ENBIS pre-conference workshop

Introduction to Kriging using R and JMP

Nicolas Durrande – Mines Saint-Étienne

durrande@emse.fr

We have seen this morning how to build Kriging models and what are the assumptions they rely on. We get into more details:

1. it is straightforward to change the prior belief
2. parameters can be included in models to get a better fit between prior belief and data
3. one must **validate** a model before using it!

Kernels

Proving that a function is psd is often intractable. However there are a lot of functions that have already been proven to be psd:

$$\begin{aligned}
\text{squared exp.} \quad & k(x,y) = \sigma^2 \exp\left(-\frac{(x-y)^2}{2\theta^2}\right) \\
\text{Matern 5/2} \quad & k(x,y) = \sigma^2 \left(1 + \frac{\sqrt{5}|x-y|}{\theta} + \frac{5|x-y|^2}{3\theta^2}\right) \exp\left(-\frac{\sqrt{5}|x-y|}{\theta}\right) \\
\text{Matern 3/2} \quad & k(x,y) = \sigma^2 \left(1 + \frac{\sqrt{3}|x-y|}{\theta}\right) \exp\left(-\frac{\sqrt{3}|x-y|}{\theta}\right) \\
\text{exponential} \quad & k(x,y) = \sigma^2 \exp\left(-\frac{|x-y|}{\theta}\right) \\
\text{Brownian} \quad & k(x,y) = \sigma^2 \min(x,y) \\
\text{white noise} \quad & k(x,y) = \sigma^2 \delta_{x,y} \\
\text{constant} \quad & k(x,y) = \sigma^2 \\
\text{linear} \quad & k(x,y) = \sigma^2 xy
\end{aligned}$$

When k is a function of $x - y$, the kernel is called **stationary**.

σ^2 is called the **variance** and θ the **lengthscale**.

For $d \geq 2$, there can be one rescaling parameter per dimension:

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

squared exp. $\quad k(x,y) = \sigma^2 \exp\left(-\frac{1}{2}||x-y||_\theta^2\right)$

Matern 5/2 $\quad k(x,y) = \sigma^2 \left(1 + \sqrt{5}||x-y||_\theta + \frac{5}{3}||x-y||_\theta^2\right) \exp\left(-\sqrt{5}||x-y||_\theta\right)$

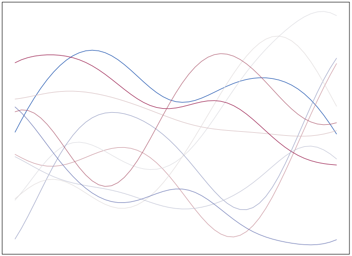
Matern 3/2 $\quad k(x,y) = \sigma^2 \left(1 + \sqrt{3}||x-y||_\theta\right) \exp\left(-\sqrt{3}||x-y||_\theta\right)$

exponential $\quad k(x,y) = \sigma^2 \exp\left(-||x-y||_\theta\right)$

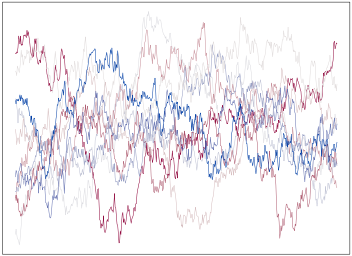
If all θ_i are equal, we say that the kernel/process is **isotropic**.

Changing kernel means changing the prior belief on f

Gaussian kernel:

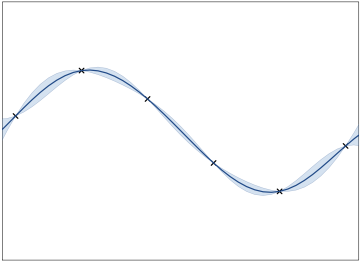


Exponential kernel:

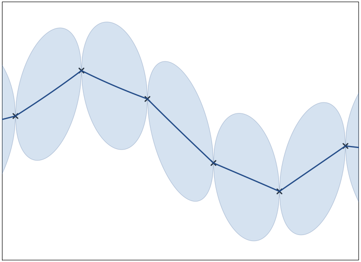


It thus has a huge impact on the model:

Gaussian kernel:



Exponential kernel:



There is no model/kernel that is intrinsically better... it depends on the data!

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

- is it continuous, differentiable, how many times?
- is it stationary ?
- Are there particular patterns (symmetry, periodicity, additivity)?
- ...

Parameter estimation

We have seen previously that the choice of the kernel and its parameters have a great influence on the model.

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

- minimising the model error
- Using maximum likelihood estimation

We will now detail the second one.

The likelihood quantifies the adequacy between a distribution and some observations.

Definition

Let f_Y be a pdf depending on some parameters p and let $y_1, \dots, y_n$ be independent observations. The **likelihood** is defined as

$$L(p) = \prod_{i=1}^{n} f_Y(y_i; p).$$

A high value of $L(p)$ indicates the observations are likely to be drawn from $f_Y(.; p)$.

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

$$L(\sigma^2, \theta) = f_{Z(X)}(F; \sigma^2, \theta) = \frac{1}{|2\pi k(X,X)|^{1/2}} \exp\left(-\frac{1}{2} F^t k(X,X)^{-1} F\right).$$

It is thus possible to maximise L with respect to the kernel's parameters in order to find a well suited prior.

In practice, the value of σ^2 can be obtained analytically. Others, such as θ, need numerical optimization.

Example

We consider 100 sample from a Matérn 5/2 process with parameters $\sigma^2 = 1$ and $\theta = 0.2$, and n observation points. We then try to recover the kernel parameters using MLE.

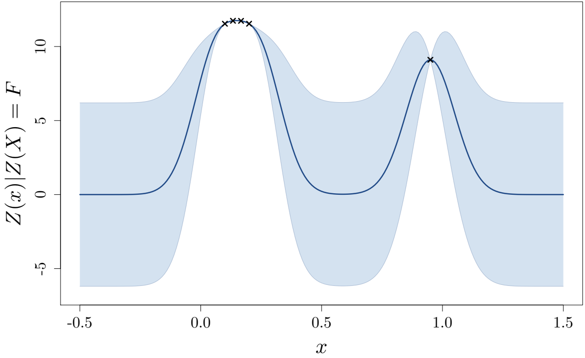
n	5	10	15	20
σ^2	1.0 (0.7)	1.11 (0.71)	1.03 (0.73)	0.88 (0.60)
θ	0.20 (0.13)	0.21 (0.07)	0.20 (0.04)	0.19 (0.03)

MLE can be applied regardless to the dimension of the input space.

Trends

We have seen that GPR models go back to zero if we consider a centred prior.

This behaviour is not always wanted



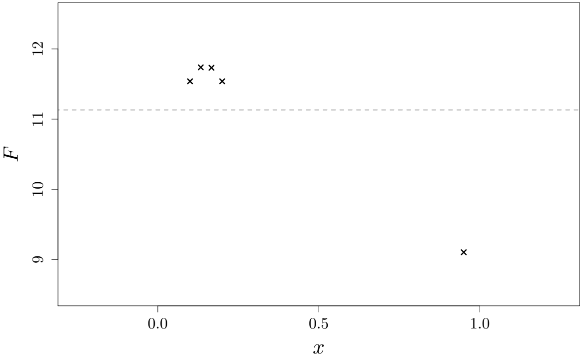
We may want to introduce some trend in models... One can distinguish:

- **simple kriging**: there is no trend or it is known
- **ordinary kriging**: the trend is a constant
- **universal kriging**: the trend is given by basis functions

The question is how to estimate the trend coefficients. Hereafter, we will focus on ordinary kriging and write t the mean value.

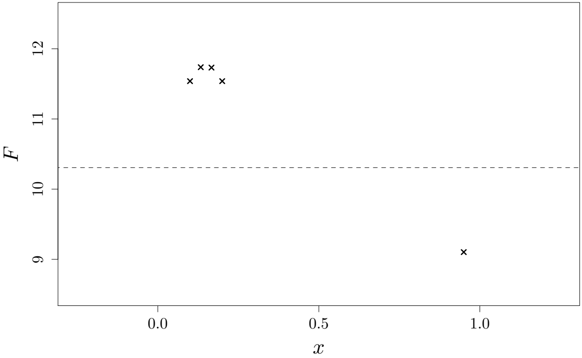
Basic least squares minimization gives

$$\hat{t} = \frac{1}{n}\sum_{i=1}^{n} F_i = \frac{\mathbf{1}^t F}{\mathbf{1}^t \mathbf{1}}$$



Generalised least squares is more appropriate

$$\hat{t} = \frac{\mathbf{1}^t k(X,X)^{-1} F}{\mathbf{1}^t k(X,X)^{-1} \mathbf{1}}$$



This can be seen as maximum likelihood estimation.

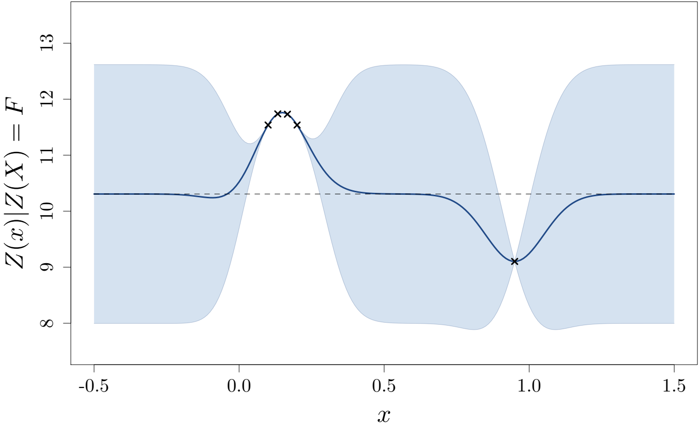
The expression of the **best predictor** is given by the usual conditioning of a GP:

$$m(x) = \mathrm{E}[Z(x)|Z(X) = F] = \hat{t} - k(x, X)k(X, X)^{-1}(F - \hat{t})$$

Regarding the **model variance**, it must account for the estimator's variance.

$$\begin{aligned} \mathrm{var}[Z(x)|Z(X)] = {} & k(x, x) - k(x, X)k(X, X)^{-1}k(X, x) \\ & + \frac{(\mathbf{1} + k(x, X)k(X, X)^{-1}\mathbf{1})^t(\mathbf{1} + k(x, X)k(X, X)^{-1}\mathbf{1})}{\mathbf{1}^t k(X, X)^{-1}\mathbf{1}} \end{aligned}$$

On the previous example we obtain:



Lab session (20 min)

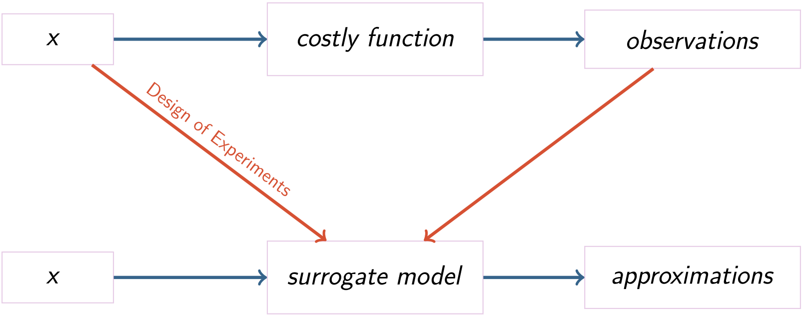
1. reload the kriging model you obtained this morning on the catapult data. (you can also use the provided one if you wish).
2. Apply the `print()` function to your model, can you understand the output?
3. Apply the `plot()` function to your model to get some cross validation diagnostics.
4. Try changing the trend and kernel to improve the model.
5. What location can you suggest for the optimum?

$$x_1 = \qquad , \ x_2 = \qquad , \ x_3 = \qquad , \ x_4 = \qquad .$$

Run the simulator at that location, is there an improvement?

Application to optimization

Framework

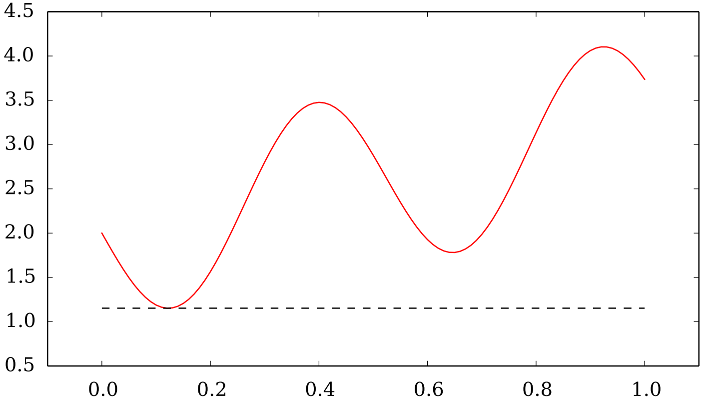


x
costly function
observations
Design of Experiments
x
surrogate model
approximations

Example : optimization

Case study

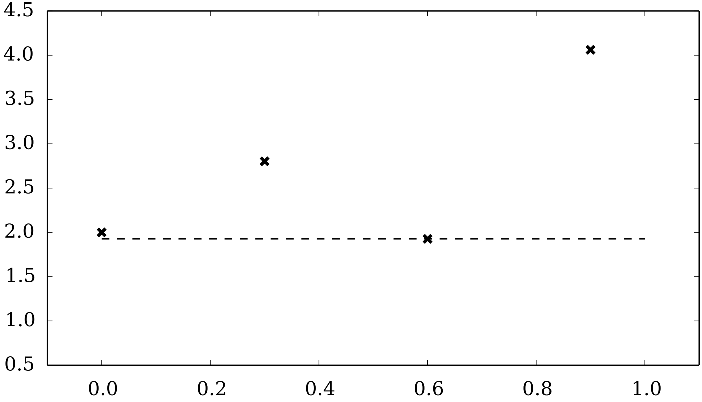
We want to minimize the following function...



Example : optimization

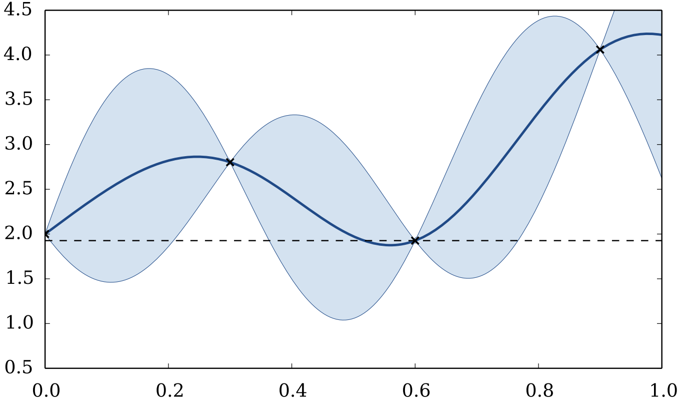
Case study

... which is costly.



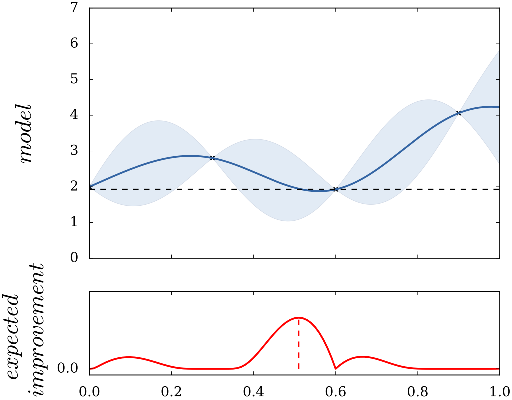
Example : optimization

We build a kriging model



Example : optimization

A common criterion is the **expected improvement**

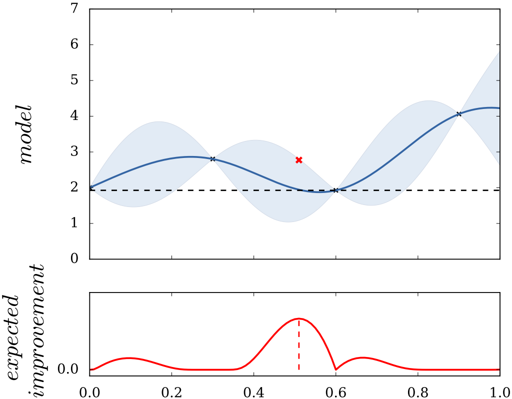


$$EI(x) = \sqrt{v(x)}\left(u(x)cdf(u(x)) + pdf(u(x))\right)$$

$$\text{with } u(x) = \frac{\min(F) - m(x)}{\sqrt{v(x)}}$$

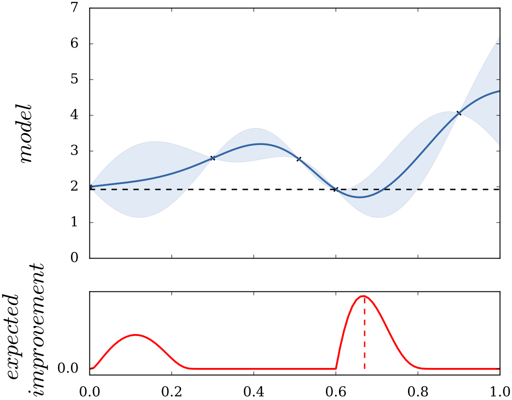
Example : optimization

We run another experiment where this criterion is maximum



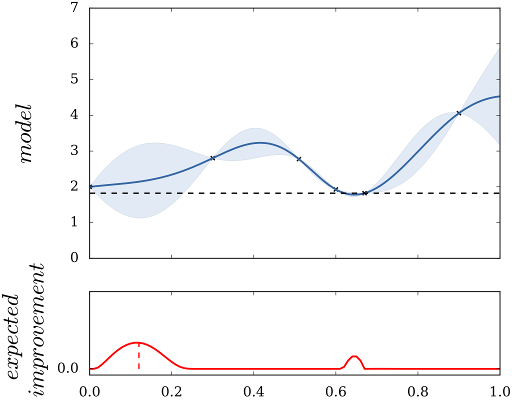
Example : optimization

Iteration 2 :



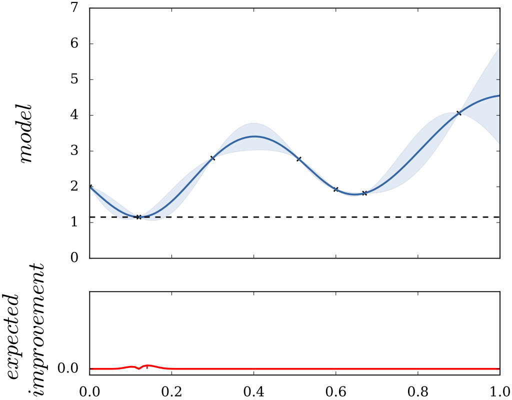
Exemple d'application : optimisation

Iteration 3 :



Exemple d'application : optimisation

Iteration 4 :



Lab session (20 min)

1. We are interested here in a maximization problem... update `Y` and the kriging model to make them suitable for usual minimization algorithms.
2. Get the new location of the point maximising the EI (see function `max_EI` from package `DiceOptim`)
3. Run the experiment at this location and update the model. Has the minimum been improved ?
4. Make 20 iterations using `EGO.nsteps`. You can look at the results using the function `visualizeEGO`.
5. What is the optimal value you obtain ?

$$x_1 = \qquad , \; x_2 = \qquad , \; x_3 = \qquad , \; x_4 = \qquad , \; Y_{min} = \qquad .$$

Conclusion

Statistical models are useful when little data is available. they allow to

- interpolate or approximate functions
- Compute quantities of interests (such as mean value, optimum, ...)
- Get some uncertainty measure

Small Recap We have seen that

- Many kernels can be used to build models.
  - Given some data, there is not one GP model but an infinity...
- Kernels encode the prior belief on the function to approximate.
  - They should be chosen accordingly
- Model/kernel parameters can tuned to the problem at hand
- GPR models do not necessarily interpolate.
- Model validation is of the utmost importance
  - mean **and** predicted (co-)variance

Limitations

The complexity of using kriging models is in building the model

- $\mathcal{O}(n^2)$ for the storage footprint
  - storage of the covariance matrix $k(X, X)$
- $\mathcal{O}(n^3)$ for the number of operations
  - inversion of the covariance matrix $k(X, X)$

In practice, the maximum number of observation for classical models lies in the range [1000, 10000].

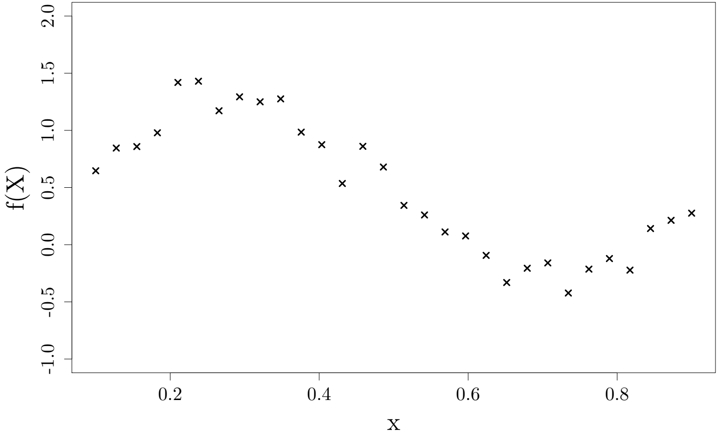
Another common issue is the numerical stability of the matrix inversion

- pseudo-inverse
- nugget (or jitter)

Appendix

Approximation

We are not always interested in models that interpolate the data.
For example, if there is some observation noise: $F = f(X) + \varepsilon$.



Approximation

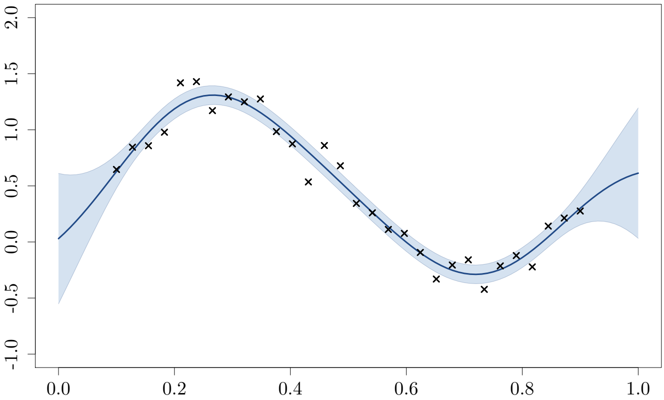
Let N be a process $\mathcal{N}(0, n)$ that represent the observation noise.
The expressions of GPR with noise are

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

$$\begin{aligned} c(x, y) &= \mathrm{cov}[Z(x), Z(y)|Z(X) + N(X){=}F] \\ &= k(x, y) - k(x, X)(k(X, X) + n(X, X))^{-1}k(X, y) \end{aligned}$$

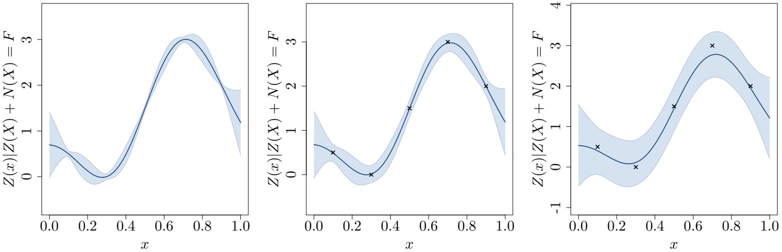
Approximation

We obtain the following model



Approximation

Influence of observation noise τ^2 (for $n(x, y) = \tau^2 \delta_{x,y}$):



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

In practice, τ^2 can be estimated with Maximum Likelihood.