Comparing distributions: ℓ_1 geometry improves kernel two-sample testing

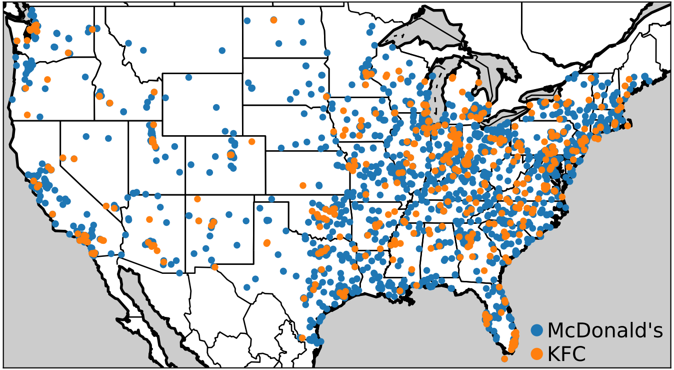
M. Scetbon[1,2] G. Varoquaux[1]

[1]Inria, Université Paris-Saclay

[2]CREST, ENSAE

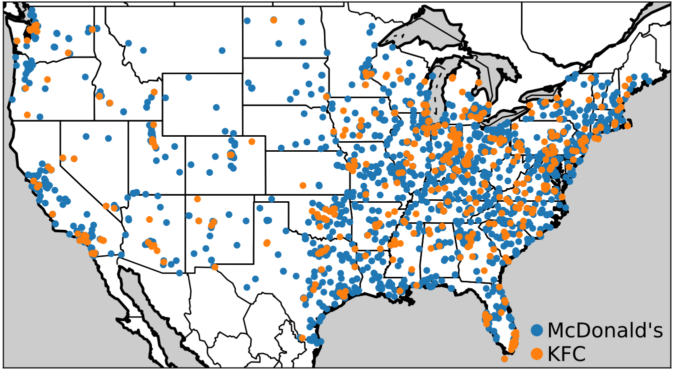
12 décembre 2019

- Two collections of samples **X**, **Y** from unknown distributions **P** and **Q**.



- **Problem** : Are the two set of observations **X** and **Y** drawn from the same distribution ?

- Two collections of samples **X**, **Y** from unknown distributions **P** and **Q**.



- **Problem** : Are the two set of observations **X** and **Y** drawn from the same distribution ?

Two-Sample Test

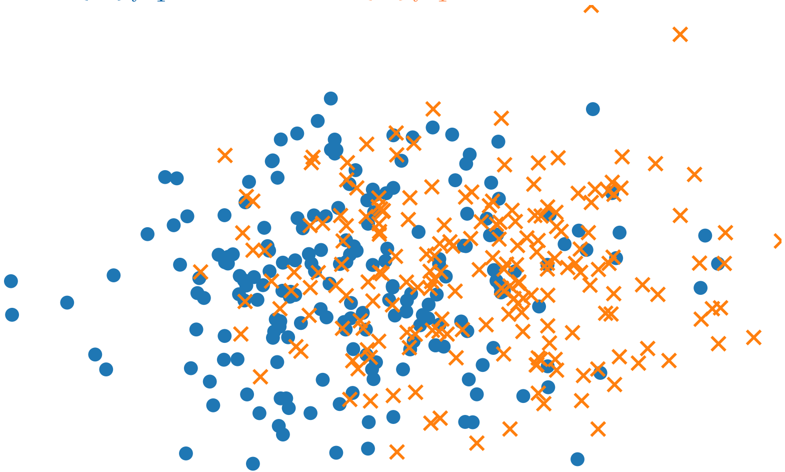
Test the null hypothesis $\mathbf{H}_0 : \mathbf{P} = \mathbf{Q}$ against $\mathbf{H}_1 : \mathbf{P} \neq \mathbf{Q}$

- Samples : $\mathbf{X} = \{x_i\}_{i=1}^n \sim \mathbf{P}$ and $\mathbf{Y} = \{y_i\}_{i=1}^n \sim \mathbf{Q}$

Two-Sample Test

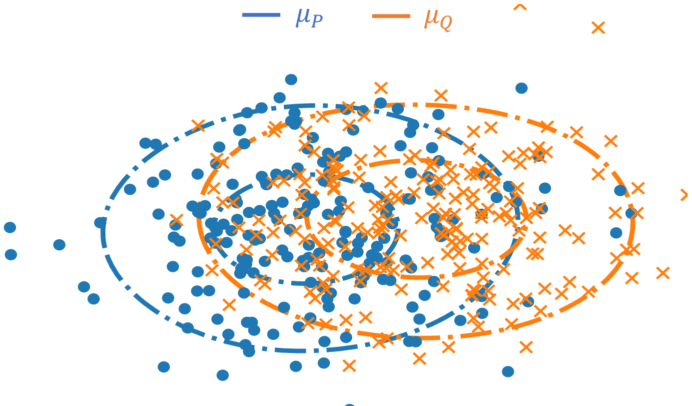
Test the null hypothesis $\mathbf{H}_0 : \mathbf{P} = \mathbf{Q}$ against $\mathbf{H}_1 : \mathbf{P} \neq \mathbf{Q}$

- Samples : $\mathbf{X} = \{x_i\}_{i=1}^n \sim \mathbf{P}$ and $\mathbf{Y} = \{y_i\}_{i=1}^n \sim \mathbf{Q}$



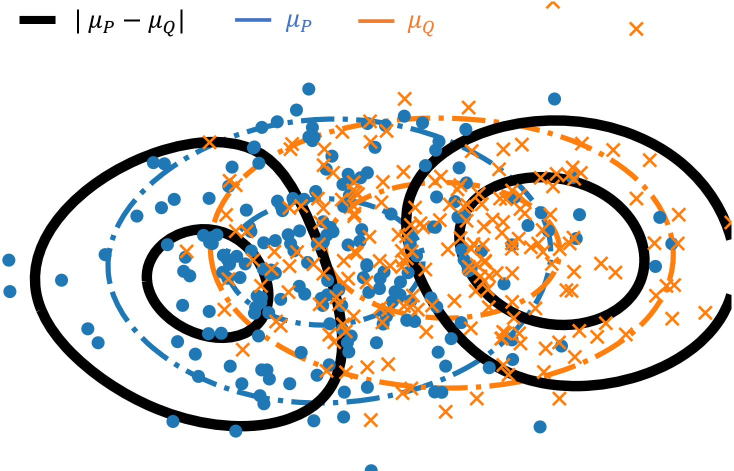
- Gaussian Kernel : $k_\sigma(x, y) = \exp\left(-\frac{\|x-y\|_2^2}{2\sigma^2}\right)$
- Empirical Mean Embeddings of $\mathbf{P}$ and $\mathbf{Q}$:

$$\widehat{\mu}_{\mathbf{P}}(\mathbf{T}) = \sum_{i=1}^{n} k(x_i, \mathbf{T}) \qquad \widehat{\mu}_{\mathbf{Q}}(\mathbf{T}) = \sum_{j=1}^{n} k(y_j, \mathbf{T})$$



- Aboslute difference of the Mean Embeddings :

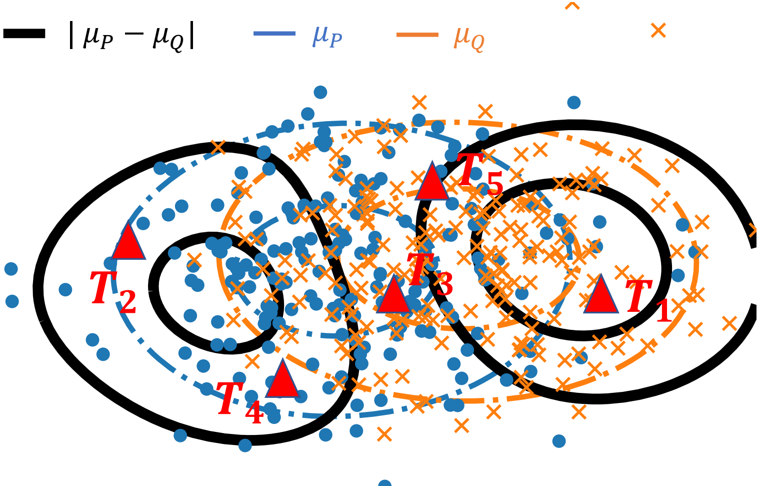
$$\widehat{\mathbf{S}}(\mathbf{T}) = |\widehat{\mu}_{\mathbf{P}}(\mathbf{T}) - \widehat{\mu}_{\mathbf{Q}}(\mathbf{T})|$$



- Aboslute difference of the Mean Embeddings :

$$\widehat{\mathbf{S}}(\mathbf{T}) = |\widehat{\mu}_{\mathbf{P}}(\mathbf{T}) - \widehat{\mu}_{\mathbf{Q}}(\mathbf{T})|$$

- Test locations : $(\mathbf{T_j})_{j=1}^{J} \sim \Gamma$



Test Statistic [1] with $p \geq 1$:

$$\left(\widehat{d}_{\ell_p,\mu,J}(\mathbf{X},\mathbf{Y})\right)^p := n^{\frac{p}{2}} \sum_{j=1}^{J} \left|\widehat{\mu}_{\mathbf{P}}(\mathbf{T_j}) - \widehat{\mu}_{\mathbf{Q}}(\mathbf{T_j})\right|^p$$

These Statistics are derived from metrics which **metrize the weak convergence** :

$$d_{L^p,\mu}(\mathbf{P},\mathbf{Q}) := \left(\int_{t\in\mathbb{R}^d} \left|\mu_{\mathbf{P}}(t) - \mu_{\mathbf{Q}}(t)\right|^p d\Gamma(t)\right)^{1/p}$$

Theorem : Weak Convergence

$$\alpha_n \xrightarrow{\mathcal{D}} \alpha \iff d_{L^p,\mu}(\alpha_n,\alpha) \to 0$$

1. The case when $p = 2$ has been studied by [1, 2]

Test Statistic [1] with $p \geq 1$:

$$\left(\widehat{d}_{\ell_p,\mu,J}(\mathbf{X},\mathbf{Y})\right)^p := n^{\frac{p}{2}} \sum_{j=1}^{J} |\widehat{\mu}_{\mathbf{P}}(\mathbf{T_j}) - \widehat{\mu}_{\mathbf{Q}}(\mathbf{T_j})|^p$$

These Statistics are derived from metrics which **metrize the weak convergence** :

$$d_{L^p,\mu}(\mathbf{P},\mathbf{Q}) := \left(\int_{t\in\mathbb{R}^d} \left|\mu_{\mathbf{P}}(t) - \mu_{\mathbf{Q}}(t)\right|^p d\Gamma(t)\right)^{1/p}$$

Theorem : Weak Convergence

$$\alpha_n \xrightarrow{\mathcal{D}} \alpha \iff d_{L^p,\mu}(\alpha_n,\alpha) \to 0$$

1. The case when $p = 2$ has been studied by [1, 2]

Test Statistic [1] with $p \geq 1$:

$$\left(\widehat{d}_{\ell_p,\mu,J}(\mathbf{X},\mathbf{Y})\right)^p := n^{\frac{p}{2}} \sum_{j=1}^{J} |\widehat{\mu}_{\mathbf{P}}(\mathbf{T_j}) - \widehat{\mu}_{\mathbf{Q}}(\mathbf{T_j})|^p$$

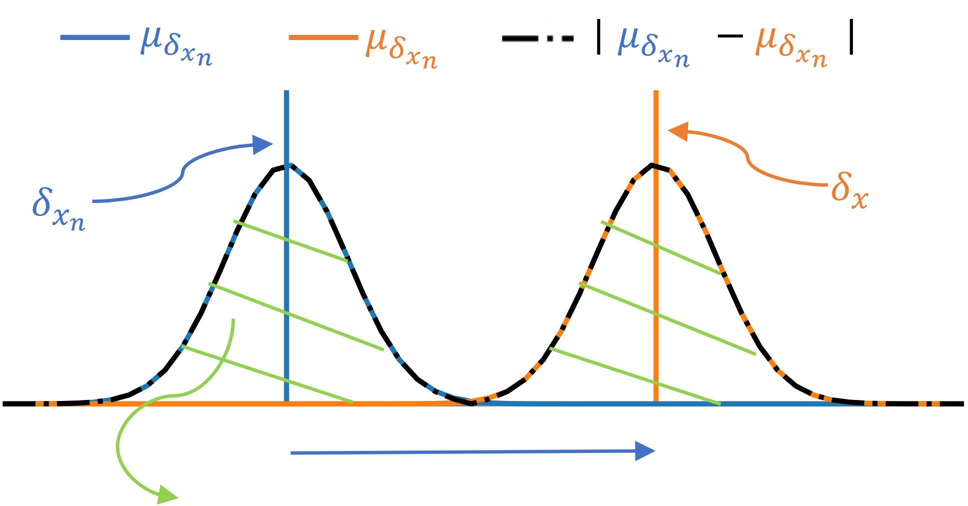
These Statistics are derived from metrics which **metrize the weak convergence** :

$$d_{L^p,\mu}(\mathbf{P},\mathbf{Q}) := \left(\int_{t\in\mathbb{R}^d} \left|\mu_{\mathbf{P}}(t) - \mu_{\mathbf{Q}}(t)\right|^p d\Gamma(t)\right)^{1/p}$$

Theorem : Weak Convergence

$$\alpha_n \xrightarrow{\mathcal{D}} \alpha \iff d_{L^p,\mu}(\alpha_n,\alpha) \to 0$$

1. The case when $p = 2$ has been studied by [1, 2]



$d_{L^1,\mu}(\delta_{x_n}, \delta_x) \to 0$ $\quad$ $TV(\delta_{x_n}, \delta_x) = 2$

Test of level α : Compute $\left(\widehat{d}_{\ell_p,\mu,J}(\mathbf{X},\mathbf{Y})\right)^p$ and reject $\mathbf{H_0}$ if $\left(\widehat{d}_{\ell_p,\mu,J}(\mathbf{X},\mathbf{Y})\right)^p > \mathbf{T_{\alpha,p}} = \mathbf{1} - \alpha$ quantile of the asymptotic null distribution.

Proposition : ℓ_1 geometry improves power

Let $\delta > 0$. Under the alternative hypothesis $\mathbf{H_1}$, almost surely there exist $N \geq 1$ such that for all $n \geq N$ with a probability $1 - \delta$:

$$\left(\widehat{d}_{\ell_2,\mu,J}(\mathbf{X},\mathbf{Y})\right)^2 > \mathbf{T_{\alpha,2}} \Rightarrow \widehat{d}_{\ell_1,\mu,J}(\mathbf{X},\mathbf{Y}) > \mathbf{T_{\alpha,1}}$$

Test of level α : Compute $\left(\widehat{d}_{\ell_p,\mu,J}(\mathbf{X},\mathbf{Y})\right)^p$ and reject $\mathbf{H}_0$ if $\left(\widehat{d}_{\ell_p,\mu,J}(\mathbf{X},\mathbf{Y})\right)^p > \mathbf{T}_{\alpha,\mathbf{p}} = \mathbf{1} - \alpha$ quantile of the asymptotic null distribution.

Proposition : ℓ_1 geometry improves power

Let $\delta > 0$. Under the alternative hypothesis $\mathbf{H}_1$, almost surely there exist $N \geq 1$ such that for all $n \geq N$ with a probability $1-\delta$:

$$\left(\widehat{d}_{\ell_2,\mu,J}(\mathbf{X},\mathbf{Y})\right)^2 > \mathbf{T}_{\alpha,\mathbf{2}} \Rightarrow \widehat{d}_{\ell_1,\mu,J}(\mathbf{X},\mathbf{Y}) > \mathbf{T}_{\alpha,\mathbf{1}}$$

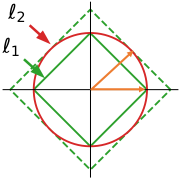
Conclusion

- Under the alternative hypothesis, Analytic Kernel (e.g Gaussian Kernel) guarantees dense differences between $\widehat{\mu}_{\mathbf{P}}$ and $\widehat{\mu}_{\mathbf{Q}}$

- We have also considered statistics based on Smooth Characteristic Functions and obtained similar results.
- Finally we have normalized the tests to obtain a simple null distribution and learn the locations where the distributions differ the most.

@ East Exhibition Hall B + C #6

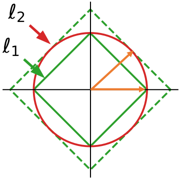
Conclusion

- Under the alternative hypothesis, Analytic Kernel (e.g Gaussian Kernel) guarantees dense differences between $\widehat{\mu}_{\mathbf{P}}$ and $\widehat{\mu}_{\mathbf{Q}}$
- 

  ℓ_1 geometry captures better these dense differences.
- We have also considered statistics based on Smooth Characteristic Functions and obtained similar results.
- Finally we have normalized the tests to obtain a simple null distribution and learn the locations where the distributions differ the most.

@ East Exhibition Hall B + C #6

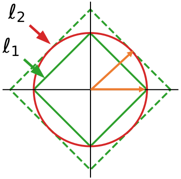
Conclusion

- Under the alternative hypothesis, Analytic Kernel (e.g Gaussian Kernel) guarantees dense differences between $\widehat{\mu}_{\mathbf{P}}$ and $\widehat{\mu}_{\mathbf{Q}}$
- 

  ℓ_1 geometry captures better these dense differences.
- We have also considered statistics based on Smooth Characteristic Functions and obtained similar results.
- Finally we have normalized the tests to obtain a simple null distribution and learn the locations where the distributions differ the most.

@ East Exhibition Hall B + C #6

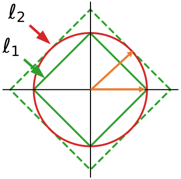
Conclusion

- Under the alternative hypothesis, Analytic Kernel (e.g Gaussian Kernel) guarantees dense differences between $\widehat{\mu}_{\mathbf{P}}$ and $\widehat{\mu}_{\mathbf{Q}}$
- 

  ℓ_1 geometry captures better these dense differences.
- We have also considered statistics based on Smooth Characteristic Functions and obtained similar results.
- Finally we have normalized the tests to obtain a simple null distribution and learn the locations where the distributions differ the most.

@ East Exhibition Hall B + C #6

Conclusion

- Under the alternative hypothesis, Analytic Kernel (e.g Gaussian Kernel) guarantees dense differences between $\widehat{\mu}_{\mathbf{P}}$ and $\widehat{\mu}_{\mathbf{Q}}$
- 

  ℓ_1 geometry captures better these dense differences.
- We have also considered statistics based on Smooth Characteristic Functions and obtained similar results.
- Finally we have normalized the tests to obtain a simple null distribution and learn the locations where the distributions differ the most.

@ East Exhibition Hall B + C #**6**

References I

[1] K. P. Chwialkowski, A. Ramdas, D. Sejdinovic, and A. Gretton. Fast two-sample testing with analytic representations of probability measures. In *Advances in Neural Information Processing Systems*, pages 1981–1989, 2015.

[2] W. Jitkrittum, Z. Szabó, K. P. Chwialkowski, and A. Gretton. Interpretable distribution features with maximum testing power. In *Advances in Neural Information Processing Systems*, pages 181–189, 2016.