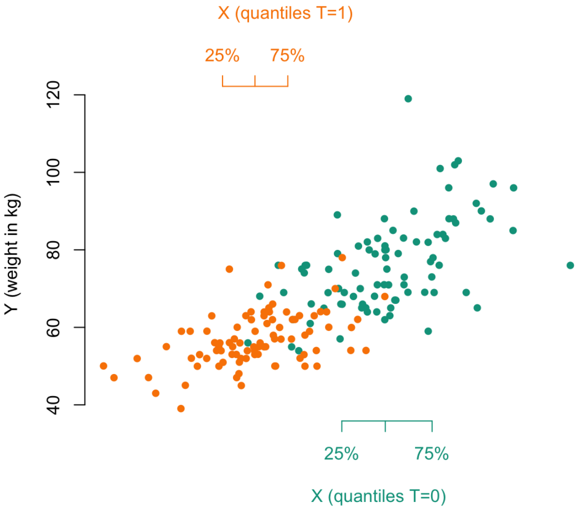
where x is considered with respect to the control group.

Quantile CATE

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

Instead of x scale, visualize

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

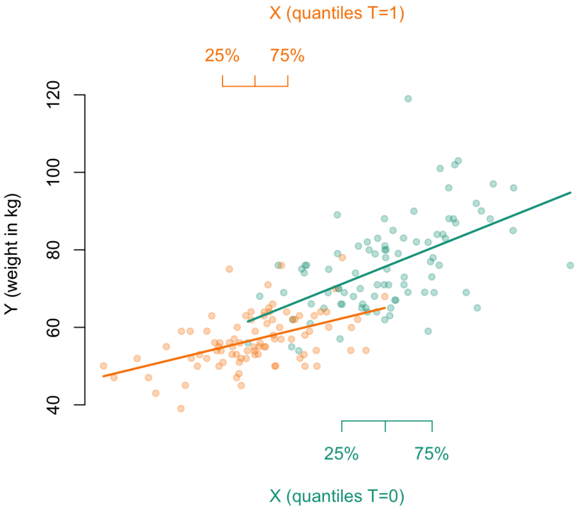


Quantile CATE

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

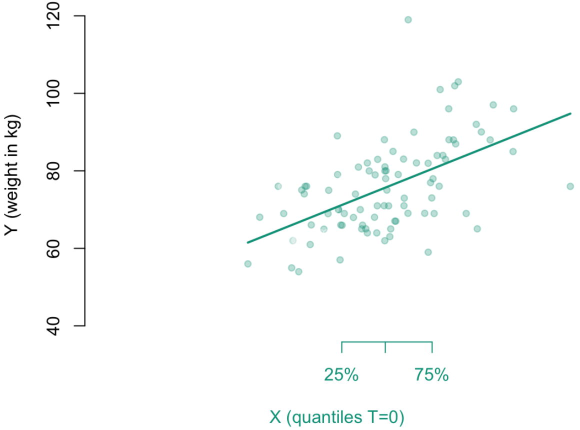
Fit two models m_0 and m_1

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



Quantile CATE

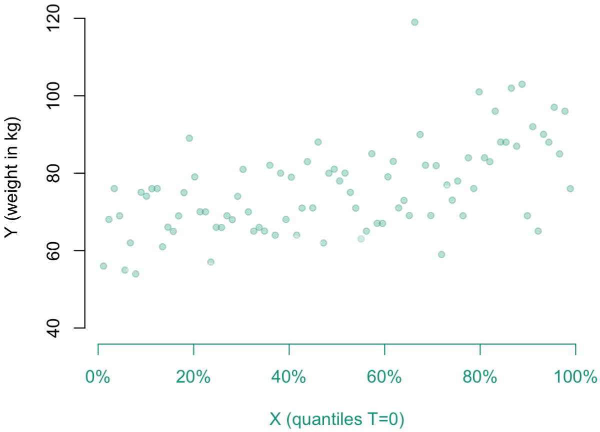
Instead of (x_i, y_i) (in $\mathcal{D}_0$)



Quantile CATE

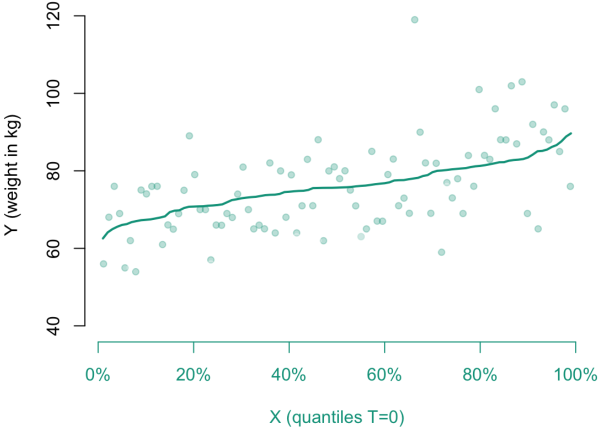
Plot $(F_0(x_i), y_i)$ (in $\mathcal{D}_0$)

Note: $F_0(x_i) \propto r_i$



Quantile CATE

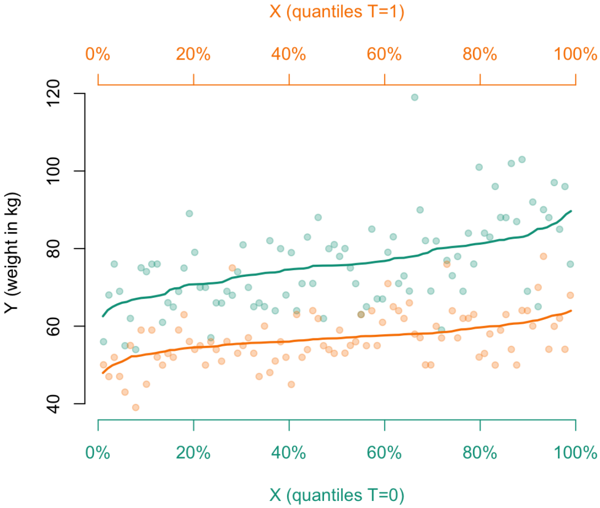
$$\mu_0(u) = \mathbb{E}[Y|X = F_0^{-1}(u), T = 0]$$
i.e. $\mu_0(u) = m_0(F_0^{-1}(u))$



Quantile CATE

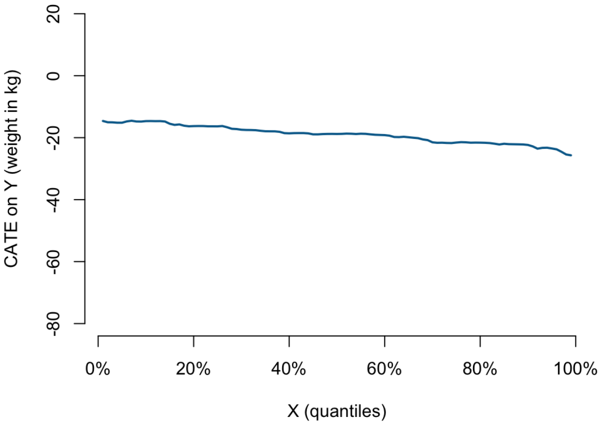
$$\mu_0(u) = \mathbb{E}[Y|X = F_0^{-1}(u), T = 0]$$
i.e. $\mu_0(u) = m_0(F_0^{-1}(u))$

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



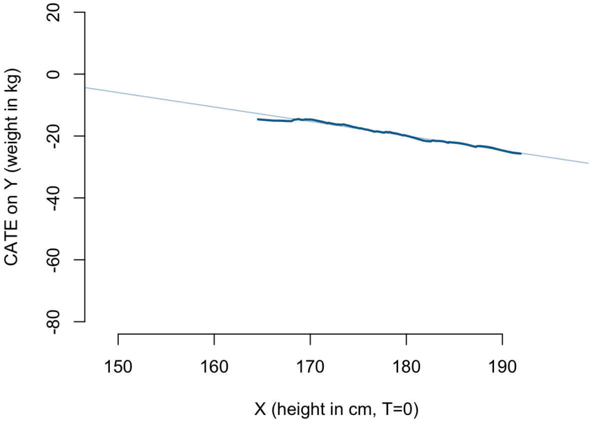
Quantile CATE

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



Quantile CATE

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



Mutatis Mutandis SCATE

This yields the following proper definition

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.

Gaussian x

Note that a simple parametric transformation can be obtained, based on the assumption that X conditional on T is Gaussian. More precisely, if

$$X_0 \overset{\mathcal{L}}{=} X|t=0 \sim \mathcal{N}(\mu_0, \Sigma_0) \text{ and } X_1 \overset{\mathcal{L}}{=} X|t=1 \sim \mathcal{N}(\mu_1, \Sigma_1)$$

,

$$\underbrace{\mu_1 + \sigma_1 \cdot \frac{X_0 - \mu_0}{\sigma_0}}_{\mathcal{T}(X_0)} \overset{\mathcal{L}}{=} X_1$$

Quite naturally, we can use a Gaussian approximation for $\mathcal{T}$,

Mutatis Mutandis Gaussian SCATE

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 $\mathbb{E}[Y|X=x,T=1]$,

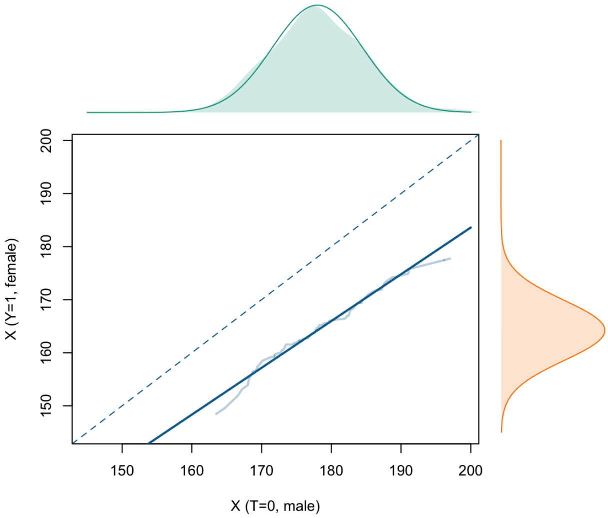
$$\text{SCATE}(x)=\widehat{m}_1(\widehat{\mathcal{T}}_{\mathcal{N}}(x))-\widehat{m}_0(x)$$

where $\widehat{\mathcal{T}}_{\mathcal{N}}(x)=\overline{x}_1+s_1s_0^{-1}(x-\overline{x}_0)$, $\overline{x}_0$ 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.

Mutatis Mutandis Gaussian SCATE

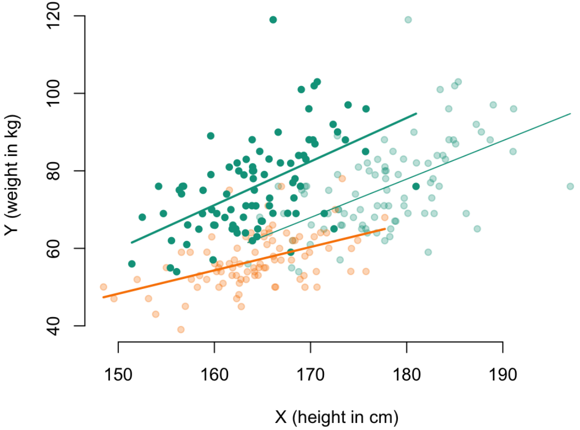
As previously,
consider $x_i \in \mathcal{D}_0$ and $x_j \in \mathcal{D}_1$

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



Mutatis Mutandis Gaussian SCATE

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

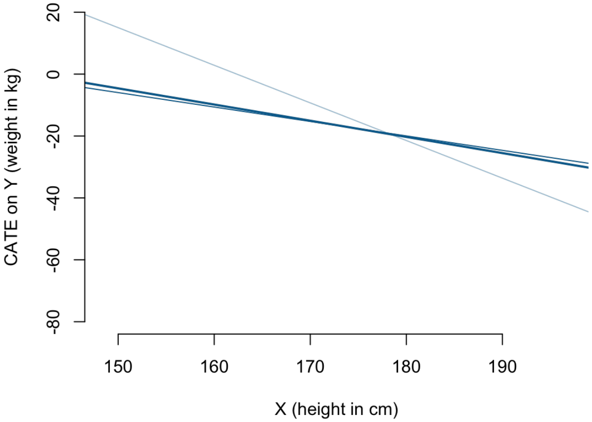


Mutatis Mutandis Gaussian SCATE

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

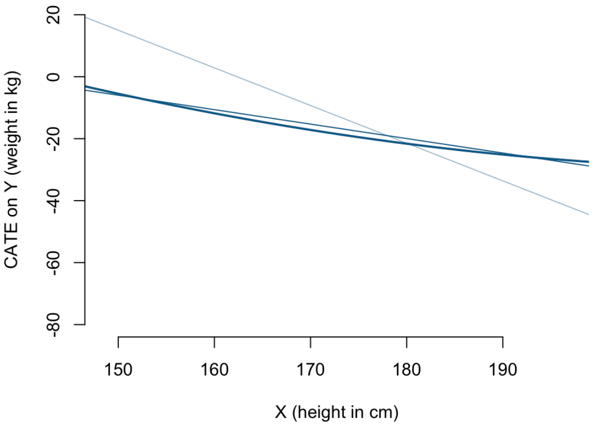


Mutatis Mutandis Gaussian SCATE

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 non-linear



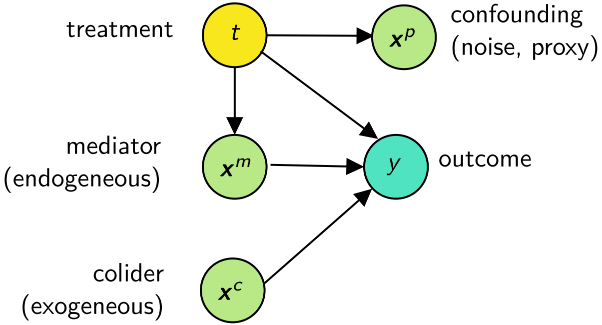
CATE($\boldsymbol{x}$)

We can consider a multivariate $\boldsymbol{x}$

But we should not be transport all components of $\boldsymbol{x}$

Mutadis Mutandis CATE would be
$m_1(\mathcal{T}(\boldsymbol{x}^m), \boldsymbol{x}^c, \cancel{\boldsymbol{x}}^p) - m_0(\boldsymbol{x}^m, \boldsymbol{x}^c, \cancel{\boldsymbol{x}}^p)$

- $\boldsymbol{x}^p$ is correlated with y
  not causal, so not in $m(\cdot)$
- $\boldsymbol{x}^c$ is not influenced by t
- $\boldsymbol{x}^m$ is influenced by t
  it should be transported



See Charpentier et al. (2023), detailed example with Gaussian Structural Causal Model.

CATE($\boldsymbol{x}$)

Given $\mathcal{T}:\mathbb{R}^k\to\mathbb{R}^k$, define the "*push-forward*" measure,

$$\mathbb{P}_1(A)=\mathcal{T}_{\#}\mathbb{P}_0(A)=\mathbb{P}_0\big(\mathcal{T}^{-1}(A)\big),\ \forall A\subset\mathbb{R}^k.$$

An optimal transport $\mathcal{T}^\star$ (in Brenier's sense, from Brenier (1991), see Villani (2009) or Galichon (2016)) from $\mathbb{P}_0$ towards $\mathbb{P}_1$ will be solution of

$$\mathcal{T}^\star\in\underset{\mathcal{T}:\mathcal{T}_{\#}\mathbb{P}_0=\mathbb{P}_1}{\text{arginf}}\left\{\int_{\mathbb{R}^k}\gamma(\boldsymbol{x}-\mathcal{T}(\boldsymbol{x}))d\mathbb{P}_0(\boldsymbol{x})\right\},$$

that we can write

$$\min_{\nu}\int_{\mathbb{R}^k\times\mathbb{R}^k}\underbrace{\gamma(\boldsymbol{x},\boldsymbol{y})}_{=C}\underbrace{\nu(d\boldsymbol{x},d\boldsymbol{y})}_{=P},\ \text{where } \nu \text{ is a coupling with margins } \mathbb{P}_0 \text{ and } \mathbb{P}_1,$$

for some cost function $\gamma:\mathbb{R}^k\times\mathbb{R}^k\to\mathbb{R}_+$.

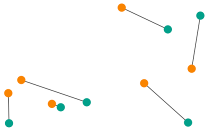
Matching & Coupling

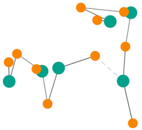
Previous problems were actually defined in any dimension. For matching, solve

$$\underset{P\in\mathcal{P}}{\text{argmin}}\Big\{\langle P, C\rangle\Big\} \text{ or } \underset{P\in\mathcal{P}}{\text{argmin}}\left\{\sum_{i,j} P_{i,j}C_{i,j}\right\}, \tag{3}$$

and, for the coupling case

$$P^\star \in \underset{P\in U_{n_0,n_1}}{\text{argmin}}\Big\{\langle P, C\rangle\Big\} \text{ or } \underset{P\in U_{n_0,n_1}}{\text{argmin}}\left\{\sum_{i=1}^{n_0}\sum_{j=1}^{n_1} P_{i,j}C_{i,j}\right\}. \tag{4}$$





Gaussian x

In the case where $\boldsymbol{X}|t=1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\boldsymbol{X}|t=0 \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, there is an explicit expression for the optimal transport, which is simply an affine map (see Villani (2003) for more details). In the univariate case, $x_1 = \mathcal{T}^\star_{\mathcal{N}}(x_0) = \mu_1 + \dfrac{\sigma_1}{\sigma_0}(x_0 - \mu_0)$, while in the multivariate case, an analogous expression can be derived:

$$\boldsymbol{x}_1 = \mathcal{T}^\star_{\mathcal{N}}(\boldsymbol{x}_0) = \boldsymbol{\mu}_1 + \boldsymbol{A}(\boldsymbol{x}_0 - \boldsymbol{\mu}_0),$$

where $\boldsymbol{A}$ is a symmetric positive matrix that satisfies $\boldsymbol{A}\boldsymbol{\Sigma}_0\boldsymbol{A} = \boldsymbol{\Sigma}_1$, which has a unique solution given by $\boldsymbol{A} = \boldsymbol{\Sigma}_0^{-1/2}\big(\boldsymbol{\Sigma}_0^{1/2}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_0^{1/2}\big)^{1/2}\boldsymbol{\Sigma}_0^{-1/2}$, where $\boldsymbol{M}^{1/2}$ is the square root of the square (symmetric) positive matrix $\boldsymbol{M}$ based on the Schur decomposition ($\boldsymbol{M}^{1/2}$ is a positive symmetric matrix), as described in Higham (2008).

Gaussian SCATE

Mutatis mutandis Gaussian CATE

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

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

where $\widehat{\mathcal{T}}_{\mathcal{N}}(\boldsymbol{x})=\overline{\boldsymbol{x}}_1+\widehat{\boldsymbol{A}}(\boldsymbol{x}-\overline{\boldsymbol{x}}_0)$, with $\overline{\boldsymbol{x}}_0$ and $\overline{\boldsymbol{x}}_1$ being, respectively, the averages of x in the two sub-populations, and $\widehat{\boldsymbol{A}}=\widehat{\boldsymbol{\Sigma}}_0^{-1/2}\big(\widehat{\boldsymbol{\Sigma}}_0^{1/2}\widehat{\boldsymbol{\Sigma}}_1\widehat{\boldsymbol{\Sigma}}_0^{1/2}\big)^{1/2}\widehat{\boldsymbol{\Sigma}}_0^{-1/2}$ where $\widehat{\boldsymbol{\Sigma}}_0$ and $\widehat{\boldsymbol{\Sigma}}_1$ denote the sample variance.

Gaussian SCATE

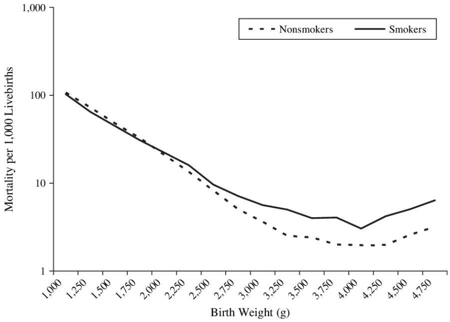
Algorithm 6 SCATE, Gaussian transport

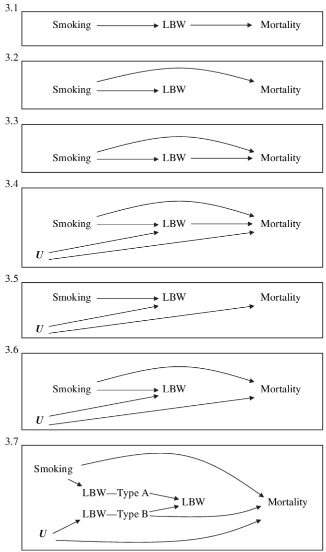
Require: dataset $\mathcal{D} = \{(y_i, \boldsymbol{x}_i, t_i)\}$

1: $\mathcal{D}_0 \leftarrow$ subset of $\mathcal{D}$ when $t = 0$ (size n_0)
2: $\mathcal{D}_1 \leftarrow$ subset of $\mathcal{D}$ when $t = 1$ (size n_1)
3: $\widehat{m}_0 \leftarrow$ model to predict y based on $\boldsymbol{x}$, trained on $\mathcal{D}_0$
4: $\widehat{m}_1 \leftarrow$ model to predict y based on $\boldsymbol{x}$, trained on $\mathcal{D}_1$
5: estimate moments of $\boldsymbol{x}_t$'s $\hat{\boldsymbol{\mu}}_0$, $\hat{\boldsymbol{\mu}}_1$, $\hat{\boldsymbol{\Sigma}}_0$ and $\hat{\boldsymbol{\Sigma}}_1$,
6: $\hat{\boldsymbol{A}} \leftarrow \hat{\boldsymbol{\Sigma}}_0^{-1/2}(\hat{\boldsymbol{\Sigma}}_0^{1/2}\hat{\boldsymbol{\Sigma}}_1\hat{\boldsymbol{\Sigma}}_0^{1/2})^{1/2}\hat{\boldsymbol{\Sigma}}_0^{-1/2}$
7: $\tilde{\boldsymbol{x}} \leftarrow \hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{A}}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_0)$
8: $\text{SCATE}_{\mathcal{N}}(\boldsymbol{x}) \leftarrow \widehat{m}_1(\tilde{\boldsymbol{x}}) - \widehat{m}_0(\boldsymbol{x})$

Application on newborn infant deliveries (natural, or not)

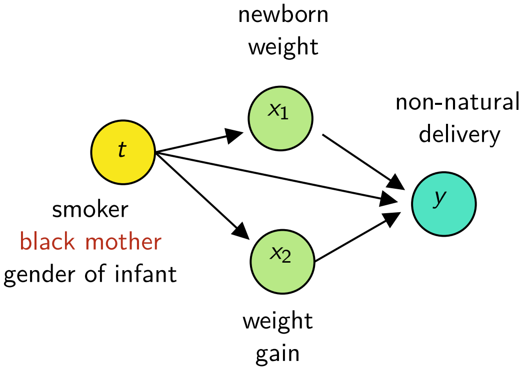
"Low birth weight paradox", introduced by Wilcox (1993, 2001): Low birth weight (LBW) of babies x is strongly associated with increased neonatal mortality y. LBW infants born to mothers who smoke $t = 1$ usually have lower mortality rates than LBW infants born to nonsmoking mothers $t = 0$.





See Hernández-Díaz et al. (2006) and Wilcox (2006)

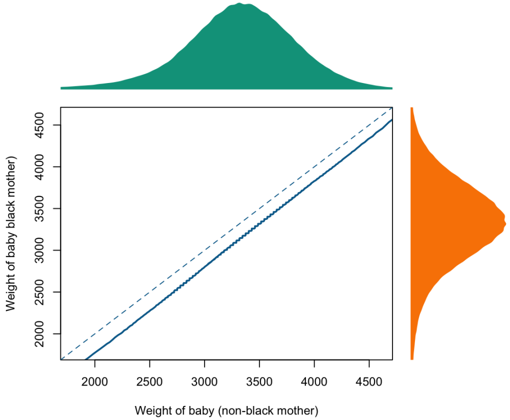
Application on newborn infant deliveries (natural, or not)

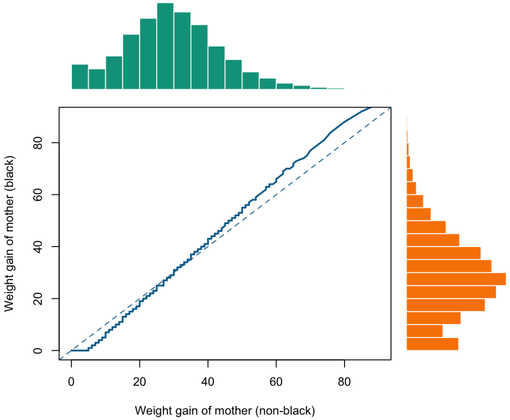


See Charpentier et al. (2023) for more details. Here y is a binary variable.

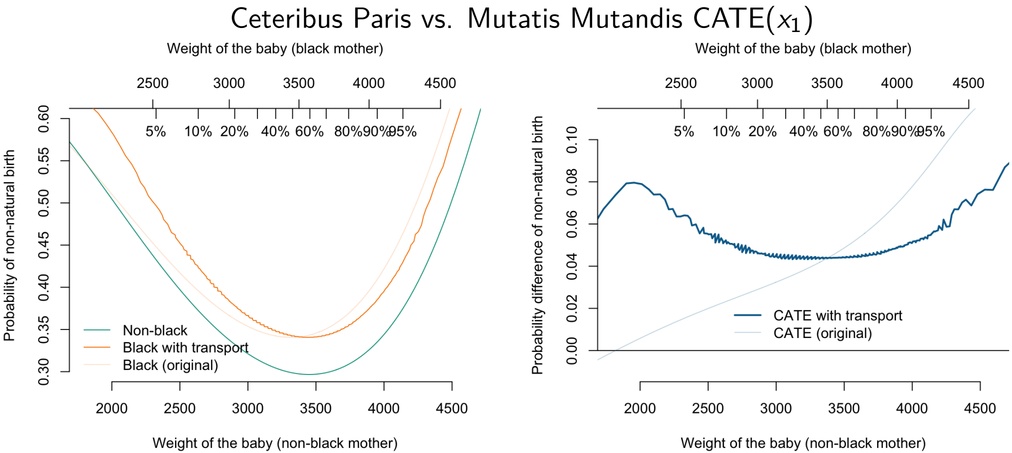
Application on newborn infant deliveries (natural, or not)

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



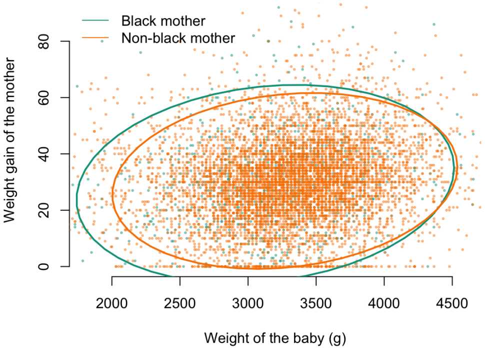


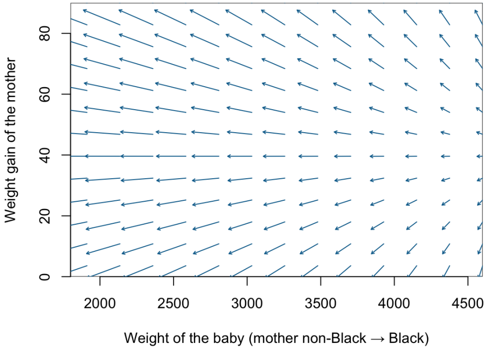
Application on newborn infant deliveries (natural, or not)



Application on newborn infant deliveries (natural, or not)

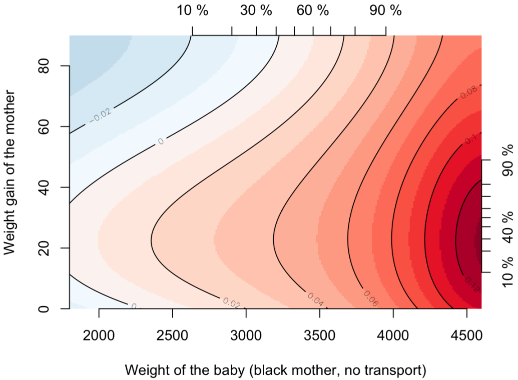
$(x_1, x_2) \leftrightarrow (x_1, x_2)$ (newborn weight, weight gain of the mother)

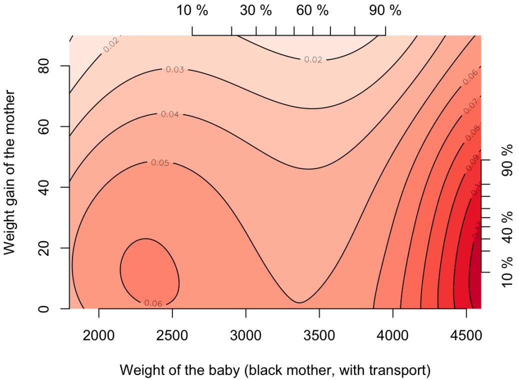




Application on newborn infant deliveries (natural, or not)

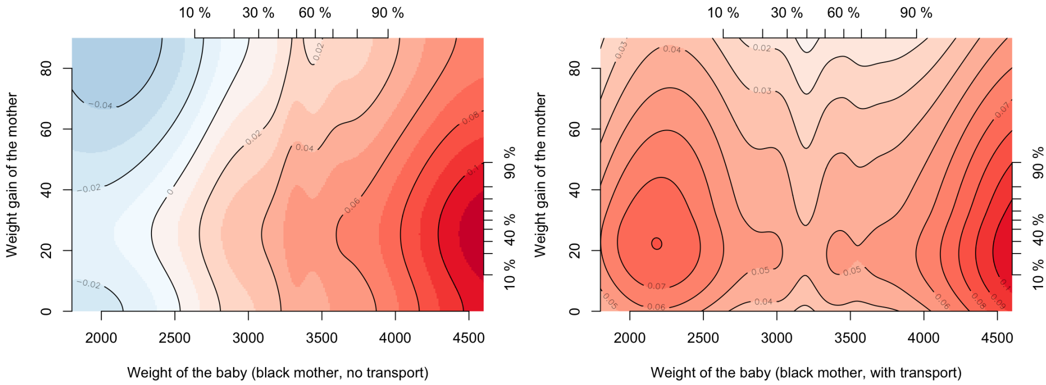
Ceteribus Paris vs. Mutatis Mutandis CATE(x_1, x_2)





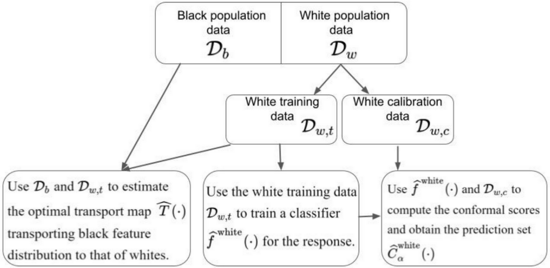
Application on newborn infant deliveries (natural, or not)

Ceteribus Paris vs. Mutatis Mutandis CATE(x_1, x_2)



Sidenote (1)

- 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 on 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.

Sidenote (2)

- Hallin et al. (2021)

Theorem 2.1 (McCann 1985) *Let* P_1 *and* P_2 *denote two distributions in* $\mathcal{P}_d$. *Then, (i) the class of functions*

$$\nabla\Psi_{\mathrm{P}_1;\mathrm{P}_2} := \{\nabla\psi \mid \psi : \mathbb{R}^d \to \mathbb{R} \text{ convex, lower semi-continuous, and such that } \nabla\psi\#\mathrm{P}_1 = \mathrm{P}_2\} \tag{2.1}$$

is not empty; (ii) if $\nabla\psi'$ *and* $\nabla\psi''$ *are two elements of* $\nabla\Psi_{\mathrm{P}_1;\mathrm{P}_2}$, *they coincide* P_1*-a.s.;*[9] *(iii) if* P_1 *and* P_2 *have finite moments of order two, any element of* $\nabla\Psi_{\mathrm{P}_1;\mathrm{P}_2}$ *is an optimal quadratic transport pushing* P_1 *forward to* P_2.

Definition 2.2 Call $\mathbf{F}_\pm := \nabla\phi$ the *center-outward distribution function* of $\mathrm{P} \in \mathcal{P}_d$.

Definition 2.3 Call *empirical center-outward distribution function* any of the mappings $\mathbf{F}_\pm^{(n)} : (\mathbf{Z}_1^{(n)}, \dots, \mathbf{Z}_n^{(n)}) \mapsto (\mathbf{F}_\pm^{(n)}(\mathbf{Z}_1^{(n)}), \dots, \mathbf{F}_\pm^{(n)}(\mathbf{Z}_n^{(n)})) =: \mathbf{F}_\pm^{(n)}(\mathbf{Z}^{(n)})$ satisfying

$$\sum_{i=1}^{n} \left\|\mathbf{Z}_i^{(n)} - \mathbf{F}_\pm^{(n)}(\mathbf{Z}_i^{(n)})\right\|^2 = \min_{T \in \mathcal{T}} \sum_{i=1}^{n} \left\|\mathbf{Z}_i^{(n)} - T(\mathbf{Z}_i^{(n)})\right\|^2 \tag{2.10}$$

or, equivalently,

$$\sum_{i=1}^{n} \left\|\mathbf{Z}_i^{(n)} - \mathbf{F}_\pm^{(n)}(\mathbf{Z}_i^{(n)})\right\|^2 = \min_{\pi} \sum_{i=1}^{n} \left\|\mathbf{Z}_{\pi(i)}^{(n)} - \mathbf{F}_\pm^{(n)}(\mathbf{Z}_i^{(n)})\right\|^2 \tag{2.10}$$

where the set $\{\mathbf{F}_\pm^{(n)}(\mathbf{Z}_i^{(n)}) \mid i = 1, \dots, n\}$ consists of the n points of the augmented grid and π ranges over the $n!$ possible permutations of $\{1, 2, \dots, n\}$.

The Annals of Statistics
2021, Vol. 49, No. 2, 1139–1165
https://doi.org/10.1214/20-AOS1996

DISTRIBUTION AND QUANTILE FUNCTIONS, RANKS AND SIGNS IN DIMENSION *d*: A MEASURE TRANSPORTATION APPROACH

BY MARC HALLIN[1], EUSTASIO DEL BARRIO[2,*], JUAN CUESTA-ALBERTOS[3] AND CARLOS MATRÁN[2,†]

[1]*ECARES and Département de Mathématique, Université libre de Bruxelles, Belgium, mhallin@ulb.ac.be*
[2]*IMUVA and Departamento de Estadística e Investigación Operativa, Universidad de Valladolid, Spain,* [*]*tasio@eio.uva.es;* [†]*carlos.matran@uva.es*
[3]*Departamento de Matemáticas, Estadística y Computación, Facultad de Ciencias, Universidad de Cantabria, Spain, juan.cuesta@unican.es*

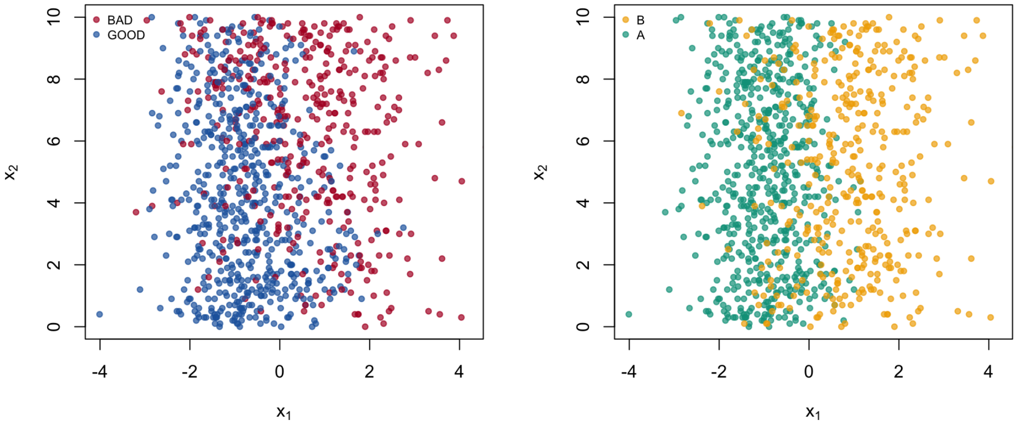
Unlike the real line, the real space $\mathbb{R}^d$, for $d \geq 2$, is not canonically ordered. As a consequence, such fundamental univariate concepts as quantile and distribution functions and their empirical counterparts, involving ranks and signs, do not canonically extend to the multivariate context. Palliating that lack of a canonical ordering has been an open problem for more than half a century, generating an abundant literature and motivating, among others, the development of statistical depth and copula-based methods. We show that, unlike the many definitions proposed in the literature, the measure transportation-based ranks and signs introduced in Chernozhukov, Galichon, Hallin and Henry (*Ann. Statist.* **45** (2017) 223–256) enjoy all the properties that make univariate ranks a successful tool for semiparametric inference. Related with those ranks, we propose a new *center-outward* definition of multivariate distribution and quantile functions, along with their empirical counterparts, for which we establish a Glivenko–Cantelli result. Our approach is based on McCann (*Duke Math. J.* **80** (1995) 309–323) and our results do not require any moment assumptions. The resulting ranks and signs are shown to be strictly distribution-free and essentially maximal ancillary in the sense of Basu (*Sankhyā* **21** (1959) 247–256) which, in semiparametric models involving noise with unspecified density, can be interpreted as a finite-sample form of semiparametric efficiency. Although constituting a sufficient summary of the sample, empirical center-outward distribution functions are defined at observed values only. A continuous extension to the entire d-dimensional space, yielding smooth empirical quantile contours and sign curves while preserving the essential monotonicity and Glivenko–Cantelli features of the concept, is provided. A numerical study of the resulting empirical quantile contours is conducted.

1. Introduction. Unlike the real line, the real space $\mathbb{R}^d$, for $d \geq 2$, is not canonically ordered. As a consequence, such fundamental concepts as quantile and distribution functions, which are strongly related to the ordering of the observation space, and their empirical counterparts—ranks and empirical quantiles—playing, in dimension $d = 1$, a fundamental role in statistical inference, do not canonically extend to dimension $d \geq 2$.

Of course, a classical concept of distribution function—the familiar one, based on marginal orderings—does exist. That concept, from a probabilistic point of view, does the job of characterizing the underlying distribution. However, the corresponding quantile function does not mean much (see, e.g., Genest and Rivest (2001)), and the corresponding empirical versions

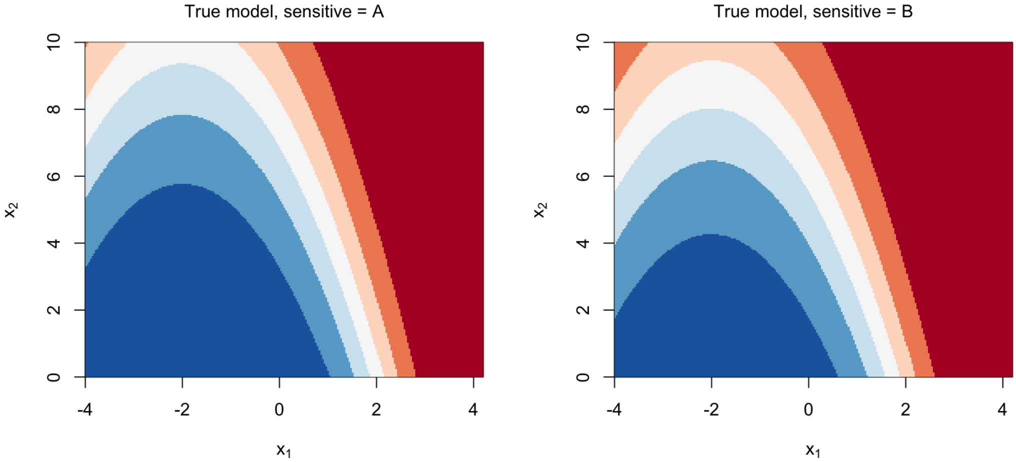
Received May 2019; revised June 2020.
MSC2020 subject classifications. Primary 62G30; secondary 62B05.
Key words and phrases. Multivariate distribution function, multivariate quantiles, multivariate ranks, multivariate signs, Glivenko–Cantelli theorem, Basu theorem, distribution-freeness, ancillarity, cyclical monotonicity.

Application on a toy example, from Charpentier (2023a)



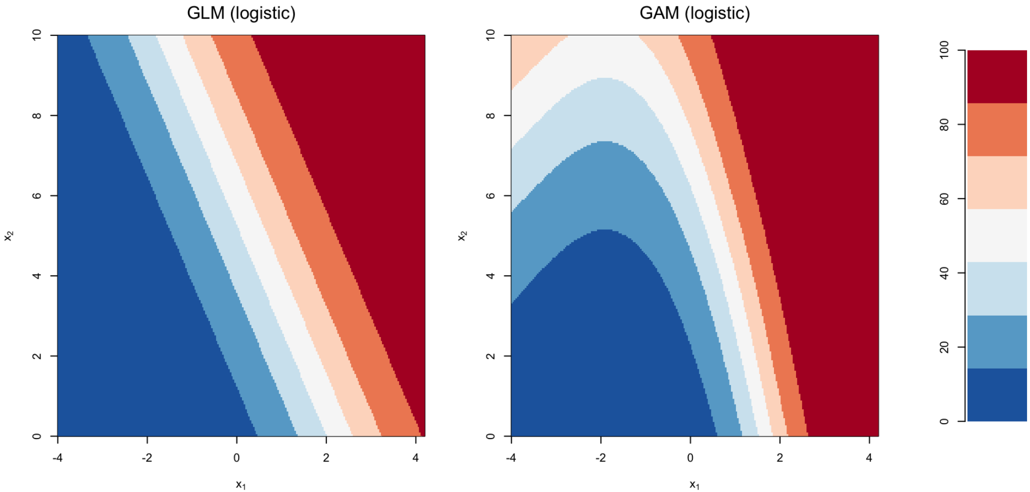
Scatterplot $(x_{1,i}, x_{2,i})$, $y_i \in \{0, 1\}$ (left) $s_i \in \{A, B\}$ (right)

Application on a toy example, from Charpentier (2023a)



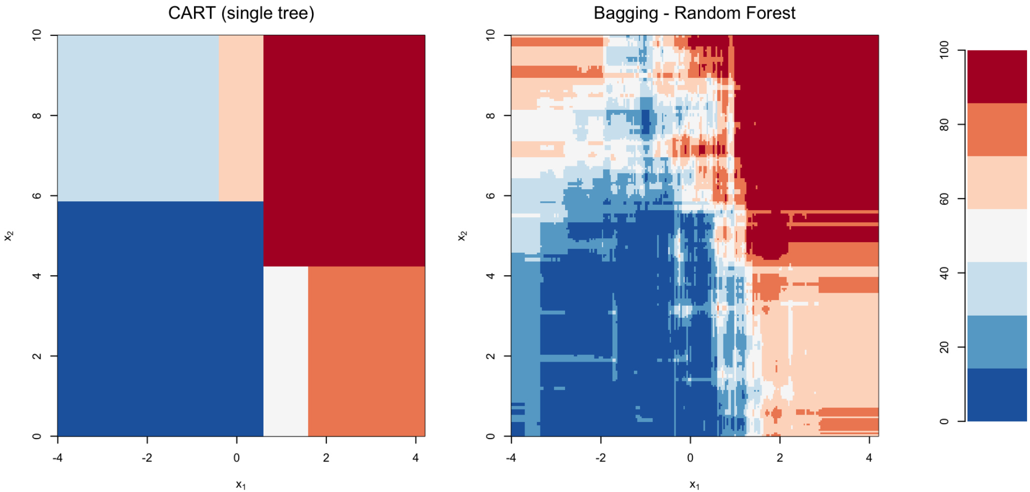
$\mu(\boldsymbol{x}, s) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}, S = s]$, with $\mu(\boldsymbol{x}, s = A)$ (left) $\mu(\boldsymbol{x}, s = B)$ (right)

Application on a toy example, from Charpentier (2023a)



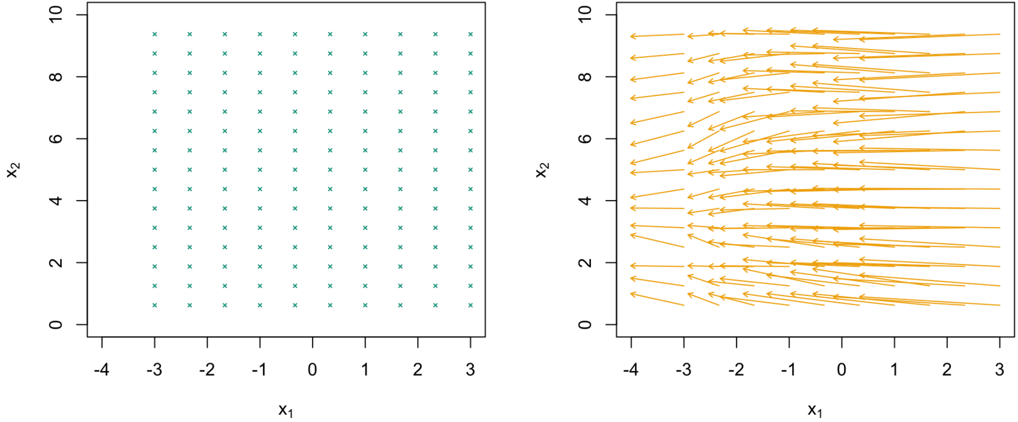
Blind fitted models $\widehat{m}(\boldsymbol{x})$, logistic regression (left) GAM (right)

Application on a toy example, from Charpentier (2023a)



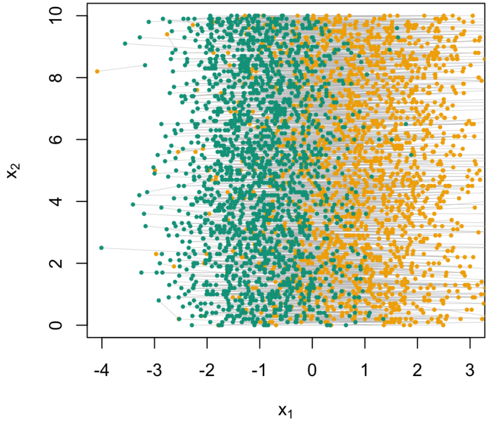
Blind fitted models $\widehat{m}(\boldsymbol{x})$, classification tree (left) random forest (right)

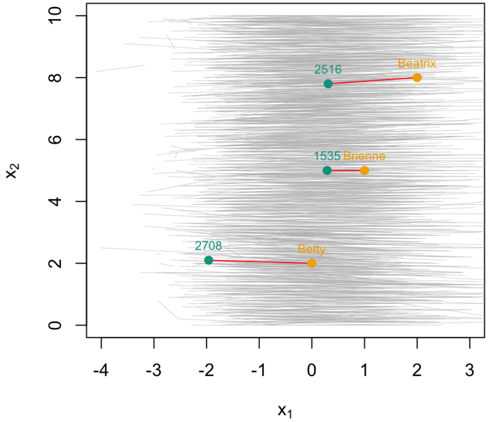
Application on a toy example, from Charpentier (2023a)



Counterfactual optimal transport, $\boldsymbol{x} = (x_1, x_2) \to \mathcal{T}(x_1, x_2)$, A (left) and B (right)

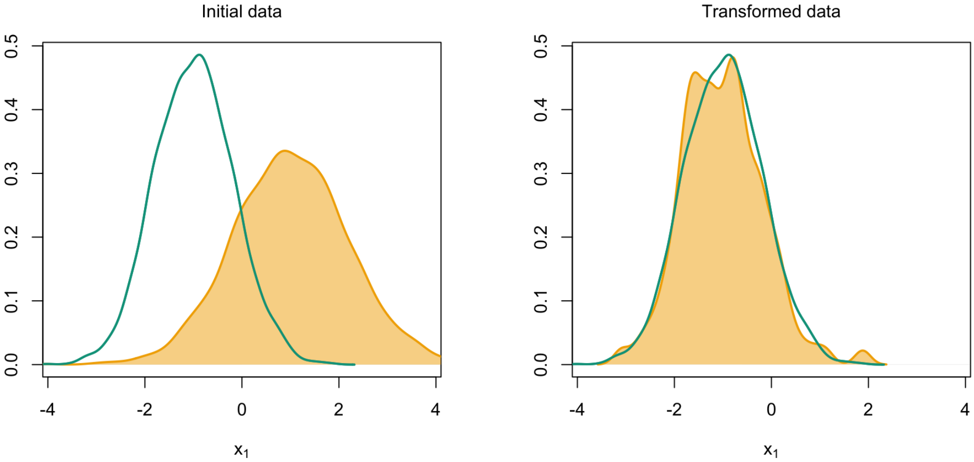
Application on a toy example, from Charpentier (2023a)





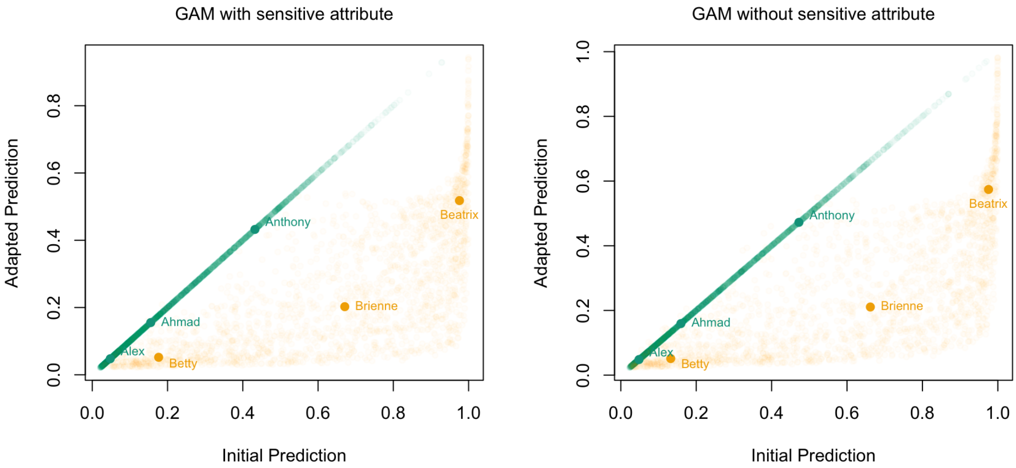
Optimal matching between individuals in group B and A.

Application on a toy example, from Charpentier (2023a)



Distribution of $x_1|s = A$, with the $x_1|s = B$ (left) and $\mathcal{T}(x_1)|s = B$ (right)

Application on a toy example, from Charpentier (2023a)



Scatterplot $\widehat{m}(\mathcal{T}(\boldsymbol{x}_i), s_i)$ against $\widehat{m}(\boldsymbol{x}_i, s_i)$ (left) $\widehat{m}(\mathcal{T}(\boldsymbol{x}_i))$ against $\widehat{m}(\boldsymbol{x}_i)$ (right)

Mitigating Discrimination

This transport can be used to quantify discrimination ("*what would have been the prediction $\widehat{y}$ for that person if that person had been a man, and not a woman?* ") but cannot be used to mitigate discrimination (Grari et al. (2022) or Hu et al. (2023a,b)).

We can therefore define some sort of average measure, solution of

$$\mathbb{P}^{\star} = \underset{\mathbb{Q}}{\text{argmin}} \left\{ \sum_{s\in\mathcal{S}} \omega_s d(\mathbb{Q}, \mathbb{P}_s)^2 \right\},$$

for some distance (or divergence) d, as in Nielsen and Boltz (2011).
Jeffreys (1946) consider the empirical case of "*averaging histograms*" (and not theoretical measures $\mathbb{P}_i$), extended in Nielsen and Nock (2009) as the Nielsen (2013) as "'*generalized Kullback–Leibler centroid*".
An alternative (see Agueh and Carlier (2011)), is to use the Wassertein distance, to define "*Wassertein barycenter*". As shown in Santambrogio (2015), if one of the measures $\mathbb{P}_i$ is absolutely continuous, the minimization problem has a unique solution.

Mitigating Discrimination

Given a reference measure, say $\mathbb{P}_1$, it is possible to write the barycenter as the "*average push-forward*" transformation of $\mathbb{P}_1$: if $\mathbb{P}_s = \mathcal{T}_\#^{1\to s}\mathbb{P}_1$ (with the convention that $\mathcal{T}_\#^{1\to 1}$ is the identity),

$$\mathbb{P}^\star = \left(\sum_{s\in\mathcal{S}} \omega_s \mathcal{T}^{1\to s}\right)_\# \mathbb{P}_1.$$

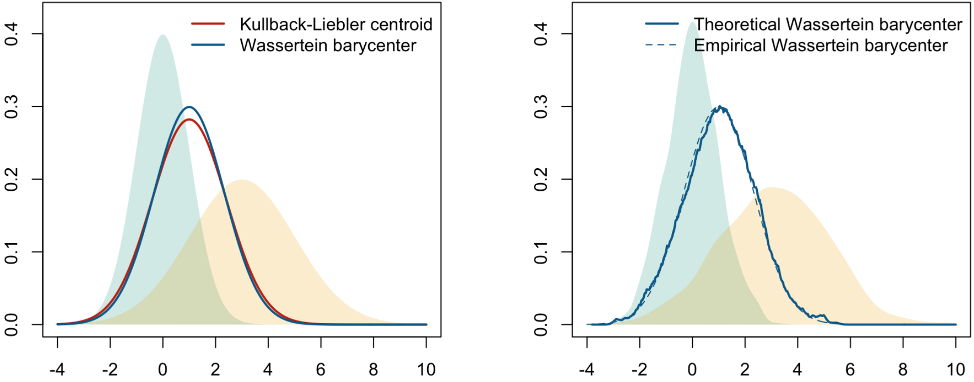
In the Gaussian case, Wassertein barycenter is

$$\mathcal{N}(\boldsymbol{\mu}^\star, \boldsymbol{\Sigma}^\star),\text{ where } \boldsymbol{\mu}^\star = \sum_{s\in\mathcal{S}} \omega_s \boldsymbol{\mu}_s, \boldsymbol{\Sigma}^\star = \sum_{s\in\mathcal{S}} \omega_s \big(\boldsymbol{\Sigma}^{\star 1/2}\boldsymbol{\Sigma}_s\boldsymbol{\Sigma}^{\star 1/2}\big)^{1/2}.$$

Jeffrey-Kullback–Leibler-centroid of those distribution would be

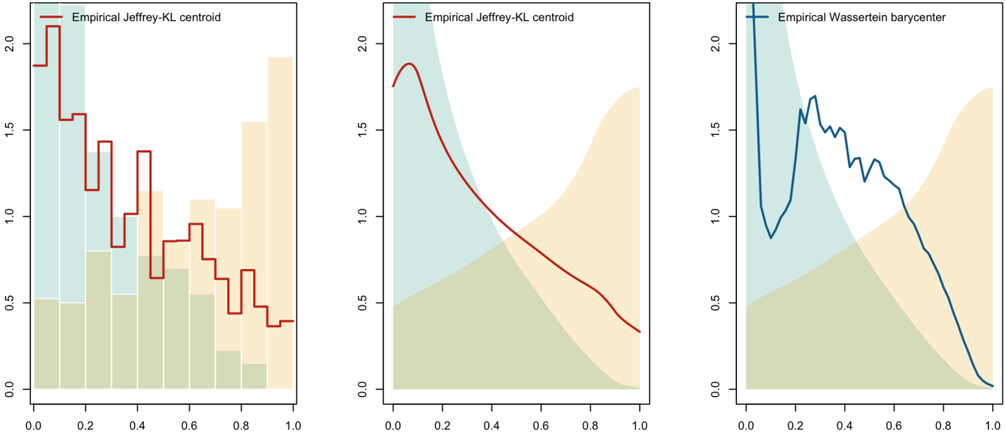
$$\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*),\text{ where } \boldsymbol{\mu}^* = \sum_{s\in\mathcal{S}} \omega_s \boldsymbol{\mu}_s \text{ and } \boldsymbol{\Sigma}^* = \sum_{s\in\mathcal{S}} \omega_s \boldsymbol{\Sigma}_s.$$

Mitigating Discrimination



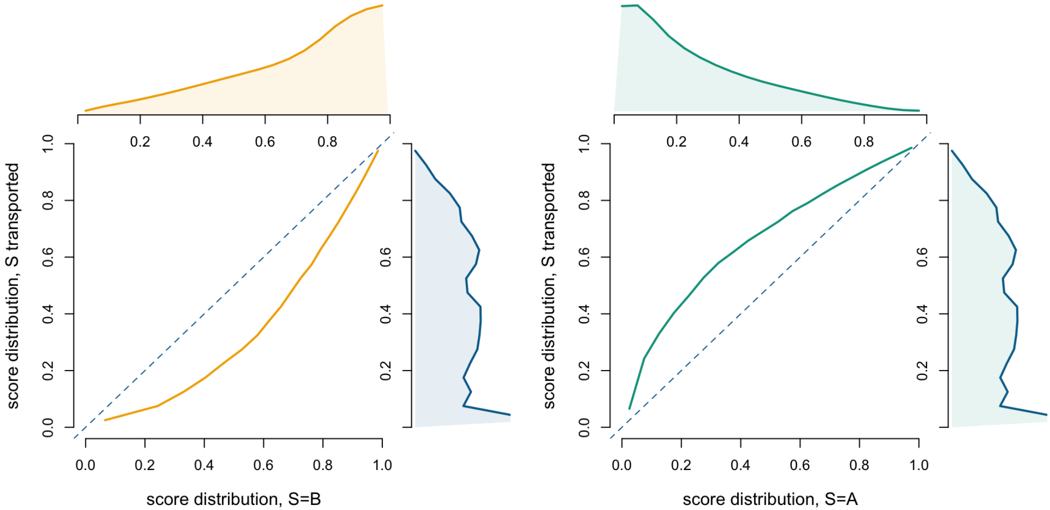
Centroid / barycenter of two Gaussian distributions

Mitigating Discrimination



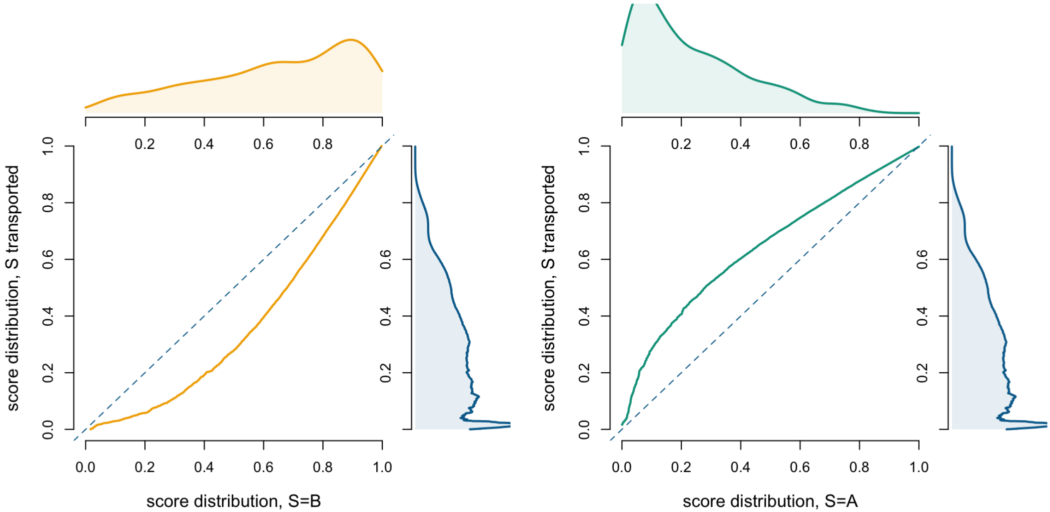
Centroid / barycenter of scoring functions $\widehat{m}(\boldsymbol{x}_i)$'s, $s_i \in \{A, B\}$

Mitigating Discrimination



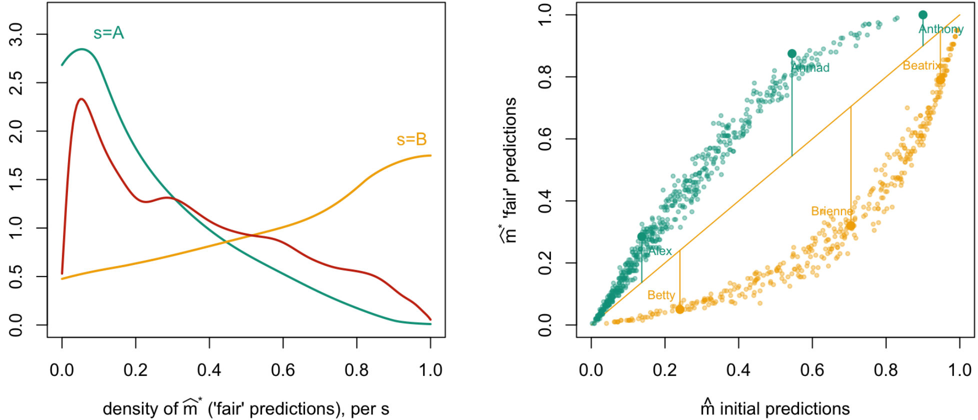
Centroid / barycenter of scoring functions $\widehat{m}(\boldsymbol{x}_i)$'s, $s_i \in \{A, B\}$ and barycenter

Mitigating Discrimination



Centroid / barycenter of scoring functions $\widehat{m}(\boldsymbol{x}_i)$'s, $s_i \in \{A, B\}$ and barycenter

Mitigating Discrimination

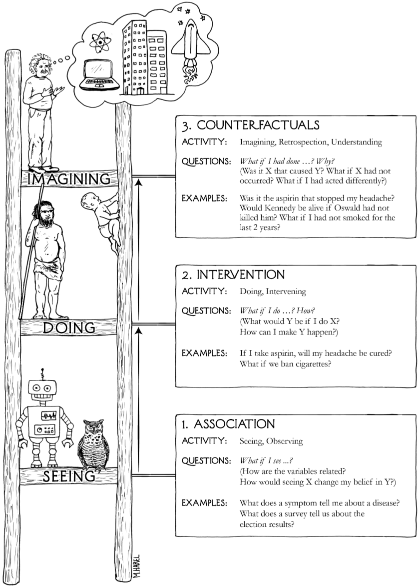


Distribution of transformed scores in the two groups $s_i \in \{A, B\}$ and barycenter

Take-away

- quantifying discrimination is hot topic
- connections with optimal transport are promising, including individual fairness
- more generally to get proper counterfactuals
  simple in dimension 1 (quantiles)
  less natural in higher dimension
  (see Hallin et al. (2021))
- see Hu et al. (2023a,b)
- see chapter 11 and 14, Charpentier (2023a)

Picture source: Pearl and Mackenzie (2018)



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.

Agueh, M. and Carlier, G. (2011). Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924.

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Angrist, J. D. and Pischke, J.-S. (2014). *Mastering'metrics: The path from cause to effect*. Princeton university press.

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

Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., Neel, S., and Roth, A. (2017). A convex framework for fair regression. *arXiv*, 1706.02409.

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.

Black, E., Yeom, S., and Fredrikson, M. (2020). Fliptest: fairness testing via optimal transport. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 111–121.

References

Brenier, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417.

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

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.

Charpentier, A. (2022). *Insurance: biases, discrimination and fairness*. Institut Louis Bachelier.

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

Charpentier, A. (2023b). Quantifying fairness and discrimination in predictive models. In Kreinovich, V., SriboonchiNa, S., and Yamaka, W., editors, *Machine Learning for Econometrics and Related Topics*. Springer Verlag.

Charpentier, A., Flachaire, E., and Gallic, E. (2023). 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.

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

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

References

de Lara, L., González-Sanz, A., Asher, N., Risser, L., and Loubes, J.-M. (2023). Transport-based counterfactual models. 2108.13025.

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.

Duivesteijn, W. and Feelders, A. (2008). Nearest neighbour classification with monotonicity constraints. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 301–316. Springer.

Galichon, A. (2016). *Optimal transport methods in economics*. Princeton University Press.

Gordaliza, P., Del Barrio, E., Fabrice, G., and Loubes, J.-M. (2019). Obtaining fairness using optimal transport theory. In *International Conference on Machine Learning*, pages 2357–2365. PMLR.

Grari, V., Charpentier, A., Lamprier, S., and Detyniecki, M. (2022). A fair pricing model via adversarial learning. *ArXiv*, 2202.12008.

Hallin, M., Del Barrio, E., Cuesta-Albertos, J., and Matrán, C. (2021). Distribution and quantile functions, ranks and signs in dimension d: A measure transportation approach. *The Annals of Statistics*, 49(2):1139–1165.

References

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323.

Hernández-Díaz, S., Schisterman, E. F., and Hernán, M. A. (2006). The birth weight "paradox" uncovered? *American journal of epidemiology*, 164(11):1115–1120.

Higham, N. J. (2008). *Functions of matrices: theory and computation*. Society for Industrial and Applied Mathematics.

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. *submitted*.

Hu, F., Ratz, P., and Charpentier, A. (2023b). Multivariate fair learning with statistical guarantees. *submitted*.

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

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461.

References

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

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.

Luong, B. T., Ruggieri, S., and Turini, F. (2011). *k*-nn as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 502–510.

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

Nielsen, F. (2013). Jeffreys centroids: A closed-form expression for positive histograms and a guaranteed tight approximation for frequency histograms. *IEEE Signal Processing Letters*, 20(7):657–660.

Nielsen, F. and Boltz, S. (2011). The burbea-rao and bhattacharyya centroids. *IEEE Transactions on Information Theory*, 57(8):5455–5466.

References

Nielsen, F. and Nock, R. (2009). Sided and symmetrized bregman centroids. *IEEE transactions on Information Theory*, 55(6):2882–2904.

Pearl, J. (2009). *Causality*. Cambridge university press.

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

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

Santambrogio, F. (2015). Optimal transport for applied mathematicians.

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

Villani, C. (2003). *Topics in optimal transportation*, volume 58. American Mathematical Society.

Villani, C. (2009). *Optimal transport: old and new*, volume 338. Springer Berlin, Heidelberg.

Wilcox, A. J. (1993). Birth weight and perinatal mortality: the effect of maternal smoking. *American journal of epidemiology*, 137(10):1098–1104.

Wilcox, A. J. (2001). On the importance—and the unimportance—of birthweight. *International journal of epidemiology*, 30(6):1233–1241.

Wilcox, A. J. (2006). Invited commentary: the perils of birth weight—a lesson from directed acyclic graphs. *American Journal of Epidemiology*, 164(11):1121–1123.