ISP IMAGE & SIGNAL PROCESSING GROUP

FFT Seminar - 7^{th} February 2022

Analyzing the structural relationship between distributions using optimal transport

PAULA GORDALIZA PASTOR

basque center for applied mathematics

EXCELENCIA SEVERO OCHOA

EUSKO JAURLARITZA GOBIERNO VASCO

Joint work with

E. del Barrio F. Gamboa JM. Loubes

P. Besse L. Risser

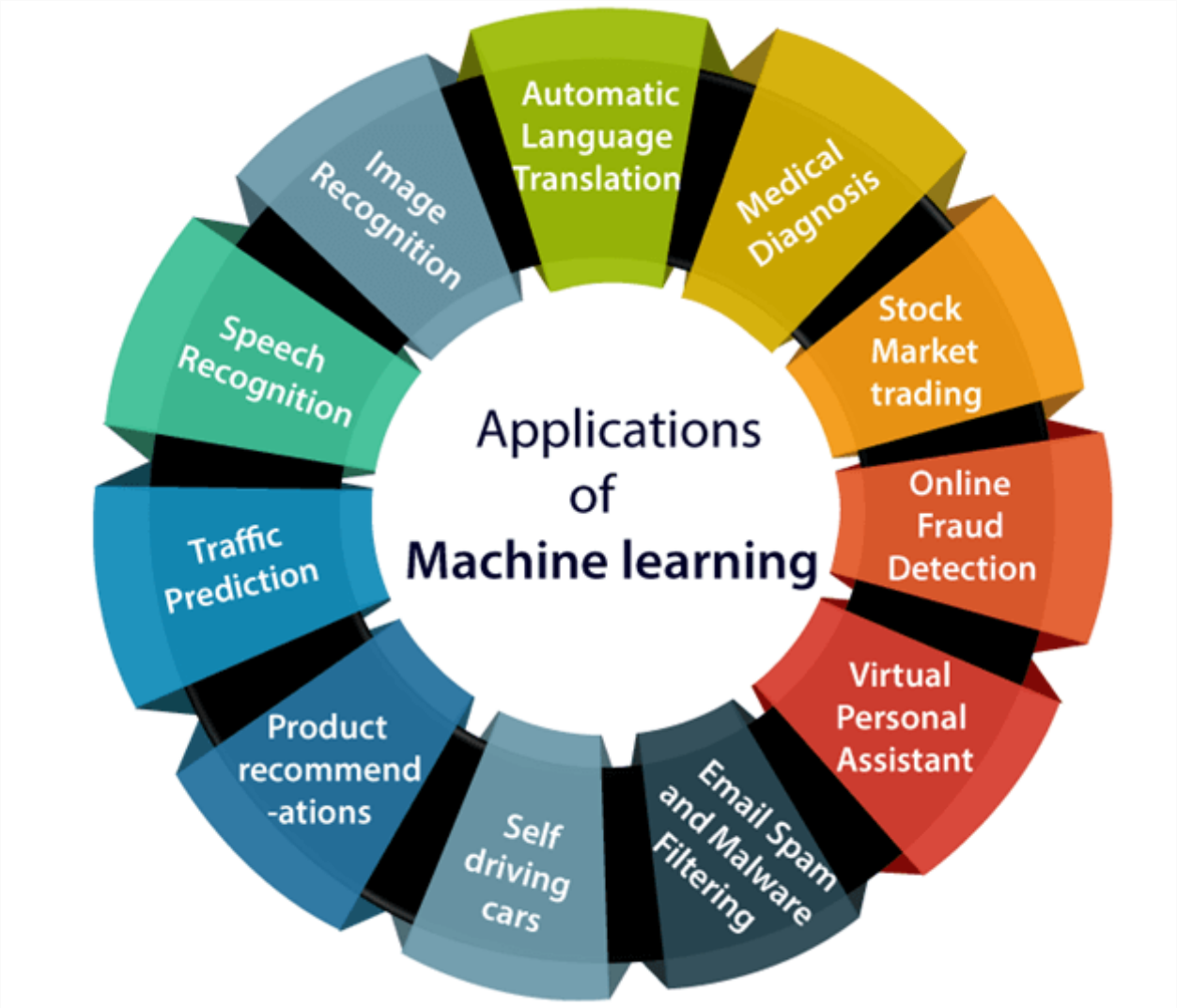
imUVa

INSTITUT de MATHEMATIQUES de TOULOUSE

elise

European Network of AI Excellence Centres

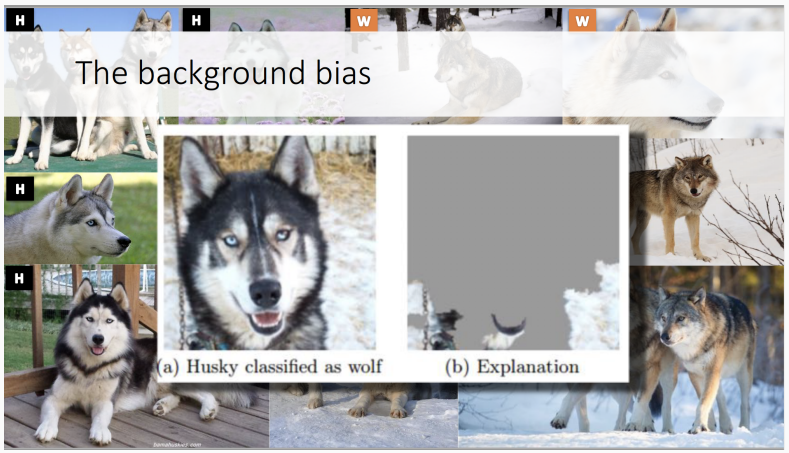
The generalization of applications based on ML models in the everyday life and the professional world has been accompanied by concerns about the ethical issues that may arise from the adoption of these technologies



AI technologies make life easier, but they are not absolutely objective...

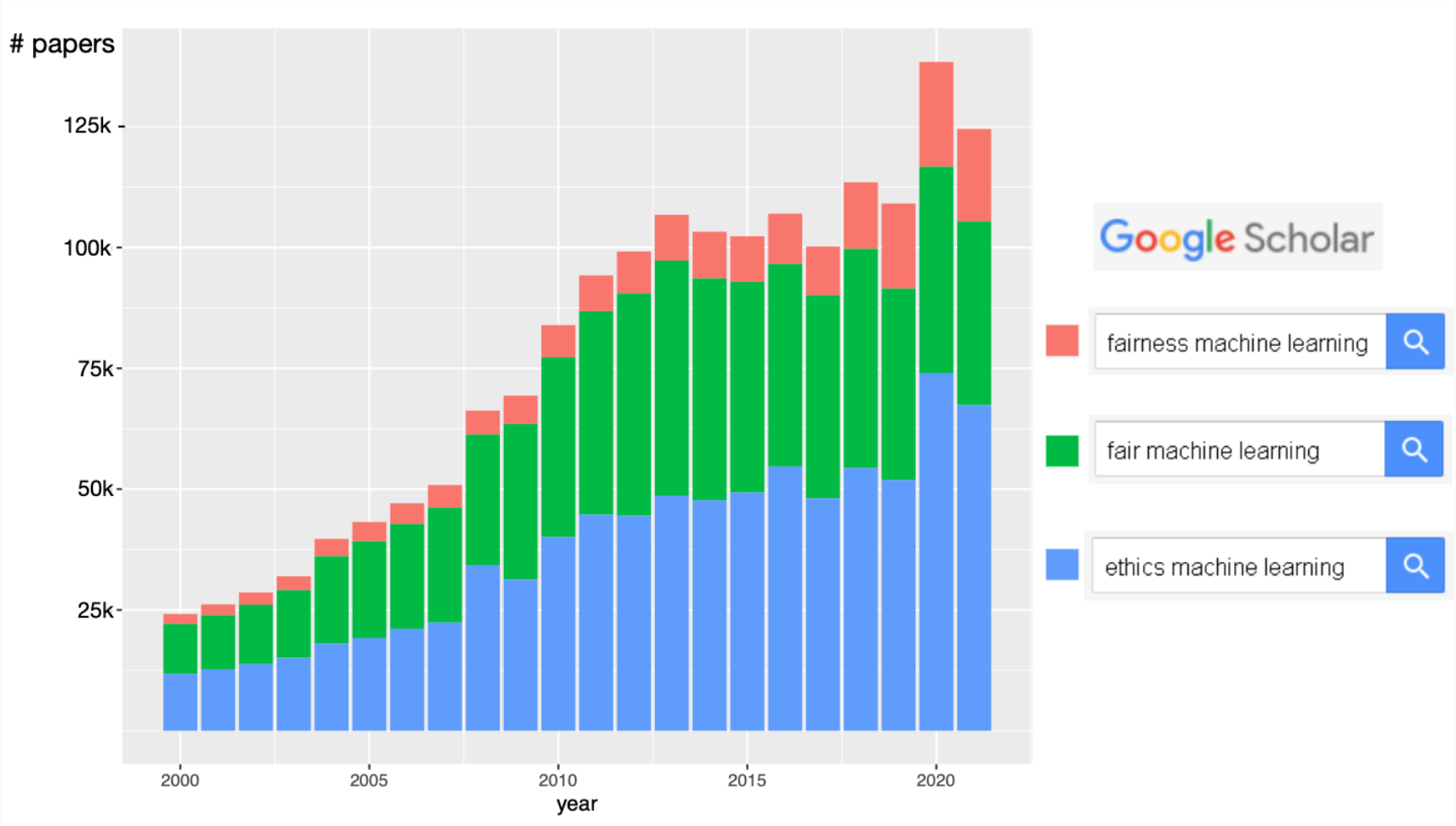
ML algorithms that are meant to automatically take accurate and efficient decisions that mimic, and even sometimes outmatch human expertise, rely heavily on potentially biased data

Inherent social bias existing in the population that is used to generate the training set



Bias without social unfairness

Fairness has become one of the most popular topics in ML over the last years and the research community is investing a **large amount of effort** in this area.



COMPAS recidivism black bias

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. *(Source: ProPublica analysis of data from Broward County, Fla.)*

Results from job platform XING

Search query	Work experience	Education experience	Profile views	Candidate	Xing ranking
Brand Strategist	146	57	12992	male	1
Brand Strategist	327	0	4715	female	2
Brand Strategist	502	74	6978	male	3
Brand Strategist	444	56	1504	female	4
Brand Strategist	139	25	63	male	5
Brand Strategist	110	65	3479	female	6
Brand Strategist	12	73	846	male	7
Brand Strategist	99	41	3019	male	8
Brand Strategist	42	51	1359	female	9
Brand Strategist	220	102	17186	female	10

TABLE II: Top k results on www.xing.com (Jan 2017) for the job serach query "Brand Strategist".

Less qualified male candidates were highly ranked

ML algorithms in banking industry: The Adult Data Set

Obtaining fairness is a more complicated task that needs mathematical models

Methods for imposing a level of fairness (Oneto and Chiappa, 2020)

Fairness through Optimal Transport

(A) Pre-processing the training data

Kamiran and Calders (2009,2010,2012)
Zemel et al. (2013)
Feldman et al. (2015)
Johndrow and Lum (2017)
Gordaliza et al. (2019)

(C) Post-processing the model outputs

Pedreschi et al. (2009)
Hardt et al. (2016)
Kusner et al. (2017)
Chzhen et al. (2019)

(B) In-processing to control the training phase of the algorithm

(i) Lagrange multipliers
Berk et al. (2017a)
Zafar et al. (2017a, 2019)
Agarwal et al. (2018)

(ii) Add penalties to the objective
Bechavod and Ligett (2017)
Dwork et al. (2018)
Donini et al. (2018)

↓

Fairness through empirical risk minimization

Recently very important : **causal approach** S. Chiappa et al. (2019) and **counterfactual approach** de Lara et al. (2021)

Consider $\left(\Omega \subset \mathbb{R}^d, \mathcal{B}, \mathbb{P}\right)$, $\mathcal{B}$ Borel $\sigma-$algebra of subsets of $\mathbb{R}^d$ and $d \geqslant 1$

Protected attribute	Visible attributes	Target	Outcome
$S \in \mathcal{S}$	$X \in \mathcal{X} \subset \mathbb{R}^d$	$Y \in \mathbb{R}^d$	$\hat{Y} = f(X, S),\ f \in \mathcal{F}$

Definition of fairness as independence criterion

Perfect fairness requires that S does not play any role in the forecast $\hat{Y}$

(I) **Statistical Parity** : $\hat{Y} \perp\!\!\!\perp S$

(II) **Equality of Odds** : $\hat{Y} \perp\!\!\!\perp S \mid Y$

Consider $\left(\Omega \subset \mathbb{R}^d, \mathcal{B}, \mathbb{P}\right)$, $\mathcal{B}$ Borel $\sigma-$algebra of subsets of $\mathbb{R}^d$ and $d \geqslant 1$

Protected attribute	Visible attributes	Target	Outcome
$S \in \mathcal{S} = \{0, 1\}$	$X \in \mathcal{X} \subset \mathbb{R}^d$	$Y \in \{0, 1\}$	$\hat{Y} = g(X, S),\ g \in \mathcal{G}$
$\begin{cases} 0 & unfavored \\ 1 & favored \end{cases}$		$\begin{cases} 0 & failure \\ 1 & success \end{cases}$	$g : \mathbb{R}^d \to \{0, 1\}$

Definition of fairness as independence criterion

Perfect fairness requires that S does not play any role in the forecast $\hat{Y}$

(I) **Statistical Parity** (SP) (Dwork et al., 2012): $\hat{Y} \perp\!\!\!\perp S$

$$\mathbb{P}(\hat{Y} = 1 \mid S = 0) = \mathbb{P}(\hat{Y} = 1 \mid S = 1)$$

(II) **Equality of Odds** (EO) (Hardt et al., 2016): $\hat{Y} \perp\!\!\!\perp S \mid Y$

$$\mathbb{P}(\hat{Y} = i \mid Y = i, S = 0) = \mathbb{P}(\hat{Y} = i \mid Y = i, S = 1),\ i = 0, 1$$

Protected attribute	Visible attributes	Target	Outcome
$S \in \mathcal{S} = \{0,1\}$	$X \in \mathcal{X} \subset \mathbb{R}^d$	$Y \in \{0,1\}$	$\hat{Y} = g(X,S),\ g \in \mathcal{G}$
$\begin{cases} 0 & unfavored \\ 1 & favored \end{cases}$		$\begin{cases} 0 & failure \\ 1 & success \end{cases}$	$g : \mathbb{R}^d \to \{0,1\}$

The **Disparate Impact** of the classifier $g \in \mathcal{G}$, with respect to (X, S) is defined as

$$DI(g, X, S) = \frac{\mathbb{P}(g(X,S) = 1 \mid S = 0)}{\mathbb{P}(g(X,S) = 1 \mid S = 1)} \in (0, 1]$$

- Ideal scenario: g achieves Statistical Parity $\Leftrightarrow DI(g, X, S) = 1$

Definition

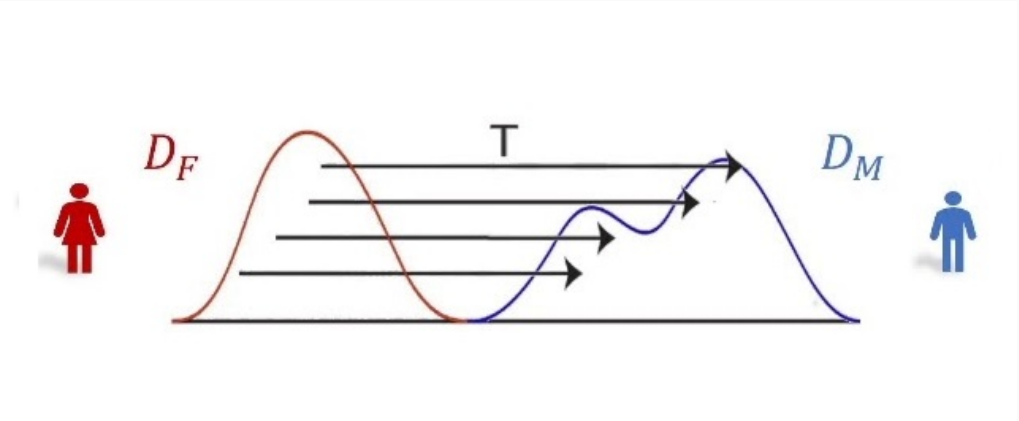
A classifier $g : \mathbb{R}^d \to \{0,1\}$ is said not to have **Disparate Impact at level** $\tau \in [0,1]$, with respect to (X, S), if $DI(g, X, S) > \tau$.

- $\tau_0 = 4/5 \dashrightarrow 80\%$ rule (*1971, State of California Fair Employment Comission*)

Obtaining fairness using Optimal Transport Theory

GORDALIZA ET AL. (2019)

Proceedings of the 36th International Conference on Machine Learning



Some preliminaries: Wasserstein distance

Kantorovich formulation: (Villani, 2003)

- A transportation plan between two probabilities P and Q on $\mathbb{R}^d$ is a joint probability π on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals P and Q
- The optimal transportation cost is the minimal value of

$$I[\pi] = \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x,y) d\pi(x,y)$$

  among all transportation plans π between P and Q

Wasserstein distance: If $c(x,y) = c_p(x,y) = \|x-y\|^p, p \geq 1$, and $\Pi(P,Q)$ denotes the set of probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals P and Q, then

$$\mathcal{W}_p(P,Q) = \left(\inf_{\pi \in \Pi(P,Q)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|^p d\pi(x,y) \right)^{1/p}$$

defines a metric in the set $\mathcal{F}_p(\mathbb{R}^d)$ of probabilities on $\mathbb{R}^d$ with finite p-th moment.

Some preliminaries: Wasserstein Variation

Wasserstein p-variation of a collection of probabilities $\mu_1, \ldots, \mu_J \in \mathcal{F}_p\left(\mathbb{R}^d\right)$ w.r.t. weights $\omega_1, \ldots, \omega_J > 0$, is defined as

$$V_p\left(\nu_1, \ldots, \nu_J\right) = \inf_{\eta \in \mathcal{F}_p\left(\mathbb{R}^d\right)} \left(\sum_{j=1}^{J} \omega_j \mathcal{W}_p^p(\mu_j, \eta) \right)^{1/p}$$

Case $p = 2$:

- Existence and uniqueness (under some smoothness assumptions) of a minimizer of $\eta \mapsto \frac{1}{J} \sum_{j=1}^{J} \mathcal{W}_2^2(\mu_j, \eta)$, called Wasserstein barycenter μ_B of the μ_j's (Agueh and Carlier, 2011)

$$V_2\left(\mu_1, \ldots, \mu_J\right) = \left(\frac{1}{J} \sum_{j=1}^{J} \mathcal{W}_2^2(\mu_j, \mu_B)\right)^{1/2}$$

- Empirical versions (Boissard et al., 2015), (Le Gouic and Loubes, 2017)

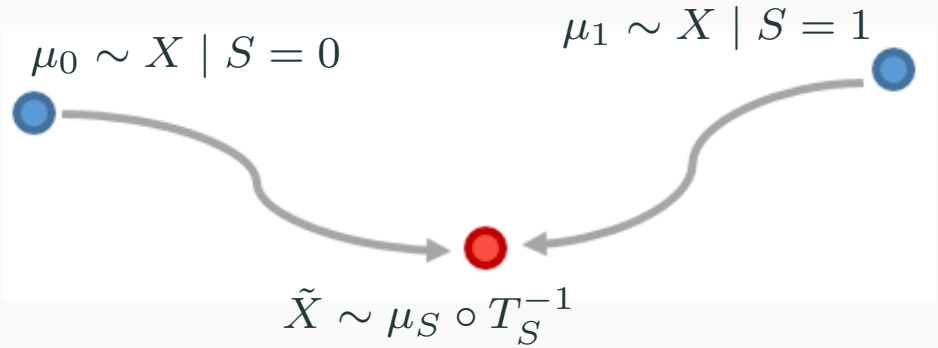
Protected attribute	Visible attributes	Target	Outcome
$S \in \mathcal{S} = \{0, 1\}$	$X \in \mathcal{X} \subset \mathbb{R}^d$	$Y \in \{0, 1\}$	$\hat{Y} = g(X, S),\ g \in \mathcal{G}$ $g : \mathbb{R}^d \to \{0, 1\}$

Find $\tilde{X} = T_S(X)$ such that $\mathcal{L}\left(T_0(X) \mid S = 0\right) = \mathcal{L}\left(T_1(X) \mid S = 1\right)$

$$\Downarrow$$

$\mathcal{L}\left(g(\tilde{X}) \mid S = 0\right) = \mathcal{L}\left(g(\tilde{X}) \mid S = 1\right)$, for all $g \in \mathcal{G}$

Fair classifier: $g \circ T_S \in \mathcal{F}_{SP}$, for all $g \in \mathcal{G}$



Questions:

a) Best choice for the distribution $\tilde{X} \sim \nu$?

b) Optimal way of transporting μ_0, μ_1 to ν?

Upper bound for the price for fairness (Gordaliza et al., 2019) If $\eta_s(x) = \mathbb{P}(Y = 1 \mid X = x, S = s), s \in \{0, 1\}$, is Lipschitz with constant $K_s > 0$ and $K = \max\{K_0, K_1\}$,

$$\mathcal{E}(T_S) := \inf_{g \in \mathcal{G}} \mathbb{P}(g(T_S(X)) \neq Y) - \inf_{g \in \mathcal{G}} R(g) \leq 2\sqrt{2}K \left(\sum_{s=0,1} \pi_s \mathcal{W}_2^2(\mu_s, {\mu_s}_\sharp T_s) \right)^{\frac{1}{2}}.$$

Upper bound for the price for statistical parity

$$\mathcal{E}(\mathcal{F}_{SP}) \leq \inf_{T_S} \mathcal{E}(T_S) \leq 2\sqrt{2}K \left(\sum_{s=0,1} \pi_s \mathcal{W}_2^2(\mu_s, \mu_B) \right)^{\frac{1}{2}}$$

Reasonable and feasible solutions:

a) Wasserstein barycenter μ_B with weights $\pi_0 = P(S = 0)$ and $\pi_1 = P(S = 1)$

$$\mu_B \in \textit{argmin}_{\nu \in \mathcal{P}_2} \left\{ \pi_0 W_2^2(\mu_0, \nu) + \pi_1 W_2^2(\mu_1, \nu) \right\}$$

b) T_S optimal transport map carrying μ_S towards μ_B

$${\mu_S}_\sharp T_S = \mu_B$$

Partial Repair with Wasserstein barycenter

Target variable	Level of repair	Transformation	
$Z \sim \mu_B$	$\lambda \in [0,1]$	o.t.m. T_S	$\mu = \mu_S{}_\sharp T_S$ $T_S^{-1}(Z) \sim \mu_S$

Geometric repair (Feldman et al., 2015)

$$\tilde{\mu}_{S,\lambda} = \mathcal{L}(\lambda T_S(X) + (1-\lambda)X)$$

Random repair (Gordaliza et al., 2019)

$B \sim \mathcal{B}(\lambda)$, independent of (X, S, Y)

$$\tilde{\mu}_{S,\lambda} = \mathcal{L}(B T_S(X) + (1-B)X)$$

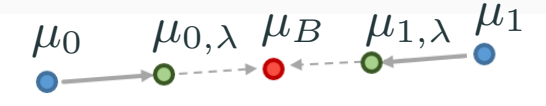
Unmodified variable	$0 \longleftarrow \lambda \longrightarrow 1$	Totally repaired variable
$\tilde{\mu}_{s,0} = \mathcal{L}(X \mid S = s)$	$\Updownarrow$	$\tilde{\mu}_{s,1} = \mathcal{L}(Z) = \mu_B$
Accuracy of $g(\tilde{X})$	$\longleftarrow$ **Trade-off** $\longrightarrow$	**Non-predictability** of S

$$d_{TV}(P,Q) = \min_{\pi \in \Pi(P,Q)} \pi(x \neq y)$$

Partial Repair with Wasserstein barycenter

Target variable	Level of repair	Transformation	
$Z \sim \mu_B$	$\lambda \in [0,1]$	o.t.m. T_S	$\mu = \mu_{S\sharp} T_S$ $R_S := T_S^{-1}(Z) \sim \mu_S$

Geometric repair (Feldman et al., 2015)



$$\tilde{\mu}_{S,\lambda} = \mathcal{L}(\lambda T_S(X) + (1-\lambda)X)$$

✗ The level of repair does not affect d_{TV}:

= in some examples

$$\begin{aligned} d_{TV}(\tilde{\mu}_{0,\lambda}, \tilde{\mu}_{1,\lambda}) \leq & \mathbb{P}(\lambda Z + (1-\lambda)R_0(Z) \\ & \neq \lambda Z + (1-\lambda)R_1(Z)) \\ = & \mathbb{P}(R_0(Z) \neq R_1(Z)). \end{aligned}$$

Random repair (Gordaliza et al., 2019)

$B \sim \mathcal{B}(\lambda)$, independent of (X, S, Y)

$$\tilde{\mu}_{S,\lambda} = \mathcal{L}(BT_S(X) + (1-B)X)$$

✓ The level of repair controls d_{TV}:

$$\begin{aligned} d_{TV}(\tilde{\mu}_{0,\lambda}, \tilde{\mu}_{1,\lambda}) \leq & 1 - \mathbb{P}(BZ + (1-B)R_0(Z) \\ = & BZ + (1-B)R_1(Z)) \\ \leq & 1 - \mathbb{P}(B=1) = 1-\lambda \end{aligned}$$

✓ The new risk is a mixture of the two errors:

$$\begin{aligned} R(g, \tilde{X}_\lambda) = & (1-\lambda)\mathbb{P}(g(X) \neq Y) \\ & + \lambda \mathbb{P}(g(T_S(X)) \neq Y) \end{aligned}$$

Application to the Adult Data Set (size 29.825)

$$Y = \begin{cases} 1 & \text{income exceeds } \$\ 50.000/year \\ 0 & \text{otherwise} \end{cases}$$

X

1) Age
2) Workclass
3) Final weight
4) Education
5) Education number
6) Marital status
7) Occupation
8) Relationship
9) **Gender**
10) Race
11) Capital gain
12) Capital loss
13) Hours per week

There exists any Disparate Impact with respect to...?

$$S = \begin{cases} 0 & female \\ 1 & male \end{cases}$$

Experiment:

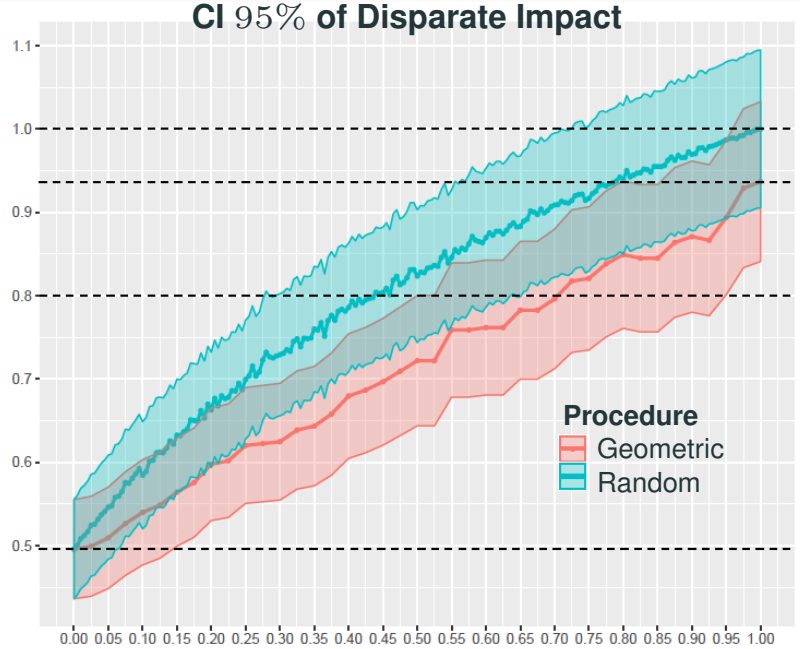
1. Split the data set : test $2.500/$ learning 27.325
2.

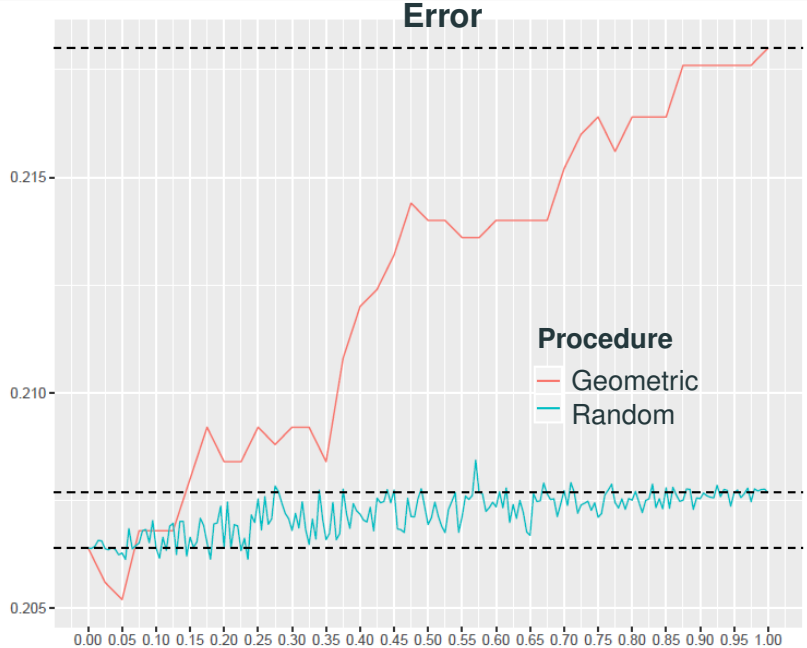
Statistical Model	Error	$\hat{DI}$	CI 95%
Logit	0.2064	0.496	$(0.437, 0.555)$
Random Forests	0.168	0.484	$(0.429, 0.54)$

3. Predict $g(X)$
4. Repair procedure $\rightarrow \tilde{X}$ $\dashrightarrow$
5. Predict $g(\tilde{X})$

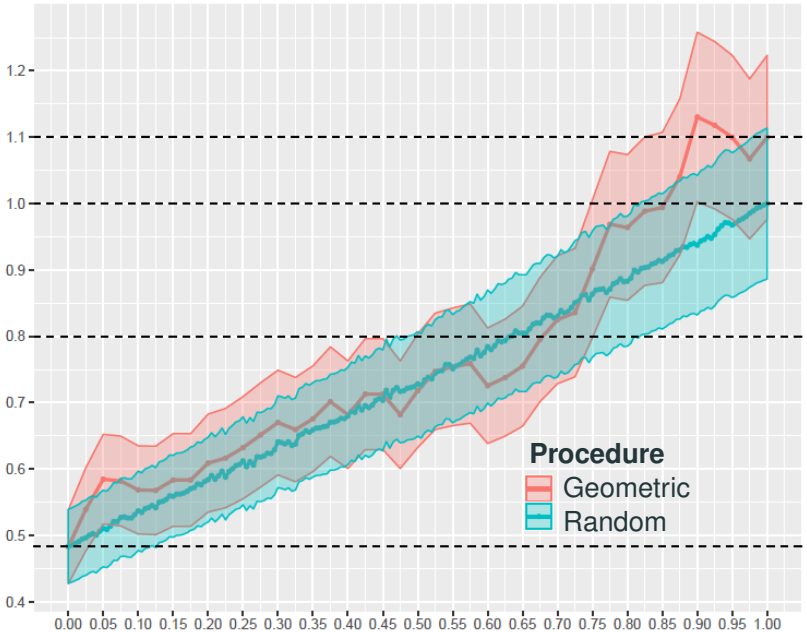
Random repair
vs.
Geometric repair
(Feldman et al., 2015)

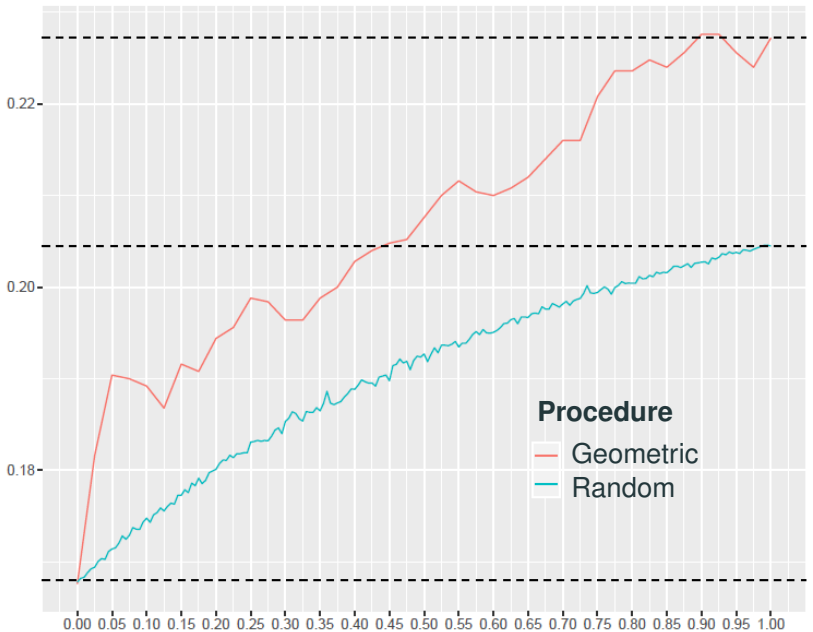
Logit





Random forest





Amount of repair λ

New criteria for statistical parity assessment

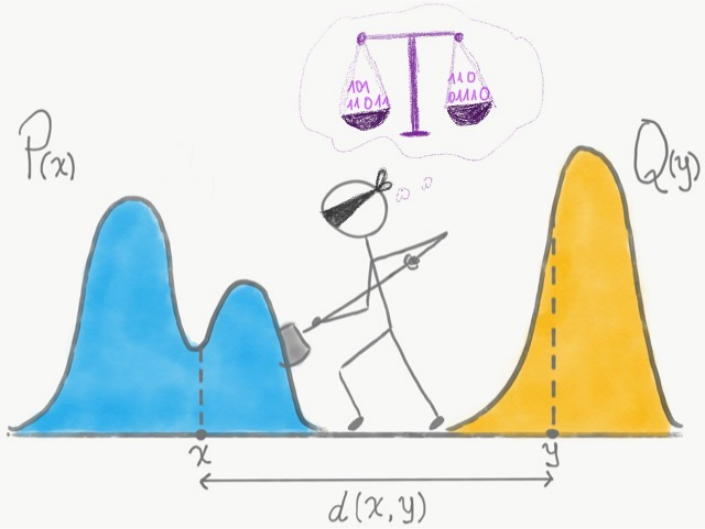
$$\varepsilon^* := \min_{g\in\mathcal{G}} BER(g, X, S) = \tfrac{1}{2}\left(1 - d_{TV}\left(\mu_0, \mu_1\right)\right),\ \mu_s = \mathcal{L}(X \mid S = s)$$

(Gordaliza et al., 2019)

Rejection of $H_0 :\ \rho(\mu_0, \mu_1) \geq \Delta_0$ $\Longrightarrow$ **Statistical certification that** $\mu_0 \approx \mu_1$

$H_0 : \mathcal{W}_p(\mu_0, \mu_1) \geq \Delta_0$ vs $H_a : \mathcal{W}_p(\mu_0, \mu_1) < \Delta_0$, for $\Delta_0 > 0$ and $p \geq 1$ ✓

Goal: CLT $\begin{cases} r_n\left(\mathcal{W}_p^p(\mu_{0,n}, \mu_1) - a_n\right) \\ r_{n,m}\left(\mathcal{W}_p^p(\mu_{0,n}, \mu_{1,m}) - a_{n,m}\right) \end{cases}$ in the case $\mu_0 \neq \mu_1$



Proposition (Central Limit Theorem for $\mathcal{W}_p$ on the real line with $p > 1$)

del Barrio et al., 2019b *Assume that $F, G \in \mathcal{F}_{2p}$ and G^{-1} is continuous on $(0,1)$ and $p > 1$.*

+ Technical assumptions

(i) If $X_1, \ldots, X_n$ are i.i.d. F and F_n is the empirical d.f. based on the X_i's

$$\sqrt{n}(\mathcal{W}_p^p(F_n, G) - \mathcal{W}_p^p(F, G)) \to_w N(0, \sigma_p^2(F, G)).$$

(ii) If, furthermore, F^{-1} is continuous, $Y_1, \ldots, Y_m$ are i.i.d. G, independent of the X_i's, G_m is the empirical d.f. based on the Y_j's and $\frac{n}{n+m} \to \lambda \in (0,1)$ then

$$\sqrt{\tfrac{nm}{n+m}}(\mathcal{W}_p^p(F_n, G_m) - \mathcal{W}_p^p(F, G)) \to_w N(0, (1-\lambda)\sigma_p^2(F, G) + \lambda\sigma_p^2(G, F)).$$

Role of the centering constants:

Kantorovich duality (Villani, 2003) $\Rightarrow \mathbb{E}(\mathcal{W}_p^p(F_n, G)) \geq \mathcal{W}_p^p(F, G)$

We can replace the centering constants in CLT provided:

$$0 \leq \sqrt{n}\Big(\mathbb{E}(\mathcal{W}_p^p(F_n, G)) - \mathcal{W}_p^p(F, G)\Big) \to 0$$

$\Uparrow$

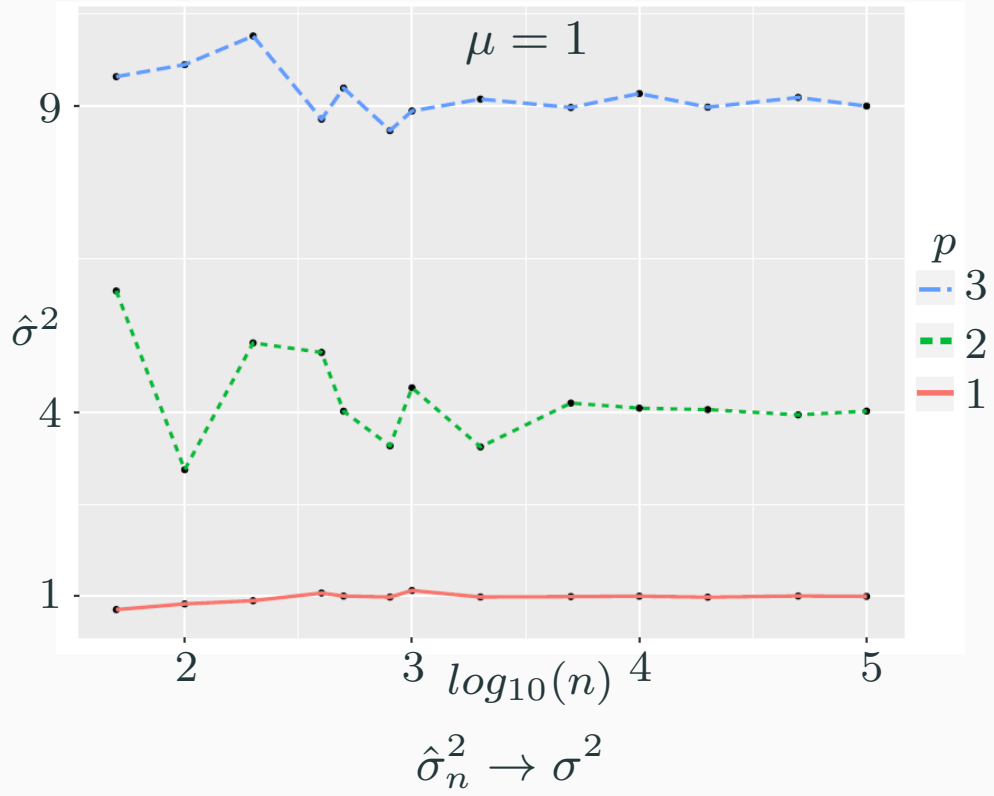
Sufficient conditions

Proposition (Consistency of variance estimates. del Barrio et al., 2019)

If $F, G \in \mathcal{F}_{2p}$, F^{-1}, G^{-1} *are continuous on* $(0,1)$ *and* $\frac{n}{n+m} \to \lambda \in (0,1)$, *then*

$$\hat{\sigma}^2_{n,m} = \frac{m}{n+m}\hat{\sigma}^2_{1,n,m} + \frac{n}{n+m}\hat{\sigma}^2_{2,n,m} \to (1-\lambda)\sigma^2_p(F,G) + \lambda\sigma^2_p(G,F) \text{ a.s.}$$

Example$(n = m)$: $F \sim N(0,1), G \sim N(\mu,1) \Rightarrow \sigma^2_p(F,G) = \sigma^2_p(G,F) = p^2\mu^{2p-2}$



n	$p=1$	$p=2$	$p=3$
50	0.03076	2.28517	79.70453
100	0.01434	1.25248	36.57057
200	0.00634	0.74908	15.10497
400	0.00290	0.32747	6.15403
500	0.00237	0.21351	5.50914
800	0.00148	0.18638	3.20970
1,000	0.00112	0.13431	2.59728
2,000	0.00054	0.0711	1.41032
5,000	0.00021	0.0304	0.52269
10,000	0.00011	0.0145	0.24127
σ^2	1	4	9

$$MSE = \frac{1}{N}\sum_{j=1}^{N}\left|\hat{\sigma}^2_j - \sigma^2\right|^2, N = 1000$$

Finite performance of the test: Normal location model ($n = m$)

$$F \sim N(0,1), G \sim N(\mu, 1)$$

$$H_0 : \mathcal{W}_p(F,G) \geq \Delta_0,$$

vs

$$H_a : \mathcal{W}_p(F,G) < \Delta_0$$

Asymptotic level $\alpha = 0.05$

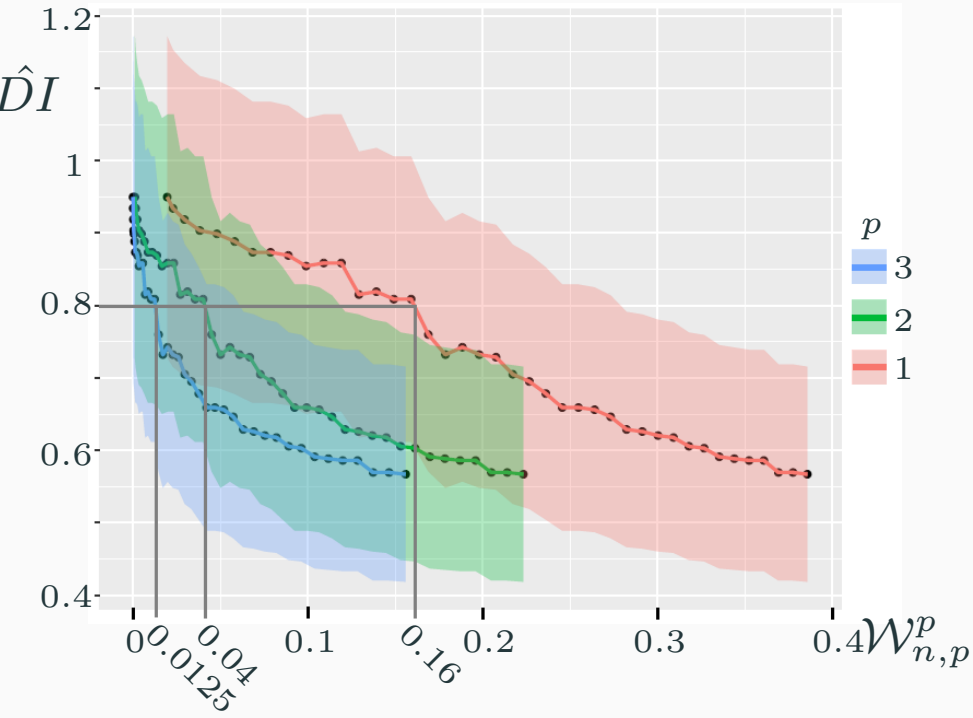
Simulations:

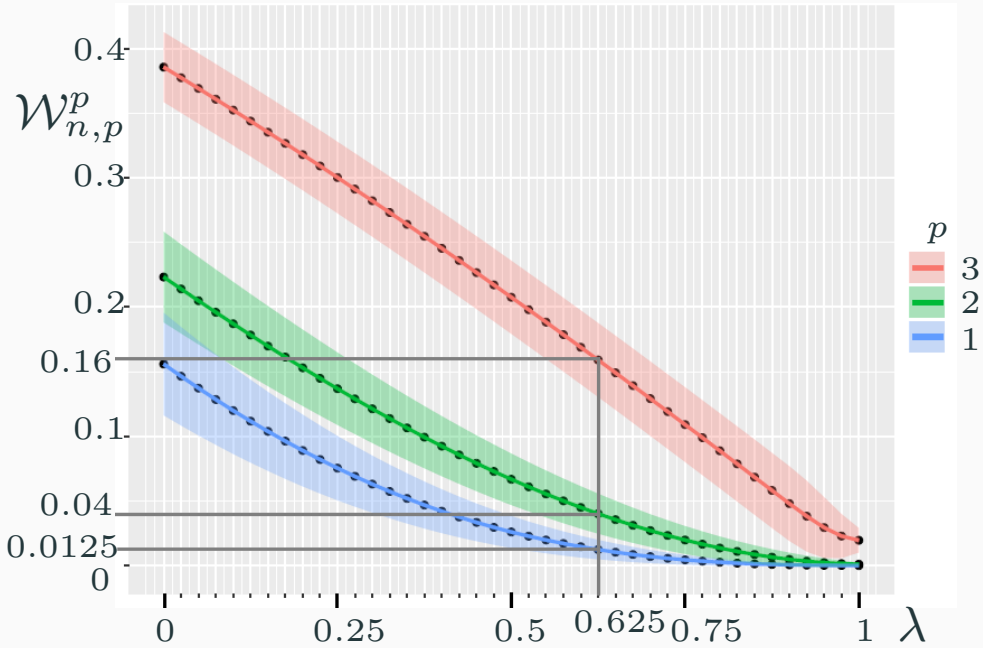
$$\Delta_0 = \mathcal{W}_p(N(0,1), N(1,1)) = 1$$

$\mu = 1 \rightarrow$ Level of the test

$\mu = 0.9, 0.7, 0.5 \rightarrow$ Power of the test

p	n	μ=1	μ=0.9	μ=0.7	μ=0.5
1	50	0.062	0.146	0.481	0.825
	100	0.055	0.193	0.698	0.974
	200	0.053	0.275	0.918	1
	400	0.051	0.413	0.995	1
	500	0.051	0.481	0.999	1
	800	0.052	0.64	1	1
	1,000	0.054	0.728	1	1
	2,000	0.047	0.937	1	1
2	50	0.074	0.167	0.513	0.839
	100	0.063	0.198	0.717	0.979
	200	0.059	0.272	0.927	1
	400	0.055	0.422	0.995	1
	500	0.05	0.484	0.999	1
	800	0.053	0.651	1	1
	1,000	0.053	0.736	1	1
	2,000	0.051	0.935	1	1
3	50	0.071	0.154	0.515	0.822
	100	0.066	0.206	0.715	0.973
	200	0.057	0.266	0.925	1
	400	0.052	0.422	0.992	1
	500	0.057	0.497	0.997	1
	800	0.053	0.652	1	1
	1,000	0.053	0.733	1	1
	2,000	0.051	0.937	1	1





DI and BER depend on a given classifier...

while $\mathcal{W}_p$ is a global condition on the fairness of the dataset

basque center for applied mathematics

EXCELENCIA SEVERO OCHOA

EUSKO JAURLARITZA GOBIERNO VASCO

Central Limit Theorem and bootstrap procedure for Wasserstein's variations with application to structural relationships between distributions

DEL BARRIO ET AL. (2019A)

Journal of Multivariate Analysis (2019)

Measure structural relationships between data

$\Downarrow$

Estimation of probability measures observed with deformations

- Registration of warped distributions (Bolstad et al., 2003), (Gallón et al., 2013)
- Comparison of distributions using optimal transport methodologies (Agueh and Carlier, 2011), (Chernozhukov et al., 2017)

We observe $X_{ij} = g_j(\varepsilon_{ij}) \sim \mu_j \quad 1 \leqslant j \leqslant J$

ε_{ij} $i.i.d.$ $\sim \mu$ unknown

$g_j \in \mathcal{G}_j$ class of warping functions

Goal: estimation of g_j $\dashrightarrow$ alignment of the estimated $\mu_j(g_j^{-1})'$s

- Extension of the functional deformation models (Gamboa et al., 2007), (Collier and Dalalyan, 2015)
- Estimation of deformations in a parametric class (Agulló-Antolín et al., 2015)
- Statistical inference on distribution deformation models (Freitag and Munk, 2005)

Deformation model for distributions

Consider $\mathcal{G} = \mathcal{G}_1 \times \cdots \times \mathcal{G}_J$, with $\mathcal{G}_j$ family of invertible functions

There exists $(\varphi_1^*, \ldots \varphi_J^*) \in \mathcal{G}$ and i.i.d. $(\varepsilon_{i,j})_{\substack{1 \le i \le n \\ 1 \le j \le J}}$ such that

$$X_{i,j} = \left(\varphi_j^*\right)^{-1}(\varepsilon_{i,j}), \quad 1 \le j \le J$$

$$\Updownarrow \qquad (*)$$

There exists $(\varphi_1^*, \ldots \varphi_J^*) \in \mathcal{G}$ such that $\mu_1(\varphi_1^*) = \cdots = \mu_J(\varphi_J^*)$,
where $\varphi_j(X_{i,j}) \sim \mu_j(\varphi_j),\ 1 \le j \le J, 1 \le i \le n$

Aligning probability distributions

$\Rightarrow \mu_j(\varphi_j)$'s should be close w.r.t their **Wasserstein variation**

$$V_p\left(\mu_1(\varphi_1),\dots,\mu_J(\varphi_J)\right) = \inf_{\eta\in\mathcal{F}_p(\mathbb{R}^d)} \left(\frac{1}{J}\sum_{j=1}^{J} \mathcal{W}_p^p(\mu_j(\varphi_j),\eta)\right)^{1/p}$$

Minimal alignment cost:

$$A_p(\mathcal{G}) := \inf_{\varphi\in\mathcal{G}} V_p^p\left(\mu_1(\varphi_1),\dots,\mu_J(\varphi_J)\right)$$

If $\mu_1(\varphi_1),\dots,\mu_J(\varphi_J) \in \mathcal{F}_p(\mathcal{R}^d)$,

$(*) \Leftrightarrow A_p(\mathcal{G}) = 0$ $\Longrightarrow$

Assesing fit to deformation models:

$$H_0 : A_p(\mathcal{G}) \geq \Delta_0$$

vs.

$$H_a : A_p(\mathcal{G}) < \Delta_0,$$

with $\Delta_0 > 0$ a fixed threshold.

Assesing fit to non-parametric deformation models ($d = 1$ and $p = 2$)

Empirical minimal aligment cost: $A_n(\mathcal{G}) := \inf_{(\varphi_1,\dots,\varphi_J)\in\mathcal{G}} U_n(\varphi)$

$$U(\varphi) := V_2^2(\mu_1(\varphi_1),\dots,\mu_J(\varphi_J)) \to U_n(\varphi) = V_2^2(\mu_{n,1}(\varphi_1),\dots,\mu_{n,J}(\varphi_J)),$$

where $\mu_{n,j}(\varphi_j)$ empirical measure on $\varphi_j(X_{i,j}),\dots,\varphi_j(X_{n,j})$.

Theorem (del Barrio et al., 2019a)

Assume that $(B_j)_{1\le j\le J}$ are independent Brownian bridges. For $\varphi\in\mathcal{G}$, set $C(\varphi) = \frac{1}{J}\sum_{j=1}^J c_j(\varphi)$, where $c_j(\varphi) = 2\int_0^1 \varphi_j'\circ F_j^{-1}(\varphi_j\circ F_j^{-1} - F_B^{-1}(\varphi))\frac{B_j}{f_j\circ F_j^{-1}}$. Then, under technical assumptions, C is a centered Gaussian process on $\mathcal{G}$ with trajectories a.s. continuous w.r.t. $\|\cdot\|_{\mathcal{H}}$. Furthermore,

$$\sqrt{n}(A_n(\mathcal{G}) - A(\mathcal{G})) \rightharpoonup \min_{\varphi\in\Gamma} C(\varphi),$$

where $\Gamma = \left\{\varphi\in\mathcal{G}:\ U(\varphi) = \inf_{\phi\in\mathcal{G}} U(\phi)\right\}$ is an nonempty compact subset of $\mathcal{G}$.

Statistical evidence against the deformation model

$$H_0 : A_p(\mathcal{G}) = 0 \quad \text{vs.} \quad H_a : A_p(\mathcal{G}) > 0$$

Under the deformation model: $\varphi_j \circ F_j^{-1} = F_B^{-1}(\varphi)$, for each $\varphi_j \in \Gamma$

$$\Rightarrow \sqrt{n} A_n(\mathcal{G}) \rightharpoonup 0 \quad \dashrightarrow \textbf{Nondegenerate limit law for } A_n(\mathcal{G})$$

Simulations

Family of scale-location deformations:

$$X_{i,j} = \mu_j^* + \sigma_j^* \varepsilon_{i,j}, 1 \le i \le n, 1 \le j \le J$$

- **1.** Estimate the **frequency of rejection** under H_0 (deformation model holds)

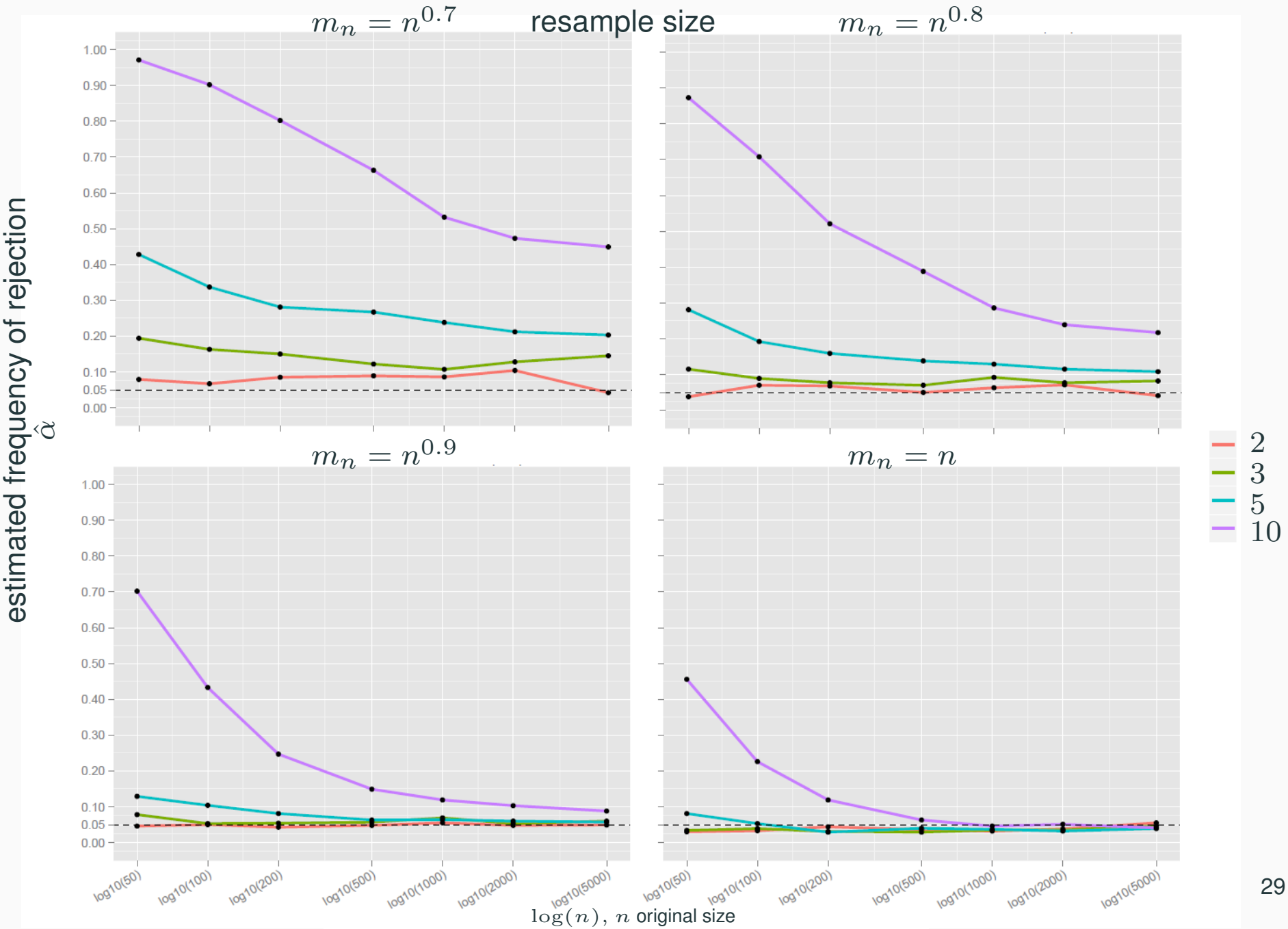
$$\varepsilon_{i,j} \text{ i.i.d. } \mathcal{N}(0,1), 1 \le i \le n, 1 \le j \le J$$

2. Estimate the **power of the test** under H_a

$$\varepsilon_{i,j} \text{ i.i.d. } \mathcal{N}(0,1), 1 \le i \le n, 1 \le j \le J-1,$$

$$\varepsilon_{i,J} \text{ i.i.d. } \gamma$$

Construction of an $\alpha-$level test



Statistical evidence against the deformation model

$$H_0 : A_p(\mathcal{G}) = 0 \quad \text{vs.} \quad H_a : A_p(\mathcal{G}) > 0$$

Under the deformation model: $\varphi_j \circ F_j^{-1} = F_B^{-1}(\varphi)$, for each $\varphi_j \in \Gamma$

$$\Rightarrow \sqrt{n} A_n(\mathcal{G}) \rightharpoonup 0 \quad \dashrightarrow \textbf{Nondegenerate limit law for } A_n(\mathcal{G})$$

Simulations

Family of scale-location deformations:

$$X_{i,j} = \mu_j^* + \sigma_j^* \varepsilon_{i,j}, 1 \le i \le n, 1 \le j \le J$$

1. Estimate the **frequency of rejection** under H_0 (deformation model holds)

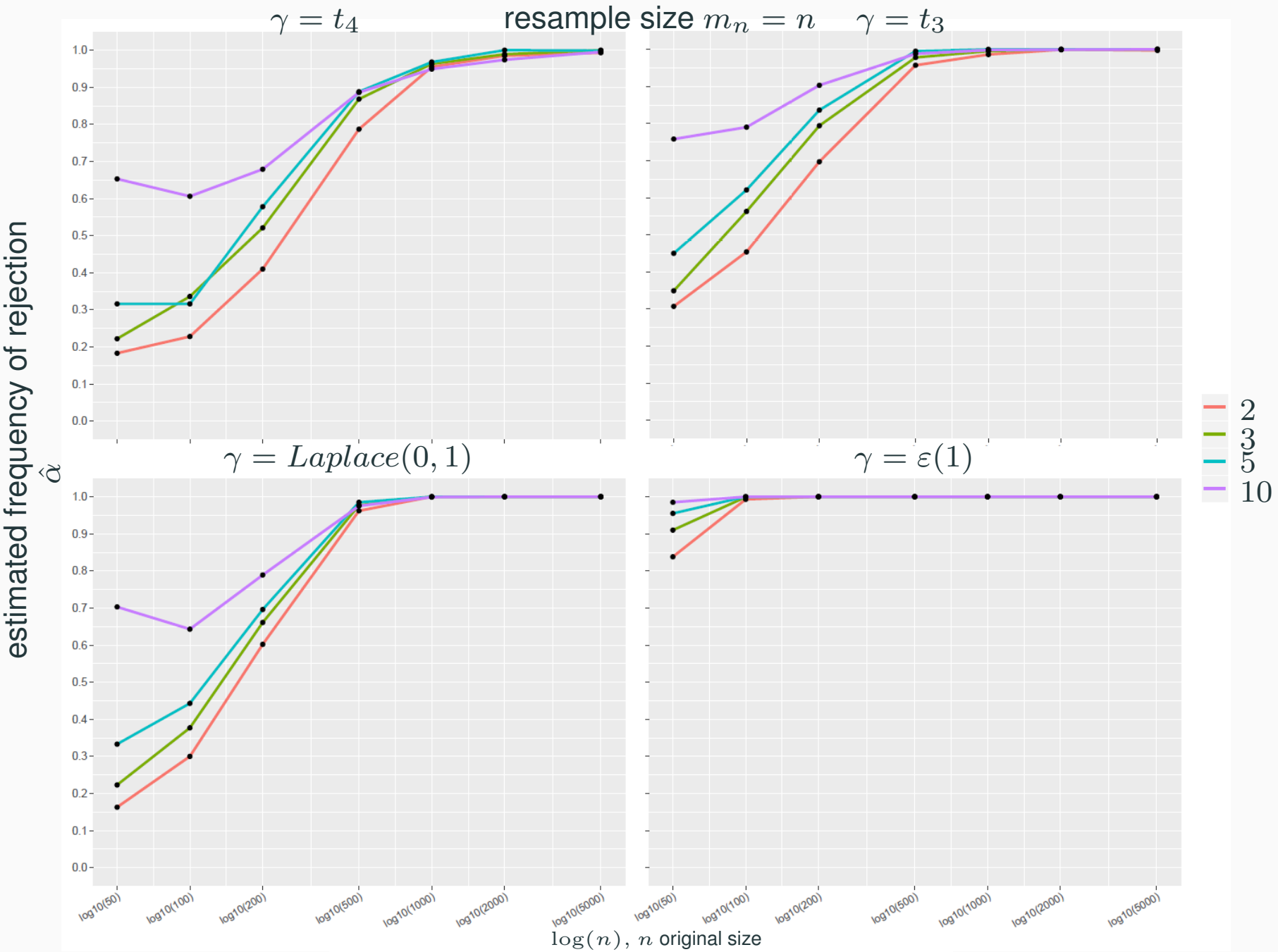
$$\varepsilon_{i,j} \text{ i.i.d. } \mathcal{N}(0,1), 1 \le i \le n, 1 \le j \le J$$

- **2.** Estimate the **power of the test** under H_a

$$\varepsilon_{i,j} \text{ i.i.d. } \mathcal{N}(0,1), 1 \le i \le n, 1 \le j \le J-1,$$

$$\varepsilon_{i,J} \text{ i.i.d. } \gamma$$

Power of the test procedure



Deformation model for fair learning

- Consider observations $(X_1, S_1, Y_1), \ldots, (X_n, S_n, Y_n)$ i.i.d. from the r.v. (X, S, Y), where $Y \in \mathbb{R}$, $X \in \mathbb{R}^d$, $d \geq 1$, and $S \in \mathcal{S} = \{1, \ldots, k\}$
- For each $s \in \mathcal{S}$ and $i \in \{1, \ldots, n\}$, $X_{s,i} := X_i$ the observations of the usable attribute such that $S_i = s$ and by n_s the size of each protected group
- the bias in the observed sample comes from the influence of the sensitive variable S, in the sense that the conditional distributions $\mu_s := \mathcal{L}(X|S = s), s \in \mathcal{S}$, are different.

There exist some warping functions $(\varphi_0^*, \ldots, \varphi_k^*) \in \mathcal{G} = \mathcal{G}_0 \times \cdots \times \mathcal{G}_k$, and some random variables $\eta_{s,1}, \ldots, \eta_{s,n_s}$, independent and equally distributed from a common but unknown distribution ν and such that, for every $s \in \mathcal{S}$,

$$X_{s,i} = (\varphi_s^*)^{-1}(\eta_{s,i}),\ 1 \leq i \leq n_s.$$

Repairing the data could be addressed through a deformation model

i) φ_S^* will be the optimal transport map pushing μ_S towards their Wasserstein barycenter μ_B, and

ii) $\tilde{X}_i := \eta_{S,i} = \varphi_S^*(X_i), i \in \{1, \ldots, n\}$, will be the repaired version of the data that we are looking for.

Publications

- E. del Barrio, P. Gordaliza, H. Lescornel and J.-M. Loubes. “Central Limit Theorem and bootstrap procedure for Wasserstein’s variations with application to structural relationships between distributions.” *JMVA*, 2019.
- P. Gordaliza, E. del Barrio, F. Gamboa and J.-M. Loubes. “Obtaining Fairness using Optimal Transport Theory.” *36th International Conference on Machine Learning,* 2019.
- E. del Barrio, P. Gordaliza and J.-M. Loubes. “A CLT for Lp transportation cost on the real line with application to fairness assessment in machine learning.” *Information and Inference: a journal of the IMA*, 2019.
- P. Besse, E. del Barrio, P. Gordaliza, J.-M. Loubes. and L. Risser. “A survey of bias in Machine Learning through the prism of Statistical Parity.” *The American Statistician*, 2021.

Working papers

- E. del Barrio, P. Gordaliza and JM. Loubes. “Review of Mathematical Frameworks for fairness in Machine Learning.”
- E. del Barrio, P. Gordaliza , JM. Loubes and A. Pérez-Suay. “Price for equalized odds in regression.”
- P. Gordaliza and H. Inouzhe. “Optimal Trimmed Matching and Trimmed Group Fairness.”
- A. Barrainkua, JA. Lozano, N. Quadrianto, P. Gordaliza and O. Thomas. “Empirical Analysis of the Stability of Fairness-Enhancing Methods Against Distribution Shift.”

elise
European Network of AI Excellence Centres

bcam basque center for applied mathematics

EXCELENCIA SEVERO OCHOA

EUSKO JAURLARITZA GOBIERNO VASCO