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

Table of contents

Context and problem

Functions defined over sets of vectors

- Let $\mathcal{F}$ be the family of considered functions.
- Suppose $f \in \mathcal{F}$ and $\mathcal{D}_f$ its domain of definition.
- $$u \in \mathcal{D}_f \Rightarrow \exists n \in \mathbb{N}, d \in \mathbb{N}, u = \{x_1, ..., x_n\}, \forall i, x_i \in \mathbb{R}^d$$
- n belongs to a finite discrete set.
- For any permutation π of the set $\{1, ..., n\}$ to a new one $\{\pi(1), ..., \pi(n)\}$, we denote by u_π the following set $\{x_{\pi(1)}, ..., x_{\pi(n)}\}$.
- Note that $\forall \pi, f(u_\pi) = f(u)$: **f is invariant under permutation**.
- The variables u will be called **clouds of points**.

Related works and domains

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.
- Confronting them numerically on a a test function mimicking the production of a windfarm.

Bayesian Approach

A Gaussian process prior

- Gaussian process is defined by a mean function m and a kernel k over the spaces of inputs $\mathcal{X}$ to approximate the functions.
- Observing $D = \{(x_1, y_1)...(x_n, y_n)\}$ where $x_i \in \mathcal{X}$ and $y \in \mathbb{R}$ as training data, the predictive mean and covariance for a new point x are given by:

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

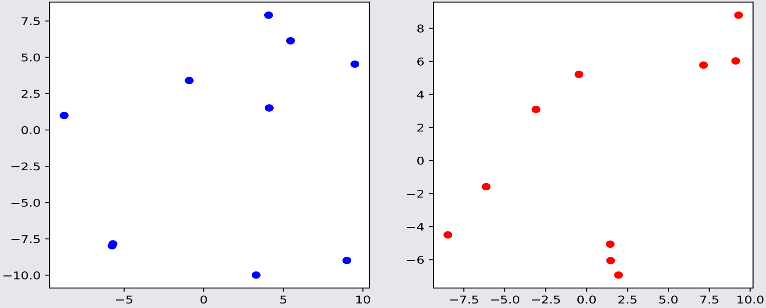
$$\Sigma(x, x; D) = K(x, x) - K(X, x)^T K(X, X)^{-1} K(X, x)$$

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



Aronszajn, Explicit, 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 $\forall x, y, k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$.

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
  - **Substitutions Kernels** as in Haasdonk and Bahlmann [8].

Substitution with Hilbertian Distance

Substitution with Exponential

- Firstly, we consider covariance kernels of the form: $k(X, Y) = \sigma^2 exp(-\frac{\Psi(X,Y)}{2\theta^2})$.
- Semi-definite positiveness is equivalent to Ψ be **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 $\Psi(X, Y) = d(\tilde{X}, \tilde{Y})^2$
- d is the distance between $\tilde{X}$ and $\tilde{Y}$ the respective images of X and Y 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 $\tilde{Y}$?

With probabilities

- Case 1 : Suppose we have two clouds $X = (x_1, ..x_n)$, $Y = (y_1, ..., y_m)$ and $P_X = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$, $P_Y = \frac{1}{m}\sum_{j=1}^{m}\delta_{y_j}$, the respective associated empirical uniform distributions.
- Case 2 : We associate to each cloud of point $X = (x_1, ..x_n)$, $Y = (y_1, ..., y_m)$, its empirical Gaussian: $\mathcal{N}_{\mathcal{X}}(m_X, \Sigma_X)$ and $\mathcal{N}_{\mathcal{Y}}(m_Y, \Sigma_Y)$. item $\mathcal{N}_{\mathcal{X}}$ is defined by $m_X = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\Sigma_X = \frac{1}{n}\sum_{i=1}^{n}(x_i - m_X)(x_i - m_X)^T$

With vectors : vectorization

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

Slice Wasserstein Distance and Gaussian approximation

Wasserstein Distances

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

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

- Let $\mathcal{S} = \{\theta \in \mathbb{R}^2, ||\theta|| = 1\}$. Consider the projected empirical measure on the line directed by $\theta \in \mathcal{S}$: $\theta^* P_X = \frac{1}{n}\sum_{i=1}^n \delta_{<x_i,\theta>}$ and $\theta^* P_Y = \frac{1}{m}\sum_{i=1}^m \delta_{<y_i,\theta>}$
- $SW_2^2(P_X, P_Y) = \int_{\mathcal{S}} \mathcal{W}_2^2(\theta^* P_X, \theta^* P_Y)\mathrm{d}\theta$. Implementation using POT [6]
- The covariance kernel $k(X,Y) = \sigma^2 exp(-\frac{SW_2^2(P_X,P_Y)}{2\theta^2})$ is symmetric and semi-definite positive as in Carriere, Cuturi, and Oudot [4].It will be denoted **Sliced Wasserstein subs**

Approximate For Gaussian Modeling (see Annex) , **Gaussian Wasserstein subs**

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

Distance between embedded laws : Maximum Mean Discrepancy

Substitution with Hilbertian distance: MMD

- There exists a Reproducing Kernel Hilbert Space, $\mathcal{H}$ with a characteristic kernel such as $k_{\mathcal{H}}(x,.) = exp(-\frac{||x-.||^2}{2\theta^2})$.
- The characteristic nature guarantees the injectivity of the embedding map Muandet et al. [11]: $P_X \longmapsto \mu_X(.) = \int P_X(x) k_{\mathcal{H}}(x,.)\mathrm{d}x$.
- $MMD^2(P_X, P_Y) = ||\mu_X - \mu_Y||^2_{\mathcal{H}}$
- For any kernel $k_{\mathcal{H}}$ of the RKHS, the uniform discrete (supported by points) laws give $MMD^2(P_X, P_Y) =$ $\frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n k_{\mathcal{H}}(x_i, x_j) + \frac{1}{m^2}\sum_{i=1}^m \sum_{j=1}^m k_{\mathcal{H}}(y_i, y_j) - 2\frac{1}{nm}\sum_{i=1}^n \sum_{j=1}^m k_{\mathcal{H}}(x_i, y_j)$
- The covariance kernel $k(X,Y) = \sigma^2 exp(-\frac{||\mu_X - \mu_Y||^2_{\mathcal{H}}}{2\theta^2})$ is symmetric and definite positive.
- We will denote the latter as **MMD**.

Constructing Features of a cloud

Relevant Features Map Kernel

- We consider a final kernel of the form $k(X,Y) = \sigma^2 \exp\left(-\sum_{j=1}^{n'} \frac{|w'_j(X)-w'_j(Y)|^2}{\theta_j'^2}\right)$ with $(w'_1(X), \ldots, w'_n(X))$ 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**.

Explicit Mappings: Probability Product Kernels and Embeddings

Explicit Mappings (see Annex)

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

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

  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. $k(x, y) =< \mu_X, \mu_Y >$ it will be called **MMK**, Mean Map Kernel, for the remainder.

A test function

Mimicking wind farms

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

$$F(\{x_1, ..., x_n\}) = \sum_{i=1}^{n} \sum_{\substack{j \\ x_{j,1} \leq x_{i,1}}} f_p(x_j, x_i) f_0(x_i)$$

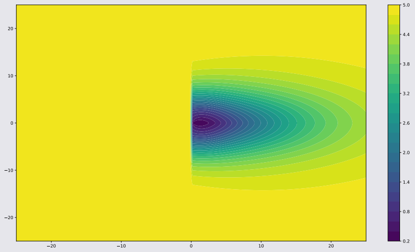
  where $f_p(x_j, x_i)$ expresses the energy loss over x_i that is caused by x_j and f_0 is a constant. $x_i \in \mathbb{R}^2$ and $\in \{10, 11, .., 20\}$
- The function $x_i \longmapsto f_p(x_j, x_i)$ can be parametrized differently:
  - It can be unidirectional with an arbitrary angle
  - It can be multi-directional

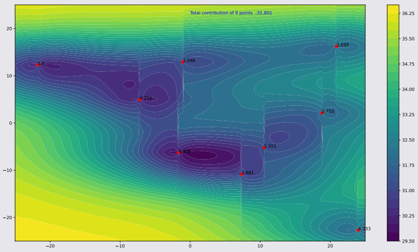
A test function

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



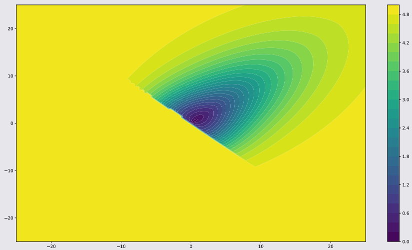


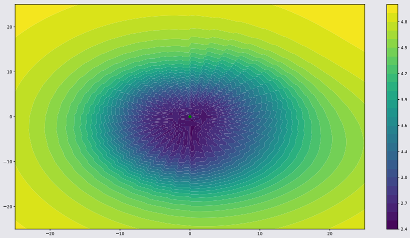
A test function

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.

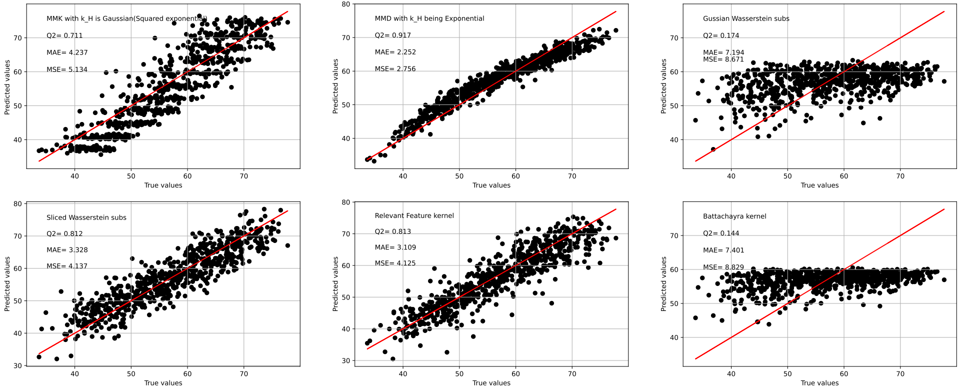
Mimicking wind farms : Illustration





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 MMD ;



Results: 45° direction Interaction

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

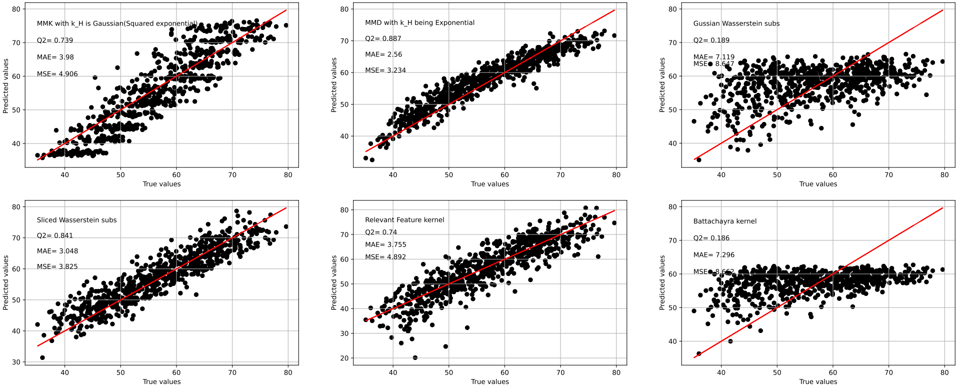


Figure: Prediction performance on 45° direction Interaction Function

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

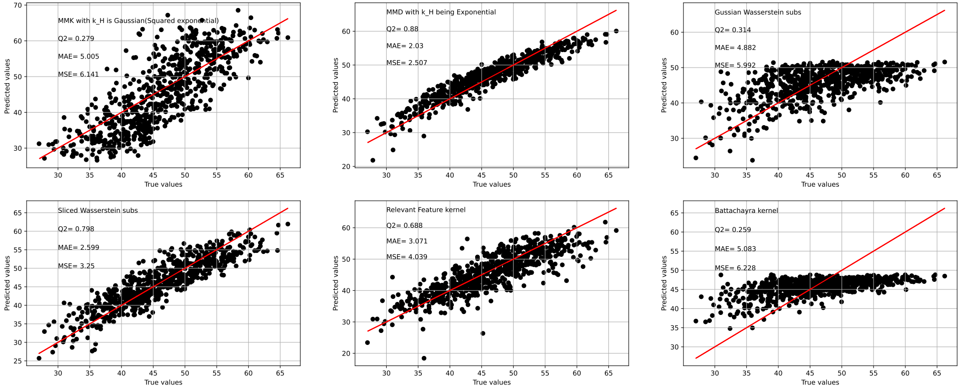


Figure: Prediction performance on 40 directions integrated function

Summary

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

Function \ Kernels	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 remains the most robust kernels. MMK fails to model a lot of directions integrated.
- Modeling clouds as Gaussian seem very poor in front of discrete uniforms modelization.
- SWS and RFK are very competitive with MMD.

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**.

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:

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

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