Analyzing the structural relationship between distributions using optimal transport

Paula Gordaliza Pastor

Joint work with
E. del Barrio  F. Gamboa  JM. Loubes
P. Besse  L. Risser
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

<table>
<thead>
<tr>
<th></th>
<th>WHITE</th>
<th>AFRICAN AMERICAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labeled Higher Risk, But Didn’t Re-Offend</td>
<td>23.5%</td>
<td>44.9%</td>
</tr>
<tr>
<td>Labeled Lower Risk, Yet Did Re-Offend</td>
<td>47.7%</td>
<td>28.0%</td>
</tr>
</tbody>
</table>

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

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

TABLE II: Top k results on [www.xing.com](http://www.xing.com) (Jan 2017) for the job search 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

- 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 \((\Omega \subset \mathbb{R}^d, \mathcal{B}, \mathbb{P})\), \(\mathcal{B}\) Borel \(\sigma\)-algebra of subsets of \(\mathbb{R}^d\) and \(d \geq 1\)

<table>
<thead>
<tr>
<th>Protected attribute</th>
<th>Visible attributes</th>
<th>Target</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>(S \in S)</td>
<td>(X \in \mathcal{X} \subset \mathbb{R}^d)</td>
<td>(Y \in \mathbb{R}^d)</td>
<td>(\hat{Y} = f(X, S), \ f \in \mathcal{F})</td>
</tr>
</tbody>
</table>

**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 \( \Omega \subset \mathbb{R}^d, \mathcal{B}, \mathbb{P} \), \( \mathcal{B} \) Borel \( \sigma \)-algebra of subsets of \( \mathbb{R}^d \) and \( d \geq 1 \)

<table>
<thead>
<tr>
<th>Protected attribute</th>
<th>Visible attributes</th>
<th>Target</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>( S \in S = {0, 1} )</td>
<td>( X \in \mathcal{X} \subset \mathbb{R}^d )</td>
<td>( Y \in {0, 1} )</td>
<td>( \hat{Y} = g(X, S), \ g \in \mathcal{G} )</td>
</tr>
<tr>
<td>0 unfavored</td>
<td>( 0 ) failure</td>
<td>( g : \mathbb{R}^d \rightarrow {0, 1} )</td>
<td></td>
</tr>
<tr>
<td>1 favored</td>
<td>( 1 ) success</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**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):** \( \hat{Y} \perp S \)

\[
P(\hat{Y} = 1 \mid S = 0) = P(\hat{Y} = 1 \mid S = 1)
\]

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

\[
P(\hat{Y} = i \mid Y = i, S = 0) = P(\hat{Y} = i \mid Y = i, S = 1), \ i = 0, 1
\]
The \textbf{Disparate Impact} of the classifier $g \in \mathcal{G}$, with respect to $(X, S)$ is defined as

\[
DI(g, X, S) = \frac{\Pr(g(X, S) = 1 \mid S = 0)}{\Pr(g(X, S) = 1 \mid S = 1)} \in (0, 1]
\]

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

\textbf{Definition}

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

- $\tau_0 = 4/5 \rightarrow 80\%$ rule (1971, State of California Fair Employment Commission)
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 $\pi$ 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 $\pi$ 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 \( \mu_1, \ldots, \mu_J \in \mathcal{F}_p (\mathbb{R}^d) \) w.r.t. weights \( \omega_1, \ldots, \omega_J > 0 \), is defined as

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

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 \( \mu_B \) of the \( \mu_j \)'s (Agueh and Carlier, 2011)

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

- Empirical versions (Boissard et al., 2015), (Le Gouic and Loubes, 2017)
Find \( \tilde{X} = T_S(X) \) such that \( \mathcal{L}(T_0(X) \mid S = 0) = \mathcal{L}(T_1(X) \mid S = 1) \)

\[
\mathcal{L}(g(\tilde{X}) \mid S = 0) = \mathcal{L}(g(\tilde{X}) \mid S = 1), \text{ for all } g \in \mathcal{G}
\]

**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 \( \mu_0, \mu_1 \) to \( \nu \)?
Upper bound for the price for fairness \cite{gordaliza2019price} 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 \( K = \max\{K_0, K_1\} \),

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

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 \( \mu_B \) with weights \( \pi_0 = P(S = 0) \) and \( \pi_1 = P(S = 1) \)

\[
\mu_B \in \arg\min_{\nu \in \mathcal{P}_2} \left\{ \pi_0 \mathcal{W}_2^2(\mu_0, \nu) + \pi_1 \mathcal{W}_2^2(\mu_1, \nu) \right\}
\]

b) \( T_S \) optimal transport map carrying \( \mu_S \) towards \( \mu_B \)

\[
\mu_S \# T_S = \mu_B
\]
Partial Repair with Wasserstein barycenter

**Target variable**
\[ Z \sim \mu_B \]

**Level of repair**
\[ \lambda \in [0, 1] \]

**Transformation**
\[ \mu = \mu_S \mathbb{#} T_S \]
\[ T_S^{-1}(Z) \sim \mu_S \]

**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)
\[ B \sim \mathcal{B}(\lambda), \text{ independent of } (X, S, Y) \]
\[ \tilde{\mu}_{S,\lambda} = \mathcal{L}(B T_S(X) + (1 - B)X) \]

**Unmodified variable**
\[ 0 \leftarrow \lambda \rightarrow 1 \]
\[ \tilde{\mu}_{s,0} = \mathcal{L}(X | S = s) \]

**Accuracy of** \( g(\tilde{X}) \)

**Trade-off**

**Totally repaired variable**
\[ \tilde{\mu}_{s,1} = \mathcal{L}(Z) = \mu_B \]

**Non-predictability** of \( S \)

\[ d_{TV}(P, Q) = \min_{\pi \in \Pi(P, Q)} \pi(x \neq y) \]
Partial Repair with Wasserstein barycenter

<table>
<thead>
<tr>
<th>Target variable</th>
<th>Level of repair</th>
<th>Transformation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Z \sim \mu_B$</td>
<td>$\lambda \in [0, 1]$</td>
<td>$\mu = \mu_S # T_S$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$R_S := T_S^{-1}(Z) \sim \mu_S$</td>
</tr>
</tbody>
</table>

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 $d_{TV}$:

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

$$= \mathbb{P}(R_0(Z) \neq R_1(Z)).$$

Random repair (Gordaliza et al., 2019)

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

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

The level of repair controls $d_{TV}$:

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

$$\leq 1 - \mathbb{P}(B = 1) = 1 - \lambda$$

The new risk is a mixture of the two errors:

$$R(g, \tilde{X}_\lambda) = (1 - \lambda)\mathbb{P}(g(X) \neq Y) + \lambda\mathbb{P}(g(T_S(X)) \neq Y)$$
Application to the Adult Data Set (size 29.825)

\[ Y = \begin{cases} 
1 & \text{income exceeds } 50,000/\text{year} \\
0 & \text{otherwise} 
\end{cases} \]

\[ X \]
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...?

\[ S = \begin{cases} 
0 & \text{female} \\
1 & \text{male} 
\end{cases} \]

Experiment:

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

<table>
<thead>
<tr>
<th>Statistical Model</th>
<th>Error</th>
<th>( \hat{DI} )</th>
<th>CI 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logit</td>
<td>0.2064</td>
<td>0.496</td>
<td>(0.437, 0.555)</td>
</tr>
<tr>
<td>Random Forests</td>
<td>0.168</td>
<td>0.484</td>
<td>(0.429, 0.54)</td>
</tr>
</tbody>
</table>

3. Predict \( g(X) \) Random repair
4. Repair procedure \( \rightarrow \tilde{X} \rightarrow \) vs.
5. Predict \( g(\tilde{X}) \) Geometric repair

(Feldman et al., 2015)
New criteria for statistical parity assessment

\[ \varepsilon^* := \min_{g \in G} \text{BER}(g, X, S) = \frac{1}{2} (1 - d_{TV}(\mu_0, \mu_1)), \mu_s = \mathcal{L}(X \mid S = s) \]

(Gordaliza et al., 2019)

Rejection of \( H_0 : \rho(\mu_0, \mu_1) \geq \Delta_0 \Rightarrow \text{Statistical certification that} \ \mu_0 \approx \mu_1 \)

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

Goal: CLT

\[ \left\{ \begin{array}{l}
  r_n \left( \mathcal{W}_p^p(\mu_{0,n}, \mu_1) - a_n \right) \\
  r_{n,m} \left( \mathcal{W}_p^p(\mu_{0,n}, \mu_{1,m}) - a_{n,m} \right)
\end{array} \right. \]

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

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

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:

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

↑

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 \((n = m)\): \( F \sim N(0, 1), G \sim N(\mu, 1) \Rightarrow \sigma^2_p(F, G) = \sigma^2_p(G, F) = p^2 \mu^{2p-2} \)
Finite performance of the test: Normal location model \((n = m)\)

\[ F \sim N(0, 1), \; G \sim N(\mu, 1) \]

\[ H_0 : W_p(F, G) \geq \Delta_0, \]

\[ H_a : W_p(F, G) < \Delta_0 \]

Asymptotic level \(\alpha = 0.05\)

### Simulations:

\[ \Delta_0 = W_p(N(0, 1), N(1, 1)) = 1 \]

\(\mu = 1 \rightarrow \text{Level of the test} \)

\(\mu = 0.9, 0.7, 0.5 \rightarrow \text{Power of the test} \)

<table>
<thead>
<tr>
<th>(p)</th>
<th>(n)</th>
<th>(\mu=1)</th>
<th>(\mu=0.9)</th>
<th>(\mu=0.7)</th>
<th>(\mu=0.5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50</td>
<td>0.062</td>
<td>0.146</td>
<td>0.481</td>
<td>0.825</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.055</td>
<td>0.193</td>
<td>0.698</td>
<td>0.974</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.053</td>
<td>0.275</td>
<td>0.918</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>0.051</td>
<td>0.413</td>
<td>0.995</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.051</td>
<td>0.481</td>
<td>0.999</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>800</td>
<td>0.052</td>
<td>0.64</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1,000</td>
<td>0.054</td>
<td>0.728</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2,000</td>
<td>0.047</td>
<td>0.937</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>50</td>
<td>0.074</td>
<td>0.167</td>
<td>0.513</td>
<td>0.839</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.063</td>
<td>0.198</td>
<td>0.717</td>
<td>0.979</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.059</td>
<td>0.272</td>
<td>0.927</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>0.055</td>
<td>0.422</td>
<td>0.995</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.05</td>
<td>0.484</td>
<td>0.999</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>800</td>
<td>0.053</td>
<td>0.651</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1,000</td>
<td>0.053</td>
<td>0.736</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2,000</td>
<td>0.051</td>
<td>0.935</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>50</td>
<td>0.071</td>
<td>0.154</td>
<td>0.515</td>
<td>0.822</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.066</td>
<td>0.206</td>
<td>0.715</td>
<td>0.973</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.057</td>
<td>0.266</td>
<td>0.925</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>0.052</td>
<td>0.422</td>
<td>0.992</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.057</td>
<td>0.497</td>
<td>0.997</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>800</td>
<td>0.053</td>
<td>0.652</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1,000</td>
<td>0.053</td>
<td>0.733</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2,000</td>
<td>0.051</td>
<td>0.937</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
DI and BER depend on a given classifier... 
while $\mathcal{W}_p$ is a global condition on the fairness of the dataset
Central Limit Theorem and bootstrap procedure for Wasserstein’s variations with application to structural relationships between distributions

DELBARRIO ET AL. (2019A)

Measure structural relationships between data

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

\[ X_{ij} = g_j(\varepsilon_{ij}) \sim \mu_j \quad 1 \leq j \leq J \]

\[ \varepsilon_{ij} \text{ i.i.d.} \sim \mu \text{ unknown} \]

\[ g_j \in G_j \text{ class of warping functions} \]

**Goal:** estimation of \( g_j \) \( \rightarrow \) 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. $(\varepsilon_{i,j})_{1 \leq i \leq n, 1 \leq j \leq J}$ such that

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

$(\ast)$

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

\[ \Rightarrow \mu_j(\varphi_j)'s \text{ should be close w.r.t their Wasserstein variation} \]

\[ V_p(\mu_1(\varphi_1), \ldots, \mu_J(\varphi_J)) = \inf_{\eta \in F_p(R^d)} \left( \frac{1}{J} \sum_{j=1}^{J} W_p^p(\mu_j(\varphi_j), \eta) \right)^{1/p} \]

**Minimal alignment cost:**

\[ A_p(G) := \inf_{\varphi \in G} V_p^p(\mu_1(\varphi_1), \ldots, \mu_J(\varphi_J)) \]

If \( \mu_1(\varphi_1), \ldots, \mu_J(\varphi_J) \in F_p(R^d) \),

\[ (*) \iff A_p(G) = 0 \]

**Assesing fit to deformation models:**

\[ H_0 : A_p(G) \geq \Delta_0 \]

vs.

\[ H_a : A_p(G) < \Delta_0, \]

with \( \Delta_0 > 0 \) a fixed threshold.
Assesing fit to non-parametric deformation models ($d = 1$ and $p = 2$)

**Empirical minimal aligment cost:** $A_n(G) := \inf_{(\varphi_1, \ldots, \varphi,J) \in G} U_n(\varphi)$

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

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

**Theorem (del Barrio et al., 2019a)**

Assume that $(B_j)_{1 \leq j \leq J}$ are independent Brownian bridges. For $\varphi \in \mathcal{G}$, set

$C(\varphi) = \frac{1}{J} \sum_{j=1}^{J} c_j(\varphi), \text{ where } c_j(\varphi) = 2 \int_0^1 \varphi_j' \circ F_j^{-1}(\varphi_j \circ F_j^{-1} - F_B^{-1}(\varphi)) \frac{B_j}{f_j \circ F_j^{-1}}.$

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,

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

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

Statistical evidence against the deformation model

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

Under the deformation model:

\[ \varphi_j \circ F_j^{-1} = F_{B}^{-1}(\varphi), \ \text{for each} \ \varphi_j \in \Gamma \]

\[ \Rightarrow \sqrt{n}A_n(G) \to 0 \quad \text{--- Nondegenerate limit law for} \ A_n(G) \]

Simulations

Family of scale-location deformations:

\[ X_{i,j} = \mu^*_j + \sigma^*_j \varepsilon_{i,j}, \ 1 \leq i \leq n, \ 1 \leq j \leq J \]

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

\[ \varepsilon_{i,j} \text{ i.i.d. } \mathcal{N}(0, 1), \ 1 \leq i \leq n, \ 1 \leq j \leq J \]

2. Estimate the power of the test under \( H_a \)

\[ \varepsilon_{i,j} \text{ i.i.d. } \mathcal{N}(0, 1), \ 1 \leq i \leq n, \ 1 \leq j \leq J - 1, \]
Construction of an $\alpha$-level test

Estimated frequency of rejection $\hat{\alpha}$ vs. log($n$), original size $n$

- $m_n = n^{0.7}$
  - Resample size
- $m_n = n^{0.8}$
- $m_n = n^{0.9}$
- $m_n = n$
Goodness-of-fit in semiparametric deformation models

Statistical evidence against the deformation model

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

Under the deformation model:

[\[ \varphi_j \circ F_j^{-1} = F_B^{-1}(\varphi), \text{ for each } \varphi_j \in \Gamma \]

\[ \Rightarrow \sqrt{n}A_n(G) \rightarrow 0 \quad \rightarrow \text{Nondegenerate limit law for } A_n(G) \]

--- Simulations ---

Family of scale-location deformations:

\[ X_{i,j} = \mu_j^* + \sigma_j^* \varepsilon_{i,j}, \quad 1 \leq i \leq n, \quad 1 \leq j \leq J \]

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

\[ \varepsilon_{i,j} \text{ i.i.d. } \mathcal{N}(0, 1), \quad 1 \leq i \leq n, \quad 1 \leq j \leq J \]

2. Estimate the power of the test under \( H_a \)

\[ \varepsilon_{i,j} \text{ i.i.d. } \mathcal{N}(0, 1), \quad 1 \leq i \leq n, \quad 1 \leq j \leq J - 1, \]
Power of the test procedure

\[ \hat{\alpha} = \gamma \]

- \( \gamma = t_4 \) resample size \( m_n = n \)
- \( \gamma = t_3 \)
- \( \gamma = \text{Laplace}(0, 1) \)
- \( \gamma = \varepsilon(1) \)

\[ \log(n), \ n \text{ original size} \]
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 \(S \in S = \{1, \ldots, k\}\)

- For each \(s \in S\) 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 \(\mu_s := \mathcal{L}(X \mid S = s), s \in S,\) are different.

There exist some warping functions \((\varphi_0^*, \ldots, \varphi_k^*) \in G = G_0 \times \cdots \times G_k,\) and some random variables \(\eta_{s,1}, \ldots, \eta_{s,n_s}\), independent and equally distributed from a common but unknown distribution \(\nu\) and such that, for every \(s \in S,\)

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

Repairing the data could be addressed through a deformation model

i) \(\varphi_S^*\) will be the optimal transport map pushing \(\mu_S\) towards their Wasserstein barycenter \(\mu_B\), 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


Working papers


P. Gordaliza and H. Inouzhe. “Optimal Trimmed Matching and Trimmed Group Fairness.”

Thanks for your attention!