Regression on distributions based on regularized Optimal Transport

Léonard Gousset
Louis Allain
Julien Heurtons

Methodological project - Groupe 37

January 2024

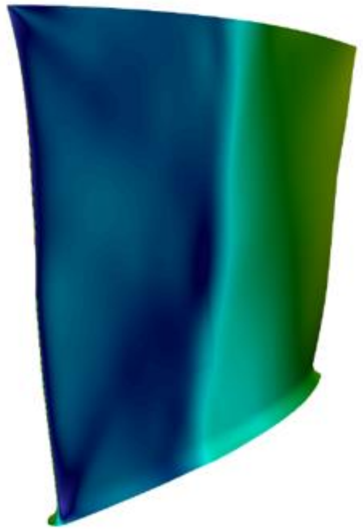
Tutor : Brian Staber

INTRODUCTION 1/3

Problem :

Find the best shape for a blade

A blade in $\mathbb{R}^3$



Fluid dynamics simulation

Aerodynamic coefficient

0.648

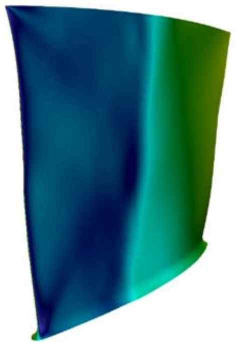
INTRODUCTION 2/3

2 majors problems

1

Simulation too heavy

x_1	0.256
x_2	14.89
…	…
x_{40}	78.02



Coefficient

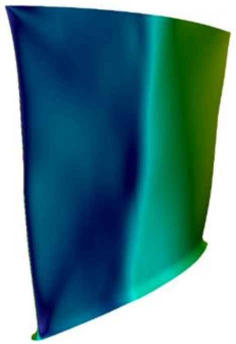
0.648

INTRODUCTION 2/3

2 majors problems

1 Simulation too heavy

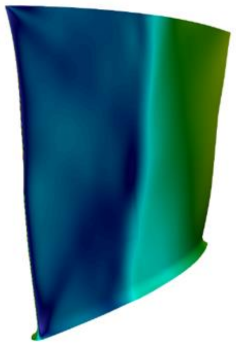
x_1	0.256
x_2	14.89
…	…
x_{40}	78.02



Coefficient

0.648

Can’t always do it 2



x_1	?
x_2	?
…	…
x_{40}	?

Coefficient

?

INTRODUCTION 3/3

Goal :

Find a model to make predictions of the aerodynamic coefficient

Machine Learning for regression

x	y
Distribution 1	0.465
Distribution 2	0.586
...	...
Distribution p	0.258

INTRODUCTION 3/3

Goal :

Find a model to make predictions of the aerodynamic coefficient

Machine Learning for regression

x	y
Distribution 1	0.465
Distribution 2	0.586
…	…
Distribution p	0.258

→ **Kernel for comparing Distributions**

SUMMARY

I. Reminder on Kernel Ridge Regression

II. Optimal Transport
   1. Monge problem
   2. Kantorovich formulation
   3. Regularized Optimal Transport

III. Suggested Kernel
   1. Definitions
   2. Properties

IV. Applications

SUMMARY

I. **Reminder on Kernel Ridge Regression**

II. **Optimal Transport**
1. Monge problem
2. Kantorovich formulation
3. Regularized Optimal Transport

III. **Suggested Kernel**
1. Definitions
2. Properties

IV. **Applications**

Kernel Ridge Regression

Ridge Regression with a kernel

$$\min_{f \in \mathcal{H}_k} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{H}_k}^2$$

Kernel Ridge Regression

Ridge Regression with a kernel

$$\min_{f \in \mathcal{H}_k} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{H}_k}^2$$

Formulation with the Representer Theorem

$$\min_{\alpha \in \mathbb{R}^n} \|Y - K\alpha\|^2 + \lambda \alpha^\intercal K \alpha$$

Kernel Ridge Regression

Kernel Ridge estimator

$$\hat{f} = \sum_{i=1}^{n} \hat{\alpha}_i k\left(\cdot, x_i\right)$$

Kernel Ridge Regression

Kernel Ridge estimator

$$\hat{f} = \sum_{i=1}^{n} \hat{\alpha}_i k\left(\cdot, x_i\right)$$

Additional remarks

If K is the linear kernel, we have: $K = X^\intercal X$ and we have the same formulation as the ridge regression

Kernel Ridge Regression

Kernel Ridge estimator

$$\hat{f} = \sum_{i=1}^{n} \hat{\alpha}_i k\left(\cdot, x_i\right)$$

Additional remarks

If K is the linear kernel, we have: $K = X^\intercal X$ and we have the same formulation as the ridge regression

With the linear Kernel and when $\lambda \to 0$, we get the usual linear regression

SUMMARY

I. Reminder on Kernel methods

II. Optimal Transport
   1. **Monge problem**
   2. Kantorovich formulation
   3. Regularized Optimal Transport

III. Suggested Kernel
   1. Definitions
   2. Properties

IV. Applications

OPTIMAL TRANSPORT - Monge

Gaspard Monge
(1746-1818)

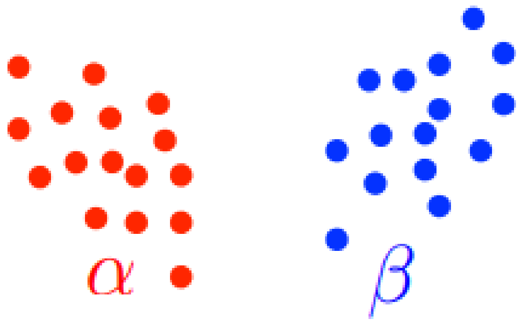
« MASS TRANSPORTATION »

OPTIMAL TRANSPORT - Monge

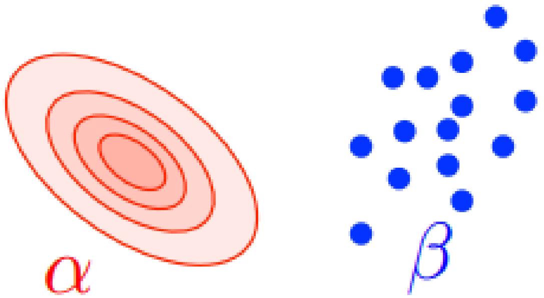
Gaspard Monge
(1746-1818)

« MASS TRANSPORTATION »

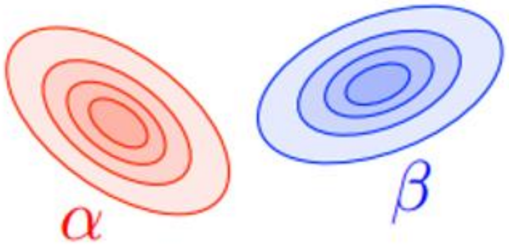
Gabriel Peyré – 2020
« Computational Optimal Transport »



Discrete-Discrete



Continuous-Discrete



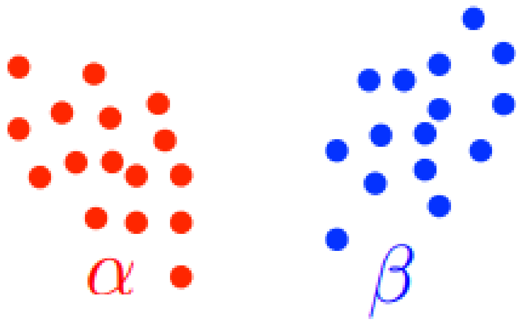
Continuous-Continous

OPTIMAL TRANSPORT - Monge

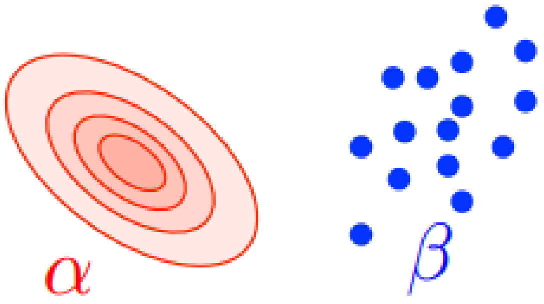
Gaspard Monge
(1746-1818)

« MASS TRANSPORTATION »

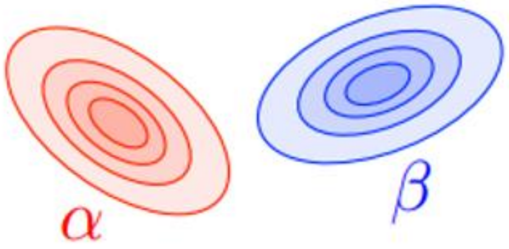
Gabriel Peyré – 2020
« Computational Optimal Transport »



Discrete-Discrete



Continuous-Discrete



Continuous-Continous

$$\min_{\sigma \in \mathrm{Perm}(n)} \sum_{i=1}^{n} C_{i,\sigma(i)}$$

OPTIMAL TRANSPORT - Monge

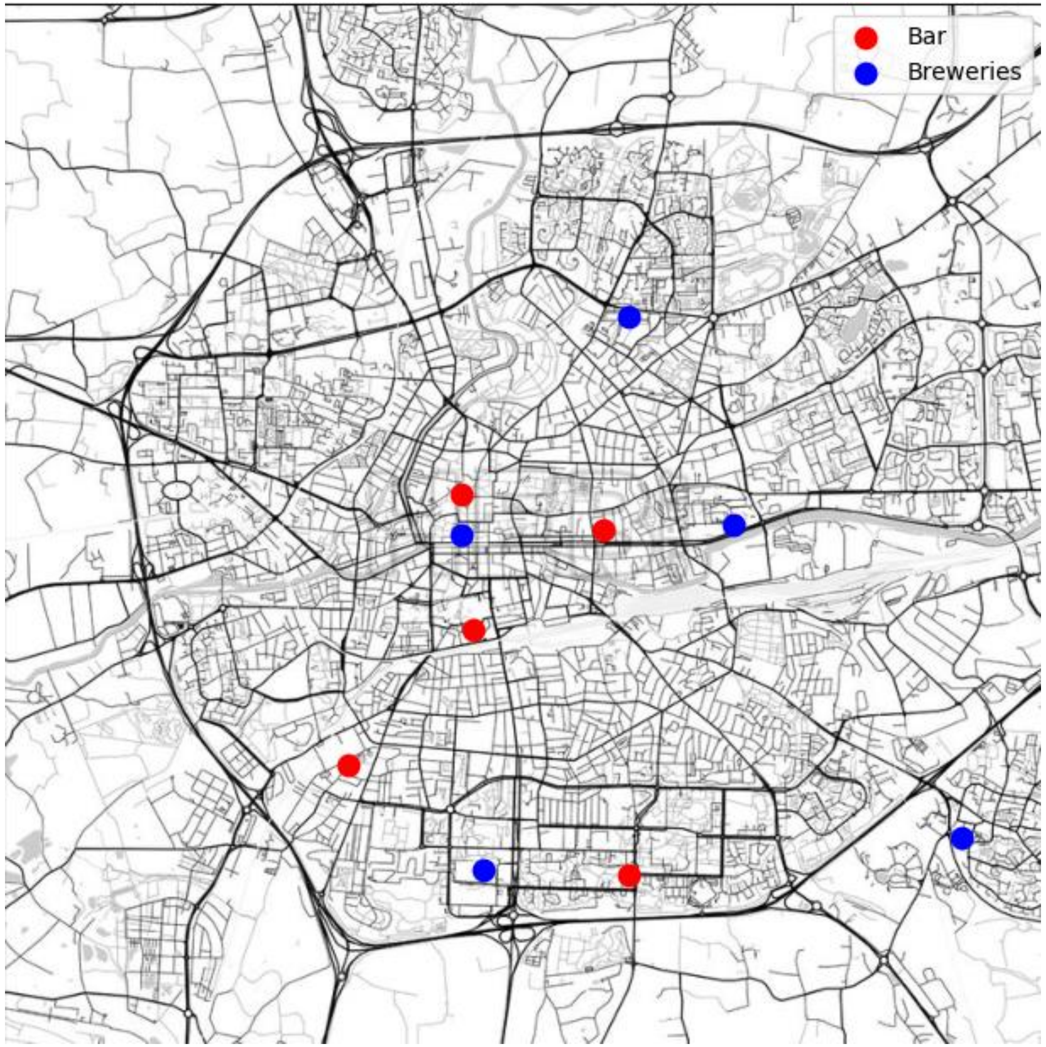
2 discretes distributions :
Breweries and Pubs

i ∈ {1, 2, 3, 4, 5}

$$\alpha = \sum_{i=1}^{5} a_i \delta_{x_i}$$

j ∈ {1, 2, 3, 4, 5}

$$\beta = \sum_{i=1}^{5} b_i \delta_{y_i}$$



OPTIMAL TRANSPORT - Monge

2 discretes distributions :
Breweries and Pubs

i $\in \{1, 2, 3, 4, 5\}$

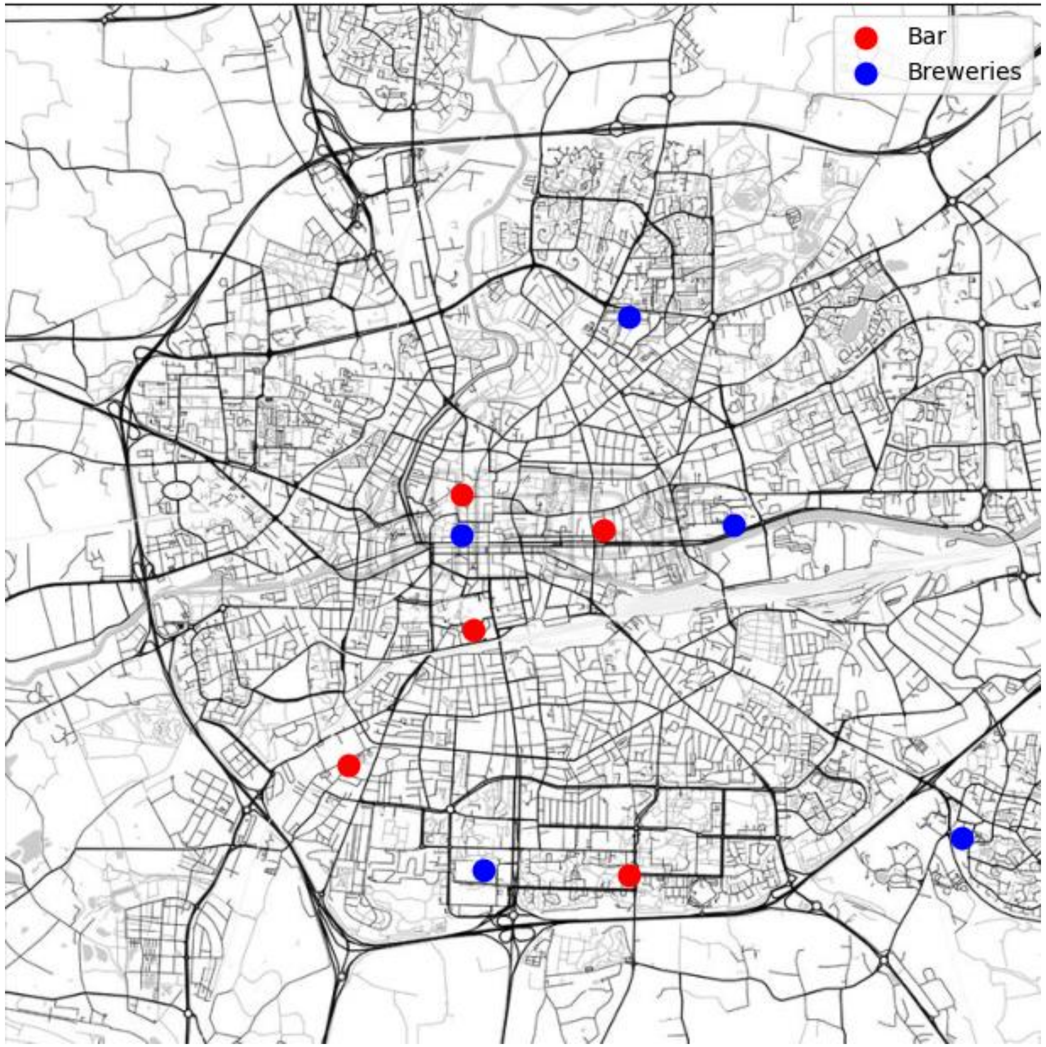
j $\in \{1, 2, 3, 4, 5\}$

$$\boldsymbol{\alpha} = \sum_{i=1}^{5} \boldsymbol{a}_i \delta_{x_i}$$

$$\boldsymbol{\beta} = \sum_{i=1}^{5} \boldsymbol{b}_i \delta_{y_i}$$

Transport map :

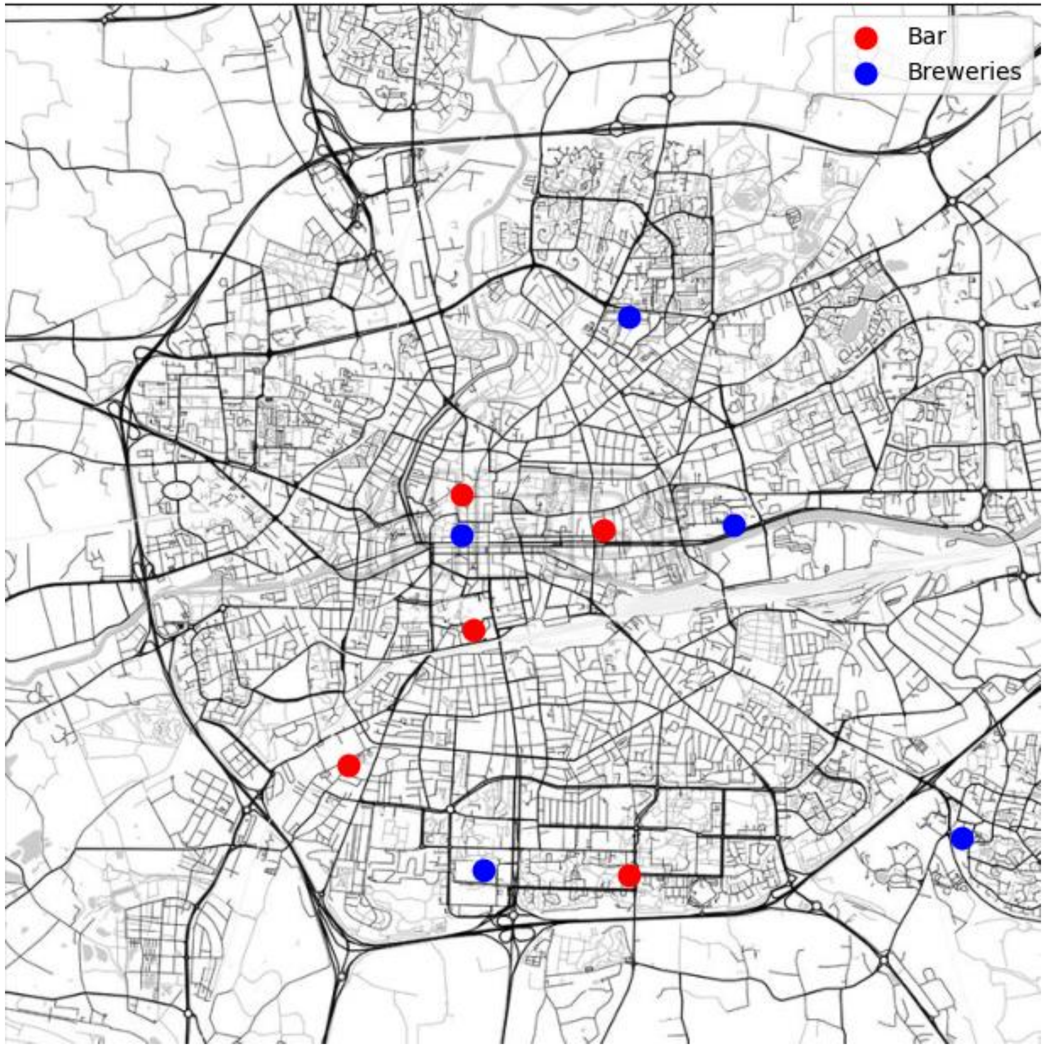
$$\mathrm{T} : \{1, \ldots, 5\} \rightarrow \{1, \ldots, 5\}$$



OPTIMAL TRANSPORT - Monge

Solve :

$$\min_T \left\{ \sum_i c(x_i, T(x_i)) \; : \; T_\sharp \alpha = \beta \right\}$$



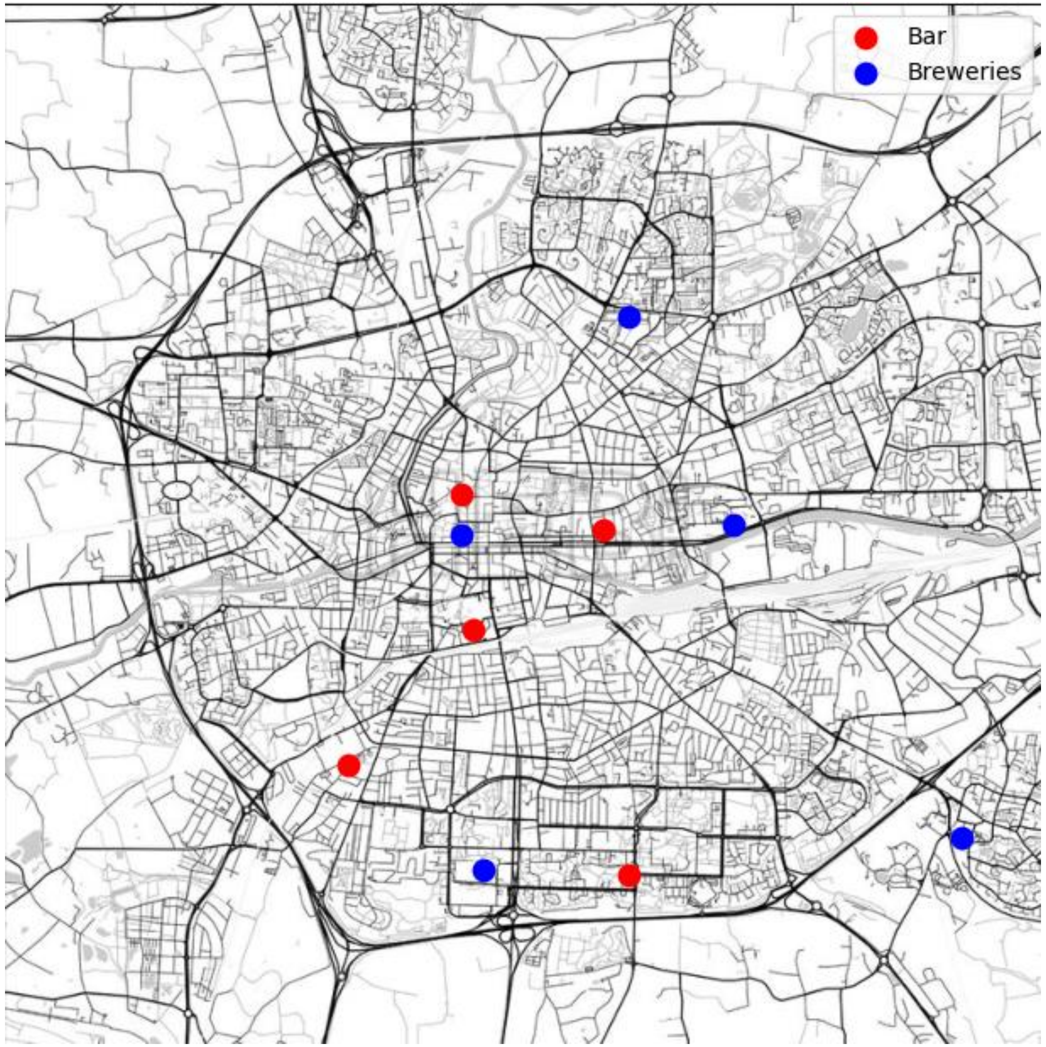
OPTIMAL TRANSPORT - Monge

Solve :

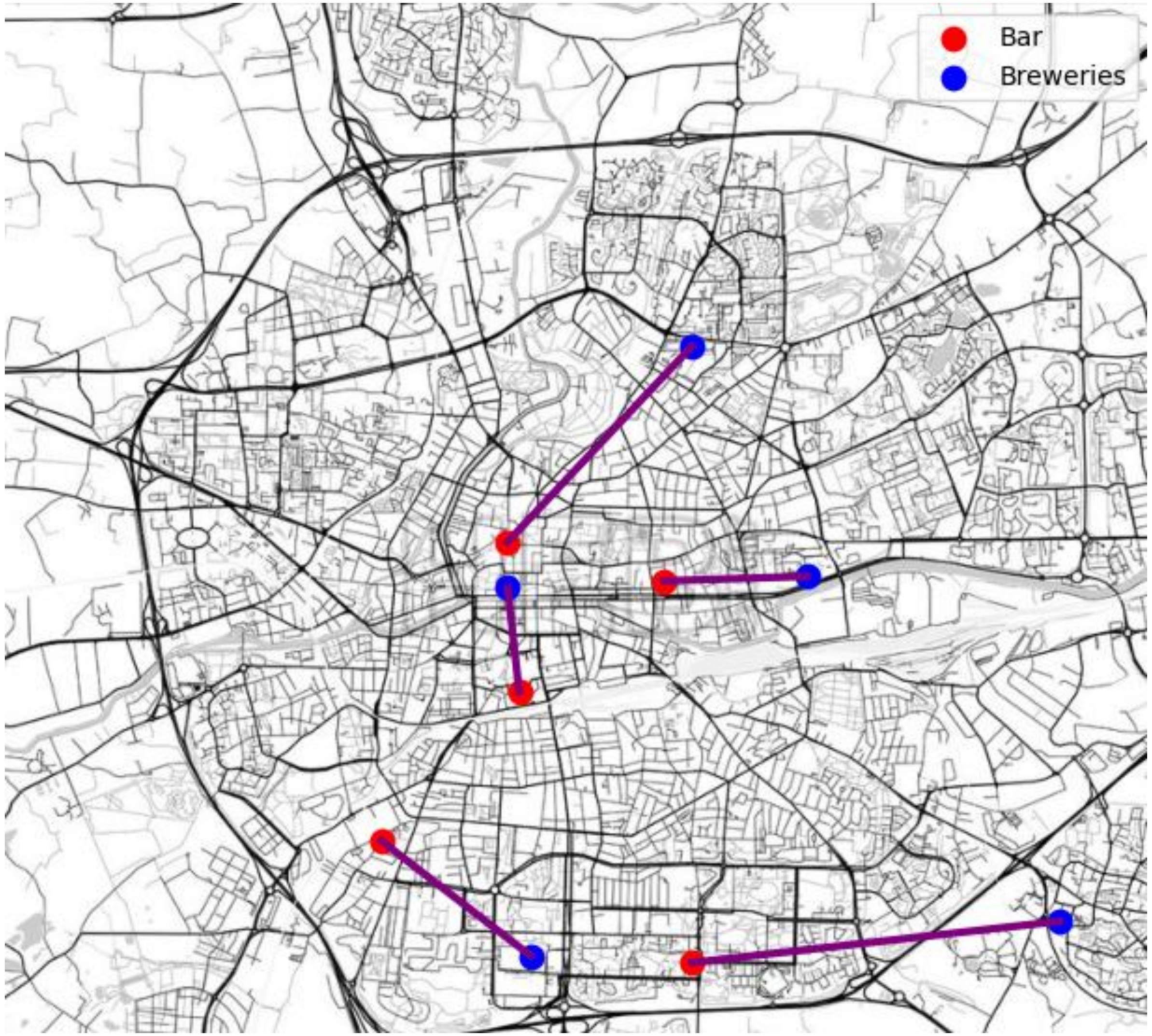
$$\min_{T} \left\{ \sum_i c(x_i, T(x_i)) \;:\; T_\sharp \alpha = \beta \right\}$$

With :

c_{ij}	y_1	y_2	y_3	y_4	y_5
x_1	8	12	22	30	25
x_2	18	2	13	29	12
x_3	12	5	13	26	17
x_4	7	18	18	16	27
x_5	17	7	6	23	10



OPTIMAL TRANSPORT - Monge



C_{ij}	y_1	y_2	y_3	y_4	y_5
x_1	8	12	22	30	25
x_2	18	2	13	29	12
x_3	12	5	13	26	17
x_4	7	18	18	16	27
x_5	17	7	6	23	10

Total Cost = 47

Optimal Transport map :

$$T : \{1,2,3,4,5\} \rightarrow \{1,5,2,4,3\}$$

OPTIMAL TRANSPORT - Monge

1

Computationally impossible

x_i : $i \in \{1, \ldots, 5\}$ y_j : $j \in \{1, \ldots, 5\}$

Permutations

$\sigma : \{1, \ldots, 5\} \rightarrow \{1, \ldots, 5\}$

$5! = 120$
permutations possibles

23 bars and breweries :

$23! = 2.585202 \times 10^{22}$
permutations possibles

OPTIMAL TRANSPORT - Monge

1 Computationally impossible

x_i : $i \in \{1, \dots, 5\}$ y_j : $j \in \{1, \dots, 5\}$

Permutations

$\sigma : \{1, \dots, 5\} \to \{1, \dots, 5\}$

5! = 120

permutations possibles

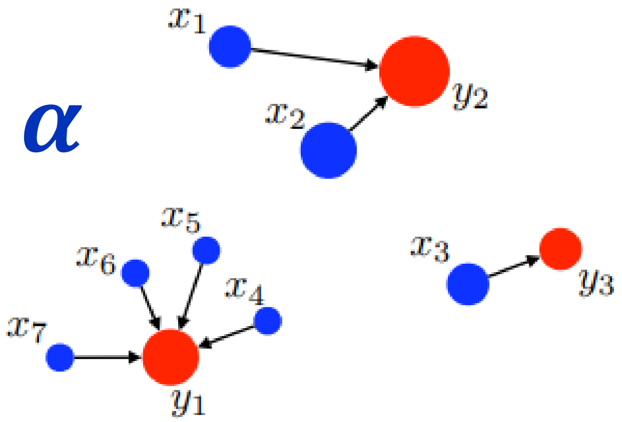
23 bars and breweries :

$23! = 2.585202 \times 10^{22}$

permutations possibles

2 $n \neq m$: not always a solution

α



β

$T_{\#}\alpha$ exists

$T_{\#}\beta$ does not exists

SUMMARY

I. Reminder on Kernel methods

II. Optimal Transport
   1. Monge problem
   2. **Kantorovich formulation**
   3. Regularized Optimal Transport

III. Suggested Kernel
   1. Definitions
   2. Properties

IV. Applications

OPTIMAL TRANSPORT - Kantorovich

Leonid Kantorovich
(1912-1986)

« MASS SPLITTING »

OPTIMAL TRANSPORT - Kantorovich

Leonid Kantorovich
(1912-1986)

« MASS SPLITTING »

$\boldsymbol{P}$: coupling matrix $\in \mathbb{R}_{+}^{n \times m}$

$\boldsymbol{P}_{\mathbf{i,j}}$: amount of mass flowing from source $\boldsymbol{x}_{\mathbf{i}}$ to destination $\boldsymbol{y}_{\mathbf{j}}$

OPTIMAL TRANSPORT - Kantorovich

Leonid Kantorovich
(1912-1986)

« MASS SPLITTING »

Kantorovich's Relaxation

$$L_{\mathbf{C}(\mathbf{a},\mathbf{b})} := \min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} \langle \mathbf{C}, \mathbf{P} \rangle := \sum_{i,j} \mathbf{C}_{i,j}\mathbf{P}_{i,j}.$$

OPTIMAL TRANSPORT - Kantorovich

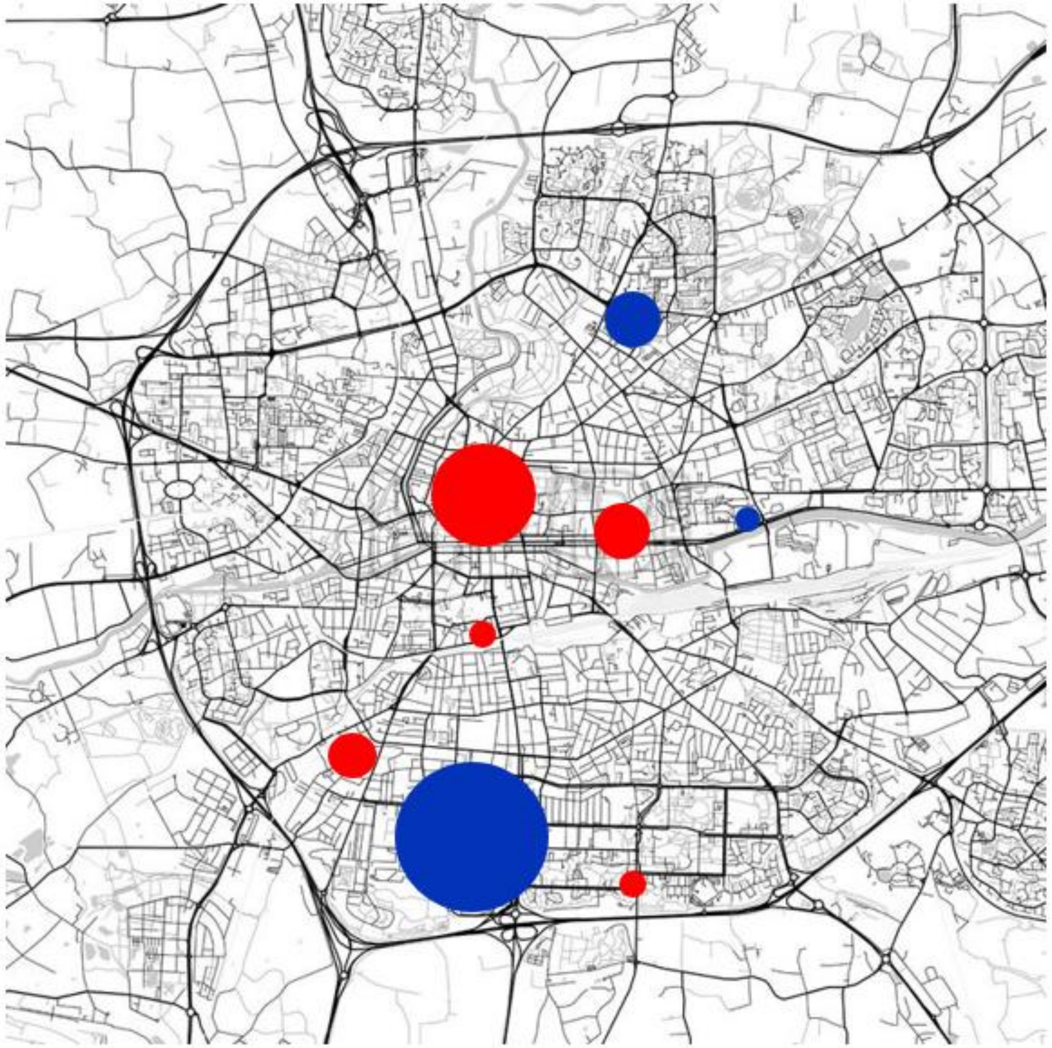
2 discretes distributions :
Breweries and **Pubs**

i ∈ {1, 2, 3}

$$\alpha = \sum_{i=1}^{3} a_i \delta_{x_i}$$

j ∈ {1, 2, 3, 4, 5}

$$\beta = \sum_{i=1}^{5} b_i \delta_{y_i}$$



OPTIMAL TRANSPORT - Kantorovich

2 discretes distributions :
Breweries and **Pubs**

i ∈ {1, 2, 3}

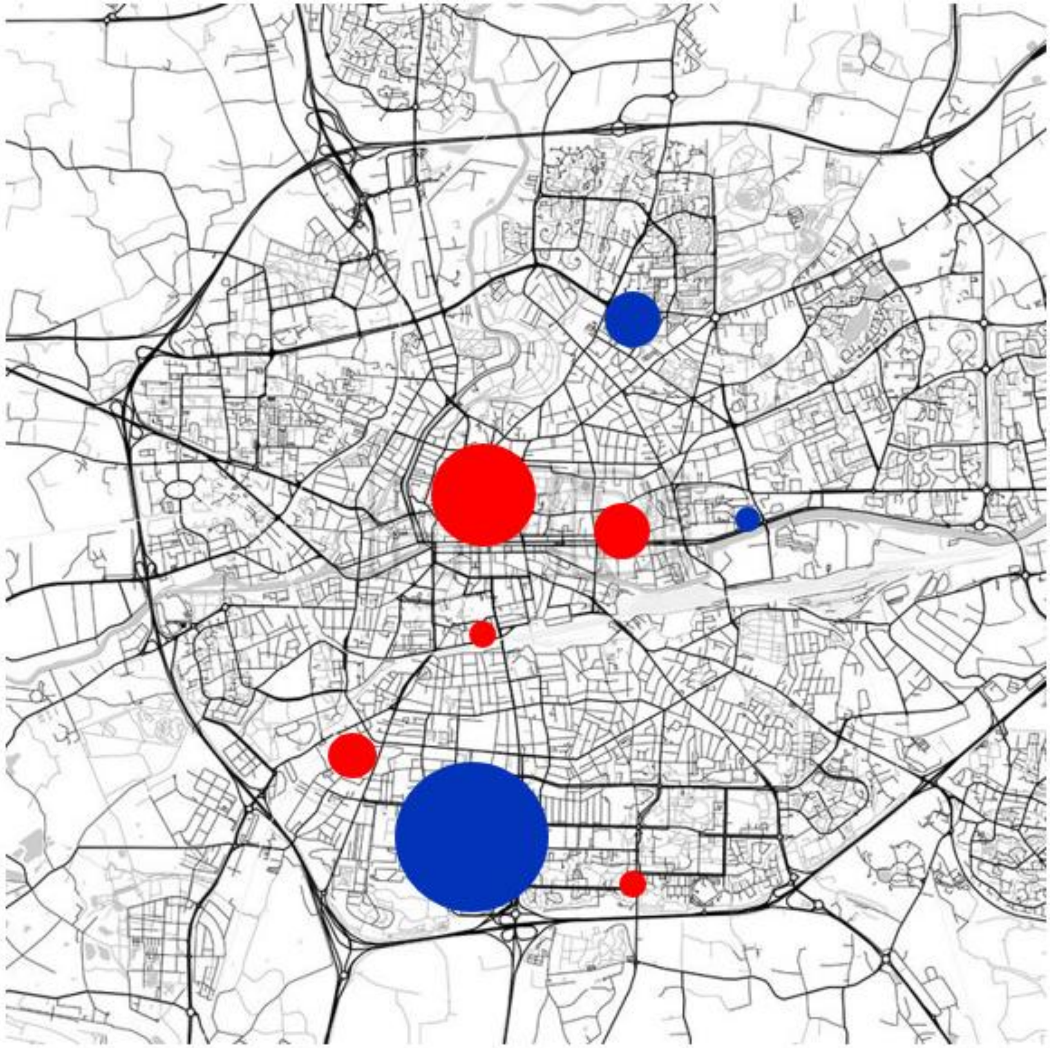
$$\alpha = \sum_{i=1}^{3} a_i \delta_{x_i}$$

j ∈ {1, 2, 3, 4, 5}

$$\beta = \sum_{i=1}^{5} b_i \delta_{y_i}$$

Mass conservation constraint :

$$\boxed{\sum_{i=1}^{3} a_i = \sum_{i=1}^{5} b_i = 1}$$



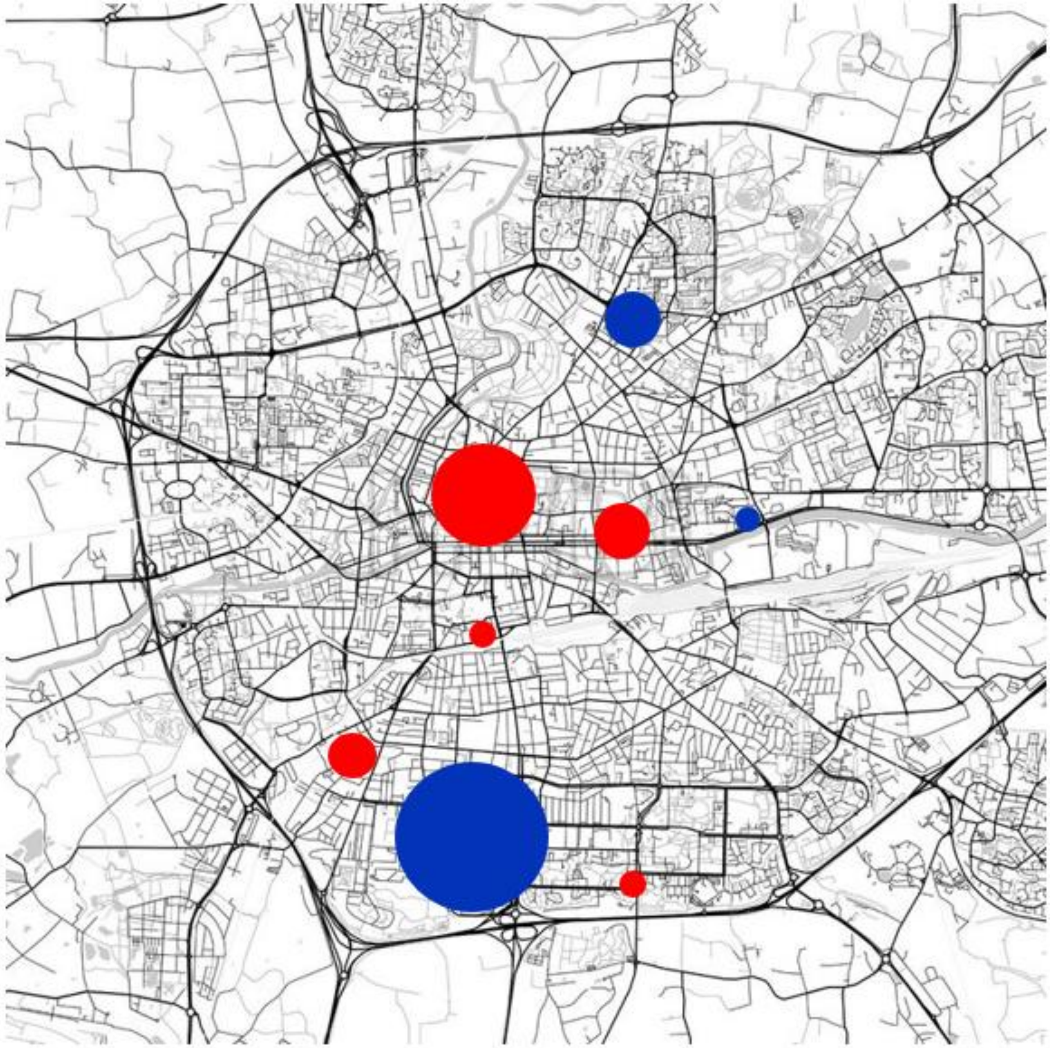
OPTIMAL TRANSPORT - Kantorovich

Solving

$$L_{\mathbf{C}(\mathbf{a},\mathbf{b})} := \min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} \langle \mathbf{C}, \mathbf{P} \rangle := \sum_{i,j} \mathbf{C}_{i,j}\mathbf{P}_{i,j}.$$

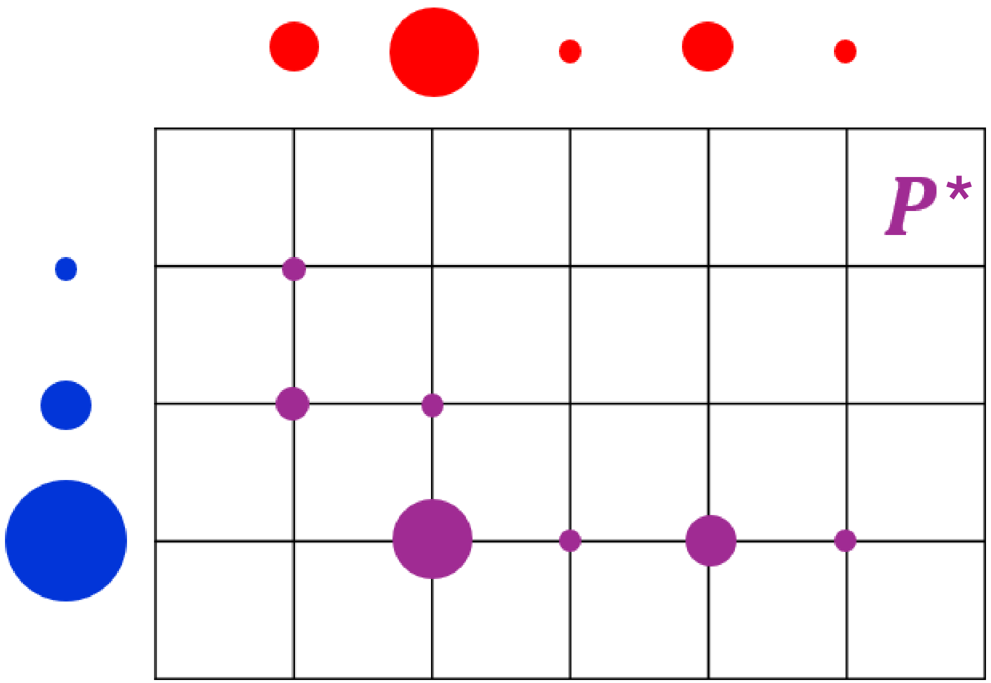
With

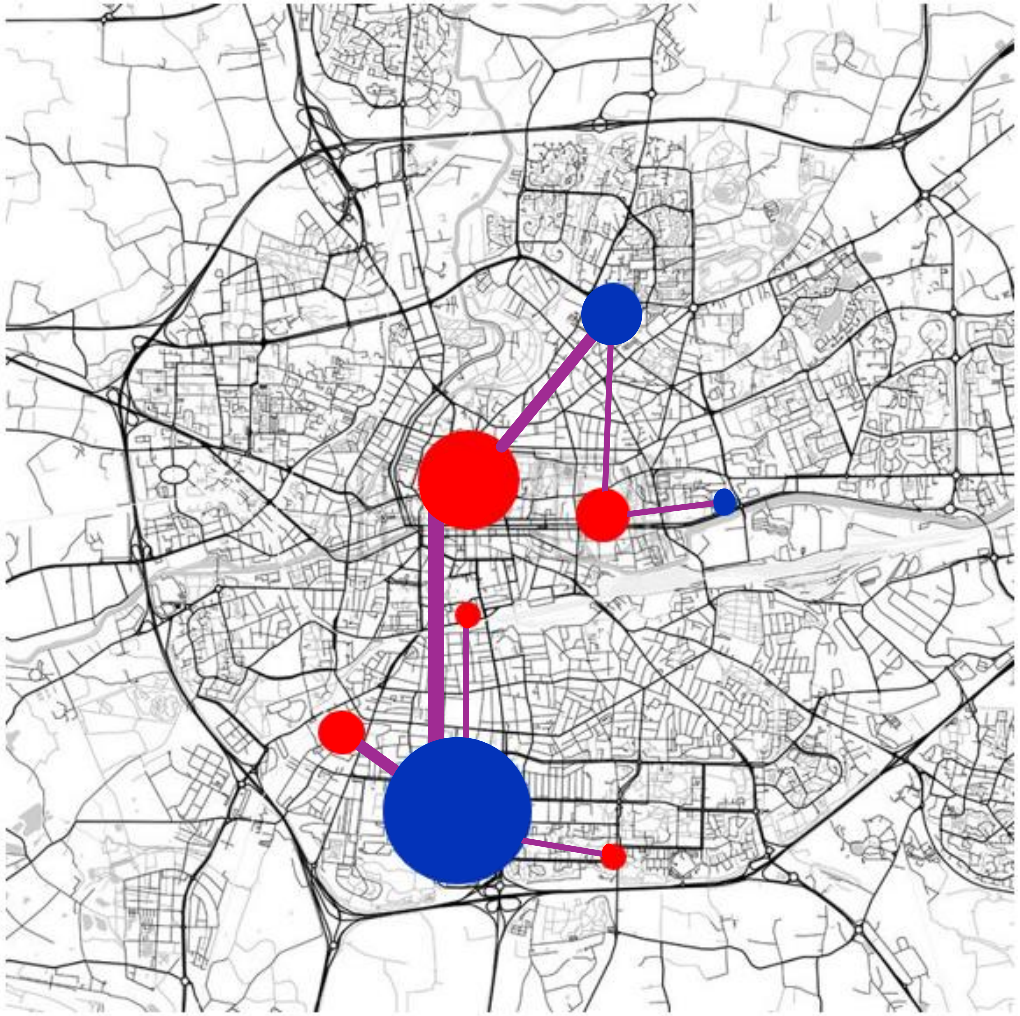
C_{ij}	y_1	y_2	y_3	y_4	y_5
X_1	8	12	22	30	25
X_2	12	5	13	26	17
X_3	17	7	6	23	10



OPTIMAL TRANSPORT - Kantorovich

Optimal Transport Plan :





SUMMARY

I. Reminder on Kernel methods

II. Optimal Transport
1. Monge problem
2. Kantorovich formulation
3. **Regularized Optimal Transport**

III. Suggested Kernel
1. Definitions
2. Properties

IV. Applications

OPTIMAL TRANSPORT – Regularized OT

Introduction of the entropy

$$H(\mathbf{P}) \stackrel{\text{def}}{=} -\sum_{i,j} \mathbf{P}_{i,j}(\log(\mathbf{P}_{i,j}) - 1)$$

OPTIMAL TRANSPORT – Regularized OT

Introduction of the entropy

$$H(\mathbf{P}) \stackrel{\text{def}}{=} -\sum_{i,j} \mathbf{P}_{i,j}(\log(\mathbf{P}_{i,j}) - 1)$$

The entropy as a regularization parameter of the Kantorovich's formulation

$$L_C^{\varepsilon}(\mathbf{a}, \mathbf{b}) \stackrel{\text{def}}{=} \min_{\mathbf{P} \in U(a,b)} \langle \mathbf{P}, \mathbf{C} \rangle - \varepsilon H(\mathbf{P})$$

OPTIMAL TRANSPORT – Regularized OT

Introduction of the entropy

$$H(\mathbf{P}) \stackrel{\text{def}}{=} -\sum_{i,j} \mathbf{P}_{i,j}(\log(\mathbf{P}_{i,j}) - 1)$$

The entropy as a regularization parameter of the Kantorovich's formulation

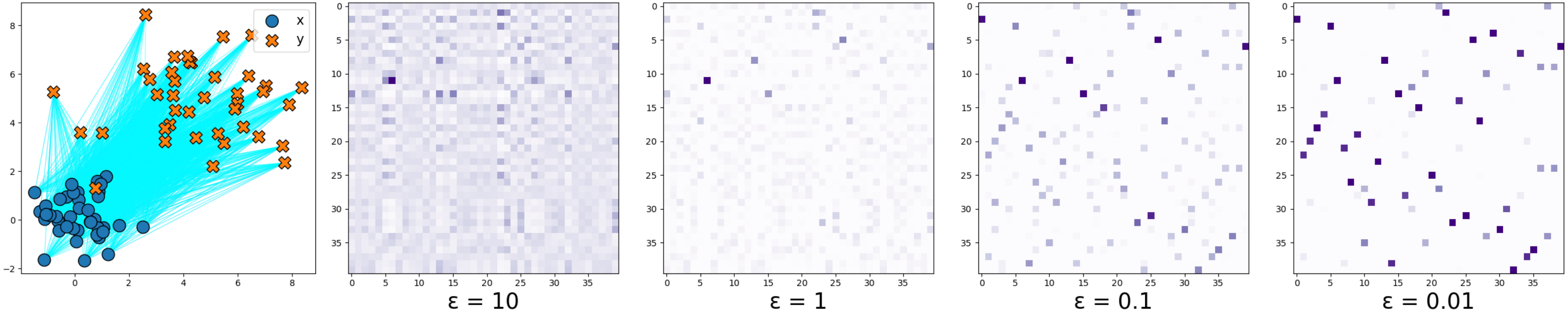
$$L_C^{\varepsilon}(\mathbf{a}, \mathbf{b}) \stackrel{\text{def}}{=} \min_{\mathbf{P} \in U(a,b)} \langle \mathbf{P}, \mathbf{C} \rangle - \varepsilon H(\mathbf{P})$$

Link between Regularized OT and Kantorovich's formulation

$$L_C^{\varepsilon}(\mathbf{a}, \mathbf{b}) \xrightarrow[\varepsilon \to 0]{} L_C(\mathbf{a}, \mathbf{b})$$

OPTIMAL TRANSPORT – Regularized OT

Impact of the regularization parameter on the solution



OPTIMAL TRANSPORT – Regularized OT

Sinkhorn's Algorithm

$$\Lambda_{\mathbf{C}}^{\varepsilon}(\mathbf{P}, \mathbf{f}, \mathbf{g}) = \langle \mathbf{P}, \mathbf{C} \rangle - \varepsilon H(\mathbf{P}) - \langle \mathbf{f}, \mathbf{P}\mathbb{1}_m - \mathbf{a} \rangle - \langle \mathbf{g}, \mathbf{P}^\intercal \mathbb{1}_n - \mathbf{b} \rangle$$

OPTIMAL TRANSPORT – Regularized OT

Sinkhorn's Algorithm

$$\Lambda_{\mathbf{C}}^{\varepsilon}(\mathbf{P}, \mathbf{f}, \mathbf{g}) = \langle \mathbf{P}, \mathbf{C} \rangle - \varepsilon H(\mathbf{P}) - \langle \mathbf{f}, \mathbf{P}\mathbb{1}_m - \mathbf{a} \rangle - \langle \mathbf{g}, \mathbf{P}^\intercal \mathbb{1}_n - \mathbf{b} \rangle$$

$$\mathbf{P}_{i,j} = e^{\frac{\mathbf{f}_i}{\varepsilon}} e^{-\frac{\mathbf{C}_{i,j}}{\varepsilon}} e^{\frac{\mathbf{g}_j}{\varepsilon}}, \qquad \forall (i,j) \in [\![n]\!] \times [\![m]\!]$$

OPTIMAL TRANSPORT – Regularized OT

Sinkhorn's Algorithm

$$\Lambda_{\mathbf{C}}^{\varepsilon}(\mathbf{P}, \mathbf{f}, \mathbf{g}) = \langle \mathbf{P}, \mathbf{C} \rangle - \varepsilon H(\mathbf{P}) - \langle \mathbf{f}, \mathbf{P}\mathbb{1}_m - \mathbf{a} \rangle - \langle \mathbf{g}, \mathbf{P}^{\intercal}\mathbb{1}_n - \mathbf{b} \rangle$$

$$\mathbf{P}_{i,j} = e^{\frac{\mathbf{f}_i}{\varepsilon}} e^{-\frac{\mathbf{C}_{i,j}}{\varepsilon}} e^{\frac{\mathbf{g}_j}{\varepsilon}}, \qquad \forall (i,j) \in [\![n]\!] \times [\![m]\!]$$

$$\mathbf{P}_{i,j} = \mathbf{u}_i \mathbf{K}_{i,j} \mathbf{v}_i, \qquad \forall (i,j) \in [\![n]\!] \times [\![m]\!]$$

OPTIMAL TRANSPORT – Regularized OT

Sinkhorn's Algorithm

$$\Lambda^{\varepsilon}_{\mathbf{C}}(\mathbf{P},\mathbf{f},\mathbf{g}) = \langle \mathbf{P},\mathbf{C}\rangle - \varepsilon H(\mathbf{P}) - \langle \mathbf{f}, \mathbf{P}\mathbb{1}_m - \mathbf{a}\rangle - \langle \mathbf{g}, \mathbf{P}^{\intercal}\mathbb{1}_n - \mathbf{b}\rangle$$

$$\mathbf{P}_{i,j} = e^{\frac{\mathbf{f}_i}{\varepsilon}} e^{-\frac{\mathbf{C}_{i,j}}{\varepsilon}} e^{\frac{\mathbf{g}_j}{\varepsilon}}, \qquad \forall (i,j) \in [\![n]\!] \times [\![m]\!]$$

$$\mathbf{P}_{i,j} = \mathbf{u}_i \mathbf{K}_{i,j} \mathbf{v}_i, \qquad \forall (i,j) \in [\![n]\!] \times [\![m]\!]$$

$$\mathbf{P}_{i,j} = \operatorname{diag}(\mathbf{u})\mathbf{K}\operatorname{diag}(\mathbf{v})$$

OPTIMAL TRANSPORT – Regularized OT

Sinkhorn's Algorithm

$$\Lambda_{\mathbf{C}}^{\varepsilon}(\mathbf{P},\mathbf{f},\mathbf{g}) = \langle \mathbf{P},\mathbf{C}\rangle - \varepsilon H(\mathbf{P}) - \langle \mathbf{f}, \mathbf{P}\mathbb{1}_m - \mathbf{a}\rangle - \langle \mathbf{g}, \mathbf{P}^{\intercal}\mathbb{1}_n - \mathbf{b}\rangle$$

$$\mathbf{P}_{i,j} = e^{\frac{\mathbf{f}_i}{\varepsilon}} e^{-\frac{\mathbf{C}_{i,j}}{\varepsilon}} e^{\frac{\mathbf{g}_j}{\varepsilon}}, \qquad \forall (i,j) \in [\![n]\!] \times [\![m]\!]$$

$$\mathbf{P}_{i,j} = \mathbf{u}_i \mathbf{K}_{i,j} \mathbf{v}_i, \qquad \forall (i,j) \in [\![n]\!] \times [\![m]\!]$$

$$\mathbf{P}_{i,j} = \mathrm{diag}(\mathbf{u})\mathbf{K}\mathrm{diag}(\mathbf{v})$$

Under the mass conservation constraints:

$$\mathbf{u} \circ (\mathbf{K}\mathbf{v}) = \mathbf{a}$$

$$\mathbf{v} \circ (\mathbf{K}^{\intercal}\mathbf{u}) = \mathbf{b}$$

OPTIMAL TRANSPORT – Regularized OT

Sinkhorn's Algorithm

$\mathbf{u} \leftarrow \mathbf{1}_n$
$\mathbf{v} \leftarrow \mathbf{1}_m$
$\mathbf{P} \leftarrow \mathrm{diag}(\mathbf{u})\mathbf{K}\mathrm{diag}(\mathbf{v})$
while P change **do**
  $\mathbf{u} \leftarrow \frac{\mathbf{a}}{\mathbf{K}\mathbf{v}}$
  $\mathbf{v} \leftarrow \frac{\mathbf{b}}{\mathbf{K}^T\mathbf{u}}$
  $\mathbf{P} \leftarrow \mathrm{diag}(\mathbf{u})\mathbf{K}\mathrm{diag}(\mathbf{v})$
end while

SUMMARY

I. Reminder on Kernel methods

II. Optimal Transport
1. Monge problem
2. Kantorovich formulation
3. **Regularized Optimal Transport**

III. Suggested Kernel
1. **Definitions**
2. Properties

IV. Applications

SUGGESTED KERNEL

Definition

Theorem 4. *Let* $F\colon [0,+\infty[\to \mathbb{R}$ *be a continuous function and* $\mathcal{U} \in \mathcal{P}_{SG}(\Omega)$*. If:*

1. $F \circ \sqrt{\cdot}$ *is completely monotonous on* $[0,+\infty[$
2. *There exist a nonnegative Borel measure* ν *on* $[0,+\infty[$ *such that for* $t > 0$, $F(t) = \int_0^{+\infty} e^{-ut^2}\, d\nu(u)$

Then

$$\begin{aligned} K\colon \mathcal{P}_{SG}(\Omega) \times \mathcal{P}_{SG}(\Omega) &\longrightarrow \mathbb{R} \\ (P,Q) &\longmapsto F\left(\|g_{\mathcal{U}}^{P} - g_{\mathcal{U}}^{Q}\|_{L^2(\mathcal{U})}\right) \end{aligned}$$

is a positive definite kernel on $\mathcal{P}_{SG}(\Omega)$.

SUMMARY

I. Reminder on Kernel methods

II. Optimal Transport
  1. Monge problem
  2. Kantorovich formulation
  3. **Regularized Optimal Transport**

III. **Suggested Kernel**
  1. Definitions
  2. **Properties**

IV. Applications

SUGGESTED KERNEL

Proprieties

Propositon 1. *Let* $s \in \mathbb{N}$. *Assume that* Ω *is compact and let* $\mathrm{P}, \mathrm{Q} \in \mathcal{P}(\Omega)$. *Then there exists a constant* c*, depending on the dimension, such that*

$$\|g_{\mathcal{U}}^{P} - g_{\mathcal{U}}^{Q}\|_{L^2(\mathcal{U})} \leq c \times \|P - Q\|_s$$

Remark: There exists another property than assures that, under specific assumptions, the distributions P and Q are equal if and only if

$$\|g_{\mathcal{U}}^{P} - g_{\mathcal{U}}^{Q}\|_{L^2(\mathcal{U})} = 0$$

SUGGESTED KERNEL

Proprieties

We consider two empirical distributions: $P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$ **and** $Q_m = \frac{1}{m}\sum_{i=1}^{m} \delta_{Y_i}$

Propositon 2. *If F is continuous, then*

$$F\left(\|g_{\mathcal{U}}^{P_n} - g_{\mathcal{U}}^{Q_m}\|_{L^2(\mathcal{U})}\right) \xrightarrow[n,m\to+\infty]{a.s} F\left(\|g_{\mathcal{U}}^{P} - g_{\mathcal{U}}^{Q}\|_{L^2(\mathcal{U})}\right)$$

Remark: This allows one to use this kernel on samples of distributions with theoretical guaranties.

SUMMARY

I. Reminder on Kernel methods

II. Optimal Transport
1. Monge problem
2. Kantorovich formulation
3. **Regularized Optimal Transport**

III. Suggested Kernel
1. Definitions
2. Properties

IV. Applications

APPLICATION

Generating the data

Training and test samples

$$(m_1, m_2) \sim (\mathcal{U}(-0.3, 0.3))^2$$

$$\sigma^2 \sim \mathcal{U}(0.0001, 0.0004)$$

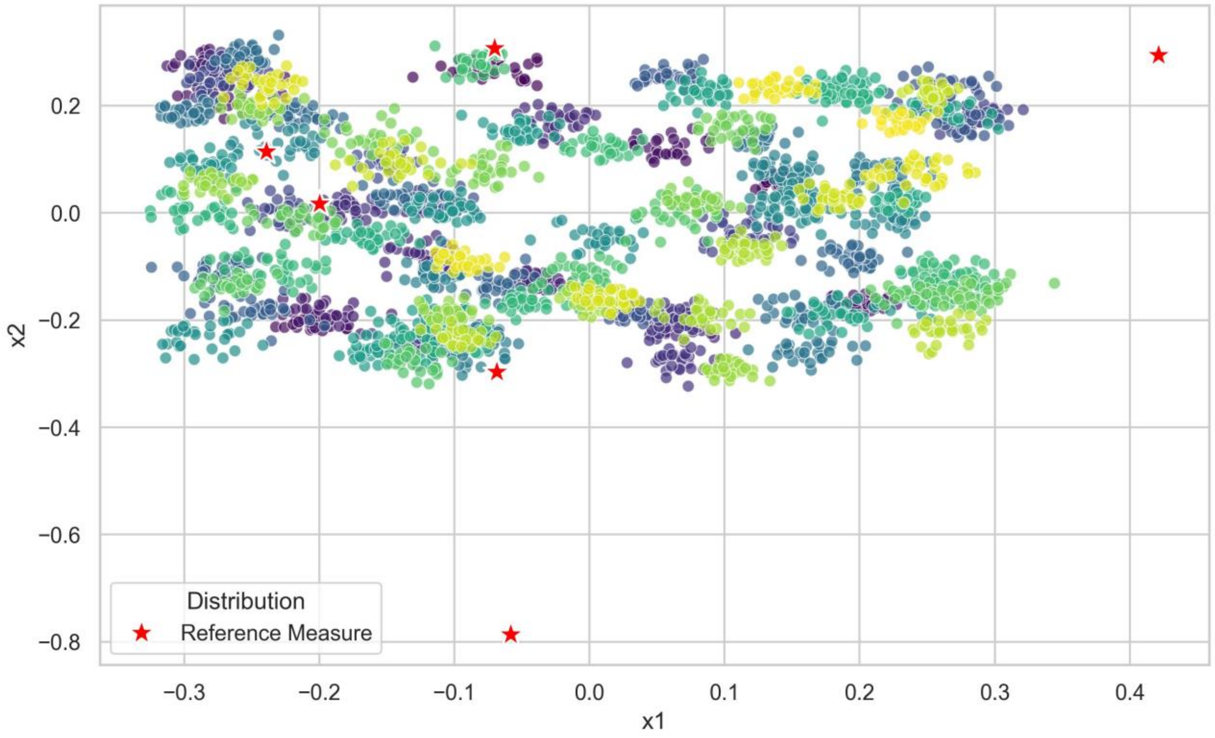
$$P \sim \mathcal{N}\left((m_1, m_2), \sigma^2 I_2\right)$$

$$Y = \frac{\left(m_1 + 0.5 - (m_2 + 0.5)^2\right)}{1 + \sigma}$$

Reference measure

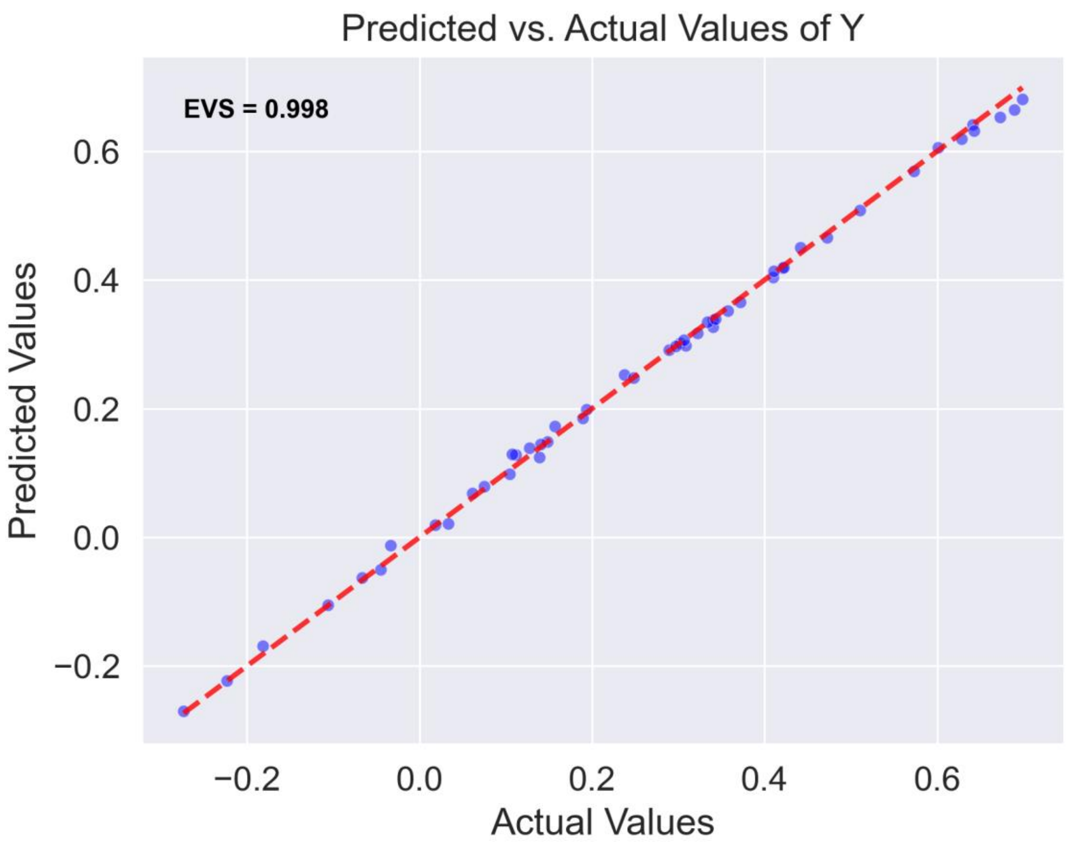
$$U \sim \mathcal{N}\left((0,0), 0.1 I_2\right)$$

Representation of all the training data and the reference measure

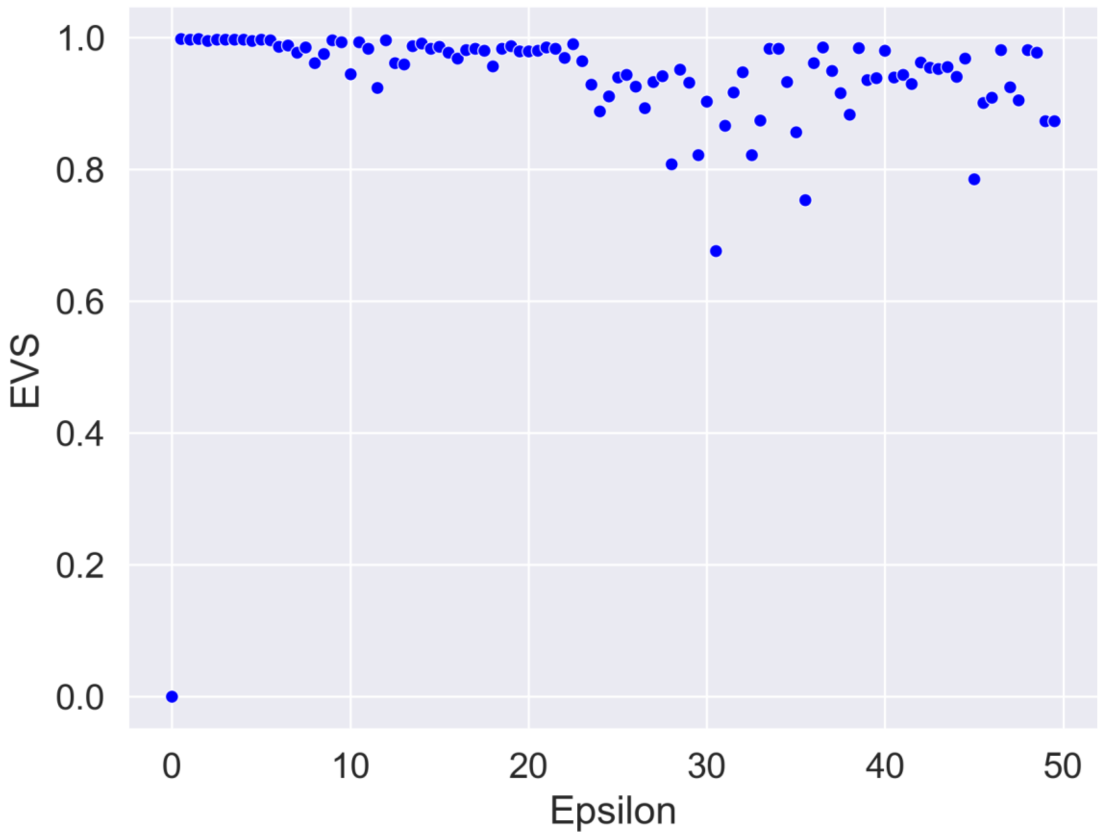


APPLICATION

Performing Kernel Ridge Regression



Performance of the KRR and the test set

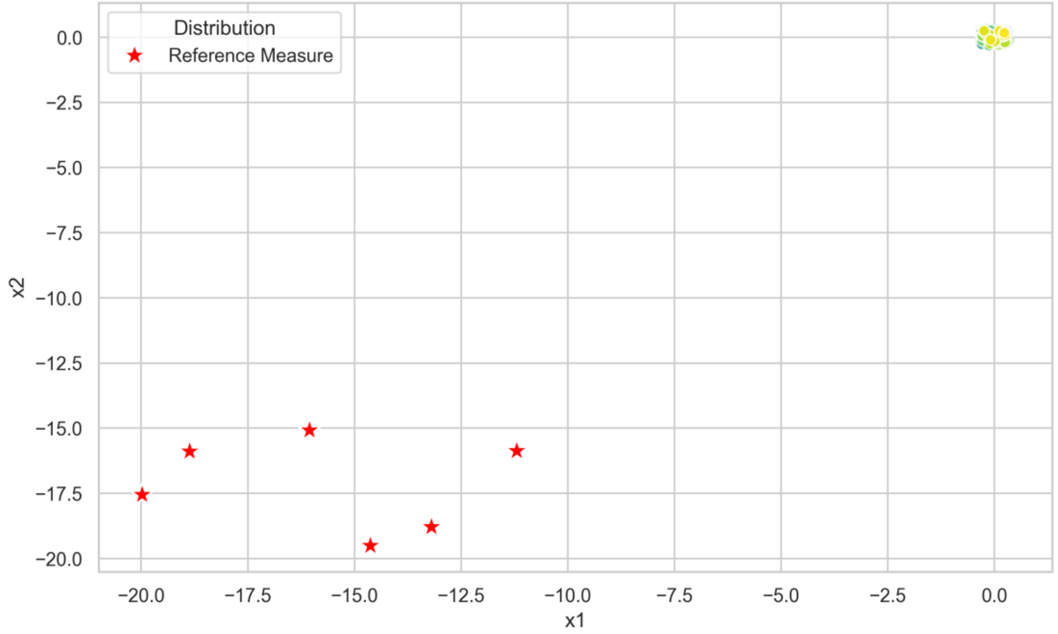


Performance of the KRR, on the test set, as a function of epsilon

APPLICATION

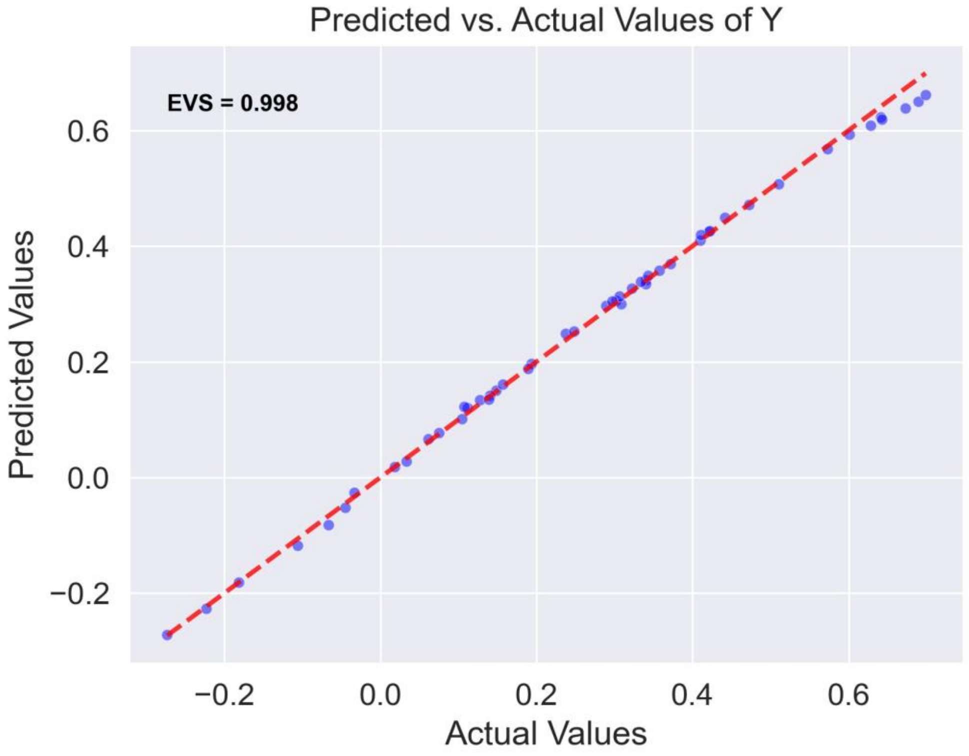
What happens with a different reference measure ?

Representation of all the training data and the reference measure



$$U \sim \mathcal{U}\,(-20,-10)^2$$

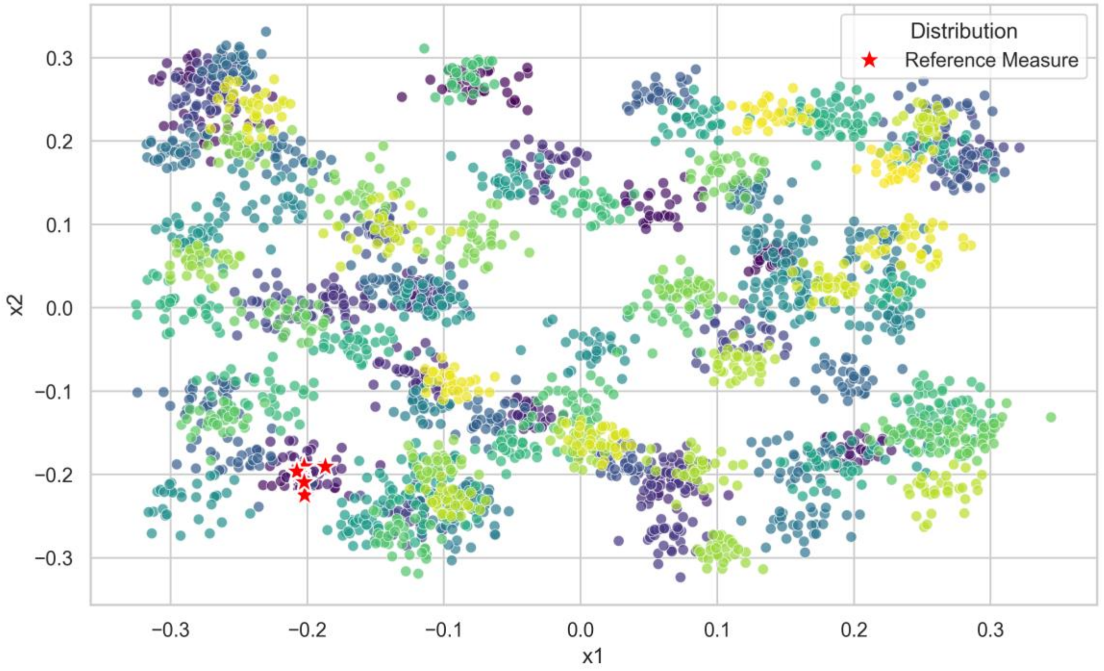
Performance of the KRR and the test set



APPLICATION

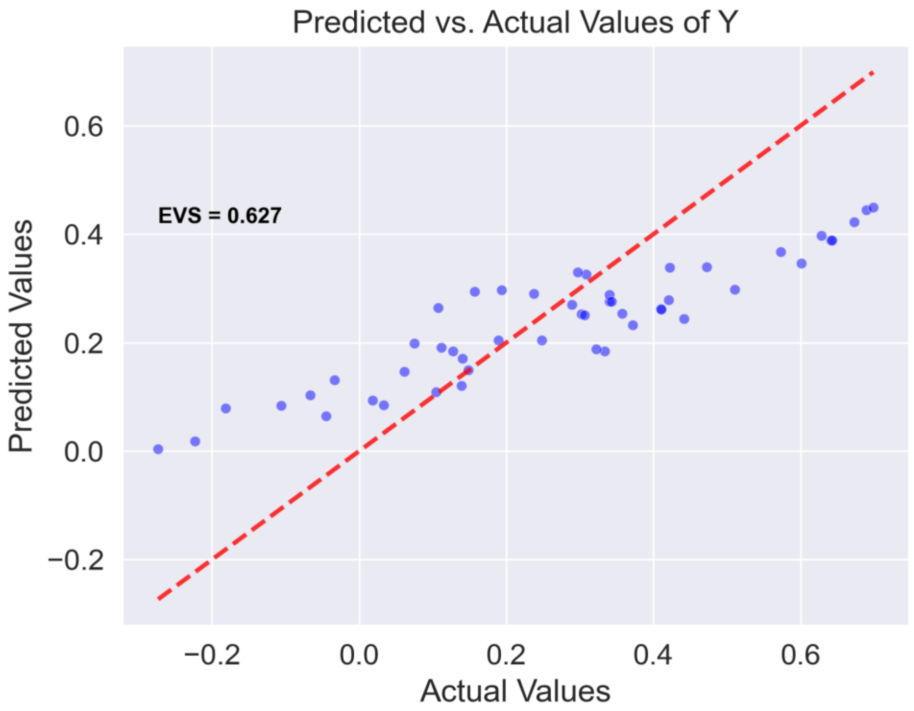
What happens with a different reference measure ?

Representation of all the training data and the reference measure



$$U \sim \mathcal{N}\left(\left(-0.2, -0.2\right), 0.0001 I_2\right)$$

Performance of the KRR and the test set



CONCLUSION

- Optimal transport is a great theoretical framework to compare distributions
- Defining a kernel on optimal transport object enables to compare distributions accurately
- All kernel methods are thereby available once such a kernel is constructed

- On a toy experiment we observe great performances (of the model and in computation term).
- Optimizing the reference measure is a key aspect to ensure good performances
- Computation does not scale well on bigger datasets

THANK YOU FOR YOUR ATTENTION