RECENT ADVANCES IN KERNEL METHODS FOR COMPUTER EXPERIMENTS

—

Sébastien Da Veiga
Safran Tech

Journées MAS 2022 – 29 août 2022

SAFRAN

Outline

Context – Computer experiments

Gaussian process regression

Design of experiments with kernels

Sensitivity analysis with kernels

Conclusion & outlook

0

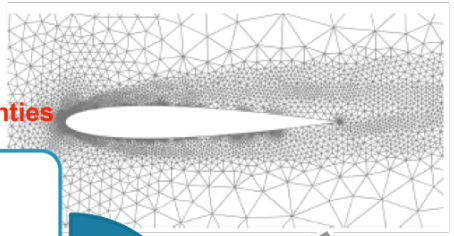
CONTEXT

COMPUTER EXPERIMENTS

Context

Simplifications

Model uncertainties



Physical phenomenon

Physicist

Physical model

Numerical approximation

Mathematician

Numerical uncertainties

Input data

Stochastic uncertainties

Computer scientist

Variables of interest

Numerical model

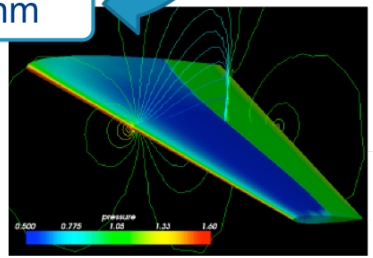
Algorithm

Statistician

Epistemic uncertainties

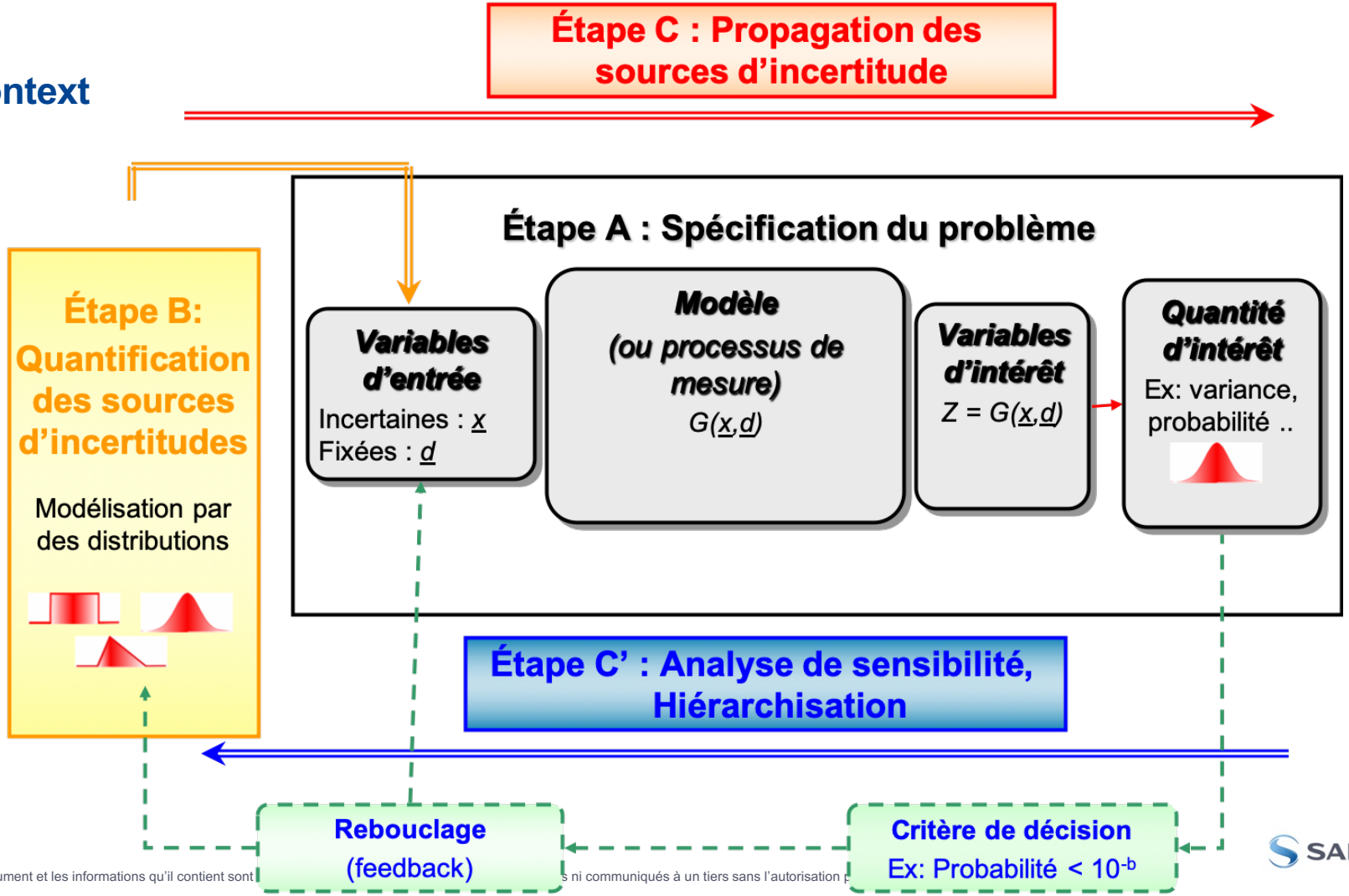
Numerical noise

Parameters



SAFRAN

Context



SAFRAN

Context – Computer experiments

The model

- Regular, with symmetries, invariances and physical constraints
- **Most of the time, very expensive to evaluate (large systems of PDEs)**
- Will then be replaced by a surrogate model for all UQ and optimization studies
  - A popular one is Gaussian Process (GP) regression → Part 1

Context – Computer experiments

The model

- Regular, with symmetries, invariances and physical constraints
- **Most of the time, very expensive to evaluate (large systems of PDEs)**
- Will then be replaced by a surrogate model for all UQ and optimization studies
  - A popular one is Gaussian Process (GP) regression → Part 1

The design of experiment

- For building the surrogate, we need a sample of inputs / outputs obtained by evaluating the expensive model
- The particularity here is that we can choose how we build this DOE
- Large literature on LHS and space-filling designs
- **Recent work on kernel embeddings of distributions provides a new paradigm for improving our practice** → Part 2

Context – Computer experiments

The model

- Regular, with symmetries, invariances and physical constraints
- **Most of the time, very expensive to evaluate (large systems of PDEs)**
- Will then be replaced by a surrogate model for all UQ and optimization studies
  - A popular one is Gaussian Process (GP) regression → Part 1

The design of experiment

- For building the surrogate, we need a sample of inputs / outputs obtained by evaluating the expensive model
- The particularity here is that we can choose how we build this DOE
- Large literature on LHS and space-filling designs
- **Recent work on kernel embeddings of distributions provides a new paradigm for improving our practice** → Part 2

Sensitivity analysis aka Feature importance

- Invaluable tool for engineers, mostly centered around Sobol' indices and Shapley effects
- **Once again, kernel methods can extend and improve previous practice** → Part 3

1

GP REGRESSION

QUICK RECAP, RECENT ADVANCES & CHALLENGES

SAFRAN

Standard GP regression

Notations

> Computer code $g : \mathbb{R}^D \to \mathbb{R}$

> Inputs $\mathbf{x} = \left(x^1, \ldots, x^D\right)$

> Output $y = g(\mathbf{x})$

> Observations $\left(\mathbf{x}_i, y_i\right)_{i=1,\ldots,n}$ $X_s = \left[\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T\right]^T$ $Y_s = \left[y_1, \ldots, y_n\right]^T$

Standard GP regression

Notations

> Computer code $g : \mathbb{R}^D \to \mathbb{R}$

> Inputs $\mathbf{x} = \left(x^1, \ldots, x^D\right)$

> Output $y = g(\mathbf{x})$

> Observations $(\mathbf{x}_i, y_i)_{i=1,\ldots,n}$ $X_s = \left[\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T\right]^T$ $Y_s = \left[y_1, \ldots, y_n\right]^T$

Model: Output seen as realization of stationary Gaussian process

$$Y(\mathbf{x}) = g_0(\mathbf{x}) + U(\mathbf{x}) \qquad g_0(\mathbf{x}) = \sum_{j=1}^{J} \beta_j g_j(\mathbf{x}) = G(\mathbf{x})\beta$$

$$C(\mathbf{x}, \mathbf{x}') = \sigma^2 R(\mathbf{x}, \mathbf{x}')$$

Standard GP regression

Notations

> Computer code $g : \mathbb{R}^D \to \mathbb{R}$

> Inputs $\mathbf{x} = \left(x^1, \ldots, x^D\right)$

> Output $y = g(\mathbf{x})$

> Observations $(\mathbf{x}_i, y_i)_{i=1,\ldots,n}$ $X_s = \left[\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T\right]^T$ $Y_s = \left[y_1, \ldots, y_n\right]^T$

Model: Output seen as realization of stationary Gaussian process

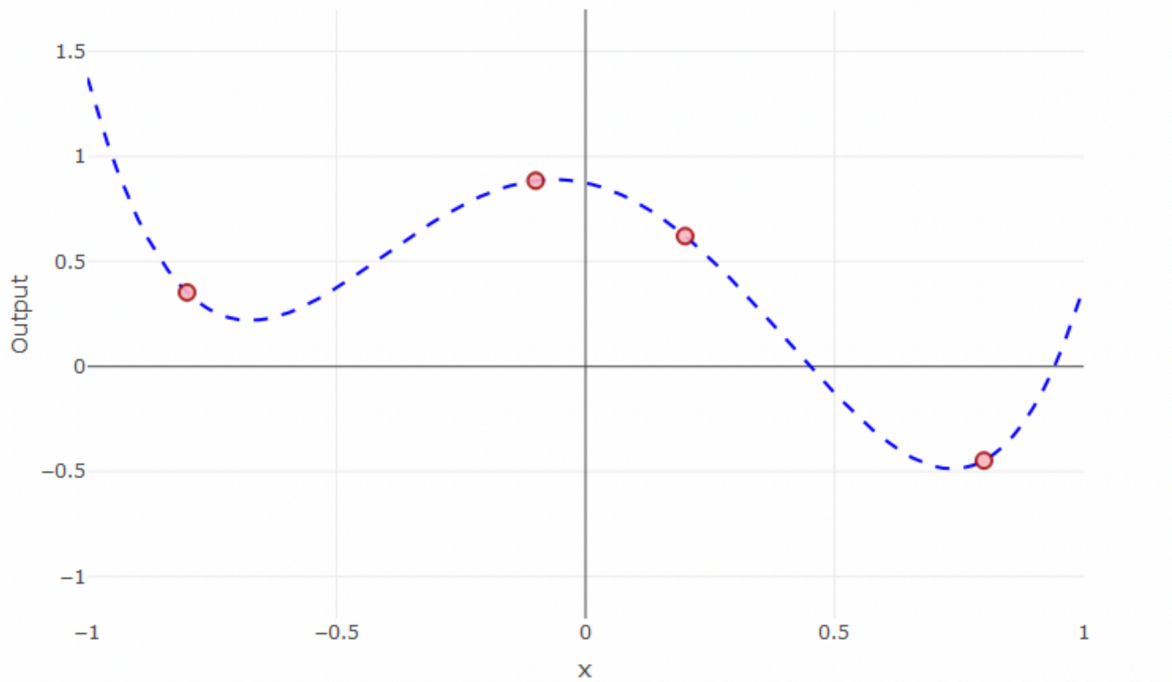
$$Y(\mathbf{x}) = g_0(\mathbf{x}) + U(\mathbf{x}) \qquad g_0(\mathbf{x}) = \sum_{j=1}^{J} \beta_j g_j(\mathbf{x}) = G(\mathbf{x})\beta$$

$$C(\mathbf{x}, \mathbf{x}') = \sigma^2 R(\mathbf{x}, \mathbf{x}')$$

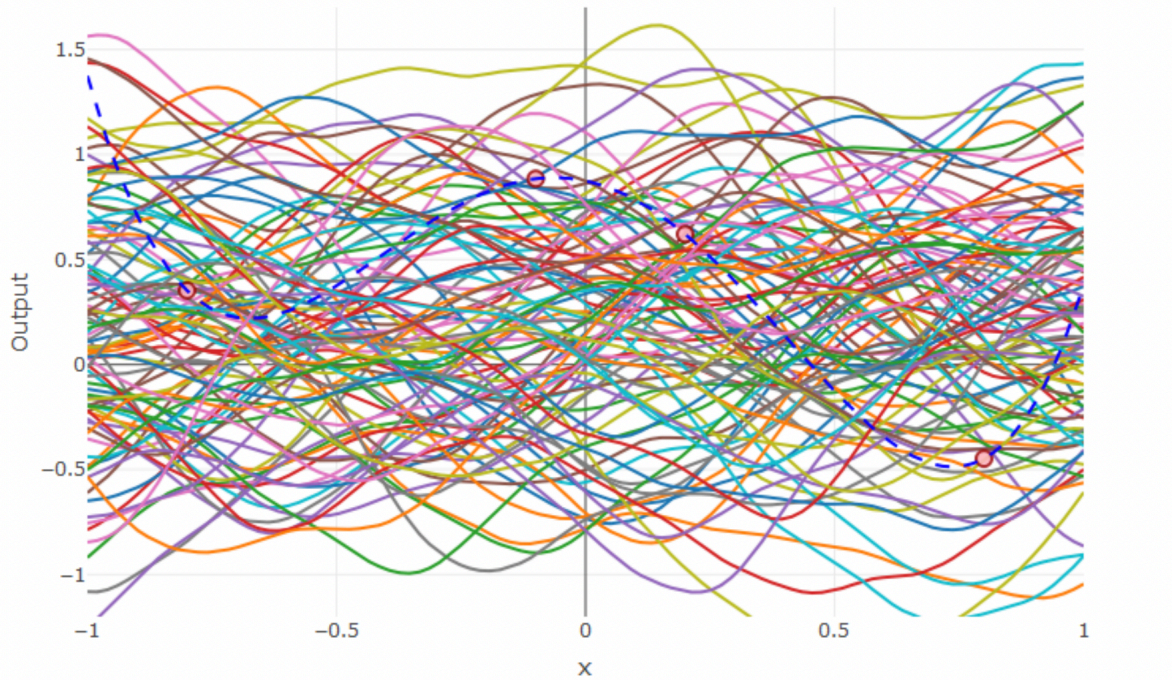
Conditioning on the observations

$$\tilde{Y}(\mathbf{x}^*) = \left[Y(\mathbf{x}^*) | Y(X_s) = Y_s\right]$$

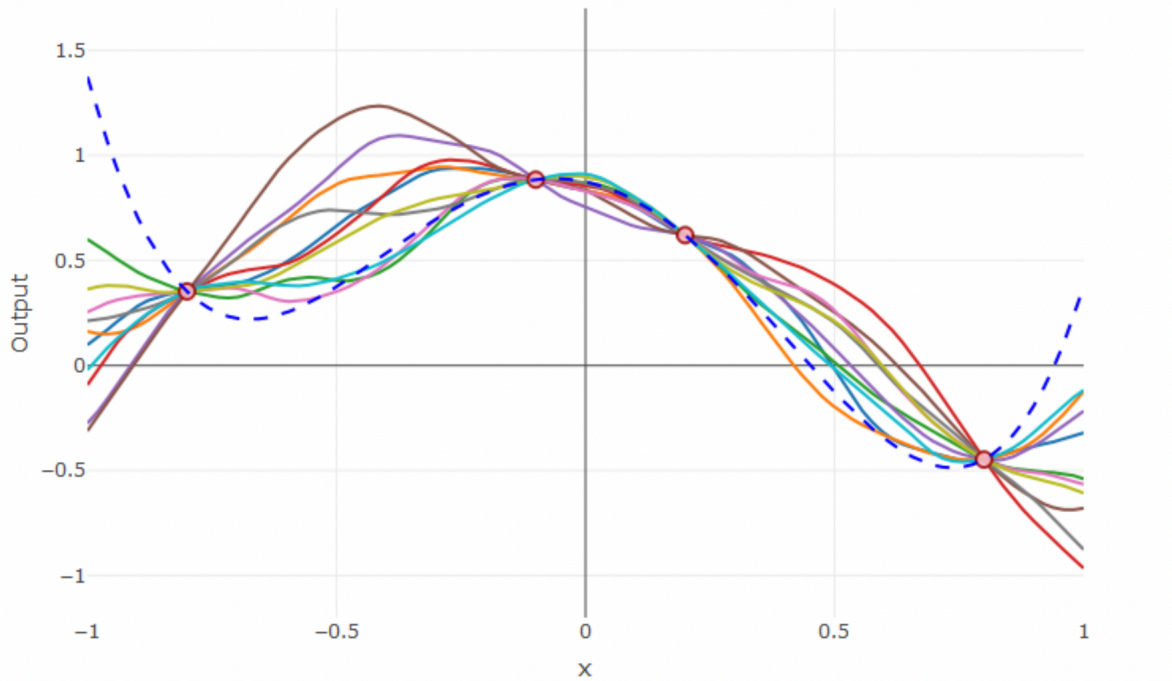
Standard GP regression



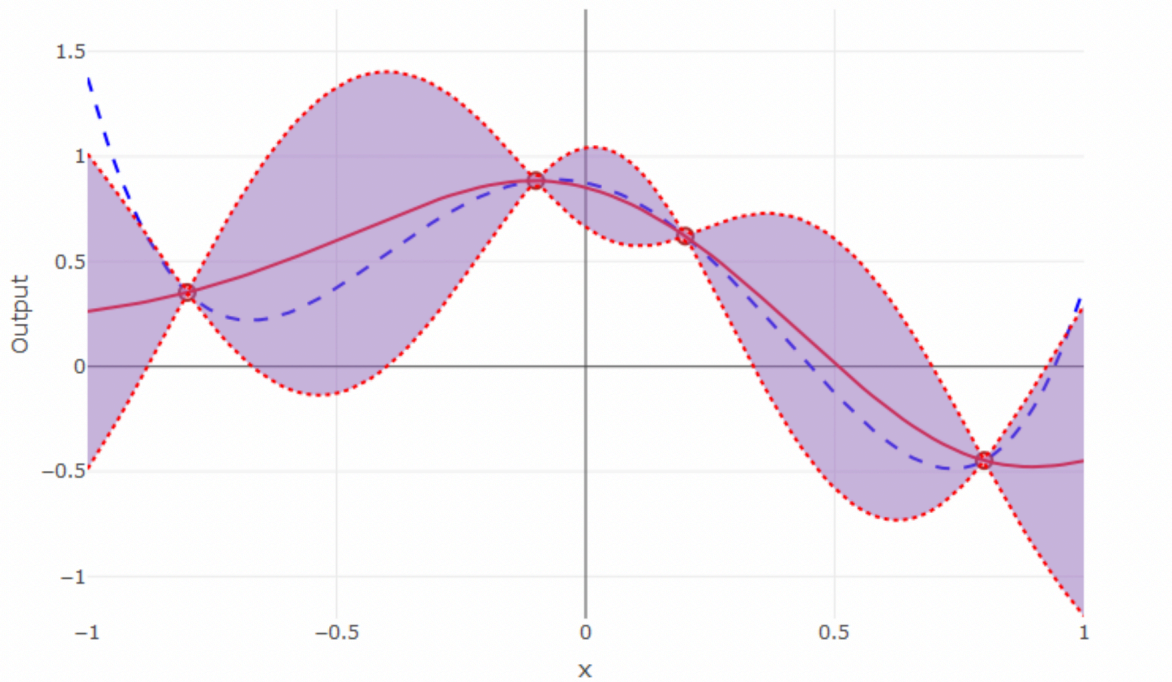
Standard GP regression



Standard GP regression



Standard GP regression



Standard GP regression

Very popular in industrial applications

- Showed great operational success on thousands of real test cases in moderate dimension (~ 15-20 inputs)
- Convenient Bayesian paradigm to propose additional simulations for sequential DOEs targeting a specific goal

Challenge 1: deal with high-dimensional inputs

- Coupling with sparsity (Yi et al. 2011, Cao et al. 2022)
- Coupling with sensitivity analysis tools from Part 3 (Iooss & Marrel 2019)

- Additive kernels with sparsity (SS-ANOVA, Cosso)
- Transformations to be as close as possible from an additive model (Lin & Joseph 2020)

- Transformations via normalizing flows (Maroñas et al. 2021)

$$y = g^{-1}\{\mu + z_1(x_1) + \ldots + z_p(x_p)\}$$
$$z_k(x_k) \sim GP(0, \tau_k^2 R_k(\cdot)),$$

Standard GP regression

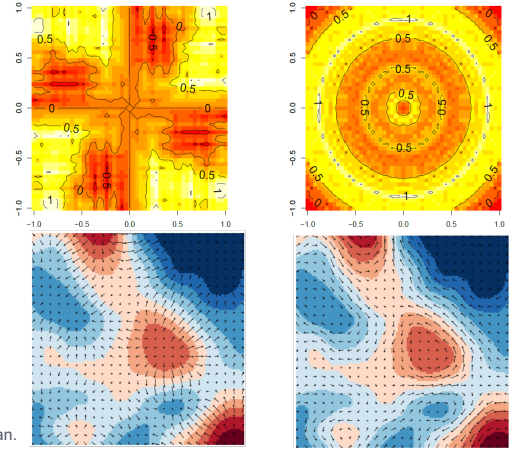
Challenge 2: incorporate constraints from physics

- Very often, computer codes simulate real physical phenomena, which usually have specific properties
  - Symmetries
  - Bound constraints (e.g. concentrations between 0 and 1, …)
  - Monotonicity w.r.t. some input variables
  - Solutions of PDEs (e.g. null Laplacian, divergence or curl free, …)

- It is of great interest to incorporate such constraints in the proxy model
  - Physics and expected behavior are respected (engineers like that !)
  - Predictions and robustness may be improved

- Linear **equality** constraints are easy to handle in the GP framework (Ginsbourger et al. 2013, Scheuerer and Schlather 2012)

- **But bounds and monotonicity are inequality ones**



GP regression with inequality constraints

To incorporate the constraints, we propose to keep the conditional expectation framework

> Predictions are equal to the expectation of the GP (conditioned at the observations) given that it respects the inequality constraints

For example, the corresponding predictor for bound or monotonicity constraints may be

$$\mathbb{E}\left(\tilde{Y}(\mathbf{x}^*)|\forall \mathbf{x}\in I, a\leq \tilde{Y}(\mathbf{x})\leq b\right) \qquad \mathbb{E}\left(\tilde{Y}(\mathbf{x}^*)|\forall \mathbf{x}\in I, \frac{\partial \tilde{Y}}{\partial x^j}(\mathbf{x})\geq 0\right)$$

> Note the link with with extrema of random fields …

$$\mathbb{E}\left(\tilde{Y}(\mathbf{x}^*)|\min_{\mathbf{x}\in I}\tilde{Y}(\mathbf{x})\geqslant a, \max_{\mathbf{x}\in I}\tilde{Y}(\mathbf{x})\leqslant b\right)$$

> … but no tractable formula exists for joint distributions in the general case

GP regression with inequality constraints

We thus propose a discrete-location approximation:

$$\mathbb{E}\left(\tilde{Y}(\mathbf{x}^*)|\forall i=1,\dots,N, a\leq\tilde{Y}(\mathbf{x}_i)\leq b\right) \quad \mathbb{E}\left(\tilde{Y}(\mathbf{x}^*)|\forall i=1,\dots,N, \frac{\partial\tilde{Y}}{\partial x^j}(\mathbf{x}_i)\geq 0\right)$$

> Same approximation in Riihimaki and Vehtari 2010, Wang and Berger 2011

This generalizes easily to other constraints

$$\mathbb{E}\left(\tilde{Y}(\mathbf{x}^*)|\ \forall k=1,\dots,K,\ \forall i=1,\dots,N_k,\ a_i^{(k)}\leq Z^{(k)}(\mathbf{x}_i^{(k)})\leq b_i^{(k)}\right)$$

$$Z^{(k)}=\mathcal{L}^{(k)}\left[\tilde{Y}\right]$$

GP regression with inequality constraints

We thus propose a discrete-location approximation:

$$\mathbb{E}\left(\tilde{Y}(\mathbf{x}^*)|\forall i=1,\dots,N, a \leq \tilde{Y}(\mathbf{x}_i) \leq b\right) \quad \mathbb{E}\left(\tilde{Y}(\mathbf{x}^*)|\forall i=1,\dots,N, \frac{\partial \tilde{Y}}{\partial x^j}(\mathbf{x}_i) \geq 0\right)$$

> Same approximation in Riihimaki and Vehtari 2010, Wang and Berger 2011

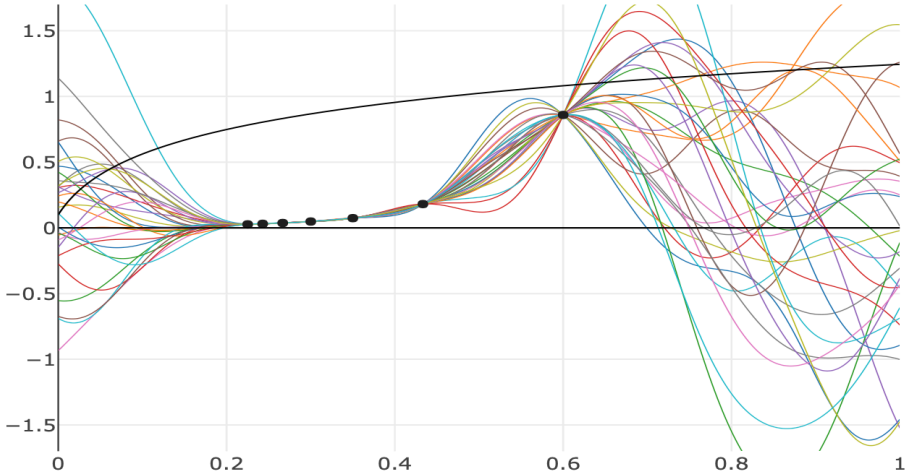
This generalizes easily to other constraints

$$\mathbb{E}\left(\tilde{Y}(\mathbf{x}^*)|\mathbf{a} \leq \mathbf{Z} \leq \mathbf{b}\right)$$

GP regression with inequality constraints

Standard framework:

- Take all trajectories which interpolate the observations
- Compute the average to get the kriging predictor
- (If desired, the variance yields a measure of accuracy)



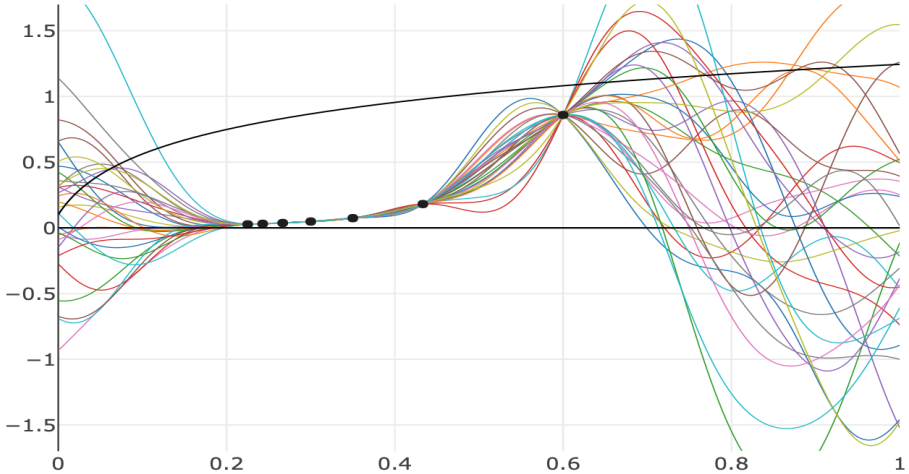
Agrell 2019

SAFRAN

GP regression with inequality constraints

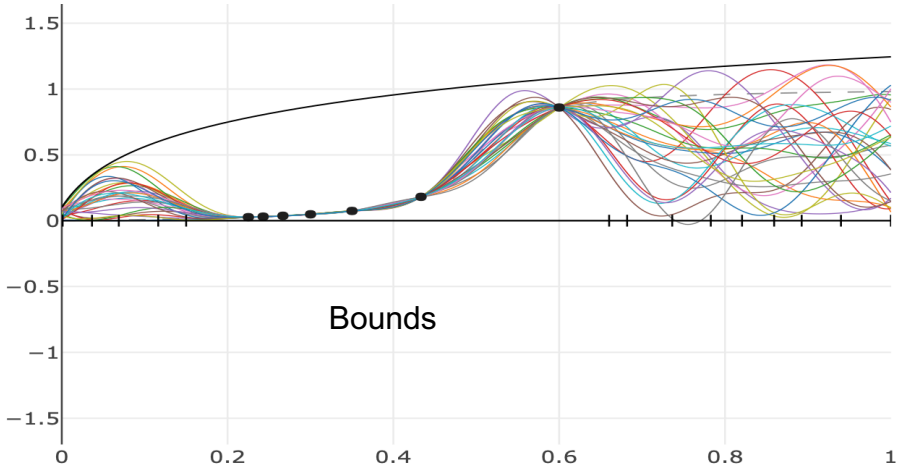
Standard framework:

- Take all trajectories which interpolate the observations
- Compute the average to get the kriging predictor
- (If desired, the variance yields a measure of accuracy)



Here:

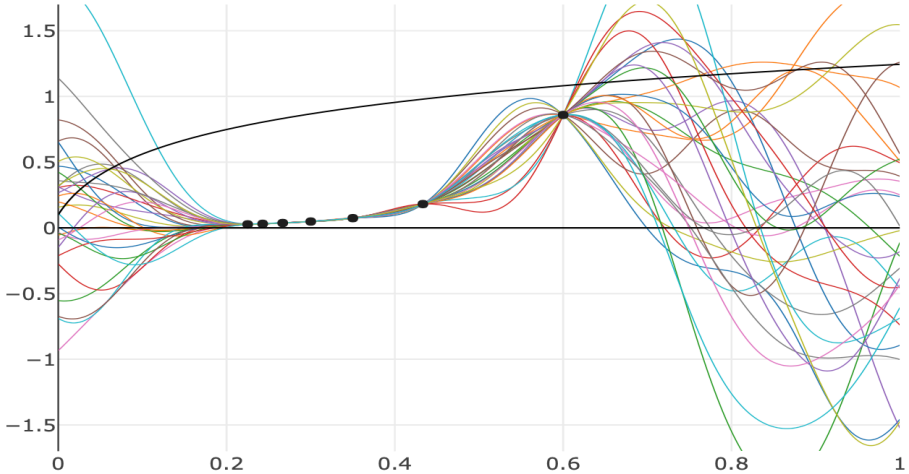
- Take all trajectories which interpolate the observations
- Select those which respect the constraints of bounds, monotonicity, …
- Compute the average to get the new kriging predictor
- (If desired, the variance yields a measure of accuracy)



GP regression with inequality constraints

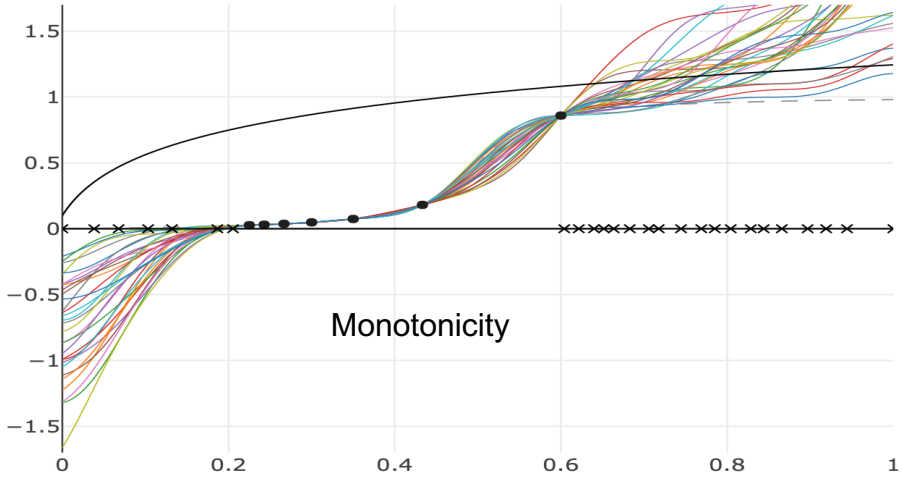
Standard framework:

- Take all trajectories which interpolate the observations
- Compute the average to get the kriging predictor
- (If desired, the variance yields a measure of accuracy)



Here:

- Take all trajectories which interpolate the observations
- Select those which respect the constraints of bounds, monotonicity, …
- Compute the average to get the new kriging predictor
- (If desired, the variance yields a measure of accuracy)



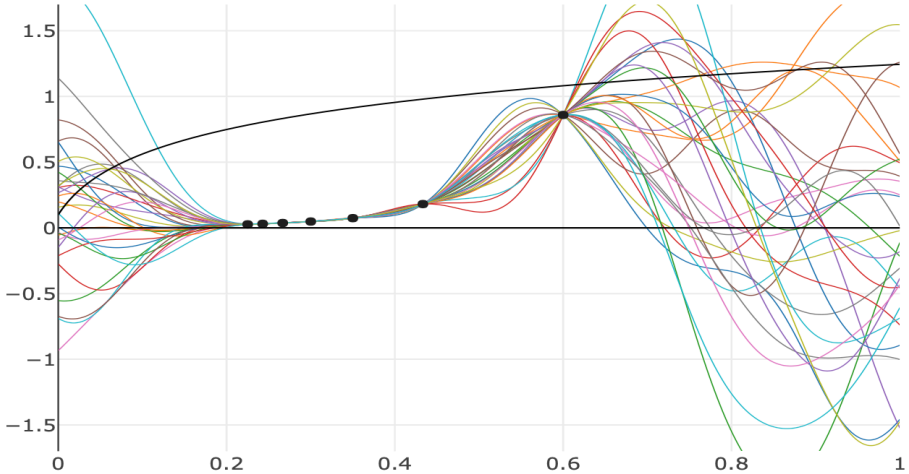
Agrell 2019

SAFRAN

GP regression with inequality constraints

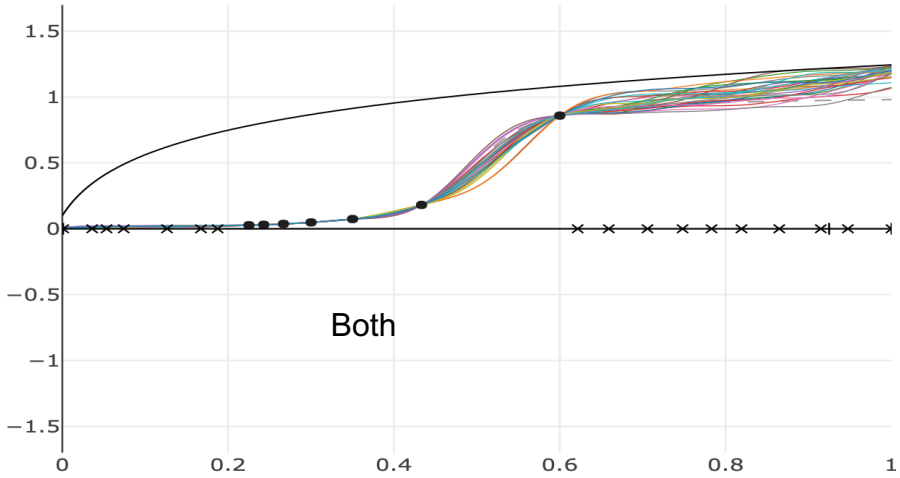
Standard framework:

- Take all trajectories which interpolate the observations
- Compute the average to get the kriging predictor
- (If desired, the variance yields a measure of accuracy)



Here:

- Take all trajectories which interpolate the observations
- Select those which respect the constraints of bounds, monotonicity, …
- Compute the average to get the new kriging predictor
- (If desired, the variance yields a measure of accuracy)



Agrell 2019

SAFRAN

GP regression with inequality constraints

But how can we compute such expectations ?

This is where the linearity assumption comes into play

- Bounds, monotonicity, integral, divergence/curl constraints are linear w.r.t. the output
- The GP obtained by stacking the output and the quantities related to the constraints is then a GP too

$$\mathbf{W} = \left(\tilde{Y}(\mathbf{x}^*), \mathbf{Z}\right) \sim \mathcal{N}\left(\begin{bmatrix} \tilde{\mu}(\mathbf{x}^*) \\ \mu_{\mathbf{Z}} \end{bmatrix}, \begin{bmatrix} \tilde{C}(\mathbf{x}^*, \mathbf{x}^*) & \Sigma_{\tilde{Y}\mathbf{Z}} \\ \Sigma_{\tilde{Y}\mathbf{Z}}^T & \Sigma_{\mathbf{Z}} \end{bmatrix}\right)$$

Kotz et al. 2010

$$\mathbb{E}\left(\tilde{Y}(\mathbf{x}^*) | \mathbf{a} \leq \mathbf{Z} \leq \mathbf{b}\right) = \tilde{\mu}(\mathbf{x}^*) + \Sigma_{\tilde{Y}\mathbf{Z}}\Sigma_{\mathbf{Z}}^{-1}\left(\nu_{\mathbf{Z}} - \mu_{\mathbf{Z}}\right)$$

$$\mathrm{Var}\left(\tilde{Y}(\mathbf{x}^*) | \mathbf{a} \leq \mathbf{Z} \leq \mathbf{b}\right) = \tilde{C}(\mathbf{x}^*, \mathbf{x}^*) - \Sigma_{\tilde{Y}\mathbf{Z}}\left(\Sigma_{\mathbf{Z}}^{-1} - \Sigma_{\mathbf{Z}}^{-1}\Gamma_{\mathbf{Z}}\Sigma_{\mathbf{Z}}^{-1}\right)\Sigma_{\tilde{Y}\mathbf{Z}}^T$$

GP regression with inequality constraints

$$\mathbb{E}\left(\tilde{Y}(\mathbf{x}^*)|\mathbf{a} \leq \mathbf{Z} \leq \mathbf{b}\right) = \tilde{\mu}(\mathbf{x}^*) + \Sigma_{\tilde{Y}\mathbf{Z}}\Sigma_{\mathbf{Z}}^{-1}\left(\nu_{\mathbf{Z}} - \mu_{\mathbf{Z}}\right)$$

$$\mathrm{Var}\left(\tilde{Y}(\mathbf{x}^*)|\mathbf{a} \leq \mathbf{Z} \leq \mathbf{b}\right) = \tilde{C}(\mathbf{x}^*, \mathbf{x}^*) - \Sigma_{\tilde{Y}\mathbf{Z}}\left(\Sigma_{\mathbf{Z}}^{-1} - \Sigma_{\mathbf{Z}}^{-1}\Gamma_{\mathbf{Z}}\Sigma_{\mathbf{Z}}^{-1}\right)\Sigma_{\tilde{Y}\mathbf{Z}}^{T}$$

$$\nu_{\mathbf{Z}} = \mathbb{E}(\mathbf{Z}|\mathbf{a} \leq \mathbf{Z} \leq \mathbf{b})$$

$$\Gamma_{\mathbf{Z}} = \mathrm{Var}(\mathbf{Z}|\mathbf{a} \leq \mathbf{Z} \leq \mathbf{b})$$

> The problem reduces to compute **moments of a multivariate normal vector subject to linear inequality constraints**

> Key object is then the **truncated normal distribution**

The truncated multivariate normal distribution

The expectation and variance are our goal here

Available formulas involve Gaussian integrals with dimensionality equal to the number of points where we impose the constraints

We thus need efficient approximations when this number is large (as it should be !)

- Genz numerical approximation of Gaussian integrals (Genz 1992)
- Sampling from a truncated Gaussian
- Correlation-free formula (« crude » covariance tapering)

$$
\begin{aligned}
\mathbb{E}(Z_i|\mathbf{a} \leq \mathbf{Z} \leq \mathbf{b}) &\approx \mathbb{E}(Z_i|a_i \leq Z_i \leq b_i) \\
&\approx \mu_i + \sigma_i \frac{\phi(\tilde{a}_i) - \phi(\tilde{b}_i)}{\Phi(\tilde{b}_i) - \Phi(\tilde{a}_i)} \\
\mathrm{Var}(Z_i|\mathbf{a} \leq \mathbf{Z} \leq \mathbf{b}) &\approx \mathrm{Var}(Z_i|a_i \leq Z_i \leq b_i) \\
&\approx \sigma_i^2 \left[1 + \frac{\tilde{a}_i \phi(\tilde{a}_i) - \tilde{b}_i \phi(\tilde{b}_i)}{\Phi(\tilde{b}_i) - \Phi(\tilde{a}_i)} - \left(\frac{\phi(\tilde{a}_i) - \phi(\tilde{b}_i)}{\Phi(\tilde{b}_i) - \Phi(\tilde{a}_i)} \right)^2 \right]
\end{aligned}
$$

Going from simple examples to realistic industrial applications

Efficient generalization to higher dimensional problems is not so easy

- From a theoretical perspective, no change in the formulas
  - However, « spanning » the subset where we impose constraints will necessitate much more constraint points in the discrete-location approximation
  - Genz numerical integration and sampling cannot be used with tens of thousands of constraints

- The idea is to use the correlation induced among the constraint points (and with the observations)
  - **It is not necessary to place constraint points where the predictor has a high probability to respect the constraints (e.g. close to another constraint point, or where the prediction variance is very low)**

Adaptive strategy for the constraint locations

This motivates the design of an adaptive strategy for choosing the constraints locations

The key here is to compute the probability that the constrained predictor does not respect the constraint at any point

- **Obviously this is not a nice and friendly normal distribution as in standard GP regression**
- It involves the CDF of a truncated normal distribution, two ways to handle it:
  - Use a truncated normal sampling algorithms (Agrell 2019, Perrin and D. 2021)
  - Make a crude but fast normal approximation (D. and Marrel 2019)

Constraint points are thus added one at a time, at locations where this probability is the highest

Recent subsequent improvements

Other adaptive criteria can be used (Perrin and D. 2021)

- A variant inspired by Expected Improvement
- Two extensions to go from point-wise to integral criteria
- Integral criteria seem to perform better

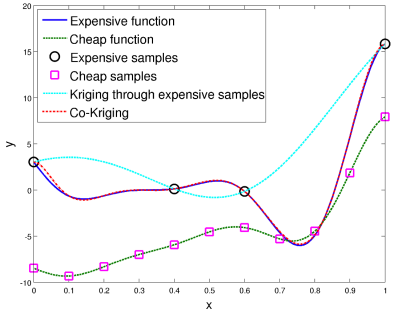
Hyperparameters estimation

- Here we estimate it on the initial GP, **so we don't account for the constraints**
- This leads to potential unexpected behaviours for the final constrained predictor
  - In particular we may need to add more locations for constraints than usual
- Recent works on MLE with constraints: do not appear to lessen this problem as much as expected
- Promising solution: write a MLE maximization with constraints (Perrin and D. 2021)

Standard GP regression

Challenge 3: make use of all available simulation data

- Sometimes, simpler but cheaper simulation models are available
  - The idea is then to combine datasets coming from different simulators to build a more predictive surrogate model, this is called a mutli-fidelity surrogate model
  - For GP, often based on co-kriging (Le Gratiet & Garnier 2014)



Zhang et al. 2013

- It also happens that during a complex design process, input variables are added sequentially
  - Consequently, several DOE with different dimensions are available
  - Can we combine them? → see T. Gonon's talk in this session!

2

DESIGN OF EXPERIMENTS WITH KERNELS

Design of experiments for computer simulations

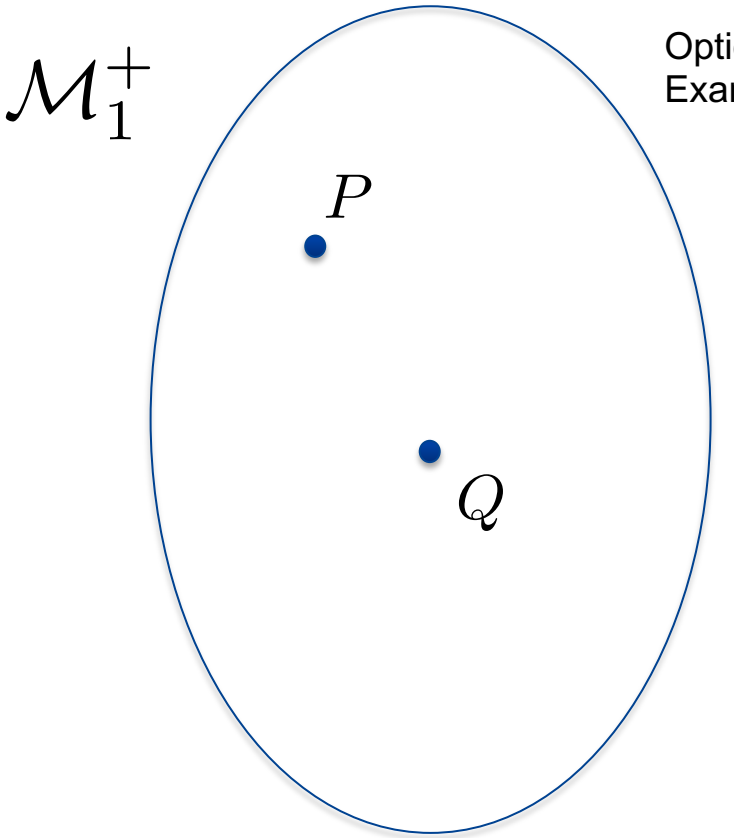
Space-filling designs are extremely popular

- QMC such as Sobol' or Halton sequences
- Optimized designs based on
  - Geometrical properties: minimax, maximin, maximum projection
  - **Discrepancy measures: distance of the DOE to the uniform distribution in the hypercube**

Generalization of standard discrepancies with kernels

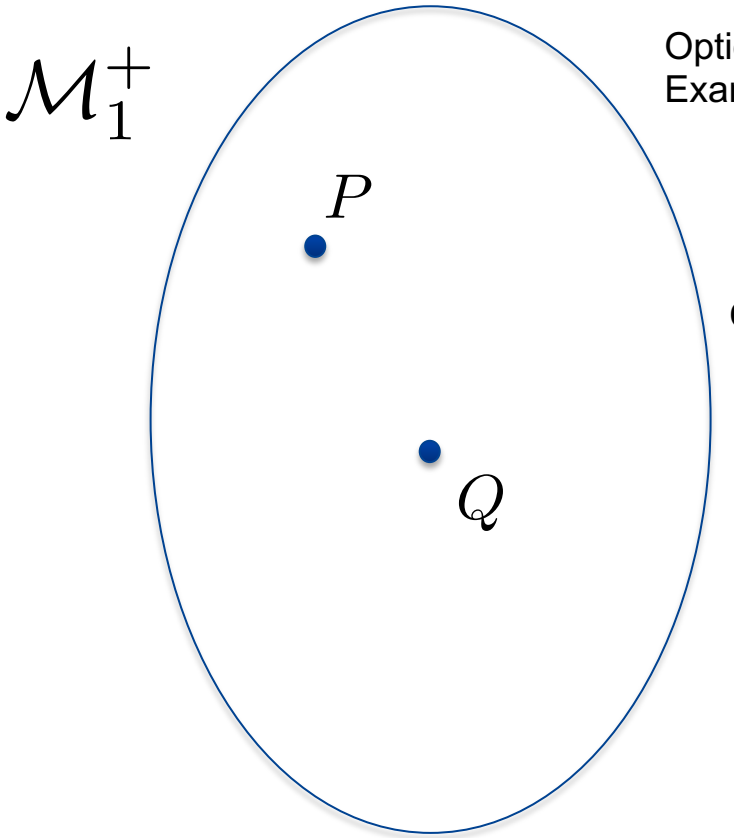
- The main idea is to change the distance between the empirical distribution of the DOE and the target uniform distribution
- This distance between probabillity distributions relies on kernel embeddings of distributions

Kernel-embedding of probability distributions



Option 1: work directly in the space of probability measures
Examples: KS, TV, KL, Hellinger, …

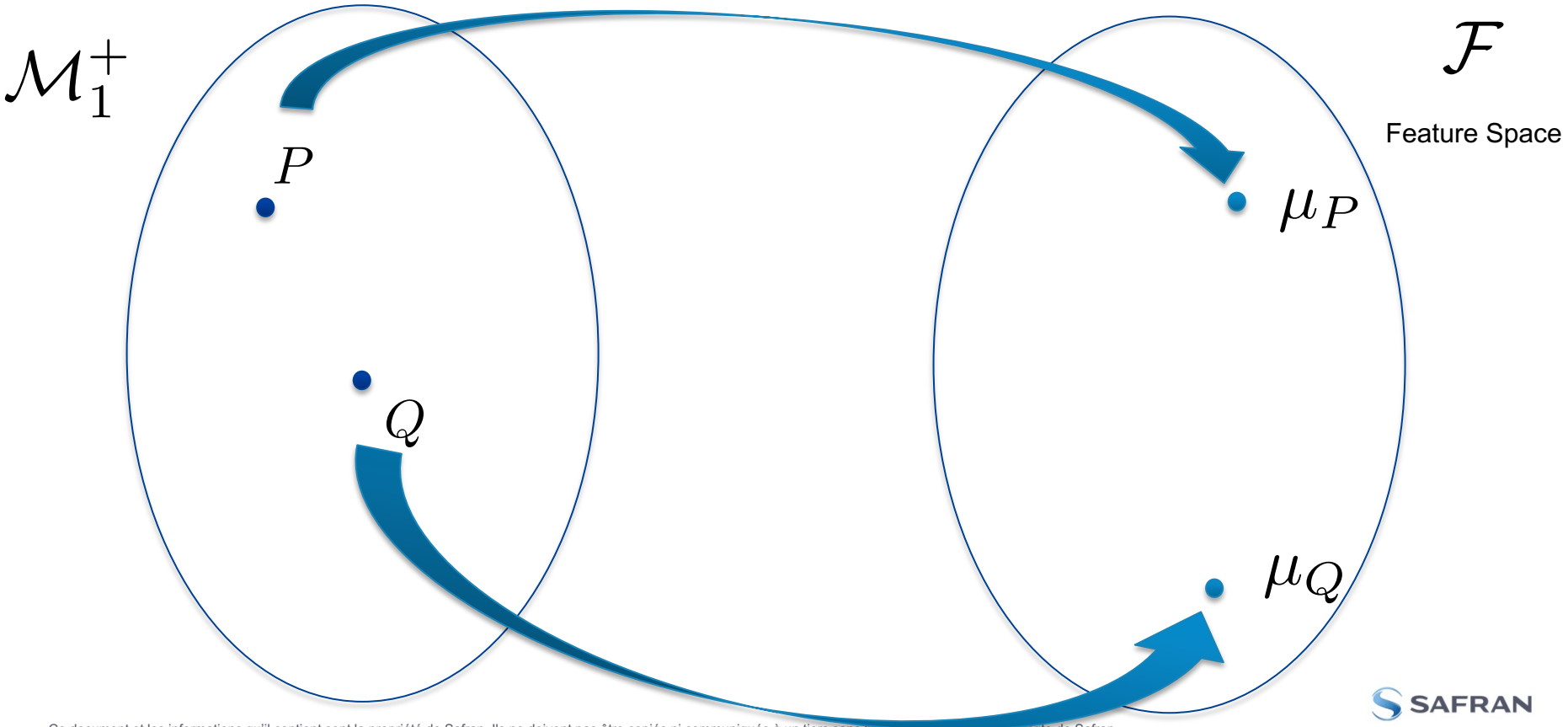
Kernel-embedding of probability distributions



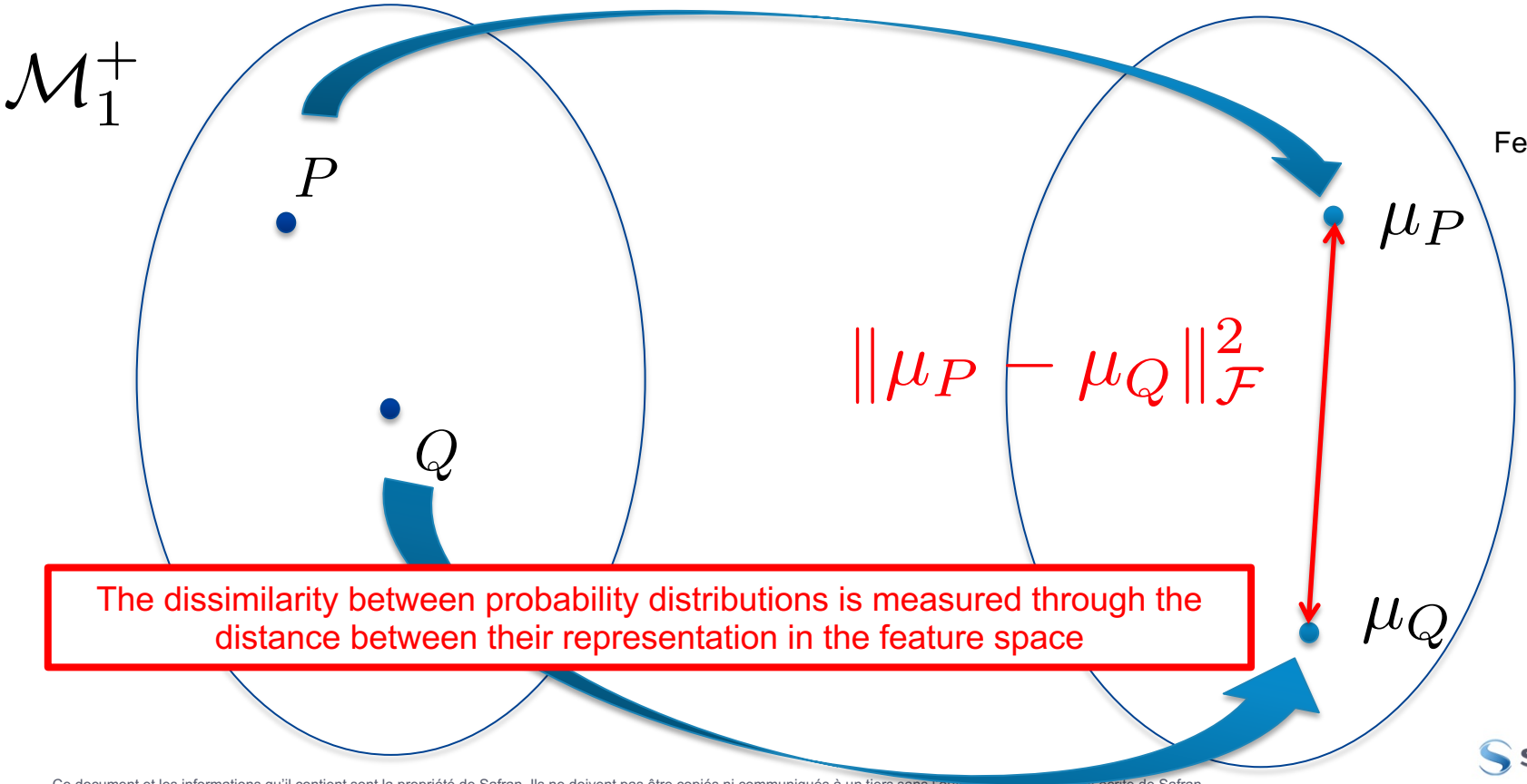
Option 1: work directly in the space of probability measures
Examples: KS, TV, KL, Hellinger, …

Option 2: represent probability measures with some features

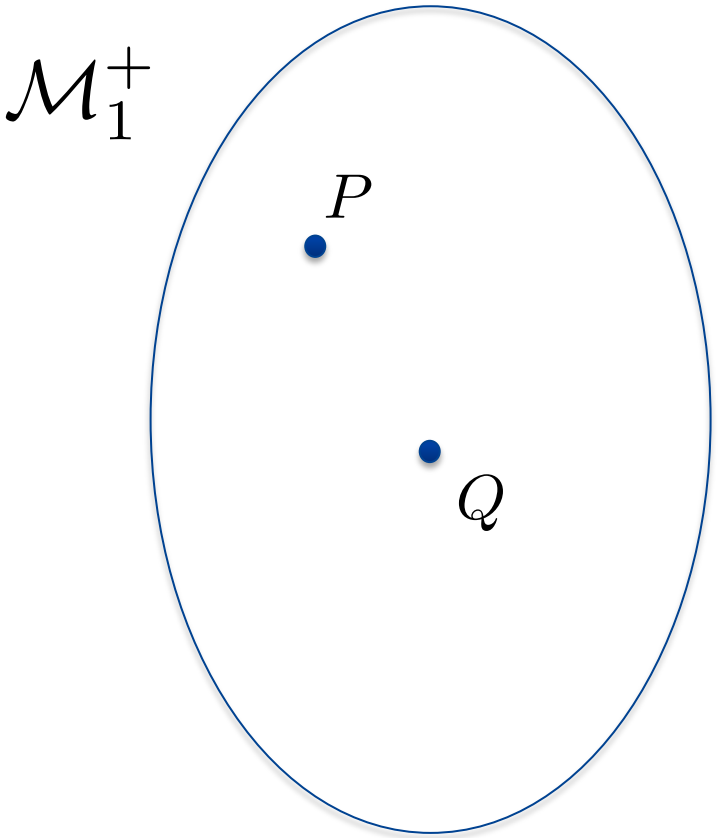
Kernel-embedding of probability distributions

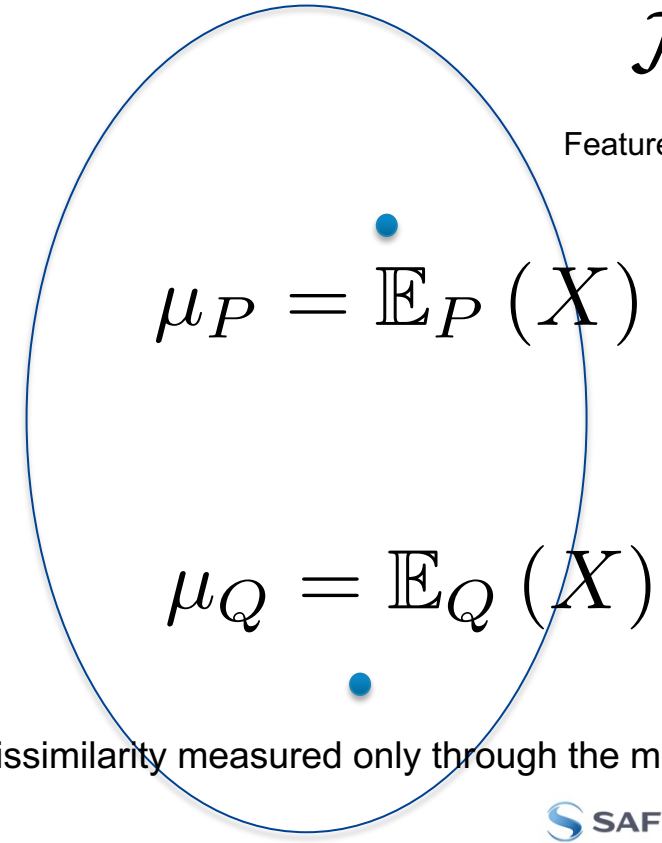


Kernel-embedding of probability distributions



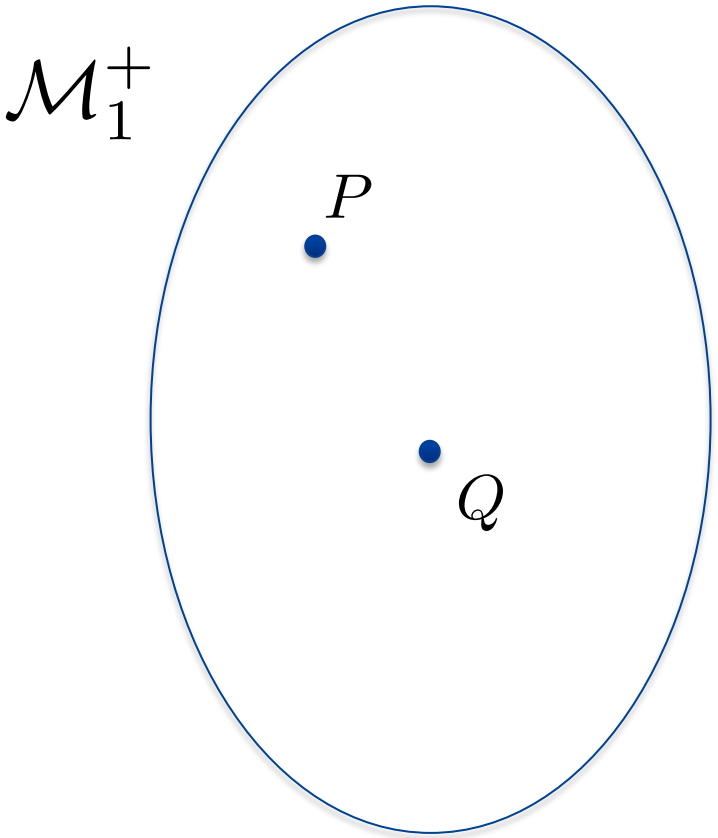
Kernel-embedding of probability distributions

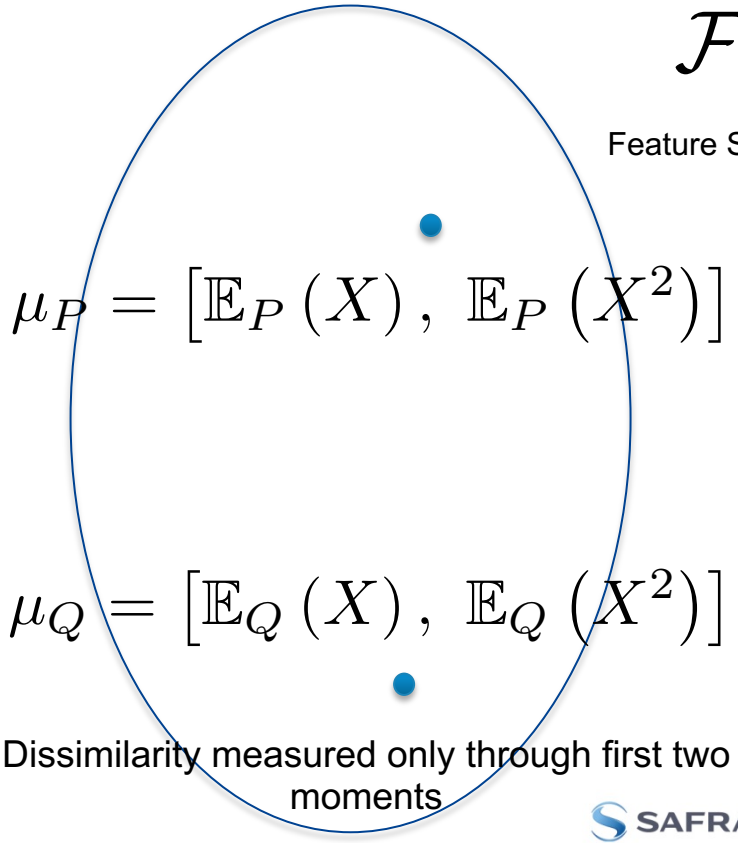




Dissimilarity measured only through the means

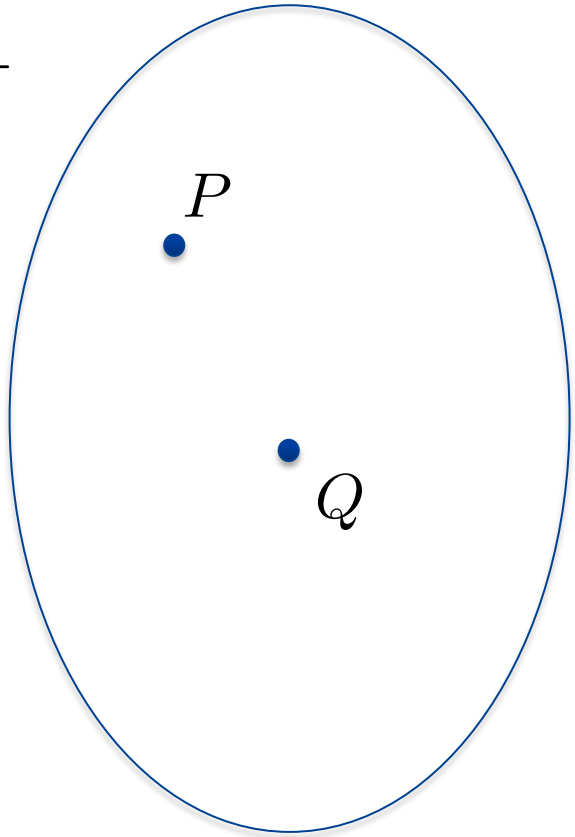
Kernel-embedding of probability distributions

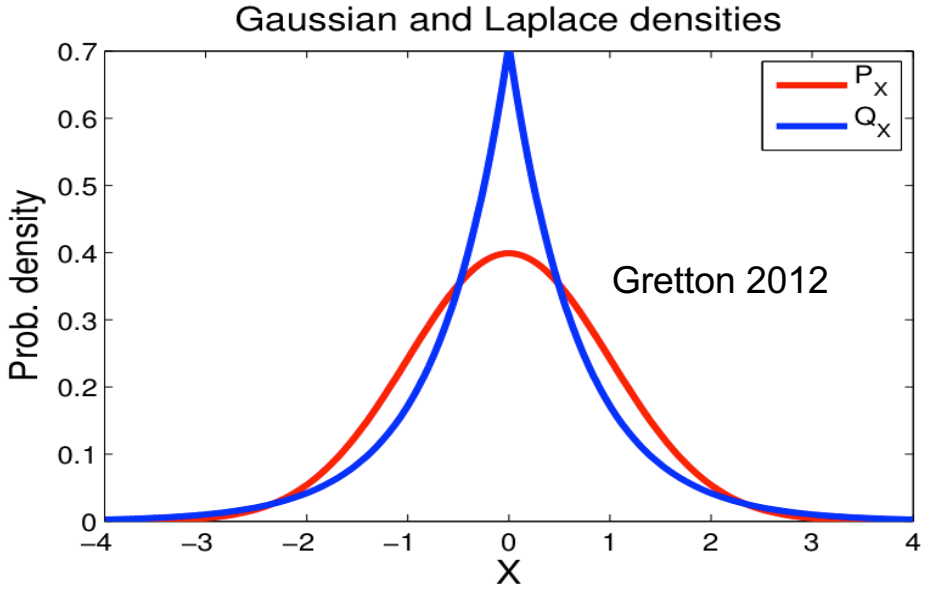




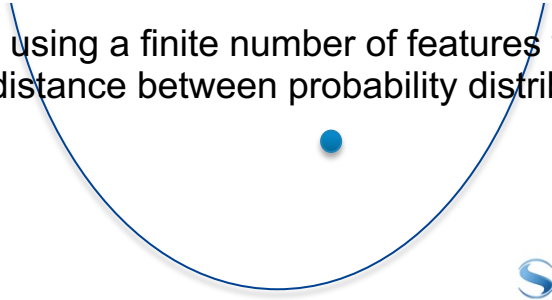
Kernel-embedding of probability distributions

$\mathcal{M}_1^+$



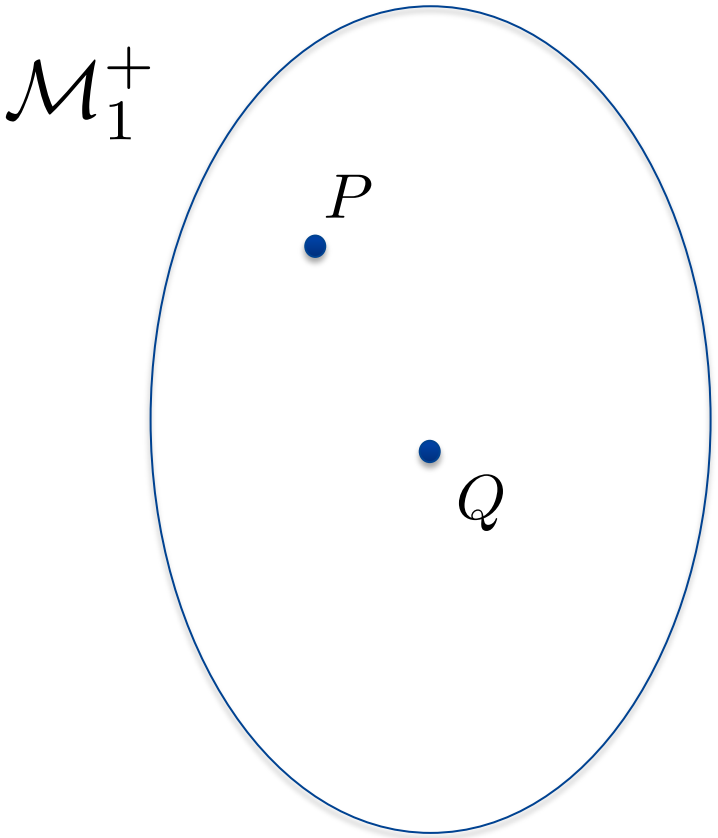


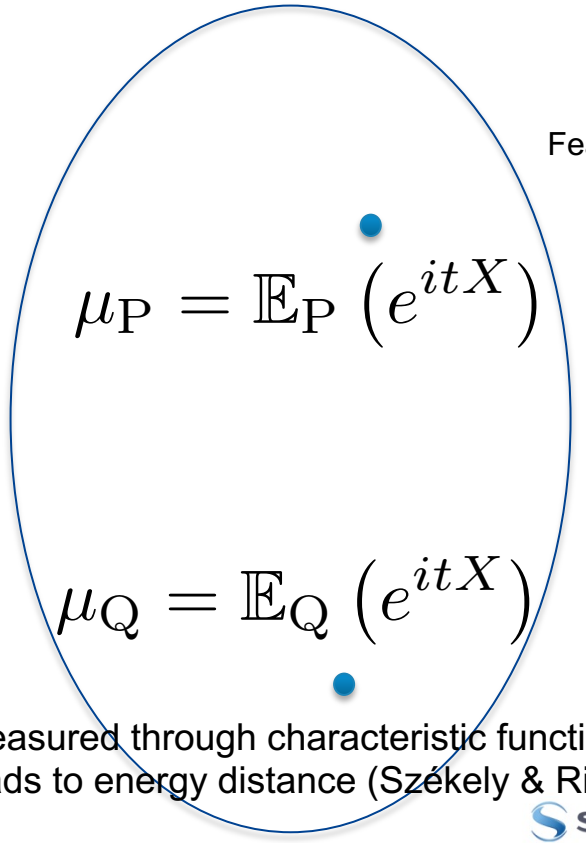
Obviously using a finite number of features will not lead to a distance between probability distributions



SAFRAN

Kernel-embedding of probability distributions

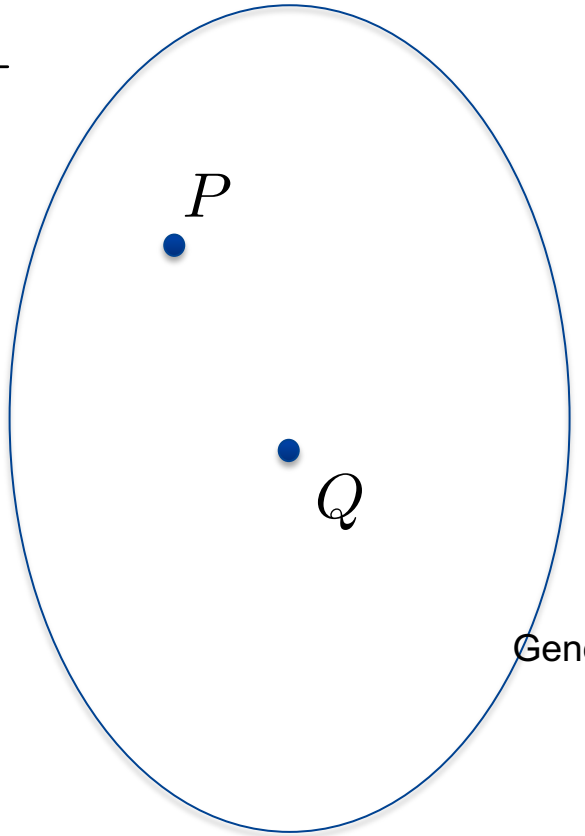




Dissimilarity measured through characteristic functions
Weighted distance leads to energy distance (Székely & Rizzo 2013)

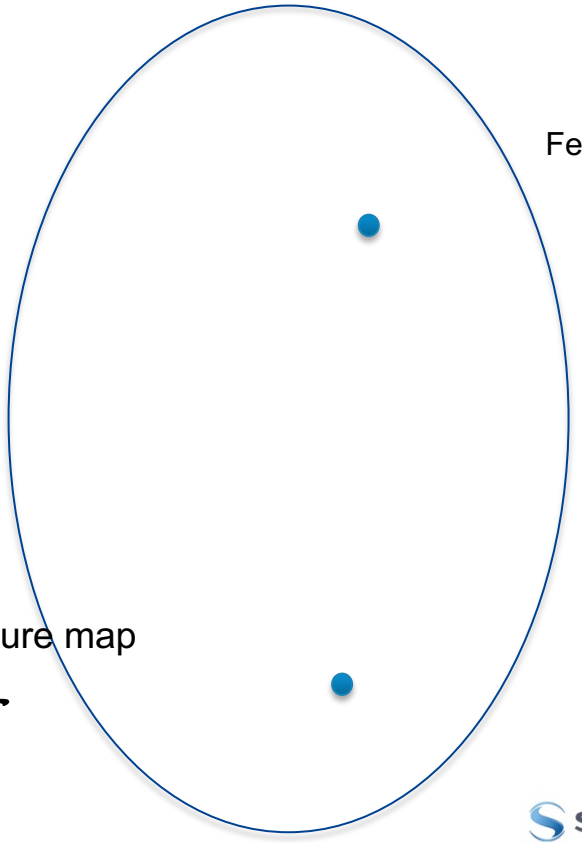
Kernel-embedding of probability distributions

$\mathcal{M}_1^+$



$\mathcal{F}$

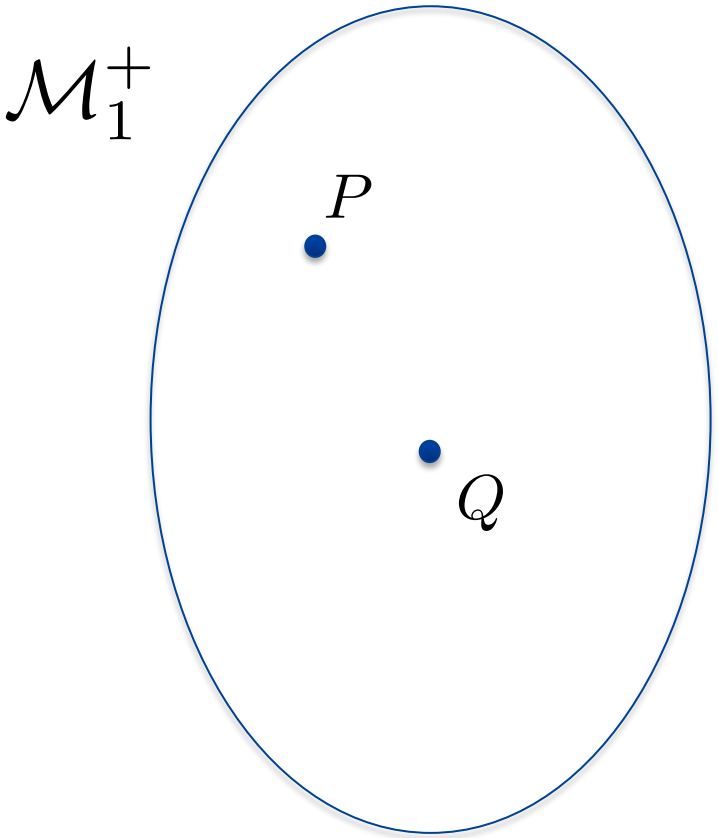
Feature Space

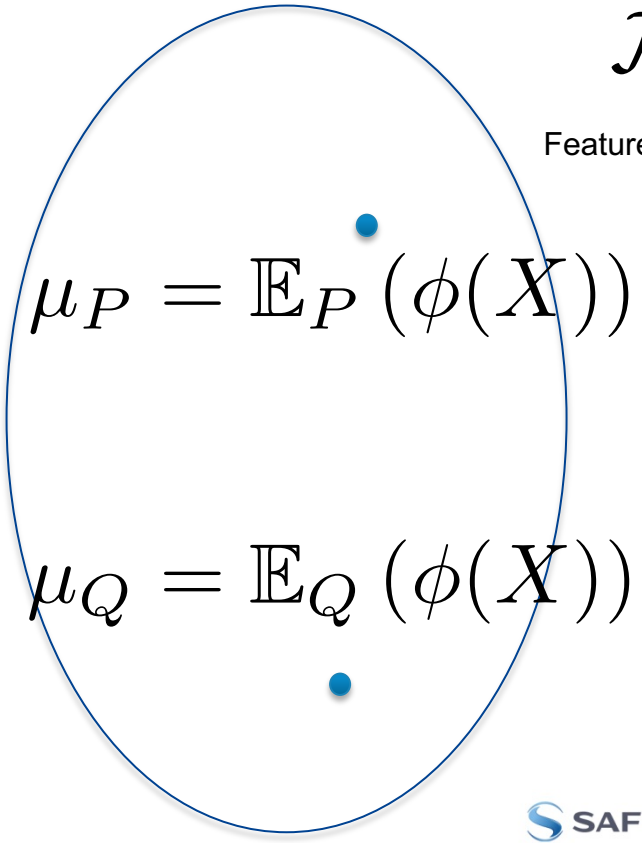


General setting: take a feature map

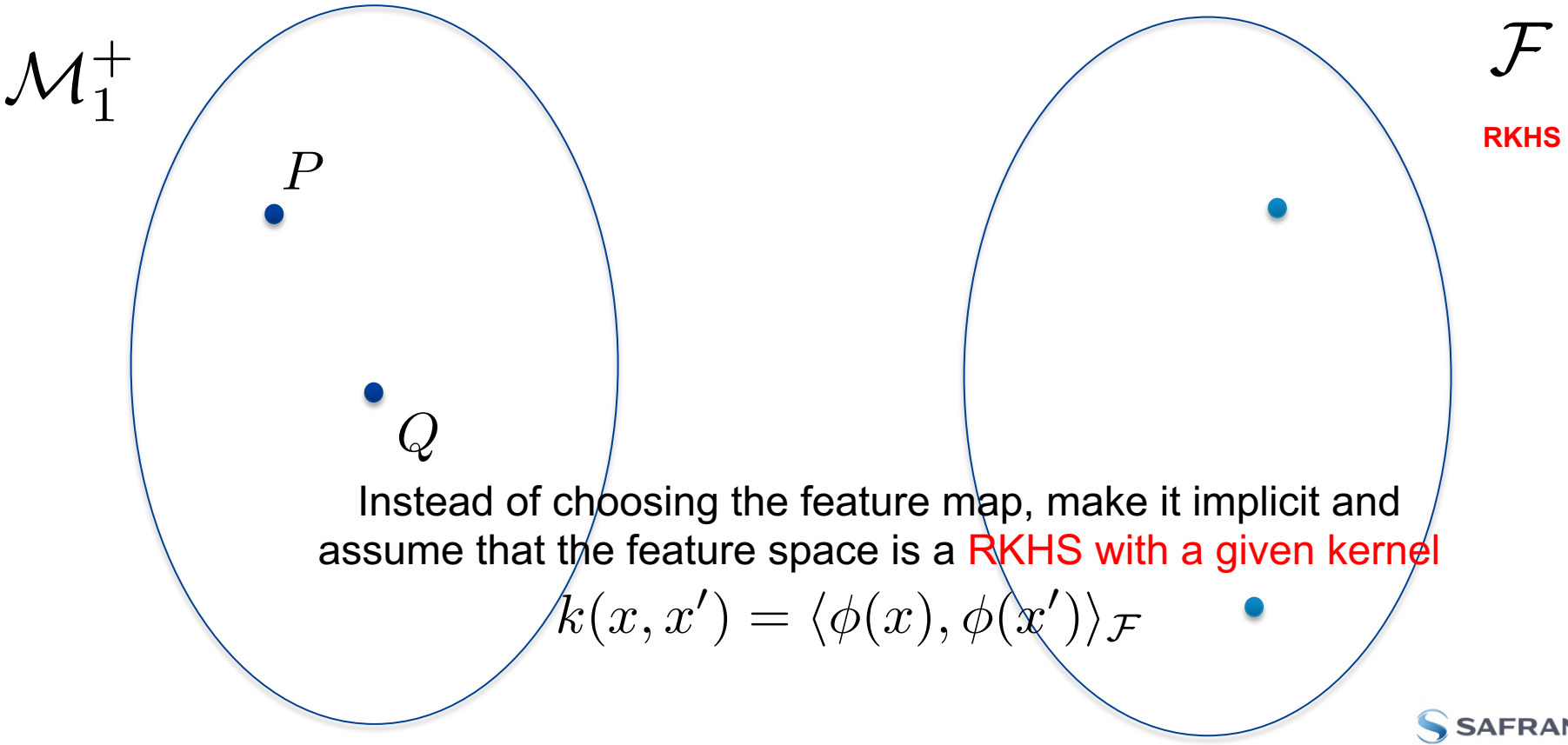
$$\phi : \Omega \to \mathcal{F}$$

Kernel-embedding of probability distributions





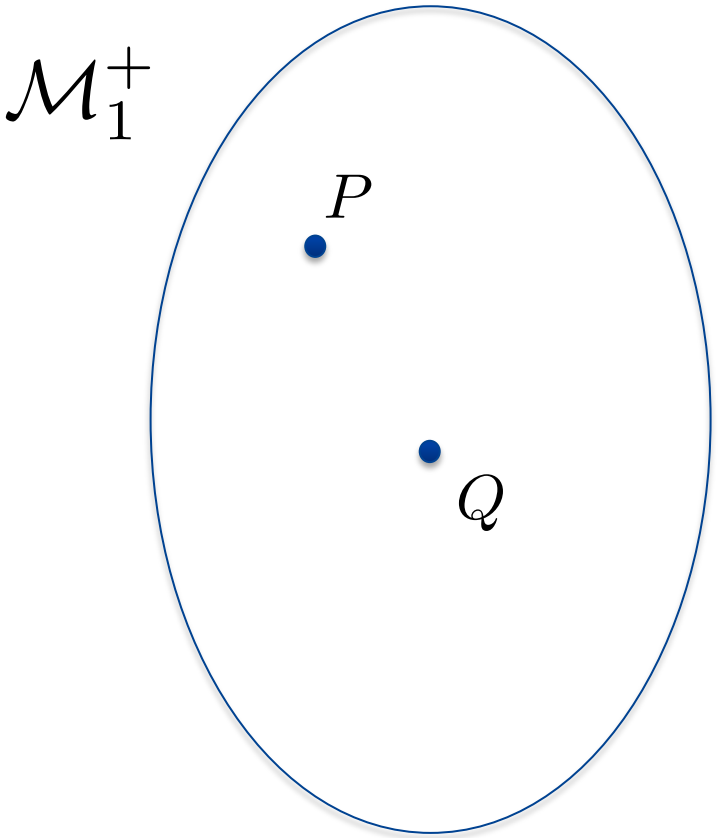
Kernel-embedding of probability distributions

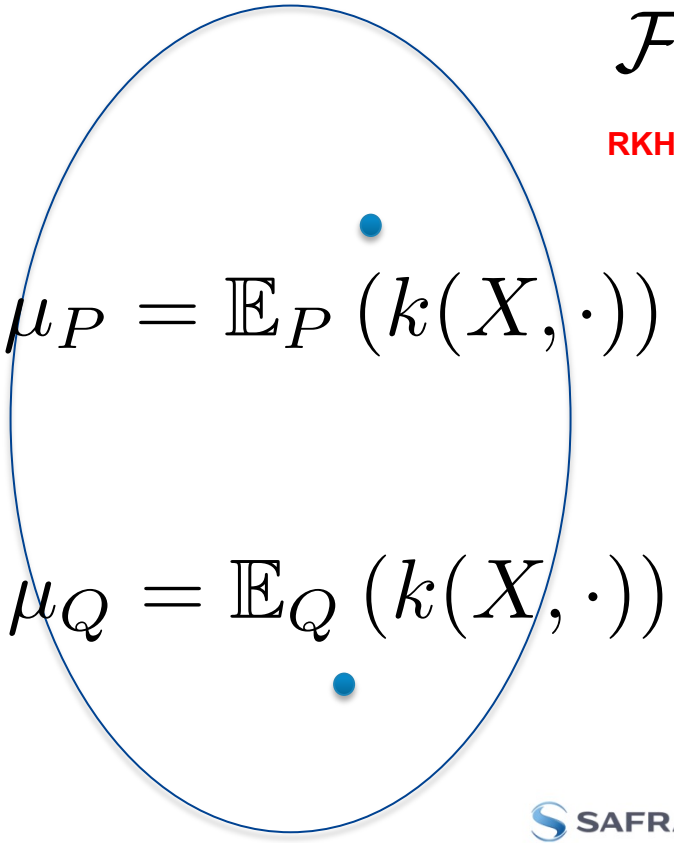


Instead of choosing the feature map, make it implicit and assume that the feature space is a RKHS with a given kernel

$$k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{F}}$$

Kernel-embedding of probability distributions





Kernel-embedding of probability distributions

The kernel mean embedding of a probability measure is defined as

$$\mu_{\mathrm{P}} = \mathbb{E}_{\xi \sim \mathrm{P}} k_{\mathcal{X}}(\xi, \cdot) = \int_{\mathcal{X}} k_{\mathcal{X}}(\xi, \cdot) d\mathrm{P}(\xi)$$

A distance between probability measures is then given by the Maximum Mean Discrepancy

$$\mathrm{MMD}(\mathrm{P}_1, \mathrm{P}_2) = \|\mu_{\mathrm{P}_1} - \mu_{\mathrm{P}_2}\|_{\mathcal{H}}$$

The reproducing property in the RKHS gives the central result

$$\mathrm{MMD}^2(\mathrm{P}_1, \mathrm{P}_2) = \mathbb{E}_{\xi,\xi'} k_{\mathcal{X}}(\xi, \xi') - 2\mathbb{E}_{\xi,\zeta} k_{\mathcal{X}}(\xi, \zeta) + \mathbb{E}_{\zeta,\zeta'} k_{\mathcal{X}}(\zeta, \zeta')$$

Smola et al. 2007, Song 2008, Song et al. 2009

SAFRAN

Kernel-embedding of probability distributions

Other major use: testing independence of random vectors

$$\mathrm{MMD}^2(\mathrm{P}_{\mathbf{UV}}, \mathrm{P}_{\mathbf{U}} \otimes \mathrm{P}_{\mathbf{V}}) = \|\mu_{\mathrm{P}_{\mathbf{UV}}} - \mu_{\mathrm{P}_{\mathbf{U}}} \otimes \mu_{\mathrm{P}_{\mathbf{V}}}\|^2_{\mathcal{H}}$$

$$\begin{aligned}
\mathrm{HSIC}(\mathbf{U}, \mathbf{V}) &= \mathrm{MMD}^2(\mathrm{P}_{\mathbf{UV}}, \mathrm{P}_{\mathbf{U}} \otimes \mathrm{P}_{\mathbf{V}}) \\
&= \mathbb{E}_{\mathbf{U},\mathbf{U}',\mathbf{V},\mathbf{V}'} k_{\mathcal{X}}(\mathbf{U}, \mathbf{U}') k_{\mathcal{Y}}(\mathbf{V}, \mathbf{V}') \\
&+ \mathbb{E}_{\mathbf{U},\mathbf{U}'} k_{\mathcal{X}}(\mathbf{U}, \mathbf{U}') \mathbb{E}_{\mathbf{V},\mathbf{V}'} k_{\mathcal{Y}}(\mathbf{V}, \mathbf{V}') \\
&- 2\mathbb{E}_{\mathbf{U},\mathbf{V}} \left[\mathbb{E}_{\mathbf{U}'} k_{\mathcal{X}}(\mathbf{U}, \mathbf{U}') \mathbb{E}_{\mathbf{V}'} k_{\mathcal{Y}}(\mathbf{V}, \mathbf{V}') \right]
\end{aligned}$$

Gretton et al. 2005a,b

Many applications: goodness-of-fit, independence tests, feature selection, …

Design of experiments for computer simulations

Now that we have access to the MMD

1. With the uniform target, specific choices of kernels give a MMD strictly equivalent to some well-known discrepancies (Hickernell 98) → **this is really a generalization**

2. We can propose an optimization algorithm which finds an empirical distribution (the DOE) which has the smallest distance with a target distribution (the uniform one in classical computer experiments)

Design of experiments for computer simulations

Now that we have access to the MMD

1. With the uniform target, specific choices of kernels give a MMD strictly equivalent to some well-known discrepancies (Hickernell 98) → **this is really a generalization**

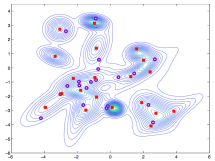
2. We can propose an optimization algorithm which finds an empirical distribution (the DOE) which has the smallest distance with a target distribution (the uniform one in classical computer experiments)

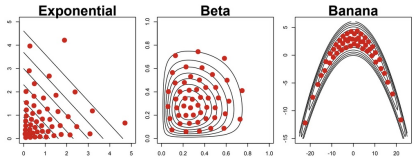
3. Several variants for optimization in the general case

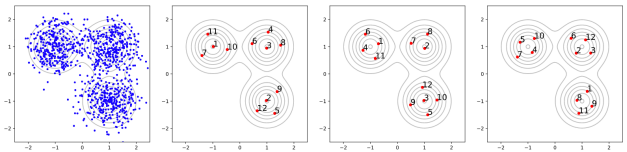
   - Greedy optimization (Pronzato 2022), with particular case *kernel herding* (Chen et al. 2010)

   - Convex-concave trick for specific kernel (*energy distance*), *support points* (Mak & Joseph 2018)

   - Integer Quadratic Programing for discrete target (Teymur et al. 2021)



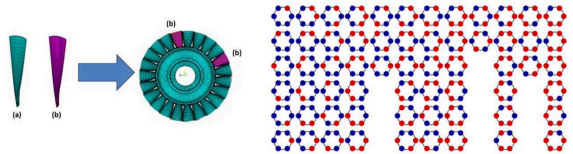




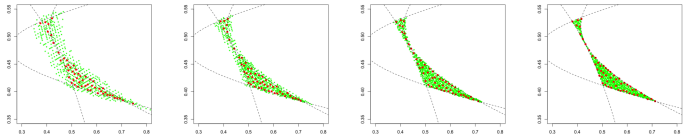
Design of experiments for computer simulations

Current hot topics using the MMD in computer experiments

> Handling categorical inputs and physical invariances in the kernel (Tran et al. 2021)



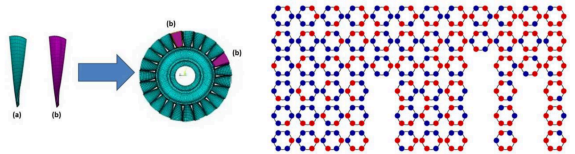
> Handling physical constraints in the input domain (Huang et al. 2021)



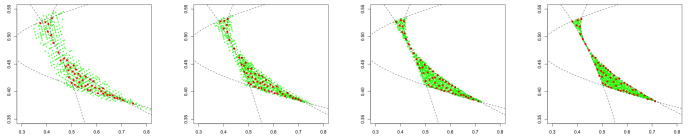
Design of experiments for computer simulations

Current hot topics using the MMD in computer experiments

> Handling categorical inputs and physical invariances in the kernel (Tran et al. 2021)

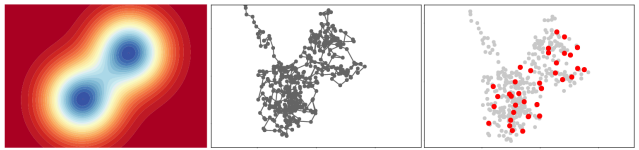


> Handling physical constraints in the input domain (Huang et al. 2021)

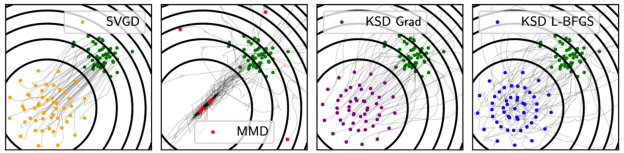


> If the target is known just up to a multiplicative constant: use of the Stein kernel, leading to the **Kernel Stein Discrepancy** (KSD) instead of the MMD

- Greedy optimization → Stein thinning (Riabiz et al. 2022)



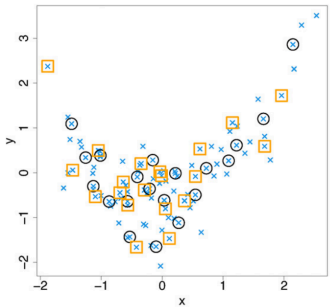
- Gradient descent (Korba et al. 2021)



Design of experiments for computer simulations

Current hot topics using the MMD in computer experiments

- Handling categorical inputs and physical invariances in the kernel (Tran et al. 2021)
- Handling physical constraints in the input domain (Huang et al. 2021)
- If the target is known just up to a multiplicative constant: use of the Stein kernel, leading to the **Kernel Stein Discrepancy** (KSD) instead of the MMD
- Use of the MMD to select train / validation / test set from given database
  - Data splitting (Joseph & Vakayil 2022)
  - Upcoming J. Muré's talk in this session!



3

SENSITIVITY ANALYSIS

SAFRAN

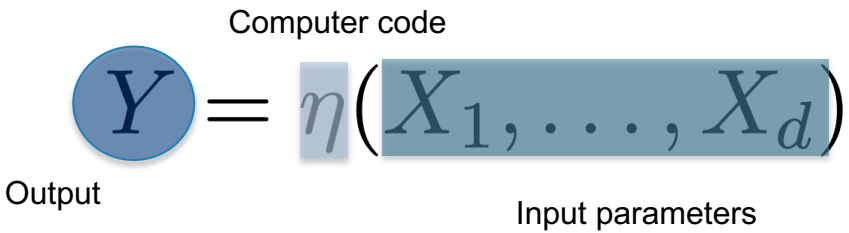
Sensitivity analysis

Goal : identify and rank the input parameters according to their impact on the output of a computer code

Why ?

- Reduce the output uncertainty efficiently by reducing the uncertainty of the main contributors
- Improve the knowledge of the physical phenomenon,
- Simplify the model

Notations



Sensitivity analysis: Sobol' indices arise from a functional ANOVA decomposition

Theorem 1 (ANOVA decomposition (Hoeffding, 1948; Antoniadis, 1984)). *Assume that* $\eta : \mathcal{X}_1 \times \ldots \times \mathcal{X}_d \to \mathcal{Y}$ *is a square integrable function of* d *independent random variables* $X_1, \ldots, X_d$. *Then* η *admits a decomposition*

$$Y = \eta(X_1, \ldots, X_d) = \sum_{A \subseteq \mathcal{P}_d} \eta_A(\mathbf{X}_A),$$

with η_A *depending only on the variables* $\mathbf{X}_A$ *and satisfying*

(a) $\eta_\emptyset = \mathbb{E}(Y)$,

(b) $\mathbb{E}_{X_l}(\eta_A(\mathbf{X}_A)) = 0$ *if* $l \in A$,

(c) $\eta_A(\mathbf{X}_A) = \sum_{B \subset A} (-1)^{|A|-|B|} \mathbb{E}(Y|\mathbf{X}_B)$.

Furthermore, (b) implies that all the terms η_A *in the decomposition are mutually orthogonal. As a consequence, the output variance can be decomposed as*

$$\operatorname{Var} Y = \sum_{A \subseteq \mathcal{P}_d} \operatorname{Var} \eta_A(\mathbf{X}_A) = \sum_{A \subseteq \mathcal{P}_d} V_A \tag{1}$$

where

$$V_A = \sum_{B \subset A} (-1)^{|A|-|B|} \operatorname{Var} \mathbb{E}(Y|\mathbf{X}_B). \tag{2}$$

Sensitivity analysis: Sobol' indices arise from a functional ANOVA decomposition

Definition 1 (Sobol' indices (Sobol', 1993)). *Under the same assumptions of Theorem 1, the Sobol' sensitivity index associated to a subset A of input variables is defined as*

$$S_A = \frac{V_A}{\operatorname{Var} Y}, \tag{3}$$

A is a subset of input variables

while the total Sobol' index associated to A is

$$S_A^T = \sum_{B \subseteq \mathcal{P}_d,\, B \cap A \neq \emptyset} S_B. \tag{4}$$

In particular, the first-order Sobol' index of an input X_l writes

$$S_l = \frac{\operatorname{Var} \mathbb{E}(Y|X_l)}{\operatorname{Var} Y}$$

Impact of an input alone

and its total Sobol' index is given by

$$S_l^T = \sum_{B \subseteq \mathcal{P}_d,\, l \in B} S_B = 1 - \frac{\operatorname{Var} \mathbb{E}(Y|\mathbf{X}_{-l})}{\operatorname{Var} Y}.$$

Impact of an input through all its potential interactions with others

Finally, the ANOVA decomposition (1) readily provides an interpretation of Sobol' indices as a percentage of explained output variance, i.e.

$$\sum_{A \subseteq \mathcal{P}_d} S_A = 1. \tag{5}$$

Interpretation as percentage

Sensitivity analysis: Sobol' indices

Sobol' indices

- The impact of each input can be quantitatively assessed
  - First-order effect
  - Total effect including also all possible interactions with other inputs
  - **Pure interactions can be properly defined**

$$S_{ll'} = \frac{\mathrm{Var}\,\mathbb{E}(Y|X_l,\,X_{l'}) - \mathrm{Var}\,\mathbb{E}(Y|X_l) - \mathrm{Var}\,\mathbb{E}(Y|X_{l'})}{\mathrm{Var}\,Y} = \frac{\mathrm{Var}\,\mathbb{E}(Y|X_l,\,X_{l'})}{\mathrm{Var}\,Y} - S_l - S_{l'}$$

First-order effects can be propertly subtracted

Sensitivity analysis: Sobol' indices

Sobol' indices

- The impact of each input can be quantitatively assessed
  - First-order effect
  - Total effect including also all possible interactions with other inputs
  - **Pure interactions can be properly defined**

$$S_{ll'} = \frac{\mathrm{Var}\,\mathbb{E}(Y|X_l,\,X_{l'}) - \mathrm{Var}\,\mathbb{E}(Y|X_l) - \mathrm{Var}\,\mathbb{E}(Y|X_{l'})}{\mathrm{Var}\,Y} = \frac{\mathrm{Var}\,\mathbb{E}(Y|X_l,\,X_{l'})}{\mathrm{Var}\,Y} - S_l - S_{l'}$$

First-order effects can be propertly subtracted

Limitations

- **Assumption of independent inputs (more on this later)**
- **Impact on output variance only**
- **Outputs may not be scalars**

Sensitivity analysis: other indices

Going beyond the variance 1: goal-oriented sensitivity analysis

- Indices based on contrast functions (Fort et al. 2014), in particular quantile-oriented indices
- Reliability-based indices
- Many industrial applications

Going beyond the variance 2: moment-independent indices

- Principle: Quantify the impact of an input parameter on the **probability distribution of the output**

$$\mathcal{S}_l^{TV} = \int |p_Y(y) - p_{Y|X_l=x}(y)| p_{X_l}(x) dx dy$$

Borgonovo 2007

$$\mathcal{S}_l^{KL} = \int p_{Y|X_l=x}(y) \ln \left(\frac{p_{Y|X_l=x}(y)}{p_Y(y)} \right) p_{X_l}(x) dx dy$$

Kraskov et al. 2001

Sensitivity analysis: general point of view

General framework for moment-independent indices

$$\mathcal{S}_l = \mathbb{E}_{X_l}\left(d(\mathrm{P}_Y, \mathrm{P}_{Y|X_l})\right)$$

Baucells & Borgonovo 2013
D. 2015

- If the output probability distribution and the conditional one are « close », the input parameter has little influence
- Example: f-divergence (D. 2015, Rahman 2016), with particular cases TV & KL

Sensitivity analysis – Moment-independent indices

Pros

- They account for the whole effect of a parameter on the output distribution
- They are density-based
  - Many methods and packages for estimation
  - Several distances can be investigated without additional cost

Sensitivity analysis – Moment-independent indices

Pros

- They account for the whole effect of a parameter on the output distribution
- They are density-based
  - Many methods and packages for estimation
  - Several distances can be investigated without additional cost

Cons

- Definition of higher-order indices means curse of dimensionality for density estimation
- No ANOVA-like decomposition
  - No access to a « natural » normalisation constant
  - No proper separation of interactions and main effects

Does this make sense ?

$$\mathcal{S}_{ll'}^{TV} = \int |p_Y(y)p_{X_l}(x)p_{X_{l'}}(x') - p_{X_l,X_{l'},Y}(x,x',y)|dxdx'dy - \mathcal{S}_l^{TV} - \mathcal{S}_{l'}^{TV}$$

Sensitivity analysis – Moment-independent indices

Pros

- They account for the whole effect of a parameter on the output distribution
- They are density-based
  - Many methods and packages for estimation
  - Several distances can be investigated without additional cost

Cons

- Definition of higher-order indices means curse of dimensionality for density estimation
- No ANOVA-like decomposition
  - No access to a « natural » normalisation constant
  - No proper separation of interactions and main effects

A promising candidate: kernel-embedding of probability distributions

$$\mathcal{S}_l = \mathbb{E}_{X_l}\left(d(\mathrm{P}_Y, \mathrm{P}_{Y|X_l})\right)$$

Kernel-embedding of probability distributions

The kernel mean embedding of a probability measure is defined as

$$\mu_{\mathrm{P}} = \mathbb{E}_{\xi \sim \mathrm{P}} k_{\mathcal{X}}(\xi, \cdot) = \int_{\mathcal{X}} k_{\mathcal{X}}(\xi, \cdot) d\mathrm{P}(\xi)$$

A distance between probability measures is then given by the Maximum Mean Discrepancy

$$\mathrm{MMD}(\mathrm{P}_1, \mathrm{P}_2) = \|\mu_{\mathrm{P}_1} - \mu_{\mathrm{P}_2}\|_{\mathcal{H}}$$

The reproducing property in the RKHS gives the central result

$$\mathrm{MMD}^2(\mathrm{P}_1, \mathrm{P}_2) = \mathbb{E}_{\xi,\xi'} k_{\mathcal{X}}(\xi, \xi') - 2\mathbb{E}_{\xi,\zeta} k_{\mathcal{X}}(\xi, \zeta) + \mathbb{E}_{\zeta,\zeta'} k_{\mathcal{X}}(\zeta, \zeta')$$

Smola et al. 2007, Song 2008, Song et al. 2009

SAFRAN

Kernel-embedding of probability distributions

Other major use: testing independence of random vectors

$$\mathrm{MMD}^2(\mathrm{P}_{\mathbf{UV}}, \mathrm{P}_{\mathbf{U}} \otimes \mathrm{P}_{\mathbf{V}}) = \|\mu_{\mathrm{P}_{\mathbf{UV}}} - \mu_{\mathrm{P}_{\mathbf{U}}} \otimes \mu_{\mathrm{P}_{\mathbf{V}}}\|^2_{\mathcal{H}}$$

$$\begin{aligned}
\mathrm{HSIC}(\mathbf{U}, \mathbf{V}) &= \mathrm{MMD}^2(\mathrm{P}_{\mathbf{UV}}, \mathrm{P}_{\mathbf{U}} \otimes \mathrm{P}_{\mathbf{V}}) \\
&= \mathbb{E}_{\mathbf{U},\mathbf{U}',\mathbf{V},\mathbf{V}'} k_{\mathcal{X}}(\mathbf{U}, \mathbf{U}') k_{\mathcal{Y}}(\mathbf{V}, \mathbf{V}') \\
&+ \mathbb{E}_{\mathbf{U},\mathbf{U}'} k_{\mathcal{X}}(\mathbf{U}, \mathbf{U}') \mathbb{E}_{\mathbf{V},\mathbf{V}'} k_{\mathcal{Y}}(\mathbf{V}, \mathbf{V}') \\
&- 2\mathbb{E}_{\mathbf{U},\mathbf{V}} \left[\mathbb{E}_{\mathbf{U}'} k_{\mathcal{X}}(\mathbf{U}, \mathbf{U}') \mathbb{E}_{\mathbf{V}'} k_{\mathcal{Y}}(\mathbf{V}, \mathbf{V}') \right]
\end{aligned}$$

Gretton et al. 2005a,b

Many applications: goodness-of-fit, independence tests, feature selection, …

Kernel-embedding of probability distributions

Pros

- Thanks to the RKHS, only involves expectations of kernels
- Less prone to the curse of dimensionality
- **Can easily handle structured objects (curves, images, graphs, probability measures, …) by using specific kernels tailored at such tasks**

Cons

- Choice of kernel / kernel hyperparameters …

Kernel-embedding of probability distributions for GSA: MMD

Remember our general GSA setting ?

$$\mathcal{S}_l = \mathbb{E}_{X_l}\left(d(\mathrm{P}_Y, \mathrm{P}_{Y|X_l})\right)$$

Kernel-embedding of probability distributions for GSA: MMD

Remember our general GSA setting ?

$$\mathcal{S}_l = \mathbb{E}_{X_l}\left(d(\mathrm{P}_Y, \mathrm{P}_{Y|X_l})\right)$$

Straightforward use of kernel-embeddings

First-order

$$\begin{aligned}
\mathcal{S}_l^{\mathrm{MMD}} &= \mathbb{E}_{X_l}\mathrm{MMD}^2(\mathrm{P}_Y, \mathrm{P}_{Y|X_l}) \\
&= \mathbb{E}_{X_l}\mathbb{E}_{\xi,\xi'\sim\mathrm{P}_Y} k_{\mathcal{Y}}(\xi,\xi') - 2\mathbb{E}_{X_l}\mathbb{E}_{\xi\sim\mathrm{P}_Y,\zeta\sim\mathrm{P}_{Y|X_l}} k_{\mathcal{Y}}(\xi,\zeta) + \mathbb{E}_{X_l}\mathbb{E}_{\zeta,\zeta'\sim\mathrm{P}_{Y|X_l}} k_{\mathcal{Y}}(\zeta,\zeta') \\
&= \mathbb{E}_{X_l}\mathbb{E}_{\zeta,\zeta'\sim\mathrm{P}_{Y|X_l}} k_{\mathcal{Y}}(\zeta,\zeta') - \mathbb{E}_{\xi,\xi'\sim\mathrm{P}_Y} k_{\mathcal{Y}}(\xi,\xi')
\end{aligned}$$

D. 2016 & 2021, Barr & Rabitz 2022

Kernel-embedding of probability distributions for GSA: MMD

Remember our general GSA setting ?

$$\mathcal{S}_l = \mathbb{E}_{X_l}\left(d(\mathrm{P}_Y, \mathrm{P}_{Y|X_l})\right)$$

Straightforward use of kernel-embeddings

First-order

$$\begin{aligned}
\mathcal{S}_l^{\mathrm{MMD}} &= \mathbb{E}_{X_l}\mathrm{MMD}^2(\mathrm{P}_Y, \mathrm{P}_{Y|X_l}) \\
&= \mathbb{E}_{X_l}\mathbb{E}_{\xi,\xi'\sim\mathrm{P}_Y} k_{\mathcal{Y}}(\xi,\xi') - 2\mathbb{E}_{X_l}\mathbb{E}_{\xi\sim\mathrm{P}_Y,\zeta\sim\mathrm{P}_{Y|X_l}} k_{\mathcal{Y}}(\xi,\zeta) + \mathbb{E}_{X_l}\mathbb{E}_{\zeta,\zeta'\sim\mathrm{P}_{Y|X_l}} k_{\mathcal{Y}}(\zeta,\zeta') \\
&= \mathbb{E}_{X_l}\mathbb{E}_{\zeta,\zeta'\sim\mathrm{P}_{Y|X_l}} k_{\mathcal{Y}}(\zeta,\zeta') - \mathbb{E}_{\xi,\xi'\sim\mathrm{P}_Y} k_{\mathcal{Y}}(\xi,\xi')
\end{aligned}$$

Group

$$\mathcal{S}_A^{\mathrm{MMD}} = \mathbb{E}_{\mathbf{X}_A}\left(\mathrm{MMD}^2(\mathrm{P}_Y, \mathrm{P}_{Y|\mathbf{X}_A})\right) = \mathbb{E}_{\mathbf{X}_A}\mathbb{E}_{\zeta,\zeta'\sim\mathrm{P}_{Y|\mathbf{X}_A}} k_{\mathcal{Y}}(\zeta,\zeta') - \mathbb{E}_{\xi,\xi'\sim\mathrm{P}_Y} k_{\mathcal{Y}}(\xi,\xi')$$

D. 2016 & 2021, Barr & Rabitz 2022

SAFRAN

Kernel-embedding of probability distributions for GSA: MMD

Links with Sobol': if we use the vanilla dot product kernel $k_{\mathcal{Y}}(y, y') = yy'$

$$
\begin{aligned}
\mathcal{S}_A^{\mathrm{MMD}} &= \mathbb{E}_{\mathbf{X}_A}\left(\mathbb{E}_{\xi\sim \mathrm{P}_Y}(\xi) - \mathbb{E}_{\zeta\sim \mathrm{P}_{Y|\mathbf{X}_A}}(\zeta)\right)^2 \\
&= \mathbb{E}_{\mathbf{X}_A}\left(\mathbb{E}Y - \mathbb{E}(Y|\mathbf{X}_A)\right)^2 \\
&= \mathrm{Var}\,\mathbb{E}(Y|\mathbf{X}_A) \qquad \text{Unnormalized Sobol'}
\end{aligned}
$$

Kernel-embedding of probability distributions for GSA: MMD

Links with Sobol': if we use the vanilla dot product kernel $k_{\mathcal{Y}}(y, y') = yy'$

$$\begin{aligned}
\mathcal{S}_A^{\mathrm{MMD}} &= \mathbb{E}_{\mathbf{X}_A}\left(\mathbb{E}_{\xi\sim \mathrm{P}_Y}(\xi) - \mathbb{E}_{\zeta\sim \mathrm{P}_{Y|\mathbf{X}_A}}(\zeta)\right)^2 \\
&= \mathbb{E}_{\mathbf{X}_A}\left(\mathbb{E}Y - \mathbb{E}(Y|\mathbf{X}_A)\right)^2 \\
&= \mathrm{Var}\,\mathbb{E}(Y|\mathbf{X}_A) \qquad \text{Unnormalized Sobol'}
\end{aligned}$$

Links with Sobol': if Mercer's theorem holds

$$k_{\mathcal{Y}}(y, y') = \sum_{r=1}^{\infty} \phi_r(y)\phi_r(y') \Rightarrow$$

$$\begin{aligned}
\mathcal{S}_A^{\mathrm{MMD}} &= \sum_{r=1}^{\infty}\left\{\mathbb{E}_{\mathbf{X}_A}\mathbb{E}_{\xi,\xi'\sim \mathrm{P}_{Y|\mathbf{X}_A}}\left(\phi_r(\xi)\phi_r(\xi')\right) - \mathbb{E}_{\zeta,\zeta'\sim \mathrm{P}}\left(\phi_r(\zeta)\phi_r(\zeta')\right)\right\} \\
&= \sum_{r=1}^{\infty}\left\{\mathbb{E}_{\mathbf{X}_A}\mathbb{E}\left(\phi_r(Y)|\mathbf{X}_A\right)^2 - \mathbb{E}\left(\phi_r(Y)\right)^2\right\} \\
&= \sum_{r=1}^{\infty}\mathrm{Var}\,\mathbb{E}\left(\phi_r(Y)|\mathbf{X}_A\right).
\end{aligned}$$

> Aggregation of Sobol' indices on a (possibly) infinite number of nonlinear transformations of the output

Kernel-embedding of probability distributions for GSA: MMD

More importantly, we have an ANOVA-like decomposition !

Theorem 3 (ANOVA decomposition for MMD). *Under the same assumptions of Theorem 1 (in particular, the random vector* $\mathbf{X}$ *has independent components) and with Assumption 1, denote* $\mathrm{MMD}^2_{\mathrm{tot}} = \mathbb{E}k_{\mathcal{Y}}(Y, Y) - \mathbb{E}k_{\mathcal{Y}}(Y, Y')$ *where* Y' *is an independent copy of* Y*. Then the total MMD can be decomposed as*

$$\mathrm{MMD}^2_{\mathrm{tot}} = \sum_{A \subseteq \mathcal{P}_d} \mathrm{MMD}^2_A$$

Kernel-embedding of probability distributions for GSA: MMD

More importantly, we have an ANOVA-like decomposition !

Theorem 3 (ANOVA decomposition for MMD). *Under the same assumptions of Theorem 1 (in particular, the random vector* $\mathbf{X}$ *has independent components) and with Assumption 1, denote* $\mathrm{MMD}^2_{\mathrm{tot}} = \mathbb{E}k_{\mathcal{Y}}(Y,Y) - \mathbb{E}k_{\mathcal{Y}}(Y,Y')$ *where* Y' *is an independent copy of* Y*. Then the total MMD can be decomposed as*

$$\mathrm{MMD}^2_{\mathrm{tot}} = \sum_{A \subseteq \mathcal{P}_d} \mathrm{MMD}^2_A$$

where each term is given by

$$\mathrm{MMD}^2_A = \sum_{B \subset A} (-1)^{|A|-|B|} \mathbb{E}_{\mathbf{X}_B} \left(\mathrm{MMD}^2(\mathrm{P}_Y, \mathrm{P}_{Y|\mathbf{X}_B}) \right).$$

- **So we can define properly normalized MMD-based sensitivity indices**
- Proof is straightforward with Mercer's theorem

Kernel-embedding of probability distributions for GSA: MMD

More importantly, we have an ANOVA-like decomposition !

Definition 2 (MMD-based sensitivity indices). *In the frame of Theorem 3, let $A \subseteq \mathcal{P}_d$. The normalized MMD-based sensitivity index associated to a subset A of input variables is defined as*

$$S_A^{\mathrm{MMD}} = \frac{\mathrm{MMD}_A^2}{\mathrm{MMD}_{\mathrm{tot}}^2},$$

Impact of a subset alone

Kernel-embedding of probability distributions for GSA: MMD

More importantly, we have an ANOVA-like decomposition !

Definition 2 (MMD-based sensitivity indices). *In the frame of Theorem 3, let $A \subseteq \mathcal{P}_d$. The normalized MMD-based sensitivity index associated to a subset A of input variables is defined as*

$$S_A^{\mathrm{MMD}} = \frac{\mathrm{MMD}_A^2}{\mathrm{MMD}_{\mathrm{tot}}^2},$$

Impact of a subset alone

while the total MMD-based index associated to A is

$$S_A^{T,\mathrm{MMD}} = \sum_{B \subseteq \mathcal{P}_d,\, B \cap A \neq \emptyset} S_B^{\mathrm{MMD}} = 1 - \frac{\mathbb{E}_{\mathbf{x}_{-A}}\left(\mathrm{MMD}^2(\mathrm{P}_Y, \mathrm{P}_{Y|\mathbf{x}_{-A}})\right)}{\mathrm{MMD}_{\mathrm{tot}}^2}.$$

Impact of a subset through all its potential interactions with others

Kernel-embedding of probability distributions for GSA: MMD

More importantly, we have an ANOVA-like decomposition !

Definition 2 (MMD-based sensitivity indices). *In the frame of Theorem 3, let $A \subseteq \mathcal{P}_d$. The normalized MMD-based sensitivity index associated to a subset A of input variables is defined as*

$$S_A^{\mathrm{MMD}} = \frac{\mathrm{MMD}_A^2}{\mathrm{MMD}_{\mathrm{tot}}^2},$$

Impact of a subset alone

while the total MMD-based index associated to A is

$$S_A^{T,\mathrm{MMD}} = \sum_{B \subseteq \mathcal{P}_d,\, B \cap A \neq \emptyset} S_B^{\mathrm{MMD}} = 1 - \frac{\mathbb{E}_{\mathbf{x}_{-A}}\left(\mathrm{MMD}^2(\mathrm{P}_Y, \mathrm{P}_{Y|\mathbf{x}_{-A}})\right)}{\mathrm{MMD}_{\mathrm{tot}}^2}.$$

Impact of a subset through all its potential interactions with others

From Theorem 3, we have the fundamental identity providing the interpretation of MMD-based indices as percentage of the explained generalized variance $\mathrm{MMD}_{\mathrm{tot}}^2$:

$$\sum_{A \subseteq \mathcal{P}_d} S_A^{\mathrm{MMD}} = 1.$$

Interpretation as percentage

Kernel-embedding of probability distributions for GSA: MMD

New MMD-based sensitivity index

- **First moment-independent index with a decomposition**
- Can handle easily structured outputs
- Close generalization of Sobol' index, which is obtained as a particular case

Kernel-embedding of probability distributions for GSA: MMD

New MMD-based sensitivity index

- **First moment-independent index with a decomposition**
- Can handle easily structured outputs
- Close generalization of Sobol' index, which is obtained as a particular case

Estimation

- We can easily recycle estimators proposed for Sobol' indices
- Monte-Carlo, Pick-freeze, Rank, k-NN
- See D. 2021 for details

Kernel-embedding of probability distributions for GSA: MMD

New MMD-based sensitivity index

- **First moment-independent index with a decomposition**
- Can handle easily structured outputs
- Close generalization of Sobol' index, which is obtained as a particular case

Estimation

- We can easily recycle estimators proposed for Sobol' indices
- Monte-Carlo, Pick-freeze, Rank, k-NN
- See D. 2021 for details

Going further by taking a step back

Kernel-embedding of probability distributions for GSA

Remember our general GSA setting ?

$$\mathcal{S}_l = \mathbb{E}_{X_l}\left(d(\mathrm{P}_Y, \mathrm{P}_{Y|X_l})\right)$$

Kernel-embedding of probability distributions for GSA

Remember our general GSA setting ?

$$\mathcal{S}_l = \mathbb{E}_{X_l}\left(d(\mathrm{P}_Y, \mathrm{P}_{Y|X_l})\right)$$

Other point of view

$$\begin{aligned}
\mathcal{S}_l^{KL} &= \int p_{Y|X_l=x}(y) \ln\left(\frac{p_{Y|X_l=x}(y)}{p_Y(y)}\right) p_{X_l}(x)dxdy \\
&= \int \ln\left(\frac{p_{Y,X_l}(y,x)}{p_Y(y)p_{X_l}(x)}\right) p_{Y,X_l}(y,x)dxdy \\
&= \mathrm{MI}(X_l, Y)
\end{aligned}$$

> **The KL-based index actually corresponds to the mutual information between one of the inputs and the output, i.e. a measure of their dependence**

Kernel-embedding of probability distributions for GSA

Remember our general GSA setting ?

$$\mathcal{S}_l = \mathbb{E}_{X_l}\left(d(\mathrm{P}_Y, \mathrm{P}_{Y|X_l})\right)$$

Other point of view

$$\mathcal{S}_l^{KL} = \int p_{Y|X_l=x}(y) \ln \left(p_{Y|} \right) x dy$$

$$= \int \ln \left(\right) (y,x) dx dy$$

Why not use HSIC instead?

> The KL-bas esponds to the mutual information between one of the inputs and the output, i.e. a measure of th ice

Kernel-embedding of probability distributions for GSA: HSIC

HSIC-based sensitivity index

$$\mathcal{S}_A^{HS} = \mathrm{HSIC}(\mathbf{X}_A, Y)$$

- Already proposed with a hand-made normalization in D. 2015
- Works very well for screening, with small sample size

Kernel-embedding of probability distributions for GSA: HSIC

HSIC-based sensitivity index

$$\mathcal{S}_A^{HS} = \mathrm{HSIC}(\mathbf{X}_A, Y)$$

- Already proposed with a hand-made normalization in D. 2015
- Works very well for screening, with small sample size

But it actually exhibits an ANOVA decomposition too

Assumption 3. *The reproducing kernel $k_{\mathcal{X}}$ of $\mathcal{F}$ is of the form*

$$k_{\mathcal{X}}(\mathbf{x}, \mathbf{x}') = \prod_{l=1}^{p} \left(1 + k_l(x_l, x_l')\right) \tag{10}$$

where for each $l = 1, \ldots, d$, $k_l(\cdot, \cdot)$ is the reproducing kernel of a RKHS $\mathcal{F}_l$ of real functions depending only on variable x_l and such that $1 \notin \mathcal{F}_l$.
In addition, for all $l = 1, \ldots, d$ and $\forall x_l \in \mathcal{X}_l$, we have

$$\int_{\mathcal{X}_l} k_l(x_l, x_l') d\mathrm{P}_{X_l}(x_l') = 0. \tag{11}$$

Kernel-embedding of probability distributions for GSA: HSIC

HSIC-based sensitivity index

$$\mathcal{S}_A^{HS} = \mathrm{HSIC}(\mathbf{X}_A, Y)$$

- Already proposed with a hand-made normalization in D. 2015
- Works very well for screening, with small sample size

But it actually exhibits an ANOVA decomposition too

Assumption 3. *The reproducing kernel $k_{\mathcal{X}}$ of $\mathcal{F}$ is of the form*

Product kernel

$$k_{\mathcal{X}}(\mathbf{x}, \mathbf{x}') = \prod_{l=1}^{p} \left(1 + k_l(x_l, x_l')\right) \tag{10}$$

where for each $l = 1, \ldots, d$, $k_l(\cdot, \cdot)$ is the reproducing kernel of a RKHS $\mathcal{F}_l$ of real functions depending only on variable x_l and such that $1 \notin \mathcal{F}_l$.
In addition, for all $l = 1, \ldots, d$ and $\forall x_l \in \mathcal{X}_l$, we have

$$\int_{\mathcal{X}_l} k_l(x_l, x_l') d\mathrm{P}_{X_l}(x_l') = 0. \tag{11}$$

Kernel-embedding of probability distributions for GSA: HSIC

HSIC-based sensitivity index

$$\mathcal{S}_A^{HS} = \mathrm{HSIC}(\mathbf{X}_A, Y)$$

- Already proposed with a hand-made normalization in D. 2015
- Works very well for screening, with small sample size

But it actually exhibits an ANOVA decomposition too

Assumption 3. *The reproducing kernel $k_{\mathcal{X}}$ of $\mathcal{F}$ is of the form*

Product kernel

$$k_{\mathcal{X}}(\mathbf{x}, \mathbf{x}') = \prod_{l=1}^{p} \left(1 + k_l(x_l, x_l')\right) \tag{10}$$

where for each $l = 1, \ldots, d$, $k_l(\cdot, \cdot)$ is the reproducing kernel of a RKHS $\mathcal{F}_l$ of real functions depending only on variable x_l and such that $1 \notin \mathcal{F}_l$. **Without constant functions**
In addition, for all $l = 1, \ldots, d$ and $\forall x_l \in \mathcal{X}_l$, we have

$$\int_{\mathcal{X}_l} k_l(x_l, x_l') d\mathrm{P}_{X_l}(x_l') = 0. \tag{11}$$

Kernel-embedding of probability distributions for GSA: HSIC

HSIC-based sensitivity index

$$\mathcal{S}_A^{HS} = \mathrm{HSIC}(\mathbf{X}_A, Y)$$

- Already proposed with a hand-made normalization in D. 2015
- Works very well for screening, with small sample size

But it actually exhibits an ANOVA decomposition too

Assumption 3. *The reproducing kernel $k_{\mathcal{X}}$ of $\mathcal{F}$ is of the form*

Product kernel

$$k_{\mathcal{X}}(\mathbf{x}, \mathbf{x}') = \prod_{l=1}^{p} \left(1 + k_l(x_l, x_l')\right) \tag{10}$$

where for each $l = 1, \ldots, d$, $k_l(\cdot, \cdot)$ is the reproducing kernel of a RKHS $\mathcal{F}_l$ of real functions depending only on variable x_l and such that $1 \notin \mathcal{F}_l$. **Without constant functions**

In addition, for all $l = 1, \ldots, d$ and $\forall x_l \in \mathcal{X}_l$, we have

Zero-mean kernel

$$\int_{\mathcal{X}_l} k_l(x_l, x_l') d\mathrm{P}_{X_l}(x_l') = 0. \tag{11}$$

Kernel-embedding of probability distributions for GSA: HSIC

HSIC-based sensitivity index

$$\mathcal{S}_A^{HS} = \mathrm{HSIC}(\mathbf{X}_A, Y)$$

- Already proposed with a hand-made normalization in D. 2015
- Works very well for screening, with small sample size

But it actually exhibits an ANOVA decomposition too

Assumption 3. *The reproducing kernel $k_{\mathcal{X}}$ of $\mathcal{F}$ is of the form*

$$k_{\mathcal{X}}(\mathbf{x}, \mathbf{x}') = \prod_{l=1}^{p} \left(1 + k_l(x_l, x_l')\right) \tag{10}$$

where for each $l = 1, \ldots, d$, $k_l(\cdot, \cdot)$ is the reproducing kernel of a RKHS $\mathcal{F}_l$ of real functions depending only on variable x_l and such that $1 \notin \mathcal{F}_l$. In addition, for all $l = 1, \ldots, d$ and $\forall x_l \in \mathcal{X}_l$, we have

$$\int_{\mathcal{X}_l} k_l(x_l, x_l') d\mathrm{P}_{X_l}(x_l') = 0. \tag{11}$$

Needed to get orthogonality inside the RKHS

Product kernel

Without constant functions

Zero-mean kernel

Kernel-embedding of probability distributions for GSA: HSIC

ANOVA-like decomposition for HSIC

Theorem 4 (ANOVA decomposition for HSIC)**.** *Under the same assumptions of Theorem 1 (in particular, the random vector* $\mathbf{X}$ *has independent components) and with Assumptions 2 and 3, the HSIC dependence measure between* $\mathbf{X} = (X_1, \dots, X_d)$ *and* Y *can be decomposed as*

$$\mathrm{HSIC}\left(\mathbf{X}, Y\right) = \sum_{A \subseteq \mathcal{P}_d} \mathrm{HSIC}_A$$

Kernel-embedding of probability distributions for GSA: HSIC

ANOVA-like decomposition for HSIC

Theorem 4 (ANOVA decomposition for HSIC)**.** *Under the same assumptions of Theorem 1 (in particular, the random vector* $\mathbf{X}$ *has independent components) and with Assumptions 2 and 3, the HSIC dependence measure between* $\mathbf{X} = (X_1, \dots, X_d)$ *and* Y *can be decomposed as*

$$\mathrm{HSIC}\left(\mathbf{X}, Y\right) = \sum_{A \subseteq \mathcal{P}_d} \mathrm{HSIC}_A$$

where each term is given by

$$\mathrm{HSIC}_A = \sum_{B \subset A} (-1)^{|A|-|B|} \mathrm{HSIC}\left(\mathbf{X}_B, Y\right)$$

and $\mathrm{HSIC}\left(\mathbf{X}_B, Y\right)$ *is defined with a product RKHS* $\mathcal{H}_B = \mathcal{F}_B \times \mathcal{G}$ *with kernel* $k_B(\mathbf{x}_B, \mathbf{x}'_B) k_{\mathcal{Y}}(y, y') = \prod_{l \in B} \left(1 + k_l(x_l, x'_l)\right) k_{\mathcal{Y}}(y, y')$ *as in (10).*

> So we can define properly normalized HSIC-based sensitivity indices

Kernel-embedding of probability distributions for GSA: HSIC

ANOVA-like decomposition for HSIC

Theorem 4 (ANOVA decomposition for HSIC)**.** *Under the same assumptions of Theorem 1 (in particular, the random vector* $\mathbf{X}$ *has independent components) and with Assumptions 2 and 3, the HSIC dependence measure between* $\mathbf{X} = (X_1, \dots, X_d)$ *and* Y *can be decomposed as*

$$\mathrm{HSIC}\left(\mathbf{X}, Y\right) = \sum_{A \subseteq \mathcal{P}_d} \mathrm{HSIC}_A$$

where each term is given by

$$\mathrm{HSIC}_A = \sum_{B \subset A} (-1)^{|A|-|B|} \mathrm{HSIC}\left(\mathbf{X}_B, Y\right)$$

and $\mathrm{HSIC}\left(\mathbf{X}_B, Y\right)$ *is defined with a product RKHS* $\mathcal{H}_B = \mathcal{F}_B \times \mathcal{G}$ *with kernel* $k_B(\mathbf{x}_B, \mathbf{x}'_B) k_{\mathcal{Y}}(y, y') = \prod_{l \in B} \left(1 + k_l(x_l, x'_l)\right) k_{\mathcal{Y}}(y, y')$ *as in (10).*

- **So we can define properly normalized HSIC-based sensitivity indices**
- Proof relies on orthogonal decompositions in RKHS (see Appendix)

Kernel-embedding of probability distributions for GSA: HSIC

ANOVA-like decomposition for HSIC

Definition 3 (HSIC-based sensitivity indices). *In the frame of Theorem 4, let* $A \subseteq \mathcal{P}_d$. *The normalized HSIC-based sensitivity index associated to a subset* A *of input variables is defined as*

$$S_A^{\mathrm{HSIC}} = \frac{\mathrm{HSIC}_A}{\mathrm{HSIC}\,(\mathbf{X}, Y)},$$

Impact of a subset alone

Kernel-embedding of probability distributions for GSA: HSIC

ANOVA-like decomposition for HSIC

Definition 3 (HSIC-based sensitivity indices). *In the frame of Theorem 4, let $A \subseteq \mathcal{P}_d$. The normalized HSIC-based sensitivity index associated to a subset A of input variables is defined as*

$$S_A^{\mathrm{HSIC}} = \frac{\mathrm{HSIC}_A}{\mathrm{HSIC}\,(\mathbf{X}, Y)},$$

Impact of a subset alone

while the total HSIC-based index associated to A is

$$S_A^{T,\mathrm{HSIC}} = \sum_{B \subseteq \mathcal{P}_d,\, B \cap A \neq \emptyset} S_B^{\mathrm{HSIC}} = 1 - \frac{\mathrm{HSIC}(\mathbf{X}_{-A}, Y)}{\mathrm{HSIC}\,(\mathbf{X}, Y)}.$$

Impact of a subset through all its potential interactions with others

Kernel-embedding of probability distributions for GSA: HSIC

ANOVA-like decomposition for HSIC

Definition 3 (HSIC-based sensitivity indices). *In the frame of Theorem 4, let $A \subseteq \mathcal{P}_d$. The normalized HSIC-based sensitivity index associated to a subset A of input variables is defined as*

$$S_A^{\mathrm{HSIC}} = \frac{\mathrm{HSIC}_A}{\mathrm{HSIC}\,(\mathbf{X}, Y)},$$

Impact of a subset alone

while the total HSIC-based index associated to A is

$$S_A^{T,\mathrm{HSIC}} = \sum_{B \subseteq \mathcal{P}_d,\, B \cap A \neq \emptyset} S_B^{\mathrm{HSIC}} = 1 - \frac{\mathrm{HSIC}(\mathbf{X}_{-A}, Y)}{\mathrm{HSIC}\,(\mathbf{X}, Y)}.$$

Impact of a subset through all its potential interactions with others

From Theorem 4, we have the fundamental identity providing the interpretation of HSIC-based indices as percentage of the explained HSIC dependence measure between $\mathbf{X} = (X_1, \ldots, X_d)$ and Y:

$$\sum_{A \subseteq \mathcal{P}_d} S_A^{\mathrm{HSIC}} = 1.$$

Interpretation as percentage

Kernel-embedding of probability distributions for GSA: HSIC

New HSIC-based sensitivity index

- Also a decomposition
- Can handle easily structured outputs

Kernel-embedding of probability distributions for GSA: HSIC

New HSIC-based sensitivity index

- Also a decomposition
- Can handle easily structured outputs
- **Generalization of MMD-based index!**

Kernel more or less converging to a dirac

Proposition 2. *For all subset $A \subseteq \mathcal{P}_d$, let us define a product RKHS $\mathcal{H}_A = \mathcal{F}_A \times \mathcal{G}$ with kernel $k_A(\mathbf{x}_A, \mathbf{x}'_A)k_{\mathcal{Y}}(y, y')$. We further assume that $\forall \mathbf{x}_A \in \mathcal{X}_A$, $p_{\mathbf{X}_A}(\mathbf{x}_A) > 0$ and that*

$$k_A(\mathbf{x}_A, \mathbf{x}'_A) = \frac{1}{\sqrt{p_{\mathbf{X}_A}(\mathbf{x}_A)}\sqrt{p_{\mathbf{X}_A}(\mathbf{x}'_A)}} \prod_{l \in A} \frac{1}{h} K\left(\frac{x_l - x'_l}{h}\right) \tag{13}$$

where $K : \mathbb{R} \to \mathbb{R}$ is a symmetric kernel function satisfying $\int_u K(u)du = 1$, and $h > 0$.

Kernel-embedding of probability distributions for GSA: HSIC

New HSIC-based sensitivity index

- Also a decomposition
- Can handle easily structured outputs
- **Generalization of MMD-based index!**

Kernel more or less converging to a dirac

Proposition 2. *For all subset $A \subseteq \mathcal{P}_d$, let us define a product RKHS $\mathcal{H}_A = \mathcal{F}_A \times \mathcal{G}$ with kernel $k_A(\mathbf{x}_A, \mathbf{x}'_A)k_{\mathcal{Y}}(y, y')$. We further assume that $\forall \mathbf{x}_A \in \mathcal{X}_A$, $p_{\mathbf{X}_A}(\mathbf{x}_A) > 0$ and that*

$$k_A(\mathbf{x}_A, \mathbf{x}'_A) = \frac{1}{\sqrt{p_{\mathbf{X}_A}(\mathbf{x}_A)}\sqrt{p_{\mathbf{X}_A}(\mathbf{x}'_A)}} \prod_{l \in A} \frac{1}{h} K\left(\frac{x_l - x'_l}{h}\right) \tag{13}$$

where $K : \mathbb{R} \to \mathbb{R}$ is a symmetric kernel function satisfying $\int_u K(u)du = 1$, and $h > 0$. Then we have $\forall A \subseteq \mathcal{P}_d$

$$\lim_{h \to 0} \mathrm{HSIC}(\mathbf{X}_A, Y) = \mathbb{E}_{\mathbf{X}_A}\left(\mathrm{MMD}^2(\mathrm{P}_Y, \mathrm{P}_{Y|\mathbf{X}_A})\right)$$

where $\mathrm{HSIC}(\mathbf{X}_A, Y)$ is defined with the product RKHS $\mathcal{H}_A = \mathcal{F}_A \times \mathcal{G}$ and $\mathrm{MMD}^2(\mathrm{P}_Y, \mathrm{P}_{Y|\mathbf{X}_A})$ with the RKHS $\mathcal{G}$.

Kernel-embedding of probability distributions for GSA: HSIC

New HSIC-based sensitivity index

- Also a decomposition
- Can handle easily structured outputs
- Generalization of MMD-based index !

Estimation

- Very easy, U-stat or V-stat, see Song et al. (2007); Gretton et al. (2008)

Kernel-embedding of probability distributions for GSA: HSIC

Wait a minute!

In addition, for all $l = 1, \dots, d$ *and* $\forall x_l \in \mathcal{X}_l$, *we have*

$$\int_{\mathcal{X}_l} k_l(x_l, x_l') d\mathrm{P}_{X_l}(x_l') = 0. \tag{11}$$

Zero-mean kernel

> **How do we build a kernel satisfying this?**

Kernel-embedding of probability distributions for GSA: HSIC

Zero-mean kernel

$$\int_{\mathcal{X}_l} k_l(x_l, x_l') d\mathrm{P}_{X_l}(x_l') = 0.$$

Easy case: inputs are uniform on [0,1]

> We can directly use famous Sobolev kernels (from SS-ANOVA, COSSO, ACOSSO, …)

$$k_l(x_l, x_l') = \frac{B_{2r}(|x_l - x_l'|)}{(-1)^{r+1}(2r)!} + \sum_{j=1}^{r} \frac{B_j(x_l) B_j(x_l')}{(j!)^2}$$

where B are Bernoulli polynomials.

> Always possible to transform independent inputs to end up with this case (via probability integral transform)
> But sensitivity index is not invariant via nonlinear transformations
> **See G. Sarazin's talk on Wednedsay (session 6A)**

Kernel-embedding of probability distributions for GSA: HSIC

Zero-mean kernel

$$\int_{\mathcal{X}_l} k_l(x_l, x_l')d\mathrm{P}_{X_l}(x_l') = 0.$$

General case 1

> Kernels built by Durrande et al. (2012) in the context of GP models with ANOVA decomposition inside

$$k_0^D(x, x') = k(x, x') - \frac{\int k(x,t)d\mathrm{P}(t) \int k(x',t)d\mathrm{P}(t)}{\iint k(s,t)d\mathrm{P}(s)d\mathrm{P}(t)}$$

> Built from any initial kernel k

> Very nice theory, but needs numerical integration to compute the second term in general

Kernel-embedding of probability distributions for GSA: HSIC

Zero-mean kernel

$$\int_{\mathcal{X}_l} k_l(x_l, x_l') d\mathrm{P}_{X_l}(x_l') = 0.$$

General case 2

> Kernels introduced in the context of Stein discrepancy in lieu of MMD

$$k_0^S(\mathbf{x}, \mathbf{x}') = \nabla_{\mathbf{x}} \nabla_{\mathbf{x}'} k(\mathbf{x}, \mathbf{x}') + \frac{\nabla_{\mathbf{x}} p(\mathbf{x})}{p(\mathbf{x})} \nabla_{\mathbf{x}'} k(\mathbf{x}, \mathbf{x}') + \frac{\nabla_{\mathbf{x}'} p(\mathbf{x}')}{p(\mathbf{x}')} \nabla_{\mathbf{x}} k(\mathbf{x}, \mathbf{x}') + \frac{\nabla_{\mathbf{x}} p(\mathbf{x})}{p(\mathbf{x})} \frac{\nabla_{\mathbf{x}'} p(\mathbf{x}')}{p(\mathbf{x}')} k(\mathbf{x}, \mathbf{x}')$$

> Built from any initial kernel k again, but must be differentiable this time
> Needs derivative of the log pdf of the inputs
> Means that we only need to know the pdf up to a constant
- Trick extensively used lately (see Chris' talk)
- **A potential interest for GSA problems where some inputs are obtained through Bayesian calibration**

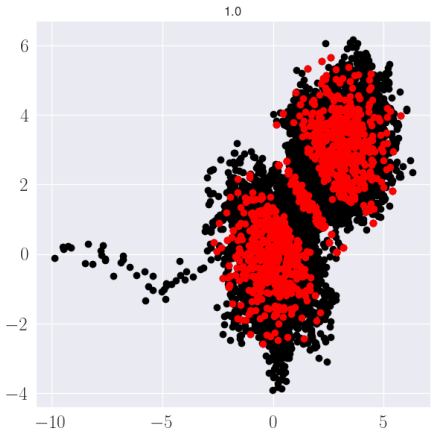
Conclusion & outlook

Kernel methods have long been used in computer experiments for surrogate models

- GP regression is very popular, with a very large literature on the subject (both theory / methodology / applications)
- There is still extensive work (beyond what I mentioned in this talk)
  - Non-stationary kernels, kernel selection
  - Criteria for sequential DOE strategies
  - Adaptation for stochastic simulators and robust optimization

For DOE and sensitivity analysis, this is still an emerging field

- Great promises illustrated only recently
- Very exciting research directions to be investigated
  - Still kernel selection, but also **fixing the flaws of the KSD**
  - Optimization algorithms for DOE, online selection
  - CLT and more efficient estimators for sensitivity analysis



Thank you for your attention

References (1/3)

Agrell, C. (2019). Gaussian Processes with Linear Operator Inequality Constraints. *Journal of Machine Learning Research*, *20*(135), 1-36.

Barr, J., & Rabitz, H. (2022). A Generalized Kernel Method for Global Sensitivity Analysis. *SIAM/ASA Journal on Uncertainty Quantification*, *10*(1), 27-54.

Baucells, M., & Borgonovo, E. (2013), 'Invariant probabilistic sensitivity analysis', *Management Science*, *59*(11), 2536-2549.

Borgonovo, E. (2007), 'A new uncertainty importance measure', Reliability Engineering & System Safety 92(6), 771–784.

Cao, J., Guinness, J., Genton, M. G., & Katzfuss, M. (2022). Scalable Gaussian-process regression and variable selection using Vecchia approximations. *arXiv preprint arXiv:2202.12981*.

Chastaing, G., Gamboa, F., Prieur, C. et al. (2012), 'Generalized hoeffding-sobol decomposition for dependent variables - application to sensitivity analysis', Electronic Journal of Statistics 6, 2420–2448.

Chen, Y., Welling, M., & Smola, A. (2012). Super-samples from kernel herding. *arXiv preprint arXiv:1203.3472*.

Cuturi, M. (2011), Fast global alignment kernels, in 'Proceedings of the 28th international conference on machine learning (ICML-11)', pp. 929–936.

Da Veiga, S., & Marrel, A. (2012). Gaussian process modeling with inequality constraints. In *Annales de la Faculté des sciences de Toulouse: Math*

Da Veiga, S. (2015), 'Global sensitivity analysis with dependence measures', Journal of Statistical Computation and Simulation 85(7), 1283–1305.

Da Veiga, S., & Marrel, A. (2020). Gaussian process regression with linear inequality constraints. *Reliability Engineering & System Safety*, *195*, 106732.

Da Veiga, S. (2021), 'Kernel-based anova decomposition and shapley effects - application to global sensitivity analysis' , https://arxiv.org/abs/2101.05487.

Da Veiga, S., Gamboa, F., Iooss, B., & Prieur, C. (2021). *Basics and Trends in Sensitivity Analysis: Theory and Practice in R*. Society for Industrial and Applied Mathematics.

Fort, J.-C., Klein, T. and Rachdi, N. (2016), 'New sensitivity analysis subordinated to a contrast', Communications in Statistics-Theory and Methods 45(15), 4349–4364.

Ginsbourger, D., Bay, X., Roustant, O., & Carraro, L. (2012). Argumentwise invariant kernels for the approximation of invariant functions. In *Annales de la Faculté des sciences de Toulouse: Mathématiques* (Vol. 21, No. 3, pp. 501-527).

References (2/3)

Gretton, A., Bousquet, O., Smola, A. and Scholkopf, B. (2005a), Measuring statistical dependence with hilbert-schmidt norms, in S. Jain, H. Simon and E. Tomita, eds, 'Algorithmic Learning Theory', Vol. 3734 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, pp. 63–77.

Gretton, A., Herbrich, R., Smola, A., Bousquet, O. and Scholkopf, B. (2005b), 'Kernel methods for measuring independence', The Journal of Machine Learning Research 6, 2075–2129.

Hickernell, F. (1998). A generalized discrepancy and quadrature error bound. *Mathematics of computation*, *67*(221), 299-322.

Hoeffding, W. (1948), 'A class of statistics with asymptotically normal distributions', Annals of Mathematical Statistics 19, 293–325.

Huang, C., Joseph, V. R., & Ray, D. M. (2021). Constrained minimum energy designs. *Statistics and Computing*, *31*(6), 1-15.

Idrissi, M. I., Chabridon, V., and Iooss, B. (2021), 'Developments and applications of Shapley effects to reliability-oriented sensitivity analysis with correlated inputs', *arXiv preprint arXiv:2101.08083*.

Iooss, B., & Marrel, A. (2019). Advanced methodology for uncertainty propagation in computer experiments with large number of inputs. *Nuclear Technology*.

Iooss, B. and Prieur, C. (2019), 'Shapley effects for sensitivity analysis with dependent inputs: comparisons with Sobol' indices, numerical estimation and applications', International Journal for Uncertainty Quantification 9, 493–514.

Joseph, V. R., & Vakayil, A. (2022). Split: An optimal method for data splitting. *Technometrics*, *64*(2), 166-176.

Korba, A., Aubin-Frankowski, P. C., Majewski, S., & Ablin, P. (2021, July). Kernel stein discrepancy descent. In *International Conference on Machine Learning* (pp. 5719-5730). PMLR.

Kuo, F., Sloan, I., Wasilkowski, G. and Wozniakowski, H. (2010), 'On decompositions of multivari- ate functions', Mathematics of computation 79(270), 953–966.

Lin, L. H., & Roshan Joseph, V. (2020). Transformation and additivity in Gaussian processes. *Technometrics*, *62*(4), 525-535.

Mak, S., & Joseph, V. R. (2018). Support points. *The Annals of Statistics*, *46*(6A), 2562-2592.

Mara, T. A., Tarantola, S. and Annoni, P. (2015), 'Non-parametric methods for global sensitivity analysis of model output with dependent inputs', Environmental modelling & software 72, 173– 183.

Maroñas, J., Hamelijnck, O., Knoblauch, J., & Damoulas, T. (2021, March). Transforming Gaussian processes with normalizing flows. In *International Conference on Artificial Intelligence and Statistics* (pp. 1081-1089). PMLR.

References (3/3)

Owen, A. (2014), ‘Sobol’ indices and Shapley value’, SIAM/ASA Journal on Uncertainty Quantification 2, 245–251.

Perrin, G., & Da Veiga, S. (2021). Constrained Gaussian process regression: an adaptive approach for the estimation of hyperparameters and the verification of constraints with high probability. *Journal of Machine Learning for Modeling and Computing*, *2*(2).

Pronzato, L. (2021). Performance analysis of greedy algorithms for minimising a Maximum Mean Discrepancy. *arXiv preprint arXiv:2101.07564*.

Rahman, S. (2016), ‘The f-sensitivity index’, SIAM/ASA Journal on Uncertainty Quantification 4(1), 130–162.

Riabiz, M., Chen, W., Cockayne, J., Swietach, P., Niederer, S. A., Mackey, L., & Oates, C. (2020). Optimal thinning of MCMC output. *arXiv preprint arXiv:2005.03952*.

Scheuerer, M., & Schlather, M. (2012). Covariance models for divergence-free and curl-free random vector fields. *Stochastic Models*, *28*(3), 433-451.

Shapley, L. (1953), A value for n-persons game, in H. Kuhn and A. Tucker, eds, ‘Contributions to the theory of games II, Annals of mathematic studies’, Princeton University Press, Princeton, NJ.

Song, E., Nelson, B. L. and Staum, J. (2016), ‘Shapley effects for global sensitivity analysis: Theory and computation’, SIAM/ASA Journal on Uncertainty Quantification 4(1), 1060–1083.

Song, L. (2008), Learning via Hilbert Space Embedding of Distributions, PhD thesis, University of Sydney.

Székely, G. J., & Rizzo, M. L. (2013), ‘Energy statistics: A class of statistics based on distances’, *Journal of statistical planning and inference*, *143*(8), 1249-1272.

Teymur, O., Gorham, J., Riabiz, M., & Oates, C. (2021, March). Optimal quantisation of probability measures using maximum mean discrepancy. In *International Conference on Artificial Intelligence and Statistics* (pp. 1027-1035). PMLR.

Tran, T., Da Veiga, S., Sinoquet, D., & Mongeau, M. (2022). Design of experiments for mixed continuous and discrete variables.

Yi, G., Shi, J. Q., & Choi, T. (2011). Penalized Gaussian Process Regression and Classification for High-Dimensional Nonlinear Data. *Biometrics*, *67*(4), 1285-1294.