

École chercheurs MEXICO, La Rochelle, Mars 2018

# Introduction to statistical modelling

Nicolas Durrande, [nicolas@prowler.io](mailto:nicolas@prowler.io)

PROWLER.io, Cambridge (UK) – Mines St-Étienne (France)





Mines St-Étienne is a French institute of Science and technology.

Main area of expertise:

- materials science & mechanical engineering
  - chemical engineering
  - applied mathematics, industrial engineering, environmental science
  - biomedical and healthcare engineering
  - microelectronics



*Small...*

- 600 students
  - 300 faculty staff
  - 180 PhD students

but *famous*: it is one of the leading French “grandes écoles”.

Since 2015, there is a student exchange program with the UTP.

The course will be over 4 days, with lectures every morning and lab sessions during afternoons.

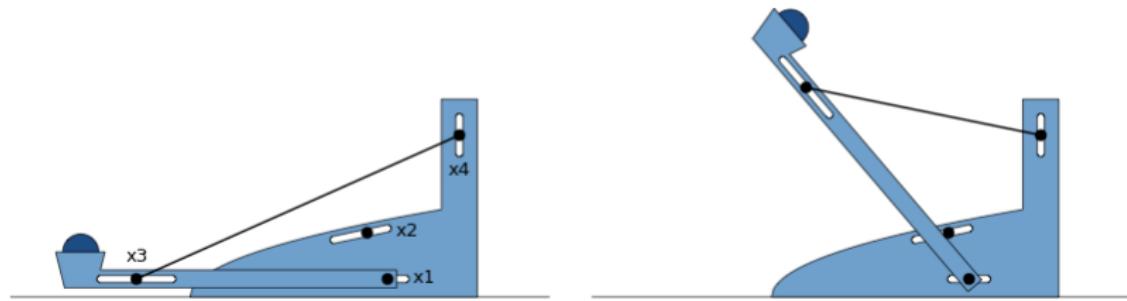
The agenda is as follow:

- day 1: Statistical models in engineering
  - day 2: Design of experiments
  - day 3: Gaussian Process regression
  - day 4: Optimization with Gaussian Process Regression

The course material is available

webpage: <https://sites.google.com/site/nicolasdurrandehomepage>  
github: user NicolasDurrande, repo Gaussian-Processes-short-Course

The lab session will be on the optimization of a catapult simulator:



**What values of  $(x_1, x_2, x_3, x_4)$  give the longest throw?**

This lab is based on a numerical simulator. Although the problem is simplistic, it shares a lot with classical engineering problems:

- Black box simulator → no analytical solution
- Noisy response (we will be interested in the breezy wind conditions)
- Experiments are “costly” → data is limited
- “High dimensional” input space → graphical interpretation is not possible

The project will be over 4 lab sessions:

- **Day 1:** data analysis, problem parametrization, region of interest
- **Day 2:** Design of experiments, data generation
- **Day 3:** statistical modelling
- **Day 4:** Optimization, new experiments

Lab sessions will be in python.

Make sure you bring computers !

## Outline of today's lecture

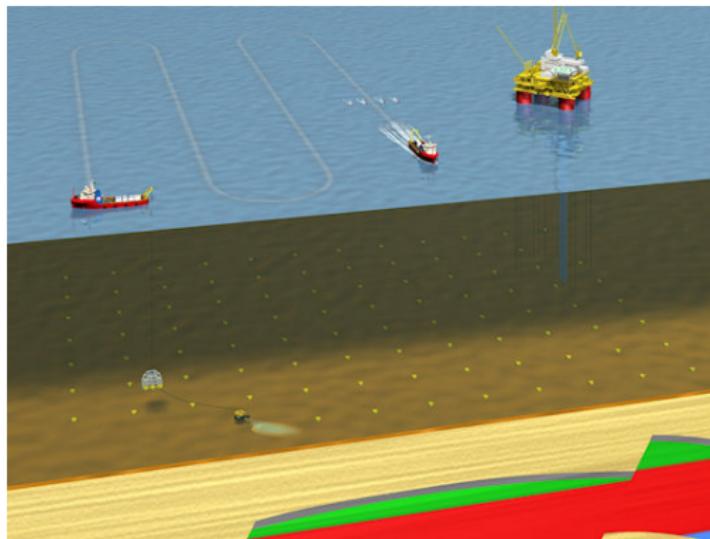
- Why (and when) statistical models can be useful in engineering?
- A basic example: Linear regression.
- Gaussian process regression

## Why are statistical models relevant in engineering?

There is a wide variety of situations where getting data about a system performance can be extremely expensive.

- real world experiments
- destructive tests
- prototyping
- numerical experiments

## Example: real world experiments



## Example: Destructive tests

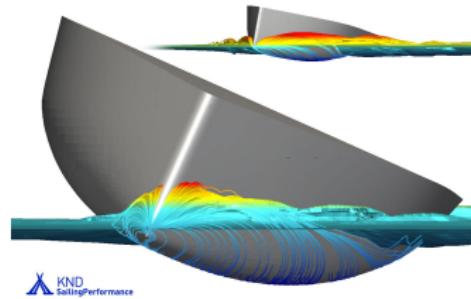


## Example: Prototyping of a boat shape



Knowing the drag for a given design requires costly experiments

## Example: Numerical experiments



Numerical experiments are less expensive but can be very time consuming!

In all these cases, the variable of interest can be seen as a function of the input parameters

$$y = f(x).$$

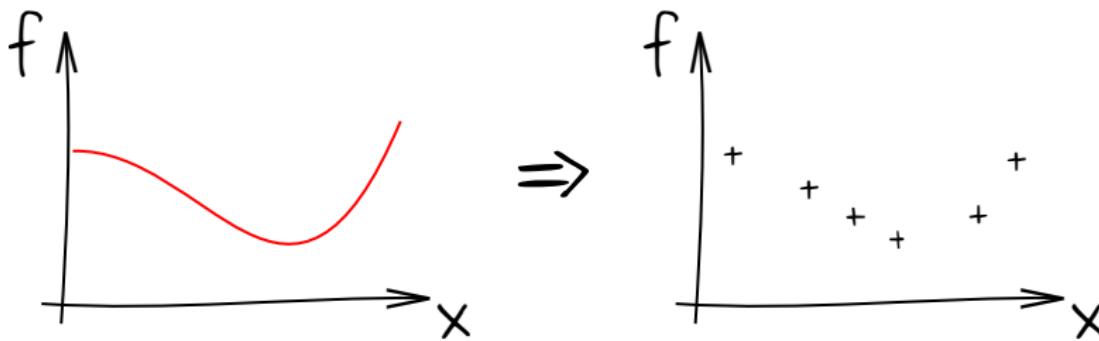
where  $f$  is a **costly to evaluate function**.

In the following, we will assume that

- $x \in \mathbb{R}^d$ : There are many input parameters
- $y \in \mathbb{R}$ : The output is a scalar.

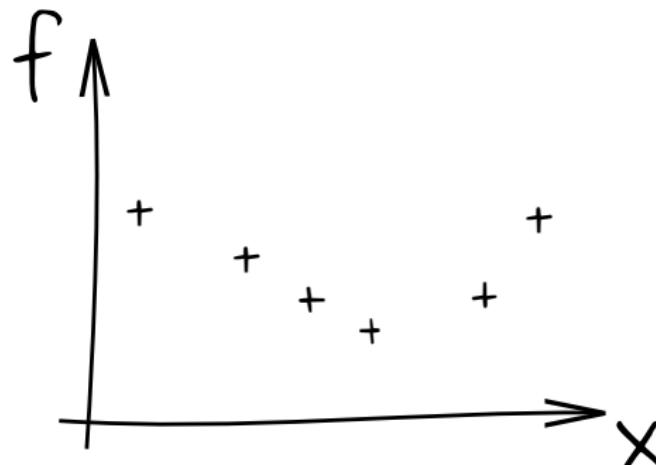
The fact that  $f$  is **costly to evaluate** changes a lot of things...

1. Representing the function is not possible...



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

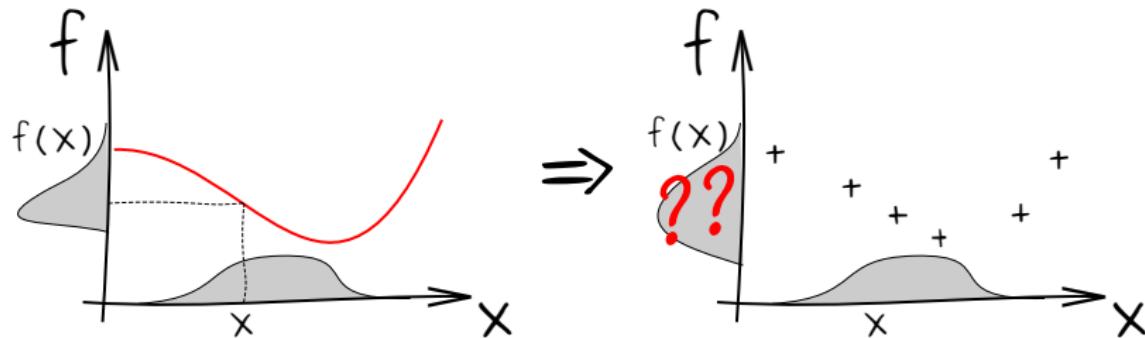
2. Computing integrals is not possible...



What is the mean value of  $f$ ?

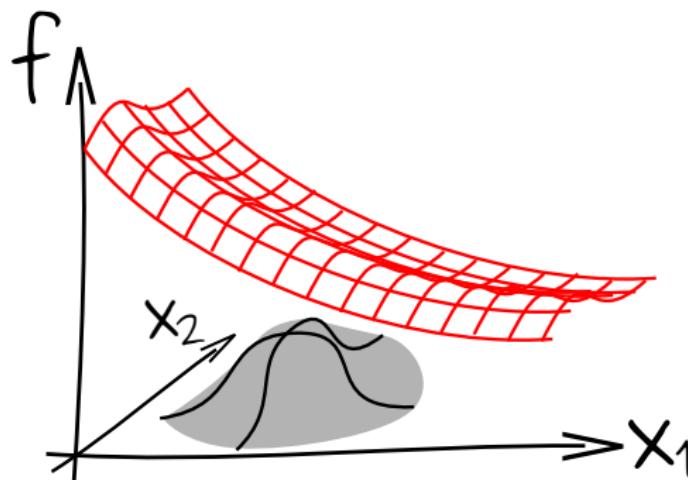
The fact that  $f$  is **costly to evaluate** changes a lot of things...

### 3. Uncertainty propagation is not possible...



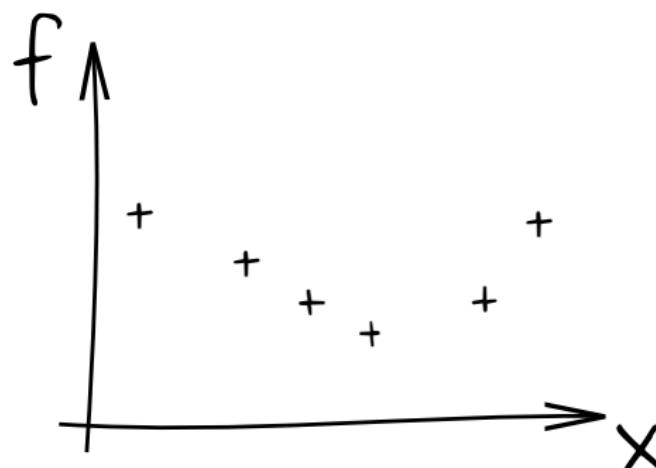
The fact that  $f$  is **costly to evaluate** changes a lot of things...

#### 4. Sensitivity analysis is not possible...



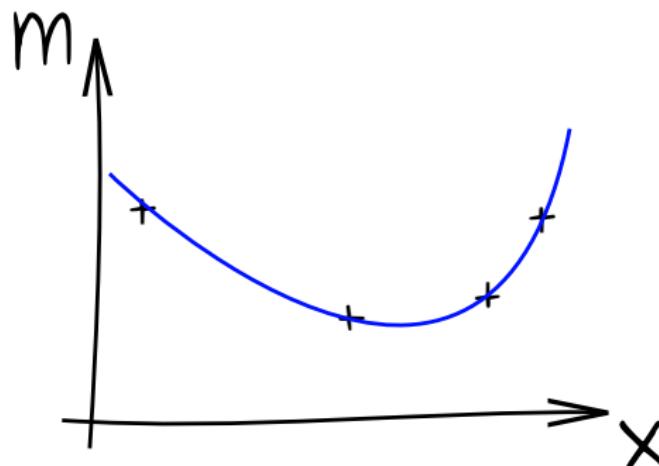
The fact that  $f$  is **costly to evaluate** changes a lot of things...

## 5. Optimisation is also tricky...



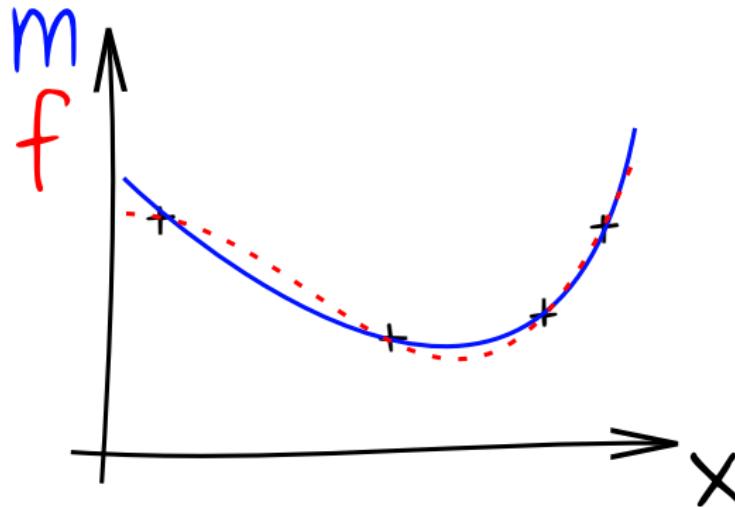
# Statistical models

The principle of statistical modelling is to use the data to build a mathematical approximation of the function.



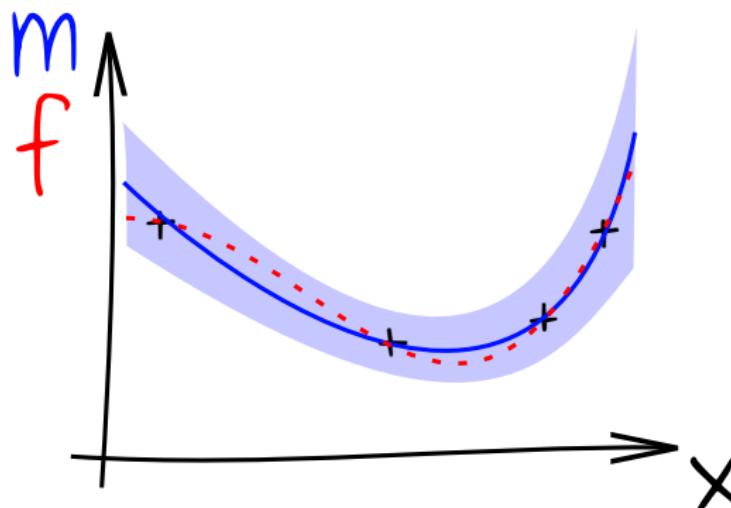
The model can then be used to answer all previous questions

Of course, there is a difference between  $f$  and  $m$ ...



## Why **statistical models**?

We want to be able to quantify the model error:



The confidence intervals can be used to obtain a **measure of uncertainty on the value of interest**.

In the sequel, we will use the following notations :

- The set of observation points will be represented by a  $n \times d$  matrix  $X = (X_1, \dots, X_n)^t$
- The vector of observations will be denoted by  $F : F_i = f(X_i)$  (or  $F = f(X)$ ).

We will now discuss two types of statistical models:

- Linear regression
- Gaussian process regression

# Linear Regression

**Linear regression** is probably the most commonly used statistical model.

Given a set of basis functions  $B = (b_0, \dots, b_p)$ , we assume that the observations come from the probabilistic model

$$F = B(X)\beta + \varepsilon \quad \left( \text{i.e. } F_i = \sum_{k=1}^p \beta_k b_k(X_i) + \varepsilon_i \right)$$

where the vector  $\beta$  is unknown,  $\varepsilon_i$  are independent and identically distributed and

$$B(X)_{(i,j)} = b_j(X_i)$$

If we consider a model of the form

$$m(x) = B(x)\hat{\beta}$$

the prediction error (Residual Sum of Square) is given by

$$RSS = (B(X)\hat{\beta} - F)^t (\hat{\beta} B(X) - F) \quad \left( \text{i.e. } \sum_{k=1}^n (B(X_i)\hat{\beta} - F_i)^2 \right)$$

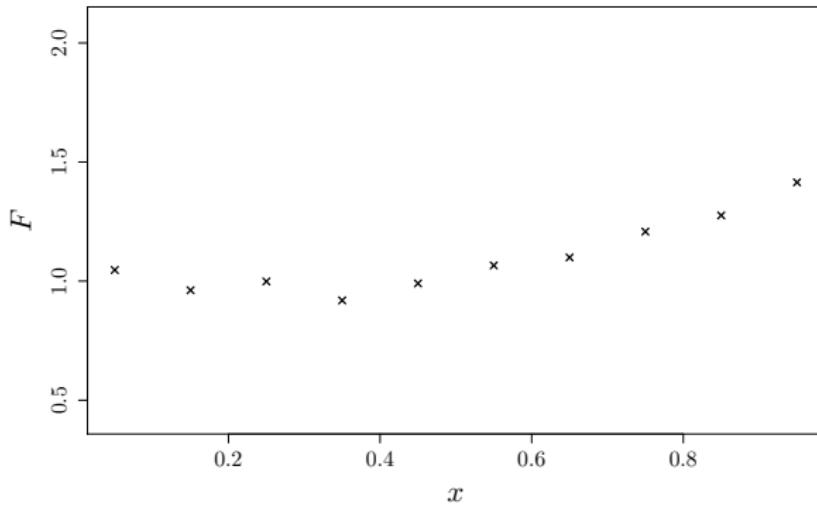
Finding the optimal value of  $\hat{\beta}$  means minimizing a quadratic form.  
This can be done analytically and we obtain  
 $\hat{\beta} = (B(X)^t B(X))^{-1} B(X)^t F.$

The associated linear regression model is thus

$$m(x) = B(x)(B(X)^t B(X))^{-1} B(X)^t F.$$

## Example

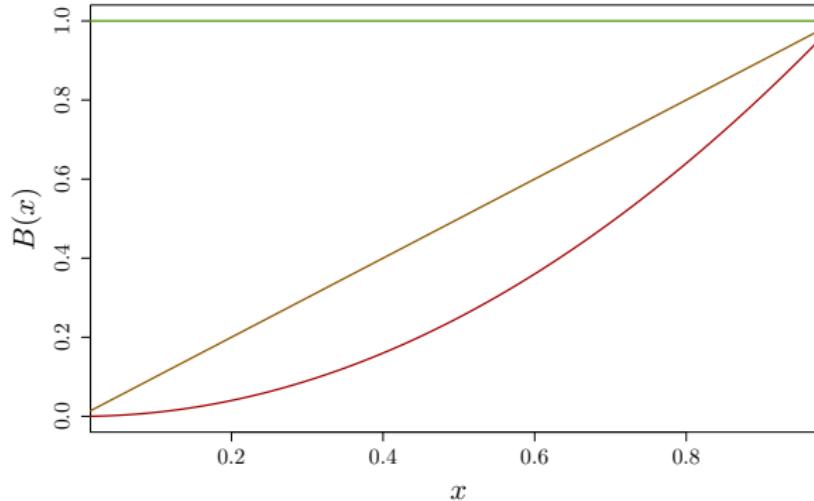
If we consider the following observations:



## Example

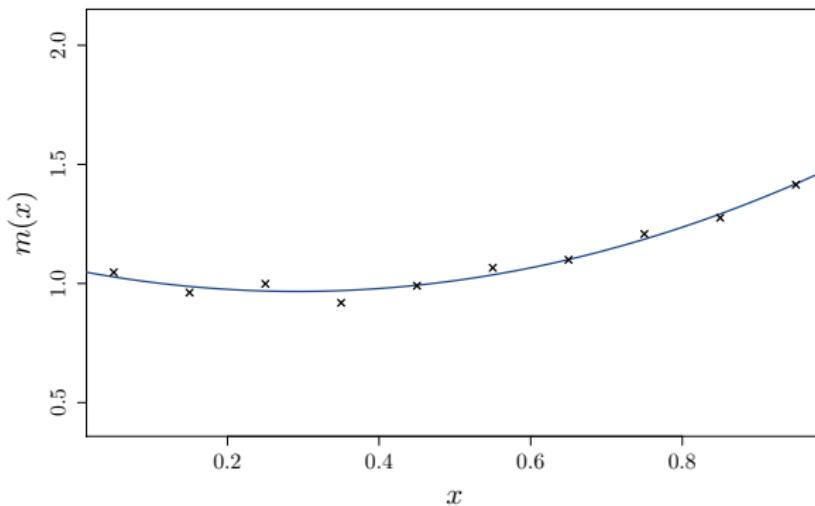
and a set of 3 basis functions:

$$b_0(x) = 1, \quad b_1(x) = x, \quad b_2(x) = x^2$$



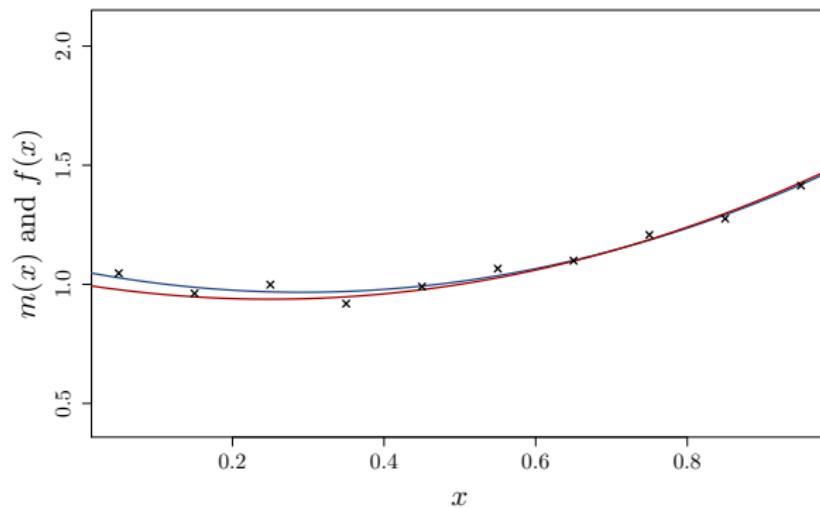
## Example

We obtain  $\hat{\beta} = (1.06, -0.61, 1.04)$  and the model is:



## Example

There is of course an error between the true generative function and the model



Can this error be quantified?

The initial assumption is  $F = B(X)\beta + \varepsilon$  and we have computed an estimator of  $\beta$ :

$$\hat{\beta} = (B(X)^t B(X))^{-1} B(X)^t F.$$

$\hat{\beta}$  can thus be seen as a sample from the random variable:

$$\hat{\beta} = (B(X)^t B(X))^{-1} B(X)^t (B(X)\beta + \varepsilon).$$

What about the distribution of  $\hat{\beta}$ ?

The initial assumption is  $F = B(X)\beta + \varepsilon$  and we have computed an estimator of  $\beta$ :

$$\hat{\beta} = (B(X)^t B(X))^{-1} B(X)^t F.$$

$\hat{\beta}$  can thus be seen as a sample from the random variable:

$$\hat{\beta} = (B(X)^t B(X))^{-1} B(X)^t (B(X)\beta + \varepsilon).$$

What about the distribution of  $\hat{\beta}$ ?

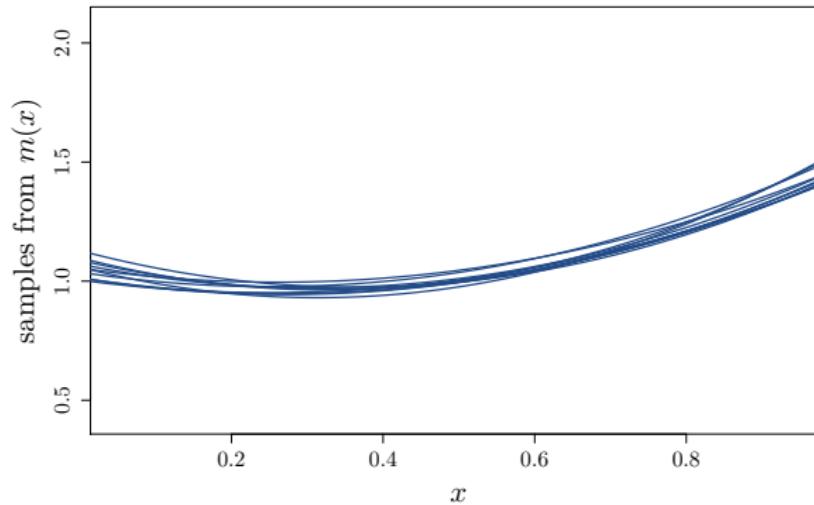
- Its expectation is  $\beta \Rightarrow$  The estimator is unbiased
- Its covariance matrix is

$$(B(X)^t B(X))^{-1} B(X)^t \text{cov}[\varepsilon, \varepsilon^t] B(X) (B(X)^t B(X))^{-1}$$

- If  $\varepsilon$  is multivariate normal, then  $\hat{\beta}$  is also multivariate normal (not necessarily i.i.d.).

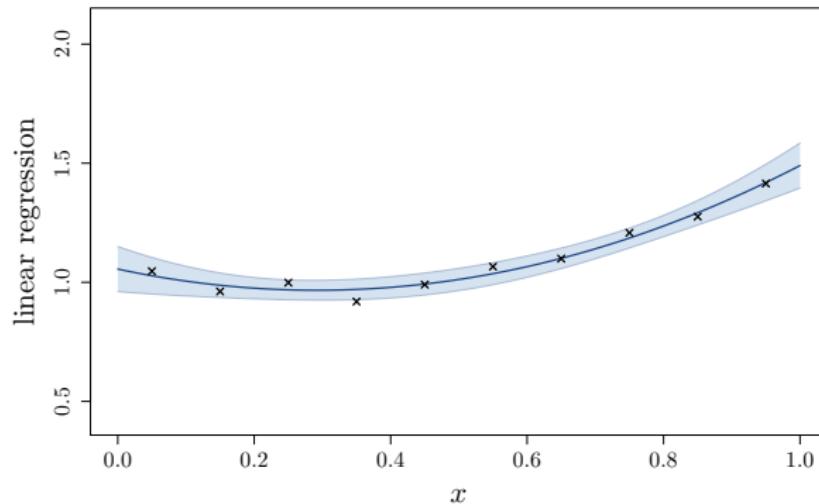
Sampling in the distribution of  $\hat{\beta}$  gives us a large variety of models which represent the uncertainty about our estimation:

## Back to the example



## Back to the example

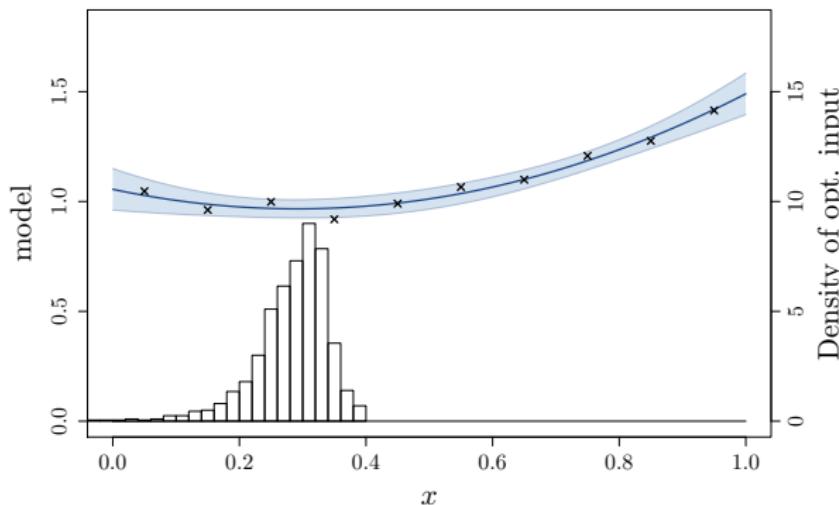
The previous picture can be summarized by showing the mean of  $m$  and 95% confidence intervals



Knowing the uncertainty on the model allows to compute an uncertainty on the quantity of interest.

## Back to the example

For example, if we are interested in the value  $x^*$  minimizing  $f(x)$ :

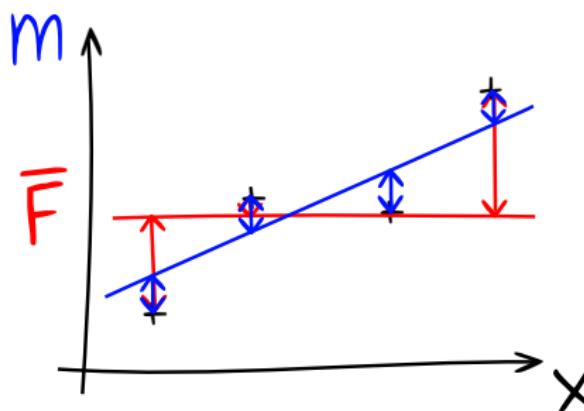


The expectation of  $x^*$  is not the input minimizing  $m(x)$ .

**Model validation** is always of upper importance.

The goodness of fit can be measured by the **coefficient of determination**:

$$R^2 = 1 - \frac{\text{var}[\text{prediction errors}]}{\text{var}[\text{data}]} = 1 - \frac{\sum_i (F_i - m(X_i))^2}{\sum_i (F_i - \text{mean}(F))^2}$$



**Be careful!** A good  $R^2$  (ie close to one) does not necessarily mean that the model is good.

We've seen that the law of  $\hat{\beta}$  is multivariate Gaussian, with its covariance depending on  $\varepsilon$ :

$$(B(X)^t B(X))^{-1} B(X)^t \text{cov}[\varepsilon, \varepsilon^t] B(X) (B(X)^t B(X))^{-1}$$

**Ordinary regression:** suppose now  $\text{cov}[\varepsilon, \varepsilon^t] = \sigma^2 I_d$  then

$$\text{cov}[\hat{\beta}, \hat{\beta}^t] = \sigma^2 (B(X)^t B(X))^{-1}$$

and from this, we now that

$$\frac{\hat{\beta}_i}{\sqrt{\hat{\sigma}^2 (B(X)^t B(X))_{(i,i)}^{-1}}} \text{ is student}$$

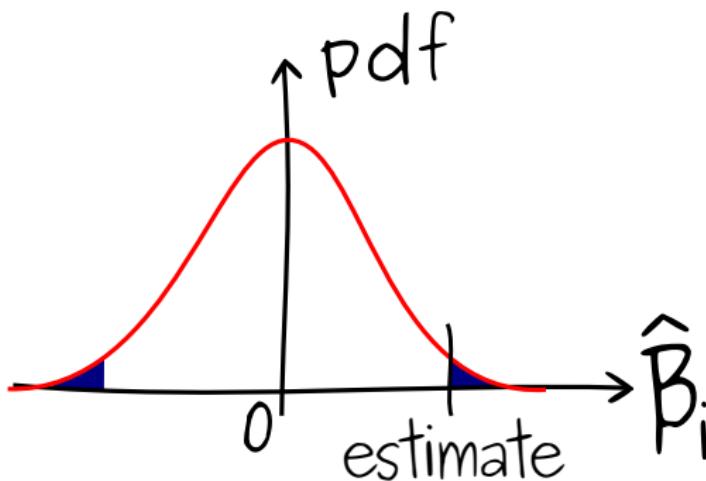
with the standard variance estimator

$$\hat{\sigma}^2 = \frac{1}{n-p} \|F - m(X)\|^2$$

The influence of one basis function can be tested using **p-values**:

$H_0$ : The basis function  $b_i$  has no influence (i.e.  $\beta_i = 0$ )

p-values: probability of observing a larger estimate



The smaller the p-value is, the less  $\hat{\beta}$  is likely to be due to chance.

We could dedicate the entire course to linear regression models...

- model validation
- choice of basis functions
- influence of input locations
- ...

We will just stress a few **pros and cons of these models:**

- + provide a good noise filtering
- + are easy to interpret
- are not flexible (need to choose the basis functions)
- do not interpolate
- may explode when using high order polynomials (over-fitting)

# Gaussian Process Regression

This section will be organised in 3 subsections:

1. Reminders on Multivariate normal distribution
2. Gaussian processes
3. Gaussian process regression

# 1. Multivariate normal distribution

The usual one dimensional normal distribution  $\mathcal{N}(\mu, \sigma^2)$  has the following pdf:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) \text{ for } x \in \mathbb{R}$$

It can be generalised to vectors:

## Definition

We say that a vector  $Y = (Y_1, \dots, Y_n)$  follows a multivariate normal distribution if any linear combination of  $Y$  follows a normal distribution:

$$\forall \alpha \in \mathbb{R}^n, \alpha^t Y \sim \mathcal{N}(m, s^2)$$

The distribution of a Gaussian vector is characterised by

- a mean vector  $\mu = (\mu_1, \dots, \mu_d)$
- a  $d \times d$  covariance matrix  $\Sigma$  :  $\Sigma_{i,j} = \text{cov}(Y_i, Y_j)$

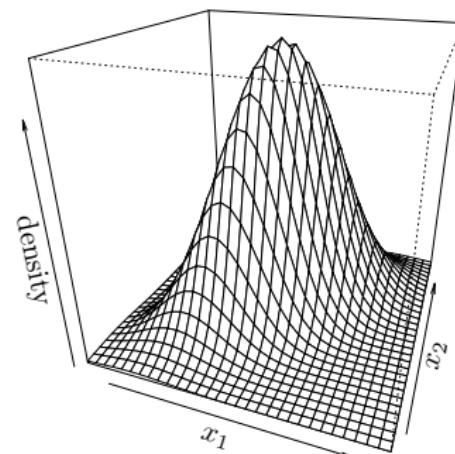
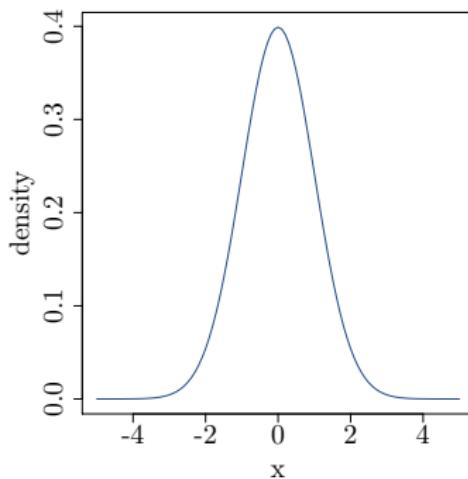
### Property:

A covariance matrix is

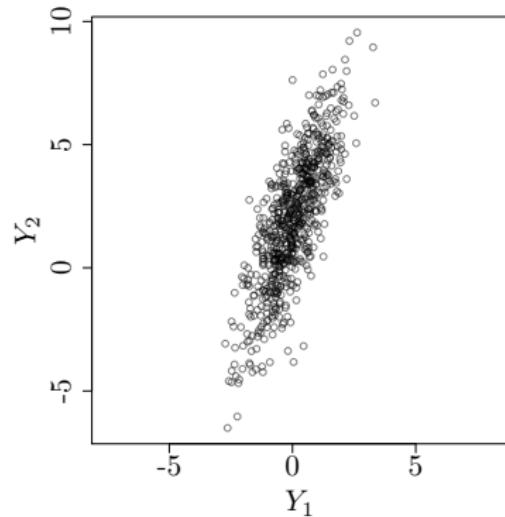
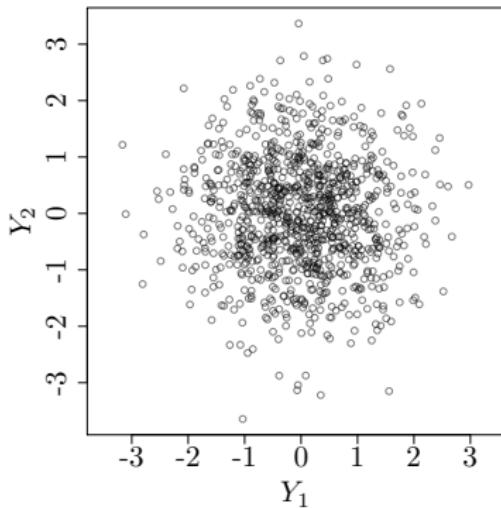
- **symmetric**  $K_{i,j} = K_{j,i}$
- **positive semi-definite**  $\forall \alpha \in \mathbb{R}^d, \alpha^t K \alpha \geq 0.$

The density of a multivariate Gaussian is:

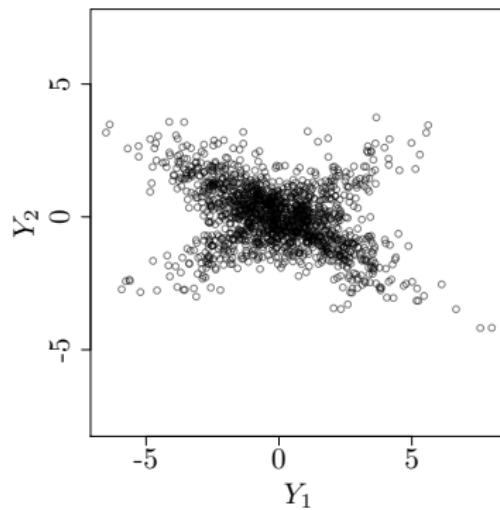
$$f_Y(x) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu)^t \Sigma^{-1} (x - \mu)\right).$$



## Example



## Counter example



$Y_1$  and  $Y_2$  are normally distributed but **the couple** ( $Y_1, Y_2$ ) is not.

## Conditional distribution

Let  $(Y, Z)$  be a Gaussian vector ( $Y$  and  $Z$  may both be vectors) with mean  $(\mu_Y, \mu_Z)^t$  and covariance matrix

$$\begin{pmatrix} \text{cov}(Y, Y) & \text{cov}(Y, Z) \\ \text{cov}(Z, Y) & \text{cov}(Z, Z) \end{pmatrix}.$$

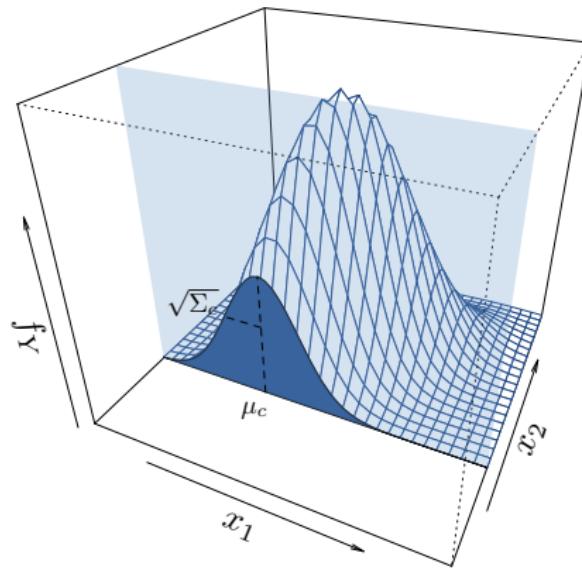
The conditional distribution of  $Y$  knowing  $Z$  is still multivariate normal  $Y|Z \sim \mathcal{N}(\mu_{cond}, \Sigma_{cond})$  with

$$\mu_{cond} = E[Y|Z] = \mu_Y + \text{cov}(Y, Z) \text{cov}(Z, Z)^{-1}(Z - \mu_Z)$$

$$\Sigma_{cond} = \text{cov}[Y, Y|Z] = \text{cov}(Y, Y) - \text{cov}(Y, Z) \text{cov}(Z, Z)^{-1} \text{cov}(Z, Y)$$

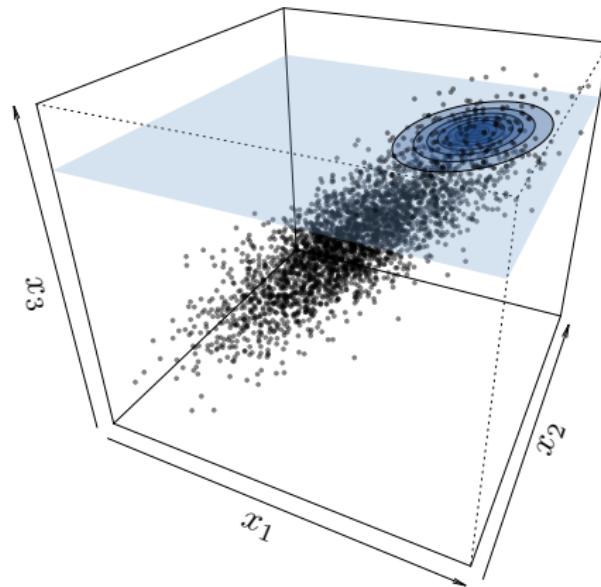
## Example 1

2D multivariate Gaussian conditional distribution:



## Example 2

3D multivariate Gaussian conditional distribution:



## Exercise

Starting from the density function, prove the previous property using the Schur bloc inverse:

$$\begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{2,1} & \Sigma_{2,2} \end{pmatrix}^{-1} = \begin{pmatrix} A & B \\ B^t & C \end{pmatrix}$$

where:  $A = (\Sigma_{1,1} - \Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1})^{-1}$

$$B = -(\Sigma_{1,1} - \Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1})^{-1}\Sigma_{1,2}\Sigma_{2,2}^{-1}$$

$$C = \Sigma_{2,2}^{-1} + \Sigma_{2,2}^{-1}\Sigma_{2,1}(\Sigma_{1,1} - \Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1})^{-1}\Sigma_{1,2}\Sigma_{2,2}^{-1}$$

## 2. Gaussian processes

The multivariate Gaussian distribution can be generalised to random processes:

### Definition

A random process  $Z$  over  $D \subset \mathbb{R}^d$  is said to be Gaussian if

$\forall n \in \mathbb{N}, \forall x_i \in D, (Z(x_1), \dots, Z(x_n))$  is a Gaussian vector.

The distribution of a GP is fully characterised by:

- its mean function  $m$  defined over  $D$
- its covariance function (or kernel)  $k$  defined over  $D \times D$ :  
 $k(x, y) = \text{cov}(Z(x), Z(y))$

We will use the notation  $Z \sim \mathcal{N}(m(.), k(., .))$ .

A kernel satisfies the following properties:

- It is symmetric:  $k(x, y) = k(y, x)$
- 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.

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

constant  $k(x, y) = 1$

white noise  $k(x, y) = \delta_{x,y}$

Brownian  $k(x, y) = \min(x, y)$

exponential  $k(x, y) = \exp(-|x - y|)$

Matern 3/2  $k(x, y) = (1 + |x - y|) \exp(-|x - y|)$

Matern 5/2  $k(x, y) = (1 + |x - y| + 1/3|x - y|^2) \exp(-|x - y|)$

squared exponential  $k(x, y) = \exp(-(x - y)^2)$

⋮

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

Can we look at the sample paths associated to these kernels?

In order to simulate sample paths from a GP  $Z \sim \mathcal{N}(m(.), k(., .))$ , we will consider samples of the GP discretised on a fine grid.

### Exercise: Simulating sample paths

Let  $X$  be a set 100 regularly spaced points over the input space of  $Z$ .

- What is the distribution of  $Z(X)$  ?
- How to simulate samples from  $Z(X)$  ?

⇒ Shiny App

Furthermore, we can include some scaling parameters into the kernels:

**Exercise:**

If  $Z$  is a GP  $\mathcal{N}(0, k(., .))$ , what is the distribution of  
 $Y(x) = \sigma Z(x/\theta)$  ?

$\sigma^2$  is called the **variance** and  $\theta$  the **length-scale**

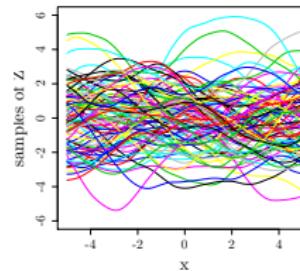
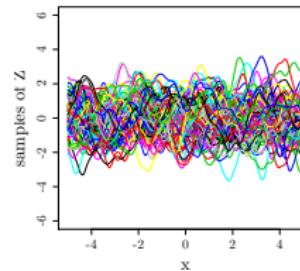
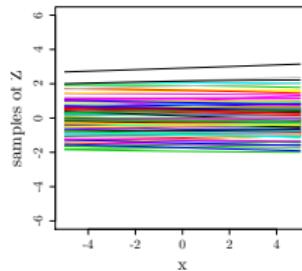
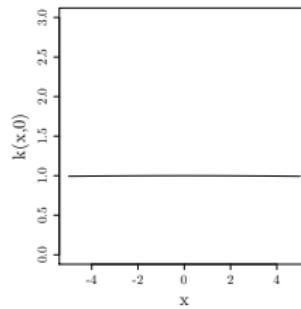
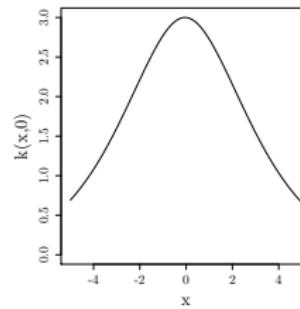
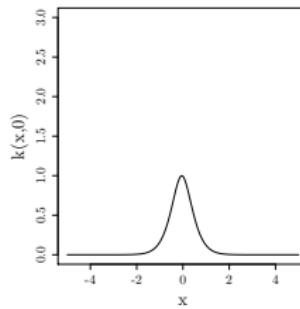
## Exercise:

The kernel is Matern 5/2. Can you put each line in the right order?

$$(\sigma^2, \theta) = (3, 3)$$

$$(\sigma^2, \theta) = (1, 0.5)$$

$$(\sigma^2, \theta) = (1, 50)$$



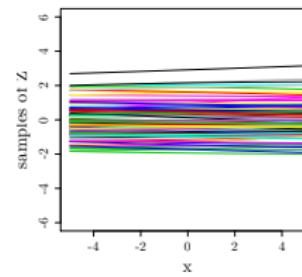
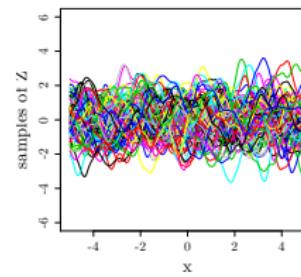
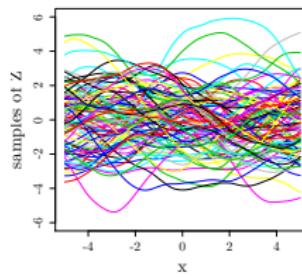
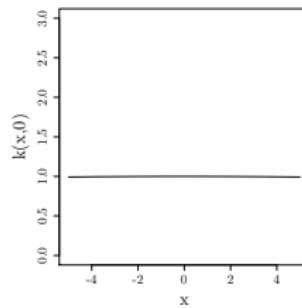
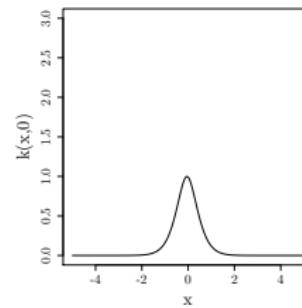
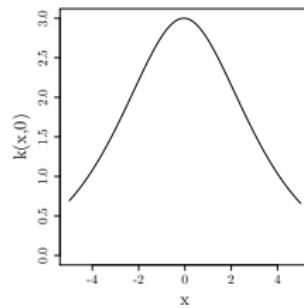
## Exercise:

Answer is:

$$(\sigma^2, \theta) = (3, 3)$$

$$(\sigma^2, \theta) = (1, 0.5)$$

$$(\sigma^2, \theta) = (1, 50)$$



In higher dimension one can introduce one length-scale parameter per dimension. The usual Euclidean distance between two points  $\|x - y\| = (\sum(x_i - y_i)^2)^{1/2}$  is thus replaced by

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

If the parameters  $\theta_i$  are equal for all the dimensions, the covariance (or the process) is called **isotropic**.

Here is a list of the most common kernels:

constant     $k(x, y) = \sigma^2$

white noise     $k(x, y) = \sigma^2 \delta_{x,y}$

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

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

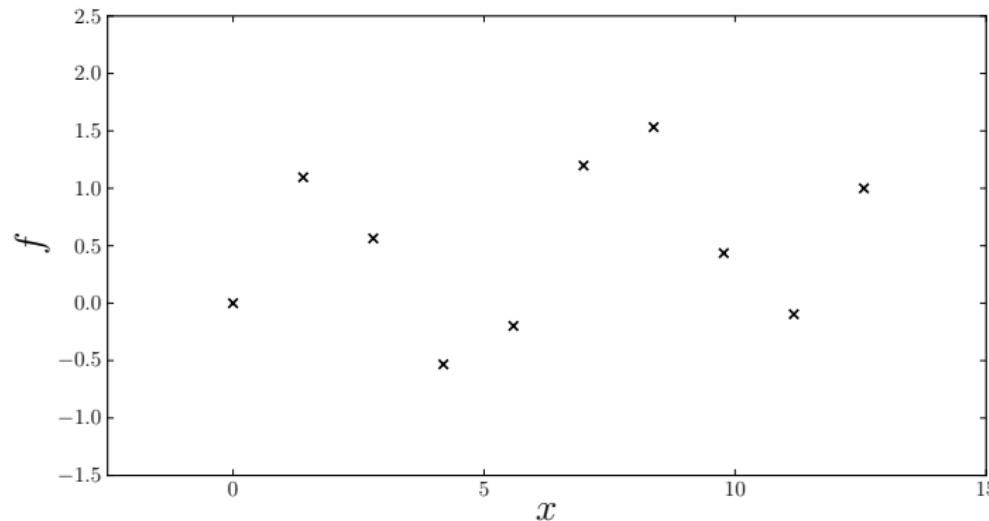
Matern 5/2     $k(x, y) = \sigma^2 \left(1 + \sqrt{5}\|x - y\|_\theta + \frac{5}{3}\|x - y\|_\theta^2\right) \exp(-\sqrt{5}\|x - y\|_\theta)$

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

### 3. Gaussian process regression

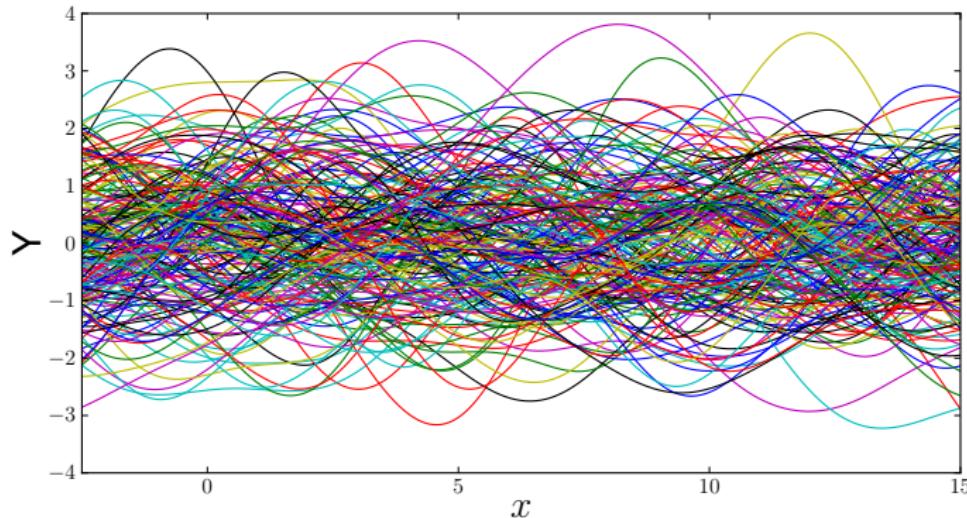
We assume we have observed a function  $f$  for a set of points

$$X = (X_1, \dots, X_n):$$



The vector of observations is  $F = f(X)$  (ie  $F_i = f(X_i)$  ).

Since  $f$  is unknown, we make the general assumption that it is the sample path of a Gaussian process  $Z \sim \mathcal{N}(0, k)$ :



What would be the next step?

We can look at the conditional distribution of  $Z$  knowing that it interpolates the data points:

## Exercise

1. What is the conditional distribution of  $Z(x)|Z(X) = F$ ?
2. Compute the conditional mean  $m$  and covariance  $c(.,.)$ .
3. Compute  $m(X_1)$  and  $c(X_1, X_1)$ .

## Solution

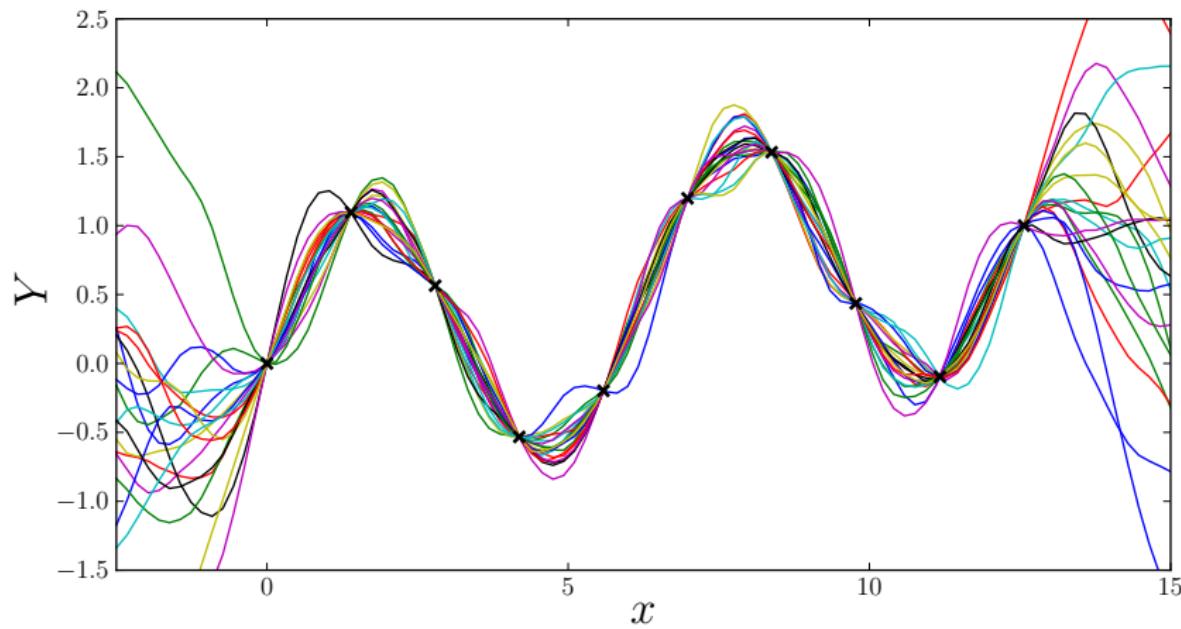
1. The conditional distribution is Gaussian.
2. It has mean and variance

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

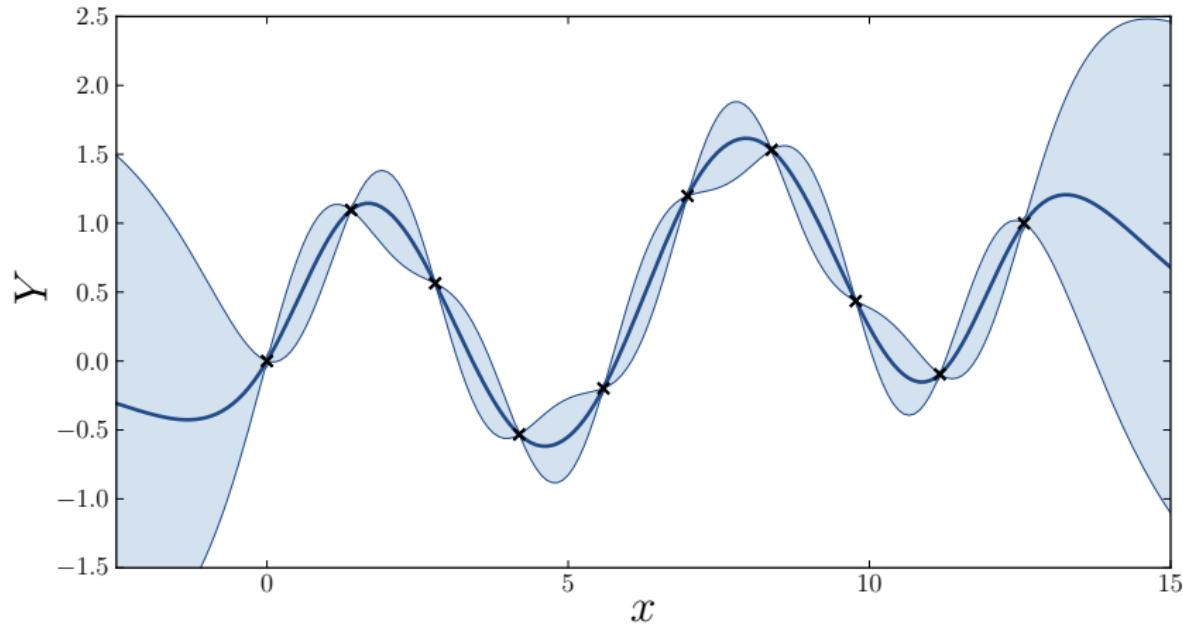
$$\begin{aligned}c(x, y) &= \text{cov}[Z(x), Z(y)|Z(X)=F] \\&= k(x, y) - k(x, X)k(X, X)^{-1}k(X, y)\end{aligned}$$

3. We have  $m(X_1) = F_1$  and  $c(X_1, X_1) = 0$

We can look at sample paths from the conditional distribution



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



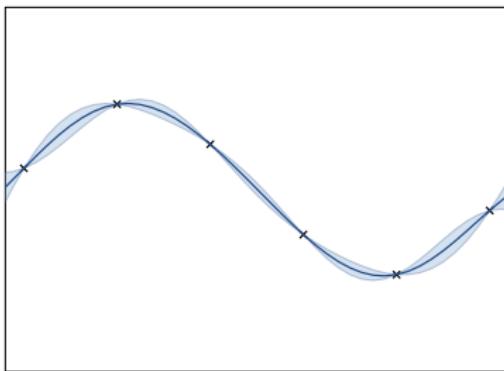
## A few remarkable properties of GPR models

- They (can) interpolate the data-points
- The prediction variance does not depend on the observations
- The mean predictor does not depend on the variance
- They (usually) come back to zero when we are far away from the observations.

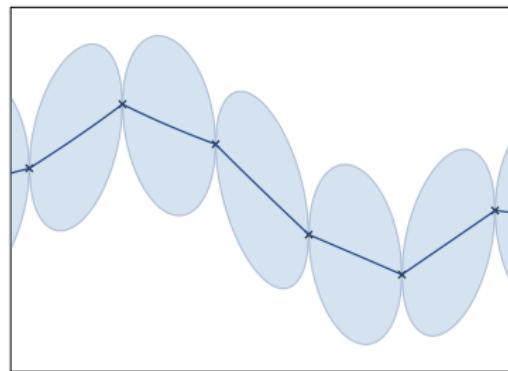
Can you prove them?

Changing the kernel has a huge impact on the model:

Gaussian kernel:

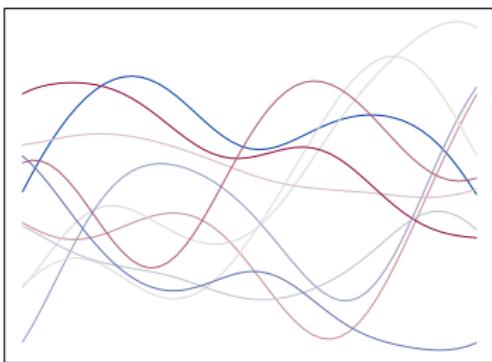


Exponential kernel:

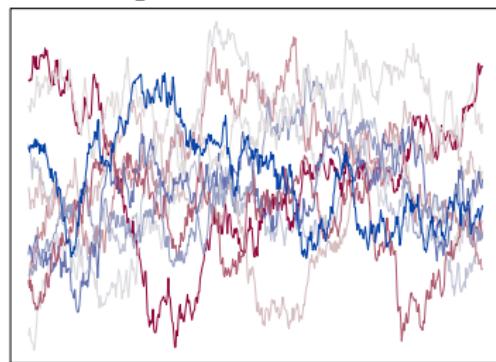


This is because changing the kernel means changing the prior on  $f$

Gaussian kernel:

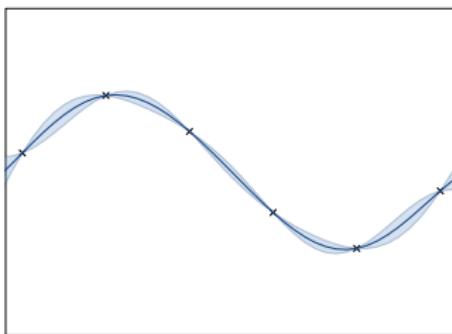


Exponential kernel:

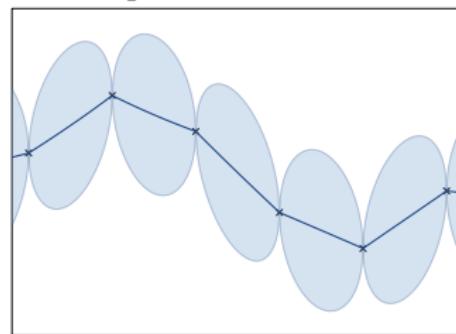


There is no model/kernel that is intrinsically better... it depends on the data!

Gaussian kernel:



Exponential kernel:



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

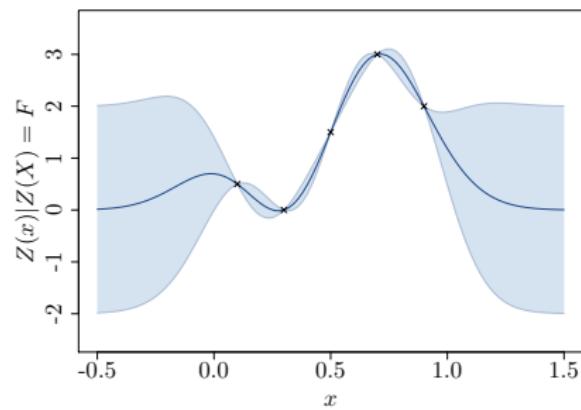
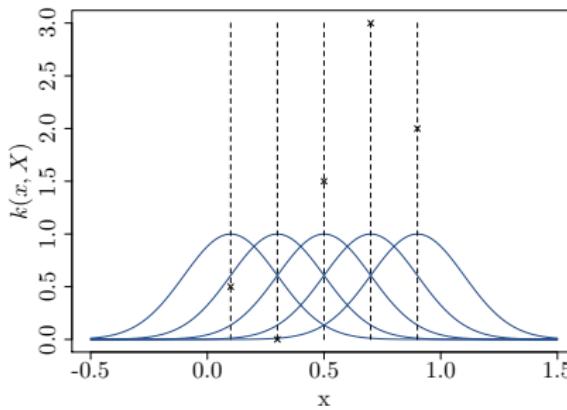
- is it continuous, differentiable, how many times?
- is it stationary ?
- ...

The best predictor can be seen either as a linear combination of

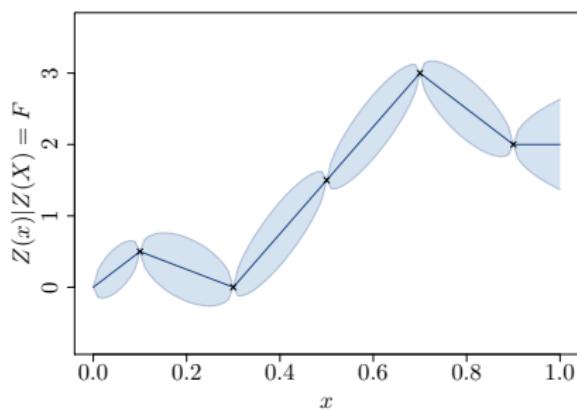
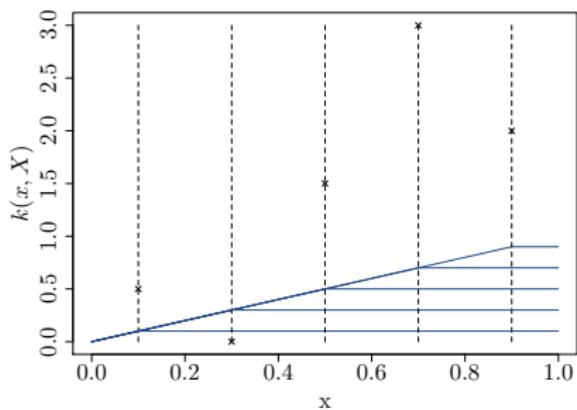
- the observations:  $m(x) = \alpha^t F$
- the kernel evaluated at  $X$ :  $m(x) = k(x, X)\beta$

The later is interesting to understand the model shape and behaviour.

For example, we have for a squared exponential kernel



and for a Brownian kernel:



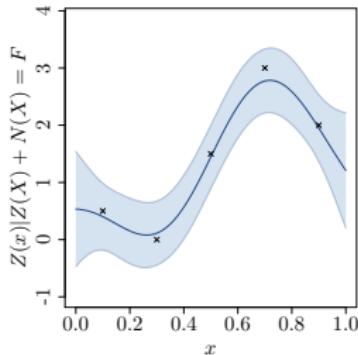
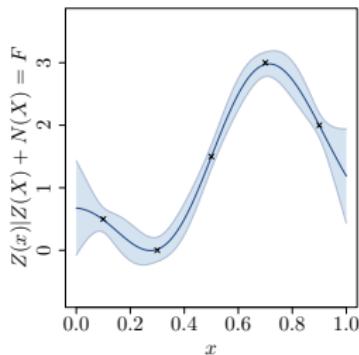
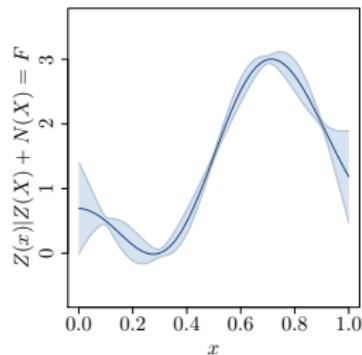
We are not always interested in models that interpolate the data.  
For example, if there is some observation noise:  $F = f(X) + \varepsilon$ .

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

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

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

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



The values of  $\tau^2$  are respectively 0.001, 0.01 and 0.1.

# Conclusion

## Three things to remember:

- 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 an error measure
- GPR is similar to linear regression but the assumption is much weaker (not a finite dimensional space)
- The GPR equations are

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

$$c(x, y) = k(x, y) - k(x, X)k(X, X)^{-1}k(X, y)$$

We still have many things to discuss about such models:

- How to choose the observation points?
- How to validate the model?
- How to estimate the model (ie kernel) parameters?

This will be discussed during the next courses.

## Reference

Carl Edward Rasmussen and Chris Williams, *Gaussian processes for machine learning*, MIT Press, 2006. (free version online).