Computer experiments
Overview of GP-based approaches

O. Roustant[a]

with acknowledgments to M. Binois, Y. Deville, N. Durrande for nice illustrations!

INSA Toulouse, September 2021

Outline

Outline

Metamodeling – Computer experiments

design

Metamodeling – Computer experiments

design → **experiment + measurements** → **meas. output**

Metamodeling – Computer experiments

design → **experiment + measurements** → **meas. output**

simulator inputs
$\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n$

→ **simulator**
physical model, numerical methods

→ **simulator output**
$f_{\text{sim}}(\mathbf{x}^1),\ f_{\text{sim}}(\mathbf{x}^2),\ \ldots,\ f_{\text{sim}}(\mathbf{x}^n)$

Metamodeling – Computer experiments

design → **experiment + measurements** → **meas. output**

simulator inputs $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n$ → **simulator** *physical model, numerical methods* → **simulator output** $f_{\text{sim}}(\mathbf{x}^1), f_{\text{sim}}(\mathbf{x}^2), \ldots, f_{\text{sim}}(\mathbf{x}^n)$

metamodel inputs $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n$ → **metamodel** $y = \mathbf{g}(\mathbf{x})^\top \beta + Z(\mathbf{x})$ → **metamodel output** $\widehat{y}(\mathbf{x}), \ldots$

Metamodeling – Computer experiments

design

experiment + measurements

meas. output

simulator inputs
$\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n$

simulator
physical model, numerical methods

simulator output
$f_{\text{sim}}(\mathbf{x}^1), f_{\text{sim}}(\mathbf{x}^2), \ldots, f_{\text{sim}}(\mathbf{x}^n)$

param.

design

metamodel inputs
$\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n$

metamodel
$y = \mathbf{g}(\mathbf{x})^\top \beta + Z(\mathbf{x})$

metamodel output
$\widehat{y}(\mathbf{x}), \ldots$

Metamodels : Why Gaussian processes are famous ?

- A metamodel is a fast model which approximates a time-consuming model. It can be any regression model, probabilistic or deterministic
  - Linear model, random forest, neural network, spline, polynomial chaos, ...

Metamodels : Why Gaussian processes are famous ?

- A metamodel is a fast model which approximates a time-consuming model. It can be any regression model, probabilistic or deterministic
  - Linear model, random forest, neural network, spline, polynomial chaos, ...
- Requirement : an uncertainty measure assessing ignorance at unknown area
  - This claims in favor of probabilistic models

Metamodels : Why Gaussian processes are famous ?

- A metamodel is a fast model which approximates a time-consuming model. It can be any regression model, probabilistic or deterministic
  - Linear model, random forest, neural network, spline, polynomial chaos, ...
- Requirement : an uncertainty measure assessing ignorance at unknown area
  - This claims in favor of probabilistic models
- Gaussian processes (GP) have nice features :
  - Their uncertainty measure have a closed-form
  - They are flexible (parameterized by two functions) and can handle prior information on data
  - They generalize splines

$\rightarrow$ In this course, we focus on GP (regression) models

Gaussian processes

Gaussian processes are stochastic processes (or random fields) s.t. every finite dimensional distribution is Gaussian $\rightarrow$ Parameterized by two functions

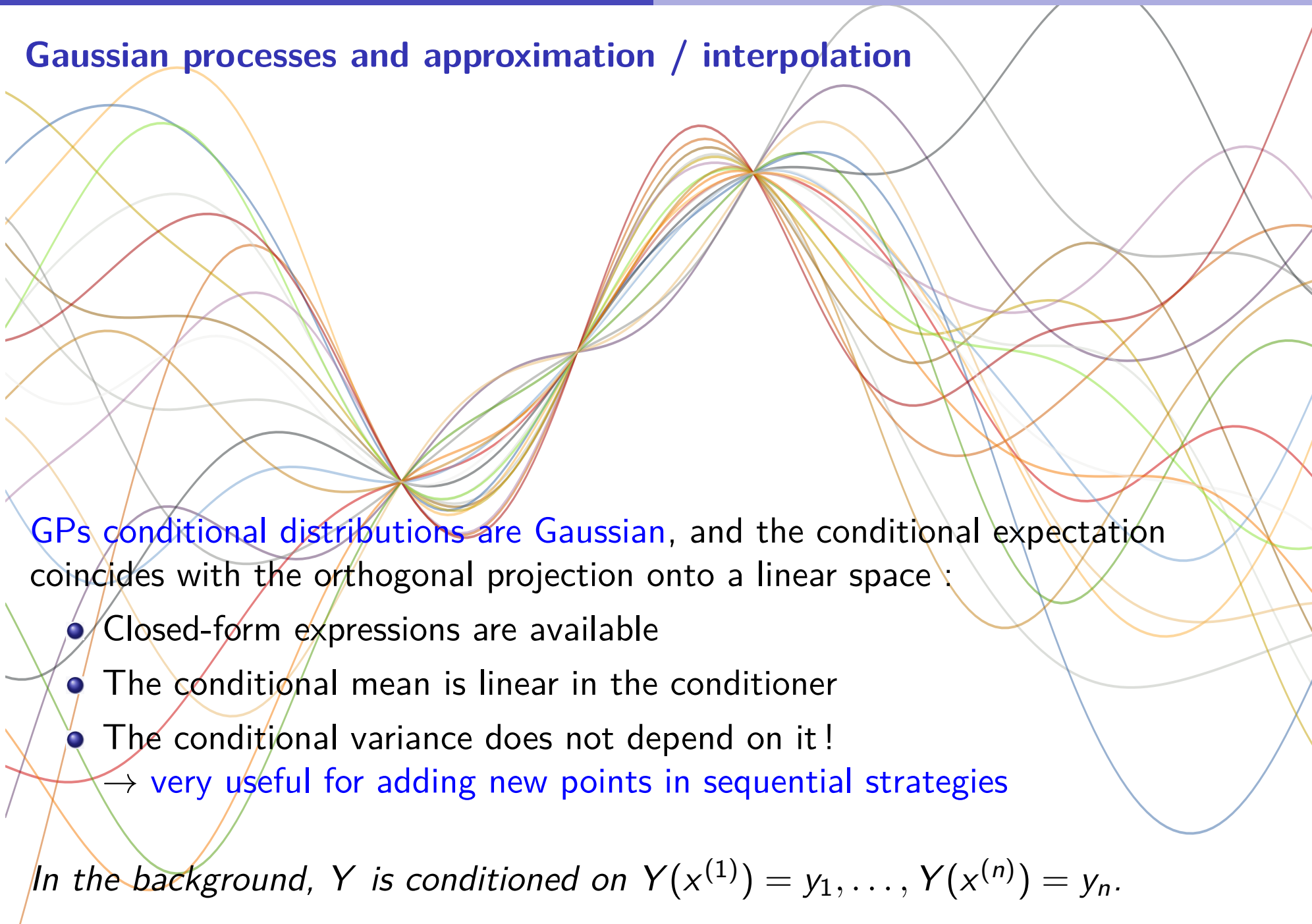
$$Y = (Y(x))_{x \in T} \sim GP(\underbrace{m(\mathbf{x})}_{trend}, \underbrace{k(\mathbf{x}, \mathbf{x}')}_{kernel})$$

- The trend can be any function.
- The kernel is positive semidefinite :

$$\forall n, \alpha_1, \ldots, \alpha_n, \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}, \qquad \sum_{i=1}^{n} \alpha_i \alpha_j k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) \geq 0.$$

It contains the spatial dependence.

Gaussian processes and approximation / interpolation



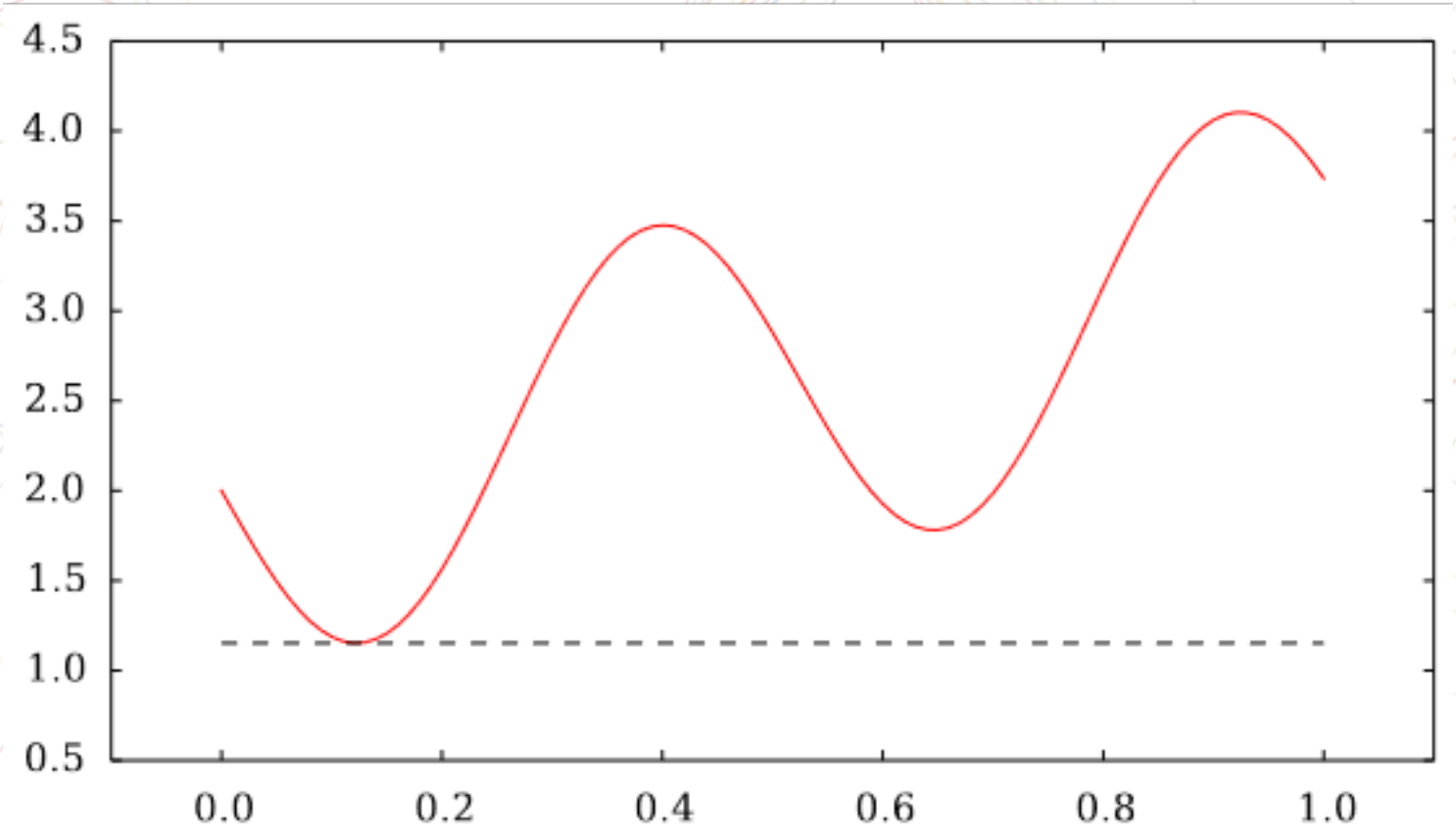
GPs conditional distributions are Gaussian, and the conditional expectation coincides with the orthogonal projection onto a linear space :

- Closed-form expressions are available
- The conditional mean is linear in the conditioner
- The conditional variance does not depend on it !
  $\rightarrow$ very useful for adding new points in sequential strategies

In the background, Y is conditioned on $Y(x^{(1)}) = y_1, \ldots, Y(x^{(n)}) = y_n$.

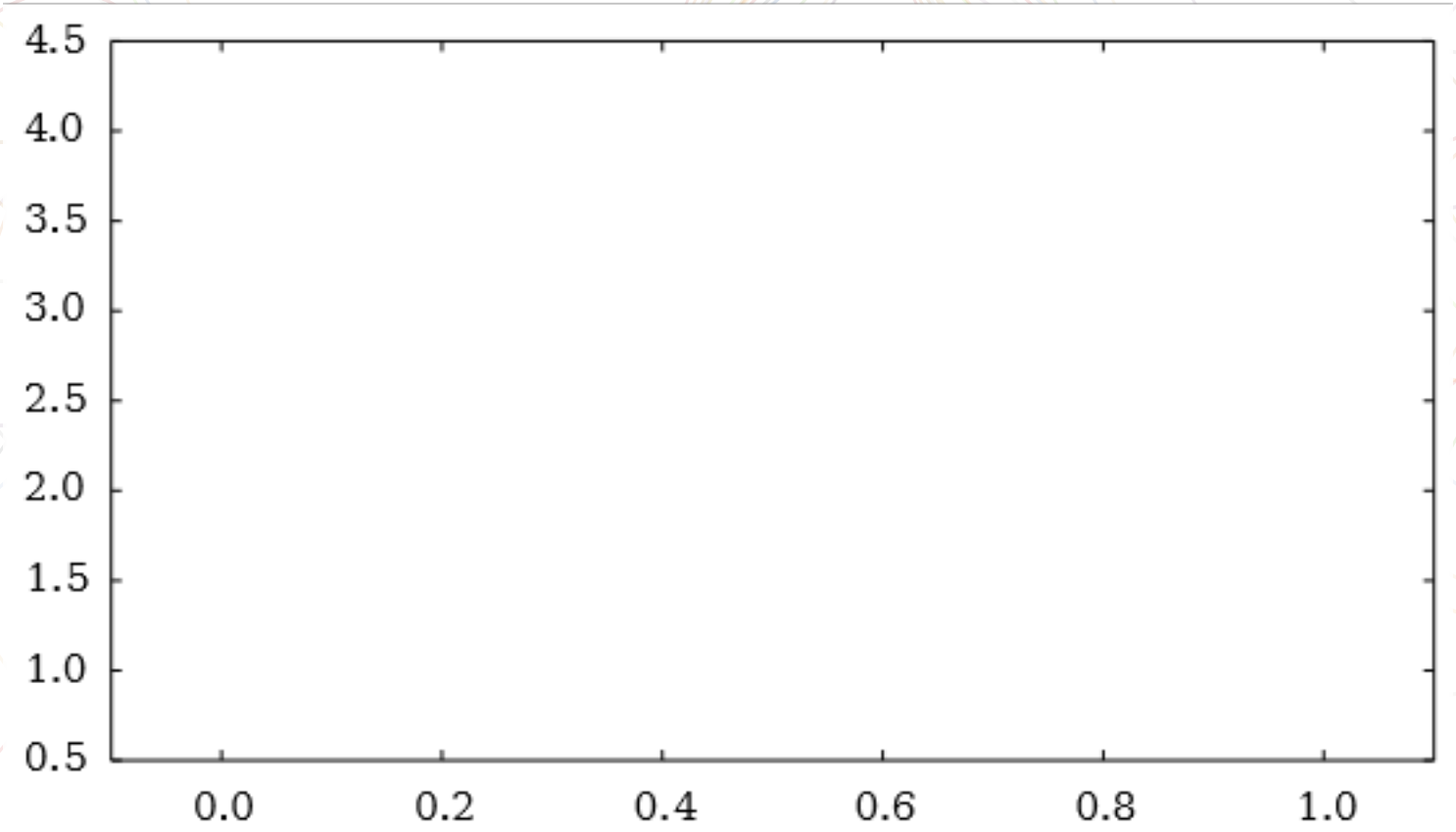
Bayesian optimization

How to find the global minimum of a function... when each evaluation is costly ?



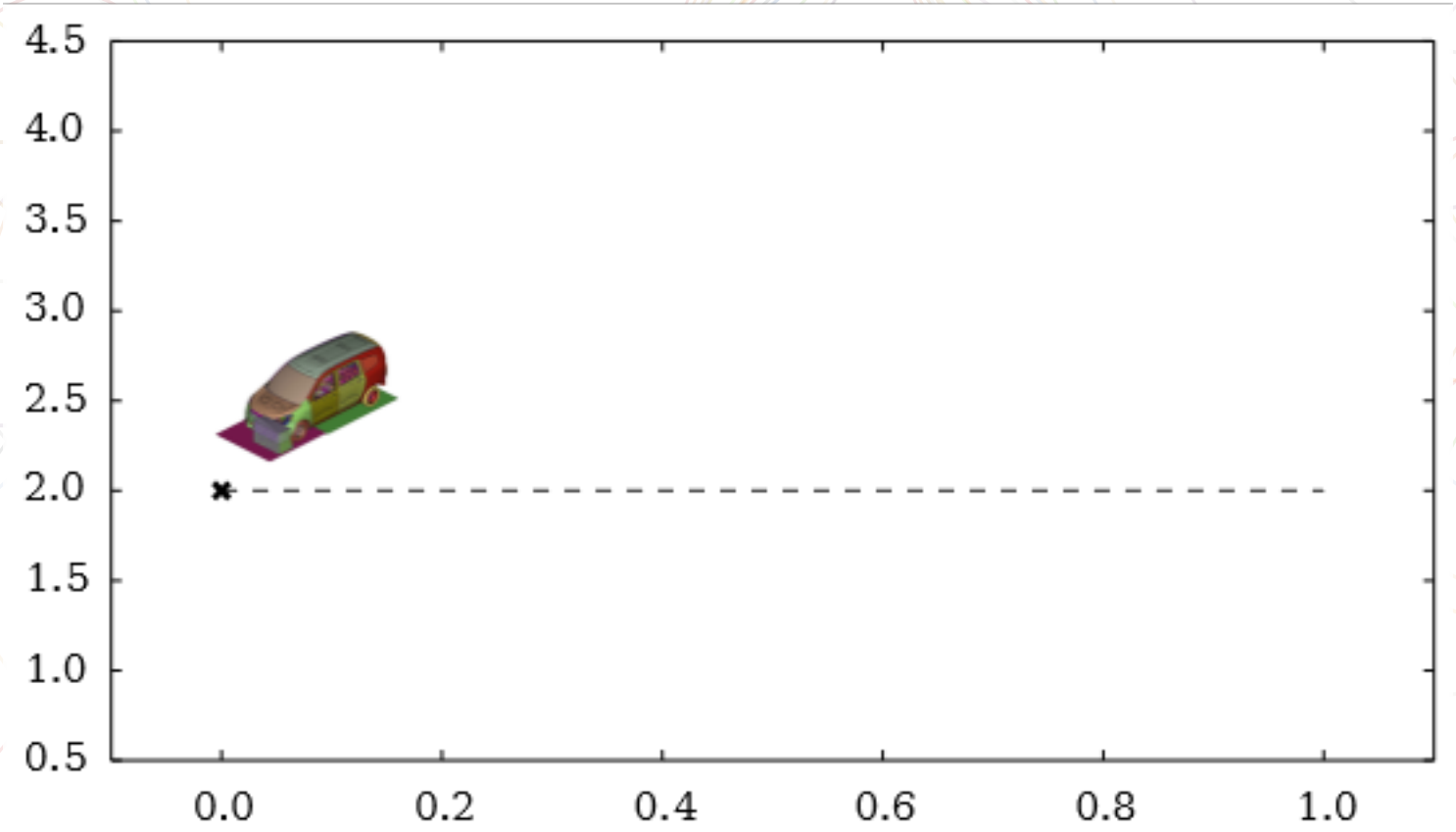
Bayesian optimization

How to find the global minimum of a function... when each evaluation is costly ?



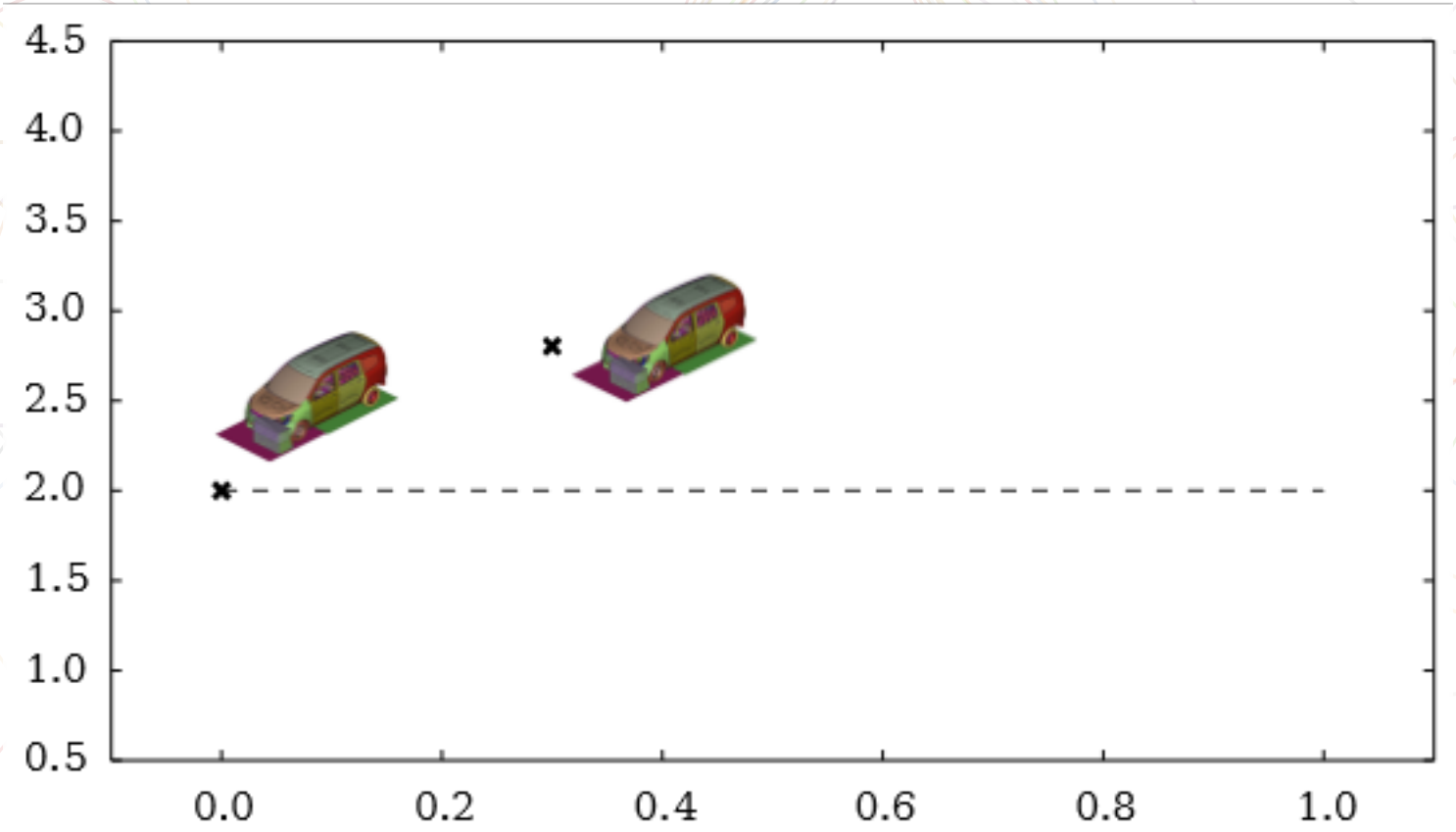
Bayesian optimization

How to find the global minimum of a function... when each evaluation is costly ?



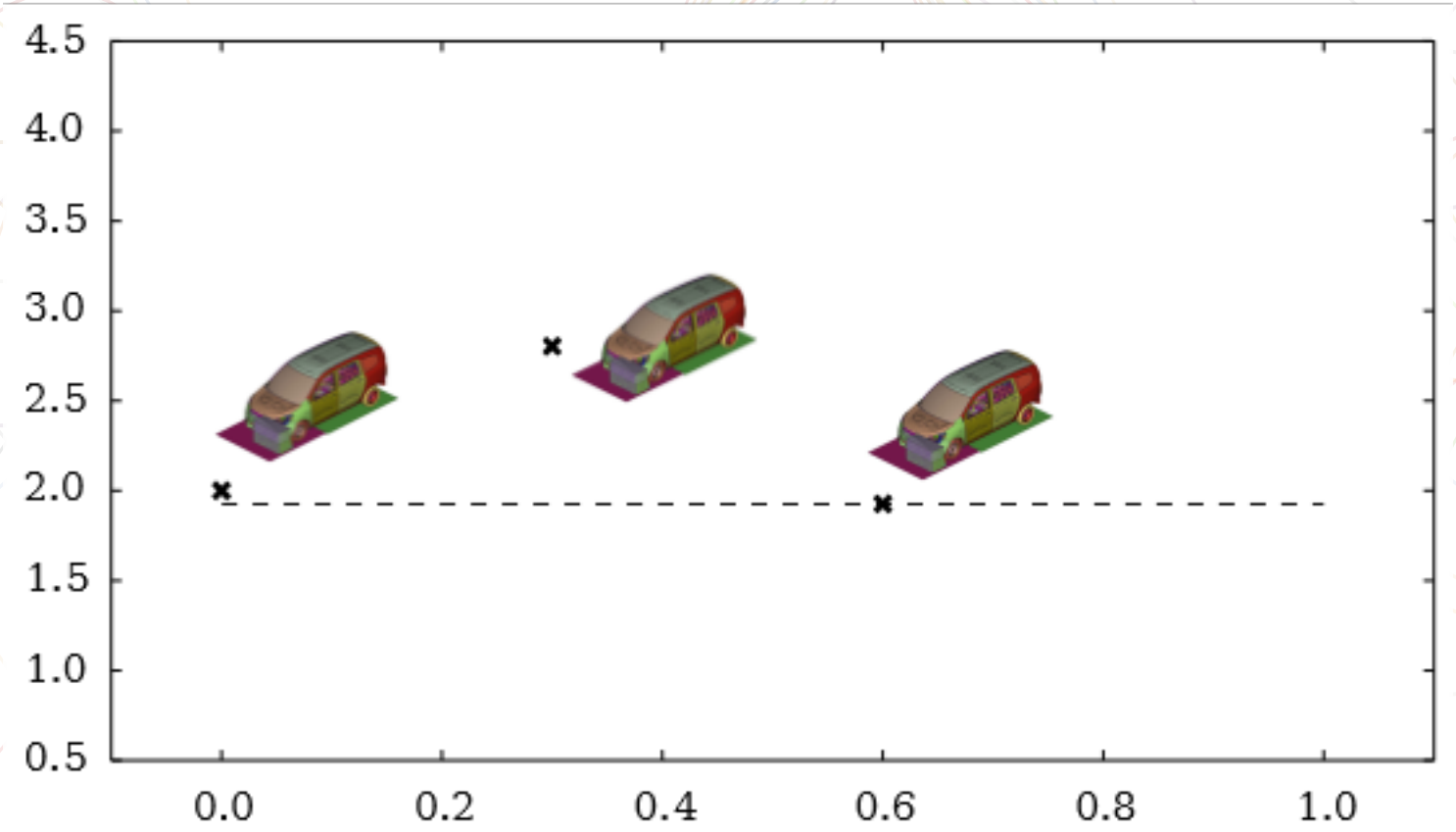
Bayesian optimization

How to find the global minimum of a function... when each evaluation is costly ?



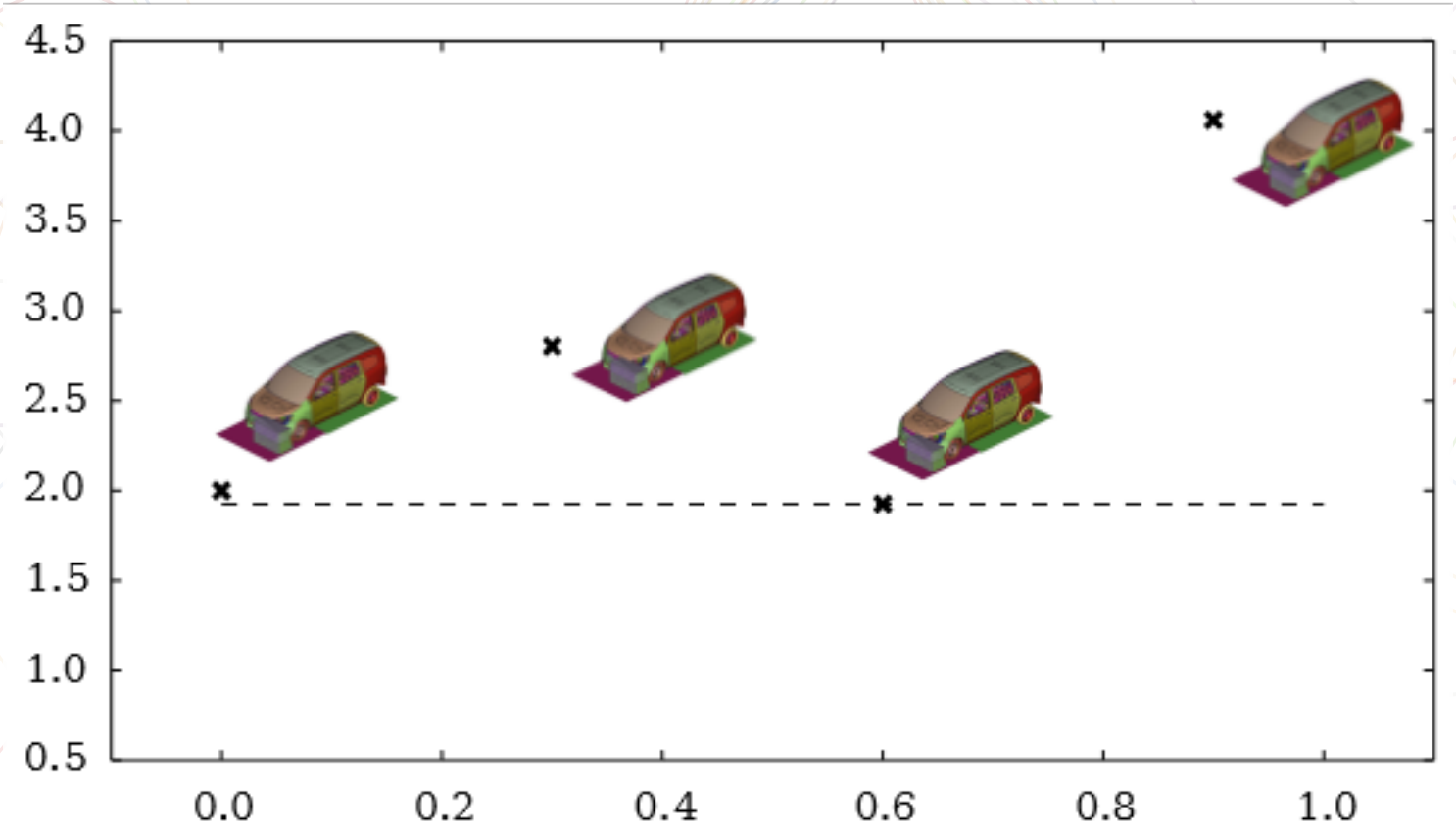
Bayesian optimization

How to find the global minimum of a function... when each evaluation is costly ?



Bayesian optimization

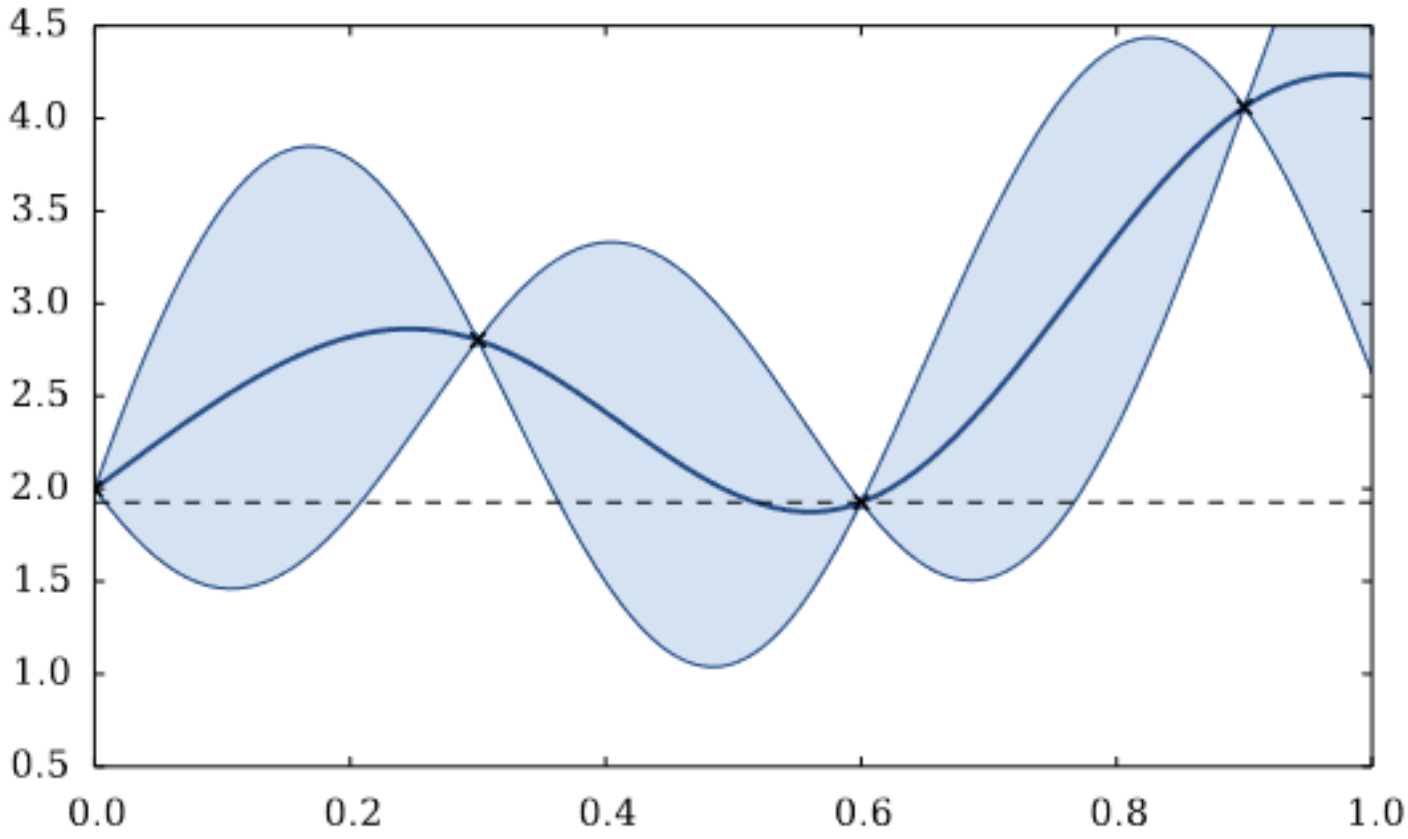
How to find the global minimum of a function... when each evaluation is costly ?



Bayesian optimization

A solution : **Bayesian optimization (BO)** [Močkus, 1975, Jones et al., 1998]

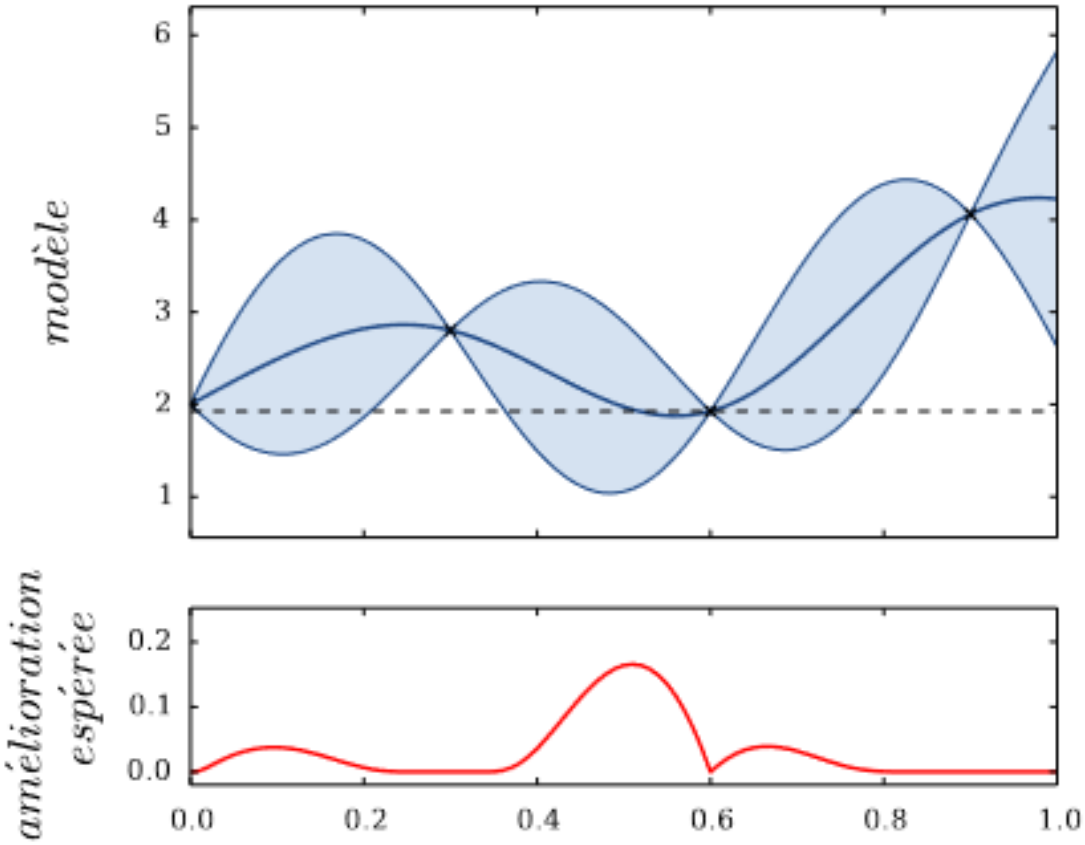
First ingredient : a GP model Y



Bayesian optimization

Second ingredient : an **easy-to-compute** criterion **accounting for uncertainty at unknown regions**, e.g. here "expected improvement"

$$EI(x) = \mathbb{E}([y_0 - Y(x)]^+ | Y(x_1), \ldots, Y(x_n)) \qquad y_0 : \textit{ current minimum}$$



Bayesian optimization

Notice that the expected improvement is indeed **easy-to-compute**.
Denote m_k, s_k the so-called conditional mean & standard deviation, i.e.

$$Y(x)|Y(x_1),\ldots,Y(x_n) \sim \mathcal{N}\left(m_k(x), s_k(x)^2\right)$$

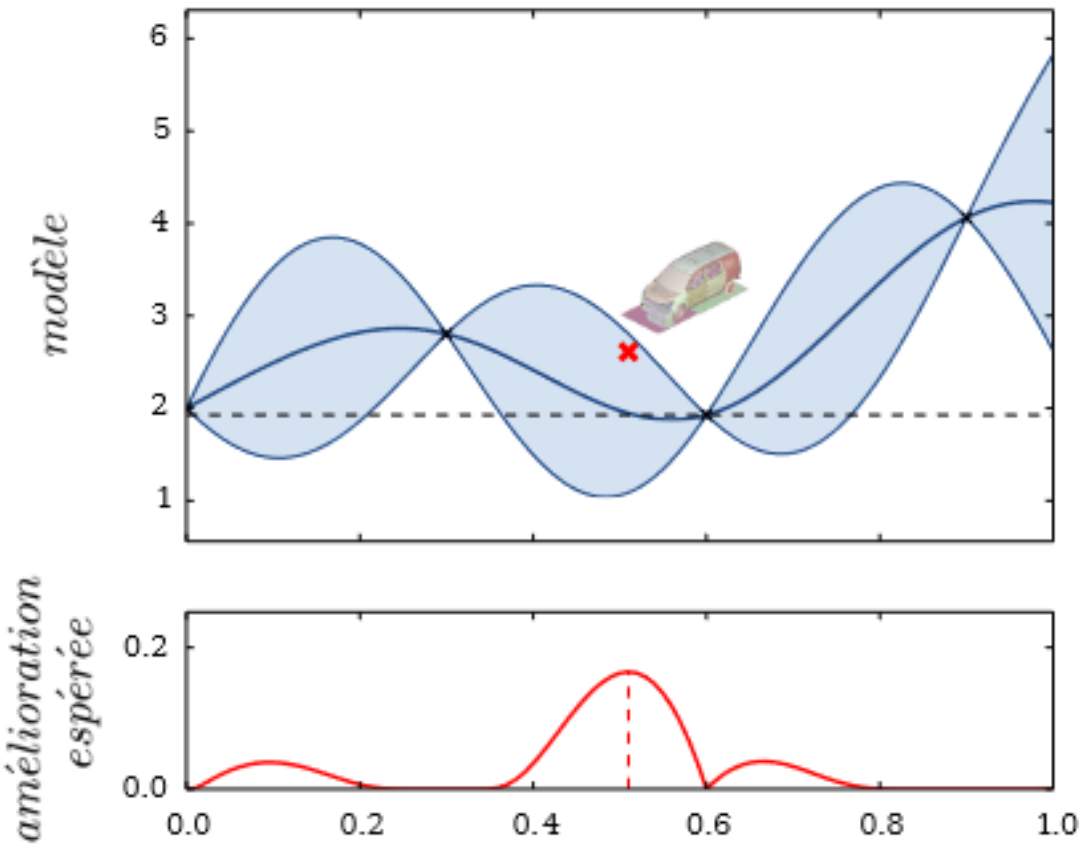
Then, we have the closed-form expression :

$$EI(x) \quad = \quad s_k(x)(z_0\Phi(z_0) + \phi(z_0))$$

with $z_0 = \dfrac{y_0 - m_k(x)}{s_k(x)}$ and ϕ, Φ the pdf, cdf of the $\mathcal{N}(0,1)$ distribution.

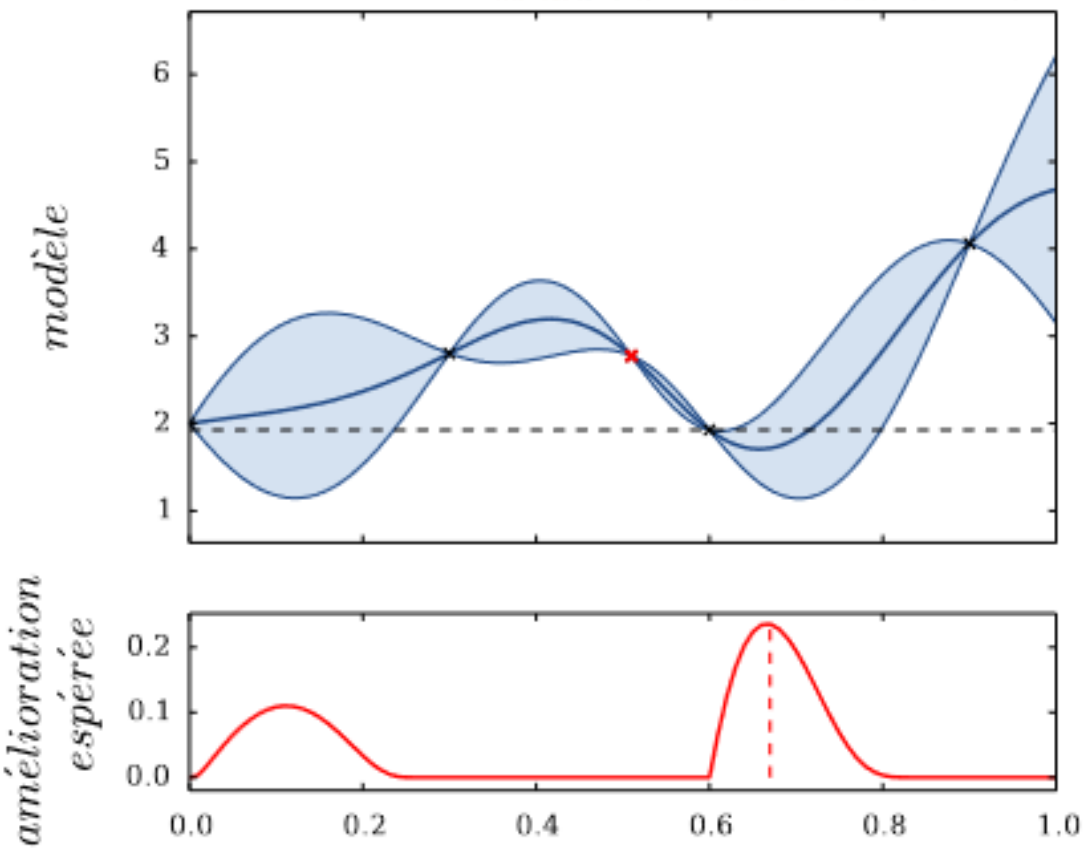
Bayesian optimization

The algorithm (here "EGO") : (1) Find the next point by maximizing the criterion $\rightarrow$ (2) Evaluate the function $\rightarrow$ (3) Update the GP model $\uparrow$



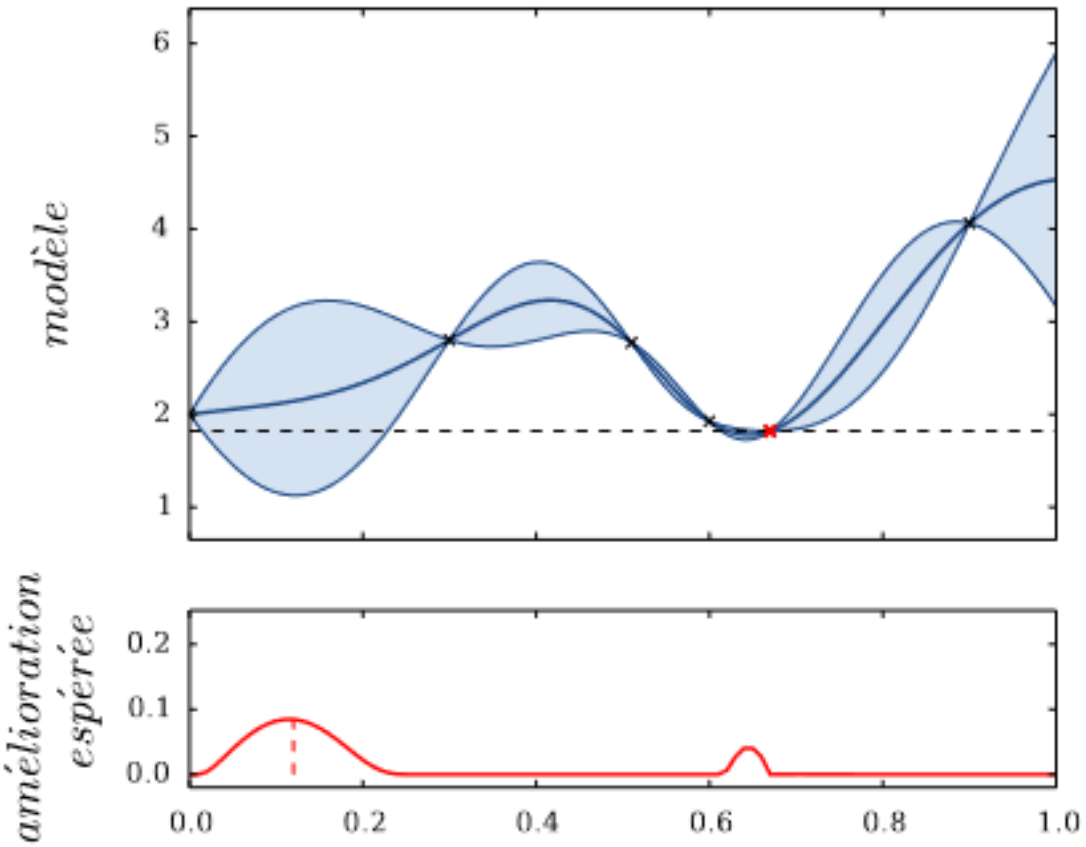
Bayesian optimization

Iteration 2 :



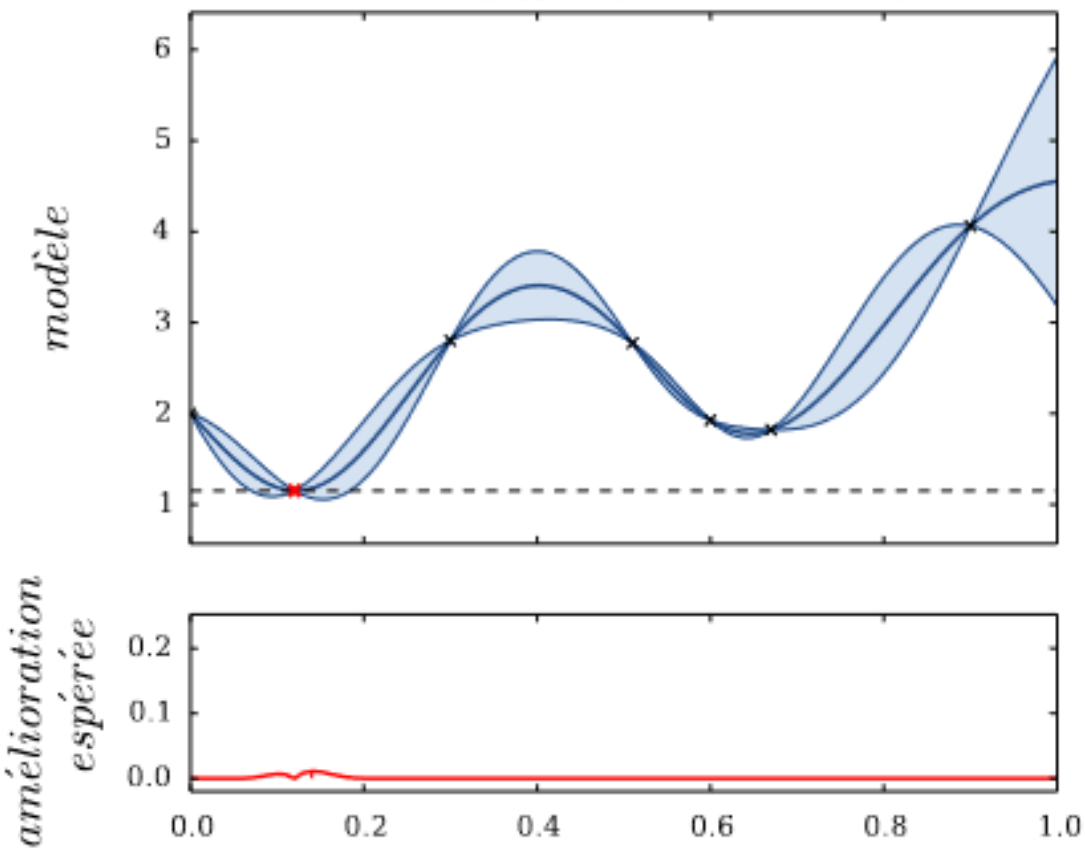
Bayesian optimization

Iteration 3 :



Bayesian optimization

Theory shows that **EGO algorithm provides a dense sequence of points**, up to a slight condition on the kernel used for GPs [Vazquez and Bect, 2010] .



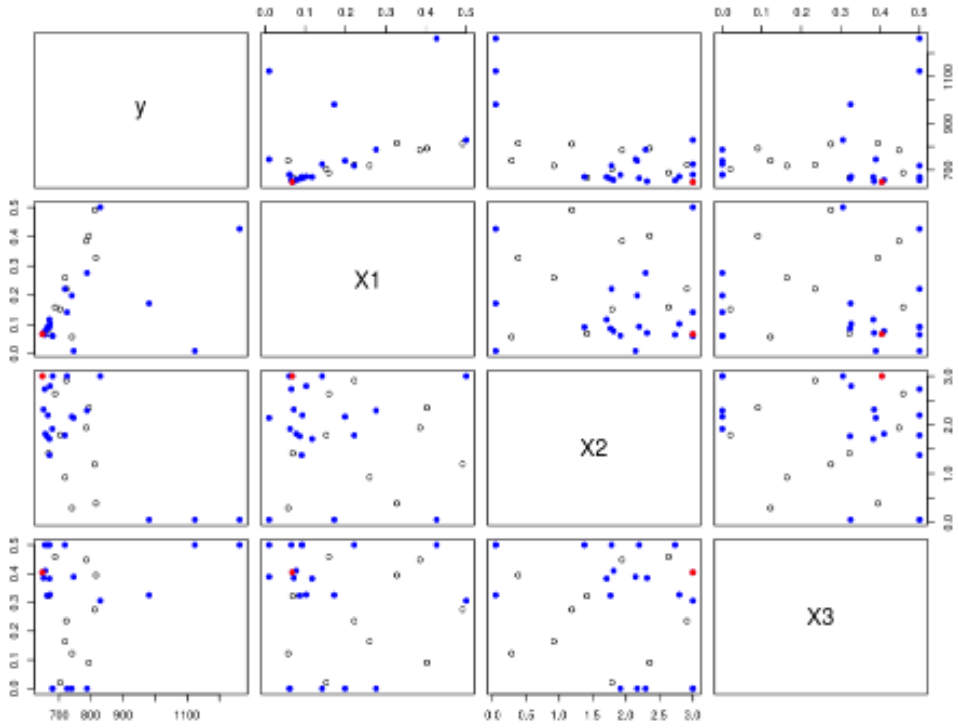
Application to algorithm tuning in machine learning

In machine learning, an algorithm can be a (time-consuming) black-box.
Example on ozone data (from computer lab) :

- ouput : k-fold cross validation error (fixed folds)
- inputs : kernel parameter, cost (regulariz. param.), epsilon (tube size)

With a small 30-point budget, BO outperforms a grid search and default tuning. Observe the tradeoff between exploration/exploitation.

Default tuning	678.8
Grid search	678.4
Marginal optim.	655.4
Bayesian optim.	647.9

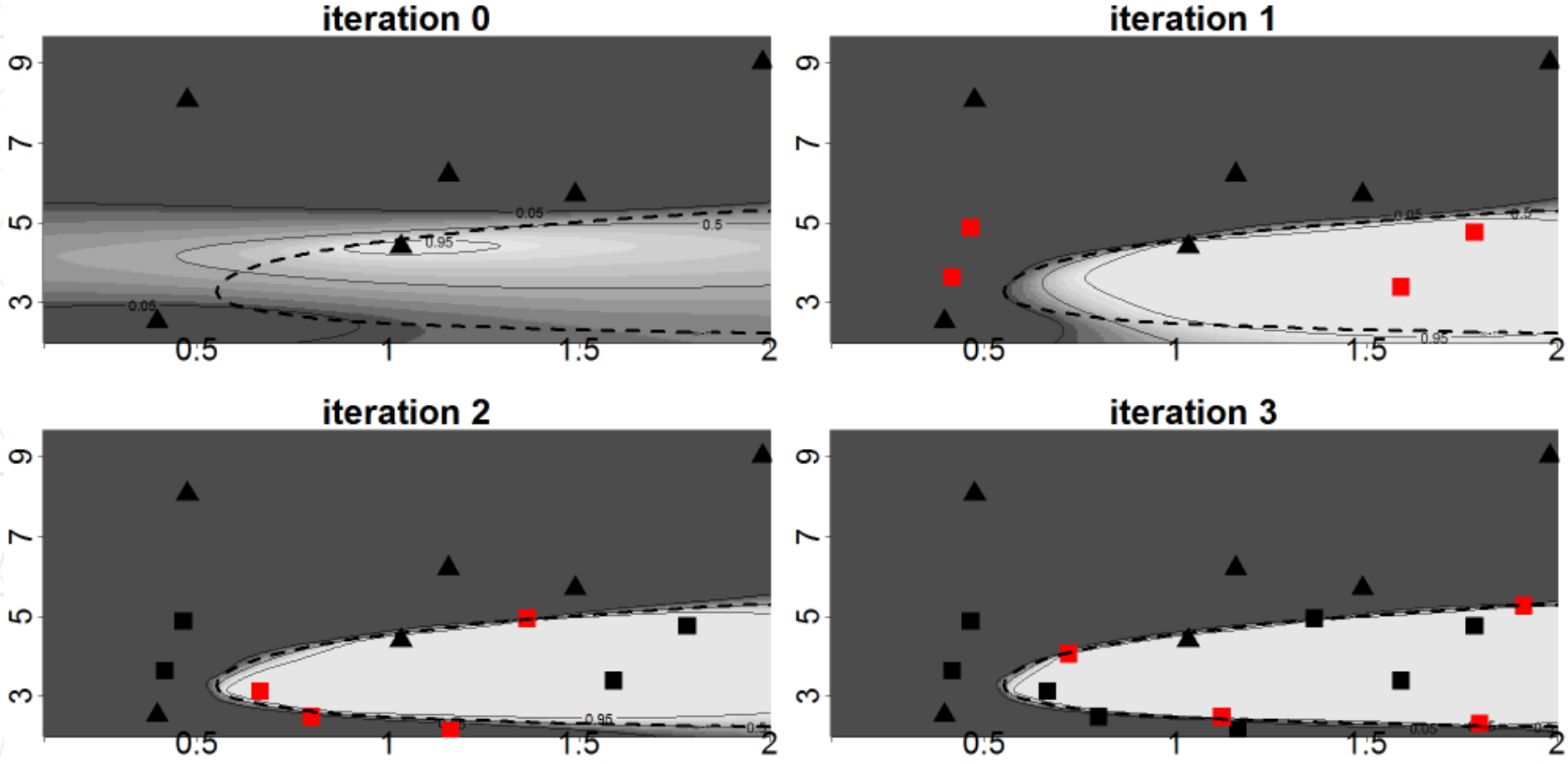


GP-based inversion

Same receipt for estimating a probability of failure ("SUR" strategy).

See [Chevalier et al., 2014] for details and [Bect et al., 2017] for a convergence analysis with supermartingales.

Illustration : Estimation of the nuclear criticity region $k_{\text{eff}} > 0.95$



Adaptation of EGO to noisy observations : the EQI criterion

For noisy observations, we assume that

$$Y_i = Y(x_i) + \varepsilon_i \quad i = 1, \ldots, n$$

where $Y \sim GP(0, k)$ and the $\varepsilon_i's$ are $N(0, \tau_i^2)$ independent mutually and of Y.
The aim is to predict $Y(x)$ given $Y_1, \ldots, Y_n \rightarrow$ filtering.

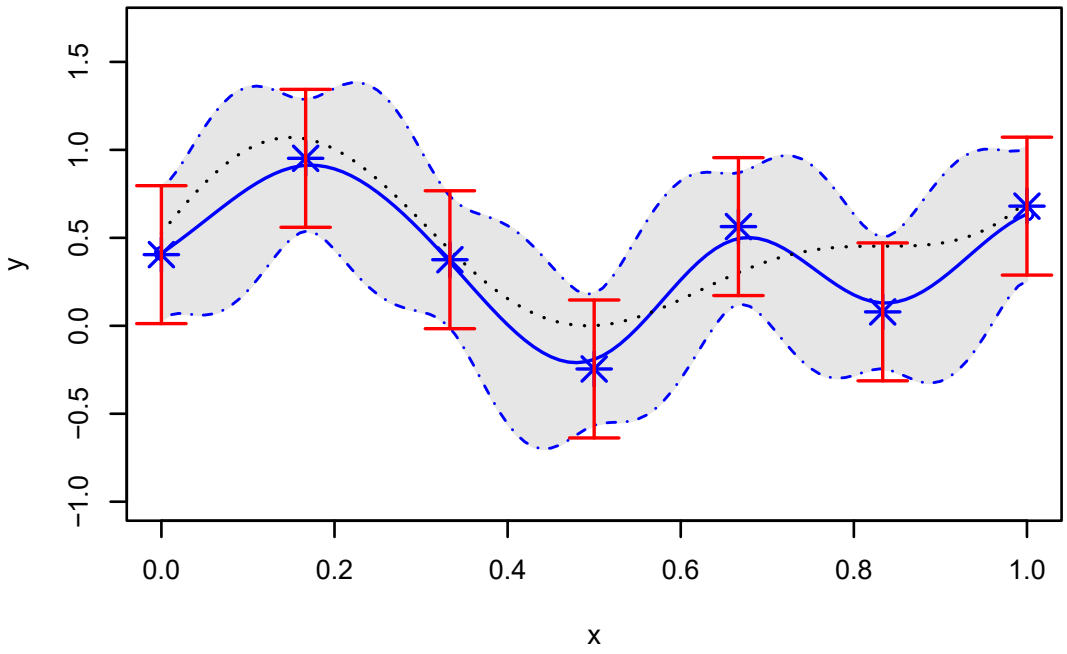


Figure – Illustration of Kriging with noisy observations.

Adaptation of EGO to noisy observations : the EQI criterion

For noisy observations, the improvement is replaced by the quantile improvement, this is the Expected Quantile Improvement criterion (EQI) [Picheny et al., 2013].

Conditionaly on $\mathcal{F}_n = \{Y_1, \ldots, Y_n\}$, given an order β (e.g. 0.9),

$$EQI(x) = \mathbb{E}([q_0 - Q(x)]^+) \qquad q_0 : \textit{ current minimum quantile}$$

where q_O is the minimum of the quantiles (at order β) of the laws of $Y(x_i)$, and $Q(x)$ is the quantile of $Y(x)$ knowing the *unobserved* $Y_{n+1} = Y(x) + \varepsilon_{n+1}$, where $\varepsilon_{n+1} \sim N(0, \tau^2_{n+1})$ for a *given* τ_{n+1}.

Although Y_{n+1} is unobserved, the law of $Q(x)$ can be computed as a normal distribution, leading to an analytical expression for EQI.

Adaptation of EGO to noisy observations : the EQI criterion

Exercise (Gaussian filtering)

Let (Z, ε) a centered Gaussian vector with Z, ε independent. Then $Z|\{Z+\epsilon\}$ is normally distributed with mean $\frac{\sigma_Z^2}{\sigma_Z^2+\sigma_\varepsilon^2}(Z+\varepsilon)$ and variance $\frac{\sigma_Z^2\sigma_\varepsilon^2}{\sigma_Z^2+\sigma_\varepsilon^2}$.

Exercise (EQI expression)

Conditionaly on $\mathcal{F}_n$:

1. The law of $Y(x)|Y_{n+1} \sim \mathcal{N}\left(M_{n+1}(x), S_{n+1}(x)^2\right)$ with
   $M_{n+1}(x) = m_k(x) + \frac{s_k(x)^2}{s_k(x)^2+\tau_{n+1}^2}(Y_{n+1} - m_k(x)), \qquad S_{n+1}(x)^2 = \frac{s_k(x)^2\tau_{n+1}^2}{s_k(x)^2+\tau_{n+1}^2}$
2. Thus $Q(x) = M_{n+1}(x) + \Phi^{-1}(\beta)S_{n+1}(x)$.
3. Moreover $Q(x) \sim \mathcal{N}\left(m_Q(x), s_Q(x)^2\right)$ with

$$m_Q = m_k(x) + \Phi_{-1}(\beta)\frac{s_k(x)\tau_{n+1}}{\sqrt{s_k(x)^2+\tau_{n+1}^2}}, \qquad s_Q^2 = \frac{s_k(x)^4}{s_k(x)^2+\tau_{n+1}^2}$$

Finally, the EQI criterion has the same analytical expression as EI criterion, replacing y_0, m_k, s_k by q_0, m_Q, s_Q.

Outline

From geostatistics to GP regression models

Time line

- 1951 : Spatial interpolation in geosciences [Krige, 1951]
  $\rightarrow$ "Kriging"
- 1963 : Foundations of geostatistics [Matheron, 1963]
- 1989 : Computer experiments, metamodelling [Sacks et al., 1989]
  $\rightarrow$ Application to dimensions ≥ 4

From geostatistics to GP regression models

Geostatiscal approach for spatial interpolation, Simple Kriging

Let Y be a centered stochastic process (or with known mean).
In geostatistics, the prediction of $Y(x)$ knowing $Y(x^{(1)}), \ldots, Y(x^{(n)})$ is computed by the Best Linear Unbiased Predictor (BLUP). It means, to find $w_1, \ldots, w_n$ s.t.

$$\hat{Y}(x) := w_0 + w_1 Y(x^{(1)}) + \cdots + w_n Y(x^{(n)})$$

minimizes $\mathrm{MSE} := \mathbb{E}([Y(x) - \hat{Y}(x)]^2)$ under $\mathbb{E}(\hat{Y}(x)) = \mathbb{E}(Y(x))$.

From geostatistics to GP regression models

Geostatiscal approach for spatial interpolation, Simple Kriging

Let Y be a centered stochastic process (or with known mean).
In geostatistics, the prediction of $Y(x)$ knowing $Y(x^{(1)}), \ldots, Y(x^{(n)})$ is computed by the Best Linear Unbiased Predictor (BLUP). It means, to find $w_1, \ldots, w_n$ s.t.

$$\hat{Y}(x) := w_0 + w_1 Y(x^{(1)}) + \cdots + w_n Y(x^{(n)})$$

minimizes $\mathrm{MSE} := \mathbb{E}([Y(x) - \hat{Y}(x)]^2)$ under $\mathbb{E}(\hat{Y}(x)) = \mathbb{E}(Y(x))$.

Link between Simple Kriging and Gaussian process interpolation

- If Y is Gaussian, the conditional expectation coincides with the orthogonal projection onto a linear space
  $\rightarrow$ BLUP = conditional expectation and min(MSE) = conditional variance
- If Y is not Gaussian, the two approaches are different in general
  $\rightarrow$ Advantage of BLUP : closed-form expressions

Gaussian processes, splines and RKHS

The 3 faces of a kernel

$GP(0, k(x, x')) \quad \Leftrightarrow \quad$ p.s.d. functions $k \quad \Leftrightarrow \quad$ RKHS : $\mathcal{H} = \overline{\text{span}\{k(., x), x \in D\}}$

where $\mathcal{H}$ is a "Reproducing Kernel" Hilbert Space with dot product :

$$\langle k(x, .), k(x', .) \rangle = k(x, x') \qquad (*)$$

RKHS can be also defined as Hilbert spaces of functions such that evaluations $f \to f(x)$ are continuous : By Riesz theorem, there exists a unique $k(., x)$ s.t.

$$f(x) = \langle f, k(., x) \rangle$$

Choosing $f = k(., x')$ gives the reproducing identity (*).

Ref : [Aronszajn, 1950], [Berlinet and Thomas-Agnan, 2011].

Gaussian processes, splines and RKHS

Correspondence between **interpolation** spline and GP conditional mean

[Kimeldorf and Wahba, 1971]

The interpolation spline is defined by the functional problem

$$(*) \qquad \min_{h \in \mathcal{H}} \|h\| \qquad s.t. \quad h(x^{(i)}) = y_i,\ i = 1, \ldots, n$$

If $\mathcal{H}$ is the RKHS of kernel k, and if $k(X, X) = (k(x^{(i)}, x^{(j)}))_{1 \leq i,j \leq n}$ is invertible, (*) has a unique solution in the finite dimensional space spanned by the $k(., x^{(i)})$:

$$\begin{aligned} h_{\text{opt}}(x) &= \mathbb{E}\left[Y(x) \,\middle|\, Y(x^{(i)}) = y_i,\ i = 1, \ldots, n \right] \\ &= k(X, x)^\top k(X, X)^{-1} y \end{aligned}$$

where $Y \sim GP(0, k)$, $k(X, x) = (k(x, x^{(i)})_{1 \leq i \leq n}$ and $y = (y_i)_{1 \leq i \leq n}$.

$\rightarrow$ In this sense, GPs are generalizing interpolation splines.

The first part (reduction to finite dimension) is know as Representer theorem.

Gaussian processes, splines and RKHS

Correspondence between approximation spline and GP conditional mean for noisy observations [Kimeldorf and Wahba, 1971]

The approximation spline is defined by the regression problem with a ridge penalty

$$(*) \qquad \min_{h \in \mathcal{H}} \quad \sum_{i=1}^{n} (h(x^{(i)}) - y_i)^2 + \lambda \|h\|^2$$

If $\mathcal{H}$ is the RKHS of kernel k, and if $k(X,X) + \lambda I_n$ is invertible, then (*) has a unique solution in the finite dimensional space spanned by the $k(.,x^{(i)})$:

$$\begin{aligned} h_{\text{opt}}(x) &= \mathbb{E}\left[Y(x) \,\middle|\, Y(x^{(i)} + \varepsilon_i = y_i,\ i = 1, \ldots, n \right] \\ &= k(X,x)^\top (k(X,X) + \lambda I_n)^{-1} y \end{aligned}$$

where $\epsilon_1, \ldots, \epsilon_n$ are independent $\mathcal{N}(0, \lambda)$, and indep. of the GP Y.

Outline

Some valid operations on kernels

A lot of flexibility can be obtained with kernels !

Building a kernel from other ones (basic examples)

Sum, tensor sum	$k_1 + k_2,\ k_1 \oplus k_2$
Product, tensor product	$k_1 \times k_2,\ k_1 \otimes k_2$
ANOVA	$(1 + k_1) \otimes (1 + k_2)$
Warping	$k(x, x') = k_1(f(x), f(x'))$
...	...

See examples in [Rasmussen and Williams, 2006]

Kernels of stationary processes

A centered GP is (second order) stationary iff $k(x, x')$ depend on $x - x'$.
$\rightarrow$ we denote $k(h) = k(x, x + h)$ (abuse of notation)

Bochner's theorem (see e.g. [Rasmussen and Williams, 2006])

The kernel of a real-valued stationary process on $\mathbb{R}^d$ is the Fourier transform of a probability distribution

$$k(h) = \int_{\mathbb{R}^d} \cos(2\pi \langle h, t \rangle) d\mu(t) \tag{1}$$

where $\langle ., . \rangle$ is the usual scalar product on $\mathbb{R}^d$.
The probability measure μ is called spectral measure.

Exercise. Using the definition of positive semidefinite functions, prove that k defined by (1) is a valid kernel.

Kernels of stationary processes

Kernel name	Kernel form	Spectral measure
cosine	$\cos(2\pi h)$	Dirac δ_1
sinc	$\frac{\sin(\pi h)}{\pi h}$	Uniform
Squared exponential	$k(h) = \exp\left(-\frac{1}{2}\frac{h^2}{\ell^2}\right)$	Gaussian
Exponential	$\exp\left(-\frac{\lvert h\rvert}{\ell}\right)$	Student $t_{1/2}$
Matérn 3/2	$\left(1+\sqrt{3}\frac{\lvert h\rvert}{\ell}\right)\exp\left(-\sqrt{3}\frac{\lvert h\rvert}{\ell}\right)$	Student $t_{3/2}$
Matérn 5/2	$\left(1+\sqrt{5}\frac{\lvert h\rvert}{\ell}+\frac{5}{3}\frac{h^2}{\ell^2}\right)\exp\left(-\sqrt{5}\frac{\lvert h\rvert}{\ell}\right)$	Student $t_{5/2}$

Table – Examples of kernels of 1-dimensional stationary processes

Remark : Characteristic length

The parameter ℓ in the previous slide is called "range" in geostatistics or "characteristic length" in machine learning. It is a scale parameter for x.

More precisely, if Y_ℓ is a centered GP with kernel of the forme

$$k_\ell(x, x') = k_1(x/\ell, x'/\ell)$$

then for all x, x', we have

$$\mathrm{Cov}(Y_\ell(x), Y_\ell(x')) = \mathrm{Cov}\left(Y_1(x/\ell), Y_1(x'/\ell)\right)$$

meaning that two Gaussian processes $(Y_\ell(x))$ and $(Y_1(x/\ell))$ have the same finite dimensional distributions.

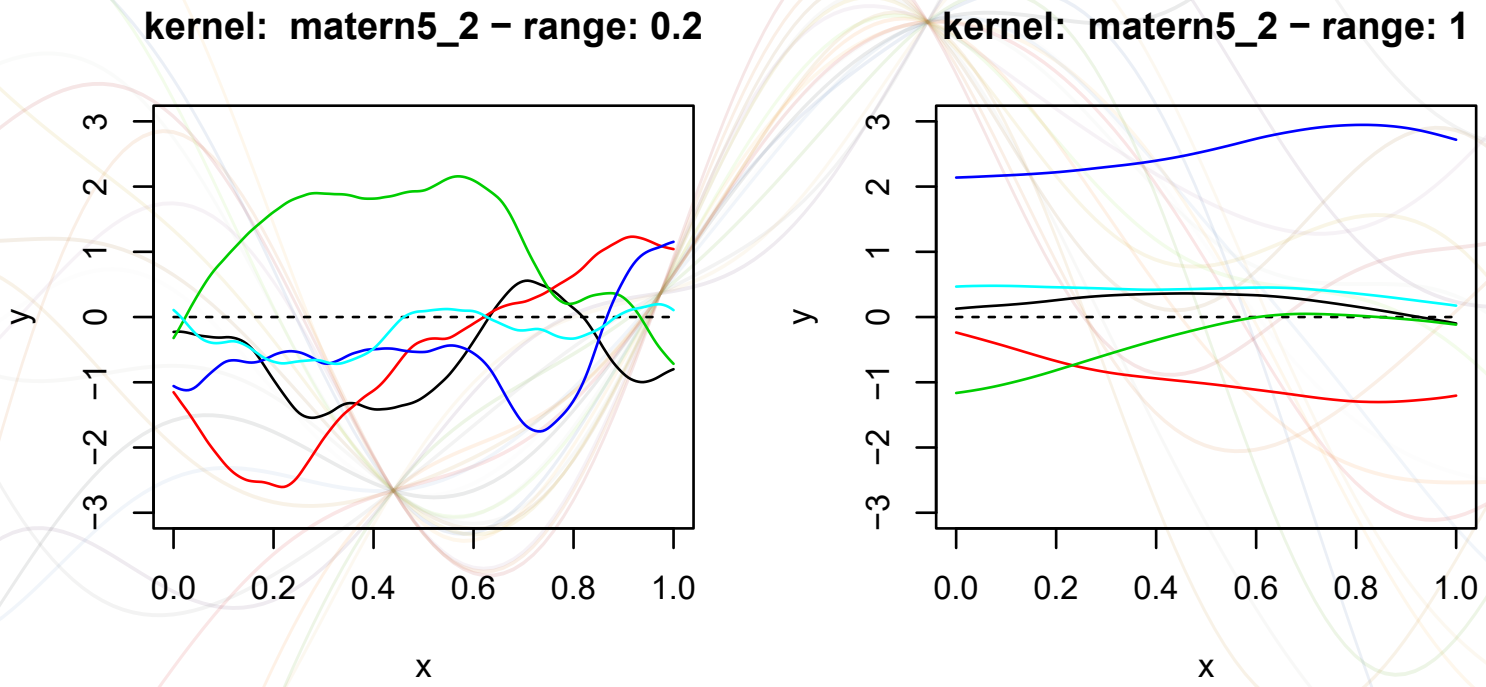


Figure – Sample paths of a GP with a Matérn 5/2 kernel for $\ell = 0.2$ and $\ell = 1$. Zooming 5 times on the first graph is equivalent to the second, in distribution.

Kernels for discrete data

- A kernel for a discrete input with L levels is a positive semi-definite matrix (p.s.d.) of size L

- The general p.s.d. matrix has $L(L+1)/2$ parameters. More parcimonious parameterizations are useful, such as exchangeable covariance matrices

$$\begin{pmatrix} v & c & \dots & c \\ c & v & \ddots & \vdots \\ \vdots & \ddots & \ddots & c \\ c & \dots & c & v \end{pmatrix}, \qquad \text{with} \quad -\frac{1}{L-1} \leq \frac{c}{v} \leq 1.$$

- If the variable is ordinal, with "1" $< \dots <$ "L", one can use a continuous stationary kernel, up to an increasing real-valued function

$$k(\text{“i”}, \text{“j”}) = k_{\text{cont}}(|f(\text{“i”}) - f(\text{“j”})|)$$

$\rightarrow h = f(\text{“i”}) - f(\text{“j”})$ represents the distance between levels "i" and "j"

Gaussian processes and linear operations

Proposition (GP and linearity)

If $Y \sim GP(0, k(s,t))$ and L is linear (acting on the sample paths of Y), then

$$LY \sim GP(0, L_s L_t k(s,t))$$

The notation L_s (resp. L_t) means that we apply L on $s \mapsto f(s,t)$ (resp. $t \mapsto k(s,t)$) when f is a function of two inputs s, t.

Formal proof

- Gaussian : Since L is linear, a linear combination from LY can be rewritten as a linear combination from Y.
- Using the bilinearity of covariance,

$$Cov(LY(s), LY(t)) = L_t \text{Cov}(LY(s), Y(t)) = L_s L_t \text{Cov}(Y(s), Y(t))$$

Playing with kernels

Example (A kernel for even functions).

$$Lf(x) = f(x) + f(-x), \qquad LY(x) = Y(x) + Y(-x)$$

- Why LY is a GP ?
  - For any $x_1, \ldots, x_n$, the linear combination of $LY(x_1), \ldots, LY(x_n)$ is a linear combination of Y values at $x_1, -x_1, \ldots x_n, -x_n$.
  - Since Y is a GP, this linear combination is Normal. Hence LY is a GP.

- Compute the kernel of LY, and observe that $x \mapsto k(x, x')$ is even for all x' :

$$\begin{aligned} \mathrm{Cov}(LY(s), LY(t)) &= L_s L_t k(s,t) = L_s(k(s,t) + k(s,-t)) \\ &= [k(s,t) + k(s,-t)] + [k(-s,t) + k(-s,-t)] \end{aligned}$$

- Conversely, if $x \mapsto k(x, x')$ is even for all x', then LY has even sample paths
  - Check that $\mathrm{var}(Y(x) - Y(-x)) = 0$, which implies that $Y(x) = Y(-x)$ a.s.

$\rightarrow$ This result can be generalized for a large class of linear operators [Ginsbourger et al., 2016].

Playing with kernels

Example (Derivatives, integrals).
Assume that Y is a centered GP with kernel k. Then, under technical conditions :

- The derivative process $(Y'(x))_x$ is a centered GP with kernel

$$k_{Y'}(s,t) = \frac{\partial^2 k}{\partial s \partial t}(s,t)$$

  Moreover, we have $\mathrm{Cov}(Y(s), Y'(t)) = \frac{\partial k}{\partial t}(s,t)$.

- The integral $\int Y(x)dx$ is a centered random variable with variance

$$\iint k(s,t)dsdt$$

  and we have for instance $\mathrm{Cov}(Y(s), \int_t Y(t)dt) = \int k(s,t)dt$.

Outline

A trended GP model

A common form of GP model is to use a linear trend

$$Y(x) = m(x) + Z(x)$$

with :

- $m(x) = \beta_1 f_1(x) + \cdots + \beta_p f_p(x)$ a linear trend (the f_i's are known functions)
- Z a centered GP with kernel $k(x, y; \Theta)$.

Here β and Θ are vectors of unknown parameters.

Model inference

Parameter estimation can be done with two classes of methods :

- Maximum likelihood. The likelihood is the pdf value at $y = (y_1; \ldots; y_n)$ of $(Y(x^{(1)}); \ldots; Y(x^{(n)})) \sim \mathcal{N}(F\beta; k(X, X; \Theta))$:

$$L(\beta, \Theta) = \frac{1}{(2\pi)^{n/2}|k(X, X; \Theta)|^{1/2}} \exp\left(-\frac{1}{2}(y - F\beta)^\top k(X, X; \Theta)^{-1}(y - F\beta)\right)$$

  where F is the $n \times p$ matrix whose row i contains $f_1(x^{(i)}), \ldots, f_p(x^{(i)})$.

- Cross validation. For instance, leave-one-out criterion

$$LOO(\beta, \Theta) = \sum_{i=1}^{n} (\hat{y}_{-i}(x^{(i)}) - y_i)^2$$

  Notice that update formula express $k(X_{-i}, X_{-i}; \Theta)$ with known expressions.

In both cases, there are no closed-form expression, the criterions are not convex and may have several local optima $\rightarrow$ optimization is done numerically.

Bayesian inference, Ordinary and Universal Kriging

- Uncertainty measured by GP variance (Simple Kriging formula) assumes that the parameters are known.
- It can be adapted to the case of unknown parameters in a Bayesian framework, i.e. assuming that the parameters themselves are random
- Prediction is obtained by integrating out the parameter distributions :

$$\begin{aligned} & \mathbb{E}\left(g(Y(x))|Y(x^{(1)}),\ldots,Y(x^{(n)})\right) \\ = & \int \mathbb{E}\left(g(Y(x))|Y(x^{(1)}),\ldots,Y(x^{(n)}),\beta,\Theta\right) f_{\beta,\Theta}(\beta,\theta)\, d\beta d\Theta \end{aligned}$$

(Use $g = Id$ for the Kriging mean, $g(y) = y^2$ for the Kriging variance)

Bayesian inference, Ordinary and Universal Kriging

Unfortunately, Bayesian inference does not provide closed-form expressions, except for special cases. The most famous one is reported below :

Universal Kriging (UK) formula, [Cressie, 1992, Helbert et al., 2008]

Assume that Θ is known, but β is unknown. Then,

1. The BLUP has the same form as for Simple Kriging, replacing β by its GLS estimate $\hat{\beta} = (F^\top K^{-1} F)^{-1} F^\top y$, with $K = k(X, X)$.
2. The UK variance is greater than SK variance, with additional term :

$$\begin{aligned} s^2_{\mathrm{UK}}(x) &= s^2_{\mathrm{SK}}(x) \\ &+ (f(x)^\top - k(x, X)K^{-1}F)^\top (F^\top K^{-1} F)^{-1} (f(x)^\top - k(x, X)K^{-1}F) \end{aligned}$$

3. These formula coincide with the Bayesian approach when choosing the improper prior for $\beta \sim \mathcal{N}(\mu, \lambda k(X, X))$, with $\lambda \to \infty$.

Vocabulary : If the trend is a constant, UK is also called Ordinary Kriging (OK)

Model validation

- Aim : Check that $(y_1, \ldots, y_n)$ is drawn from a multivariate normal dist.
- At least, we check graphically that the leave-one-out (LOO) predictions of $Y(x^{(i)})$ are normal (removing $x^{(i)}$ from the learning set)

$$Y(x^{(i)})|\{Y(x^{(j)}) = y_j, \forall j \neq i\} \sim \mathcal{N}\left(m_{k;-i}(x^{(i)}), s^2_{k,-i}(x^{(i)})\right)$$

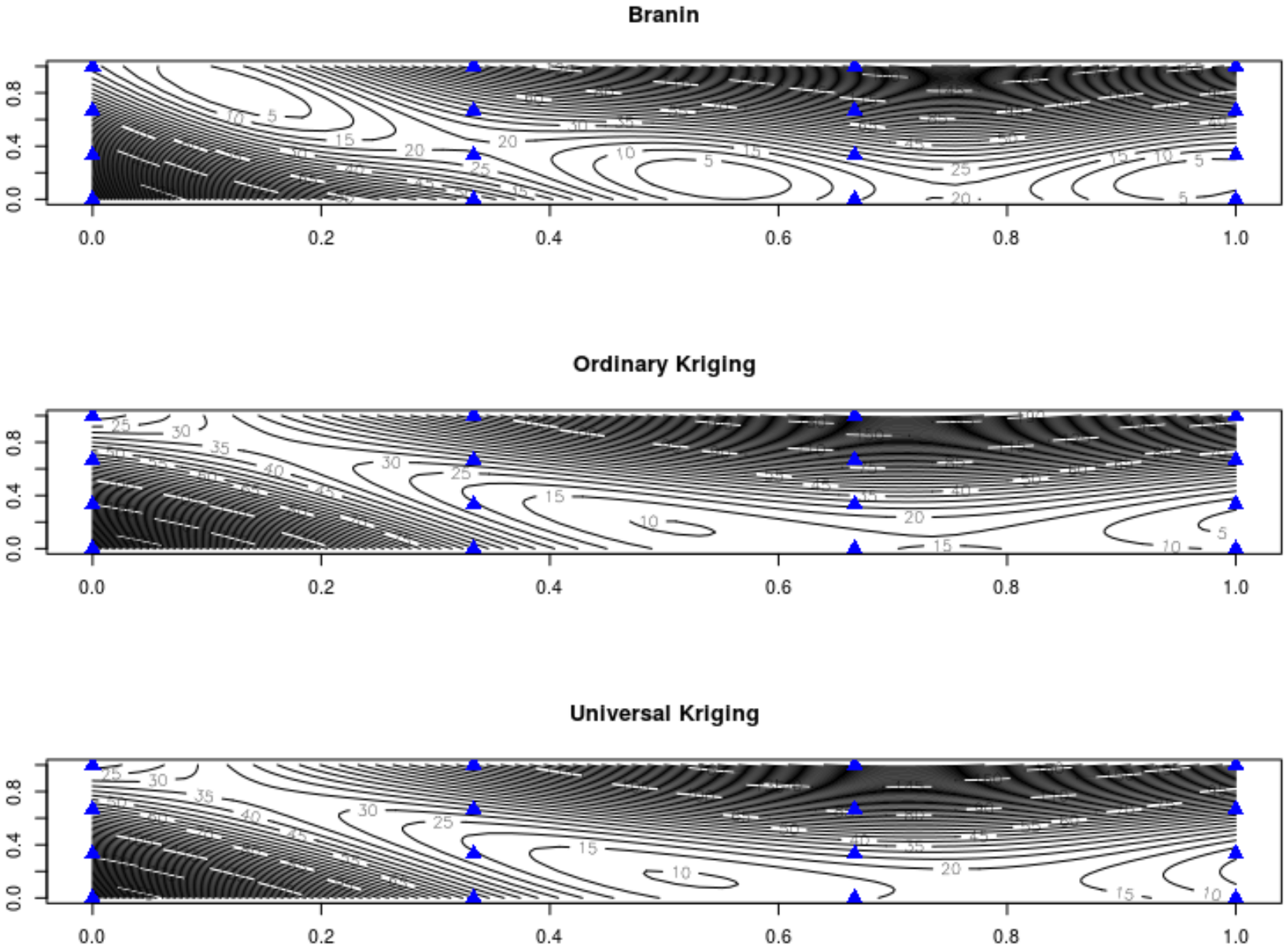
Diagnostic : plot the standardized LOO residuals

$$\frac{y_i - m_{k;-i}(x^{(i)})}{s_{k,-i}(x^{(i)})}$$

Under the GP assumption, they are drawn from a $N(0,1)$ distribution. Notice that they are correlated.

Remark. This is true if all parameters are known.

Illustration, with the DiceKriging R package [Roustant et al., 2012]



```
Call:
km(formula = ~.^2, design = design, response = y, multistart = 100)

Trend  coeff.:
               Estimate
 (Intercept)   195.1009
          x1  -310.9559
          x2  -210.3665
       x1:x2   514.2816

Covar. type   : matern5_2
Covar. coeff.:
               Estimate
   theta(x1)     0.2141
   theta(x2)     0.4025

Variance estimate: 1197.468
```

Figure – Details of parameter estimation for the UK Kriging model

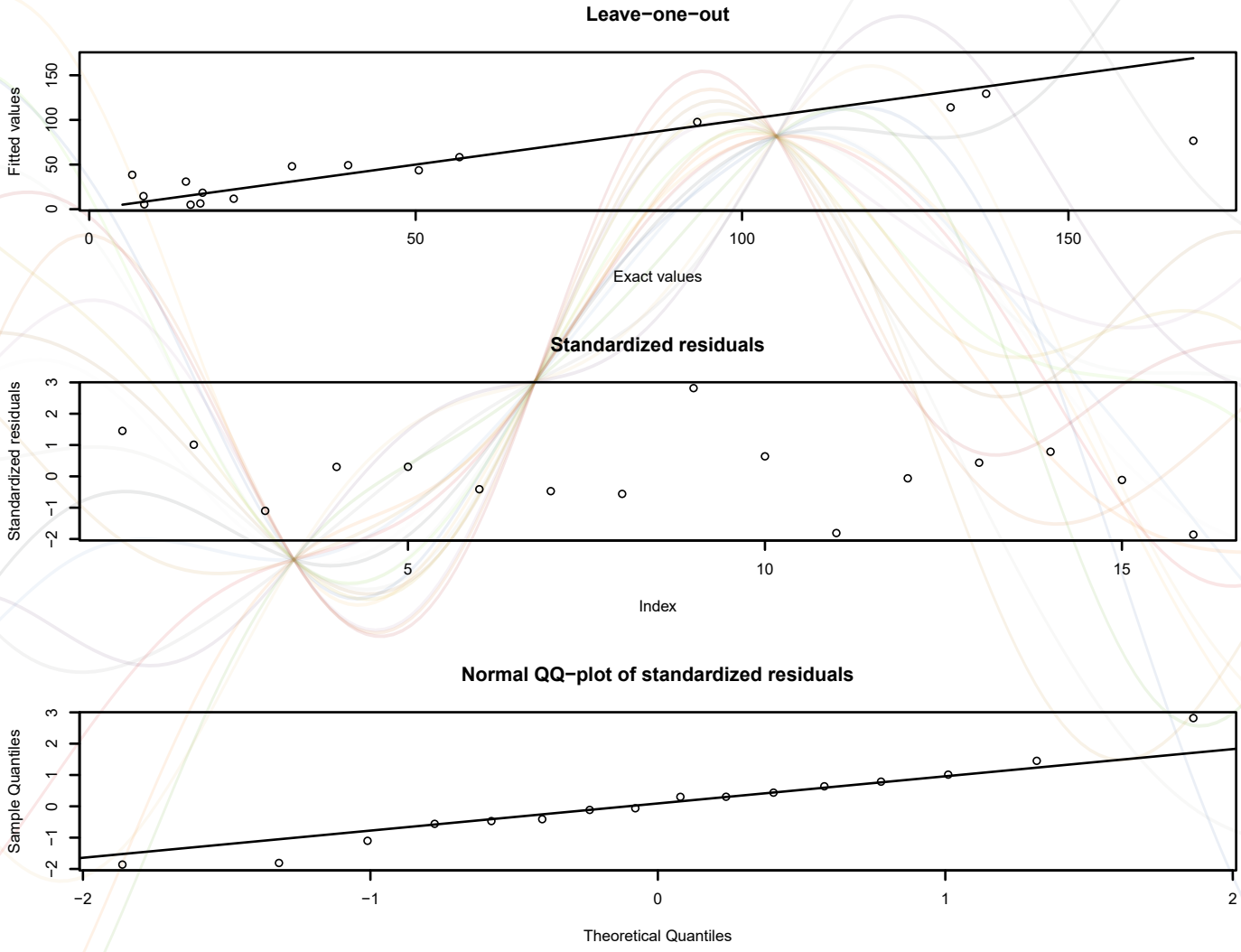


Figure – Leave-one-out validation for the UK Kriging model

Références I

Aronszajn, N. (1950).

Theory of reproducing kernels.

Transactions of the American mathematical society, 68(3) :337–404.

Bect, J., Bachoc, F., and Ginsbourger, D. (2017).

A supermartingale approach to Gaussian process based sequential design of experiments.

working paper or preprint.

Berlinet, A. and Thomas-Agnan, C. (2011).

Reproducing kernel Hilbert spaces in probability and statistics.

Springer Science & Business Media.

Chevalier, C., Bect, J., Ginsbourger, D., Vazquez, E., Picheny, V., and Richet, Y. (2014).

Fast parallel kriging-based stepwise uncertainty reduction with application to the identification of an excursion set.

Technometrics, 56(4) :455–465.

Cressie, N. (1992).

Statistics for spatial data.

Terra Nova, 4(5) :613–617.

Références II

Ginsbourger, D., Roustant, O., and Durrande, N. (2016).

On degeneracy and invariances of random fields paths with applications in Gaussian process modelling.

Journal of Statistical Planning and Inference, 170 :117 – 128.

Helbert, C., Dupuy, D., and Carraro, L. (2008).

Assessment of uncertainty in computer experiments from universal to Bayesian Kriging.

ASMBI, 25 :99–113.

Jones, D. R., Schonlau, M., and Welch, W. J. (1998).

Efficient global optimization of expensive black-box functions.

Journal of Global Optimization, 13(4) :455–492.

Kimeldorf, G. and Wahba, G. (1971).

Some results on Tchebycheffian spline functions.

Journal of mathematical analysis and applications, 33(1) :82–95.

Krige, D. G. (1951).

A statistical approach to some basic mine valuation problems on the witwatersrand.

Journal of the Chemical, Metallurgical and Mining Society of South Africa, 52(6) :119–139.

Références III

Matheron, G. (1963).

Principles of geostatistics.

Economic Geology, 58 :1246–1266.

Močkus, J. (1975).

On Bayesian methods for seeking the extremum.

In Marchuk, G. I., editor, *Optimization Techniques IFIP Technical Conference Novosibirsk, July 1–7, 1974*, pages 400–404, Berlin, Heidelberg. Springer Berlin Heidelberg.

Picheny, V., Ginsbourger, D., Richet, Y., and Caplin, G. (2013).

Quantile-based optimization of noisy computer experiments with tunable precision.

Technometrics, 55(1) :2–13.

Rasmussen, C. E. and Williams, C. K. (2006).

Gaussian processes for machine learning.

the MIT Press.

Roustant, O., Ginsbourger, D., and Deville, Y. (2012).

Dicekriging, diceoptim : Two r packages for the analysis of computer experiments by kriging-based metamodeling and optimization.

Journal of Statistical Software, Articles, 51(1) :1–55.

Références IV

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989).

Design and analysis of computer experiments.

Statistical Science, 4(4) :409–435.

Vazquez, E. and Bect, J. (2010).

Convergence properties of the expected improvement algorithm with fixed mean and covariance functions.

Journal of Statistical Planning and Inference, 140(11) :3088 – 3095.