Gaussian process regression model for distribution inputs

4th Conference of the International Society for Nonparametric Statistics

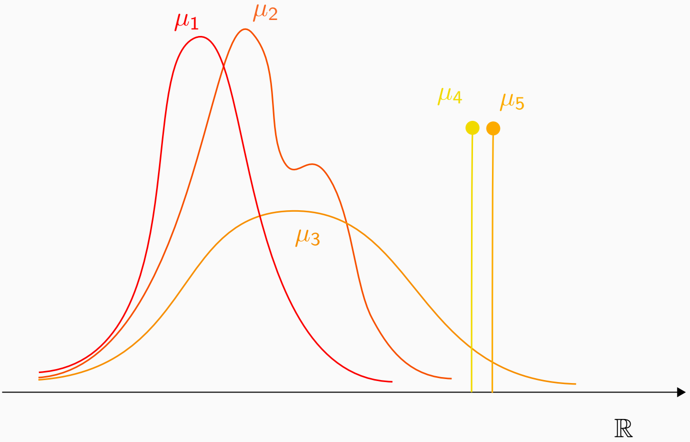
N. Venet*, F. Bachoc⋆, F. Gamboa⋆, J.-M. Loubes⋆

July the 15th, 2018, Salerno

*Università degli Studi di Bergamo, ⋆Institut de Mathématiques de Toulouse

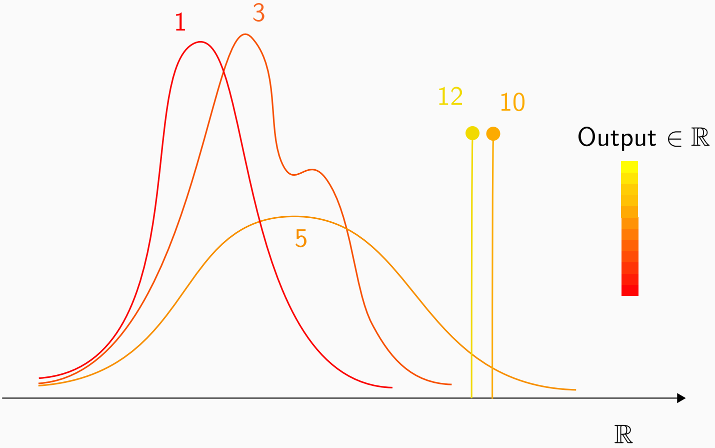
The regression problem for distribution inputs

We are given n *input/output* couples $(\mu_i, y_i) \in \mathcal{P}(\mathbb{R}) \times \mathbb{R}$, and we are looking to associate an output to a new input μ_{n+1}.



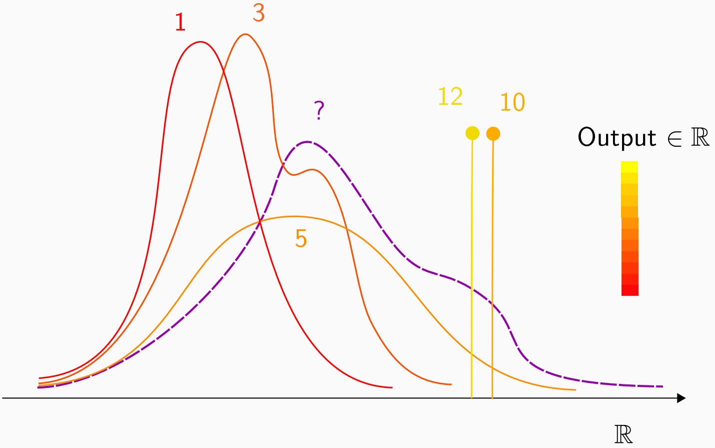
The regression problem for distribution inputs

We are given n *input/output* couples $(\mu_i, y_i) \in \mathcal{P}(\mathbb{R}) \times \mathbb{R}$, and we are looking to associate an output to a new input μ_{n+1}.



The regression problem for distribution inputs

We are given n *input/output* couples $(\mu_i, y_i) \in \mathcal{P}(\mathbb{R}) \times \mathbb{R}$, and we are looking to associate an output to a new input μ_{n+1}.



Motivations

Our motivations are twofold: we want to deal with regression problems which inputs are

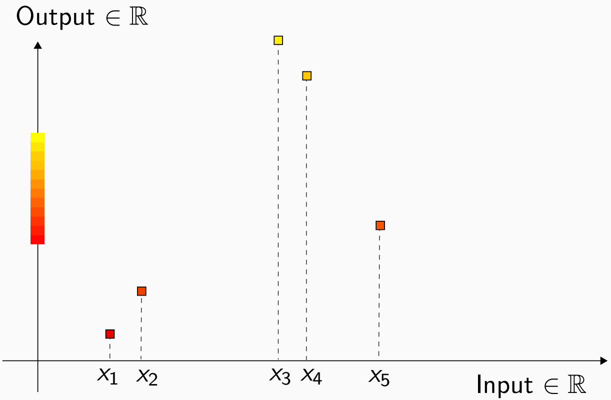
1. probability distributions (ex: blood sampling problem, anonymised data, ...)
2. functional objects (spectra, histograms, ...)
    - with the nonnegative values and mass 1 restrictions
    - ... which in turn allow the use of tools such as the Wasserstein distance

Outline of the presentation

1. Gaussian Process Regression
2. Existence of models – Stationary kernels on the Wasserstein space
3. Maximum-likelihood model selection – Asymptotic results
4. Numerical performances

Gaussian Process Regression

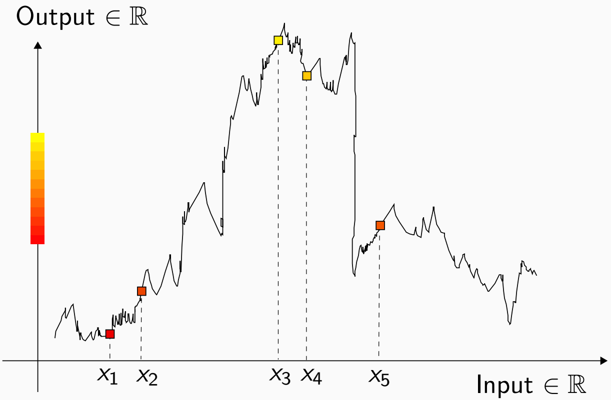
Gaussian Process Regression (Kriging)



We chose a random process $(Y_x)_{x\in\mathbb{R}}$ and consider

$$\hat{Y}(x) := \mathbb{E}(Y_x | Y_{x_1} = y_1, \cdots, Y_{x_n} = y_n)$$

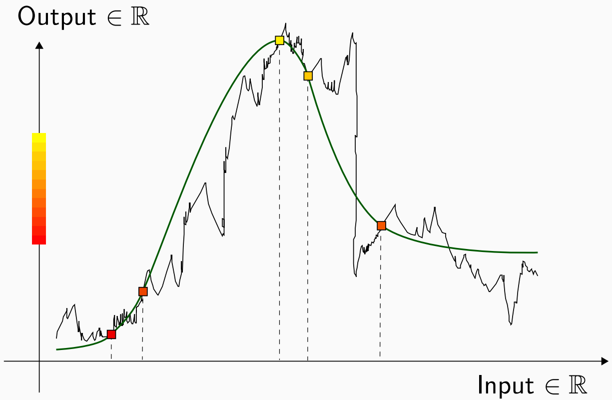
Gaussian Process Regression (Kriging)



We chose a random process $(Y_x)_{x \in \mathbb{R}}$ and consider

$$\hat{Y}(x) := \mathbb{E}(Y_x | Y_{x_1} = y_1, \cdots, Y_{x_n} = y_n)$$

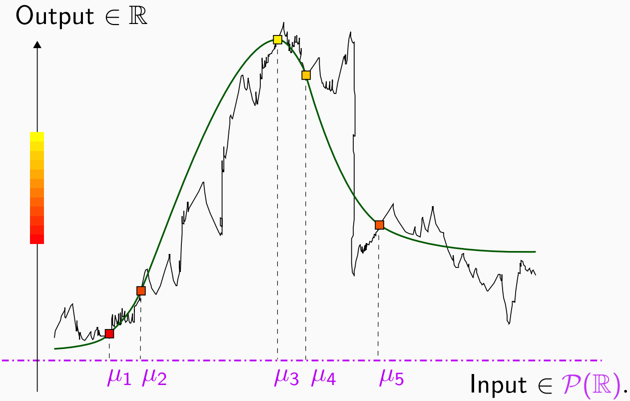
Gaussian Process Regression (Kriging)



We chose a random process $(Y_x)_{x\in\mathbb{R}}$ and consider

$$\hat{Y}(x) := \mathbb{E}(Y_x | Y_{x_1} = y_1, \cdots, Y_{x_n} = y_n)$$

Gaussian Process Regression (Kriging)



Here we need a random process $(Y_\mu)_{\mu \in \mathcal{P}(\mathbb{R})}$ to consider

$$\hat{Y}(\mu) := \mathbb{E}(Y_\mu | Y_{\mu_1} = y_1, \cdots, Y_{\mu_n} = y_n)$$

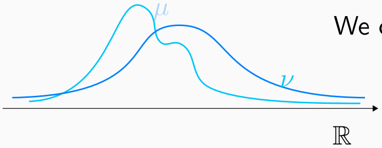
Existence of models – Stationary kernels on the Wasserstein space

The Wasserstein distance

The *Wasserstein distance* between two probability distributions μ and ν that admit a second order moment is defined by:

$$W_2(\mu,\nu) := \left(\inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathbb{R}^2} |x-y|^2 \, d\pi(x,y) \right)^{1/2},$$

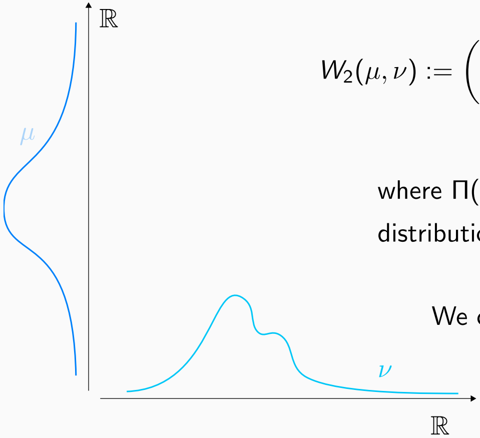
where $\Pi(\mu,\nu)$ is the set of probability distributions on $\mathbb{R}^2$ with margins μ and ν.



We obtain a metric space $\mathcal{W}_2(\mathbb{R})$.

The Wasserstein distance

The *Wasserstein distance* between two probability distributions μ and ν that admit a second order moment is defined by:



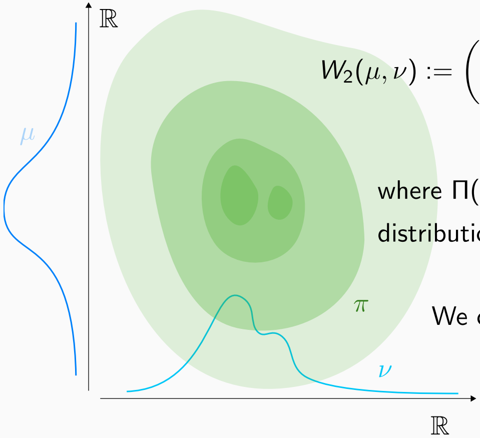
$$W_2(\mu,\nu) := \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^2} |x-y|^2 \, d\pi(x,y) \right)^{1/2},$$

where $\Pi(\mu,\nu)$ is the set of probability distributions on $\mathbb{R}^2$ with margins μ and ν.

We obtain a metric space $\mathcal{W}_2(\mathbb{R})$.

The Wasserstein distance

The *Wasserstein distance* between two probability distributions μ and ν that admit a second order moment is defined by:



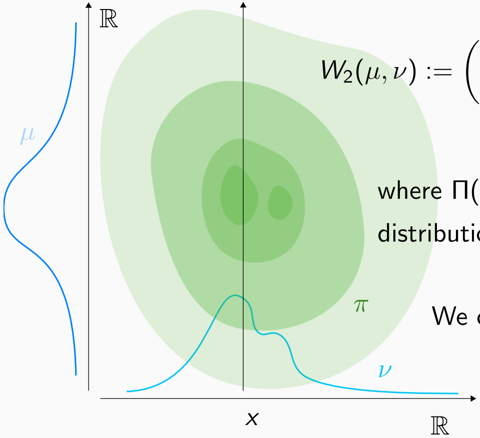
$$W_2(\mu, \nu) := \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^2} |x-y|^2 \, d\pi(x,y) \right)^{1/2},$$

where $\Pi(\mu, \nu)$ is the set of probability distributions on $\mathbb{R}^2$ with margins μ and ν.

We obtain a metric space $\mathcal{W}_2(\mathbb{R})$.

The Wasserstein distance

The *Wasserstein distance* between two probability distributions μ and ν that admit a second order moment is defined by:



$$W_2(\mu,\nu) := \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^2} |x-y|^2 \, d\pi(x,y) \right)^{1/2},$$

where $\Pi(\mu,\nu)$ is the set of probability distributions on $\mathbb{R}^2$ with margins μ and ν.

We obtain a metric space $\mathcal{W}_2(\mathbb{R})$.

A core remark in $\mathcal{W}_2(\mathbb{R})$

For $\mu, \nu \in \mathcal{W}_2(\mathbb{R})$ and F_μ^{-1}, F_ν^{-1} the associated quantile functions,

$$W_2(\mu,\nu) = \left(\int_{[0,1]} \left(F_\mu^{-1}(u) - F_\nu^{-1}(u) \right)^2 du \right)^{1/2}. \tag{1}$$

- This optimal coupling, which is specific to the dimension 1 case, allows the numerical evaluation of Wasserstein distances.
- It is also the main ingredient of the proofs of Theorems 1 and 2.

Existence of Wasserstein-indexed models i

Theorem 1 (Fractional Brownian fields)

For ever $0 \leq H \leq 1$ *and* $\sigma_0 \in \mathcal{W}_2(\mathbb{R})$*,*

$$K^{H,\sigma}(\mu,\nu) = \frac{1}{2}\left(W_2^{2H}(\sigma_0,\mu) + W_2^{2H}(\sigma_0,\nu) - W_2^{2H}(\mu,\nu)\right) \qquad (2)$$

is a covariance function on $\mathcal{W}_2(\mathbb{R})$*. Moreover, it is nondegenerated if and only if* $0 < H < 1$*.*

- We get a fractional Brownian field indexed by $\mathcal{W}_2(\mathbb{R})$. It is a generalisation of the time-indexed fractional Brownian motion, which inherits many enjoyable properties:
- Statistical auto-similarity, path-regularity and long distance memory that are governed by the *Hurst parameter* H.

Existence of Wasserstein-indexed models ii

Theorem 2 (Stationary processes)

For every completely monotone $F : \mathbb{R}^+ \to \mathbb{R}^+$ and $0 < H \leq 1$,

$$(\mu, \nu) \mapsto F\left(W_2^{2H}(\mu, \nu)\right) \tag{3}$$

is a stationary covariance function on $\mathcal{W}_2(\mathbb{R})$.

- Recall that $F \in C^\infty(\mathbb{R}^+, \mathbb{R}^+)$ is completely monotone if $(-1)^n F^{(n)}$ is nonnegatively valued for every $n \in \mathbb{N}$.
- In particular for every $\sigma^2, \ell > 0$ and $0 \leq H \leq 1$,

$$K_{\sigma^2,\ell,H}(\nu_1, \nu_2) = \sigma^2 \exp\left(-\frac{W_2(\nu_1, \nu_2)^{2H}}{\ell}\right) \tag{M}$$

is a valid covariance.

Maximum-likelihood model selection – Asymptotic results

Conditions for our results i

Condition 1 (Asymptotic expansion framework)

We consider a triangular array of observation points $\{\mu_1, ..., \mu_n\} = \{\mu_1^{(n)}, ..., \mu_n^{(n)}\}$ *so that for all* $n \in \mathbb{N}$ *and* $1 \leq i \leq n$, μ_i *has support in* $[i, i + K]$ *with a fixed* $K < \infty$.

Condition 2 (Parametric stationary model)

The model of covariance functions $\{K_\theta, \theta \in \Theta\}$ *satisfies*

$$\forall \theta \in \Theta, \; K_\theta(\mu, \nu) = F_\theta \left(W_2(\mu, \nu) \right),$$

with $F_\theta : \mathbb{R}^+ \to \mathbb{R}$ *and* $\sup_{\theta \in \Theta} |F_\theta(t)| \leq \frac{A}{1+|t|^{1+\tau}}$ *with a fixed* $A < \infty$, $\tau > 1$.

Conditions for our results ii

Condition 3 (Well-specified case)

We have observations $y_i = Y(\mu_i)$, $i = 1, \cdots, n$ of the centered Gaussian Process Y with covariance function K_{θ_0} for some $\theta_0 \in \Theta$.

Condition 4 (Asymptotical nondegeneracy)

The sequence of matrices $R_\theta = (K_\theta(\mu_i, \mu_j))_{1 \le i,j \le n}$ satisfies

$$\lambda_{\inf}(R_\theta) \ge c$$

for a fixed $c > 0$, where $\lambda_{\inf}(R_\theta)$ denotes the smallest eigenvalue of R_θ.

Conditions for our results iii

Condition 5 (First sampling condition)

$\forall \alpha > 0,$

$$\liminf_{n\to\infty} \inf_{\|\theta-\theta_0\|\geq\alpha} \frac{1}{n}\sum_{i,j=1}^{n}\left[K_\theta(\mu_i,\mu_j) - K_{\theta_0}(\mu_i,\mu_j)\right]^2 > 0.$$

Consistency of the maximum-likelihood estimator

Theorem 3 (Consistency of MLE)

Under conditions 1 to 5, the maximum-likelihood estimator is consistent, that is to say:

$$\hat{\theta}_{ML} \xrightarrow[n \to \infty]{\mathbb{P}} \theta_0.$$

Supplementary conditions

Condition 6 (Model regularity)

- *$\forall t \geq 0$, $F_\theta(t)$ is $\mathcal{C}^1$ with respect to θ and verifies $\sup_{\theta\in\Theta} \max_{i=1,\cdots,p} \left|\frac{\partial}{\partial\theta_i} F_\theta(t)\right| \leq \frac{A}{1+t^{1+\tau}}$, where A, τ are defined in Condition 2.*
- *For every $t \geq 0$, $F_\theta(t)$ is $\mathcal{C}^3$ with respect to θ and $\forall q \in \{2,3\}$, $\forall i_1 \cdots i_q \in \{1, \cdots p\}$,*

$$\sup_{\theta\in\Theta} \max_{i=1,\cdots,p} \left|\frac{\partial}{\partial\theta_{i_1}} \cdots \frac{\partial}{\partial\theta_{i_q}} F_\theta(t)\right| \leq \frac{A}{1+|t|^{1+\tau}}.$$

Condition 7 (Second sampling condition)

$\forall(\lambda_1 \cdots, \lambda_p) \neq (0, \cdots, 0)$,

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{i,j=1}^{n} \left(\sum_{k=1}^{p} \lambda_k \frac{\partial}{\partial_{\theta_k}} K_{\theta_0}(\mu_i, \mu_j) \right)^2 > 0.$$

Asymptotic normality of the maximum-likelihood estimator

Theorem 4

Let M_{ML} be the $p \times p$ matrix defined by

$$(M_{ML})_{i,j} = \frac{1}{2n} Tr \left(K_{\theta_0}^{-1} \frac{\partial K_{\theta_0}}{\partial \theta_i} K_{\theta_0}^{-1} \frac{\partial K_{\theta_0}}{\partial \theta_j} \right).$$

Under conditions 1 to 6, the maximum-likelihood estimator is asymptotically normal:

$$\sqrt{n}\ M_{ML}^{1/2} \left(\hat{\theta}_{ML} - \theta_0 \right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, I_p).$$

Moreover
$0 < \liminf_{n\to\infty} \lambda_{min}(M_{ML}) \leq \limsup_{n\to\infty} \lambda_{max}(M_{ML}) < +\infty.$

Sampling conditions are reasonable

Proposition 1

Assume that Conditions 2 and 6 hold, that for $\theta \neq \theta_0$, F_θ and F_{θ_0} are not equal everywhere on $\mathbb{R}^+$, and that there does not exist $(\lambda_1, ..., \lambda_p) \neq (0, ..., 0)$ so that $\sum_{i=1}^{p} \lambda_i (\partial / \partial \theta_i) F_{\theta_0}$ is the zero function on $\mathbb{R}^+$.

Let $(Z_i)_{i \in \mathbb{Z}}$ be iid, centred Gaussian processes on $\mathbb{R}$ with continuous trajectories, and stationary covariance $C_0(u - v)$.

Assume that $\hat{C}_0(w)|w|^{2p}$ is bounded away from 0 and ∞ as $|w| \to \infty$.

Let $K > 1$ be fixed. For $i \in \mathbb{N}$, let μ_i be the measure with density

$$f_i(t) = \frac{e^{Z_i(t-i)}}{\int_i^{i+K} e^{Z_i(t-i) dt}} 1_{[i, i+K]}(t).$$

Then, almost surely, with the sequence of random probability measures $\{\mu_1, ..., \mu_n\}$, Conditions 5 and 7 hold.

Kriging under the ML-estimated parameter

Theorem 5

Under conditions 1 to 6, the Kriging estimator under the ML-estimated parameter $\hat{\theta}_{ML}$ is asymptotically optimal:

$$\forall \mu \in \mathcal{W}_2(\mathbb{R}), \ \left| \hat{Y}_{\hat{\theta}_{ML}}(\mu) - \hat{Y}_{\theta_0}(\mu) \right| = o_{\mathbb{P}}(1).$$

Numerical performances

Comparison with projection-based covariances

- Denote by $m_k(\nu)$ the order k moment of ν. We consider

$$F : \mathcal{W}_2(\mathbb{R}) \to \mathbb{R}$$

$$F(\nu) = \frac{m_1(\nu)}{0.05 + \sqrt{m_2(\nu) - m_1(\nu)^2}}, \tag{4}$$

which we are going to regress.

Comparison with projection-based covariances

- Denote by $m_k(\nu)$ the order k moment of ν. We consider

$$F : \mathcal{W}_2(\mathbb{R}) \to \mathbb{R}$$

$$F(\nu) = \frac{m_1(\nu)}{0.05 + \sqrt{m_2(\nu) - m_1(\nu)^2}}, \tag{4}$$

which we are going to regress.

- Let us generate normal random variables $\nu_1, \cdots, \nu_{100}$, with means and variances drawn uniformly at random, randomly perturbed to exhibit irregularities.

Comparison with projection-based covariances

- Denote by $m_k(\nu)$ the order k moment of ν. We consider

$$F : \mathcal{W}_2(\mathbb{R}) \to \mathbb{R}$$

$$F(\nu) = \frac{m_1(\nu)}{0.05 + \sqrt{m_2(\nu) - m_1(\nu)^2}}, \tag{4}$$

  which we are going to regress.
- Let us generate normal random variables $\nu_1, \cdots, \nu_{100}$, with means and variances drawn uniformly at random, randomly perturbed to exhibit irregularities.
- We estimate $\hat{\sigma}^2, \hat{\ell}, \hat{H}$ by maximising the maximum likelihood for the parametric model:

$$K_{\sigma^2,\ell,H}(\nu_1, \nu_2) = \sigma^2 \exp\left(-\frac{W_2(\nu_1, \nu_2)^{2H}}{\ell}\right). \tag{5}$$

Comparison with projection-based covariances

- We evaluate the method on a test dataset $(\nu_{t,i})_{i=1}^{500}$ which is generated in a same way as the ν_i,

Comparison with projection-based covariances

- We evaluate the method on a test dataset $(\nu_{t,i})_{i=1}^{500}$ which is generated in a same way as the ν_i, with the criteria:

$$RMSE^2 = \frac{1}{500}\sum_{i=1}^{500}\left(F(\nu_{t,i}) - \hat{F}(\nu_{t,i})\right)^2,$$

$$CIR_\alpha = \frac{1}{500}\sum_{i=1}^{500}\mathbf{1}\left\{\left|F(\nu_{t,i}) - \hat{F}(\nu_{t,i})\right| \leq q_\alpha \hat{\sigma}(\nu_{t,i})\right\}.$$

Comparison with projection-based covariances

- We evaluate the method on a test dataset $(\nu_{t,i})_{i=1}^{500}$ which is generated in a same way as the ν_i, with the criteria:

$$RMSE^2 = \frac{1}{500}\sum_{i=1}^{500}\left(F(\nu_{t,i}) - \hat{F}(\nu_{t,i})\right)^2,$$

$$CIR_\alpha = \frac{1}{500}\sum_{i=1}^{500}\mathbf{1}\left\{\left|F(\nu_{t,i}) - \hat{F}(\nu_{t,i})\right| \leq q_\alpha \hat{\sigma}(\nu_{t,i})\right\}.$$

modèle	RMSE	$CIR_{0.9}$
“Wasserstein”	0.094	0.92
“Legendre” ordre 5	0.49	0.92
“Legendre” ordre 10	0.34	0.89
“Legendre” ordre 15	0.29	0.91
“PCA” ordre 5	0.63	0.82
“PCA” ordre 10	0.52	0.87
“PCA” ordre 15	0.47	0.93

Two stage sampling

- In [5], Poczos and al. try to learn the skewness $S(P)$ of beta distributions P from some samplings.
  1. Starting with a kernel smoothing $\hat{P}, \hat{P}_1, \cdots, \hat{P}_n$ of the empirical distributions corresponding to the samplings.
  2. Then the prediction $\hat{S}(\hat{P})$ of $S(P)$ is obtained by a weighted average of $S(P_1), ..., S(P_n)$. Weights are obtained by applying some kernel to the L^1 distance between densities $\hat{P}$ and $\hat{P}_1, \cdots, \hat{P}_n$.

Two stage sampling

- In [5], Poczos and al. try to learn the skewness $S(P)$ of beta distributions P from some samplings.
  1. Starting with a kernel smoothing $\hat{P}, \hat{P}_1, \cdots, \hat{P}_n$ of the empirical distributions corresponding to the samplings.
  2. Then the prediction $\hat{S}(\hat{P})$ of $S(P)$ is obtained by a weighted average of $S(P_1), ..., S(P_n)$. Weights are obtained by applying some kernel to the L^1 distance between densities $\hat{P}$ and $\hat{P}_1, \cdots, \hat{P}_n$.
- We add a nugget term to our covariance to accomodate for the differenve between $S(P)$ and $S(\hat{P})$:

$$K_{\sigma^2,\ell,H,\delta}(\nu_1,\nu_2) = \sigma^2 \exp\left(-\frac{W_2(\nu_1,\nu_2)^{2H}}{\ell}\right) + \delta \mathbf{1}_{\{W_2(\nu_1,\nu_2)=0\}},$$

and obtain the following results:

model	RMSE	$CIR_{0.9}$
"distribution"	0.21	0.91
"kernel regression"	0.93	

Thank you for your attention

F. Bachoc.
Asymptotic analysis of the role of spatial sampling for covariance parameter estimation of Gaussian processes.
Journal of Multivariate Analysis, 125:1–35, 2014.

F. Bachoc, F. Gamboa, J.-M. Loubes, and N. Venet.
Gaussian process regression model for distribution inputs.
arXiv preprint arXiv:1701.09055, to appear in IEEE Transactions on Information Theory, 2018.

C. Berg, J. P. R. Christensen, and P. Ressel.
Harmonic analysis on semigroups.
Springer-Verlag, 1984.

J. Istas.
Manifold indexed fractional fields.
ESAIM Probab. Stat., 16:222–276, 2012.

B. Póczos, A. Singh, A. Rinaldo, and L. Wasserman.
Distribution-free distribution regression.
In *In Proceedings of the 16th International Conference on Artificial Intelligence and Statistics, volume 31 of JMLR Proceedings*, pages 507–515, 2013.

N. Venet.
Nonexistence of fractional brownian fields indexed by cylinders.
arXiv preprint, 2016.

N. Venet.
On the existence of fractional brownian fields indexed by manifolds with closed geodesics.
arXiv preprint, 2016.

C. Villani.
Optimal transport: old and new, volume 338.
Springer Science & Business Media, 2009.

Wanted:

Dataset with inputs on the cylinder $\mathbb{S}^1 \times \mathbb{R}$.