An overview of kriging for researchers

Rodolphe Le Riche, Nicolas Durrande

An overview of kriging for researchers

Rodolphe Le Riche[1], Nicolas Durrande[2]

[1] CNRS LIMOS at Mines Saint-Etienne, France
[2] Prowler.io, UK

September 2019
Modeling and Numerical Methods for Uncertainty Quantification (MNMUQ 2019)
French-German Summer School / Ecole Thématique CNRS
Porquerolles

Goal of the class, acknowledgements

- This 1h30 course is an overview of kriging = conditional Gaussian process (GP), GP regression (GPR)
- with openings towards research items.
- Material partly recycled from two previous classes, one given with Nicolas Durrande [Durrande and Le Riche, 2017] and the previous edition of this class [Le Riche, 2014].
- A few new slides on RKHS coming from discussions with Xavier Bay.

Content

Context

- Kriging is most often used in the context of expensive (numerical) experiments (simulators, e.g. PDE solvers):
- The experiment can be seen as a function of the input parameters

$$y = f(x)$$

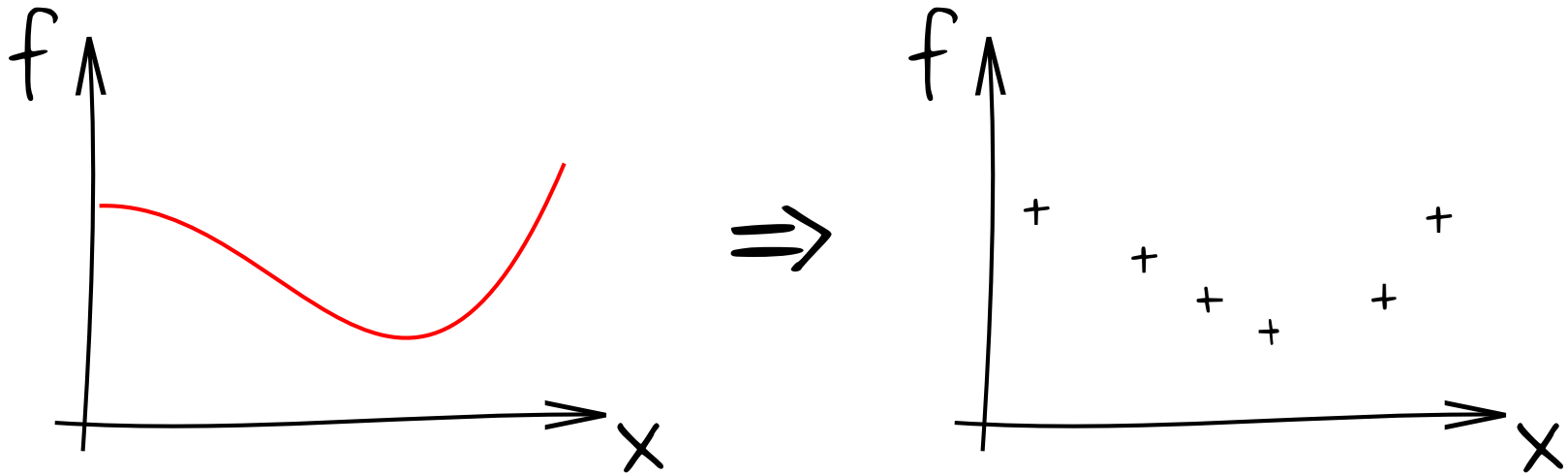
where f is a **costly to evaluate function**.

In the following, we will assume that

- $x \in \mathcal{X}$: There are d input variables. Usually (but not necessarily) $\mathcal{X}$ is $\mathbb{R}^d$.
- $y \in \mathbb{R}$: The output is a scalar. But extensions to GP regression with multiple outputs exist.

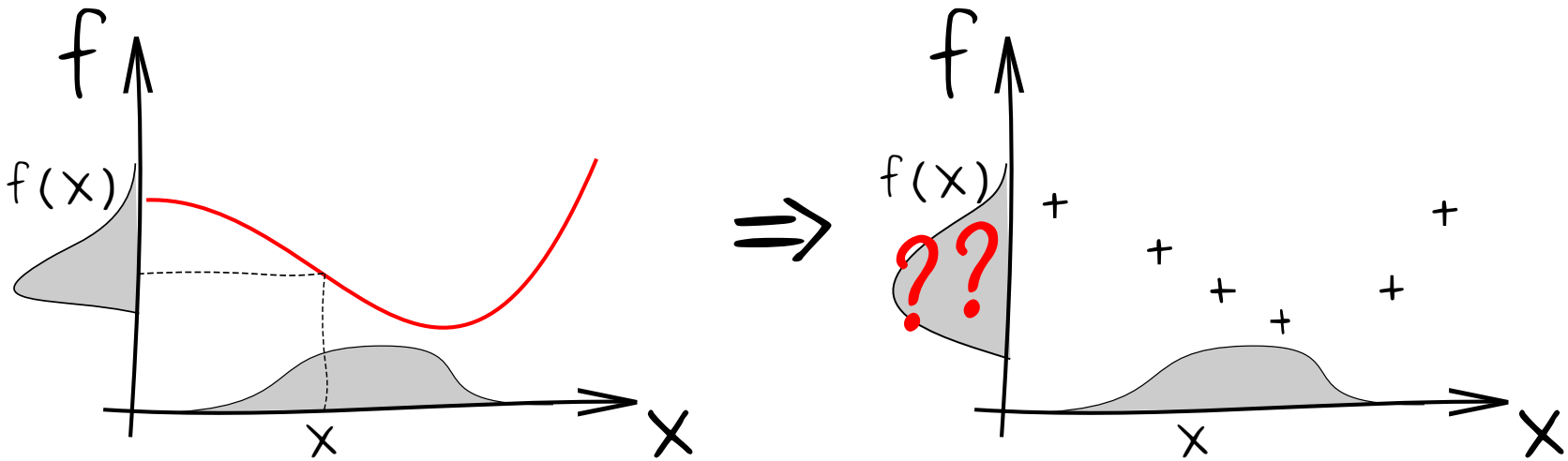
The fact that f is **costly to evaluate** changes a lot of things...

1. Representing the function is not possible...



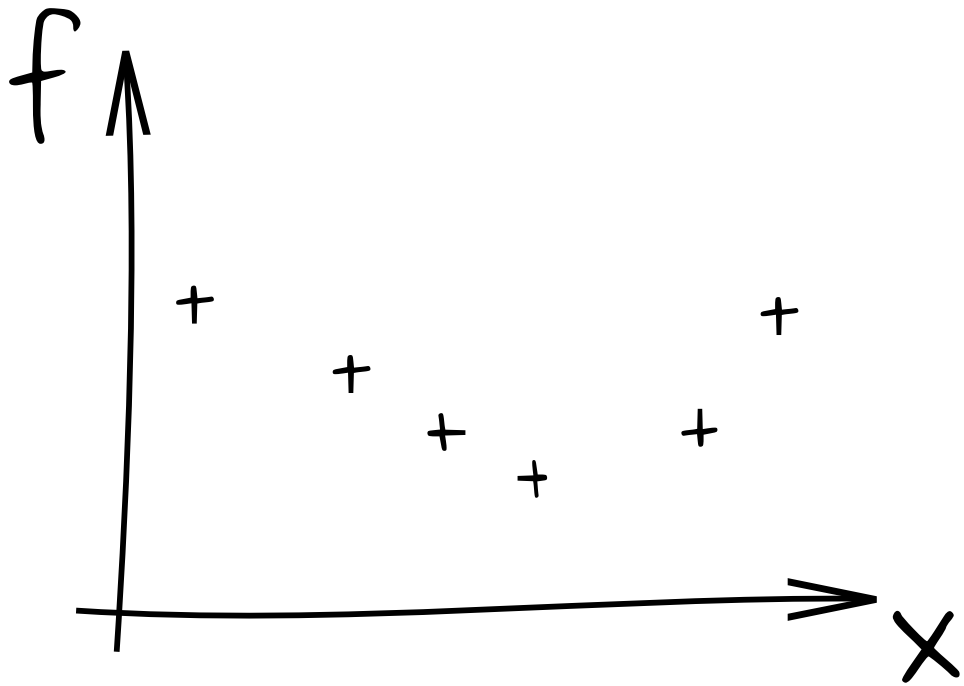
The fact that f is **costly to evaluate** changes a lot of things...

2. Uncertainty propagation is not possible...



The fact that f is **costly to evaluate** changes a lot of things...

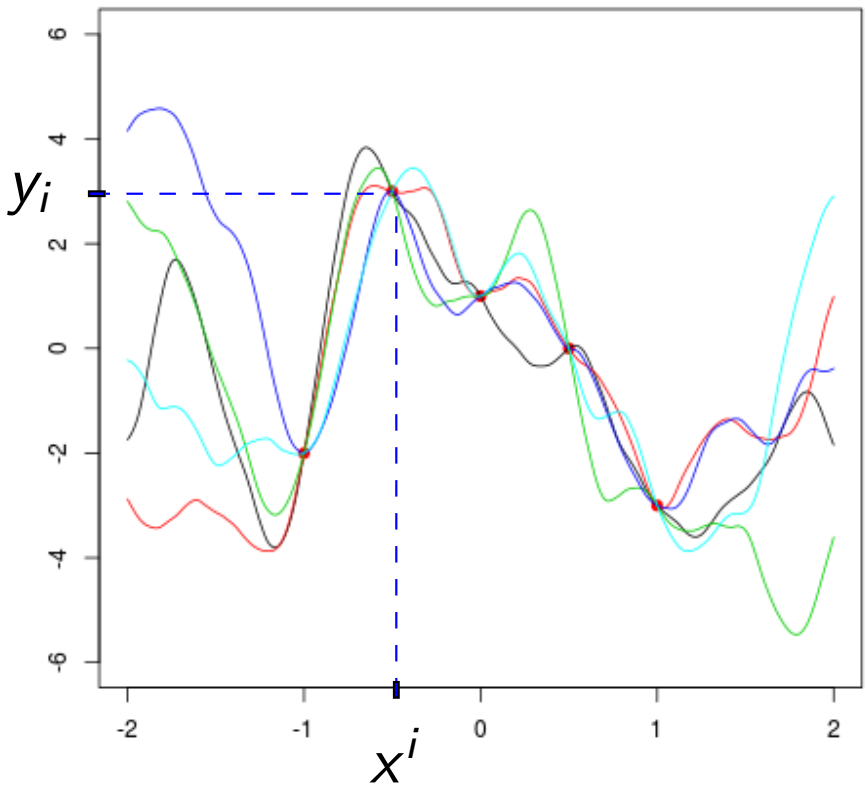
3. Optimisation is also tricky...



4. Computing integrals is not possible...
5. Sensitivity analysis is not possible...

Statistical modelling

We know an initial Design of Experiments (DoE) of n points (x^i, y_i), $y_i = f(x^i)$.
What can be said about possible y at any x using probabilities?
$\Rightarrow$ kriging for regression (conditional GP) [Krige, 1951, Matheron, 1963] = a family of surrogates (metamodels) with embedded uncertainty.



General bibliography: [Rasmussen and Williams, 2006, Durrande and Le Riche, 2017, Le Riche, 2014]

Content

Random Process (1/2)

Random variable Y

random event $\omega \in \Omega$
(e.g., throw a dice)

$\Longrightarrow$

get an instance y. Ex:

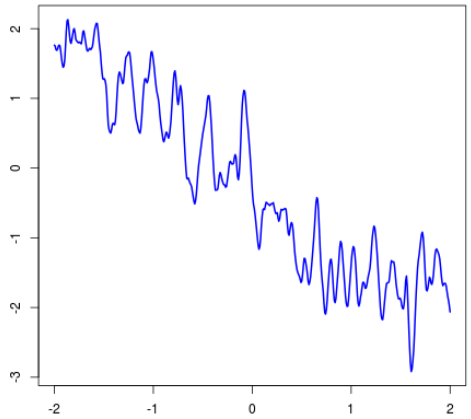
- if dice ≤ 3 , $y = 1$
- if $4 \geq$ dice ≤ 5 , $y = 2$
- if dice $= 6$, $y = 3$

Random process $Y(x)$

A set of RV's indexed by x
random event $\omega \in \Omega$
(e.g., weather)

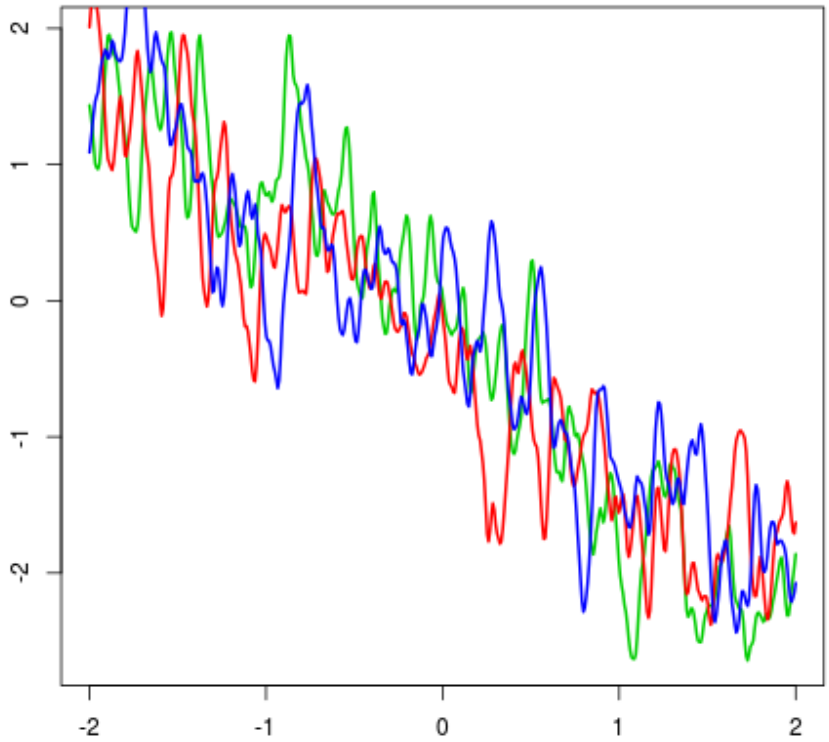
$\Longrightarrow$

get a function $y(x)$. Ex:



Random Process (2/2)

Repeat the random event (say 3 times):



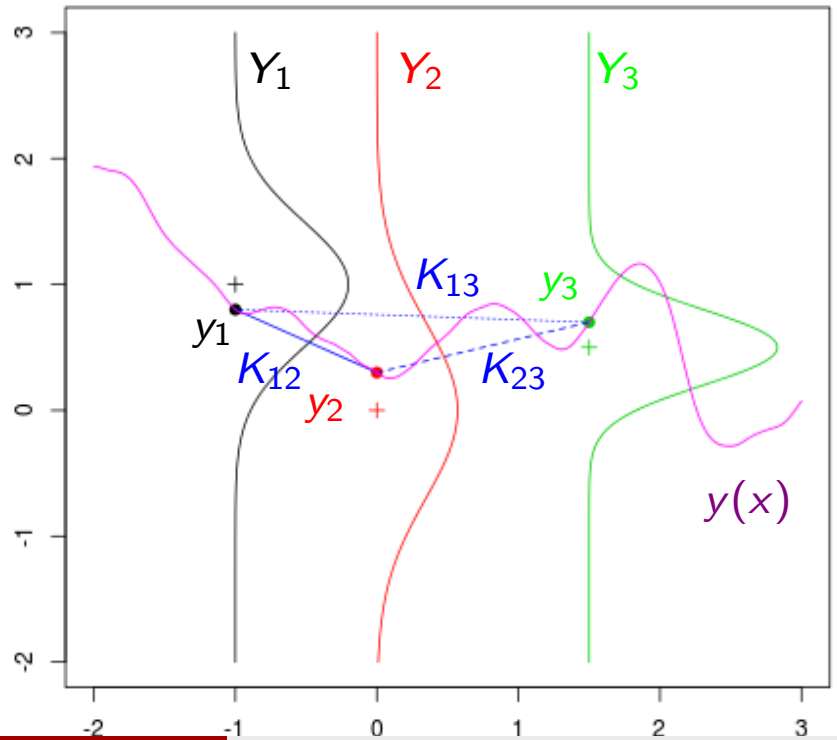
3 $y(x)$'s. They are different, yet bear strong similarities.

Gaussian Process (1/2)

Assume $Y()$ is a GP, $Y(x) \sim \mathcal{N}(\mu(x), k(x,x)) \Leftrightarrow$

$$\forall X = \begin{pmatrix} x^1 \\ \dots \\ x^n \end{pmatrix} \in \mathcal{X}^{n \times d} \quad , \quad Y(X) = \begin{pmatrix} Y(x^1) \\ \dots \\ Y(x^n) \end{pmatrix} \sim \mathcal{N}(\mu(X), K)$$

where $K_{ij} = \text{Cov}(Y(x^i), Y(x^j)) = k(x^i, x^j)$ depends only on the x's



$+$: $\mu(x^i)$

Gaussian Process (2/2)

The distribution of a GP is fully characterised by:

- its mean function $\mu(.)$ defined over $\mathcal{X}$
- its covariance function (or kernel) $k(.,.)$ defined over $\mathcal{X} \times \mathcal{X}$: $k(x, x') = \text{Cov}(Y(x), Y(x'))$

$\Rightarrow$ Example path simulation: Say $k(x, x') = \sigma^2 \exp\left(-(x - x')^2/\theta^2\right)$ and in pseudo-R, build a fine grid `X`, choose mean function `mu()`, build the covariance matrix, `K[i,j]=k(X[i],X[j])`, eigenanalysis, `Keig = eigen(K)`, and sample,

```
y = mu[X] + Keig$vectors %*% diag(sqrt(Keig$values))
                                %*% matrix(rnorm(n))
```

$\Rightarrow$ See also Shiny App:

`https://github.com/NicolasDurrande/shinyApps`

Valid kernels

A kernel satisfies the following properties:

- It is symmetric: $k(x, x') = k(x', x)$
- It is positive semi-definite (psd):

$$\forall n \in \mathbb{N}, \forall x_i \in D, \forall \alpha \in \mathbb{R}^n, \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) \geq 0$$

Furthermore any symmetric psd function can be seen as the covariance of a Gaussian process. This equivalence is known as the Loeve theorem.

Popular kernels in 1D

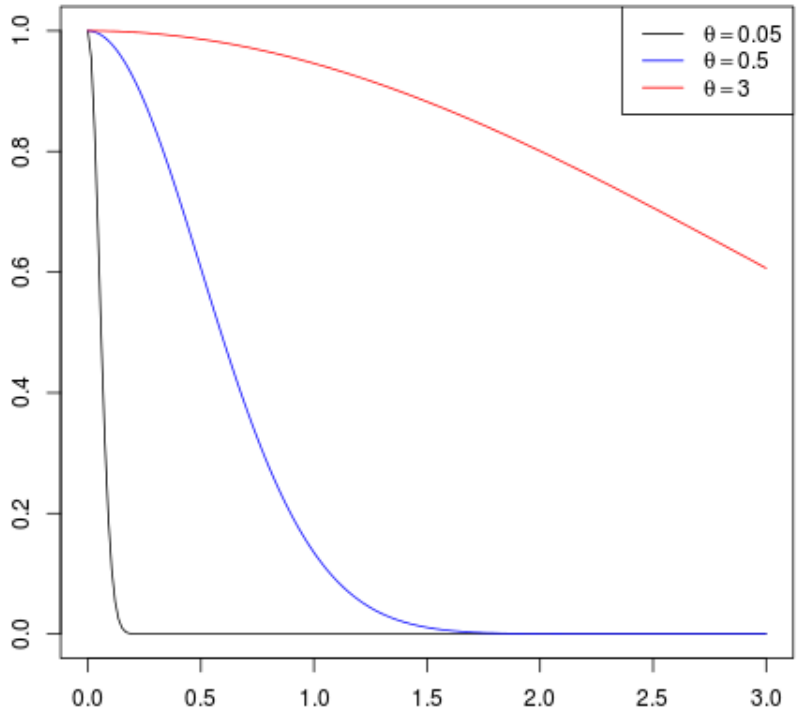
There are a lot of functions that have already been proven psd:

constant	$k(x,x') = \sigma^2$
white noise	$k(x,x') = \sigma^2 \delta_{x,x'}$ (Kronecker delta function)
Brownian	$k(x,x') = \sigma^2 \min(x,x')$
power-exponential	$k(x,x') = \sigma^2 \exp\left(-\lvert x-x'\rvert^p/\theta\right)$, $0 < p \leq 2$
Matérn 3/2	$k(x,x') = \sigma^2 \left(1+\lvert x-x'\rvert\right)\exp\left(-\lvert x-x'\rvert/\theta\right)$
Matérn 5/2	$k(x,x') = \sigma^2 \left(1+\lvert x-x'\rvert/\theta+1/3\lvert x-x'\rvert^2/\theta^2\right)\times \exp\left(-\lvert x-x'\rvert/\theta\right)$
squared exponential	$k(x,x') = \sigma^2 \exp\left(-(x-x')^2/\theta^2\right)$
linear	$k(x,x') = \sigma^2 x x'$
$\vdots$	

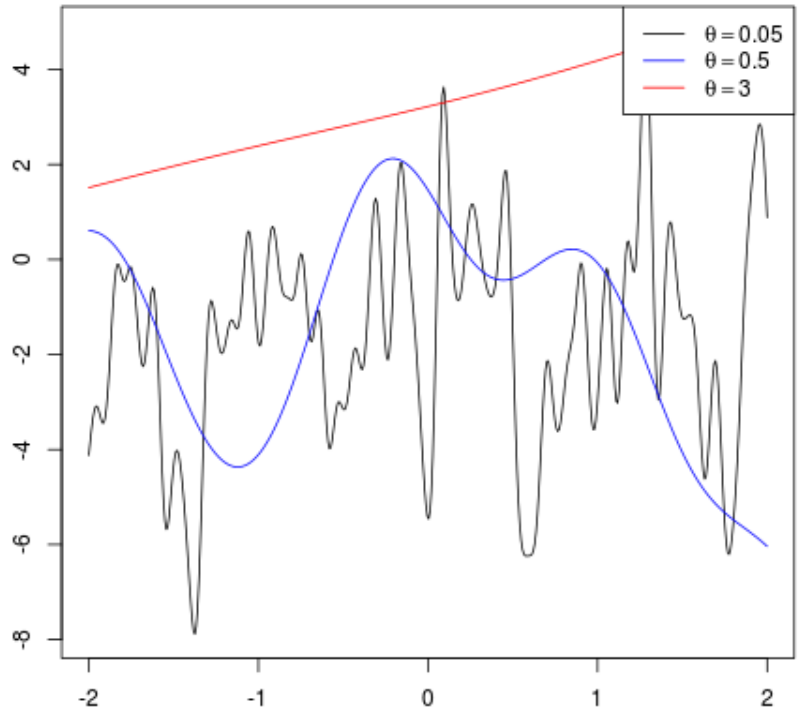
The parameter σ^2 is called the **variance** and θ the **length-scale**.
General factorized form: $k(x,x') = \sigma^2 r(x,x')$, $r(,)$ the correlation function.

Effect of θ, squared exponential kernel

$k(|x - x'|)$

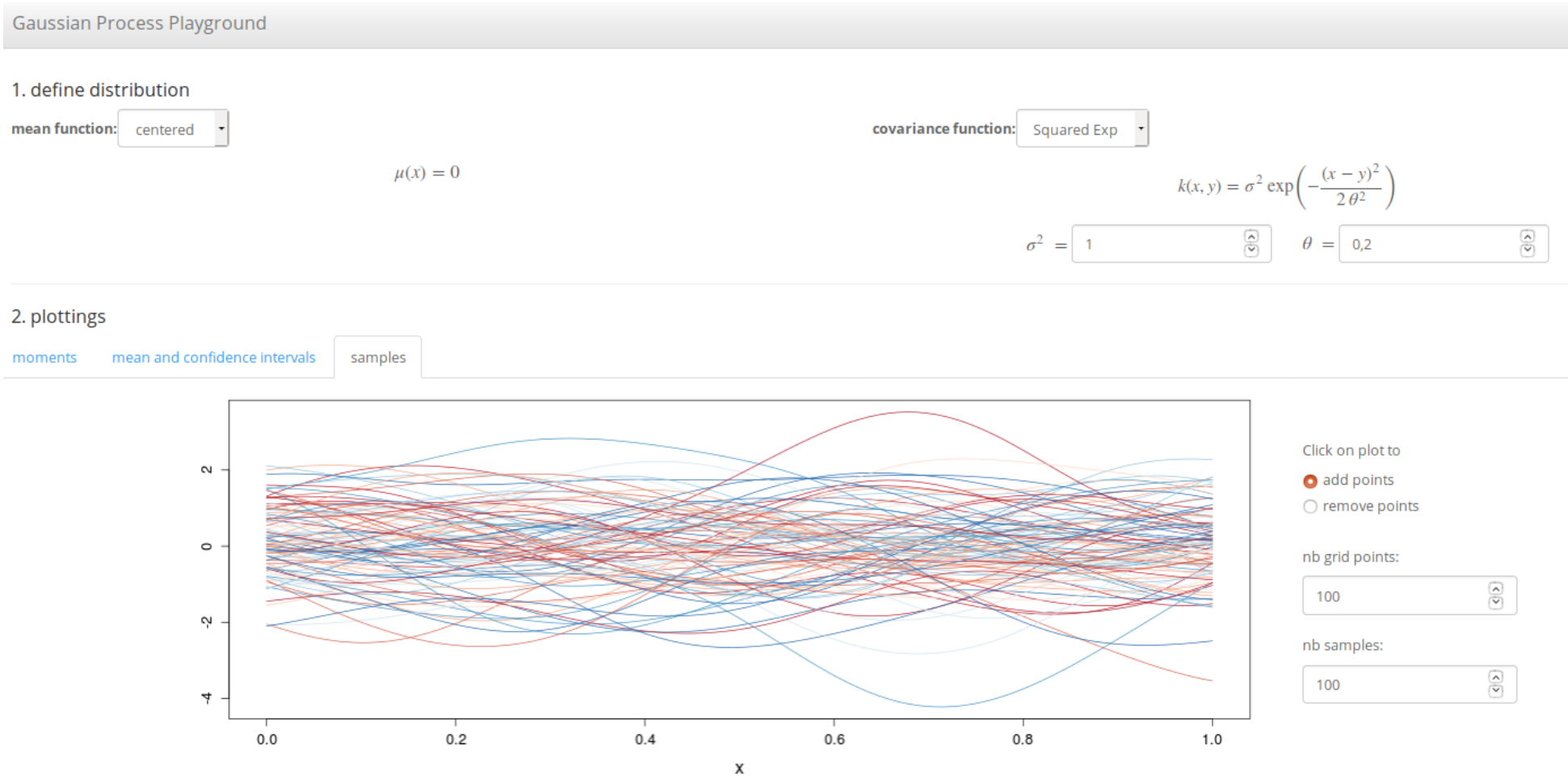


trajectories $y(x)$



Trajectories with squared exponential kernel

and the Shiny App @ `https://github.com/NicolasDurrande/shinyApps`



Trajectories with the Brownian kernel

Gaussian Process Playground

1. define distribution

mean function: centered

$$\mu(x) = 0$$

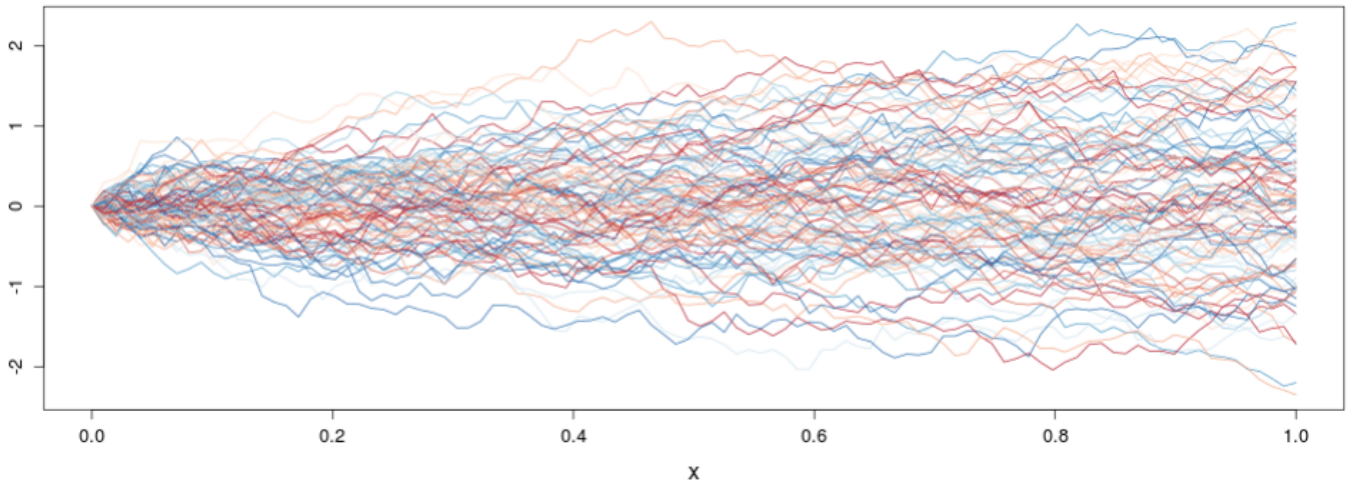
covariance function: Brownian

$$k(x, y) = \sigma^2 \min(x, y)$$

$\sigma^2 =$ 1

2. plottings

moments | mean and confidence intervals | samples



Click on plot to

- add points
- remove points

nb grid points: 100

nb samples: 100

Trajectories with the Matérn 3/2 kernel

Gaussian Process Playground

1. define distribution

mean function: centered

$$\mu(x) = 0$$

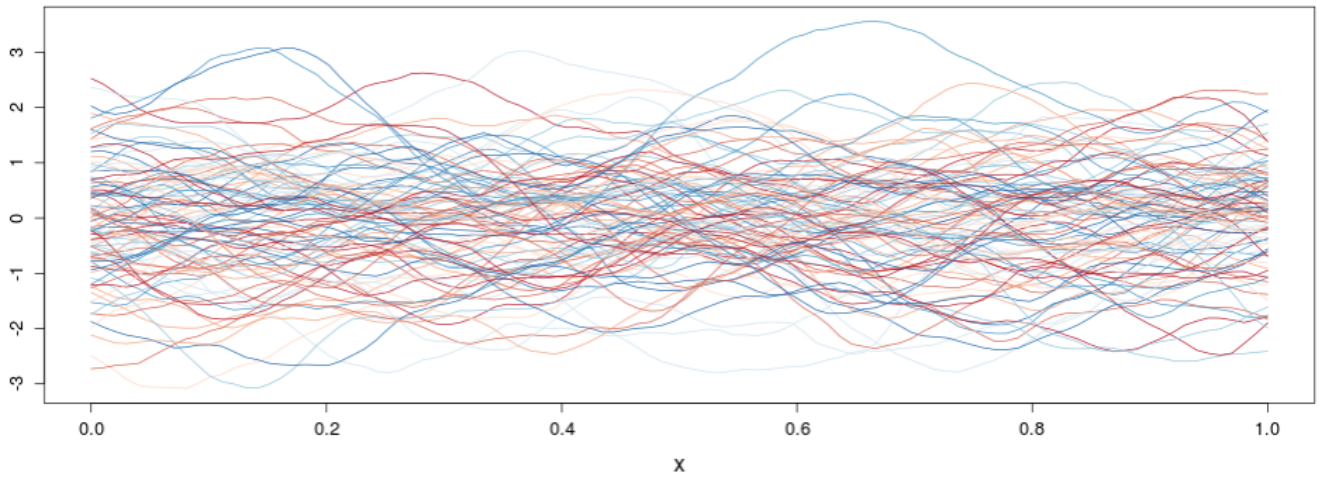
covariance function: Matern 3/2

$$k(x, y) = \sigma^2 \left(1 + \sqrt{3}\frac{|x-y|}{\theta}\right) \exp\left(-\frac{|x-y|}{\theta}\right)$$

$\sigma^2 =$ 1 $\theta =$ 0,2

2. plottings

moments | mean and confidence intervals | samples



Click on plot to
- add points
- remove points

nb grid points: 100

nb samples: 100

Trajectories with the exponential kernel

Gaussian Process Playground

1. define distribution

mean function: centered

$$\mu(x) = 0$$

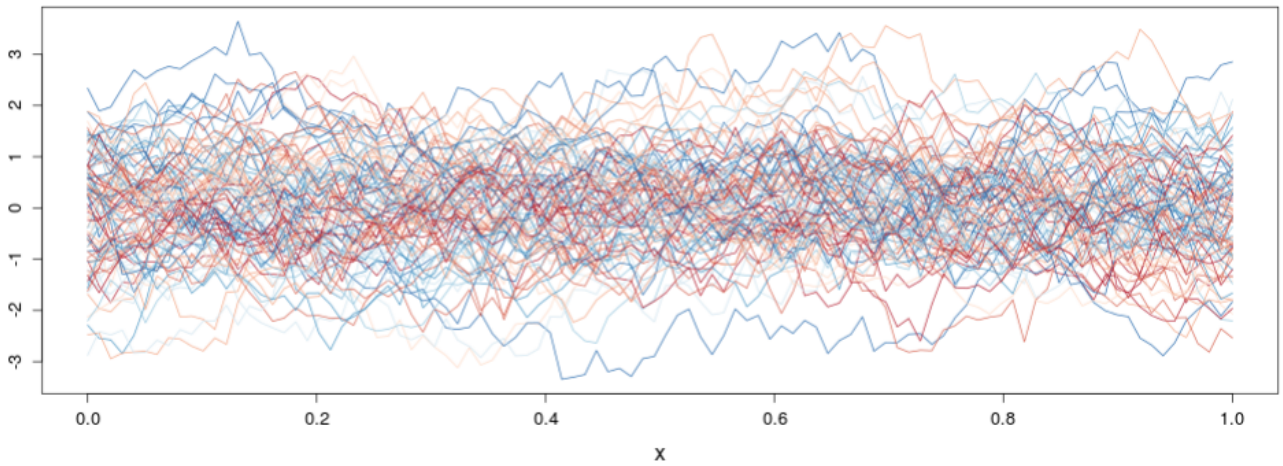
covariance function: Exponential

$$k(x, y) = \sigma^2 \exp\left(-\frac{|x - y|}{\theta}\right)$$

σ^2 = 1 θ = 0,2

2. plottings

moments | mean and confidence intervals | samples



Click on plot to
- add points
- remove points

nb grid points: 100

nb samples: 100

Regularity and covariance function

- The regularity and frequency content of the $y(x)$ are controlled by the kernel (and its length-scale)
- For stationary processes (depend on $\tau = x - x'$ only), the trajectories are p times differentiable (in the mean square sense) if $k(\tau)$ is $2p$ times differentiable at $\tau = 0 \Rightarrow$ the property of $k(\tau)$ at $\tau = 0$ define the regularity of the process.
- Examples:
  - trajectories with squared exponential kernels are infinitely differentiable = very (unrealistically?) smooth.
  - trajectories with Matérn $5/2$ and $3/2$ kernels are twice and once differentiable.
  - trajectories with power-exponential are not differentiable excepted when $p = 2$.

Popular multi-dimensional kernels (1/2)

$$
\begin{aligned}
\text{constant} \quad & k(x,x') = \sigma^2 \\
\text{white noise} \quad & k(x,x') = \sigma^2 \delta_{x,x'} \\
\text{exponential} \quad & k(x,x') = \sigma^2 \exp\left(-||x-x'||_\theta\right) \\
\text{Matérn } 3/2 \quad & k(x,x') = \sigma^2 \left(1+\sqrt{3}||x-x'||_\theta\right) \exp\left(-\sqrt{3}||x-x'||_\theta\right) \\
\text{Matérn } 5/2 \quad & k(x,x') = \sigma^2 \left(1+\sqrt{5}||x-x'||_\theta + \frac{5}{3}||x-x'||_\theta^2\right) \times \\
& \qquad\qquad \exp\left(-\sqrt{5}||x-x'||_\theta\right) \\
\text{sq. exp.} \quad & k(x,x') = \sigma^2 \exp\left(-\frac{1}{2}||x-x'||_\theta^2\right)
\end{aligned}
$$

$$
\text{where } ||x-x'||_\theta = \left(\sum_{i=1}^{d} \frac{(x_i - x_i')^2}{\theta_i^2}\right)^{1/2}.
$$

Popular multi-dimensional kernels (2/2)

A common general recipee: a product of univariate kernels,

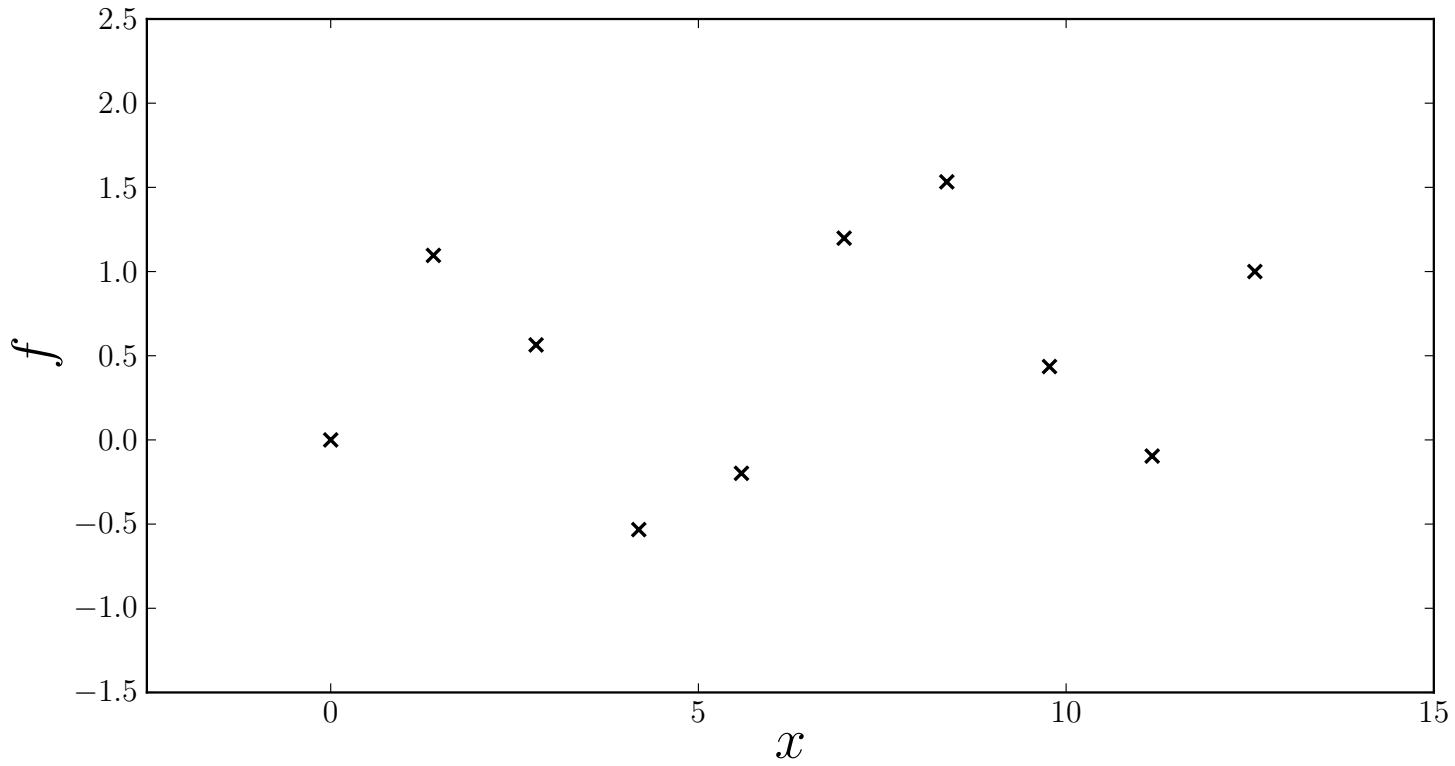
$$k(x, x') = \sigma^2 \prod_{i=1}^{d} r_i(x_i, x_i')$$

which has $d + 1$ parameters.

(more on kernel design later)

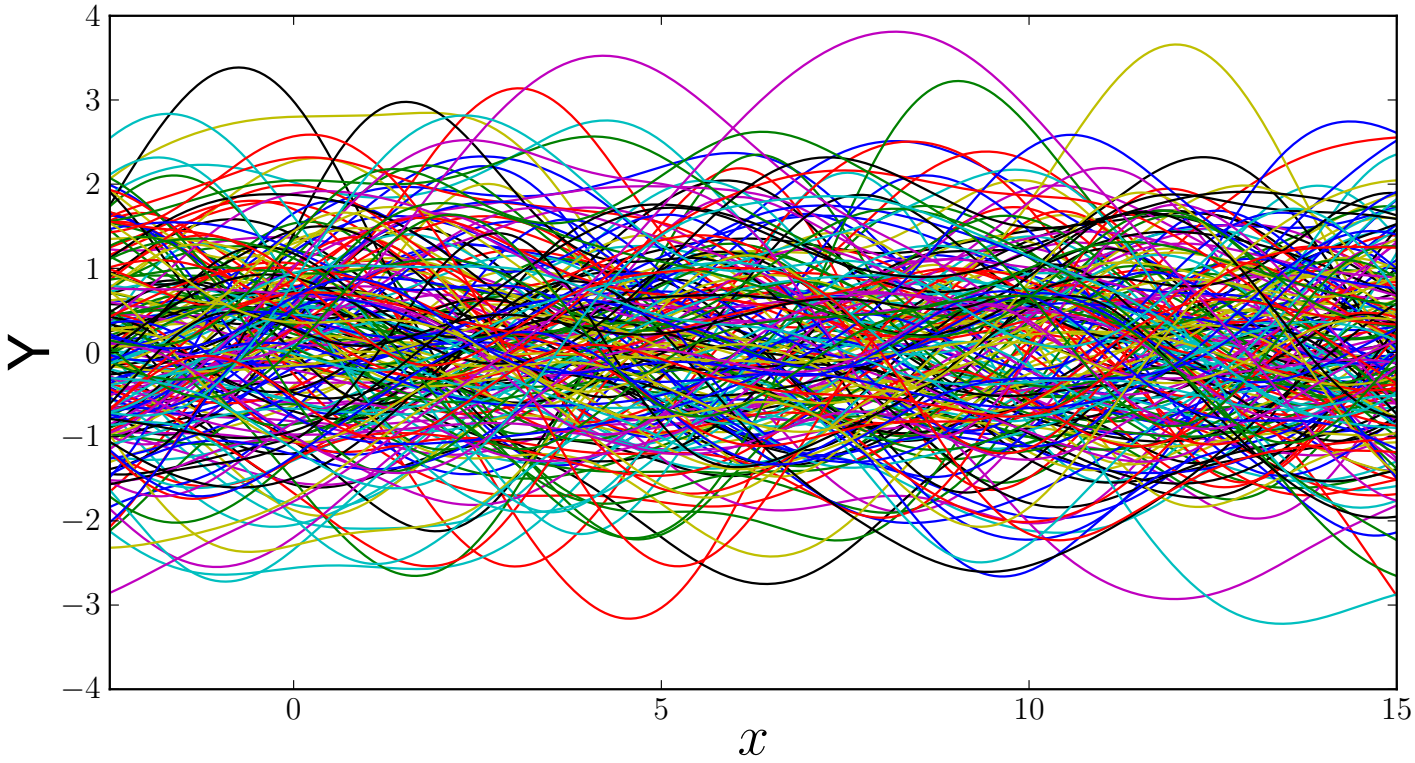
Gaussian process regression

Assume we have observed a function $f()$ over a set of points $X = (x^1, \ldots, x^n)$:



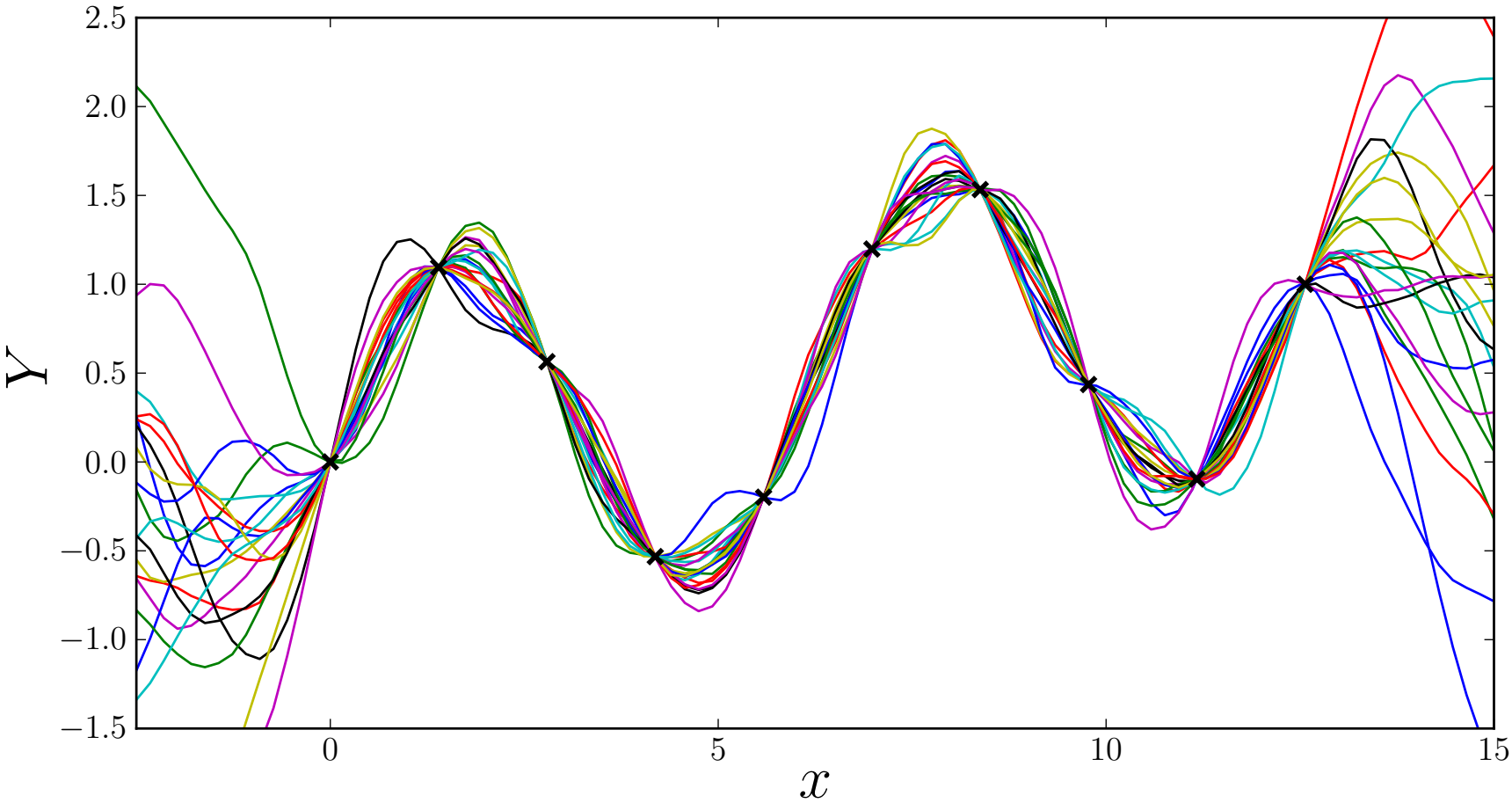
The vector of observations is $F = f(X)$ (ie $F_i = f(x^i)$).

Since $f()$ in unknown, we make the general assumption that it is the sample path of a Gaussian process $Y \sim \mathcal{N}(\mu(), k(,))$:

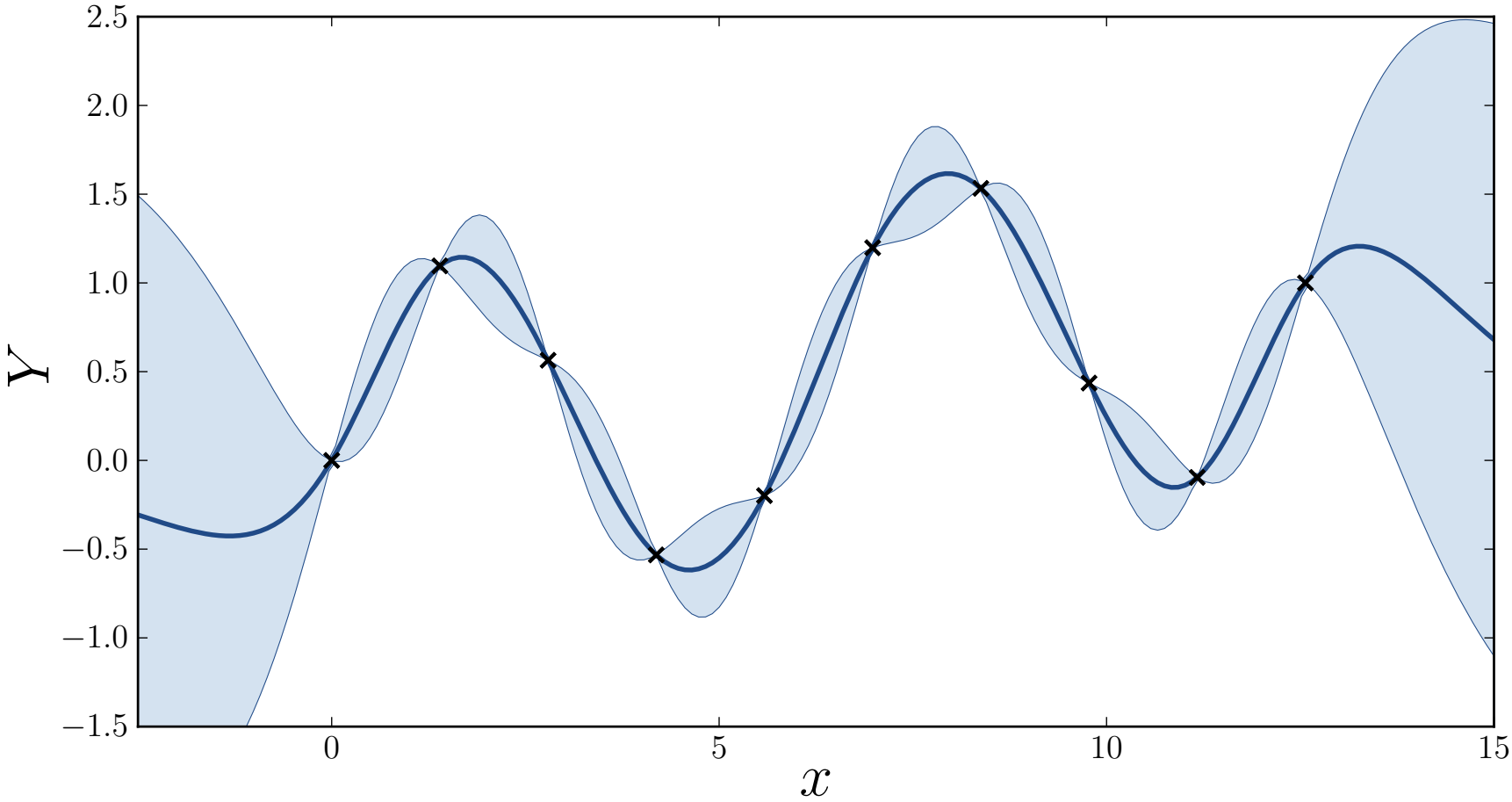


(here $\mu(x) = 0$)

If we remove all the samples that do not interpolate the observations we obtain:



It can summarized by a mean function and 95% confidence intervals.



Kriging equations (1/2)

The conditional distribution can be obtained analytically:

By definition, $(Y(x), Y(X))$ is multivariate normal. Formulas on the conditioning of Gaussian vectors give the distribution of $Y(x)|Y(X) = F$. It is $\mathcal{N}(m(.), c(.,.))$ with :

$$\begin{aligned} m(x) &= \mathbb{E}[Y(x)|Y(X)=F] \\ &= \mu(x) + k(x,X)k(X,X)^{-1}(F - \mu(X)) \\ c(x,x') &= \text{Cov}[Y(x), Y(x')|Y(X)=F] \\ &= k(x,x') - k(x,X)k(X,X)^{-1}k(X,x') \end{aligned}$$

Kriging equations (2/2)

The distribution of $Y(x)|Y(X) = F$ is $\mathcal{N}(m(.), c(.,.))$ with:

$$\begin{aligned} m(x) &= \mathbb{E}[Y(x)|Y(X)=F] \\ &= \mu(x) + k(x,X)k(X,X)^{-1}(F - \mu(X)) \\ c(x,x') &= \mathrm{Cov}[Y(x), Y(x')|Y(X)=F] \\ &= k(x,x') - k(x,X)k(X,X)^{-1}k(X,x') \end{aligned}$$

- $k(X,X) = [k(x^i, x^j)]$: covariance matrix, Gram matrix in SVM.
- $k(x,X) = [k(x,x^1), \ldots, k(x,x^n)]$: covariance vector, only dependance on x beside $\mu(x)$.
- It is a Gaussian distribution: gives confidence intervals, can be sampled, this is actually how the previous slides were generated.
- Bayesian: $Y(x)|Y(X) = F$ is the posterior distribution of $Y(x)$ once $Y(X) = F$ is observed.

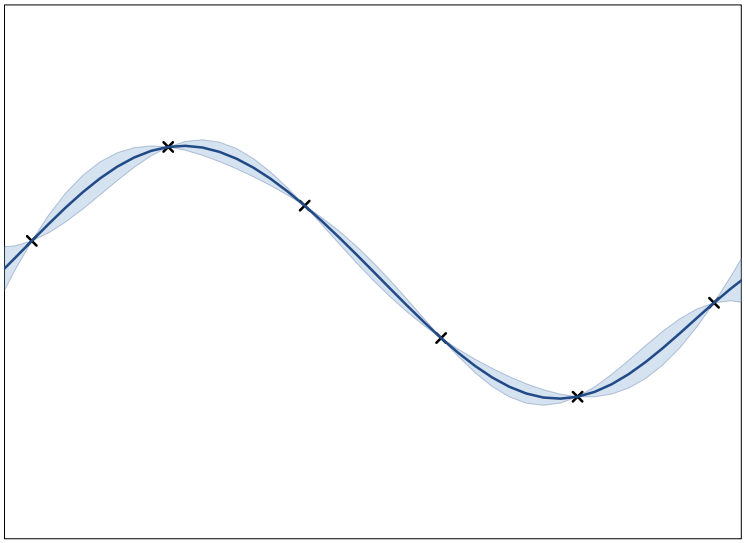
A few remarkable properties of GPR models

- They (can) interpolate the data-points
- The prediction variance does not depend on the observations
- The mean predictor does not depend on the variance parameter
- They (usually) come back to the a priori trend $\mu(x)$ when we are far away from the observations.

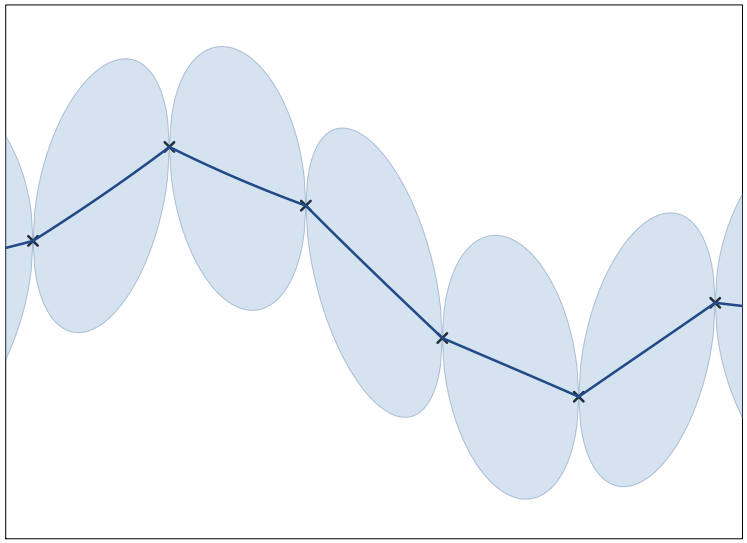
(proofs left as exercise)

Changing the kernel has a huge impact on the model:

Gaussian kernel:

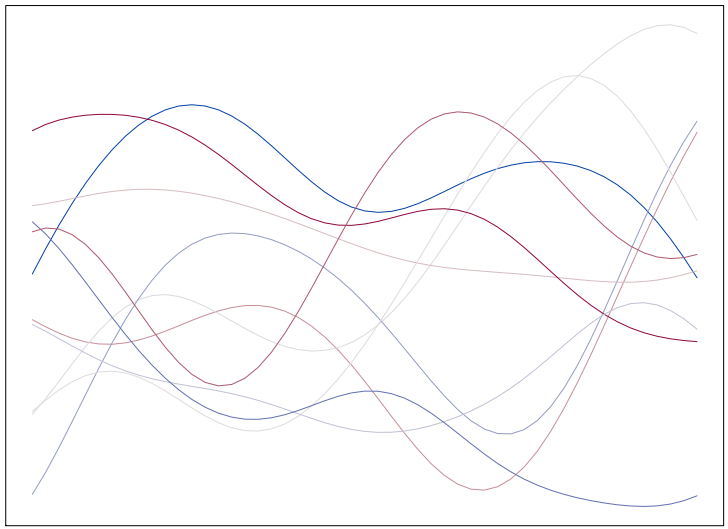


Exponential kernel:

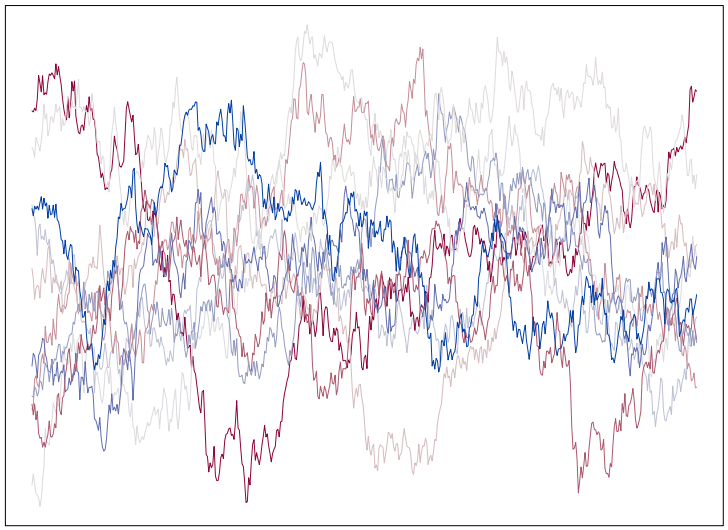


This is because changing the kernel means changing the prior on f

Gaussian kernel:

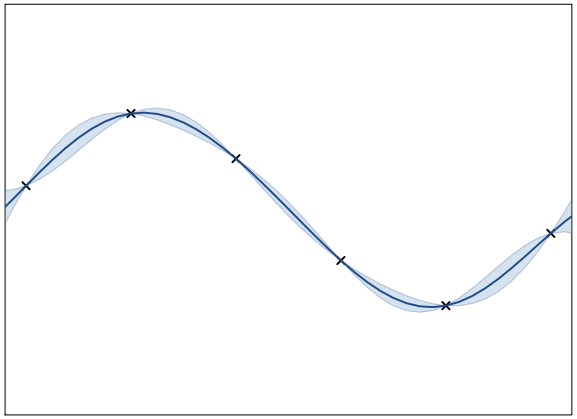


Exponential kernel:

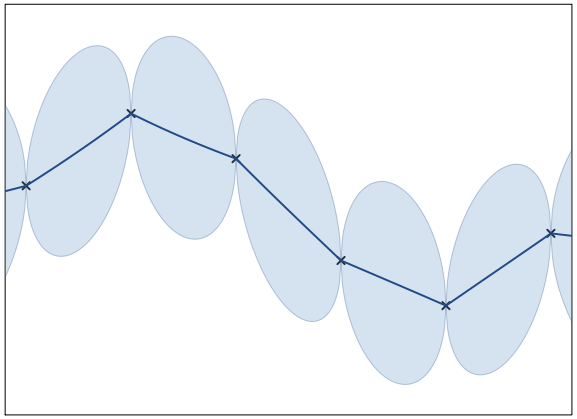


There is no kernel that is intrinsically better... it depends!

Gaussian kernel:



Exponential kernel:



The kernel has to be chosen according to the prior belief on the behaviour of the function to study:

- is it continuous, differentiable, how many times?
- is it stationary ?
- is it monotonous, bounded? Cf. [López-Lopera et al., 2018]
- ... (more on this in the kernel design section later)
- Default: constant trend μ (empirical mean or $\widehat{\mu}$ from max likelihood [Roustant et al., 2012]) and Matérn 5/2 kernel.

Kriging of noisy data

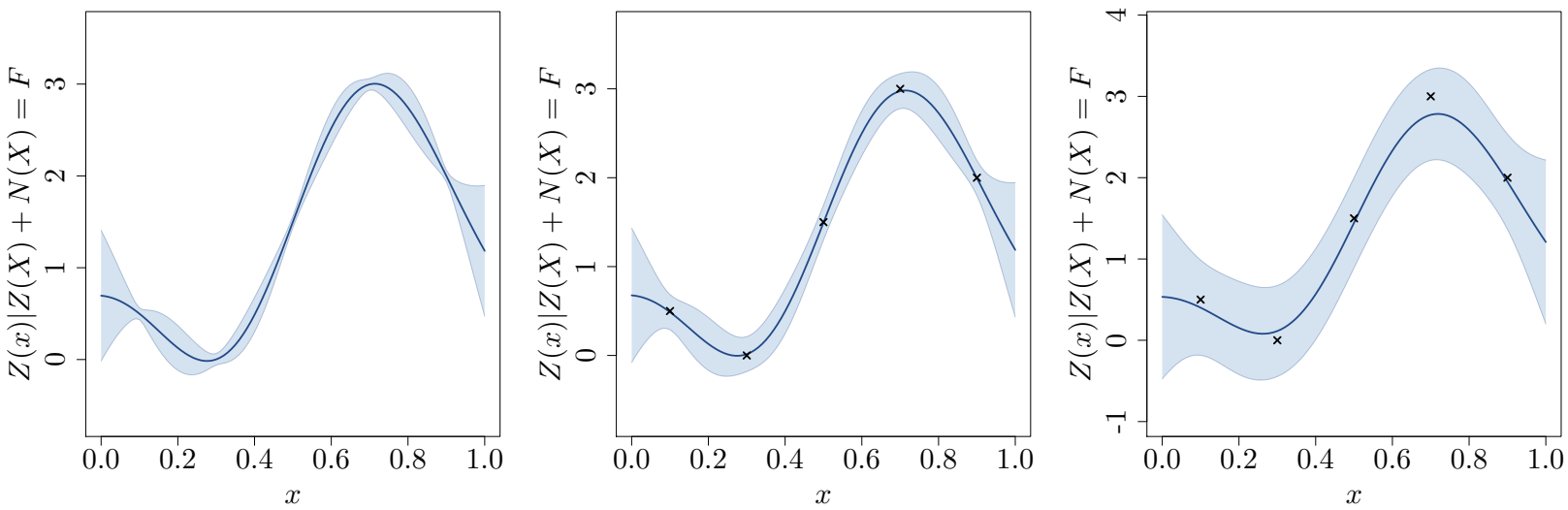
An important special case, noisy data $F = f(X) + \varepsilon$.
Model F with $Y(x) + N(x)$ where $N(x) \sim \mathcal{N}(0, n(.,.))$ independent of $Y(x)$. Then,

$$\begin{aligned} \text{Cov}(Y(x^i) + N(x^i), Y(x^j) + N(x^j)) &= k(x^i, x^j) + n(x^i, x^j) \\ \text{Cov}(Y(x), Y(x^i) + N(x^i)) &= k(x, x^i) \end{aligned}$$

The expressions of GPR with noise become (just apply Gaussian vector conditioning with the above)

$$\begin{aligned} m(x) &= \mathbb{E}[Z(x)|Z(X) + N(X){=}F] \\ &= \mu(x) + k(x, X)(k(X, X) + n(X, X))^{-1}(F - \mu(X)) \\ c(x, x') &= \text{Cov}[Z(x), Z(x')|Z(X) + N(X){=}F] \\ &= k(x, x') - k(x, X)(k(X, X) + n(X, X))^{-1}k(X, x') \end{aligned}$$

Examples of models with observation noise for $n(x, x') = \tau^2 \delta_{x,x'}$:



The values of τ^2 are respectively 0.001, 0.01 and 0.1.

Kriging with noise kernel (nugget) does not interpolate the data.

A small τ^2 (e.g., 10^{-10}) often used to make the covariance matrix invertible (more on regularization of GPs in [Le Riche et al., 2017]).

Parameter estimation

We have seen previously that the choice of the kernel and its parameters (σ^2, the θ's, the trend and other parameters) have a great influence on the model.

In order to choose a prior that is suited to the data at hand, we can:

- minimise the model error
- maximize the model likelihood

We now detail the second approach.

Definition: The **likelihood** of a distribution with a density p_U given observations $u^1, \ldots, u^p$ is:

$$L = \prod_{i=1}^{p} p_U(u^i)$$

The likelihood measures the adequacy between observations and a distribution.

In the GPR context, we often have only **one observation** of the vector F. The likelihood is then:

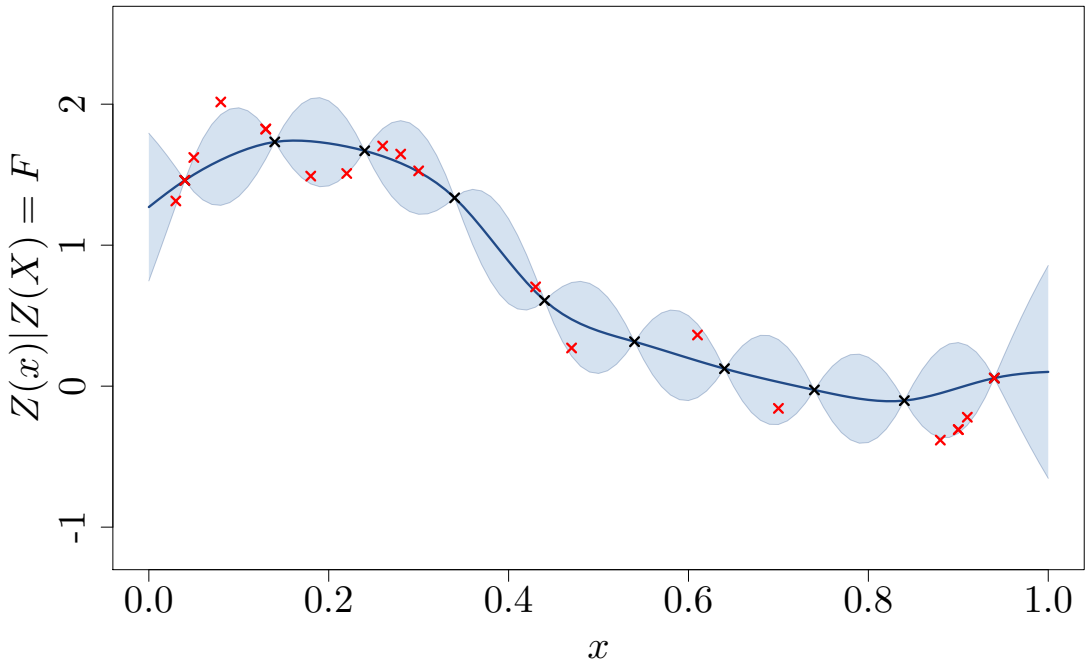
$$L = p_{Y(X)}(F) = \frac{1}{(2\pi)^{n/2}\det(k(X,X))^{1/2}} \times \exp\left(-\frac{1}{2}(F-\mu(X))^\top k(X,X)^{-1}(F-\mu(X))\right).$$

It is thus possible to maximise L – or $\log(L)$ – with respect to the kernel and model parameters in order to find a well suited prior. The likelihood in a multi-modal function in θ's and must be optimized with global optimization algorithms.

(more details on likelihood such as concentration in, e.g. [Le Riche, 2014])

We have seen that given some observations $F = f(X)$, it is very easy to build lots of models, either by changing the kernel parameters or the kernel itself.

The question is now how to measure the quality of a model to build the best one at the end.
Principle: introduce new data and to compare them to the model prediction.



Let X_t be the test set and $F_t = f(X_t)$ be the associated observations.

The accuracy of the mean can be measured by computing:

$$\text{Mean Square Error} \qquad MSE = \text{mean}((F_t - m(X_t))^2)$$

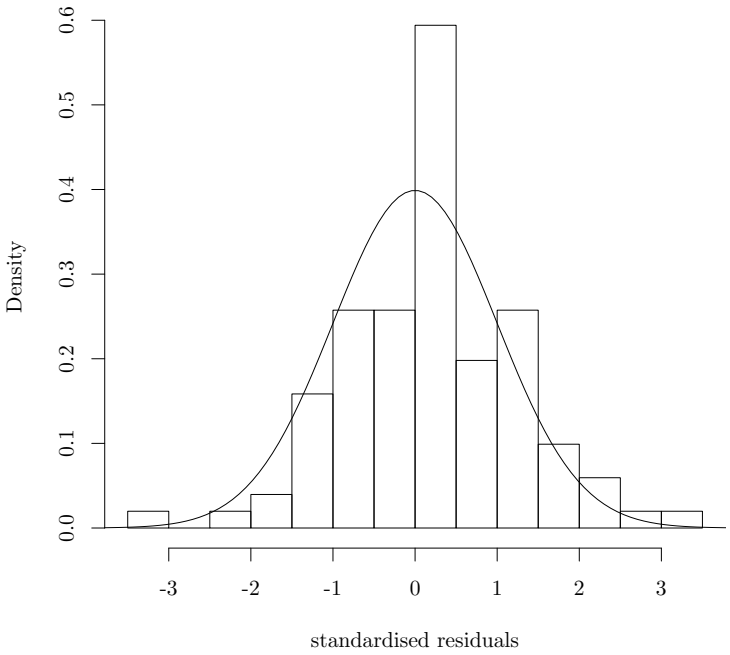
$$\text{A "normalized" criterion} \qquad Q_2 = 1 - \frac{\sum (F_t - m(X_t))^2}{\sum (F_t - \text{mean}(F_t))^2}$$

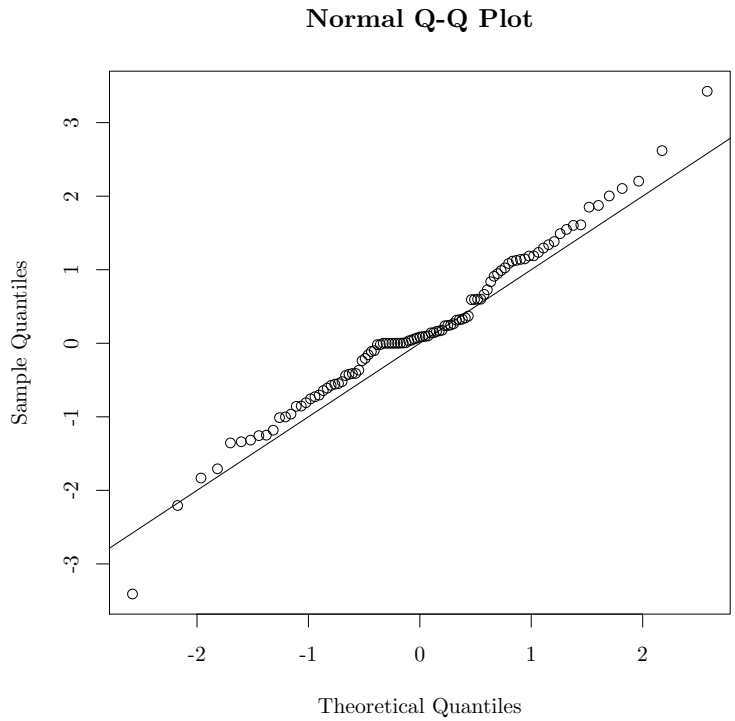
On the above example we get $MSE = 0.038$ and $Q_2 = 0.95$.

The predicted distribution can be tested by normalizing the residuals.

According to the model, $F_t \sim \mathcal{N}(m(X_t), c(X_t, X_t))$.

$c(X_t, X_t)^{-1/2}(F_t - m(X_t))$ should thus be independents $\mathcal{N}(0, 1)$:





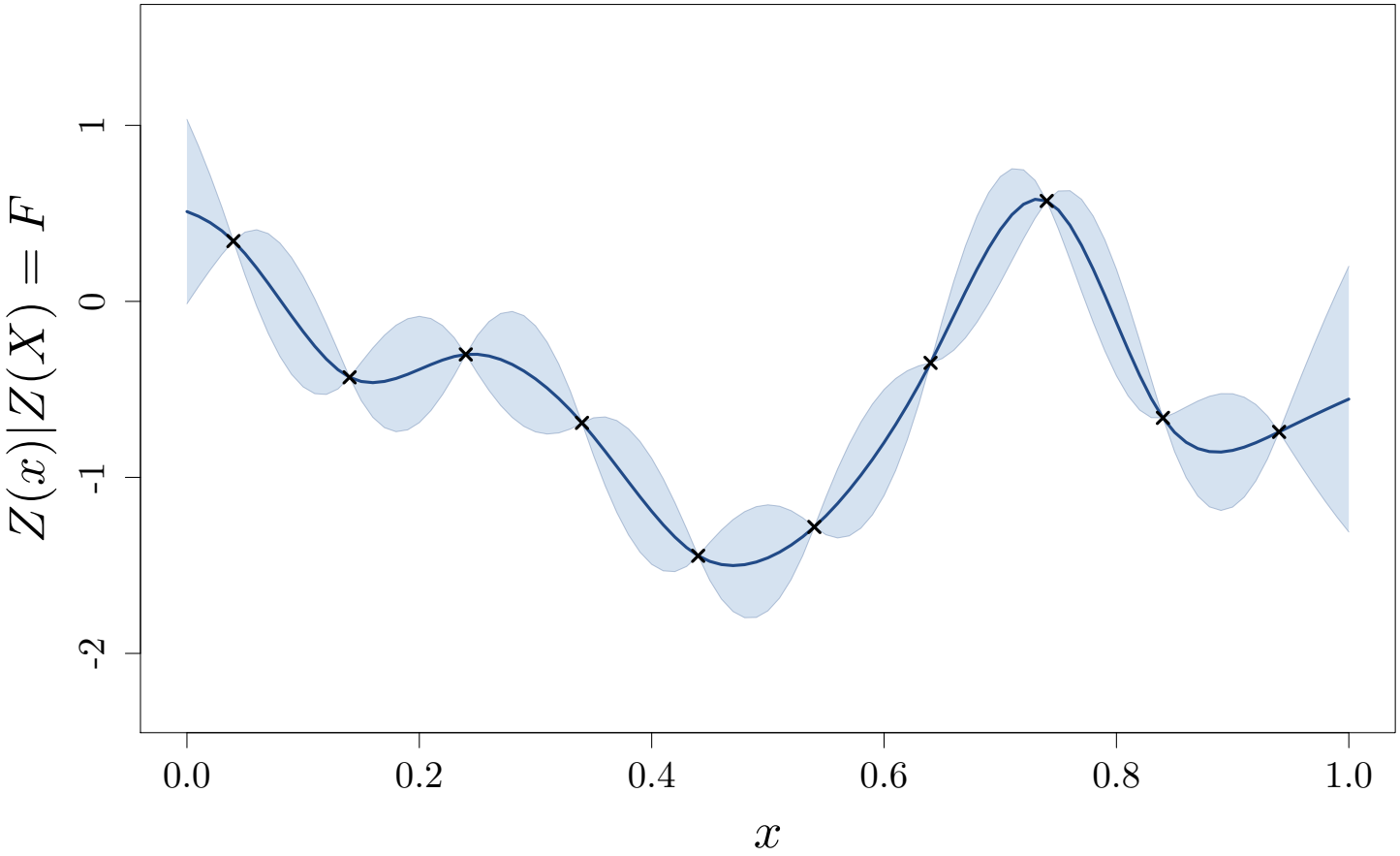
When no test set is available, another option is to consider cross validation methods such as leave-one-out.

The steps are:

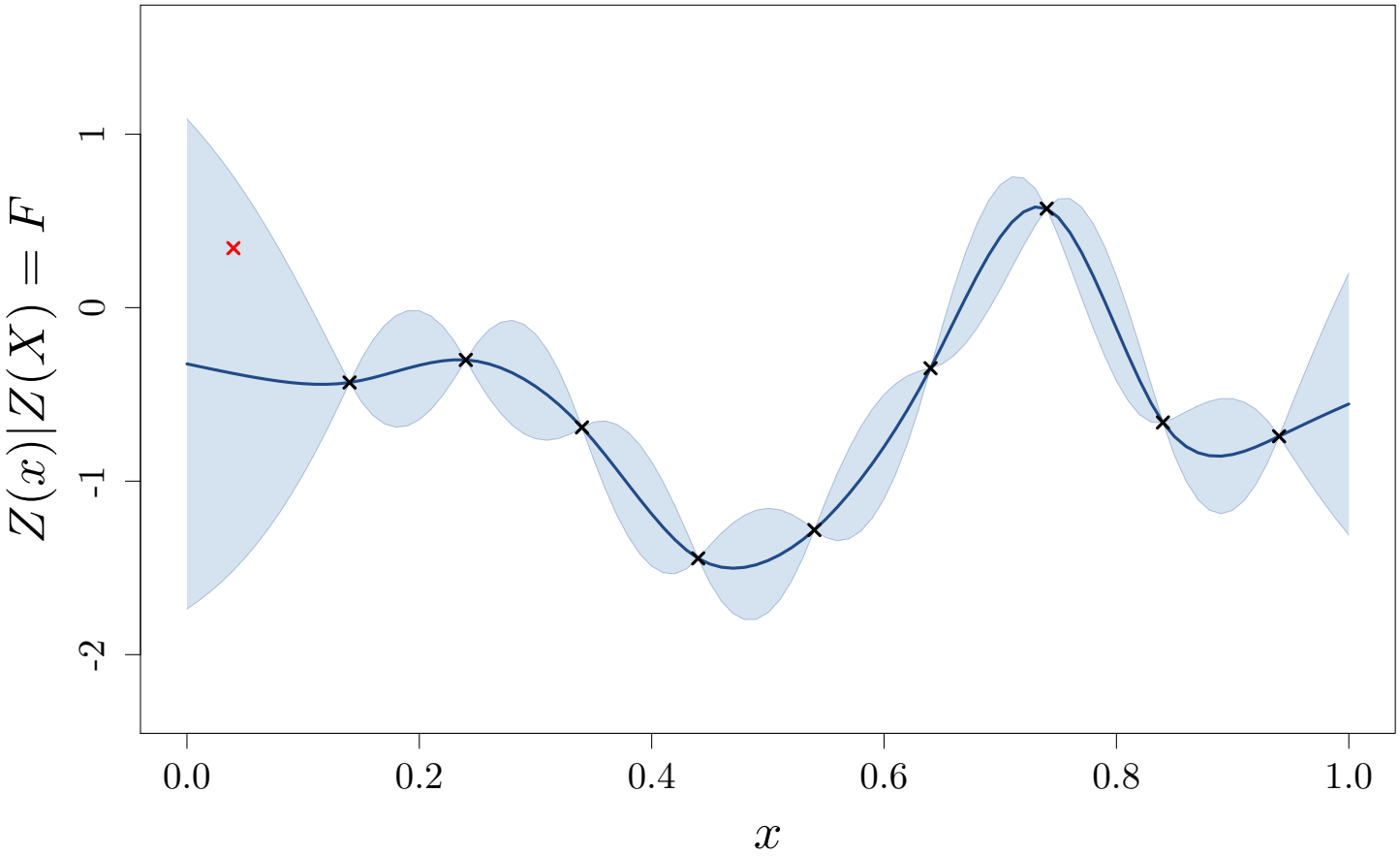
1. build a model based on all observations except one
2. compute the model error at this point

This procedure can be repeated for all the design points in order to get a vector of error.

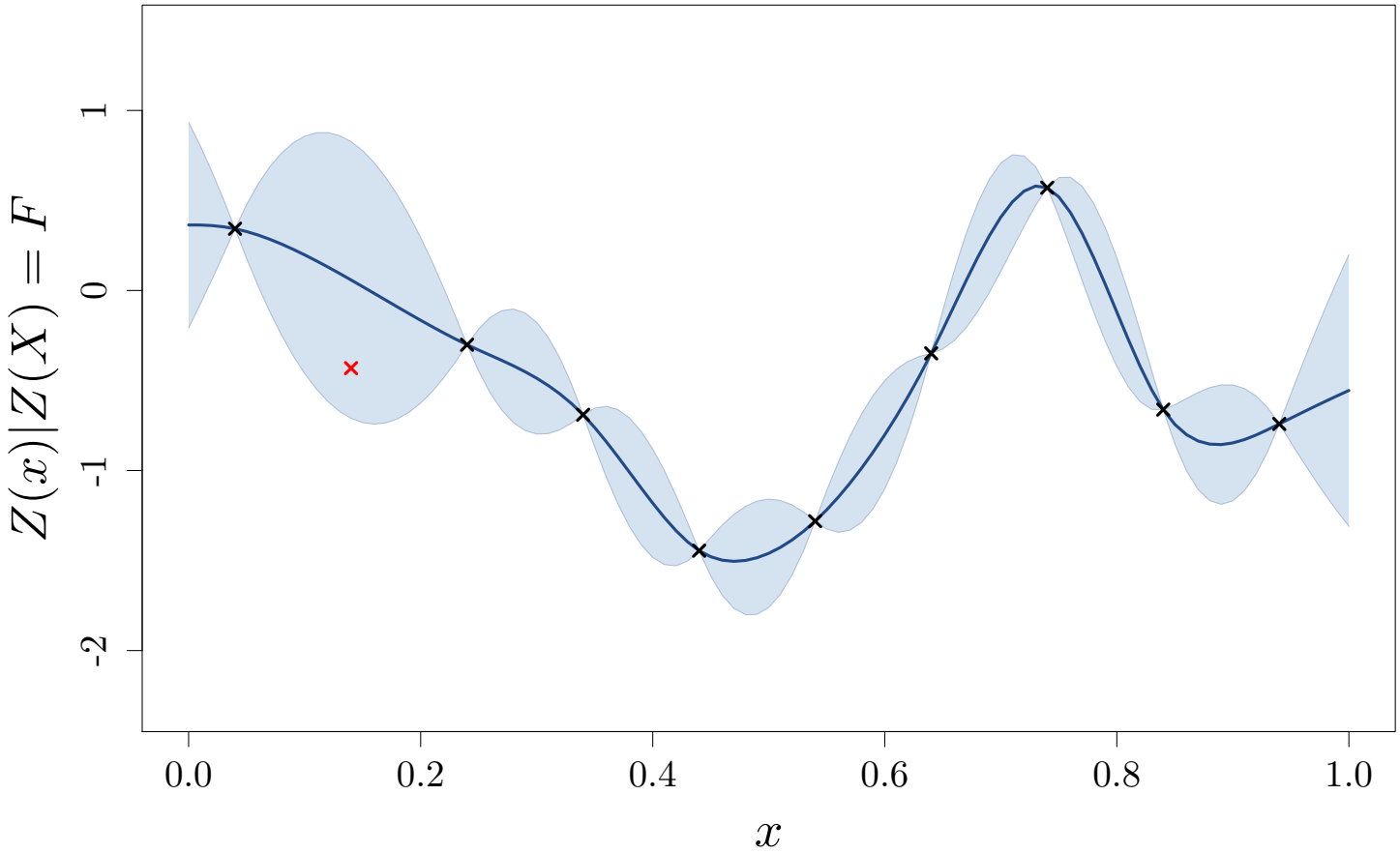
Model to be tested:



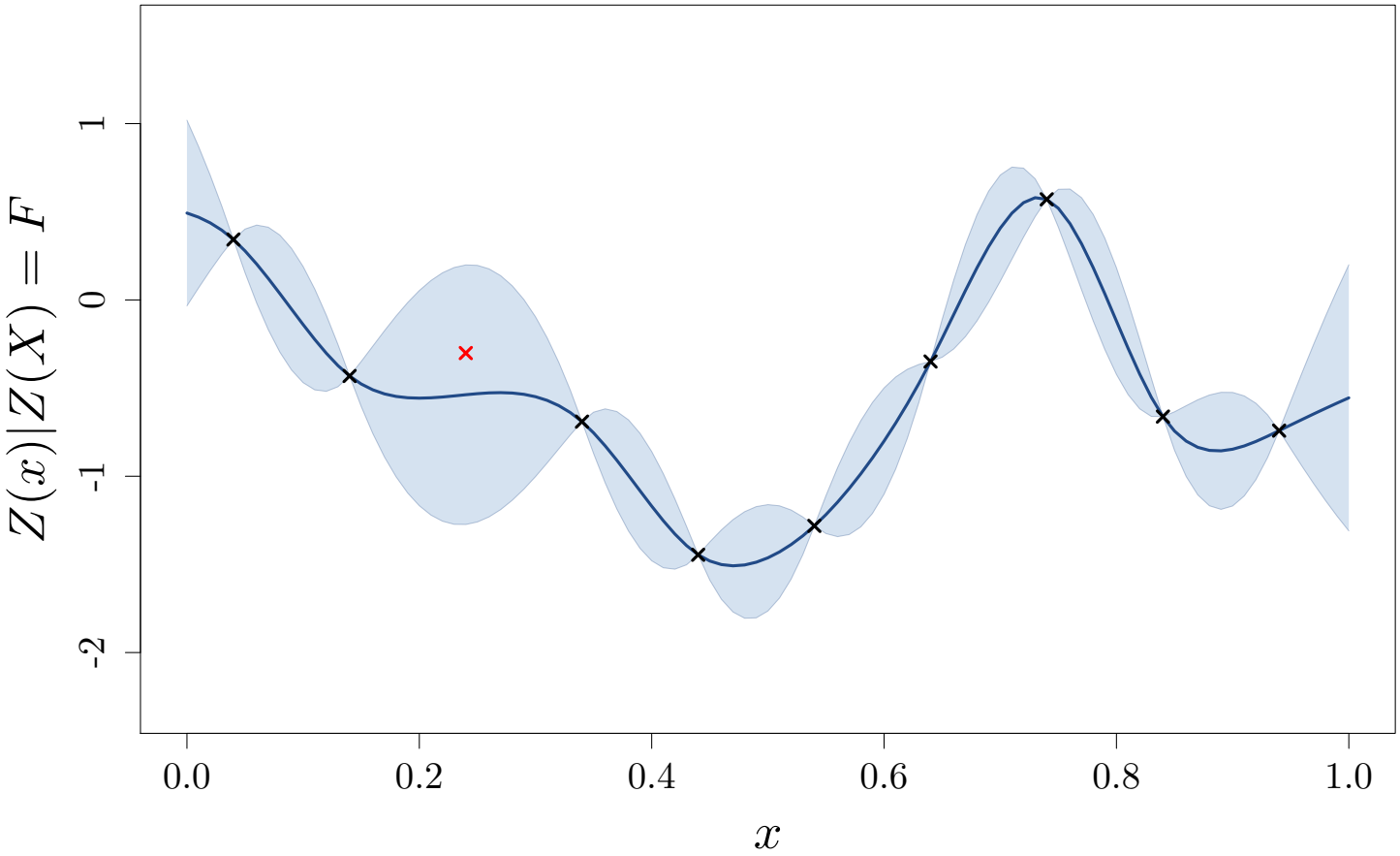
Step 1:



Step 2:



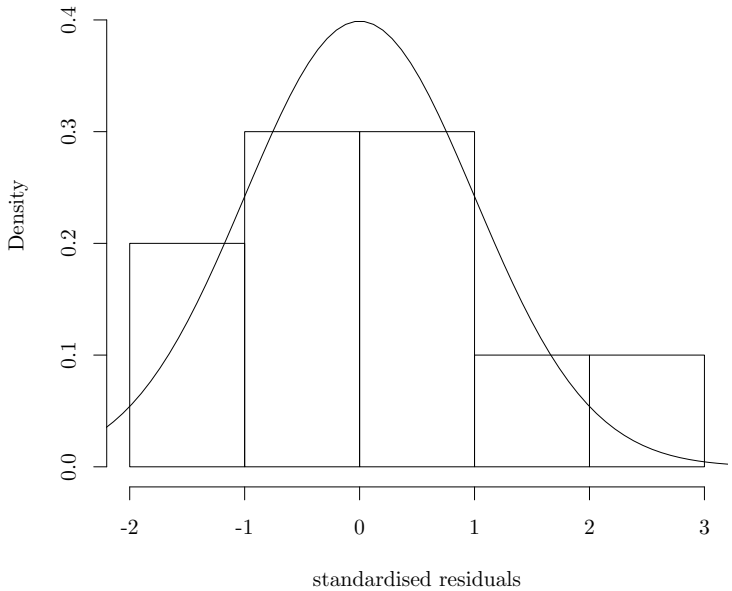
Step 3:

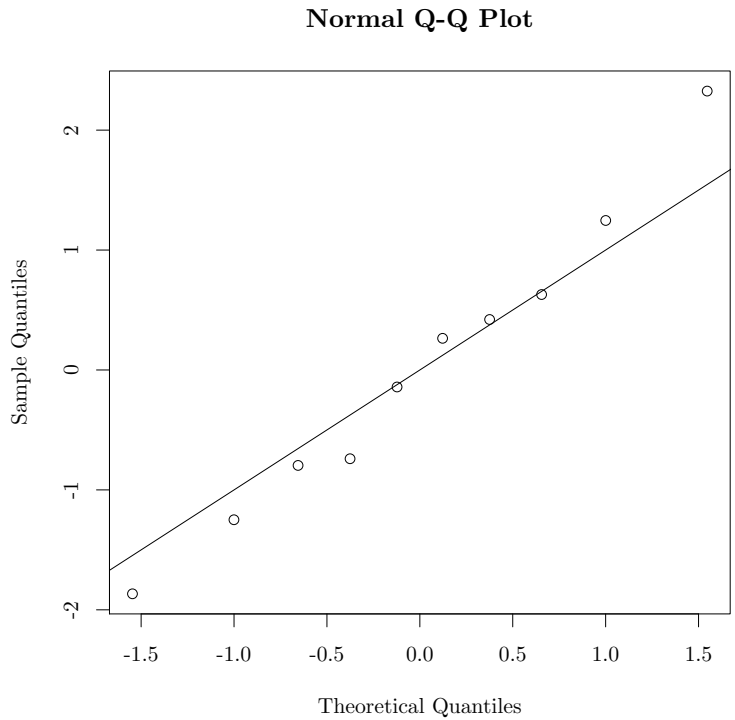


We finally obtain:

$$MSE = 0.24 \text{ and } Q_2 = 0.34.$$

We can also look at the residual distribution. For leave-one-out, there is no joint distribution for the residuals so they have to be standardized independently.

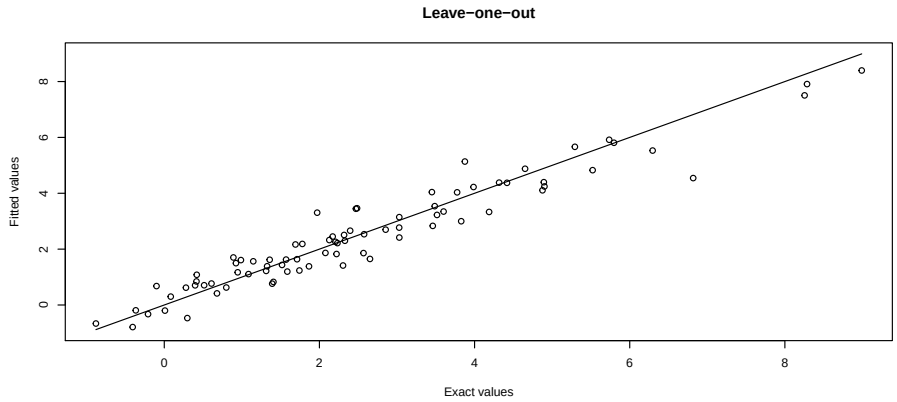


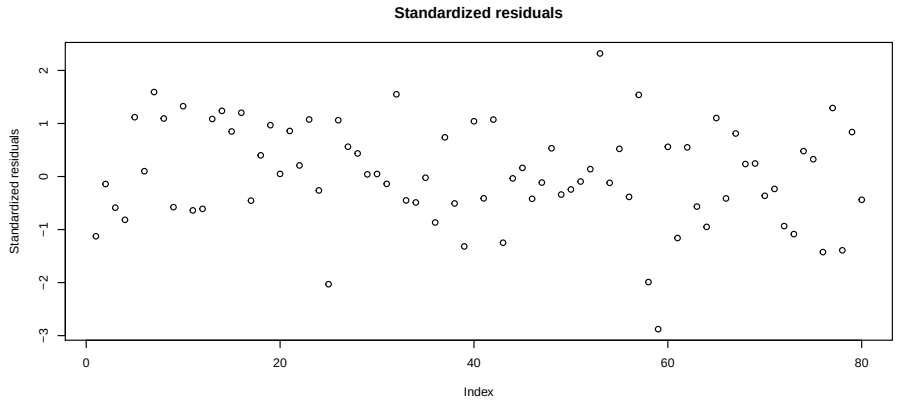


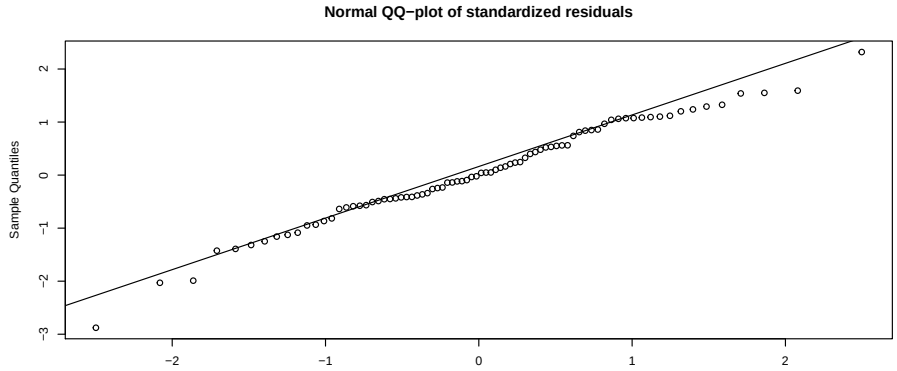
Sample code in R

(with 6D Hartman function)

```
library(DiceKriging)
library(DiceDesign)
X <- lhsDesign(n=80,...
    dimension=6)$design
X <- data.frame(X)
y <- apply(X, 1, hartman6)
mlog <- km(design = X,
response = -log(-y))
plot(mlog)
```







Content

Kernel design: making new from old

Many operations can be applied to psd functions while retaining this property

Kernels can be:

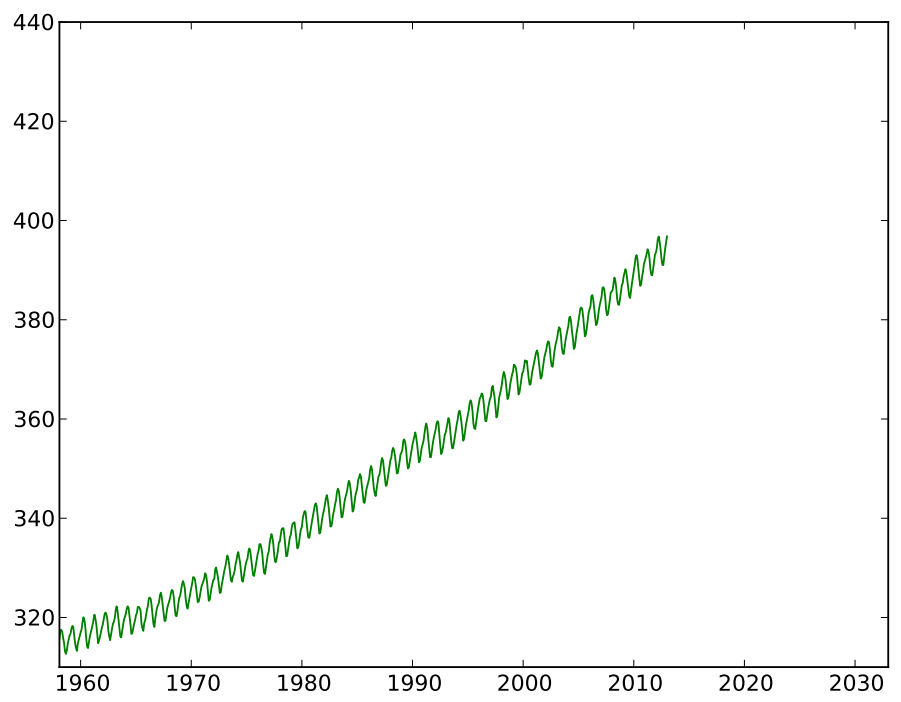
- Summed
  - On the same space $k(x, x') = k_1(x, x') + k_2(x, x')$
  - On the tensor space $k(x, x') = k_1(x_1, x_1') + k_2(x_2, x_2')$
- Multiplied
  - On the same space $k(x, x') = k_1(x, x') \times k_2(x, x')$
  - On the tensor space $k(x, x') = k_1(x_1, x_1') \times k_2(x_2, x_2')$
- Composed with a function
  - $k(x, x') = k_1(h(x), h(x'))$

to create new (non stationary) kernels, increase their dimension. All these transformations can be combined. Examples ...

Sum of kernels over the same space

Example (The Mauna Loa observatory dataset)

This famous dataset compiles the monthly CO_2 concentration in Hawaii since 1958.

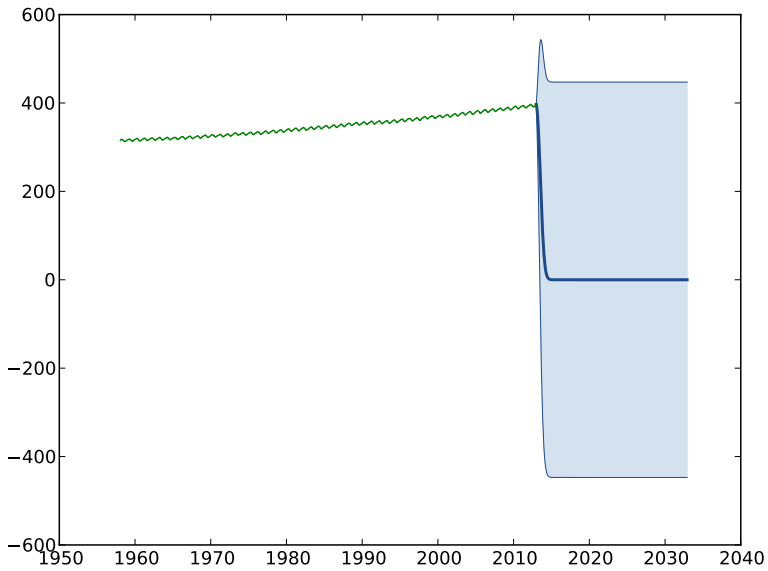


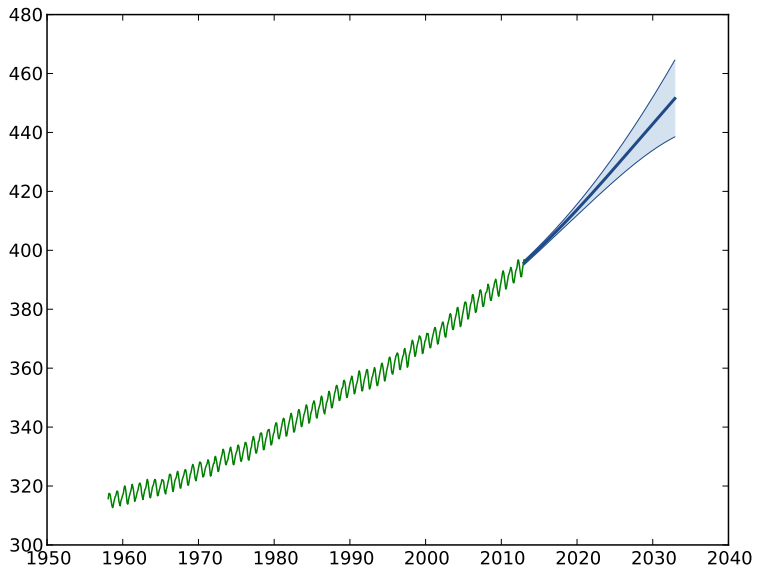
Let's try to predict the concentration for the next 20 years.

Sum of kernels over the same space

We first consider a squared-exponential kernel with a small and a large length-scale:

$$k_{se}(x, x') = \sigma^2 \exp\left(-\frac{(x - x')^2}{\theta^2}\right)$$



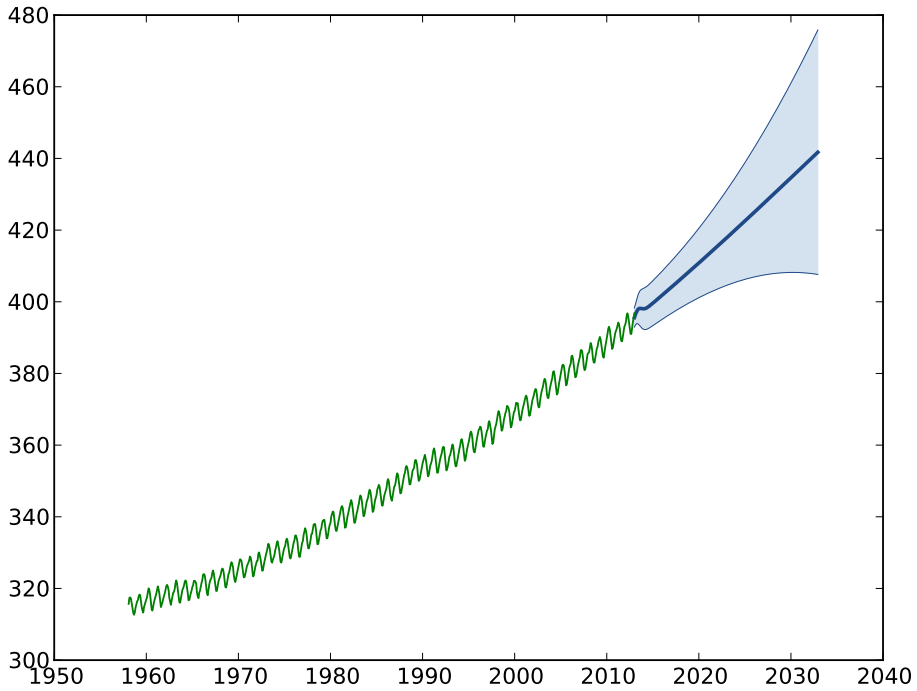


The results are terrible!

Sum of kernels over the same space

What happens if we sum both kernels?

$$k(x, x') = k_{se1}(x, x') + k_{se2}(x, x')$$



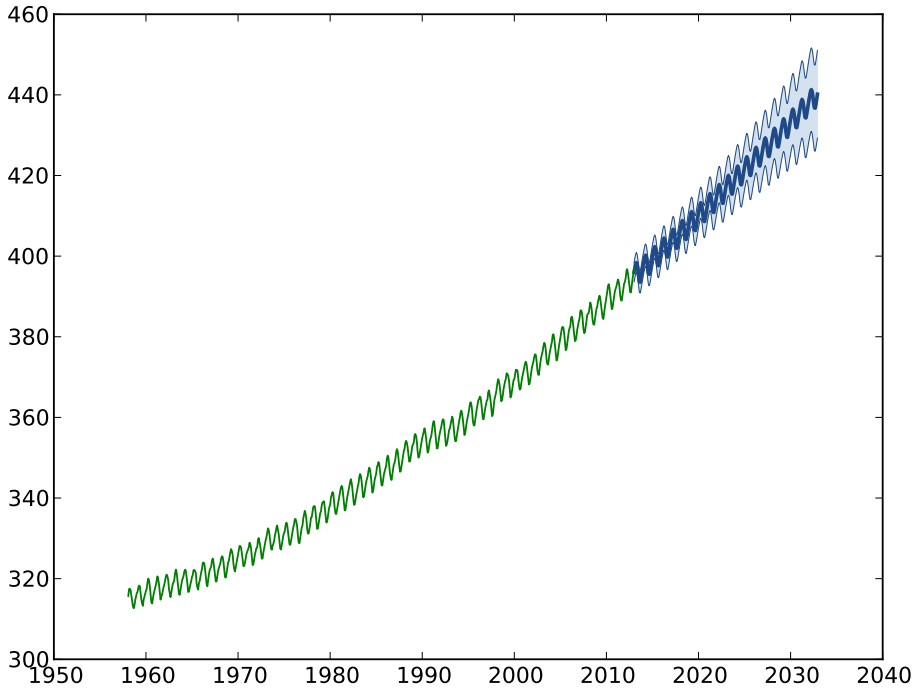
The model is drastically improved

Sum of kernels over the same space

We can try the following kernel:

$$k(x, x') = \sigma_0^2 x^2 x'^2 + k_{se1}(x, x') + k_{se2}(x, x') + k_{per}(x, x')$$

The first term is a product of linear kernels. The periodic kernel is $k_{per}(x, x') = -\sigma^2 \exp\left(-\frac{\sin^2(\pi|x-x'|/p)}{\theta}\right)$



Once again, the model is significantly improved.

Composition with a function

Let k_1 be a kernel over $\mathcal{X}_1 \times \mathcal{X}_1$ and h be an arbitrary function $\mathcal{X} \to \mathcal{X}_1$, then

$$k(x, x') = k_1(h(x), h(x'))$$

is a kernel over $\mathcal{X} \times \mathcal{X}$.

proof

$$\sum\sum a_i a_j k(x_i, x_j) = \sum\sum a_i a_j k_1(\underbrace{h(x_i)}_{y_i}, \underbrace{h(x_j)}_{y_j}) \geq 0 \ \square$$

Remarks:

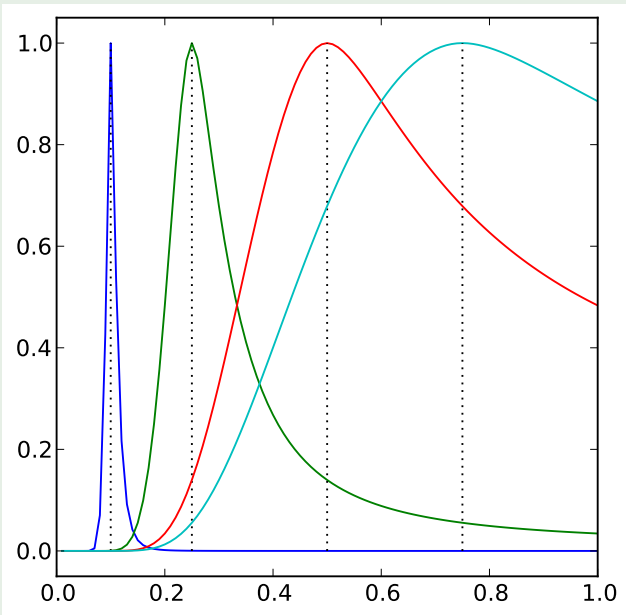
- k corresponds to the covariance of $Z(x) = Z_1(h(x))$
- This can be seen as a (non-linear) rescaling of the input space. A way to make non-stationary kernels.

Example

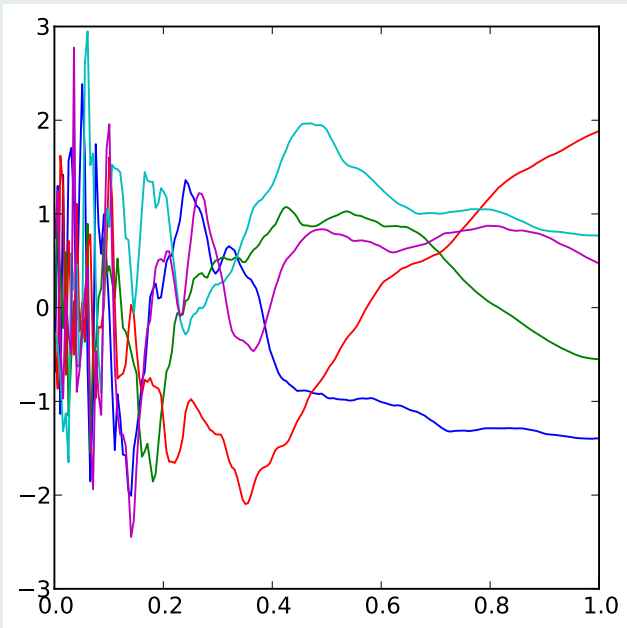
We consider $h(x) = \frac{1}{x}$ and a Matérn 3/2 kernel $k_1(x, y) = (1 + |x - y|)e^{-|x-y|}$.

We obtain:

Kernel



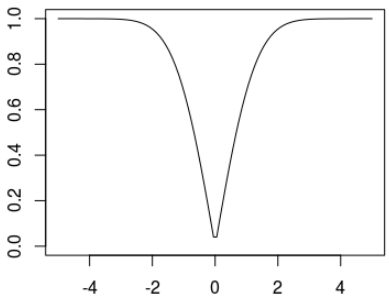
Sample paths

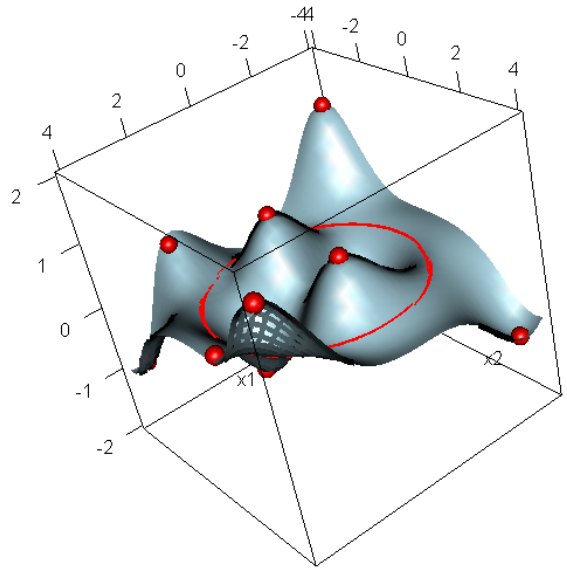


$k(x, x') = h(x)h(x')k_1(x, x')$ is a valid kernel.
Can be seen as $k_1(,)\times$ composition of function and linear kernel.
Better, see it as the covariance of $Y(x) = h(x)Y_1(x)$.

- Trajectories of $Y()$ and $Y_1()$ can be obtained from each other:
  $Y(x) \mid Y(X) = F$ sampled through $Y_1()$ with
  $h(x)Y_1(x) \mid Y_1(X) = F/h(X)$ (component-wise division)
- Boundary conditions: say you want to impose that all trajectories go through $Y(x) = 0$ for all (infinite number) x's such that $a(x) = 0$. Use $h(x) = d(a(x))$

$d()$





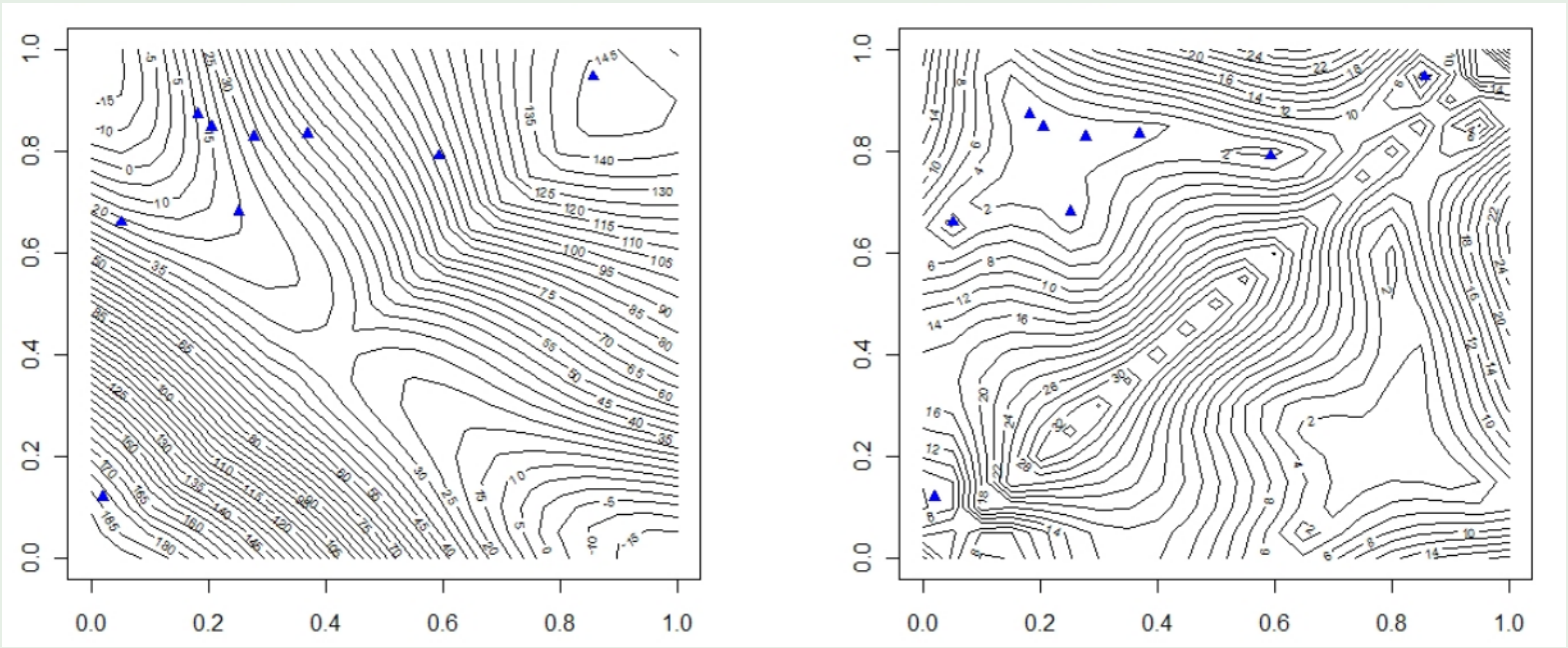
from Durrande & Gauthier ENBIS09

- Symmetric kernel: to have $Y\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = Y\begin{pmatrix} x_2 \\ x_1 \end{pmatrix}$, use

$$k\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \begin{pmatrix} x_1' \\ x_2' \end{pmatrix}\right) = k'\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \begin{pmatrix} x_1' \\ x_2' \end{pmatrix}\right) + k'\left(\begin{pmatrix} x_2 \\ x_1 \end{pmatrix}, \begin{pmatrix} x_1' \\ x_2' \end{pmatrix}\right)$$

Example

Symmetrical kriging, mean and std. deviation.



Note how the variance is null symmetrically to observations.

from [Ginsbourger, 2009]

Content

The statistical point of view

Kriging is often introduced as best linear interpolator:

- linear: $\widehat{Y}(x) = \sum_{i=1}^{n} \lambda_i(x) Y(x^i) = \boldsymbol{\lambda}(x)^\top Y(X)$
- unbiased: $\mathbb{E}\widehat{Y}(x) = \boldsymbol{\lambda}(x)^\top \mathbb{E}Y(X) = \mathbb{E}Y(x) = \mu(x)$
- best: $\boldsymbol{\lambda}(x) = \arg\min_{\boldsymbol{\lambda} \in \mathbb{R}^n} \mathbb{E}\| \widehat{Y}(x) - Y(x)\|^2$

This constrained optimization problem is solved in $\boldsymbol{\lambda}(x)$ and the kriging equations are recovered from

- $m(x) = \mathbb{E}(\widehat{Y}(x) \mid Y(X) = F)$
- and $c(x, x') = \mathbb{E}\left((\widehat{Y}(x) - Y(x))(\widehat{Y}(x') - Y(x'))\right)$

but the link with the GP interpretation is typically not discussed,

- $\mathbb{E}(Y(x) \mid Y(X) = F) \stackrel{?}{=} \mathbb{E}(\widehat{Y}(x) \mid Y(X) = F)$
- and $\mathbb{E}((Y(x) - \mu(x))^2 \mid Y(X) = F) \stackrel{?}{=} \mathbb{E}(\widehat{Y}(x) - Y(x))^2$.

The functional point of view (thanks Xavier Bay)

The kernel $k(.,.)$ defines a space of functions, a RKHS, $\mathcal{H}_k := \text{span}\{k(x,.), x \in \mathcal{X}\}$ with an inner product[a] $\langle .,. \rangle_{\mathcal{H}}$ such that there is a linear evaluation functional $\langle f(.), k(x,.) \rangle_{\mathcal{H}} = f(x)$.

[a] The inner product between 2 functions is: $f(.) = \sum_{i=1}^{M} \alpha_i k(x_i,.)$, $g(.) = \sum_{j=1}^{N} \beta_j k(x'_j,.)$, $\langle f(.), g(.) \rangle_{\mathcal{H}} = \sum_{i,j} \alpha_i \beta_j k(x_i, x'_j)$, which implies the evaluation functional.

Associated to the psd $k(.,.)$ are eigenvalues and eigenfunctions

$$\int_{\mathcal{X}} k(.,t)\phi_i(t)dt = \lambda_i \phi_i(.) \quad , \quad \lambda_1 \geq \lambda_2 \geq \ldots \geq 0$$

The $\phi_i(.)$'s form an orthonormal basis of $\mathcal{H}$ w.r.t. the usual scalar product. All this is a generalization of the eigendecomposition of symmetric positive definite matrices to infinite dimensions.

$\Rightarrow$ Another way to make kernels (Mercer): choose the $\phi_i(.)$'s,

$$k(x, x') = \sum_{i=1}^{N} \lambda_i \phi_i(x) \phi_i(x')$$

Degenerated $k(.,.)$ if $N < +\infty$ (the covariance matrix becomes non-invertible beyond N observations)

Proof: $k(x,.) \in L^2(\mathcal{X})$, $k(x,.) = \sum_i \langle k(x,.), \phi_i(.) \rangle_{L^2} \phi_i(.)$
$= \sum_i \int_{\mathcal{X}} k(x,t)\phi_i(t)dt\phi_i(.) = \sum_i \lambda_i \phi_i(x)\phi_i(.)$ $\square$

Alternative definition of the RKHS:
$\mathcal{H} = \{f(.) \in L^2(\mathcal{X}) : f(.) = \sum_{i=1}^{\infty} c_i \phi_i(.)$ and $\sum_{i=1}^{\infty} \frac{c_i^2}{\lambda_i} < +\infty\}$

i.e., impose a sufficiently fast decrease in eigencomponents, a kind of regularization.

Intuition behind the alternative definition of the RKHS, $\mathcal{H}$:

1. construct an orthonormal basis of $\mathcal{H}$

The $\phi_i(.)$'s are bi-orthogonal w.r.t. $\langle ., . \rangle_{L^2}$ and $\langle ., . \rangle_{\mathcal{H}}$ but need to be normalized in $\mathcal{H}$: we use the integral equation of slide 61, which gives an intuition that the $\phi_i()$'s belong to $\mathcal{H}$ (think of the integral as a sum). Then,

$$\begin{aligned}
\langle \phi_i, \phi_j \rangle_{\mathcal{H}} &= \frac{1}{\lambda_i \lambda_j} \langle \int_{\mathcal{X}} k(., t) \phi_i(t) dt, \int_{\mathcal{X}} k(., t') \phi_j(t') dt' \rangle \\
&= \frac{1}{\lambda_i \lambda_j} \int_{\mathcal{X}} \int_{\mathcal{X}} \phi_i(t) \phi_j(t') k(t, t') dt dt' = \frac{1}{\lambda_i \lambda_j} \int_{\mathcal{X}} \phi_i(t) \left[\int_{\mathcal{X}} k(t, t') \phi_j(t') dt' \right] dt \\
&= \frac{\lambda_j}{\lambda_i \lambda_j} \langle \phi_i, \phi_j \rangle_{L^2} = \frac{\lambda_j}{\lambda_i \lambda_j} \delta_{ij} \quad \Rightarrow \quad \widetilde{\phi}_i() = \sqrt{\lambda_i} \phi_i() \text{ is an orthonormal basis of } \mathcal{H}
\end{aligned}$$

2. f is in the RKHS if its coefficients are a converging series,

$$f() = \sum_i c_i \phi_i() = \sum_i \frac{c_i}{\sqrt{\lambda_i}} \widetilde{\phi}_i()$$

$$\|f()\|_{\mathcal{H}} = \sum_i \frac{c_i^2}{\lambda_i} < +\infty \qquad \square$$

Trajectories can be generated with (Karhunen-Loève),

$$Y(x) = \sum_{i=1}^{N} \sqrt{\lambda_i}\xi_i\phi_i(x) \quad , \quad \xi_i \sim \mathcal{N}(0,1) \text{ i.i.d}$$

$\Rightarrow$ in general the trajectories are not in the RKHS:

N finite, $Y(x) \in \mathcal{H}$, N infinite, $Y(x) \notin \mathcal{H}$.

Proof: $\sum_{i=1}^{N} \frac{(\sqrt{\lambda_i}\xi_i)^2}{\lambda_i} = \sum_{i=1}^{N} \xi_i^2 \xrightarrow{N\nearrow} N$ □

But the GP mean is in the RKHS.

Proof: $m(x) = k(x,X)k(X,X)^{-1}F = \sum_{i=1}^{n} \beta_i k(x,x^i)$ □

GPR and complexity control

If complexity is measured as the norm of the function, the Representer Theorem [Schölkopf et al., 2001] says that $m()$ is the least complex interpolator:

$$m(.) \;=\; \begin{cases} \arg\min_{h\in\mathcal{H}} \|h\|^2_{\mathcal{H}} \\ \text{such that } h(x^i) = f(x^i)\ ,\ i = 1,\dots,n \end{cases}$$

Proof: $h = h_k + h'$ where $h_k = \sum_{i=1}^n c_i k(x^i,.)$ and $h' \perp h_k$. Then,

$f(x^i) = \langle k(x^i,.), h_k + h'\rangle = \langle k(x^i,.), \sum_{j=1}^n c_j k(x^j,.)\rangle + \underbrace{0}_{h'\perp h_k} = \sum_{j=1}^n c_j k(x^j,x^i) = k(x^i,X)c,$

c the vector of n c_i's. The problem becomes, $\min_{c\in\mathbb{R}^n, h'\perp \text{span}\{k(x^i,.)\}} c^\top k(X,X)c$ such that $k(X,X)c = F$ whose solution is $h' = 0$, $c = k(X,X)^{-1}F$, i.e.,

$m(x) = h_k(x) = k(x,X)k(X,X)^{-1}F$ □

A regularization can also be seen in the likelihood in Slide 38 with $\det(k(X,X))$ which must be as small as possible ($\Rightarrow$ large θ's).

Content

Other GPR variants

- Universal kriging: account for trend parameters in the GPR equations. Cf. [Le Riche, 2014].
- Multiple outputs: cokriging. Cf. [Garland et al., 2019],[Fricker et al., 2013], with gradient as 2nd output [Laurent et al., 2019].
- Discrete x variables: Cf. [Roustant et al., 2019], mixed variables and optimization [Pelamatti et al., 2019].

Other names, (almost) same equations

$$m(x) \;=\; k(x,X)k(X,X)^{-1}F \qquad \text{is ubiquitous}$$

- Bayesian linear regression: the posterior distribution is identical to the GPR equations under conditions on the kernel, cf. [Le Riche, 2014] slide 35 and [Rasmussen and Williams, 2006], slide 20 of [Rosić, 2019].
- Kalman filter, see slide 21 of [Rosić, 2019].
- LS-SVR: same functional form of predictor (sum of kernels centered), but explicit regularization control (C, whereas GPR is implicit in likelihood), no uncertainty.
- RBF (Radial Basis Functions) [Broomhead and Lowe, 1988]: same prediction, no uncertainty (hence no likelihood).

Kriging issues

- Too large n: $k(X, X)$ is $n \times n$ and takes $\mathcal{O}(n^3)$ operations for its inversion $\Rightarrow$ not directly applicable beyond $n = 1000$. Solutions: inducing points [Hensman et al., 2013], nested kriging [Rullière et al., 2018].
- $k(X, X)$ is ill-conditioned: regularize it, 3 variants in [Le Riche et al., 2017] (nugget, pseudo-inverse and distribution-wise GP).
- Maximizing the likelihood (for infering the GP parameters) or minimizing the cross-validation error are multi-modal problems in $\mathcal{O}(d)$ dimensions: use global optimization algorithms.

References I

Broomhead, D. S. and Lowe, D. (1988).
Radial basis functions, multi-variable functional interpolation and adaptive networks.
Technical report, Royal Signals and Radar Establishment Malvern (United Kingdom).

Durrande, N. and Le Riche, R. (2017).
Introduction to Gaussian Process Surrogate Models.
Lecture at 4th MDIS form@ter workshop, Clermont-Fd, France.
HAL report cel-01618068.

Fricker, T. E., Oakley, J. E., and Urban, N. M. (2013).
Multivariate gaussian process emulators with nonseparable covariance structures.
Technometrics, 55(1):47–56.
cokrigeage : convolution, LMC, ...

Garland, N., Le Riche, R., Richet, Y., and Durrande, N. (2019).
Aerospace System Analysis and Optimization in uncertainty, chapter Cokriging for multifidelity analysis and optimization.
Springer.
submitted for publication.

References II


Ginsbourger, D. (2009).
Multiples métamodèles pour l'approximation et l'optimisation de fonctions numériques multivariables.
PhD thesis, Mines de Saint-Etienne.


Hensman, J., Fusi, N., and Lawrence, N. D. (2013).
Gaussian processes for big data.
arXiv preprint arXiv:1309.6835.


Krige, D. G. (1951).
A statistical approach to some basic mine valuation problems on the witwatersrand.
Journal of the Southern African Institute of Mining and Metallurgy, 52(6):119–139.


Laurent, L., Le Riche, R., Soulier, B., and Boucard, P.-A. (2019).
An overview of gradient-enhanced metamodels with applications.
Archives of Computational Methods in Engineering, 26(1):61–106.


Le Riche, R. (2014).
Introduction to Kriging.
Lecture at mnmuq2014 summer school, Porquerolles, France.
HAL report cel-01081304.

References III

Le Riche, R., Mohammadi, H., Durrande, N., Touboul, E., and Bay, X. (2017).
A Comparison of Regularization Methods for Gaussian Processes.
slides of talk at siam conference on optimization op17 and accompanying technical report hal-01264192, Vancouver, BC, Canada.
https://www.emse.fr/~leriche/op17_R_LeRiche_slides_v2.pdf and https://hal.archives-ouvertes.fr/hal-01264192.

López-Lopera, A. F., Bachoc, F., Durrande, N., and Roustant, O. (2018).
Finite-dimensional gaussian approximation with linear inequality constraints.
SIAM/ASA Journal on Uncertainty Quantification, 6(3):1224–1255.

Matheron, G. (1963).
Principles of geostatistics.
Economic geology, 58(8):1246–1266.

Pelamatti, J., Brevault, L., Balesdent, M., Talbi, E.-G., and Guerin, Y. (2019).
Efficient global optimization of constrained mixed variable problems.
Journal of Global Optimization, 73(3):583–613.

Rasmussen, C. E. and Williams, C. K. (2006).
Gaussian Processes for Machine Learning.
The MIT Press.

References IV

Rosić, B. (2019).
Inverse problems.
slides of the MNMUQ2019 course.
presented at the French-German summer school Modeling and Numerical Methods for Uncertainty Quantification, Porquerolles.

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, 51(1).

Roustant, O., Padonou, E., Deville, Y., Clément, A., Perrin, G., Giorla, J., and Wynn, H. (2019).
Group kernels for gaussian process metamodels with categorical inputs.
SIAM/ASA Journal on Uncertainty Quantification.

Rullière, D., Durrande, N., Bachoc, F., and Chevalier, C. (2018).
Nested kriging predictions for datasets with a large number of observations.
Statistics and Computing, 28(4):849–867.

Schölkopf, B., Herbrich, R., and Smola, A. J. (2001).
A generalized representer theorem.
In *International conference on computational learning theory*, pages 416–426. Springer.