A central limit theorem for L_p transportation cost on the real line with application to fairness assessment in machine learning

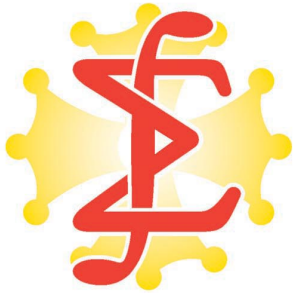
Stochastic processes and statistical machine learning II
Toulouse - Potsdam International Workshop

$13 - 15$ march 2019, Toulouse

Paula Gordaliza

IMT (Université de Toulouse) and IMUVa (Universidad de Valladolid)

Joint work with Jean-Michel Loubes (IMT) and Eustasio del Barrio (IMUVa)



INSTITUT
de MATHEMATIQUES
de TOULOUSE

Introduction

M. Sommerfeld and A. Munk (2018) : *"Transportation cost distance is an attractive tool for data analysis but statistical inference is hindered by the lack of distributional limits"*

Kantorovich formulation:

- 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

- If $c(x,y) = c_p(x,y) = \|x-y\|^p, p \geq 1$, the optimal transportation cost is

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

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

Introduction

Consider $X_1, \ldots, X_n$ i.i.d. P, P_n; and $Y_1, \ldots, Y_m$ i.i.d. Q, Q_m
Assuming that P and Q have finite p-th moment,

$$\mathcal{W}_p^p(P_n, Q) \to \mathcal{W}_p^p(P, Q), \ as\ n \to \infty, \text{ a.s.}$$
$$\mathcal{W}_p^p(P_n, Q_m) \to \mathcal{W}_p^p(P, Q), \ as\ n, m \to \infty, \text{ a.s.}$$

Distributional limit theorem?

a) case $P = Q$: goodness-of-fit problems

- $d \geq 1$
  - M. Atjai et al. (1984) and M. Talagrand and J.E. Yuckich (1993) : $P = Q$ uniform distribution on the unit hypercube
  - V. Dobrić and J.E. Yuckich (1995), N. Fournier and A. Guillin (2015): rates of convergence
- $d = 1$
  - $p = 1$: E. Del Barrio, E. Giné and C. Matrán (1999) (integrability conditions) $\mathcal{W}_1(P_n, P) = O_P(n^{-1/2})$ with $\sqrt{n}\mathcal{W}_1(P_n, P) \to_w$ non-Gaussian
  - $p = 2$: E. del Barrio, J.A. Cuesta-Albertos, C. Matrán and J.M. Rodríguez-Rodríguez (1999), E. del Barrio, E. Giné and F. Utzet (2005) (integrability + smoothness conditions on P) $\sqrt{n}\mathcal{W}_p(P_n, P) \to_w$

Introduction

Provide statistical certification that the data are not too far from a model $P = Q$

× Not rejecting the null $H_0 : P = Q$

✓ Rejection of the null $H_0 : \rho(P, Q) \geq \Delta_0$ for some distance ρ

b) case $P \neq Q$

$$H_0 : \mathcal{W}_p(P, Q) \geq \Delta_0$$

CLT : $\left.\begin{array}{c} r_n\big(\mathcal{W}_p^p(P_n, Q) - a_n\big) \\ r_{n,m}\big(\mathcal{W}_p^p(P_n, Q_m) - a_{n,m}\big) \end{array}\right\} \Rightarrow$ Computation of approximate $p - values$

- $d = 1, p = 2$: A. Munk and C. Czado (1998) $\mathcal{W}_2$ or trimmed version
- $d \geq 1, p \geq 1$:
  - M. Sommerfeld and A. Munk (2018): P, Q finitely supported
  - A. Tameling, M. Sommerfeld and A. Munk (2018): P, Q countable support
- $d \geq 1, p = 2$: E. del Barrio and J.-M. Loubes (2017): P and Q continuous, **CLT in general dimension**: if Q has a positive density in the interior of its convex support and P and Q have finite moments of order $4 + \delta$ for some $\delta > 0$ then

$$\sqrt{n}\big(\mathcal{W}_2^2(P_n, Q) - E(\mathcal{W}_2^2(P_n, Q))\big) \rightarrow_w N(0, \sigma^2(P, Q))$$

for some $\sigma^2(P, Q)$, which is not null if and only if $P \neq Q$.

+ two-sample version

Main contributions of the paper

$$d = 1\ ,\ p \geq 1$$

→ **CLT** for general cost on the real line:

$$\sqrt{n}\Big(\mathcal{W}_p^p(P_n, Q) - E(\mathcal{W}_p^p(P_n, Q))\Big) \to_w N(0, \sigma^2(P, Q))$$

→ $p > 1$: under sharp moment and smoothness assumptions

→ $p = 1$: when strict convexity of the cost function is lost, non-normal limits can occur, even in the case $P \neq Q$

\+ **two sample version**: if $\frac{n}{n+m} \to \lambda \in (0, 1)$

$$\sqrt{n}\Big(\mathcal{W}_p^p(P_n, Q_m) - E(\mathcal{W}_p^p(P_n, Q_m))\Big) \to_w N(0, (1-\lambda)\sigma^2(P, Q) + \lambda\sigma^2(Q, P))$$

→ General conditions under which $E(\mathcal{W}_p^p(P_n, Q))$ can be replaced by $\mathcal{W}_p^p(P, Q)$ as centering constant

→ Consistent estimator of the asymptotic variance in the CLT

→ Confidence interval for $\mathcal{W}_p^p(P, Q)$ of asymptotic level $1 - \alpha$

→ Consistent test $H_0 : \mathcal{W}_p(P, Q) \geq \Delta_0$ vs $H_a : \mathcal{W}_p(P, Q) < \Delta_0$

Outline

CLT for L_p transportation cost on the real line

P and Q probabilities on $\mathbb{R}$, F, G d.f.'s

$$\mathcal{W}_p^p(F,G) = \int_0^1 |F^{-1}(t) - G^{-1}(t)|^p dt, \text{ (C. Villani, 2003)}$$

Set $h_p(x) = |x|^p$, $x \in \mathbb{R}$, $p > 1$, and

$$c_p(t; F, G) := \int_{F^{-1}(\frac{1}{2})}^{F^{-1}(t)} h_p'\Big(s - G^{-1}(F(s))\Big) ds, \quad 0 < t < 1$$

$$\bar{c}_p(t; F, G) := c_p(t; F, G) - \int_0^1 c_p(s; F, G) ds, \quad 0 < t < 1$$

Lemma

If $F, G \in \mathcal{F}_{2p}$, $p > 1$, then $c_p(\cdot; F, G) \in L_2(0,1)$ and $\bar{c}_p(\cdot; F, G) \in L_2(0,1)$. Furthermore, if $F_m, G_m \in \mathcal{F}_{2p}$ satisfy $\mathcal{W}_{2p}(F_m, F) \to 0$, $\mathcal{W}_{2p}(G_m, G) \to 0$ and G^{-1} is continuous on $(0,1)$ then $\bar{c}_p(\cdot; F_m, G_m) \to \bar{c}_p(\cdot; F, G)$ in $L_2(0,1)$ as $m \to \infty$.

CLT for L_p transportation cost on the real line

$$\sigma_p^2(F,G) = \int_0^1 \bar{c}_p^2(t;F,G)dt$$

- $F, G \in \mathcal{F}_{2p} \Rightarrow \sigma_p^2(F,G) < \infty$
- $F = G \Rightarrow \sigma_p^2(F,G) = 0$
- $F \neq G \Rightarrow G^{-1} \circ F \neq Id$ on a set of positive measure ($G^{-1} \circ F$ = o.t.m. $F \to G$) and $\sigma_p^2(F,G) > 0$ if F is not a Dirac measure
- $\sigma_p^2(F,G)$ is not symmetric in F and G

Theorem (Central Limit Theorem for $\mathcal{W}_p$ with $p > 1$)

Assume that $F, G \in \mathcal{F}_{2p}$ and G^{-1} is continuous on $(0,1)$ and $p > 1$. Then

(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

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

(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

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

The assumptions in the CLT are sharp

1. $\mathcal{W}_p^p(F_n, G) = \int_0^1 |F_n^{-1}(t) - G^{-1}(t)|^p dt = \sum_{i=1}^n \int_{\frac{i-1}{n}}^{\frac{i}{n}} |X_{(i)} - G^{-1}(t)|^p \Rightarrow G \in \mathcal{F}_p$
2. F satisfies (i) for every $G \in \mathcal{F}_p$
   G Dirac's measure on 0, $\mathcal{W}_p^p(F_n, G) = \frac{1}{n}\sum_{i=1}^n |X_i|^p \Rightarrow F \in \mathcal{F}_{2p}$
3. $\sigma_p^2(F, G) < \infty$ for all $F \in \mathcal{F}_{2p} \Leftrightarrow G \in \mathcal{F}_{2p}$

$\Longrightarrow$ $F, G \in \mathcal{F}_{2p}$ minimal requirement for *(i)* to hold

- E. del Barrio and J.-M. Loubes (2017): $p = 2 \longrightarrow F, G \in \mathcal{F}_{4+\delta}, \delta > 0$

4. Continuity of G^{-1}
   - S. Bobkov and M. Ledoux (2014) : $F = G \longrightarrow$ absolute continuity of F^{-1} is a necessary condition for $E(\mathcal{W}_p(F_n, F)) = O(\frac{1}{\sqrt{n}})$
   - E. del Barrio and J.-M. Loubes (2017) $\longrightarrow G$ is supported in a (possibly unbounded) interval and G^{-1} is differentiable in the interior of that interval
   - M. Sommerfeld and A. Munk (2018): finitely supported probabilities on $\mathbb{R}$ $\longrightarrow$ nonnormal limiting distributions

Role of the centering constants in the CLT

Kantorovich duality: (C. Villani, 2003)

$$\mathcal{W}_p^p(F,G) = \sup_{(\varphi,\psi)\in\Phi_p} \int \varphi dF + \int \psi dG,$$

Φ_p set of pairs of integrable functions (with respect to F and G, respectively) satisfying $\varphi(x) + \psi(y) \leq |x-y|^p$

$$\begin{aligned} E(\mathcal{W}_p^p(F_n,G)) &\geq \sup_{(\varphi,\psi)\in\Phi_p} E\Big(\int \varphi dF_n\Big) + \int \psi dG \\ &= \sup_{(\varphi,\psi)\in\Phi_p} \int \varphi dF + \int \psi dG = \mathcal{W}_p^p(F,G) \end{aligned}$$

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

$\Rightarrow$ we can replace the centering constants in CLT:

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

Sufficient conditions for $\sqrt{n}\big(E(\mathcal{W}_p^p(F_n, G)) - \mathcal{W}_p^p(F, G)\big) \to 0$ with $p \geq 2$

F is twice differentiable

f nonvanishing density in the interior of $\text{supp}(F) = \text{cl}\{x : F(x) \notin \{0, 1\}\}$

I) $\sup_{t\in(0,1)} \dfrac{t(1-t)|f'(F^{-1}(t))|}{f^2(F^{-1}(t))} < \infty$

II) for some $s \in (\frac{p}{4}, \frac{p}{2}), \quad n^s E\mathcal{W}_p^p(F_n, F) \to 0 \quad$ as $n \to \infty$,

III) $\dfrac{1}{\sqrt{n}} \displaystyle\int_{\frac{1}{n}}^{1-\frac{1}{n}} \frac{(t(1-t))^{1/2}}{f^2(F^{-1}(t))} dt \to 0,$

IV) $\displaystyle\int_0^1 \int_0^1 \frac{(s \wedge t - st)^2}{f^2(F^{-1}(s))f^2(F^{-1}(t))} dsdt < \infty.$

V) $\displaystyle\int_0^1 \frac{(t(1-t))^{p/2}}{f^p(F^{-1}(t))} dt < \infty \quad \Rightarrow \quad$ II), III), IV)

Proposition

Assume $p \geq 2$. Under the assumptions of the CLT,

(i) if F satisfies I) to IV) then $\sqrt{n}(\mathcal{W}_p^p(F_n, G) - \mathcal{W}_p^p(F, G)) \to_w N(0, \sigma_p^2(F, G))$.

(ii) if, furthermore, G satisfies I) to IV) and $\frac{n}{n+m} \to \lambda \in (0, 1)$ then

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

Sufficient conditions for $\sqrt{n}\big(E(\mathcal{W}_p^p(F_n, G)) - \mathcal{W}_p^p(F, G)\big) \to 0$ with $p \geq 2$

F is twice differentiable

f nonvanishing density in the interior of $\text{supp}(F) = \text{cl}\{x : F(x) \notin \{0, 1\}\}$

I) $\sup_{t\in(0,1)} \dfrac{t(1-t)|f'(F^{-1}(t))|}{f^2(F^{-1}(t))} < \infty$

II) for some $s \in (\frac{p}{4}, \frac{p}{2}), \quad n^s E\mathcal{W}_p^p(F_n, F) \to 0$ as $n \to \infty$,

III) $\dfrac{1}{\sqrt{n}} \displaystyle\int_{\frac{1}{n}}^{1-\frac{1}{n}} \frac{(t(1-t))^{1/2}}{f^2(F^{-1}(t))} dt \to 0,$

IV) $\displaystyle\int_0^1 \int_0^1 \frac{(s \wedge t - st)^2}{f^2(F^{-1}(s))f^2(F^{-1}(t))} dsdt < \infty.$

V) $\displaystyle\int_0^1 \frac{(t(1-t))^{p/2}}{f^p(F^{-1}(t))} dt < \infty \quad \Rightarrow \quad$ II), III), IV)

Proposition

Assume $p \geq 2$. Under the assumptions of the CLT,

(i) if F satisfies I) to IV) then $\sqrt{n}(\mathcal{W}_p^p(F_n, G) - \mathcal{W}_p^p(F, G)) \to_w N(0, \sigma_p^2(F, G))$.

(ii) if, furthermore, G satisfies I) to IV) and $\frac{n}{n+m} \to \lambda \in (0, 1)$ then

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

Statistical application of the CLT: two sample case

P and Q probabilities on $\mathbb{R}$

$X_1, \dots, X_n$ i.i.d. P, F, F_n; and $Y_1, \dots, Y_m$ i.i.d. Q, G, G_m, independent of the X_i's

Recall $\begin{cases} h_p(x) = |x|^p, & x \in \mathbb{R}, p > 1 \\ c_p(t; F, G) = \int_{F^{-1}(\frac{1}{2})}^{F^{-1}(t)} h_p'\big(s - G^{-1}(F(s))\big) ds, & 0 < t < 1 \end{cases}$

Define:

$$\begin{cases} d_{i,n,m}(X,Y) := \sum_{j=2}^{i} \left[\left| X_{(j)} - G_m^{-1}(\frac{j-1}{n}) \right|^p - \left| X_{(j-1)} - G_m^{-1}(\frac{j-1}{n}) \right|^p \right], & i = 2, \dots, n \\ d_{1,n,m}(X,Y) := 0 \end{cases}$$

$$\Rightarrow \hat{\sigma}^2_{1,n,m} = \frac{1}{n} \sum_{i=1}^{n} d^2_{i,n,m}(X,Y) - \left(\frac{1}{n} \sum_{i=1}^{n} d_{i,n,m}(X,Y) \right)^2$$

$\hat{\sigma}^2_{2,n,m}$ similarly exchanging the roles of the X_i's and the Y_j's

Proposition (Consistency of variance estimation)

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

$$\hat{\sigma}^2_{n,m} = \frac{m}{n+m} \hat{\sigma}^2_{1,n,m} + \frac{n}{n+m} \hat{\sigma}^2_{2,n,m} \to (1-\lambda)\sigma_p^2(F,G) + \lambda \sigma_p^2(G,F)$$

almost surely.

Statistical application of the CLT

If, additionally, $F \neq G$ and F (or G) is not a Dirac measure then

$$\sqrt{\tfrac{nm}{n+m}}\frac{(\mathcal{W}_p^p(F_n, G_m) - \mathcal{W}_p^p(F, G))}{\hat{\sigma}_{n,m}} \to_w N(0,1)$$

$\to$ Confidence interval for $\mathcal{W}_p^p(F, G)$ with asymptotic confidence level $1 - \alpha$

$$\left[\mathcal{W}_p^p(F_n, G_m) \pm \sqrt{\tfrac{n+m}{nm}}\hat{\sigma}_{n,m}\Phi^{-1}(1 - \tfrac{\alpha}{2})\right]$$

$\to$ Testing problem with asymptotic level α

$$H_0 : \mathcal{W}_p(F, G) \geq \Delta_0, \quad \text{vs} \quad H_a : \mathcal{W}_p(F, G) < \Delta_0,$$

where Δ_0 is some threshold

$\Rightarrow$ Rejection of the null if

$$\mathcal{W}_p^p(F_n, G_m) < \Delta_0^p - \sqrt{\tfrac{n+m}{nm}}\hat{\sigma}_{n,m}\Phi^{-1}(1 - \alpha)$$

Normal model: variance estimates

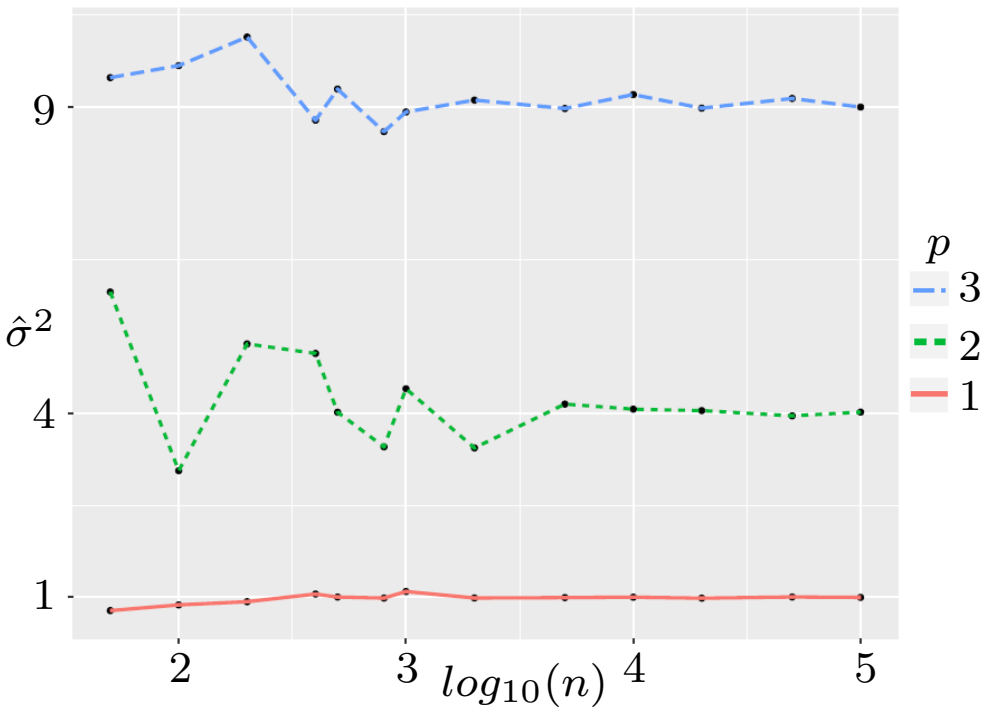
$$F \sim N(0,1), G \sim N(\mu,1)$$

$$\sigma_p^2(F,G) = \sigma_p^2(G,F) = p^2\mu^{2p-2}$$

Example: $n = m, \quad \mu = 1$

$$\hat{\sigma}_n^2 \to \sigma^2$$

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



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

Normal location model: finite performance of the test

(i) $1,000$ data sets: $P = N(0,1), Q = N(\mu,1)$ with $\Delta_0 = \mathcal{W}_p(N(0,1), N(1,1)) = 1$

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

Normal location-scale model: finite performance of the test

(ii) $1,000$ data sets: $P = N(0,1), Q = N(\mu, \lambda)$ with $\Delta_0 = \mathcal{W}_p(N(0,1), N(1,2))$

p	n	$\mu = 1$ $\lambda = 2$	$\mu = 1$ $\lambda = 1.5$	$\mu = 0$ $\lambda = 2$	$\mu = 0$ $\lambda = 1.5$
1	50	0.047	0.165	0.535	0.996
	100	0.045	0.195	0.8	1
	200	0.036	0.323	0.974	1
	400	0.052	0.532	1	1
	500	0.056	0.614	1	1
	800	0.035	0.810	1	1
	1,000	0.045	0.895	1	1
	2,000	0.050	0.994	1	1
2	50	0.078	0.376	0.595	0.998
	100	0.067	0.551	0.823	1
	200	0.062	0.786	0.976	1
	400	0.055	0.969	1	1
	500	0.059	0.985	1	1
	800	0.052	1	1	1
	1,000	0.056	1	1	1
	2,000	0.05	1	1	1
3	50	0.091	0.569	0.571	0.997
	100	0.093	0.762	0.758	1
	200	0.072	0.935	0.939	1
	400	0.06	1	0.996	1
	500	0.064	0.999	0.997	1
	800	0.069	1	1	1
	1,000	0.06	1	1	1
	2,000	0.049	1	1	1

Distances $\mathcal{W}_p(N(0,1), N(\mu, \lambda))$

p	1	2	3
$\mu = 1$ $\lambda = 2$	1.16664	1.41421	1.61120
$\mu = 1$ $\lambda = 1.5$	1.00849	1.11803	1.20538
$\mu = 0$ $\lambda = 2$	0.79788	1	1.16858
$\mu = 0$ $\lambda = 1.5$	0.39894	0.5	0.58429

Fair Learning setting

- $Y = \begin{cases} 0 & failure \\ 1 & success \end{cases}$ target class
- $X \in \mathbb{R}^d,\ d \geqslant 1$, visible attributes
- $S = \begin{cases} 0 & unfavored \\ 1 & favored \end{cases}$ protected attribute
- $\mathcal{G}$ family of binary classifiers $g : \mathbb{R}^d \to \{0, 1\}$

Criteria of fairness

- Disparate Impact

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

$\rightarrow$ g is said not to have Disparate Impact at level $\tau \in (0, 1]$ if $DI(g, X, S) > \tau$

- Balanced Error Rate

$$BER(g, X, S) = \frac{\mathbb{P}\left(g(X) = 0 \mid S = 1\right) + \mathbb{P}\left(g(X) = 1 \mid S = 0\right)}{2}$$

$\rightarrow$ Given $\varepsilon > 0$, S is not $\varepsilon-$predictable from X if $BER(g, X, S) > \varepsilon$, for all $g \in \mathcal{G}$

Application to Fair Learning

E. del Barrio, F. Gamboa, P. Gordaliza and J.-M. Loubes (2018):

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

$\Rightarrow$ S is not $\varepsilon-$predictable from X for all $\varepsilon < \varepsilon^*$

$\Rightarrow$ the maximal value of ε^* is $1/2 \Leftrightarrow d_{TV}\left(\mu_0, \mu_1\right) = 0$
$\Leftrightarrow$ total confusion between μ_0 and μ_1
$\Leftrightarrow$ complete absence of bias in the training data

Fairness assessment:

× $H_0 : d_{TV}\left(\mu_0, \mu_1\right) \geq \Delta_0$ vs $H_a : d_{TV}\left(\mu_0, \mu_1\right) < \Delta_0$, for $\Delta_0 > 0$ (Barron, 1989)

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

- Confidence intervals for $\mathcal{W}_p(\mu_0, \mu_1)$ using CLT (two-sample version)
- Application to high-dimensional data:
  - score $f : \mathbb{R}^d \to \mathbb{R} \dashrightarrow \mathcal{W}_p(\mathcal{L}(f(X) \mid S = 0), \mathcal{L}(f(X) \mid S = 1))$
  - f logistic regression (other regression models or machine learning techniques: SVM, random forest...)

Application to Fair Learning

E. del Barrio, F. Gamboa, P. Gordaliza and J.-M. Loubes (2018):

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

$\Rightarrow$ S is not $\varepsilon-$predictable from X for all $\varepsilon < \varepsilon^*$

$\Rightarrow$ the maximal value of ε^* is $1/2 \Leftrightarrow d_{TV}\left(\mu_0, \mu_1\right) = 0$
$\Leftrightarrow$ total confusion between μ_0 and μ_1
$\Leftrightarrow$ complete absence of bias in the training data

Fairness assessment:

× $H_0 : d_{TV}\left(\mu_0, \mu_1\right) \geq \Delta_0$ vs $H_a : d_{TV}\left(\mu_0, \mu_1\right) < \Delta_0$, for $\Delta_0 > 0$ (Barron, 1989)

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

- Confidence intervals for $\mathcal{W}_p(\mu_0, \mu_1)$ using CLT (two-sample version)
- Application to high-dimensional data:
  - score $f : \mathbb{R}^d \to \mathbb{R} \dashrightarrow \mathcal{W}_p(\mathcal{L}(f(X) \mid S = 0), \mathcal{L}(f(X) \mid S = 1))$
  - f logistic regression (other regression models or machine learning techniques: SVM, random forest...)

Application to Fair Learning

E. del Barrio, F. Gamboa, P. Gordaliza and J.-M. Loubes (2018):

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

$\Rightarrow$ S is not $\varepsilon-$predictable from X for all $\varepsilon < \varepsilon^*$

$\Rightarrow$ the maximal value of ε^* is $1/2 \Leftrightarrow d_{TV}\left(\mu_0, \mu_1\right) = 0$
$\Leftrightarrow$ total confusion between μ_0 and μ_1
$\Leftrightarrow$ complete absence of bias in the training data

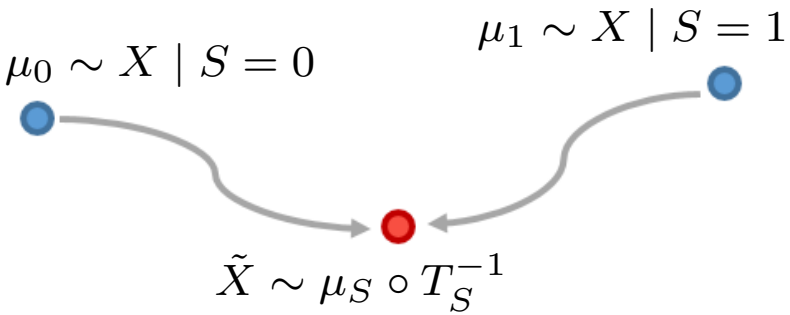
Fairness assessment:

× $H_0 : d_{TV}\left(\mu_0, \mu_1\right) \geq \Delta_0$ vs $H_a : d_{TV}\left(\mu_0, \mu_1\right) < \Delta_0$, for $\Delta_0 > 0$ (Barron, 1989)

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

- Confidence intervals for $\mathcal{W}_p(\mu_0, \mu_1)$ using CLT (two-sample version)
- Application to high-dimensional data:
  - score $f : \mathbb{R}^d \to \mathbb{R} \dashrightarrow \mathcal{W}_p(\mathcal{L}(f(X) \mid S = 0), \mathcal{L}(f(X) \mid S = 1))$
  - f logistic regression (other regression models or machine learning techniques: SVM, random forest...)

Repairing the data with Wasserstein barycenter

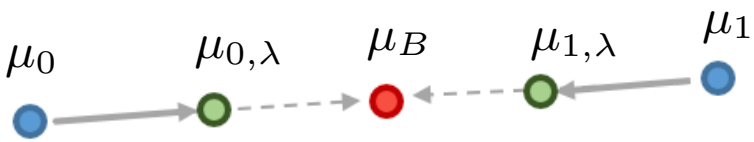


Goal: $X \longrightarrow \tilde{X}$ such that
$\mathcal{L}\left(\tilde{X} \mid S = 0\right) = \mathcal{L}\left(\tilde{X} \mid S = 1\right)$

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

$$\Rightarrow DI(g, \tilde{X}, S) = 1$$

M. Feldman et al. (2015) $\rightarrow$ Geometric Repair: move μ_0, μ_1 only part towards μ_B along Wasserstein's geodesic



$\mu_{s,\lambda} = \mathcal{L}(\lambda T_s(X) + (1-\lambda)X \mid S = s)$,
$\lambda \in [0,1]$ amount of repair desired for X

E. del Barrio, F. Gamboa, P. Gordaliza and J.-M. Loubes (2018): under some regularity conditions, $\mathcal{E}(\tilde{X}) := R_B(\tilde{X}) - R_B(X, S)$

$$\mathcal{E}(\tilde{X}) \leq 2\sqrt{2}K \left(\sum_{s=0,1} \pi_s \mathcal{W}_2^2(\mu_s, \mu_{s\sharp} T_s) \right)^{\frac{1}{2}}, K > 0$$

Application to a real data example: Adult Income data

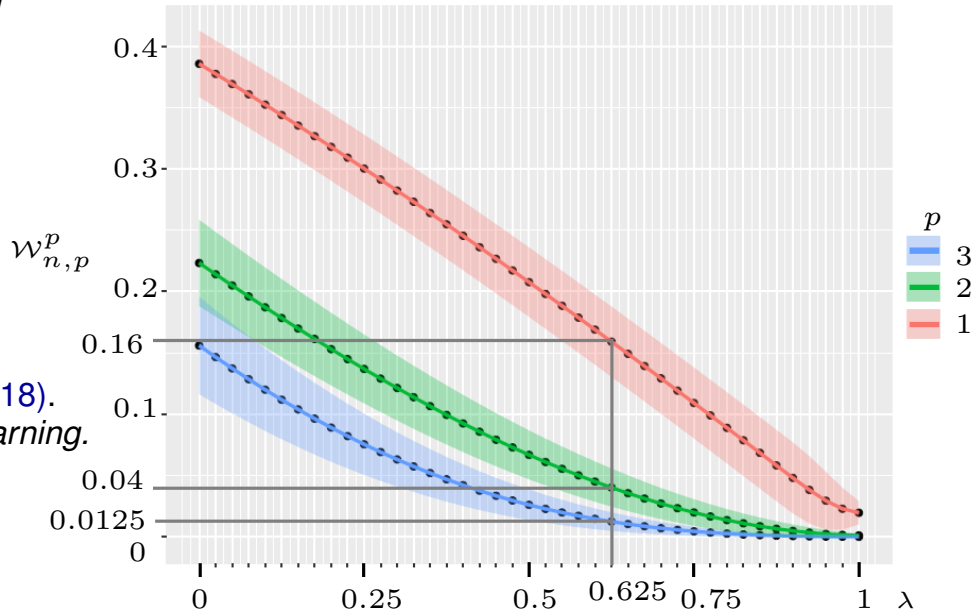
$$Y = \begin{cases} 1 & \text{income exceeds } \$\ 50.000/year \\ 0 & \text{otherwise} \end{cases}$$

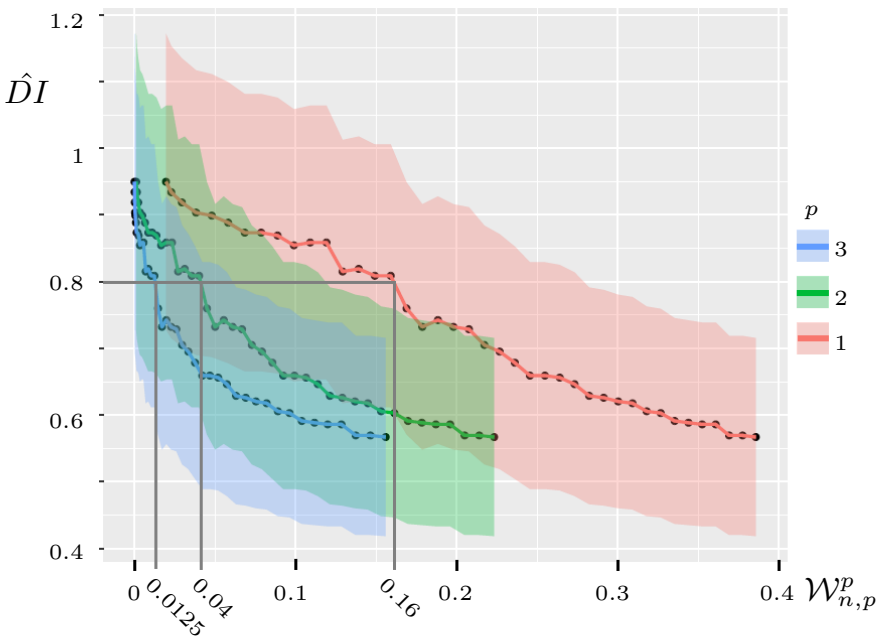
$X =$(age, education number, capital gain, capital loss, worked hours/ week)

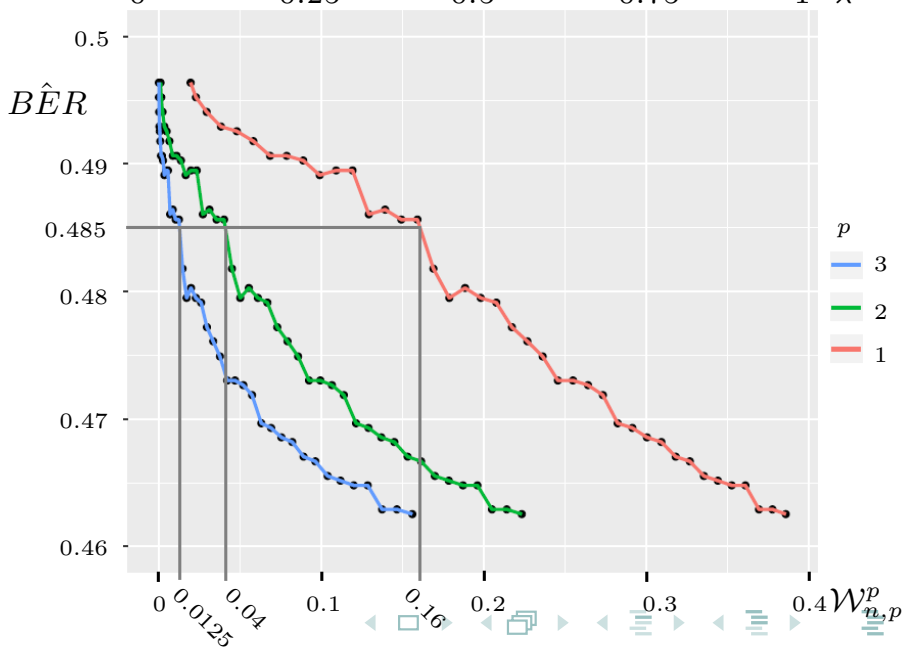
$$S = \text{gender} \begin{cases} 0 & female \\ 1 & male \end{cases}$$

P. Besse, E. del Barrio, P. Gordaliza and J.-M. Loubes (2018). *Confidence intervals for testing disparate impact in fair learning.* arXiv

↓







E. del Barrio, P. Gordaliza and J.-M. Loubes. (December, 2018)
A central limit theorem on the real line with application to fairness assessment in machine learning.
Accepted for publication in Information and Inference.

Thanks for the attention!