Fairness and discrimination in actuarial pricing

Arthur Charpentier[1], Laurence Barry[2], Vincent Grari[3],
Lamprier Sylvain[3], Detyniecki Marcin[4]

[1] Univeristé du Québec à Montréal [2] Chaire Pari [3] Sorbonne Université [4] AXA

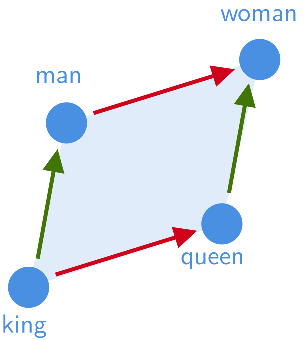
CIRM - MLISTRAL - September 2022

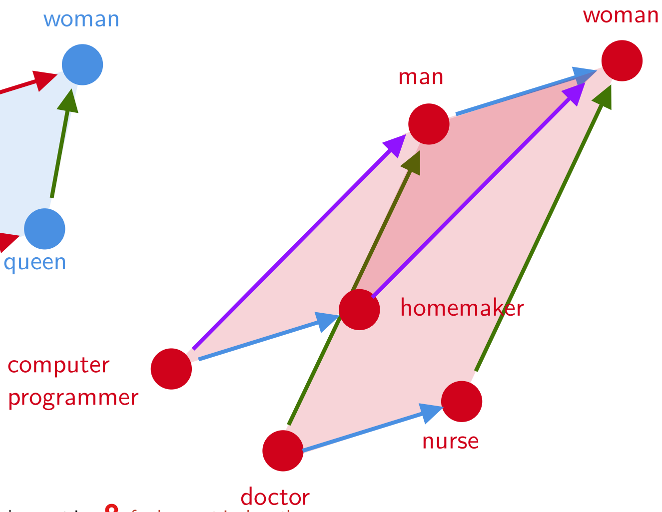
Agenda

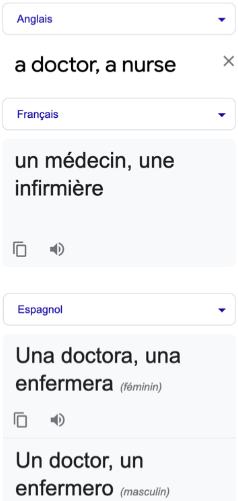
- Charpentier (2022) Insurance: Discrimination, Biases and Fairness, *Institut Louis Bachelier*
- Grari et al. (2022) A fair pricing model via adversarial learning, *ArXiv:2202.12008*
- "*Technology is neither good nor bad; nor is it neutral* " , Kranzberg (1986)
- "*Machine learning won't give you anything like gender neutrality 'for free' that you didn't explicitely ask for* ", Kearns and Roth (2019)

Motivation

- Accuracy : $\pi(\boldsymbol{x}) = \mathbb{E}_{\mathbb{P}}[Y|\boldsymbol{X} = \boldsymbol{x}]$ ($\mathbb{P}$ historical probability) (is)
- Fairness : $\pi^{\star}(\boldsymbol{x}) = \mathbb{E}_{\mathbb{P}^{\star}}[Y|\boldsymbol{X} = \boldsymbol{x}]$ ($\mathbb{P}^{\star}$ targeted probability) (ought, Hume (1739))







Actual Pricing I

- "*at the core of insurance business lies discrimination between risky and non-risky insureds*", Avraham (2017)

$$
\begin{cases}
\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^d : \text{'explanatory' variables} \\
\begin{cases}
\boldsymbol{x}_c \in \mathcal{X}_c \subset \mathbb{R}^{d_c} : \text{car-vehicule 'explanatory' variables} \\
x'_c = m_c(\boldsymbol{x}_c; \boldsymbol{x}) \in \mathbb{R} : \text{car-vehicle scoring}
\end{cases} \\
\begin{cases}
\boldsymbol{x}_g \in \mathcal{X}_g \subset \mathbb{R}^{d_g} : \text{geographic 'explanatory' variables} \\
x'_g = m_g(\boldsymbol{x}_g; \boldsymbol{x}) \in \mathbb{R} : \text{geographic scoring}
\end{cases} \\
y \in \{0,1\}, \mathbb{N} \text{ or } \mathbb{R}_+ : \text{variable of interest} \\
\widehat{y} = m(\boldsymbol{x}, x'_c, x'_g) : \text{prediction (or score)}
\end{cases}
$$

Classically, a two-stage approach is considered to create a geographic score

- Fit a model $\widehat{m}$ to predict y based on $\boldsymbol{x}$, compute residuals $\widehat{\varepsilon}$
- Fit a model $\widehat{m}_g(\cdot; \boldsymbol{x})$ to predict $\widehat{\varepsilon}$ based on $\boldsymbol{x}_g$
- define a score $x'_g = \widehat{m}_g(\boldsymbol{x}_g; \boldsymbol{x})$ (that is function of $\boldsymbol{x}$, too)

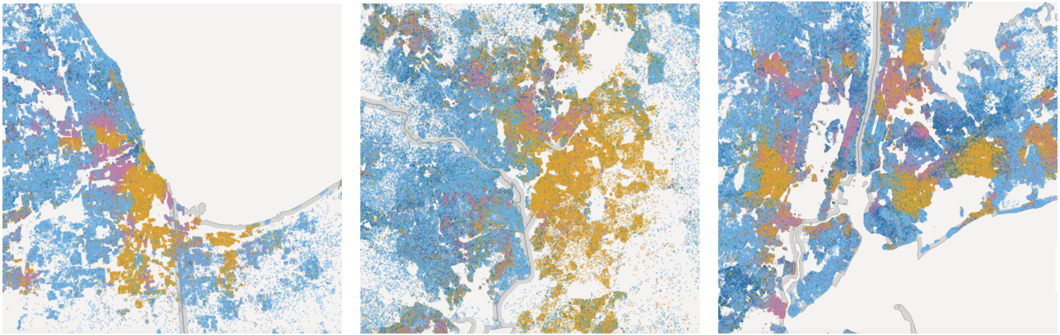
Protected Attributes ? Motor insurance in North America

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

CA: Californie, HI: Hawaii, GA: Georgia, NC: Caroline du nord, NY: New York, MA: Massachusetts, PA: Pennsylvanie, FL: Floride, TX: Texas, AL: Alberta, ON: Ontario, NB: Nouveau-Brunswick, NL: Terre-Neuve-et-Labrador, QC: Québec

Protected Attributes ? Motor insurance in North America

Spatial information and racial bias (redlining)



Defining Group Fairness when y is binary I

$$
\begin{cases}
\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^d : \text{'explanatory' variables} \\
p \in \{0,1\} : \text{protected variable or sensitive attribute} \\
\begin{cases}
\boldsymbol{x}_c \in \mathcal{X}_c \subset \mathbb{R}^{d_c} : \text{car-vehicule 'explanatory' variables} \\
x'_c = m_c(\boldsymbol{x}_c; \boldsymbol{x}, p) \in \mathbb{R} : \text{car-vehicle scoring}
\end{cases} \\
\begin{cases}
\boldsymbol{x}_g \in \mathcal{X}_g \subset \mathbb{R}^{d_g} : \text{geographic 'explanatory' variables} \\
x'_g = m_g(\boldsymbol{x}_g; \boldsymbol{x}, p) \in \mathbb{R} : \text{geographic scoring}
\end{cases} \\
y \in \{0,1\} : \text{variable of interest} \quad \text{(binary)} \\
s = m(\boldsymbol{x}, p, x'_c, x'_g) : \text{score} \\
\widehat{y} = \mathbf{1}(s > \text{threshold}) : \text{prediction}
\end{cases}
$$

Fairness Through Unawareness, Kusner et al. (2017)

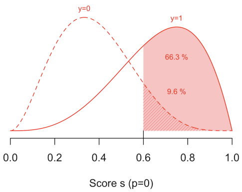
Protected attribute p is not explicitly used in decision function $\widehat{y}$.

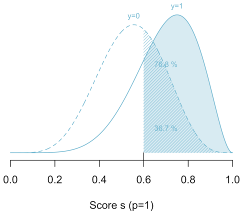
Defining Group Fairness when y is binary II

Demographic Parity, (Corbett-Davies et al. (2017), Agarwal (2021))
Decision function $\widehat{y}$ satisfies demographic parity if $\widehat{Y} \perp\!\!\!\perp P$, i.e.

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

We can compare $s(\boldsymbol{X})$ conditional on Y, but also on P





Defining Group Fairness when y is binary III

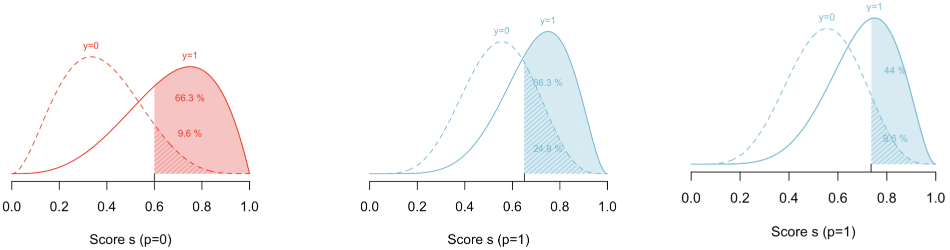
Equal Opportunity, Hardt et al. (2016)

True positive parity

$$\mathbb{P}[\widehat{Y}=1|P=0,Y=1]=\mathbb{P}[\widehat{Y}=1|P=1,Y=1]$$

or false positive parity

$$\mathbb{P}[\widehat{Y}=1|P=0,Y=0]=\mathbb{P}[\widehat{Y}=1|P=1,Y=0]$$



Defining Group Fairness when y is binary IV

statistical parity	Dwork et al. (2012)	$\mathbb{P}[\widehat{Y}=1\mid P=p]=\text{cst},\ \forall p$	independence
conditional statistical parity	Corbett-Davies et al. (2017)	$\mathbb{P}[\widehat{Y}=1\mid P=p,X=x]=\text{cst}_x,\ \forall p,y$	$\widehat{Y}\perp\!\!\!\perp P$
equalized odds	Hardt et al. (2016)	$\mathbb{P}[\widehat{Y}=1\mid P=p,Y=y]=\text{cst}_y,\ \forall p,y$	separation
equalized opportunity	Hardt et al. (2016)	$\mathbb{P}[\widehat{Y}=1\mid P=p,Y=1]=\text{cst},\ \forall p$	
predictive equality	Corbett-Davies et al. (2017)	$\mathbb{P}[\widehat{Y}=1\mid P=p,Y=0]=\text{cst},\ \forall p$	$\widehat{Y}\perp\!\!\!\perp P\mid Y$
balance (positive)	Kleinberg et al. (2017)	$\mathbb{E}[S\mid P=p,Y=1]=\text{cst},\ \forall p$	$S\perp\!\!\!\perp P\mid Y$
balance (negative)	Kleinberg et al. (2017)	$\mathbb{E}[S\mid P=p,Y=0]=\text{cst},\ \forall p$	
conditional accuracy equality	Berk et al. (2017)	$\mathbb{P}[Y=y\mid P=p,\widehat{Y}=y]=\text{cst}_y,\ \forall p,y$	sufficiency
predictive parity	Chouldechova (2017)	$\mathbb{P}[Y=1\mid P=p,\widehat{Y}=1]=\text{cst},\ \forall p$	
calibration	Chouldechova (2017)	$\mathbb{P}[Y=1\mid P=p,S=s]=\text{cst}_s,\ \forall p,s$	$Y\perp\!\!\!\perp P\mid\widehat{Y}$
well-calibration	Chouldechova (2017)	$\mathbb{P}[Y=1\mid P=p,S=s]=s,\ \forall p,s$	
accuracy equality	Berk et al. (2017)	$\mathbb{P}[\widehat{Y}=Y\mid P=p]=\text{cst},\ \forall p$	
treatment equality	Berk et al. (2017)	$\dfrac{\text{FN}_p}{\text{FP}_p}=\text{cst}_p,\ \forall p$	

Dependence measures I

Group fairness is characterized be independence or conditional independence properties. Given two random variables U and V,

$$C(U,V)=\begin{cases}\text{corr}[U,V] \text{ Pearson's correlation}\\ \text{corr}[F_U(U),F_V(V)] \text{ Spearman's rank correlation}\\ \tau[U,V] \text{ Kendall's tau}\end{cases}$$

that can be extended to conditional measures, as Lawrance (1976), since

$$\text{corr}[U,V]=\mathbb{E}\big[UV\big] \text{ when } \begin{cases}\mathbb{E}[U]=\mathbb{E}[V]=0\\ \mathbb{E}[U^2]=\mathbb{E}[V^2]=1\end{cases}$$

$$\begin{cases}\textbf{Demographic Parity}: \widehat{Y}\perp\!\!\!\perp P & \Longrightarrow C(\widehat{Y},P)=0\\ \textbf{Equalized Odds}: \widehat{Y}\perp\!\!\!\perp P|Y & \Longrightarrow C(\widehat{Y},P|Y=y)=0,\ \forall y\end{cases}$$

Dependence measures II

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

$$HGR(U,V)=\max\left\{\text{corr}[f(U),g(V)]\right\}=\max_{f\in\mathcal{S}_U,g\in\mathcal{S}_V}\left\{\mathbb{E}[f(U)g(V)]\right\}$$

where $\mathcal{S}_U=\{f:\mathcal{U}\to\mathbb{R}:\mathbb{E}[f(U)]=0 \text{ and } \mathbb{E}[f(U)^2]=1\}$ and similarly $\mathcal{S}_V$.
One can also consider a conditional version,

$$HGR(U,V|Z)=\max_{f\in\mathcal{S}_{U|Z},g\in\mathcal{S}_{V|Z}}\left\{\mathbb{E}[f(U)g(V)|Z]\right\}$$

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

This measure can be used to characterize independence,

$$U\perp\!\!\!\perp V \Longleftrightarrow HGR(U,V)=0,$$

and if (U,V) is a Gaussian vector, $HGR(U,V)=|\text{corr}(U,V)|$.

Dependence measures III

Thus, this measure can be used to characterize fairness,

$$\begin{cases} \textbf{Demographic Parity}: & \widehat{Y} \perp\!\!\!\perp P & \Longleftrightarrow HGR(\widehat{Y}, P) = 0 \\ \textbf{Equalized Odds}: & \widehat{Y} \perp\!\!\!\perp P|Y & \Longleftrightarrow HGR(\widehat{Y}, P|Y = y) = 0, \ \forall y \end{cases}$$

HGR can be difficult to estimate, but one can use some Neural Networks

$$HGR_{NN}(U, V) = \max_{\boldsymbol{\omega}_f, \boldsymbol{\omega}_g} \left\{ \mathbb{E}[f_{\boldsymbol{\omega}_f}(U) g_{\boldsymbol{\omega}_g}(V)] \right\}$$

See also Breiman and Friedman (1985) on the estimation of this maximal correlation, in the context of regression

Dependence measures IV

More generally, $\boldsymbol{V}$ can be a vector on $\mathcal{V} \subset \mathbb{R}^k$, then

$$HGR(U, \boldsymbol{V}) = \max_{h:\mathcal{V}\to\mathbb{R}} \big\{HGR[U, h(\boldsymbol{V})]\big\} = \max_{f\in\mathcal{S}_U, g\in\mathcal{S}_V} \big\{\mathbb{E}[f(U)g(\boldsymbol{V})]\big\}$$

where $\mathcal{S}_V = \{g : \mathcal{V} \to \mathbb{R} : \mathbb{E}[g(\boldsymbol{V})] = 0 \text{ and } \mathbb{E}[g(\boldsymbol{V})^2] = 1\}$. A conditional version exists, and one can estimate that measure using a neural network,

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

$$\begin{cases} \textbf{Demographic Parity}: & \widehat{Y} \perp\!\!\!\perp \boldsymbol{P} & \Longleftrightarrow HGR(\widehat{Y}, \boldsymbol{P}) = 0 \\ \textbf{Equalized Odds}: & \widehat{Y} \perp\!\!\!\perp \boldsymbol{P} | Y & \Longleftrightarrow HGR(\widehat{Y}, \boldsymbol{P} | Y = y) = 0,\ \forall y \end{cases}$$

Adversarial Approach I

In a classical ML or econometric pricing model, solve

$$\underset{\boldsymbol{\theta}}{\text{argmin}}\left\{\mathcal{L}\big(\widehat{y},y\big)\right\},\text{ where }\mathcal{L}\big(\widehat{y},y\big)=\sum_{i=1}^{n}\ell\big(\widehat{y}_i,y_i\big)\text{ and }\widehat{y}=h_{\boldsymbol{\theta}}(\boldsymbol{x})$$

either related to some loss, or some log-likelihood,
To avoid over-fit: penalize complexity (penalty $\mathcal{P}$)

$$\underset{\boldsymbol{\theta}}{\text{argmin}}\left\{\mathcal{L}\big(h_{\boldsymbol{\theta}}(\boldsymbol{x}),y\big)+\lambda\mathcal{P}(h_{\boldsymbol{\theta}})\right\}$$

Adversarial Approach II

Inspired by Goodfellow et al. (2018) (but also Bechavod and Ligett (2017) or Cho et al. (2020)), to avoid un-fairness: penalize according to $HGR(\widehat{y},p)$ (for demographic parity),

$$\underset{\boldsymbol{\theta},\boldsymbol{\omega}_f,\boldsymbol{\omega}_g}{\text{argmin}}\left\{\mathcal{L}\big(h_{\boldsymbol{\theta}}(\boldsymbol{x}),y\big)+\lambda\text{HGR}_{\boldsymbol{\omega}_f,\boldsymbol{\omega}_g}(h_{\boldsymbol{\theta}}(\boldsymbol{x}),p)\right\}$$

i.e.

$$\underset{\boldsymbol{\theta}}{\text{argmin}}\left\{\max_{\boldsymbol{\omega}_f,\boldsymbol{\omega}_g}\left\{\mathcal{L}(h_{\boldsymbol{\theta}}(\boldsymbol{X}),Y)+\lambda\mathbb{E}_{(\boldsymbol{X},S)\sim\mathcal{D}}\big(\widehat{f}_{\boldsymbol{\omega}_f}(h_{\boldsymbol{\theta}}(\boldsymbol{X}))\widehat{g}_{\boldsymbol{\omega}_g}(P)\right\}\right\}$$

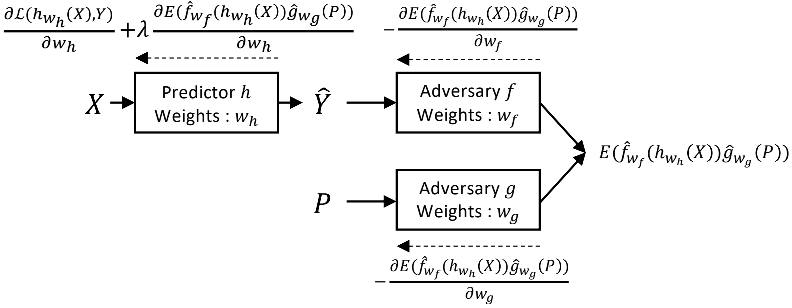
or $HGR(\widehat{y},p|y)$ (for equalized odds), i.e. when $y\in\{0,1\}$

$$\begin{aligned}\underset{\boldsymbol{\theta}}{\text{argmin}}\Big\{\max_{\boldsymbol{\omega}_{f0},\boldsymbol{\omega}_{g0},\boldsymbol{\omega}_{f1},\boldsymbol{\omega}_{g1}}\Big\{\mathcal{L}(h_{\boldsymbol{\theta}}(\boldsymbol{X}),Y)\quad&+\quad\lambda_0\mathbb{E}_{(\boldsymbol{X},P)\sim\mathcal{D}_0}\big(\widehat{f}_{\boldsymbol{\omega}_{f_0}}(h_{\boldsymbol{\theta}}(\boldsymbol{X}))\widehat{g}_{\boldsymbol{\omega}_{g_0}}(P)\big)\\&+\quad\lambda_1\mathbb{E}_{(\boldsymbol{X},P)\sim\mathcal{D}_1}\big(\widehat{f}_{\boldsymbol{\omega}_{f_1}}(h_{\boldsymbol{\theta}}(\boldsymbol{X}))\widehat{g}_{\boldsymbol{\omega}_{g_1}}(P)\big)\Big\}\Big\}\end{aligned}$$

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

Adversarial Approach III

$$\underset{\boldsymbol{\theta}}{\text{argmin}}\left\{\max_{\boldsymbol{\omega}_{f0},\boldsymbol{\omega}_{g0},\cdots,\boldsymbol{\omega}_{fk},\boldsymbol{\omega}_{gk}}\left\{\mathcal{L}(h_{\boldsymbol{\theta}}(\boldsymbol{X}),Y)+\sum_{y\in\Omega_y}\lambda_y\mathbb{E}_{(\boldsymbol{X},P)\sim\mathcal{D}_y}(\widehat{f}_{\boldsymbol{\omega}_{f_y}}(h_{\boldsymbol{\theta}}(\boldsymbol{X}))\widehat{g}_{\boldsymbol{\omega}_{g_y}}(P))\right\}\right\}$$



Dealing with high dimension I

- geographic / spatial information, $\boldsymbol{X}_g$
- car type / make / model, $\boldsymbol{X}_c$
- other classical ratemaking variables, $\boldsymbol{X}$ (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)

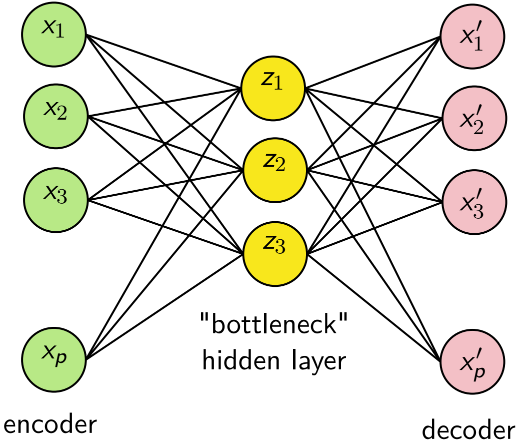
$$\min_{\boldsymbol{P}\in\Pi}\left\{\|\boldsymbol{X}-\boldsymbol{P}^\top\boldsymbol{P}\boldsymbol{X}\|_F^2\right\}\text{ s.t. }\text{rank}(\boldsymbol{P})=k$$

where Π is the set of projection matrices.

$$\begin{tikzcd}\boxed{\boldsymbol{X}\in\mathcal{X}} \arrow[r, bend right, "\boldsymbol{P}"'] & \boxed{\boldsymbol{Z}\in\mathcal{Z}} \arrow[r, bend right, "\boldsymbol{P}^\top"'] & \boxed{\boldsymbol{X}'\in\mathcal{X}}\end{tikzcd}$$

$$\min\|\boldsymbol{X}-\boldsymbol{X}'\|^2=\min\|\boldsymbol{X}-\boldsymbol{P}^\top\boldsymbol{P}\boldsymbol{X}\|^2$$

$$=\min\sum_{i=1}^{n}(\boldsymbol{P}^\top\boldsymbol{P}\boldsymbol{x}_i-\boldsymbol{x}_i)^\top(\boldsymbol{P}^\top\boldsymbol{P}\boldsymbol{x}_i-\boldsymbol{x}_i)$$



Dealing with high dimension III

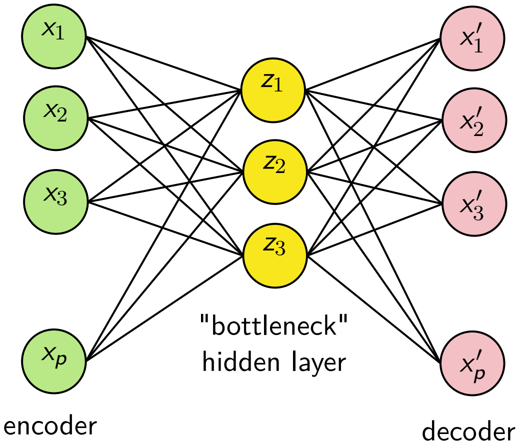
Autoencoder

$$\min_{\psi}\left\{\|\boldsymbol{X}-\psi\circ\varphi\boldsymbol{X}\|_F^2\right\}$$

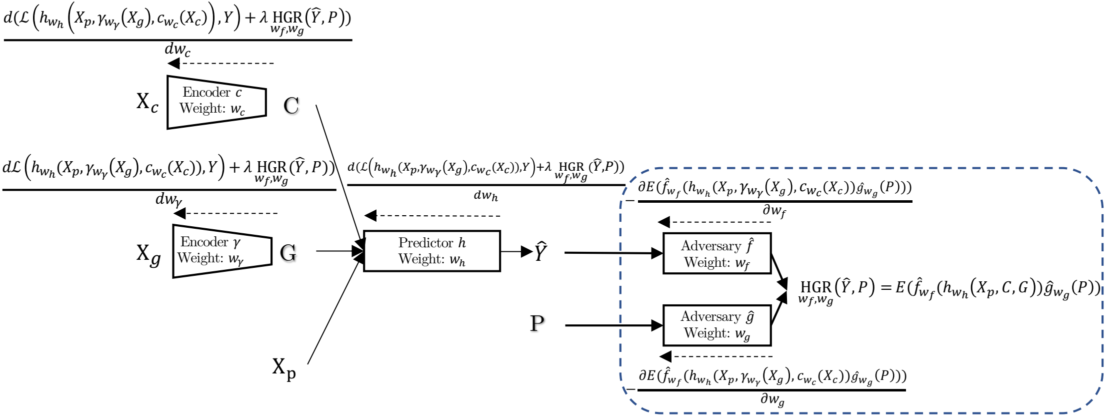
$$\boxed{\boldsymbol{X}\in\mathcal{X}} \xrightarrow{\varphi} \boxed{\boldsymbol{Z}\in\mathcal{Z}} \xrightarrow{\psi} \boxed{\boldsymbol{X}'\in\mathcal{X}}$$

$$\min\|\boldsymbol{X}-\boldsymbol{X}'\|^2=\min\|\boldsymbol{X}-\psi\circ\varphi(\boldsymbol{X})\|^2$$

$$\min\sum_{i=1}^{n}\big(\psi\circ\varphi(\boldsymbol{x}_i)-\boldsymbol{x}_i\big)^\top\big(\psi\circ\varphi(\boldsymbol{x}_i)-\boldsymbol{x}_i\big)$$

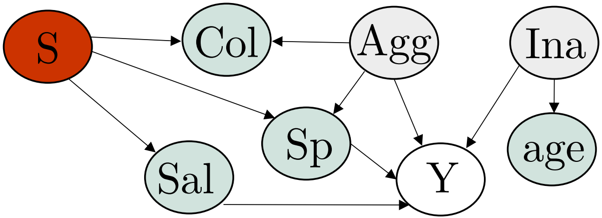


Dealing with high dimension IV

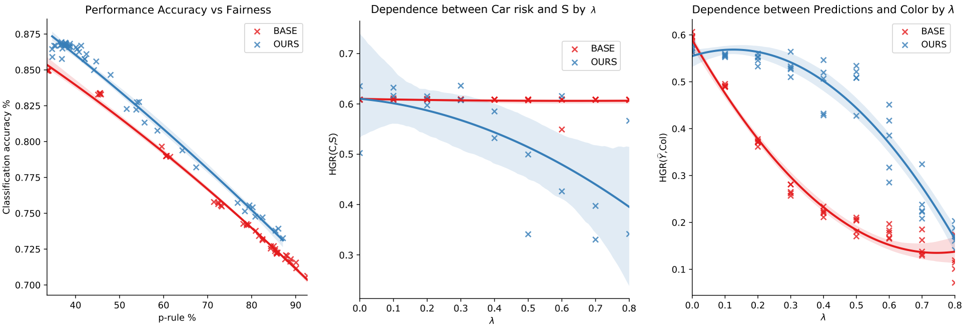


Application on synthetic data I

- 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



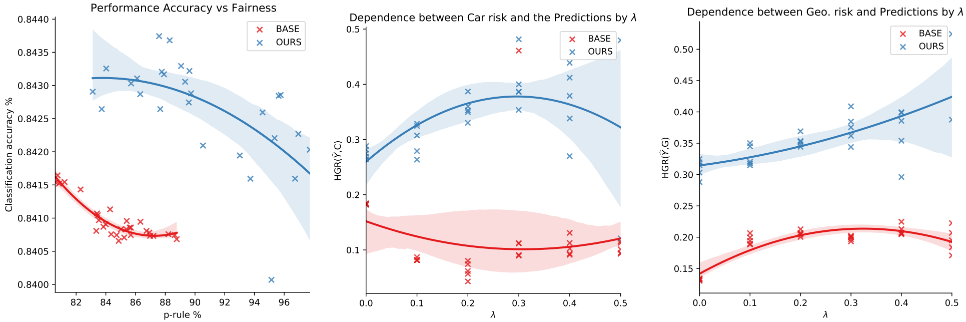
Application on synthetic data II



- λ: fairness penalty
- p-rule: $\min\left\{\dfrac{\mathbb{P}(\widehat{Y}=1|P=1)}{\mathbb{P}(\widehat{Y}=1|P=0)},\dfrac{\mathbb{P}(\widehat{Y}=1|P=0)}{\mathbb{P}(\widehat{Y}=1|P=1)}\right\}$

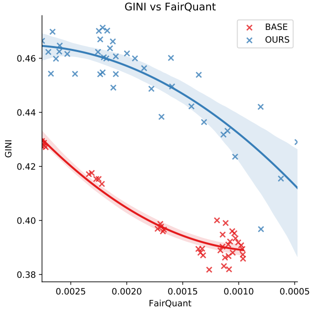
Application on real data (pricing game 2015) I

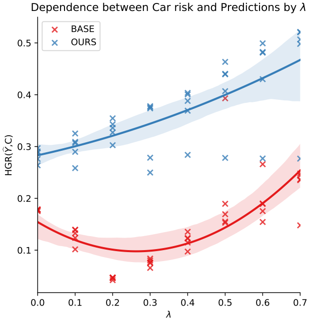
$y \in \{0, 1\}$ (claim occurrence), and p is the (binary) gender

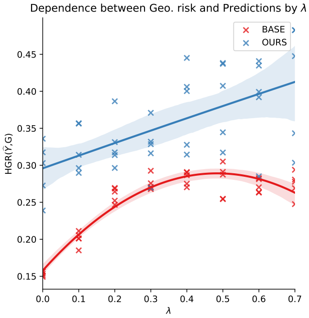


Application on real data (pricing game 2015) II

$y \in \{0, 1, 2+\}$ (claim frequency), and p is the (binary) gender



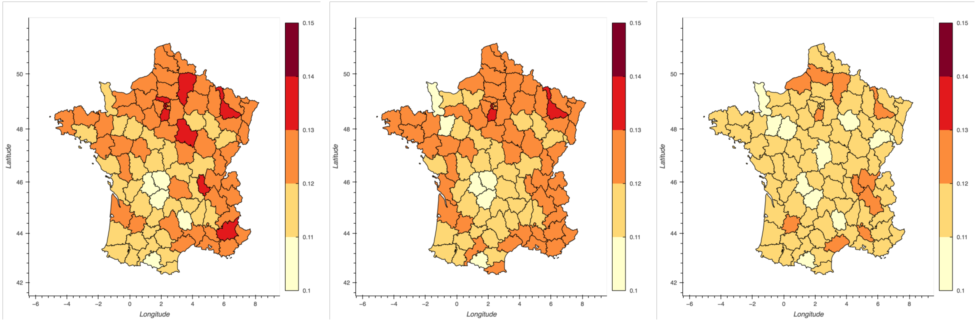




see Grari et al. (2022) for more examples (including the case where $y \in \mathbb{R}^+$)

Application on real data (pricing game 2015) III

$y \in \{0, 1, 2+\}$ (claim frequency), and p is spatial information (redlining)



From correlation to causality I

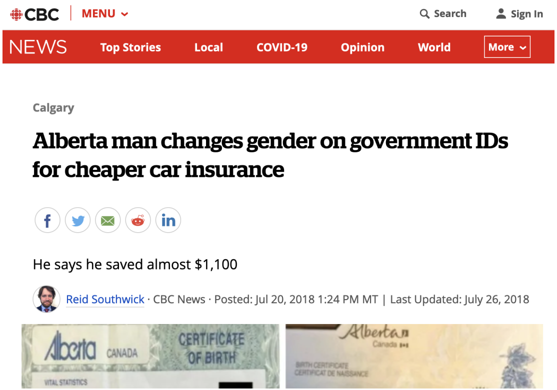
- "*classifying projection methods as using demographic/actuarial models or non-demographic/causal models*"

  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



CBC | MENU   Search   Sign In

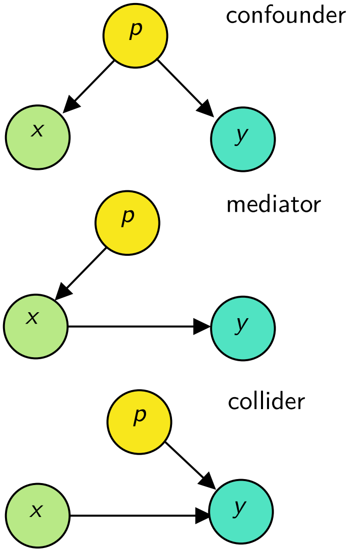
NEWS   Top Stories   Local   COVID-19   Opinion   World   More

Calgary

Alberta man changes gender on government IDs for cheaper car insurance

He says he saved almost $1,100

Reid Southwick · CBC News · Posted: Jul 20, 2018 1:24 PM MT | Last Updated: July 26, 2018



- DAGs are important
- Looking for a counterfactual

From correlation to causality III

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)

$$D(\widehat{y}_i, \widehat{y}_j) \text{ or } D(s_i, s_j) \leq d(\boldsymbol{x}_i, \boldsymbol{x}_j),\ \forall i, j = 1, \cdots, n.$$

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.

$$\mathbb{P}\big[Y^\star_{P\leftarrow p} = y | \boldsymbol{X} = \boldsymbol{x}\big] = \mathbb{P}\big[Y^\star_{P\leftarrow p'} = y | \boldsymbol{X} = \boldsymbol{x}\big],\ \forall p', \boldsymbol{x}, y.$$

From correlation to causality IV

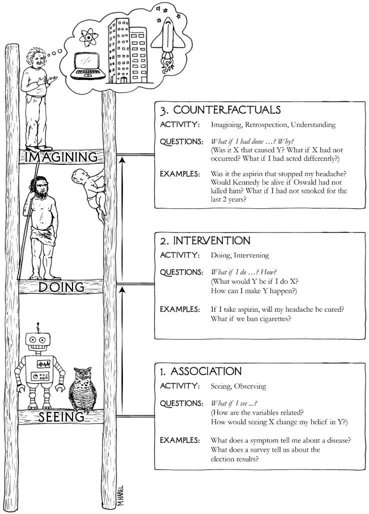
- counterfactuals

  (*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 V

Causal inference litterature,

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

	treatement	outcome			age	gender	height	weight
	t_i	y_i	$y^\star_{i,T\leftarrow 1}$	$y^\star_{i,T\leftarrow o}$	$x_{1,i}$	$x_{2,i}$	$x_{3,i}$	$x_{4,i}$
1	1	121	121	?	37	F	160	56
2	0	109	?	109	28	F	156	54
3	1	162	162	?	53	M	190	87

There will be a significant impact of treatment t on y if $y^\star_{T\leftarrow 0} \neq y^\star_{T\leftarrow 1}$ (see Rubin (1974), Hernán and Robins (2010) or Imai (2018)).
The causal effect for individual i is $\tau_i = y^\star_{i,T\leftarrow 1} - y^\star_{i,T\leftarrow 0}$

From correlation to causality VI

One can define the sample average treatment effect (SATE)

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

the average treatment effect (ATE)

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

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

$$\tau(\boldsymbol{x}) = \text{CATE}(\boldsymbol{x}) = \mathbb{E}\big[Y^\star_{i,T\leftarrow 1} - Y^\star_{i,T\leftarrow 0}\big|\boldsymbol{X} = \boldsymbol{x}\big]$$

References I

,

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

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

Barry, L. and Charpentier, A. (2022). The Fairness of Machine Learning in Insurance: New Rags for an Old Man? . *ArXiv*.

Bechavod, Y. and Ligett, K. (2017). Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044*.

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.

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association*, 80(391):580–598.

Charpentier, A. (2022a). *Assurance: biais, discrimination et équité*. Institut Louis Bachelier.

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

References II

Cho, J., Hwang, G., and Suh, C. (2020). A fair classifier using mutual information. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2521–2526. IEEE.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163.

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

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.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.

Gebelein, H. (1941). Das statistische problem der korrelation als variations- und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung. *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik*, 21(6):364–379.

Gomery, S., Renault, O., and John, N. (2011). Gender neutral. *Canadian Underwriter*.

Goodfellow, I., McDaniel, P., and Papernot, N. (2018). Making machine learning robust against adversarial inputs. *Communications of the ACM*, 61(7):56–66.

References III

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

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

Hernán, M. A. and Robins, J. M. (2010). Causal inference.

Hirschfeld, H. O. (1935). A connection between correlation and contingency. *Mathematical Proceedings of the Cambridge Philosophical Society*, 31(4):520524.

Hudson, R. (2007). Mortality projections and unisex pricing of annuities in the uk. *Journal of Financial Regulation and Compliance*.

Hume, D. (1739). *A Treatise of Human Nature*. Cambridge University Press Archive.

Imai, K. (2018). *Quantitative social science: an introduction*. Princeton University Press.

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

Keilman, N. (2003). Types of models for projecting mortality. *Perspectives on mortality forecasting*, 1:19–27.

References IV

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2017). Human Decisions and Machine Predictions. *The Quarterly Journal of Economics*, 133(1):237–293.

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.

Lawrance, A. (1976). On conditional and partial correlation. *The American Statistician*, 30(3):146–149.

Rényi, A. (1959). On measures of dependence. *Acta mathematica hungarica*, 10(3-4):441–451.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Schmeiser, H., Störmer, T., and Wagner, J. (2014). Unisex insurance pricing: consumers perception and market implications. *The Geneva Papers on Risk and Insurance-Issues and Practice*, 39(2):322–350.

Shi, P. and Shi, K. (2021). Nonlife insurance risk classification using categorical embedding. *SSRN*, 3777526.

Thiery, Y. and Van Schoubroeck, C. (2006). Fairness and equality in insurance classification. *The Geneva Papers on Risk and Insurance-Issues and Practice*, 31(2):190–211.