Kernel over sets of vectors.

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), Merlin KELLER (EDF), Sanaa ZANNANE (EDF), Julien PELAMATTI (EDF)

Table of contents

1. Context and Problem
2. Bayesian Approach and Kernels between clouds of points
3. Focus on MMD
4. Perspectives
5. Bibliography

Context and problem formulation

Functions defined over sets of vectors

In this presentation, we consider functions having inputs in the form of sets of vectors (or points). 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 $n_1 \leq n \leq n_2$.
- $X \in \mathcal{X}$ is a set of points and will be referred to as a cloud of points. Compared to an (ordered) list of points, X is invariant with respect to any point permutation.
- $\mathcal{F}$: set of functions over clouds of points, $F \ : \ \mathcal{X} \to \mathbb{R}$, $X \mapsto F(X)$.

Related works and topics

Learning functions defined over sets of objects with kernels

- Kernels on bags of vectors, applied to SVM Classification on images in [7].
- Same technique to define kernel on graphs by averaging over kernels between paths in [13] to measure similarity between shapes.
- Classification on text data with a set representation view in [14].
- A Kernel between sets of points is used in [5] to optimize the layout of a wind farm.

Focus of this presentation

- In this presentation,we discuss some general methods to construct such kernels.
- Compare them numerically on a a test function mimicking the production of a windfarm.

Bayesian Approach

A Gaussian process prior

- Gaussian processes are defined by a mean function m and a covariance kernel k over the input spaces $\mathcal{X}$; it can be used as a prior law to approximate a costly function $F \in \mathcal{F}$.
- Observing $D = \{(X_1, y_1)...(X_N, y_N)\}$ where $X_i \in \mathcal{X}$ and $y \in \mathbb{R}$ as training data with $y_i = f(X_i)$, the predictive mean and covariance $F(X)$ at a new point X are given by:

$$\mu(X; D) = m(X) + K(X, X)^T K(\mathbb{X}, \mathbb{X})^{-1}(y - m(\mathbb{X}))$$

$$\Sigma(X, X; D) = K(X, X) - K(\mathbb{X}, X)^T K(\mathbb{X}, \mathbb{X})^{-1} K(\mathbb{X}, X)$$

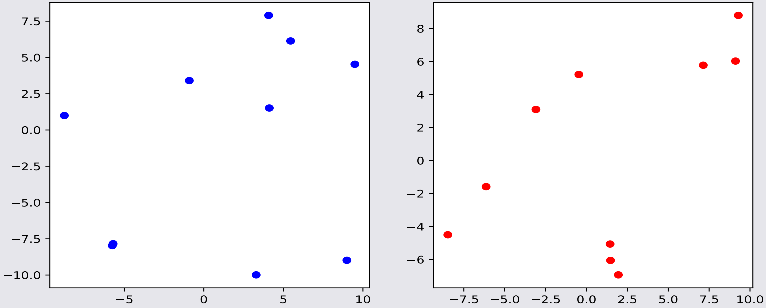
with $\mathbb{X} = [X_1, ..., X_N]$ and $y = [y_1, ..., y_N]$

Necessary Conditions on k

- k must be symmetric and positive definite, i.e, for any M distinct clouds of points, for any vector $c \in \mathbb{R}^M$, the following inequality must hold: $\sum_{i=1}^{M} \sum_{j=1}^{M} c_i c_j k(X_i, X_j) \geq 0$

Bayesian approach: kernel trick and mapping

Comparing two clouds of points



Aronszajn Theorem, explicit and implicit mappings

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 $\forall X, X', k(X, X') = \langle \phi(X), \phi(X') \rangle_{\mathcal{H}}$.

Explicit and implicit mappings

- Explicit mapping: in some cases ϕ and the scalar product, $\langle ., . \rangle_{\mathcal{H}}$ are known by definition or by construction
- Implicit mapping : in some cases, we just use the compact formula of k

Substitution with Hilbertian Distance

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 semi-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: $\sum_{i=1}^{M} \sum_{j=1}^{M} c_i c_j \Psi(X_i, X_j) \leq 0$

Metric Cases

- We consider cases where $\Psi(X, X') = d(\tilde{X}, \tilde{X}')^2$
- d is the distance between $\tilde{X}$ and $\tilde{X}'$ the respective images of X and X' into a known metric Space.
- The above conditions are equivalent to ensuring that the metric be **Hilbertian**, as stated in Haasdonk and Bahlmann [8].

How to construct $\tilde{X}$ and $\tilde{X}'$?

With probabilities

Suppose we have two clouds $X = (x_1, ..x_n)$, $X' = (x'_1, ..., x'_m)$

- Case 1 : 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 associated empirical uniform distributions.
- Case 2 : Define $\tilde{X} = \mathcal{N}_{\mathcal{X}}(m_X, \Sigma_X)$ and $\tilde{X}' = \mathcal{N}'_{\mathcal{X}}(m'_X, \Sigma'_X)$ with $m_X = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\Sigma_X = \frac{1}{n}\sum_{i=1}^{n}(x_i - m_X)(x_i - m_X)^T$ and likewise for m'_X, Σ'_X.

With vectors : vectorization

- $\tilde{X}$ and $\tilde{X}'$ can be two vectors of characteristic features of the clouds.

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

We discuss in the following the candidates distances to define $d(\tilde{X}, \tilde{X}')$?

Sliced Wasserstein Distance and Gaussian approximation

Wasserstein Distances

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 : $W_p^p = \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{M}\times\mathcal{M}} \rho(x,x')^p \mathrm{d}\pi(x,x')$

Substitution with Hilbertian distance : Sliced Wasserstein Distance (see Annex)

- 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 [6].
- The covariance kernel $k(X,X') = \sigma^2 exp(-\frac{SW_2^2(P_X,P_{X'})}{2\theta^2})$ is symmetric and positive semi-definite as in Carriere, Cuturi, and Oudot [4]. It will be denoted $\mathbf{k}_{SWS}$.

Approximate For Gaussian Modeling (see Annex) , $\mathbf{k}_{SWG}$

$W_2^2 \approx ||m_X - m_{X'}||^2 + ||\Sigma_X^{1/2} - \Sigma_{X'}^{1/2}||^2_{Frobenius}$ as in Bui et al. [3] (= if $\Sigma_X^{1/2}\Sigma_{X'}^{1/2} = \Sigma_{X'}^{1/2}\Sigma_X^{1/2}$

Distance between embedded laws : Maximum Mean Discrepancy

Substitution with Hilbertian distance: MMD

- Suppose there exists a Reproducing Kernel Hilbert Space, $\mathcal{H}$ with a characteristic kernel.
- The characteristic nature guarantees the injectivity of the embedding map Muandet et al. [11]: $P_X \longmapsto \mu_X(.) = \int P_X(x) k_{\mathcal{H}}(x,.)\mathrm{d}x$.
- $MMD^2(P_X, P_{X'}) = ||\mu_X - \mu'_X||^2_{\mathcal{H}}$
- For any kernel $k_{\mathcal{H}}$ of the RKHS, and any uniform discrete laws: $MMD^2(P_X, P_{X'}) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} k_{\mathcal{H}}(x_i, x_j) + \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} k_{\mathcal{H}}(x'_i, x'_j) - 2\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} k_{\mathcal{H}}(x_i, x'_j)$
- The covariance kernel $k(X, X') = \sigma^2 exp(-\frac{||\mu_X - \mu_{X'}||^2_{\mathcal{H}}}{2\theta^2})$ is symmetric and positive definite.
- We will denote the latter as $\mathrm{k_{MMD}}$.

Constructing Features of a cloud

Relevant Features Map Kernel

- We consider a final kernel of the form $k(X, X') = \sigma^2 \exp\big(-\sum_{j=1}^{o} \frac{|w_j(X)-w_j(X')|^2}{\theta_j'^2}\big)$ with $(w_1(X), ..., w_o(X))$ a vector of features.
- As features we consider:
  - The coordinates of the mean
  - the eigenvalues and eigenvectors of the empirical covariance matrix.
  - the number of points in the set
  - Greatest and shortest distances between points of the set.
- This kernel will be called **Relevant Feature Kernel** and denoted k_{RFK}

Explicit Mappings: Probability Product Kernels and Embeddings

Explicit Mappings (see Annex)

- Recall $k(X, X') =< \phi(X), \phi(X') >$
- We consider first a case where the mapping $\phi(X) = P_X^\rho$ with $\rho \in]0, 1]$ where P_X is an underlying empirical distribution.
  - A possible kernel: $k(X, X') = \int_\Omega P(x)^\rho P'^\rho(x) \mathrm{d}x$, Jebara and Kondor [9]. This family of kernels are called Probability Product Kernels. For two Gaussians $P_X = \mathcal{N}(\mu, \Sigma)$ and $P_{X'} = \mathcal{N}(\mu', \Sigma')$, one gets:

    $$k(X, X') = (2\pi)^{(1-2\rho)D/2} |\Sigma^+|^{1/2} |\Sigma|^{-\rho/2} |\Sigma|^{-\rho/2} \exp\big(-\frac{\rho}{2}\mu^\top\Sigma^{-1}\mu - \frac{\rho}{2}\mu'^\top\Sigma'^{-1}\mu' + \frac{1}{2}\mu^{+\top}\Sigma^{+\top}\mu^+$$

    where $\Sigma^+ = (\rho\Sigma^{-1} + \rho\Sigma^{-1})^{-1}$ and $\mu^+ = \rho\Sigma^{-1}\mu + \rho\Sigma'^{-1}\mu'$
  - If $\rho = \frac{1}{2}$, it is called the **Bhattacharrya Kernel** and when $\rho = 1$ Expected Likelihood Kernel.
- $\phi(X) = \mu_X$ where μ_X is the embedding of the underlying empirical distribution into an RKHS. $k(X, X') =< \mu_X, \mu_{X'} >$ it will be called **MMK**, Mean Map Kernel and denoted k_{MMK} for the remainder.

A test function

Mimicking wind farms

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

$$F(\{x_1, ..., x_n\}) = \sum_{i=1}^{n} \left(\prod_{j, j \neq i} f_p(x_j, x_i) \right) f_0(x_i) \tag{1}$$

  where $f_p(x_j, x_i)$ expresses the energy loss over x_i that is caused by x_j and f_0 is a constant. $x_i \in \mathbb{R}^2$ and $n \in \{10, 11, .., 20\}$

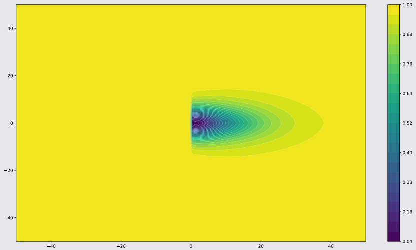
- The function $x_i \longmapsto f_p(x_j, x_i)$ can be parameterized differently:
  - It can be unidirectional with an arbitrary angle.
  - It can be multi-directional

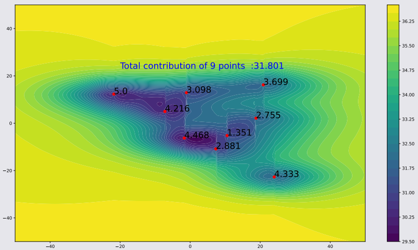
A test function

Mimicking wind farms :Example

In the following we represent: $x_i \longmapsto f_p(x_0, x_i)$ on the left, f with a one varying point on the right. We note F with f_p on left F_0.

Mimicking wind farms : Illustration



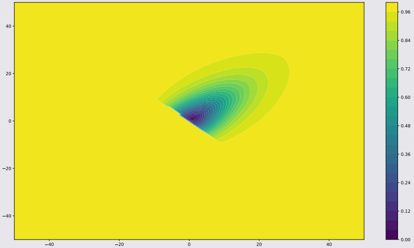


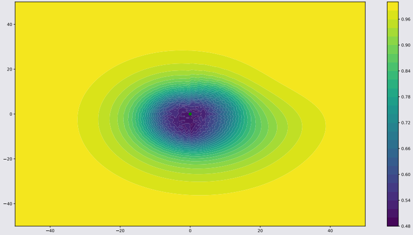
A test function

Mimicking wind farms : Example

In the following we represent: $x_i \longmapsto f_p(x_0, x_i)$ with $\pi/4$ rotated direction, and 40 directions on the right.We note F with f_p on left F_{45} and F_{40d} for the f_p on the right.

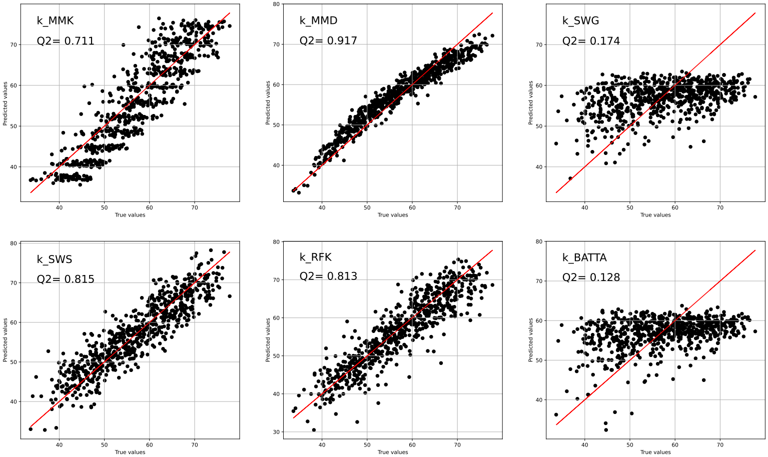
Mimicking wind farms : Illustration





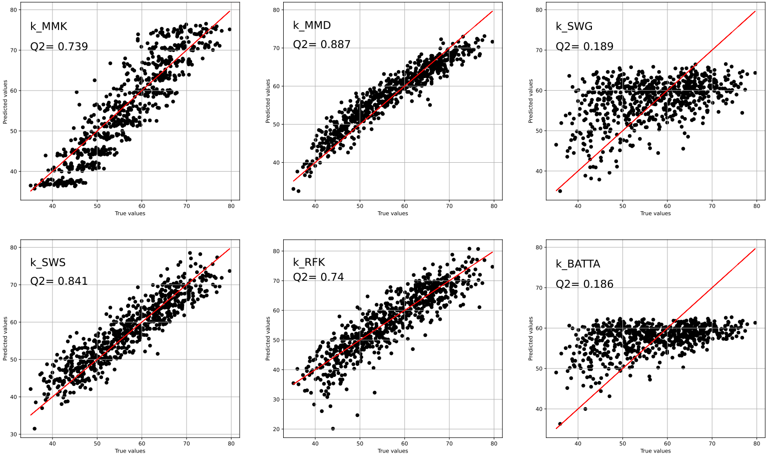
Preleminary Results: 0°Interaction Function

- Modeling with Gaussians distributions is weaker than with discrete uniform ones for this function.
- Sliced Wasserstein Kernel is very competitive with Relevant Feature Kernel. MMD works best!



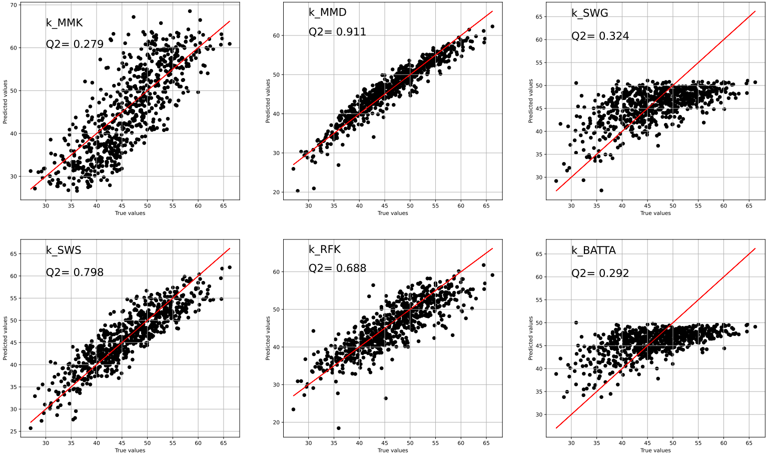
Results: 45° direction Interaction

- 45° direction does not change performance for lot of kernels but Feature Map Kernel .



Preleminary Results: 40 directions integrated

- 40 directions integrated Function improves slightly Gaussian based kernels.
- MMD shows better results than Relevant Feature kernel and Sliced Wasserstein



Summary

Table: Summary of the Q2 observed : Battacha refers to Bhattacharrya kernel, RFK (Relevant Feture kernel), SWS (Sliced Wasserstein subs), GWS (Gaussian Wasserstein subs)

Kernels Function	MMD	MMK	Battacha	RFK	SWS	GWS
F_0	0.917	0.711	0.144	0.813	0.812	0.174
F_{45}	0.887	0.739	0.186	0.74	0.841	0.189
F_{40d}	0.88	0.279	0.314	0.688	0.798	0.259

- MMD-based kernels remain the most robust. MMK fails to model a lot of directions integrated.
- Modeling clouds as Gaussian distributions performs poorly when dealing with discrete uniform modelization.
- SWS and RFK are very competitive with MMD.

Focus on MMD :The choice of the embedding kernel $k_{\mathcal{H}}$

- We have $P_X \longmapsto \mu_X(.) = \int P_X(x) k_{\mathcal{H}}(x,.)\mathrm{d}x$
- Candidates $k_{\mathcal{H}}$: Squared Exponential(S Exp), Exponential(Exp), Matern32, Matern52
- Test performance predictions on unseen clouds after training.
- Same methodology from a lower dimension has not a great effect .

Table: Result about the influence of $k_{\mathcal{H}}$

$k_{\mathcal{H}}$	S Exp	Exp	Matern32	Matern52
Q_2 on F_0	0.894	0.917	0.911	0.906

Making k_{MMD} translation invariant

- We make k_{MMD} invariant under translation.
- For this we center the clouds before computing k_{MMD}.
- The objective is to force k_{MMD} to be as F_0
- The results are the following with different $k_{\mathcal{H}}$
- The performances are approximately the same.

Table: Results : centered (c) vs non-centered (nc)

$k_{\mathcal{H}}$	S Exp	Exp	Matern32	Matern52
Q_2 on F_0 (c)	0.894	0.917	0.911	0.906
Q_2 on F_0 (nc)	0.899	0.912	0.911	0.908

Embedding sensitivity and F_0

- We take $k_{\mathcal{H}}(x_i, x_j) = \exp\left(-\frac{|x_{i,1} - x_{j,1}|}{2\theta_1^2} - \frac{|x_{i,2} - x_{j,2}|}{2\theta_2^2} \right)$
- We compare the sensitivity of $\langle \mu_X, \mu_X \rangle$ with respect to F_0 concerning horizontal and vertical dilatations.
- The hyper-parameters are estimated by maximizing Log-likelihood allow $\langle \mu_X, \mu_X \rangle$ to behave differently under horizontal and vertical dilatations.

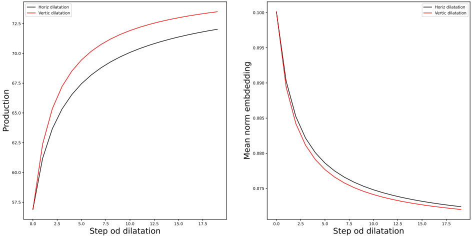


Figure: Representation of the sensitivity of F_0 (left) and Norm of Embedding (MMD) (right) with respect to horizontal and vertical expansion of the clouds. The average observations over 300 clouds is represented.

Embedding sensitivity and F_{45}

- The particularity of F_{45} is that the angle of interaction is $\pi/4$.
- The function has the same sensitivity to horizontal and vertical dilatations of clouds.
- We find anew this property in $\langle \mu_X, \mu_X \rangle$

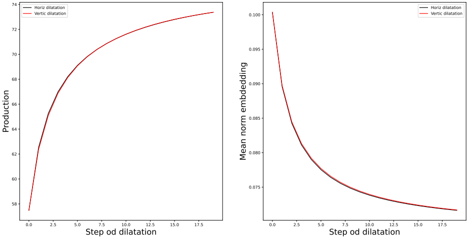


Figure: Representation of the sensitivity of F_{45} (left) and Norm of Embedding (MMD) (right) with respect to horizontal and vertical expansion of the clouds. The average observations over 300 clouds is represented.

Embedding sensitivity and F_{40d}

- We have the same observations on F_{40d} as previously.
- **Conclusion regarding MMD** $\langle \mu_X, \mu_X \rangle$ *is sensitive in the same way as the functions of interest.*

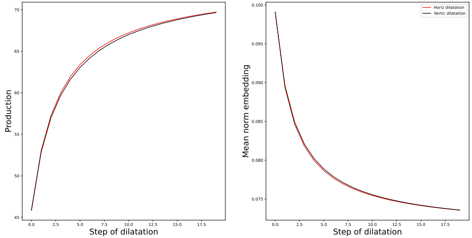


Figure: Representation of the sensitivity of F_{40d} (left) and Norm of Embedding (MMD) (right) with respect to horizontal and vertical expansion of the clouds. The average observations over 300 clouds is represented.

Perspectives

Scientific Perspectives

- Concerning Relevant Feature kernel, find automatically the most relevant features for a given function
- For MMD and MMK, model with **non uniform probabilities**. Considering different weights on points could allow giving more importance to some specific points of the cloud.
- Define the directions of **Sliced Wasserstein Distance by Log Likelihood**.
- Apply to TOPFARM industrial usecase.
- Perform Bayesian optimization on the layout of the windfarm.

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] Thi Thien Trang Bui et al. "Distribution regression model with a Reproducing Kernel Hilbert Space approach". In: *arXiv preprint arXiv:1806.10493* (2018).

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

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

Bibliography II

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

[7] 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.

[8] Bernard Haasdonk and Claus Bahlmann. “Learning with distance substitution kernels”. In: *Joint pattern recognition symposium*. Springer. 2004, pp. 220–227.

[9] Tony Jebara and Risi Kondor. “Bhattacharyya and expected likelihood kernels”. In: *Learning theory and kernel machines*. Springer, 2003, pp. 57–71.

[10] 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.

Bibliography III

[11] 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.

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

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

[14] 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 [4] and Kolouri, Zou, and Rohde [10]
- 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 $\mathcal{W}_2^2(\mu, \nu) = ||\mathcal{C}_\mu^{-1} - \mathcal{C}_\nu^{-1}||_{L^p([0,1])}^2$, see Peyré, Cuturi, et al. [12]
- $\mathcal{W}_2^2(\mu, \nu)$ is symmetric and conditionally negative definite. (Kolouri, Zou, and Rohde [10])
- If μ and ν are defined in $\mathbb{R} \times \mathbb{R}$, the above condition is no longer guaranteed.

Distance between laws: Wasserstein Distance between Gaussians

Substitution with Hilbertian distance: Wasserstein Distance Between Gaussians

- 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 : $W_p^p = \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{M}\times\mathcal{M}} \rho(x,x')^p \mathrm{d}\pi(x,x')$
- We consider the case 2
- For an Euclidean cost in 2D , the Wasserstein distance of two Gaussians is given in a closed form as : $W_2^2 = ||m_X - m_{X'}||^2 + tr(\Sigma_X + \Sigma_{X'} - 2(\Sigma_X^{1/2}\Sigma_{X'}\Sigma_X^{1/2})^{1/2})$
- Consider the version $W_2^2 = ||m_X - m_{X'}||^2 + ||\Sigma_X^{1/2} - \Sigma_{X'}^{1/2}||_{Frobenius}^2$ as in Bui et al. [3]
- The above distance is conditionally negative definite and $k(X,X') = \sigma^2 exp(-\frac{W_2^2}{2\theta^2})$ is therefore a valid kernel.