Learning functions defined over clouds of points with kernel methods

Phd Student: Babacar SOW

Contract duration: From 01/11/2021 to 31/10/2024

Project: ANR SAMOURAI

University: Ecole Des Mines de Saint-Etienne

Supervisors: Rodolphe LE RICHE (CNRS/LIMOS), Julien PELAMATTI, Merlin KELLER, Sanaa ZANNANE (EDF)

Table of contents

Context and problem formulation

Functions defined over clouds of points

- Metamodel functions assumed to be time consuming.
- In this presentation, we consider functions having inputs in the form of **bag of vectors** (or point clouds).
- These types of functions are encountered in many domains, such as: **image processing**, **design of experiments** and **optimization**, ...

In the following we consider the notations below:

- $\mathcal{X}$: space of all sets of n unordered points $\{\mathrm{x}_1, \dots, x_n\}$ where $x_i \in \mathbb{R}^d$, $i = 1, \dots, n$ and $n_{\text{min}} \leq n \leq n_{\text{max}}$.
- $X \in \mathcal{X}$ is a set of points and will be referred to as a **cloud of points**.

Mixed aspect: no order and varying size

Comparing two clouds of points with different sizes

The functions of interest are **permutation-invariant** with respect to their inputs.

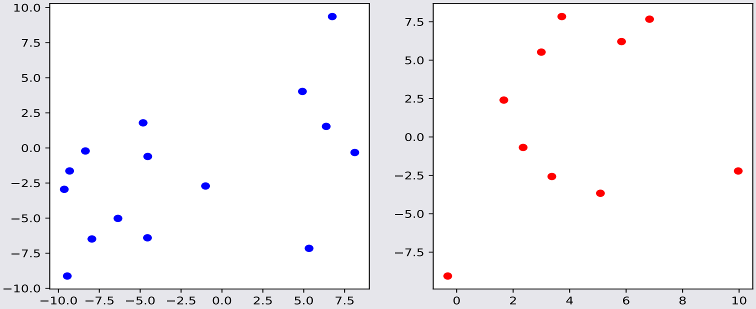

Figure: Two clouds of points in $d = 2$ dimensions with $n = 15$ points for the blue cloud and $n = 10$ points for the red one.

Example of a related industrial problem

A set of points model

- **Each point** (vector) represents the positions of a turbine.
- **The set of points** corresponds to the positions of all the turbines.
- Find an **optimal layout** of turbines minimizing the wake effects.

Use of kernels methods, related works and topics

Kernel methods

- Need of regression on such complex input functions.
- Use of kernel methods for their capacity of extending many statistical inference tools on non-vectorial data.

Learning functions defined over sets of objects with kernels

- Kernels on bags of vectors, applied to SVM Classification on images in [6].
- Same technique to define kernel on graphs by averaging over kernels between paths in [10] to measure similarity between shapes.
- Classification on text data with a set representation view in [11].
- A kernel between sets of points is used in [4] to optimize the layout of a wind farm.

The main content of this presentation

Focus of this presentation

- Deal with **varying-size** clouds of points as **the global input of interest**.
- Adopt Gaussian process regression: using kernel trick, closed form expression of lot of statistics (variance of prediction).
- Show **numerical performances of Gaussian processes** depending on the kernel on new test functions.
- Discuss the ability of extrapolation of the predictors: testing on rare data.

Semi-definite positive kernels

Feature Mapping, Aronszajn (1950)

Theoreme, Aronszajn [1]

k is a positive definite kernel if and only if there exists a Hilbert space $\mathcal{H}$, and a function $\phi : \mathcal{X} \longmapsto \mathcal{H}$ such that $\forall X, X', k(X, X') = \langle \phi(X), \phi(X') \rangle_{\mathcal{H}}$.

Substitution with Exponential

- Firstly, we consider covariance kernels of the form: $k(X, X') = \sigma^2 exp(-\frac{\Psi(X,X')}{2\theta^2})$.
- Semi-definite positiveness is equivalent to Ψ being **Hermitian** (symmetric in the real case) and **conditionally negative definite** [2].
- In other words, for any M distinct points and $c \in R^M$ with $\sum_{i=1}^{M} c_i = 0$, the following inequality must hold: $\sum_{i=1}^{M} \sum_{j=1}^{M} c_i c_j \Psi(X_i, X_j) \leq 0$
- We can define **Gaussian process over clouds of points** with any kernel satisfying the above conditions.

Modeling clouds of points

Through measures

Suppose we have two clouds $X = \{x_1, ..x_n\}$, $X' = \{x'_1, ..., x'_m\}$

- Define $\tilde{X} := P_X = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$ and $\tilde{X}' := P'_X = \frac{1}{m}\sum_{j=1}^{m} \delta_{x'_j}$, the respective **discrete measures**.
- Note that **this mapping is bijective** and there is no ambiguity in the modeling.

Through vectors

- $\tilde{X} := (w_1(X), ..., w_o(X))$ and $\tilde{X}' := (w_1(X'), ..., w_o(X'))$ can be two vectors of **characteristic features** of the clouds.

What distances between $\tilde{X}$ and $\tilde{X}'$ or mappings ?

With appropriate distances between $\tilde{X}$ and $\tilde{X}'$ (or mapping into an RKHS), we can define kernels between X and X'.

Kernel through measures and vectors

Wasserstein Distance

For two measures μ and ν defined over a space $\mathcal{M}$, the Wasserstein distance of positive cost function ρ and order p is defined as follows: $W_p^p = \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{M}\times\mathcal{M}} \rho(x,x')^p \mathrm{d}\pi(x,x')$. In the following ρ is the Euclidean distance and $p = 2$.

Sliced Wasserstein Distance (see Appendix)

- Let $\mathcal{S} = \{\alpha \in \mathbb{R}^2, ||\alpha|| = 1\}$. Consider the projected empirical measure of P_X on the line directed by $\alpha \in \mathcal{S}$ denoted $\alpha * P_X$ with: $\alpha * P_X = \frac{1}{n}\sum_{i=1}^{n} \delta_{<x_i,\alpha>}$
- $SW_2^2(P_X, P_{X'}) = \int_{\mathcal{S}} \mathcal{W}_2^2(\alpha * P_X, \alpha * P_{X'})\mathrm{d}\alpha$. Implementation using POT [5].
- The covariance kernel $k(X,X') = \sigma^2 exp(-\frac{SW_2^2(P_X,P_{X'})}{2\theta^2})$ is symmetric and semi-definite positive as in Carriere, Cuturi, and Oudot [3]. It will be denoted **SWS**.

Kernel through measures and vectors

Embedding based kernel: MMD and n-MeanMap

- Consider the embedding map into an RKHS $\mathcal{H}$ (equipped with $k_{\mathcal{H}}$) as defined in Muandet et al. [8] $P_X \longmapsto \mu_X(.) = \int P_X(x) k_{\mathcal{H}}(x, .)\mathrm{d}x$.
- $k(X, X') = \langle \frac{\mu_X}{||\mu_X||}, \frac{\mu_{X'}}{||\mu_{X'}||} \rangle_{\mathcal{H}}$ is a semi-definite positive (s.d.p) kernel denoted **n-MeanMap**.
- It is the same case with $k(X, X') = \sigma^2 exp(-\frac{||\mu_X - \mu_{X'}||^2_{\mathcal{H}}}{2\theta^2})$ denoted as **MMD**.
- Note that $k_{\mathcal{H}}$ is defined over the space of the vectors. It can belong to Matèrn kernels family for instance.

Vector-based, relevant-features kernel

- We consider a last kernel of the form $k(X, X') = \sigma^2 \exp\left(- \sum_{j=1}^{o} \frac{|w_j(X) - w_j(X')|^2}{\theta_j'^2} \right)$ with $(w_1(X), ..., w_o(X))$ a **vector of features**. Among the features we can have the mean, the diameter, the number of points or spectral information. It is denoted **RFK**.

Design of experiments, learning process and test clouds

Random cloud design and learning process

The **random cloud design of experiments** is implemented as follows:

- The size of the design is chosen proportionally to the average of the cloud of points sizes.
- The size of each cloud is randomly picked in $n \in \{n_{\mathrm{min}}, \ldots, n_{\mathrm{max}}\}$ and each point is uniformly sampled in the domain of the function.
- The hyper-parameters of the kernels are found by maximizing the log-likelihood of the observed (design) data.

Test on random (normal) and geometrically modified clouds of points (extrapolation)

- The random test is of the same type as the design. The size of test clouds is 1000.
- To assess the exploratory abilities of the kernels, we evaluate their prediction performance on clouds of points modified geometrically with dilation and rotation.

Illustration of geometrical transformation

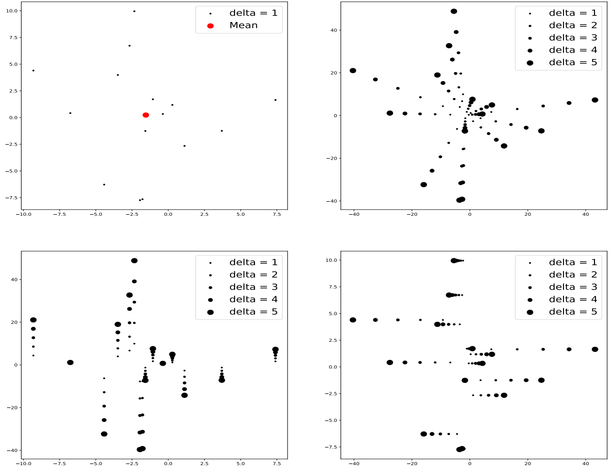

Figure: Illustration of the dilation transformation of clouds: initial cloud at top left with its mean (red bullet), the 5 isotropic dilations at top right, 5 vertical dilations at bottom left and 5 horizontal dilations at bottom right. Note that the horizontal and vertical ranges vary between the plots.

Test function

Inspired from wind-farms (see Appendix for other test functions)

Mimicking wind farms

- We consider the following family of test functions mimicking wind-farms productions

$$F_\theta(\{\mathrm{x_1}, \ldots, \mathrm{x_n}\}) = \sum_{i=1}^{n} \Big(\prod_{j, j \neq i} f_{x_j,\theta}(\mathrm{x_i}) \Big) f_0(\mathrm{x_i}) \tag{1}$$

  where :

  - $f_{x_j,\theta}(\mathrm{x_i})$ expresses the energy loss over $\mathrm{x_i}$ that is caused by $\mathrm{x_j}$. Its parametrized by x_j and $\theta \in (0, 2\pi)$ (the direction of wind)
  - f_0 is a constant and corresponds to maximal production of x_i (if it was alone)
  - $\mathrm{x_i} \in \mathbb{R}^2$ and $n \in \{10, 11, .., 20\}$

Some examples of $f_{x,\theta}(.)$

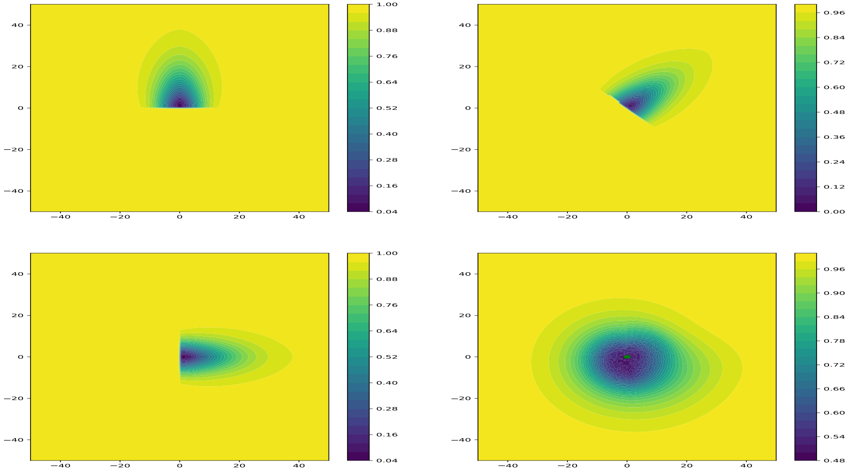

Figure: Representation of f_p with $\theta = 90^\circ$ at top left, $\theta = 45^\circ$ top right, $\theta = 0^\circ$ bottom left, and averaged directions at bottom right. We denote the corresponding functions respectively F_{90}, F_{45}, F_0, F_{40d}.

Q^2 values on wind-farm proxy: random and rotated clouds of points.

The Q^2 are high on random and rotated clouds of points.

Function \ Kernels	MMD	n-MeanMap	RFK	Slice-Wass
F_0	**0.906**	0.647	0.897	0.828
F_{45}	0.868	0.623	0.893	0.821
F_{90}	**0.899**	0.639	0.871	0.843
F_{40d}	**0.906**	0.734	0.799	0.824

Table: Q^2 of 4 kernels on all the wind farm proxy functions, the testing clouds come from a random design.

Function \ Kernels	MMD	RFK	Slice-Wass
F_0	0.808	**0.863**	0.780
F_{45}	0.780	**0.877**	0.802
F_{90}	0.800	**0.881**	0.797
F_{40d}	0.701	0.771	**0.775**

Table: Q^2 observed on rotated clouds of points: lot of **RFK** features are rotation-invariant.

Q^2 values on wind-farm proxy: isotropic and horizontal dilation

Function \ Kernels	MMD	RFK	Slice-Wass
F_0	0.933	0.952	0.893
F_{45}	0.939	0.954	0.933
F_{90}	0.942	0.931	0.879
F_{40d}	0.940	0.974	0.975

Table: Q^2 observed on isotropically dilated clouds of points

Function \ Kernels	MMD	RFK	Slice-Wass
F_0	**0.05**	-15.535	-10.033
F_{45}	**0.519**	-0.879	0.397
F_{90}	**0.518**	0.711	0.631
F_{40d}	**0.103**	-2.415	-0.827

Table: Q^2 observed on horizontally dilated clouds of points

Note the poor performances yielded on horizontally dilated clouds of points !

Predictors vs functions: horizontal dilation

The functions do not vary a lot with horizontal dilations.

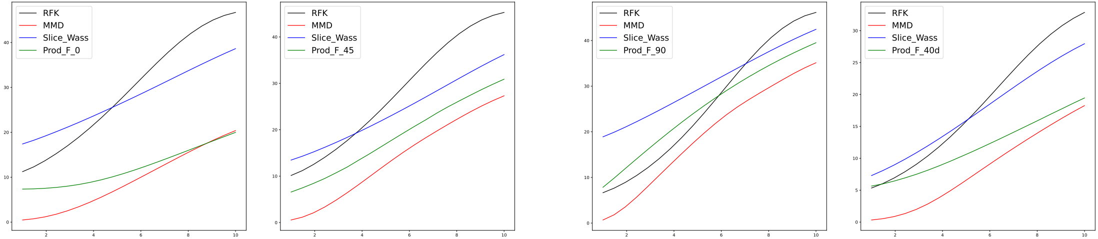

Figure: Wind farm proxy outputs as a function of the horizontal dilation δ: function output in green, Gaussian process with RFK, MMD and Slice-Wass kernels in black, red, and blue. Wind orientations are 0°, 45°, 90° and the 40 directions (i.e., F_0, F_{45}, F_{90}, F_{40d}) from left to right and top to bottom. The curves are averaged over 50 clouds.

MMD-based kernels: hyper-parameters adaptation

MMD-based kernels adapt to geometrical properties of wind-farms functions through the hyper-parameters of the embedding kernel.

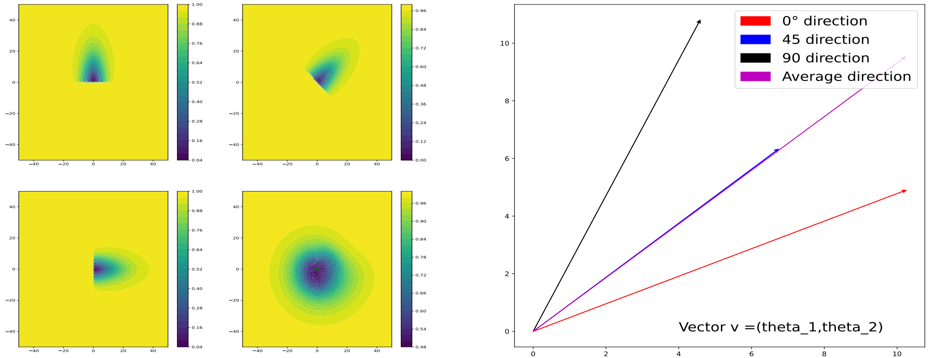

Figure: Vectors of length scales of the MMD embedding Matérn 5/2 kernel learned by maximum likelihood on the wind farm proxy for various wind directions. Left: reminder of the turbine contributions for winds at 90°,45°,0° and 40 directions (left to right, top to bottom). Right: $(\theta_1, \theta_2)^\top$ vectors of length scales of the embedding kernel.

Conclusions

Modeling a cloud as a discrete measure

- Modeling a cloud as a discrete measure helps having more possibilities of defining kernels and can yield interpretable results.

Different kernels

- MMD based kernels yield the best results on many examples based on their embedding representation.
- In clouds of points context, MMD based kernels seem to be more adapted to functions with different directions of variations whereas the others are not.
- The extrapolation shows that based on the anisotropic variation of functions, the performances of prediction are very different.

Design of experiments

- It can be interesting to define new criteria for design of experiments over clouds of points depending on the application.

Perspectives

- Test for other dimensions $d \geq 3$.
- Study the size of the design vs performances.
- Define criteria of design of experiments over clouds of points.
- Extend metamodeling to other related problems such as Bayesian optimization.

Thanks For Your Attention !

Bibliography I

[1] Nachman Aronszajn. "Theory of reproducing kernels". In: *Transactions of the American mathematical society* 68.3 (1950), pp. 337–404.

[2] Christian Berg, Jens Peter Reus Christensen, and Paul Ressel. *Harmonic analysis on semigroups: theory of positive definite and related functions*. Vol. 100. Springer, 1984.

[3] Mathieu Carriere, Marco Cuturi, and Steve Oudot. "Sliced Wasserstein kernel for persistence diagrams". In: *International conference on machine learning*. PMLR. 2017, pp. 664–673.

[4] Tinkle Chugh and Endi Ymeraj. "Wind Farm Layout Optimisation using Set Based Multi-objective Bayesian Optimisation". In: *arXiv preprint arXiv:2203.17065* (2022).

[5] Rémi Flamary et al. "Pot: Python optimal transport". In: *Journal of Machine Learning Research* 22.78 (2021), pp. 1–8.

Bibliography II

[6] Philippe H Gosselin, Matthieu Cord, and Sylvie Philipp-Foliguet. "Kernels on bags for multi-object database retrieval". In: *Proceedings of the 6th ACM international conference on Image and video retrieval*. 2007, pp. 226–231.

[7] Soheil Kolouri, Yang Zou, and Gustavo K Rohde. "Sliced Wasserstein kernels for probability distributions". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 5258–5267.

[8] Krikamol Muandet et al. "Kernel mean embedding of distributions: A review and beyond". In: *Foundations and Trends® in Machine Learning* 10.1-2 (2017), pp. 1–141.

[9] Gabriel Peyré, Marco Cuturi, et al. "Computational optimal transport". In: *Center for Research in Economics and Statistics Working Papers* 2017-86 (2017).

[10] Frédéric Suard, Alain Rakotomamonjy, and Abdelaziz Bensrhair. "Kernel on Bag of Paths For Measuring Similarity of Shapes.". In: *ESANN*. Citeseer. 2007, pp. 355–360.

Bibliography III

[11] Yuya Yoshikawa et al. "Cross-domain matching for bag-of-words data via kernel embeddings of latent distributions". In: *Advances in Neural Information Processing Systems* 28 (2015).

Distance between laws: Wasserstein Distance

Substitution with Hilbertian distance: Wasserstein Distance in 1D Case

- Definition and properties see Carriere, Cuturi, and Oudot [3] and Kolouri, Zou, and Rohde [7]
- Let μ and ν be two nonnegative measures in $\mathbb{R}$ with $\mu(\mathbb{R}) = \nu(\mathbb{R}) = 1$. The Wasserstein distance of order 2 between μ and ν is defined as folllows:

$$\mathcal{W}_2^2(\mu,\nu) = \inf_{P \in \Pi(\mu,\nu)} \int\int_{\mathbb{R}\times\mathbb{R}} |x - x'|^2 P(dx, dx')$$

- Let $\mathcal{C}_\mu(x) = \int_{-\infty}^{x} d\mu$, $\mathcal{C}_\nu(x) = \int_{-\infty}^{x} d\nu$ their cumulative distribution function.
- Pseudo-inverse: $\forall r \in [0,1], \mathcal{C}_\mu^{-1}(r) = \min_x \{x \in \mathbb{R} \cup \{-\infty\} : \mathcal{C}_\mu(r) \geq x\}$
- Then $\mathcal{W}_2^2(\mu,\nu) = ||\mathcal{C}_\mu^{-1} - \mathcal{C}_\nu^{-1}||^2_{L^p([0,1])}$, see Peyré, Cuturi, et al. [9]
- $\mathcal{W}_2^2(\mu,\nu)$ is symmetric and conditionally negative definite. (Kolouri, Zou, and Rohde [7])
- If μ and ν are defined in $\mathbb{R} \times \mathbb{R}$, the above condition is no longer guaranteed.

Other test functions

Mindist and Inertia

Mindist Function: returns the shortest distance between points as value

- $F_{minDist}(\{\mathrm{x}_1, ..., \mathrm{x}_\mathrm{n}\}) = \min_{i \neq j} ||\mathrm{x}_i - \mathrm{x}_j||$.

Inertia function

- $F_{inert}(\{\mathrm{x}_1, ..., \mathrm{x}_\mathrm{n}\}) = \sum_{i=1}^{n} ||\mathrm{x}_i - \bar{X}||^2$ with $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{x}_\mathrm{i}$

Geometrically modified clouds of points

For a given cloud of points $X = \{x_1, ..., x_n\}$, we note respectively, X_r^θ, X_d^δ, X_{dh}^δ, X_{dv}^δ its rotated, isotropically dilated, horizontally dilated and vertically dilated transformations. We have

$$\begin{aligned}
X_r^\theta &= \{R_\theta x_1 + (I - R_\theta)\bar{X}, ..., R_\theta x_n + (I - R_\theta)\bar{X}\} \ , \\
X_d^\delta &= \{D_\delta x_1 + (I - D_\delta)\bar{X}, ..., D_\delta x_n + (I - D_\delta)\bar{X}\} \ , \\
X_{dh}^\delta &= \{D_{\delta h} x_1 + (I - D_{\delta h})\bar{X}, ..., D_{\delta h} x_n + (I - D_{\delta h})\bar{X}\} \ , \\
X_{dv}^\delta &= \{D_{\delta v} x_1 + (I - D_{\delta v})\bar{X}, ..., D_{\delta v} x_n + (I - D_{\delta v})\bar{X}\} \ .
\end{aligned}$$

Rotations and dilations are done with respect to the point cloud means, $\bar{X}$. In addition,

$$R_\theta = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}, D_\delta = \begin{bmatrix} \delta & 0 \\ 0 & \delta \end{bmatrix}, D_{\delta h} = \begin{bmatrix} \delta & 0 \\ 0 & 1 \end{bmatrix}, D_{\delta v} = \begin{bmatrix} 1 & 0 \\ 0 & \delta \end{bmatrix} \ ,$$

where θ and δ are the rotation and dilation factors.

Q^2 values on Inertia and Mindist: random, dilated and rotated clouds of points

Function \ Kernels	MMD	n-MeanMap	RFK	Slice-Wass
F_{inert}	0.734	0.506	0.988	0.905
$F_{minDist}$	-0.051	0.035	0.997	0.587

Table: Summary of the Q^2 observed on $F_{minDist}$ and F_{inert}

Function \ Kernels	MMD	RFK	Slice-Wass
F_{inert}	0.901	0.982	0.845
$F_{minDist}$	-0.802	0.998	0.280

Table: Summary of the Q^2 observed on $F_{minDist}$ and F_{inert}

Function \ Kernels	MMD	RFK	Slice-Wass
F_{inert}	0.422	0.988	0.854
$F_{minDist}$	-0.025	0.998	0.206

Table: Q^2 observed on rotated clouds of points for the F_{inert} and $F_{minDist}$ functions.