Fairness and discrimination in actuarial pricing

Arthur Charpentier\textsuperscript{1}, Laurence Barry\textsuperscript{2}, Vincent Grari\textsuperscript{3}, Lamprier Sylvain\textsuperscript{3}, Detyniecki Marcin\textsuperscript{4}

\textsuperscript{1} Univeristé du Québec à Montréal \textsuperscript{2} Chaire Pari \textsuperscript{3} Sorbonne Université \textsuperscript{4} AXA

Séminaire de la chaire DIALog (Digital Insurance And Long-term risks) 2022
Agenda

- Charpentier (2022) Insurance: Discrimination, Biases and Fairness, Institut Louis Bachelier
- Grari et al. (2022) A fair pricing model via adversarial learning, ArXiv:2202.12008

Ethics, fairness and discrimination
Protected Attributes?
Big Data and Proxies
Big/Small Data and Possible Pitfalls

Group fairness
From correlation to causality
Counterfactual and optimal transport
Adversarial Approach
Discriminating algorithms: 5 times AI showed prejudice

Artificial intelligence is supposed to make life easier for us all – but it is also prone to amplify sexist and racist biases from the real world.

By Daniel Cossins
Machine Bias
There’s software used across the country to predict future criminals. And it’s biased against blacks.
by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid’s blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.
Ethics, Fairness and Discrimination III

**Ethics Through Awareness**

Cynthia Dwork* Moritz Hardt† Toniann Pitassi‡ Omer Reingold§
Richard Zenev¶
November 30, 2011

**Algorithmic decision making and the cost of fairness**

Sam Corbett-Davies
Stanford University
sccorbett@stanford.edu

Emma Pierson
Stanford University
emmapi@stanford.edu

Avi Feller
Univ. of California, Berkeley
afeller@berkeley.edu

Sharad Goel
Stanford University
sgoel@stanford.edu

Aziz Huq
University of Chicago
huq@uchicago.edu

**Equality of Opportunity in Supervised Learning**

Moritz Hardt Eric Price Nathan Srebro
October 11, 2016

**Fairness in Criminal Justice Risk Assessments: The State of the Art**

Richard Berk,*, Hoda Heidari*, Shahin Jabbari*, Michael Kearns*, Aaron Roth*

**Fair prediction with disparate impact: A study of bias in recidivism prediction instruments**

Alexandra Chouldechova *

**The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning**

Sam Corbett-Davies
Stanford University
sccorbett@stanford.edu

Sharad Goel
Stanford University
sgoel@stanford.edu

August 14, 2018

**HUMAN DECISIONS AND MACHINE PREDICTIONS**

Jon Kleinberg
Harvard University

Himabindu Lakkaraju
University of Washington

Julie Leskiwic
University of California, Berkeley

Jens Ludwig
University of Chicago

Sendhil Mullainathan
University of Chicago

**The Frontiers of Fairness in Machine Learning**

Alexandra Chouldechova*
October 23, 2018

**A Survey on Bias and Fairness in Machine Learning**

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan, USC-ISI

**FAIRNESS IN MACHINE LEARNING: A SURVEY**

A Preprint

Simon Caton
University College Dublin
Dublin, Ireland
simon.caton@ucd.ie

Christian Haas
University of Nebraska at Omaha
Omaha, US
christianhaas@unomaha.edu
“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 explicitly ask for “, Kearns and Roth (2019)

It is a complex problem...

Accuracy : $\pi(x) = \mathbb{E}_P[Y|X = x]$ ($P$ historical probability) (is)

Fairness : $\pi^*(x) = \mathbb{E}_{P^*}[Y|X = x]$ ($P^*$ targeted probability) (ought, Hume (1739))
Word embedding exhibiting gender bias, Bolukbasi et al. (2016)
Spatial information and racial bias (redlining)
<table>
<thead>
<tr>
<th>Protected Attributes ?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>CA</td>
</tr>
<tr>
<td>-------------------------</td>
</tr>
<tr>
<td>Gender</td>
</tr>
<tr>
<td>Age</td>
</tr>
<tr>
<td>Driving experience</td>
</tr>
<tr>
<td>Credit history</td>
</tr>
<tr>
<td>Education</td>
</tr>
<tr>
<td>Occupation</td>
</tr>
<tr>
<td>Employment status</td>
</tr>
<tr>
<td>Marital status</td>
</tr>
<tr>
<td>Housing situation</td>
</tr>
<tr>
<td>Address/ZIP code</td>
</tr>
<tr>
<td>Insurance history</td>
</tr>
</tbody>
</table>

More and more features, possibly (strongly) correlated with protected variables

- **location** (policyholder home address)

Jean et al. (2016), Seresinhe et al. (2017), Gebru et al. (2017), Law et al. (2019), Ilic et al. (2019), Kita and Kidziński (2019), see also redlining
Bohren et al. (2019) on statistical discrimination, or efficient discrimination, as in Becker (1957) (inspired by Edgeworth (1922) up to Phelps (1972)).

Becker (2005) says “if young Moslem Middle Eastern males were in fact much more likely to commit terrorism against U.S. than were other groups, putting them through tighter security clearance would reduce current airport terrorism”.

“racial profiling” is “effective”, even though “such profiling is ‘unfair’ to the many young male Moslems who are not terrorists, and to the many minority shoppers who are honest ...”
“... That could be made up in part by compensating groups who are forced to go through more careful airport screening through putting them in shorter security lines, or in other ways. Similarly, innocent shoppers who are stopped and searched could be compensated for their embarrassment and time”

See also Boczar et al. (2021) (Insurance Data Science Conference in London) phrenology (Lombroso (1876) and Bertillon and Chervin (1909)) and ugly laws (TenBroek (1966) and Burgdorf and Burgdorf Jr (1974))

Missing an important covariate is an important issue

\[ y_i = \beta_0 + x_1^\top \beta_1 + x_2^\top \beta_2 + \varepsilon_i: \text{true model} \]

\[ y_i = b_0 + x_1^\top b_1 + \eta_i: \text{estimated model} \]

Maximum likelihood estimator of \( b_1 \) is

\[
\hat{b}_1 = (X_1^\top X_1)^{-1}X_1^\top y \\
= (X_1^\top X_1)^{-1}X_1^\top [X_1\beta_1 + X_2\beta_2 + \varepsilon] \\
= (X_1^\top X_1)^{-1}X_1^\top X_1\beta_1 + (X_1^\top X_1)^{-1}X_1^\top X_2\beta_2 + (X_1^\top X_1)^{-1}X_1^\top \varepsilon \\
= \beta_1 + \underbrace{(X_1^\top X_1)^{-1}X_1^\top X_2\beta_2}_{\beta_{12}} + \underbrace{(X_1^\top X_1)^{-1}X_1^\top \varepsilon}_{\nu_i}
\]

so that \( E[\hat{b}_1] = \beta_1 + \beta_{12} \neq \beta_1 \).
## Big/Small Data and Possible Pitfalls II

From Bickel et al. (1975) (see also Alipourfard et al. (2018))

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Men</th>
<th>Women</th>
<th>Proportions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>5233/12763    ~ 41%</td>
<td>3714/8442    ~ 44%</td>
<td>1512/4321   ~ 35%</td>
<td>66%-34%</td>
</tr>
<tr>
<td>Top 6</td>
<td>1745/4526    ~ 39%</td>
<td>1198/2691    ~ 45%</td>
<td>557/1835    ~ 30%</td>
<td>59%-41%</td>
</tr>
<tr>
<td>A</td>
<td>597/933     ~ 64%</td>
<td>512/825      ~ 62%</td>
<td>89/108      ~ 82%</td>
<td>88%-12%</td>
</tr>
<tr>
<td>B</td>
<td>369/585     ~ 63%</td>
<td>353/560      ~ 63%</td>
<td>17/25       ~ 68%</td>
<td>96%- 4%</td>
</tr>
<tr>
<td>C</td>
<td>321/918     ~ 35%</td>
<td>120/325      ~ 37%</td>
<td>202/593     ~ 34%</td>
<td>35%-65%</td>
</tr>
<tr>
<td>D</td>
<td>269/792     ~ 34%</td>
<td>138/417      ~ 33%</td>
<td>131/375     ~ 35%</td>
<td>53%-47%</td>
</tr>
<tr>
<td>E</td>
<td>146/584     ~ 25%</td>
<td>53/191       ~ 28%</td>
<td>94/393      ~ 24%</td>
<td>33%-67%</td>
</tr>
<tr>
<td>F</td>
<td>43/714      ~ 6%</td>
<td>22/373       ~ 6%</td>
<td>24/341      ~ 7%</td>
<td>52%-48%</td>
</tr>
</tbody>
</table>
Big/Small Data and Possible Pitfalls III

As mentioned in Cohen (1986),

\[ P[T \leq 1|X = \text{Costa Rica}] < P[T \leq 1|X = \text{Sweden}] \]

\[ P[T \leq 1|X = (\text{Costa Rica}, x)] > P[T \leq 1|X = (\text{Sweden}, x)], \ \forall x \]
Defining Group Fairness I

\[
\begin{align*}
\{ y \in \{0, 1\} \} & \quad \text{variable of interest (classically binary)} \\
\{ p \in \{0, 1\} \} & \quad \text{protected variable (sensitive)} \\
\{ x \in \mathbb{R}^d \} & \quad \text{‘explanatory’ variables} \\
\{ s \in [0, 1] \} & \quad \text{score, classically } s = s(x, p) \\
\{ \hat{y} \in \{0, 1\} \} & \quad \text{classifier, classically } \hat{y} = 1(s > t)
\end{align*}
\]

**Fairness Through Unawareness**, Kusner et al. (2017)
Protected attribute \( p \) is not explicitly used in decision function \( \hat{y} \).

▶ what is \( y \)?
▶ how to define a fair pricing?
Defining Group Fairness II

policyholder

$$(x_i, p_i)$$
Defining Group Fairness III

- Policyholder: \((x_i, p_i)\)
- Infered Data: \(m_d(x, p)\)
- Possible Scores
Defining Group Fairness IV

- commercial discount
- multiple decisions from policyholder (related to various personal features)
- multiple decisions from agent / broker

\[ m_u(x, p) \]

\[ m_d(x, p) \]
Defining Group Fairness V

- **policyholder** $(x_i, p_i)$
- **underwriting** $m_u(x, p)$
- **claim a loss** $\ell(x, p)$
- **inferred data** $m_d(x, p)$

(true value of the loss)
Defining Group Fairness VI

- policyholder \((x_i, p_i)\)
- underwriting \(m_u(x, p)\)
- claim a loss \(\ell(x, p)\)
- inferred data \(m_d(x, p)\)
- claims manager \(m_c(x, p)\)

- go to court?
- possible discrimination
- internal process
Defining Group Fairness VII

- policyholder \((x_i, p_i)\)
- underwriting \(m_u(x, p)\)
- claim a loss \(\ell(x, p)\)
- inferred data \(m_d(x, p)\)
- claims manager \(m_c(x, p)\)
- fraud detection \(m_f(x, p, x'(p))\)
- fraud scores
Defining Group Fairness VIII

- Policyholder $(x_i, p_i)$
- Underwriting $m_u(x, p)$
- Claim a loss $\ell(x, p)$
- Indemnity $y_i$
- Inferred data $m_d(x, p)$
- Claims manager $m_c(x, p)$
- Fraud detection $m_f(x, p, x'(p))$
Defining Group Fairness IX

- **Policyholder**: $(x_i, p_i)$
- **Underwriting**: $m_u(x, p)$
- **Claim a Loss**: $\ell(x, p)$
- **Indemnity**: $y_i$

**Inferred Data**
- $m_d(x, p)$

**Claims Manager**
- $m_c(x, p)$

**Fraud Detection**
- $m_f(x, p, x'(p))$

**Dataset**
- $(y_i, x_i, p_i)$ (training data)
Defining Group Fairness X

Not to mention multiple biases, from feedback bias (telematics and gamification) to selection bias

(see Hand (2020) on dark data)
Defining Group Fairness XI

\[
\begin{align*}
    \begin{cases}
    y \in \{0, 1\} & \text{variable of interest (classically binary)} \\
    p \in \{0, 1\} & \text{protected variable (sensitive)} \\
    x \in \mathbb{R}^d & \text{‘explanatory’ variables} \\
    s \in [0, 1] & \text{score, classically } s = s(x, p) \\
    \hat{y} \in \{0, 1\} & \text{classifier, classically } \hat{y} = 1(s > t)
    \end{cases}
\end{align*}
\]

**Demographic Parity**, (Corbett-Davies et al. (2017), Agarwal (2021))

Decision function \(\hat{y}\) satisfies demographic parity if \(\hat{Y} \perp P\), i.e.

\[
\mathbb{P}[\hat{Y} = y|P = 0] = \mathbb{P}[\hat{Y} = y|P = 1], \quad \forall y \text{ or } \mathbb{E}[\hat{Y}|P = 0] = \mathbb{E}[\hat{Y}|P = 1]
\]

Since \(y\) is binary, a classical intermediary component is the score \(s(x)\) and a classical tool is the ROC curve (obtained by changing threshold \(t\))
Defining Group Fairness XII

\[ \hat{y} = 0 \]
\[ \hat{y} = 1 \]

\[ \hat{y} = 0 \]
\[ \hat{y} = 1 \]
\[ y = 0 \]
\[ y = 1 \]

negatives
positives
false negative
false positives
TPR
FPR
Defining Group Fairness XIII

\[
\hat{y} = 1 \\
\hat{y} = 0 \\
\hat{y} = 1 \\
\hat{y} = 0 \\
y = 0 \\
y = 1
\]

false positives

positives

false negative

negatives

positives

FPR

TPR
Defining Group Fairness XIV

\[ \hat{y} = 1 \]
\[ \hat{y} = 0 \]

\[ x_1 \]
\[ x_2 \]

positive
negatives

score \[ s(x) \]
Defining Group Fairness XV

\[ \hat{y} = 1 \]
\[ \hat{y} = 0 \]

positives

negatives

\[ s(x)|y = 1 \]

\[ s(x)|y = 0 \]
Defining Group Fairness XVI

\[ \hat{y} = 1 \]
\[ \hat{y} = 0 \]

\[ s(x) \mid y = 1 \] true positives
\[ s(x) \mid y = 0 \] false positives

negatives
positives
We can compare $s(\mathbf{X})$ conditional on $Y$, but also on $P$.
Defining Group Fairness XVIII

**Equal Opportunity**, Hardt et al. (2016)

True positive parity

\[
P[\hat{Y} = 1 | P = 0, Y = 1] = P[\hat{Y} = 1 | P = 1, Y = 1]
\]

or false positive parity

\[
P[\hat{Y} = 1 | P = 0, Y = 0] = P[\hat{Y} = 1 | P = 1, Y = 0]
\]
<table>
<thead>
<tr>
<th>Type of Fairness</th>
<th>Definition</th>
<th>Authors (Year)</th>
<th>Independence/Certainty Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical Parity</td>
<td>$P[\hat{Y} = 1</td>
<td>P = p] = \text{cst, } \forall p$</td>
<td>Dwork et al. (2012)</td>
</tr>
<tr>
<td>Conditional Statistical Parity</td>
<td>$P[\hat{Y} = 1</td>
<td>P = p, X = x] = \text{cst}_x, \forall p, y$</td>
<td>Corbett-Davies et al. (2017)</td>
</tr>
<tr>
<td>Equalized Odds</td>
<td>$P[\hat{Y} = 1</td>
<td>P = p, Y = y] = \text{cst}_y, \forall p, y$</td>
<td>Hardt et al. (2016)</td>
</tr>
<tr>
<td>Equalized Opportunity</td>
<td>$P[\hat{Y} = 1</td>
<td>P = p, Y = 1] = \text{cst, } \forall p$</td>
<td>Hardt et al. (2016)</td>
</tr>
<tr>
<td>Predictive Equality</td>
<td>$P[\hat{Y} = 1</td>
<td>P = p, Y = 0] = \text{cst, } \forall p$</td>
<td>Corbett-Davies et al. (2017)</td>
</tr>
<tr>
<td>Balance (Positive)</td>
<td>$E[S</td>
<td>P = p, Y = 1] = \text{cst, } \forall p$</td>
<td>Kleinberg et al. (2017)</td>
</tr>
<tr>
<td>Balance (Negative)</td>
<td>$E[S</td>
<td>P = p, Y = 0] = \text{cst, } \forall p$</td>
<td>Kleinberg et al. (2017)</td>
</tr>
<tr>
<td>Conditional Accuracy Equality</td>
<td>$P[Y = y</td>
<td>P = p, \hat{Y} = y] = \text{cst}_y, \forall p, y$</td>
<td>Berk et al. (2017)</td>
</tr>
<tr>
<td>Predictive Parity</td>
<td>$P[Y = 1</td>
<td>P = p, \hat{Y} = 1] = \text{cst, } \forall p$</td>
<td>Chouldechova (2017)</td>
</tr>
<tr>
<td>Calibration</td>
<td>$P[Y = 1</td>
<td>P = p, S = s] = \text{cst}_s, \forall p, s$</td>
<td>Chouldechova (2017)</td>
</tr>
<tr>
<td>Well-Calibration</td>
<td>$P[Y = 1</td>
<td>P = p, S = s] = s, \forall p, s$</td>
<td>Chouldechova (2017)</td>
</tr>
<tr>
<td>Accuracy Equality</td>
<td>$P[\hat{Y} = Y</td>
<td>P = p] = \text{cst, } \forall p$</td>
<td>Berk et al. (2017)</td>
</tr>
<tr>
<td>Treatment Equality</td>
<td>$\begin{align*} \text{FN}_p &amp;= \text{cst}_p, \forall p \ \text{FP}_p &amp;= \text{cst}_p, \forall p \end{align*}$</td>
<td>Berk et al. (2017)</td>
<td></td>
</tr>
</tbody>
</table>

@freakonometrics  •  freakonometrics  •  freakonometrics.hypotheses.org
From correlation to causality I

▶ “classifying projection methods as using demographic/actuarial models or non-demographic/causal models”
Keilman (2003) and Hudson (2007)

▶ “Article 5(2) allowed Member States to Permit proportionate differences in individuals premiums and benefits where the use of sex is a determining factor in the assessment of risk based on relevant and accurate actuarial and statistical data.”
Thiery and Van Schoubroeck (2006) and Schmeiser et al. (2014)

▶ “Two judges on the Supreme Court dissented in the Zurich case. In their view, an insurer must not only prove a statistical correlation between a particular group and higher risk, but a causal connection”
Gomery et al. (2011)
From correlation to causality II

- DAGs are important
- Looking for a counterfactual
Consider some distances $D$ on $\{0, 1\} \times \{0, 1\}$ or $[0, 1] \times [0, 1]$, and $d$ on $\mathbb{R}^p \times \mathbb{R}^p$.

**Lipschitz property**, Duivesteijn and Feelders (2008)

$$D(\hat{y}_i, \hat{y}_j) \text{ or } D(s_i, s_j) \leq d(x_i, x_j), \ \forall i, j = 1, \ldots, n.$$  

**Conterfactual fairness**, Kusner et al. (2017) If the prediction in the real world is the same as the prediction in the counterfactual world where the individual would have belonged to a different demographic group, we have counterfactual equity, i.e.

$$\mathbb{P}[Y^*_p = y|X = x] = \mathbb{P}[Y^*_p' = y|X = x], \ \forall p', x, y.$$
From correlation to causality IV

▸ counterfactuals
  (what if I had done...?)

▸ intervention

▸ association
  (what if I see...?)

what would have happened if this person had had treatment 1 instead of treatment 0?

(picture Pearl & Mackenzie (2018))
From correlation to causality

Causal inference literature,

- $t$ some binary treatment ($t \in \{0, 1\}$)
- $x$ some covariates
- $y$ denote the observed outcome, $y_{i,T\leftarrow 1}^{\ast}$ and $y_{i,T\leftarrow 0}^{\ast}$ the potential outcomes

<table>
<thead>
<tr>
<th>treatment</th>
<th>outcome</th>
<th>age</th>
<th>gender</th>
<th>height</th>
<th>weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_i$</td>
<td>$y_i$</td>
<td>$y_{i,T\leftarrow 1}^{\ast}$</td>
<td>$y_{i,T\leftarrow 0}^{\ast}$</td>
<td>$x_{1,i}$</td>
<td>$x_{2,i}$</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>121</td>
<td>121</td>
<td>?</td>
<td>37</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>109</td>
<td>?</td>
<td>109</td>
<td>28</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>162</td>
<td>162</td>
<td>?</td>
<td>53</td>
</tr>
</tbody>
</table>

There will be a significant impact of treatment $t$ on $y$ if $y_{T\leftarrow 0}^{\ast} \neq y_{T\leftarrow 1}^{\ast}$ (see Rubin (1974), Hernán and Robins (2010) or Imai (2018)).

The causal effect for individual $i$ is $\tau_i = y_{i,T\leftarrow 1}^{\ast} - y_{i,T\leftarrow 0}^{\ast}$
From correlation to causality VI

One can define the sample average treatment effect (SATE)

$$\text{SATE} = \frac{1}{n} \sum_{i=1}^{n} y^*_{i,T \leftarrow 1} - y^*_{i,T \leftarrow 0}$$

the average treatment effect (ATE)

$$\tau = \text{ATE} = \mathbb{E}[Y^*_{i,T \leftarrow 1} - Y^*_{i,T \leftarrow 0}]$$

and, for possibly heterogeneous effects, sample average treatment effect (CATE)

$$\tau(x) = \text{CATE}(x) = \mathbb{E}[Y^*_{i,T \leftarrow 1} - Y^*_{i,T \leftarrow 0} | X = x]$$
From correlation to causality VII

Naive difference-in-means estimators, but Splawa-Neyman et al. (1990)

\[ \hat{\tau} = \frac{1}{n_1} \sum_{i=1}^{n} t_i y_i - \frac{1}{n_0} \sum_{i=1}^{n} (1 - t_i) y_i, \quad \text{where } n_j = \sum_{i=1}^{n} t_i \]

or the regression-based estimators (also called single learner)

\[ \hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x) \]

where \( \hat{\mu}_t(x) = \mathbb{E}[Y|T = t, X = x] \)

One can also use the propensity score \( \pi_i = \mathbb{P}[T = 1|X = x_i] \)

\[ \hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \frac{t_i y_i}{\pi_i} - \frac{(1 - t_i) y_i}{1 - \pi_i} \]

to derive a inverse-propensity score weighted estimator.

This can be use to derive a doubly-robust estimator
One can also use matching technique,

- for treated individual $i$ ($t_i = 1$), match that individual to someone in the non-treated group ($t_j = 0$), $j^* = \arg\min_{j: t_j = 0} \{d(x_i, x_j)\}$,

E.g. $x_i$ is the height of a person, $t_i$ is the gender of that person, then

$$\hat{\tau} = \frac{1}{n_1} \sum_{i=1}^{n_1} t_i \cdot (y_i - y_{j^*_i})$$
Counterfactual and optimal transport II

- **Graphs**
  - Left: Illustration of 'MEN' and 'WOMEN' with height distribution from 150 to 190 cm.
  - Right: Frequency distribution of 'Height (cm) of men' from 160 to 200 cm, showing 'MATCHED SAMPLE' and 'ORIGINAL SAMPLE'.

- **Axes**
  - X-axis: Height (cm)
  - Y-axis: Frequency

---

@freakonometrics  🐦 freakonometrics  🤝 freakonometrics.hypotheses.org
Consider a permutation of all treated individuals \((t_i = 1)\),

for treated individual \(i \ (t_i = 1)\), match that individual to someone in the non-treated group \((t_j = 0)\), \(j^* = \arg\min_{j: t_j = 0} \{d(x_i, x_j)\}\),

then we remove the untreated observation from the database, and iterate (so as to match all treated individuals with an untreated person
Counterfactual and optimal transport IV
Counterfactual and optimal transport V

![Diagram](image)
Counterfactual and optimal transport VI
Counterfactual and optimal transport VII
Counterfactual and optimal transport VIII

![Diagram showing height distribution for men and women](image)

**Height (cm)**

150 160 170 180 190

@freakonometrics  🐒  freakonometrics  📍  freakonometrics.hypotheses.org
Counterfactual and optimal transport IX
Counterfactual and optimal transport $X$

See Monge (1781)’s problem (déblais/remblais - excavation/infill)

We want to go from $\mathbb{P}_0$ to $\mathbb{P}_1$ (distributions on $\mathbb{R}^k$). Given $T : \mathbb{R}^k \to \mathbb{R}^k$, define the “push-forward” measure

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

An optimal transport $T^*$ (in Brenier’s sense) from $\mathbb{P}_0$ towards $\mathbb{P}_1$ will be solution of

$$T^* \in \arg \inf_{T : T_\#\mathbb{P} = \mathbb{Q}} \int_{\mathbb{R}^k} ||x - T(x)||^2 d\mathbb{P}(x).$$
Counterfactual and optimal transport XI
Counterfactual and optimal transport XII

One can prove (see) that $T^* = \nabla \psi$ where $\psi$ is a convex function. If $k = 1$, $T$ is an increasing function, i.e. if $F_0(x) = \mathbb{P}_0[X \leq x]$ and $F_1(x) = \mathbb{P}_1[X \leq x]$, then $T^*(x) = F_1^{-1} \circ F_0(x)$ satisfies $T^* \# \mathbb{P}_0 = \mathbb{P}_1$ (since $F_1(x) = F_0(T^*^{-1}(x))$) and $T^*$ is optimal.
Counterfactual and optimal transport XIII

That is an optimal coupling problem, that we can write

$$\min_{\text{copula } C} \int_{\mathbb{R}^2} c(x, y) dF(x, y), \text{ where } F(x, y) = C(F_0(x), F_1(y))$$

As Gordaliza et al. (2019) and Black et al. (2020), a classifier $m$ obtained from score $s$ ($m(x, p) = 1(s(x, p) > \text{seuil})$) is unfair if $m(x, 0) \neq m(T^*(x), 1)$. Black et al. (2020) define the FlipSet as

$$\mathcal{X}_F(m, T^*) = \{x \in \mathcal{X} : m(x, 0) \neq m(T^*(x), 1)\}.$$  

$$\begin{cases} 
\mathcal{X}_F^+(m, T^*) = \{x \in \mathcal{X} : m(x, 0) > m(T^*(x), 1)\}. \\
\mathcal{X}_F^-(m, T^*) = \{x \in \mathcal{X} : m(x, 0) < m(T^*(x), 1)\}. 
\end{cases}$$

and if $T^*\mathbb{P}_0 = \mathbb{P}_1$, the average causal effect is

$$ACE = \mathbb{E}[Y^*_{T \leftarrow 1}] - \mathbb{E}[Y^*_{T \leftarrow 0}] = \mathbb{P}_0[\mathcal{X}_F^-(m, T^*)] - \mathbb{P}_0[\mathcal{X}_F^+(m, T^*)]$$
Counterfactual and optimal transport XIV
Adversarial Approach I

Hirschfeld (1935), Gebelein (1941) and Rényi (1959)

\[ HGR(U, V) = \max \{ \text{corr}[f(U), g(V)] \} = \max_{f \in S_U, g \in S_V} \left\{ \mathbb{E}[f(U)g(V)] \right\} \]

where \( S_U = \{ f : U \to \mathbb{R} : \mathbb{E}[f(U)] = 0 \text{ and } \mathbb{E}[f(U)^2] = 1 \} \) and similarly \( S_V \).

One can also consider a conditional version,

\[ HGR(U, V \mid Z) = \max_{f \in S_{U|Z}, g \in S_{V|Z}} \left\{ \mathbb{E}[f(U)g(V) \mid Z] \right\} \]

where \( S_{U|Z} = \{ f : U \to \mathbb{R} : \mathbb{E}[f(U) \mid Z] = 0 \text{ and } \mathbb{E}[f(U)^2 \mid Z] = 1 \} \).

\[
\begin{cases}
\text{Demographic Parity} : \quad \hat{Y} \perp \perp P \quad \text{i.e. } HGR(\hat{Y}, P) = 0 \\
\text{Equalized Odds} : \quad \hat{Y} \perp \perp P \mid Y \quad \text{i.e. } HGR(\hat{Y}, P \mid Y) = 0
\end{cases}
\]
Adversarial Approach II

\( HGR \) can be difficult to estimate, but one can use some Neural Net,

\[
HGR_{NN}(U, V) = \max_{\omega_f, \omega_g} \{ \mathbb{E}[f_{\omega_f}(U)g_{\omega_g}(V)] \}
\]

In a classical ML or econometric pricing model, solve

\[
\arg\min_{\theta} \{ L(\hat{y}, y) \}, \text{ where } L(\hat{y}, y) = \sum_{i=1}^{n} \ell(\hat{y}_i, y_i) \text{ and } \hat{y} = h_\theta(x)
\]

To avoid over-fit: penalize complexity (penalty \( P \))

\[
\arg\min_{\theta} \{ L(h_\theta(x), y) + \lambda P(h_\theta) \}
\]
Adversarial Approach III

Inspired by Goodfellow et al. (2018), to avoid un-fairness: penalize according to $HGR(\hat{y}, p)$ (for demographic parity)

$$\argmin_{\theta, \omega_f, \omega_g} \left\{ \mathcal{L}(h_\theta(x), y) + \lambda HGR_{\omega_f, \omega_g}(\hat{y}, p) \right\}$$

i.e.

$$\argmin_{\theta} \left\{ \max_{\omega_f, \omega_g} \left\{ \mathcal{L}(h_\theta(X), Y) + \lambda \mathbb{E}_{(X,S) \sim D}(f_{\omega_f}(h_\theta(X)) \hat{g}_{\omega_g}(P)) \right\} \right\}$$

or $HGR(\hat{y}, p|y)$ (for equalized odds), i.e. when $y \in \{0, 1\}$

$$\argmin_{\theta} \left\{ \max_{\omega_{f0}, \omega_{g0}, \omega_{f1}, \omega_{g1}} \left\{ \mathcal{L}(h_\theta(X), Y) + \lambda_0 \mathbb{E}_{(X,P) \sim D_0}(f_{\omega_{f0}}(h_\theta(X)) \hat{g}_{\omega_{g0}}(P)) + \lambda_1 \mathbb{E}_{(X,P) \sim D_1}(f_{\omega_{f1}}(h_\theta(X)) \hat{g}_{\omega_{g1}}(P)) \right\} \right\}$$
Adversarial Approach IV

or, more generally when \( y \in \Omega_Y \) (e.g. \( \{0, 1, 2, 3+\} \)), if \( k = \#\Omega_y \)

\[
\arg\min_{\theta} \left\{ \max_{\omega_f, \omega_g, \omega_k} \left\{ \mathcal{L}(h_\theta(X), Y) + \sum_{y \in \Omega_y} \lambda_y \mathbb{E}(X,P) \sim D_y \left( \hat{f}_{\omega_f} (h_\theta(X)) \hat{g}_{\omega_g} (P) \right) \right\} \right\}
\]

\[
\frac{\partial \mathcal{L}(h_w(X),Y)}{\partial w_h} + \lambda \frac{\partial E(\hat{f}_{w_f}(h_w(X))\hat{g}_{w_g}(S))}{\partial w_h} \quad \frac{\partial E(\hat{f}_{w_f}(h_w(X))\hat{g}_{w_g}(S))}{\partial w_f}
\]

\[
\frac{\partial E(\hat{f}_{w_f}(h_w(X))\hat{g}_{w_g}(S))}{\partial w_f} \quad \frac{\partial E(\hat{f}_{w_f}(h_w(X))\hat{g}_{w_g}(S))}{\partial w_g}
\]
Dealing with high dimension I

- geographic / spatial information, $X_g$
- car type / make / model, $X_g$
- other classical ratemaking variables, $X_p$ (non protected)

Some features can be in high dimension, natural solution would be PCA or autoencoders (see Shi and Shi (2021) about feature embedding in high dimension)
Dealing with high dimension II

Principal Component Analysis (PCA)

$$\min_{P \in \Pi} \{ \| X - P^T P X \|_F^2 \} \text{ s.t. } \text{rank}(P) = k$$

where $\Pi$ is the set of projection matrices.

$$\min_{P \in \Pi} \{ \| X - P^T P X \|_F^2 \} = \min \sum_{i=1}^{n} (P^T P x_i - x_i)^T (P^T P x_i - x_i)$$
Dealing with high dimension III

Autoencoder

\[
\min_\psi \left\{ \| X - \psi \circ \varphi X \|_F^2 \right\}
\]

\[
X \in \mathcal{X} \quad \quad Z \in \mathcal{Z} \quad \quad X' \in \mathcal{X}
\]

\[
\min \| X - X' \|^2 = \min \| X - \psi \circ \varphi (X) \|^2
\]

\[
\min \sum_{i=1}^{n} (\psi \circ \varphi (x_i) - x_i)^\top (\psi \circ \varphi (x_i) - x_i)
\]

encoder

"bottleneck" hidden layer

decoder
Dealing with high dimension IV
Application on synthetic data I

- S: sensitive / protected (gender)
- Col: color of the car
- Sp: maximum speed of the car
- Sal: average salary of the policyholders area
- Age: age of the driver
- Ina: inattention
- Agg: aggressivity
- Y: total cost
Application on synthetic data II

- $\lambda$: fairness penalty
- $p$-rule: \[
\min \left\{ \frac{\mathbb{P}(\hat{Y} = 1|P = 1)}{\mathbb{P}(\hat{Y} = 1|P = 0)}, \frac{\mathbb{P}(\hat{Y} = 1|P = 0)}{\mathbb{P}(\hat{Y} = 1|P = 1)} \right\}
\]
Application on real data (pricing game 2015) I

\[ y \in \{0, 1\} \text{ (claim occurrence)} \]
Application on real data (pricing game 2015) II

\( y \in \{0, 1, 2+\} \) (claim frequency)

see Grari et al. (2022) for more examples (including the case where \( y \in \mathbb{R}^+ \))


References II


References IV


References VII


