Short course Statistical Modelling for Optimization – lecture 1/4

Statistical models in engineering

June 2016, Universidad Tecnológica de Pereira (Colombia)

Nicolas Durrande 2 (durrande@emse.fr)
Jean-Charles Croix (jean-charles.croix@emse.fr)
Mines St-Étienne (France)

Introduction

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:

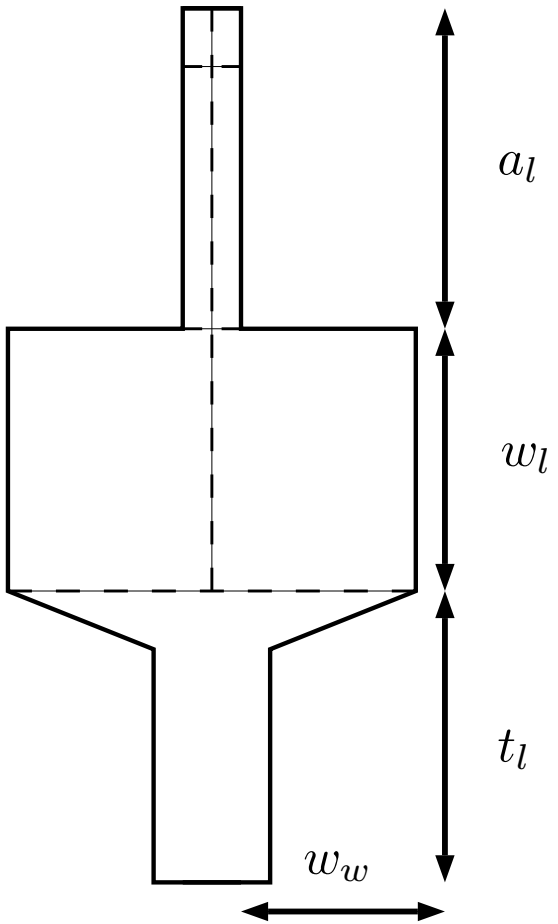
- day 1: Statistical models in engineering
- day 2: Design of experiments
- day 3: Gaussian Process regression
- day 4: Optimization with Gaussian Process Regression

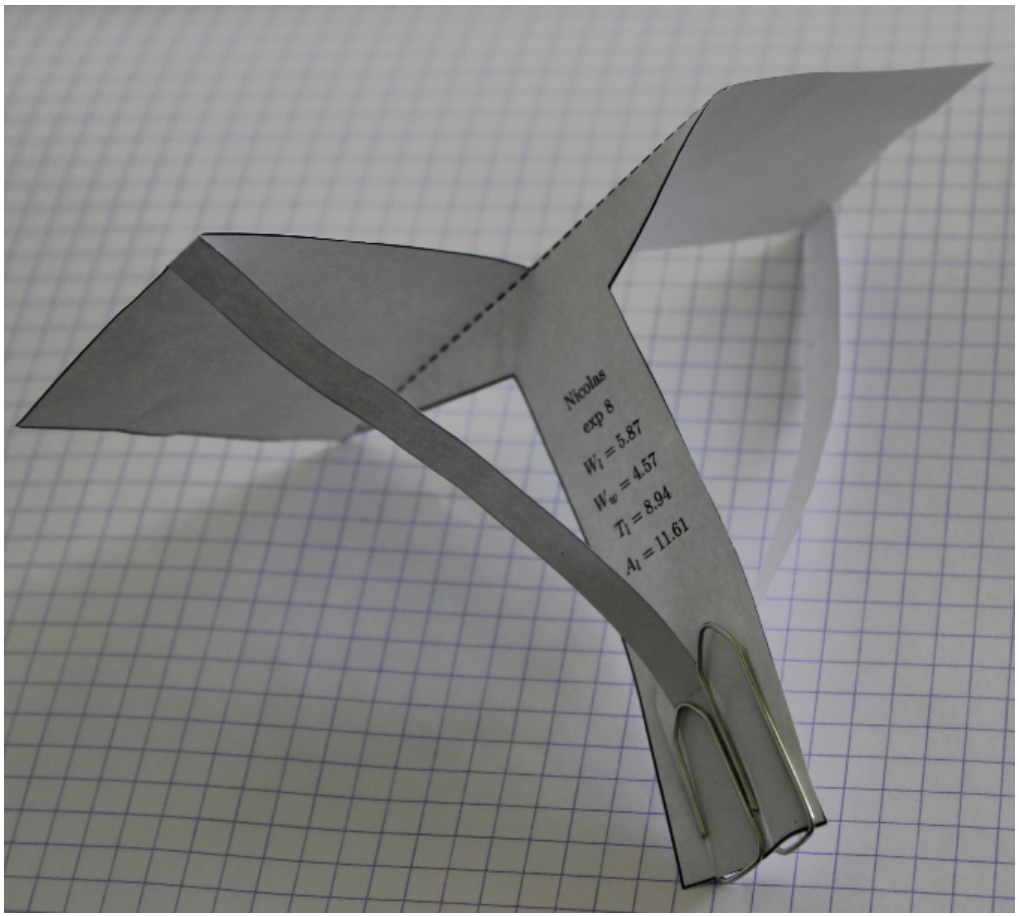
The course material is available

github: user NicolasDurrande

webpage : https://sites.google.com/site/nicolasdurrandehomepage

The lab session will be on the optimization of paper helicopters:





What values of $(a_l,\ w_l,\ t_l,\ w_w)$ give the longest falling time?

This lab is based on a course of Victor Picheny (INRA) at Mines St-Étienne:

> V. Picheny, R. Le Riche, *Revisiting the paper helicopter project using an adaptive surrogate-based approach*, hal-01116601, 2015.

Also the problem is simplistic, it shares a lot with classical engineering problems:

- the physics is complex $\rightarrow$ no analytical solution
- Experiments are "costly" $\rightarrow$ data is limited
- "High dimensional" input space $\rightarrow$ 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 scissors tomorrow!

Outline of today's lecture

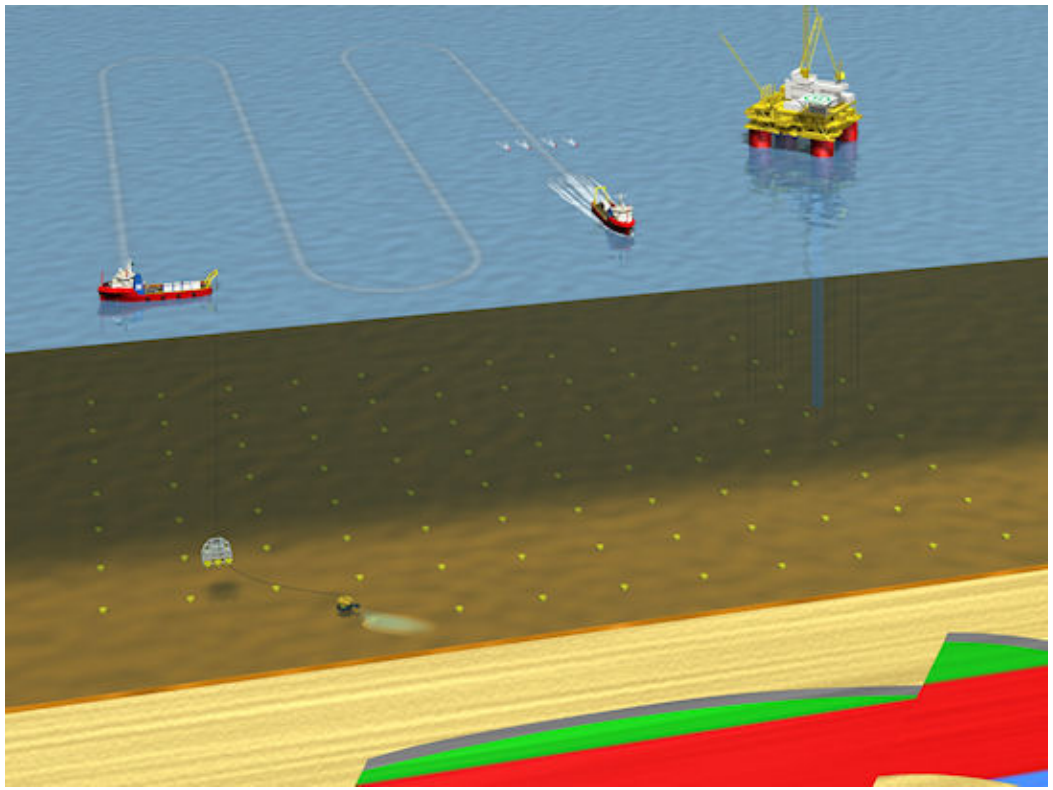
- Why (and when) statistical models can be useful in engineering?
- A basic example: Linear regression.
- Gaussian process regression

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.

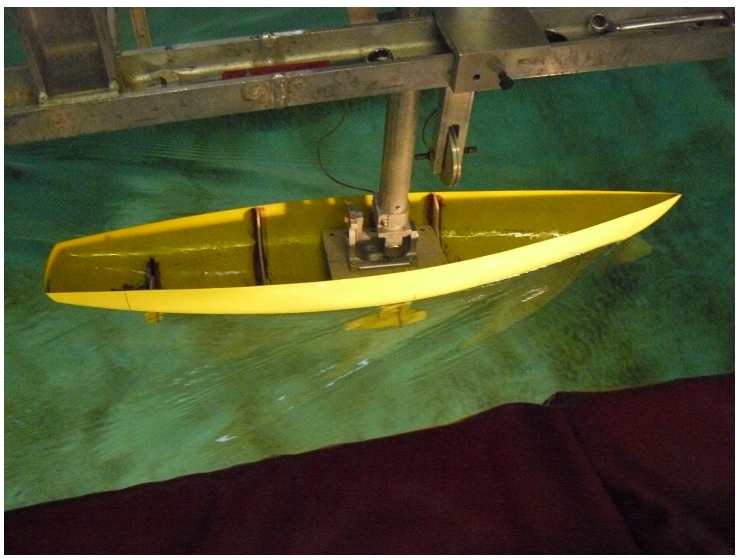
- real world experiments
- destructive tests
- prototyping
- numerical experiments

Example: real world experiments



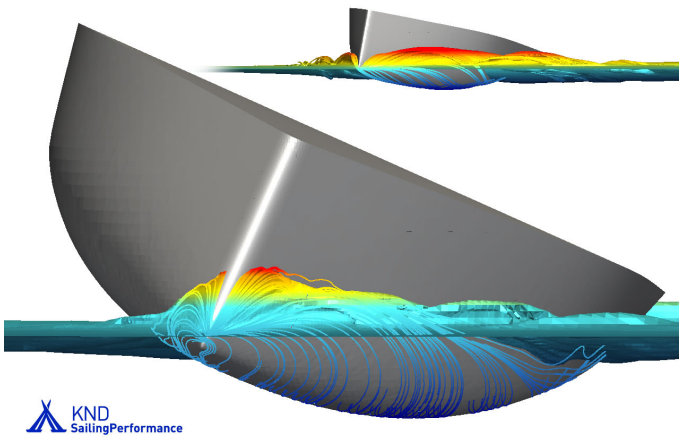
Example: Destructive tests

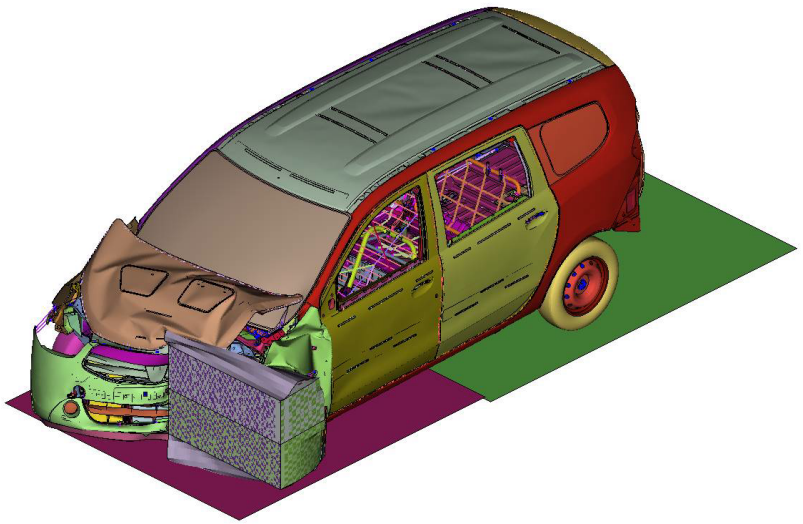
Example: Prototyping of a boat shape



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

$$y = f(x).$$

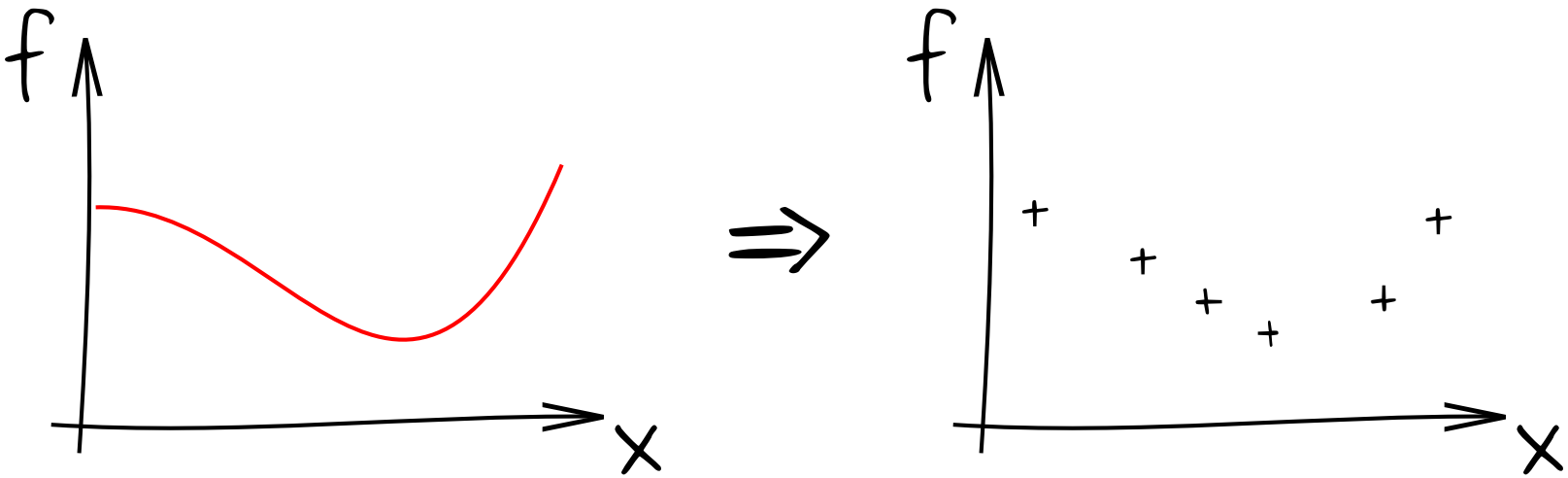
where f is a **costly to evaluate function**.

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.

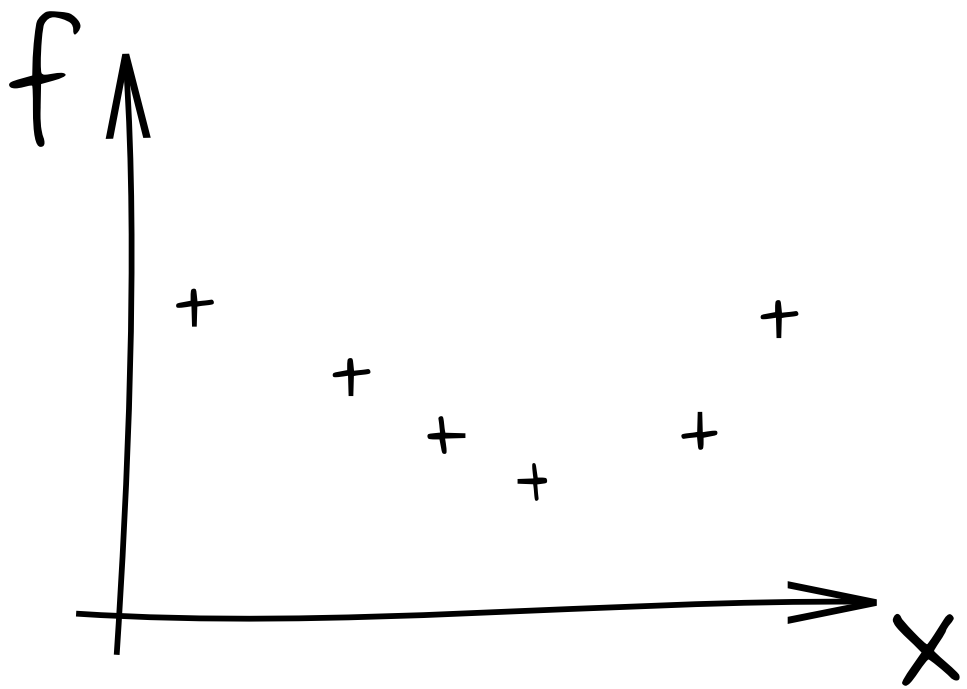
The fact that f is **costly to evaluate** changes a lot of things...

1. Representing the function is not possible...



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

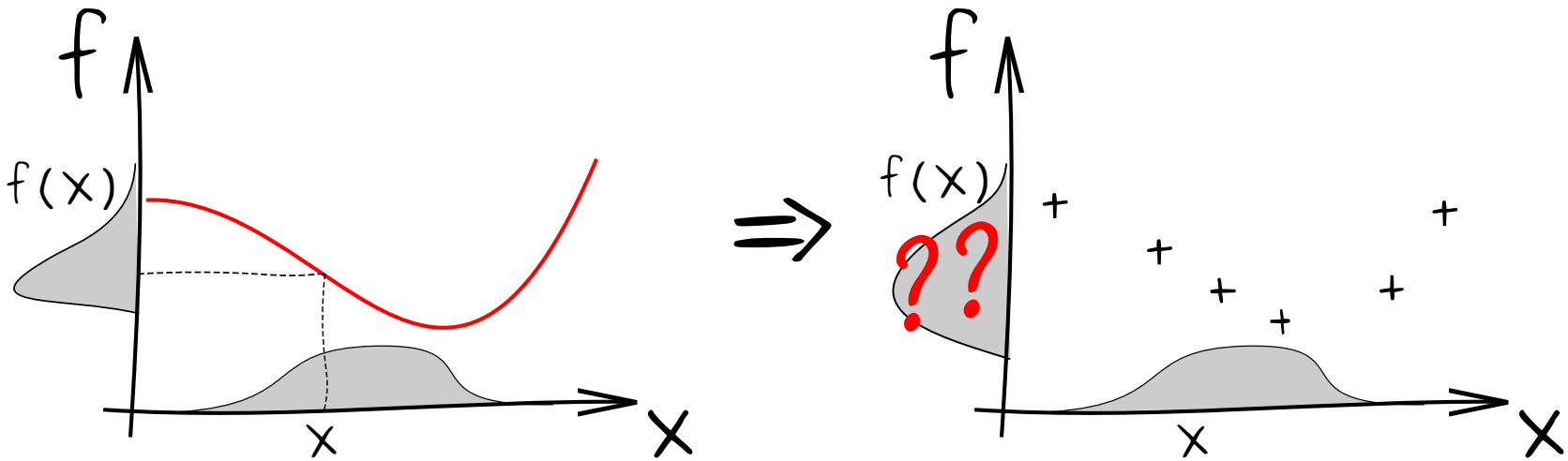
2. Computing integrals is not possible...



What is the mean value of f?

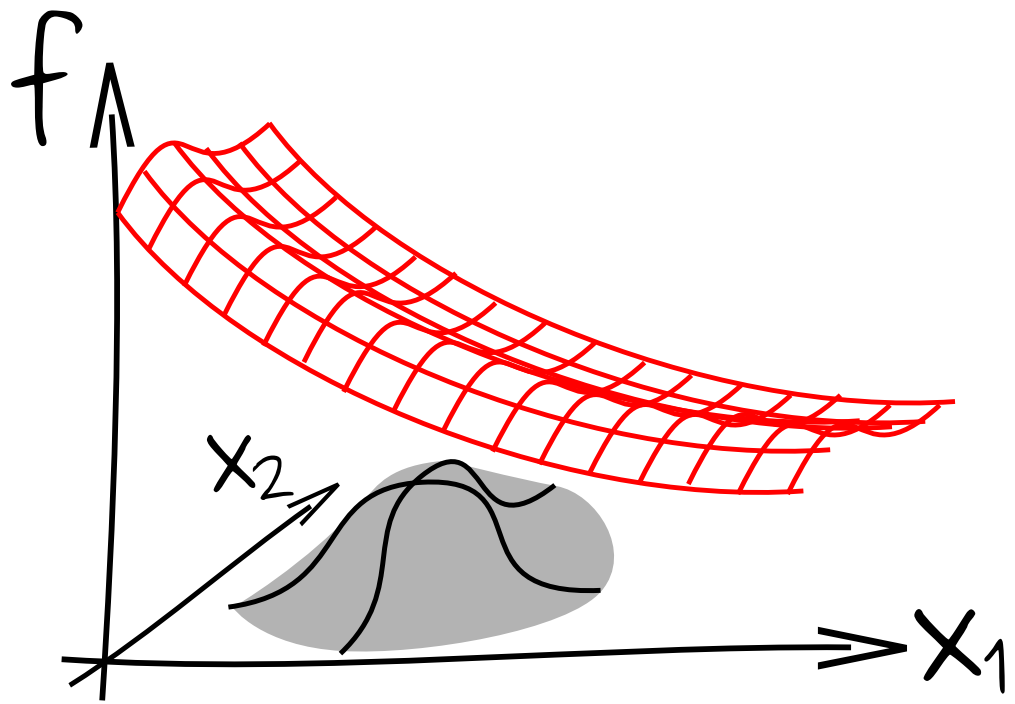
The fact that f is **costly to evaluate** changes a lot of things...

3. Uncertainty propagation is not possible...



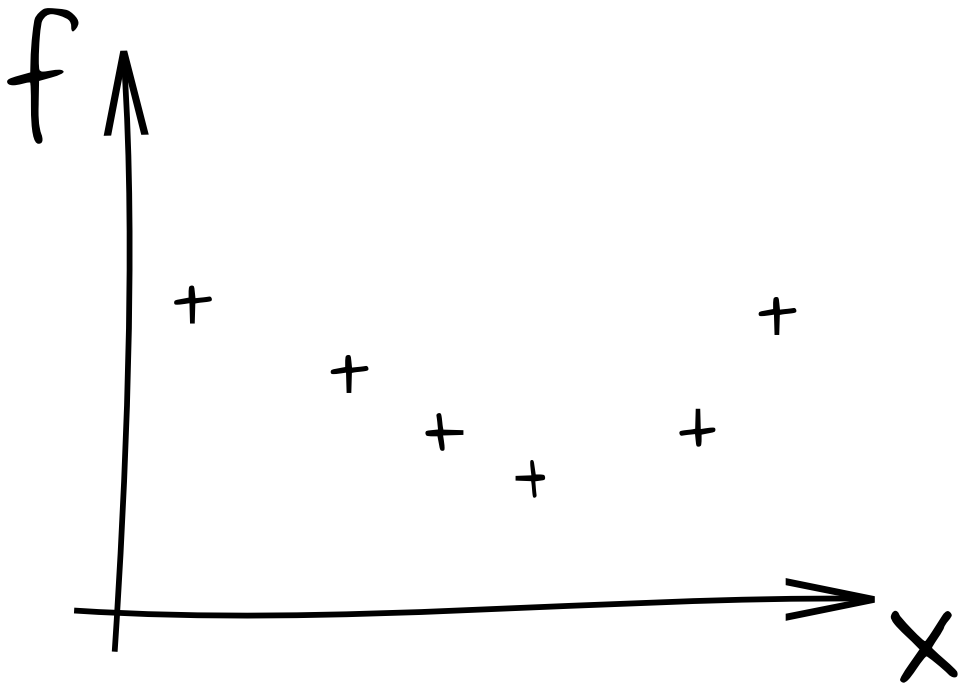
The fact that f is **costly to evaluate** changes a lot of things...

4. Sensitivity analysis is not possible...



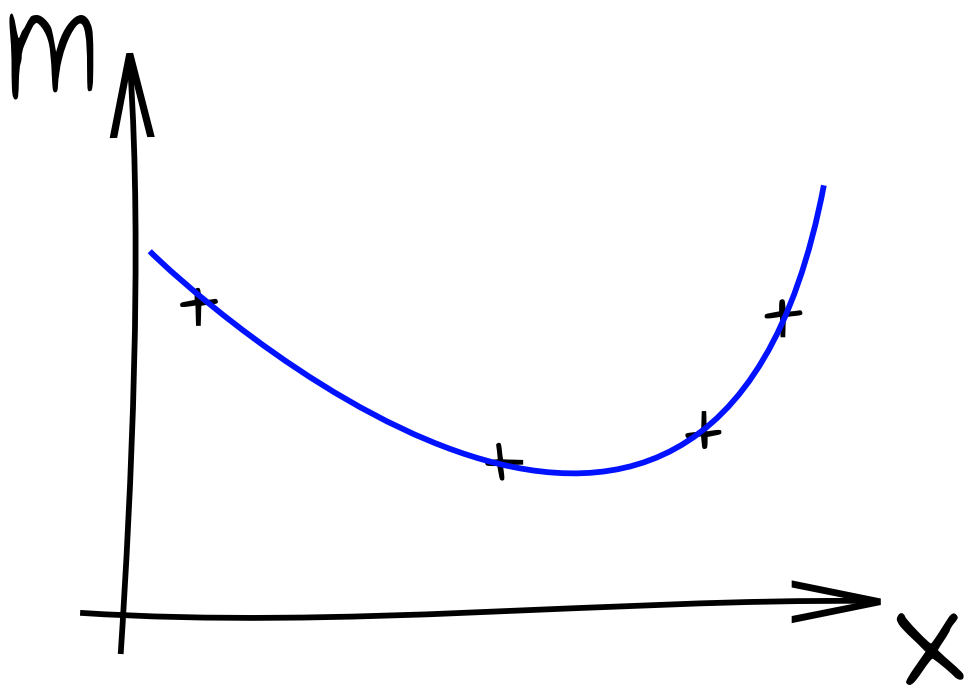
The fact that f is **costly to evaluate** changes a lot of things...

5. Optimisation is also tricky...



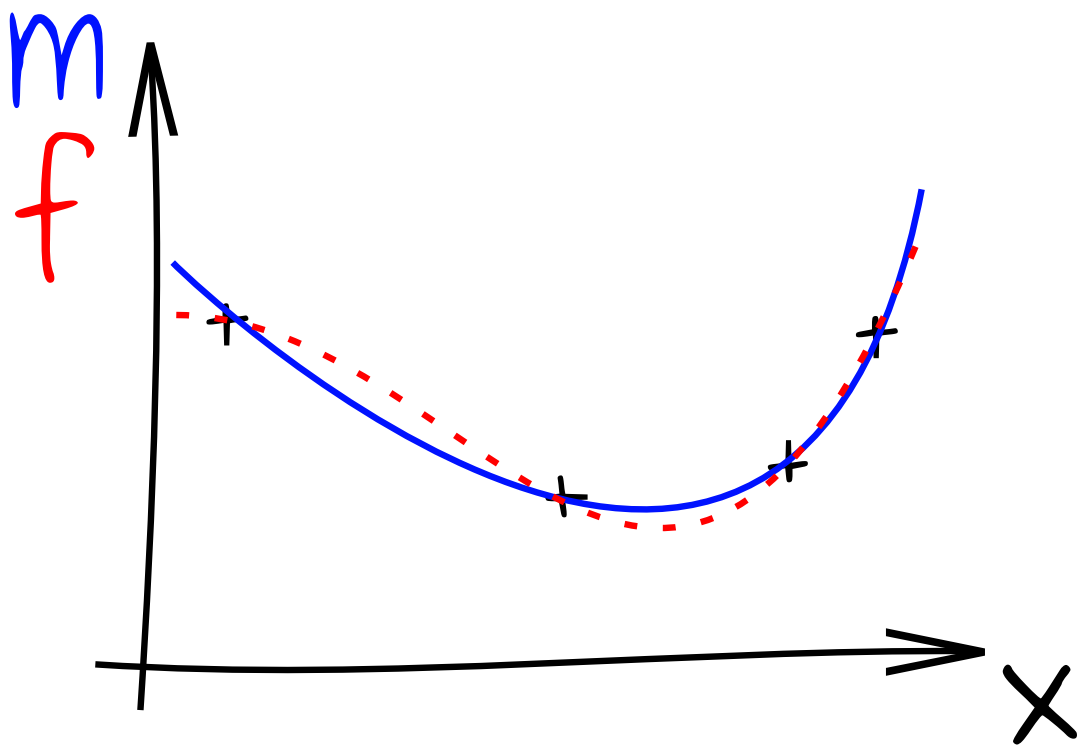
Statistical models

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

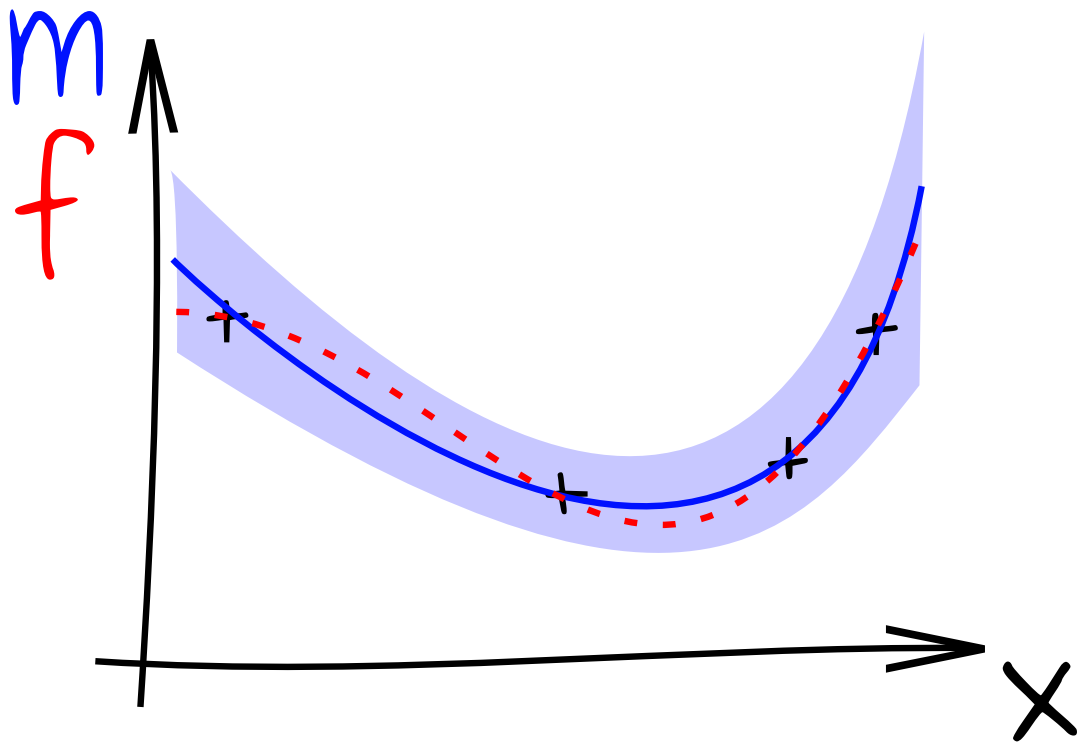


The model can then be used to answer all previous questions

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 $X = (X_1, ..., X_n)^t$
- 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 $B = (b_0, \ldots, b_p)$, we assume that the observations come from the probabilistic model

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

where the vector β is unknown and the ε_i are independent and identically distributed.

If we consider a model of the form

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

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

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

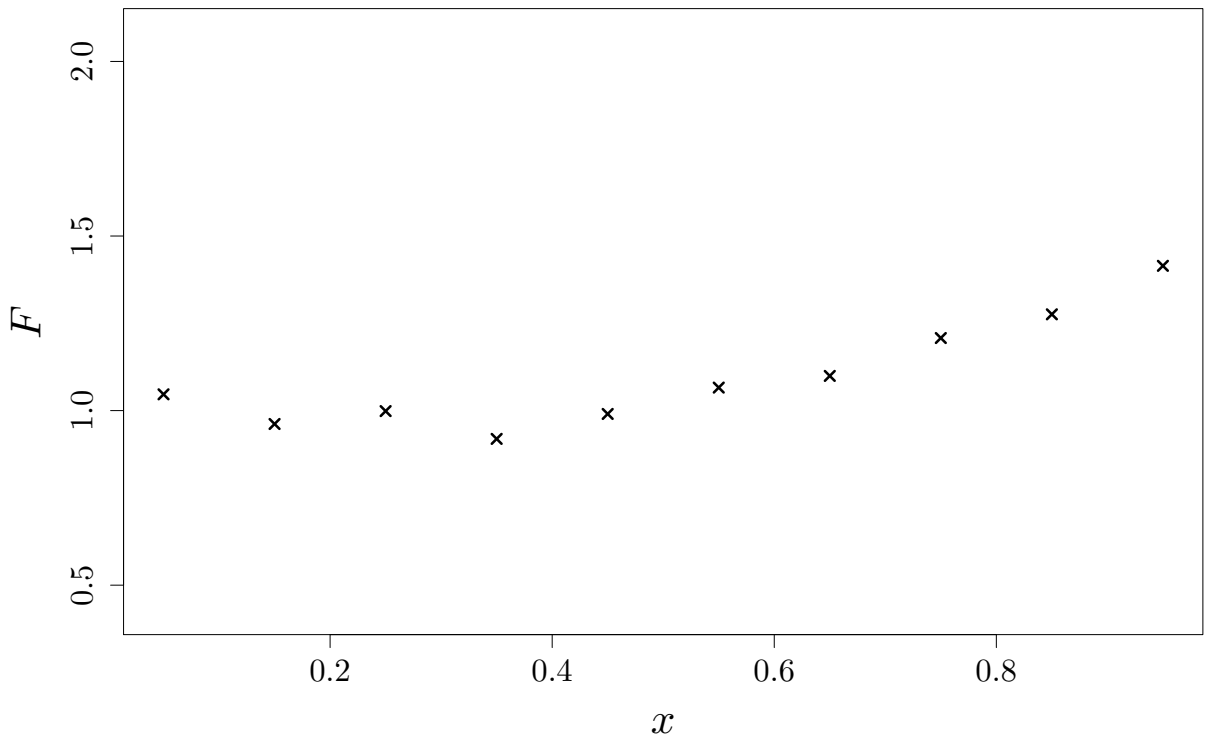
Finding the optimal value of $\hat{\beta}$ means minimizing a quadratic form. This can be done analytically and we obtain $\hat{\beta} = (B(X)^t B(X))^{-1} B(X)^t F$.

The associated linear regression model is thus

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

Example

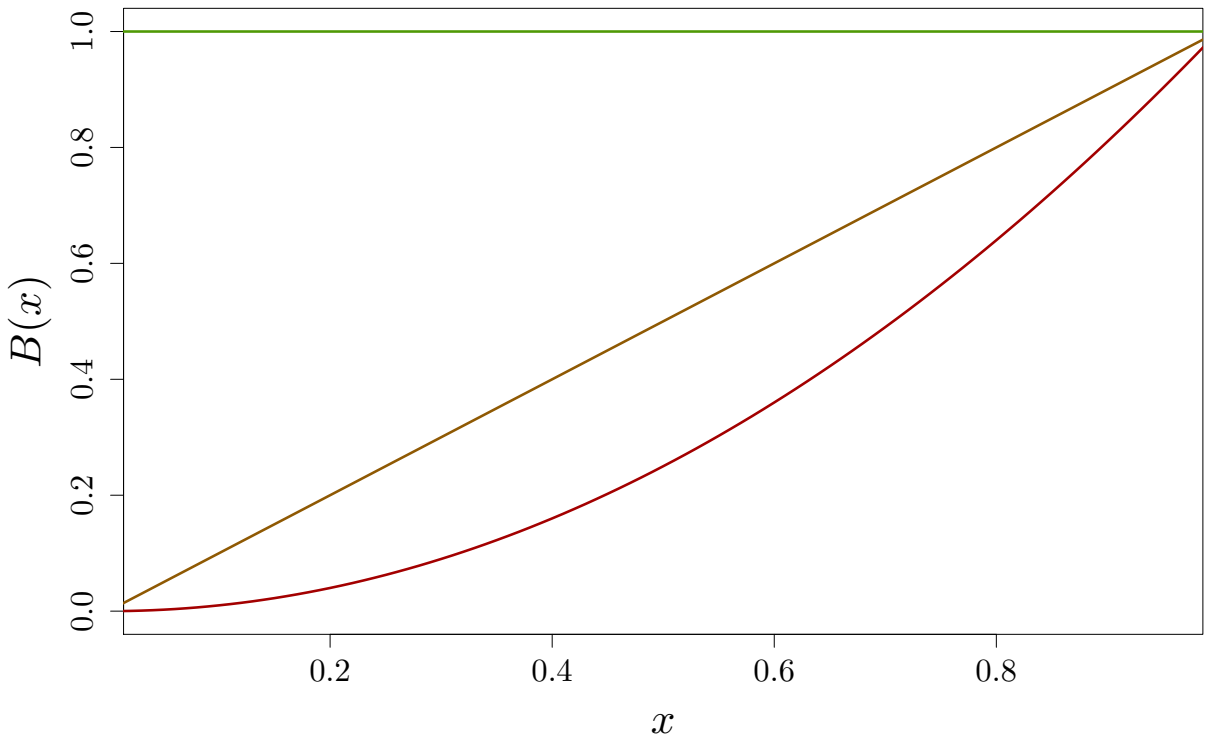
If we consider the following observations:



Example

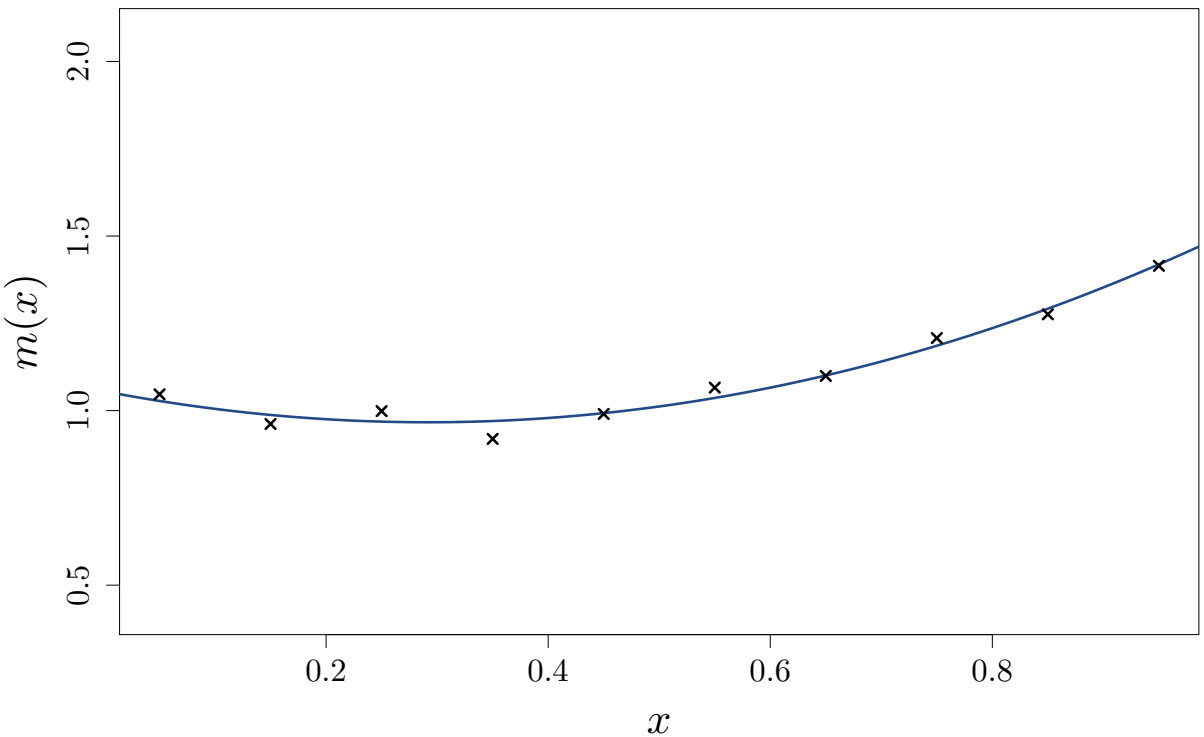
and a set of 3 basis functions:

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



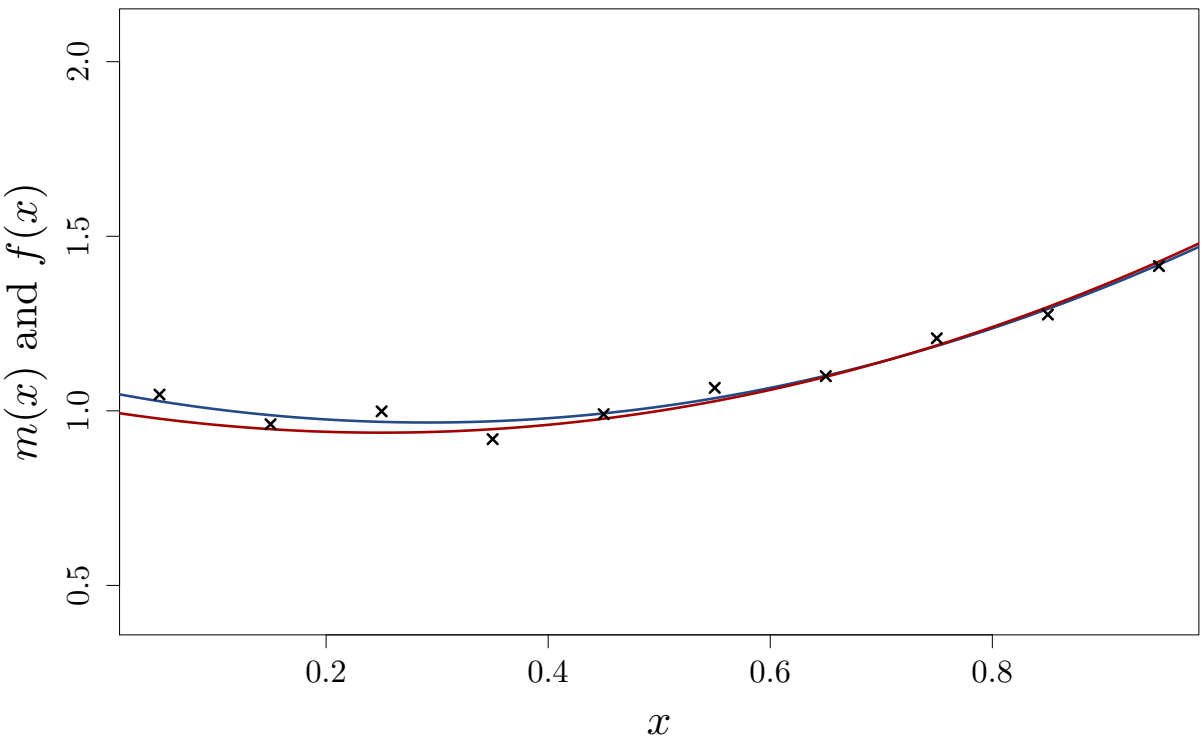
Example

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 β:

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

$\hat{\beta}$ 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 β:

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

$\hat{\beta}$ 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 $\hat{\beta}$?

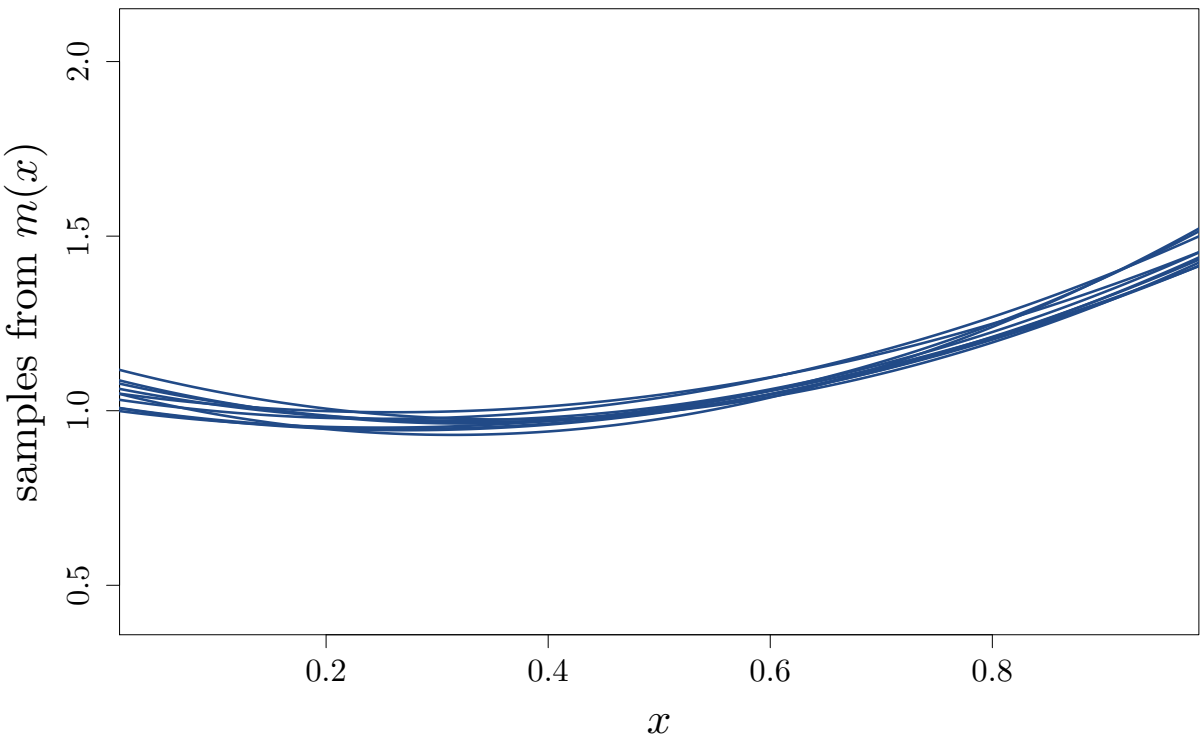
- Its expectation is $\beta \Rightarrow$ The estimator is unbiased
- Its covariance matrix is

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

- If ε is multivariate normal, then $\hat{\beta}$ is also multivariate normal.

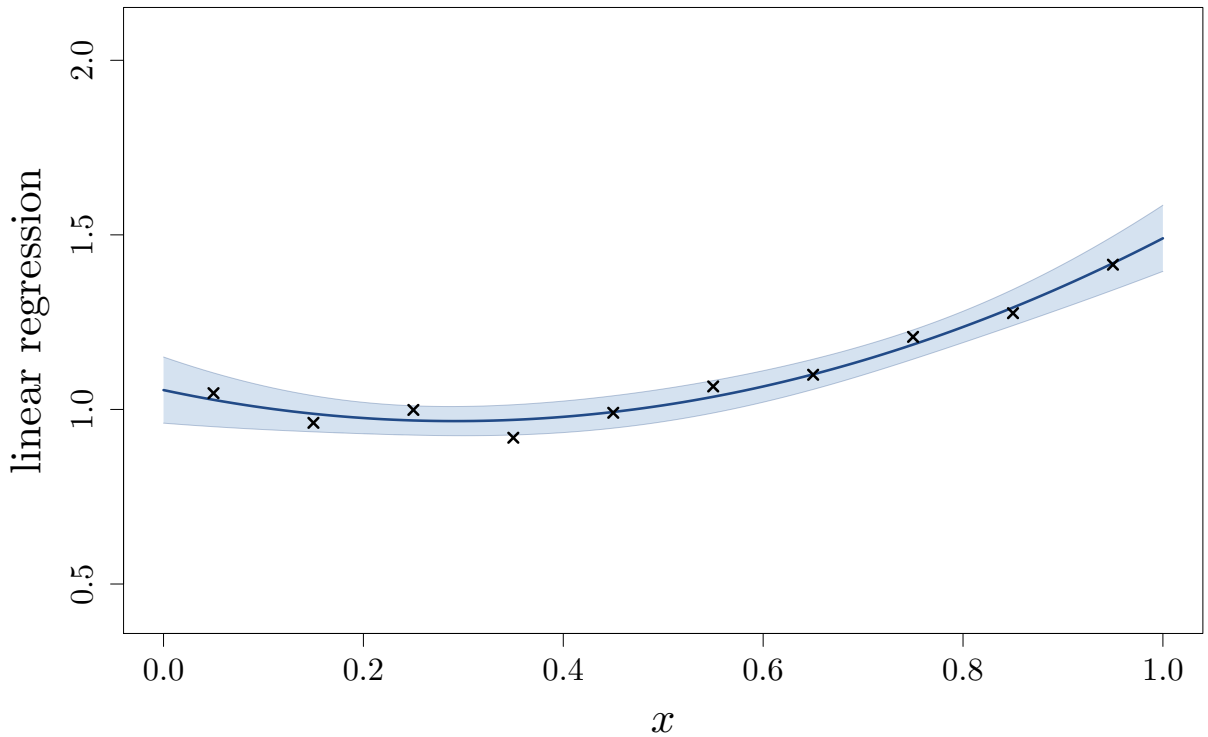
Sampling in the distribution of $\hat{\beta}$ gives us a large variety of models which represent the uncertainty about our estimation:

Back to the example



Back to the example

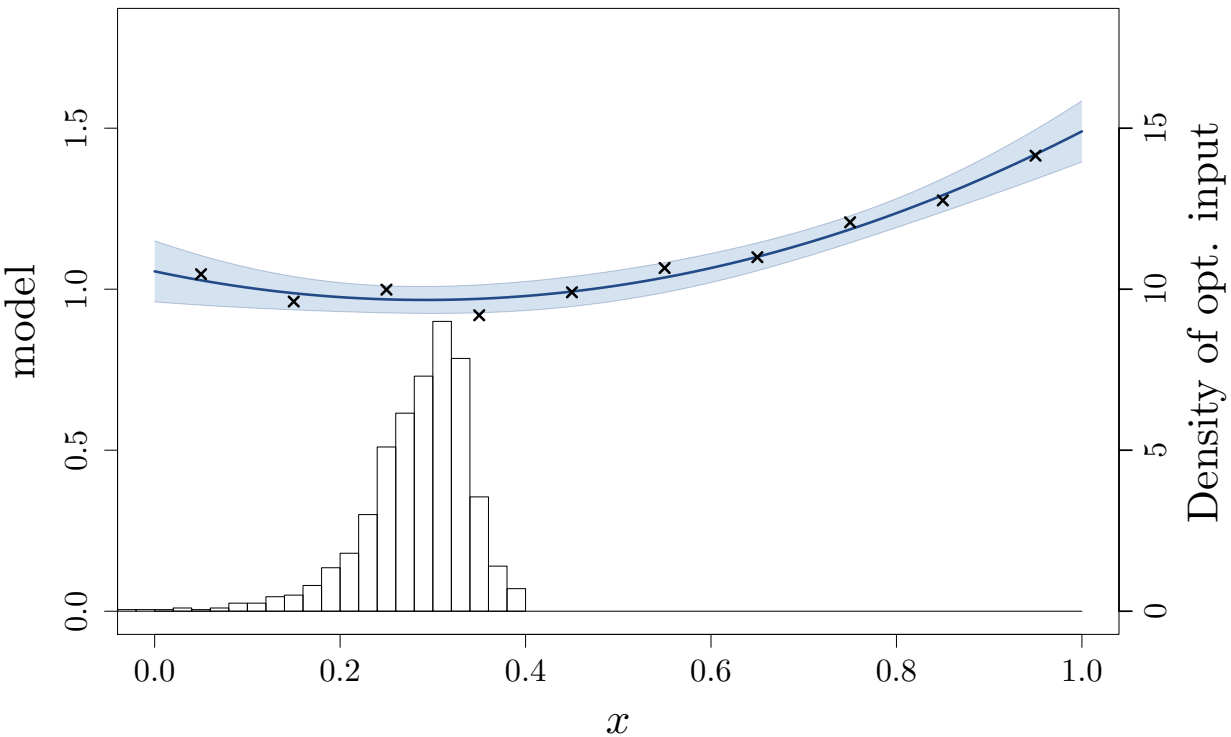
The previous picture can be summarized by showing the mean of m and 95% confidence intervals



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 $x^\star$ minimizing $f(x)$:

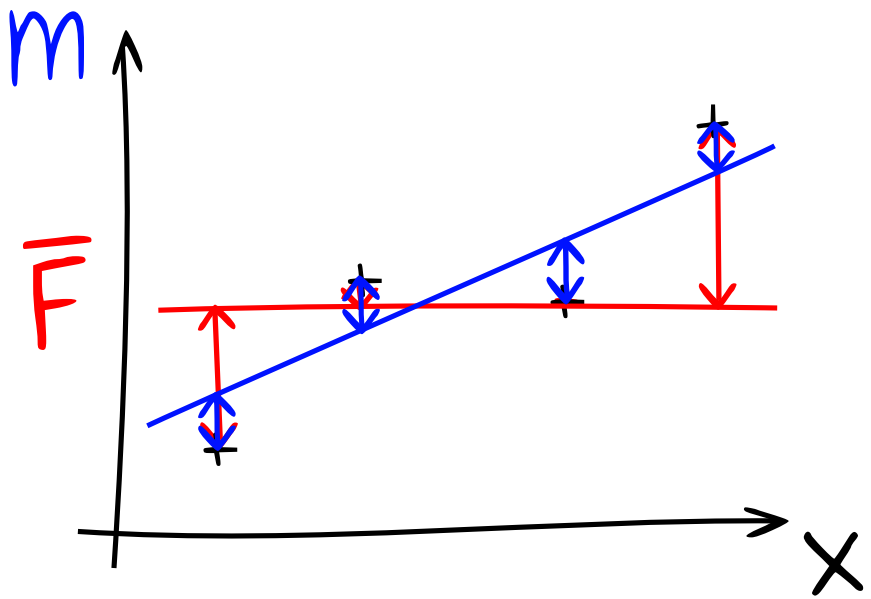


The expectation of $x^\star$ **is not** the input minimizing $m(x)$.

Model validation is always of upper importance.

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

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

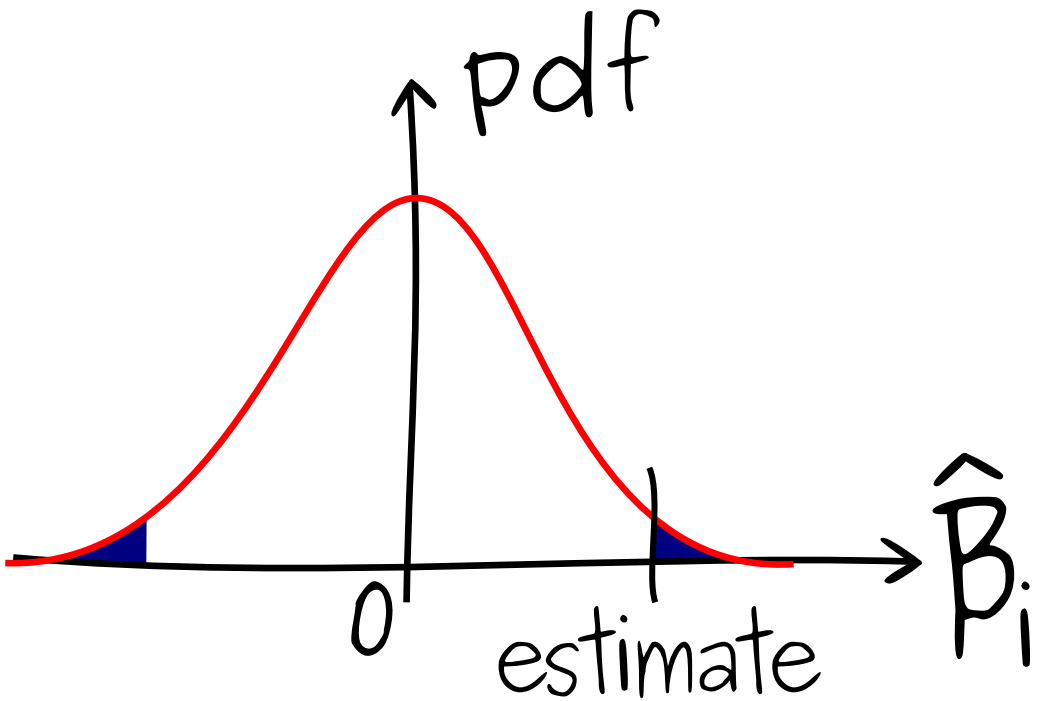


Be careful! A good R^2 (ie close to one) does not necessarily mean that the model is good.

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

H_0: The basis function b_i has no influence

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 (overfitting)

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 $Y = (Y_1, \ldots, Y_n)$ 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 $\mu = (\mu_1, \ldots, \mu_d)$
- a $d \times d$ covariance matrix Σ : $\Sigma_{i,j} = \text{cov}(Y_i, Y_j)$

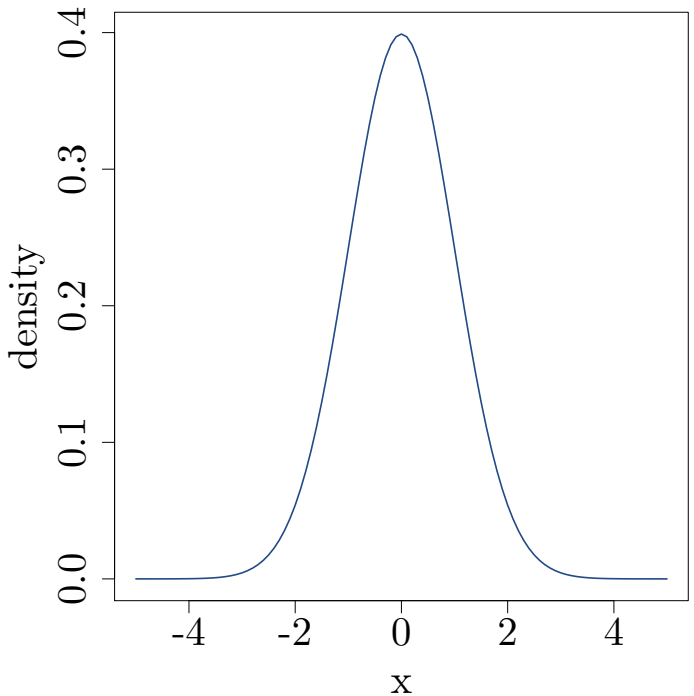
Property:

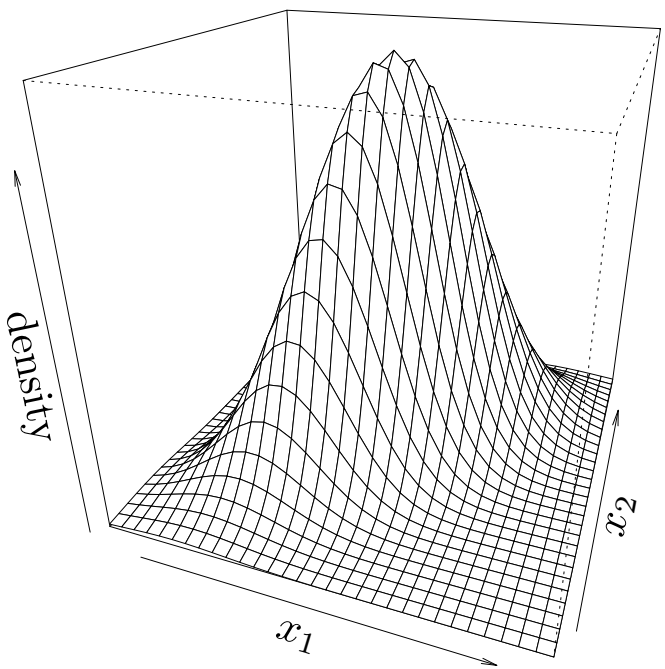
A covariance matrix is

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

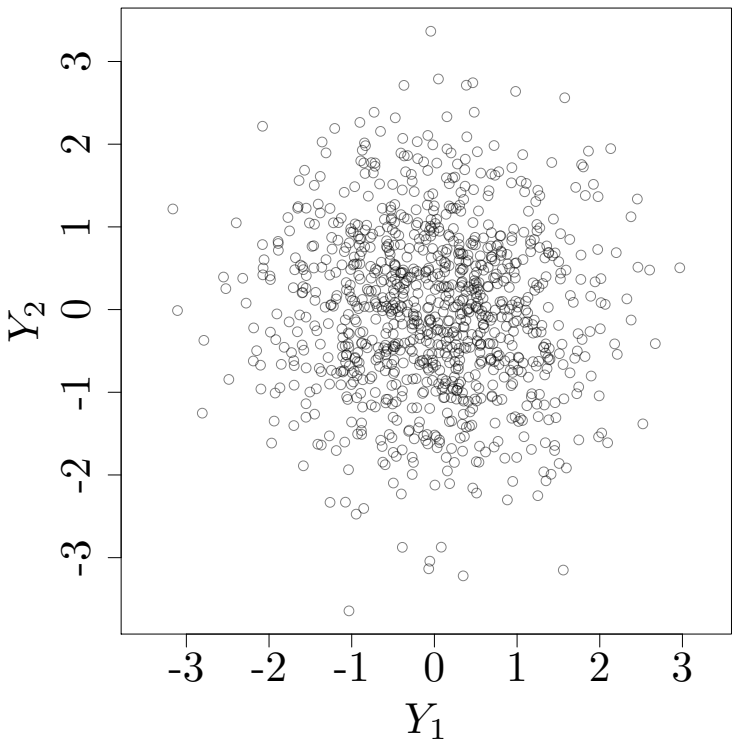
The density of a multivariate Gaussian is:

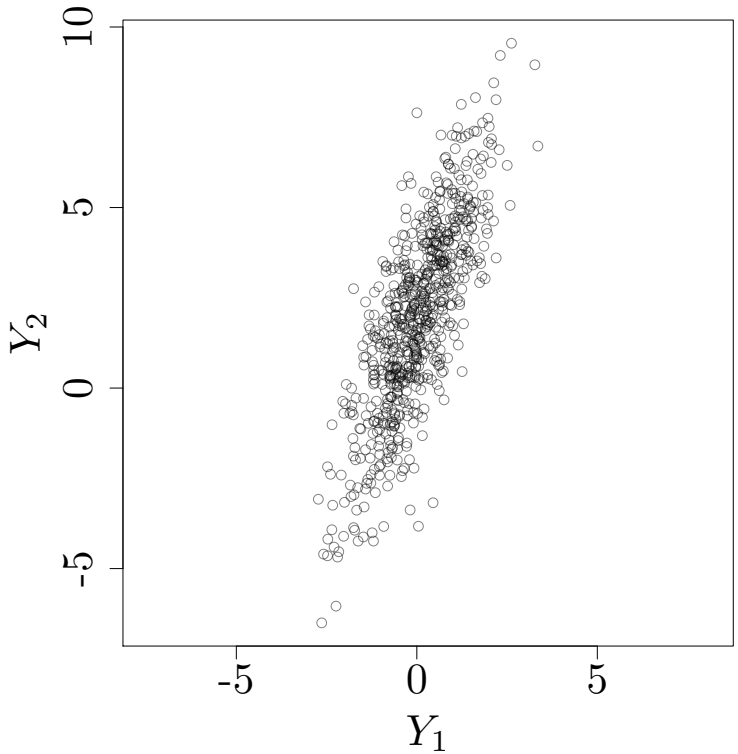
$$f_Y(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^t\Sigma^{-1}(x-\mu)\right).$$



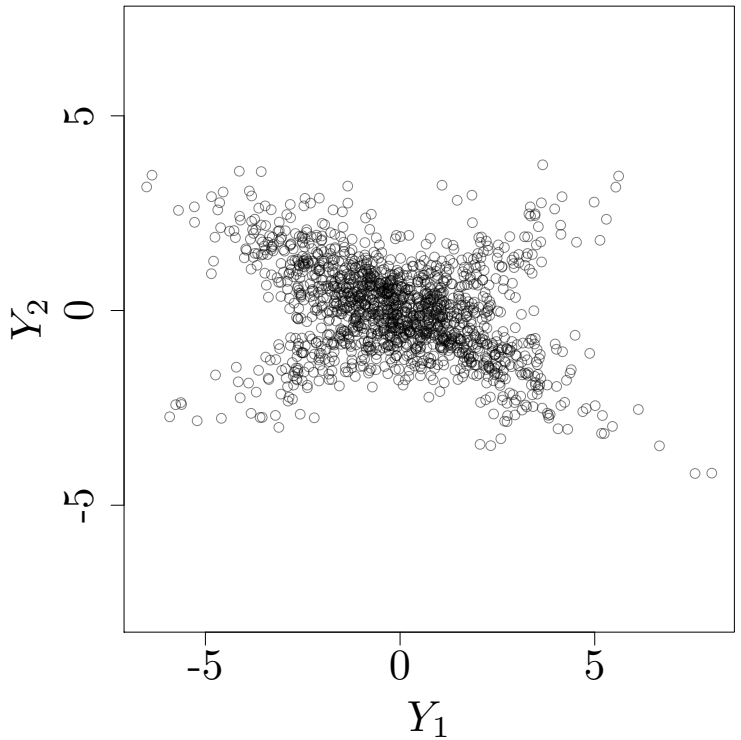


Example





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

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

Exercise

Starting from the density function, prove the previous property using the Schur block 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

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

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:

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

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.

Exercice: 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 $Z(X)$?

$\Rightarrow$ This will be illustrated this afternoon during the lab session

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

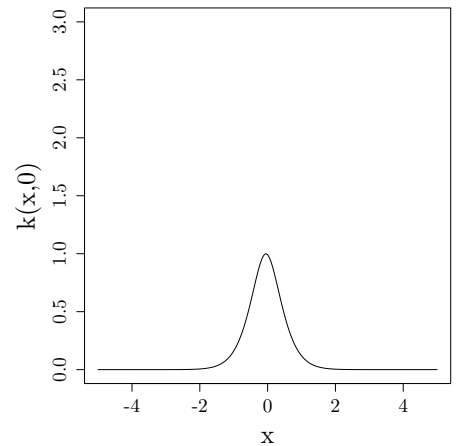
Exercice:
If Z is a GP $\mathcal{N}(0, k(.,.))$, what is the distribution of $Y(x) = \sigma Z(x/\theta)$

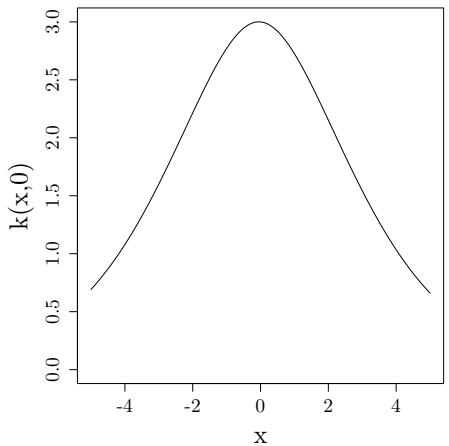
σ^2 is called the **variance** and θ the **lengthscale**

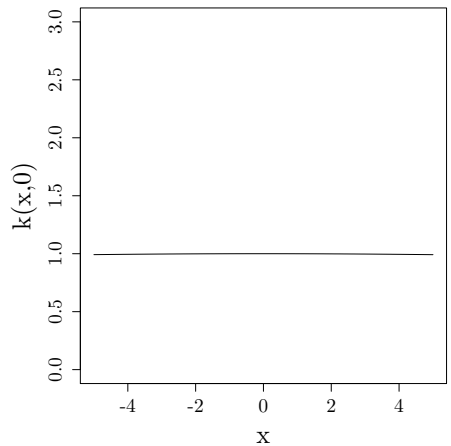
Exercice:

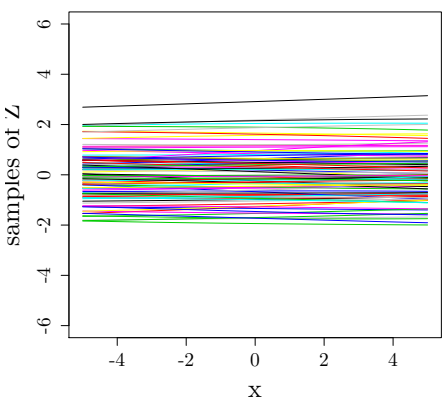
The kernel is Matern 5/2. Can you put each line in the right order?

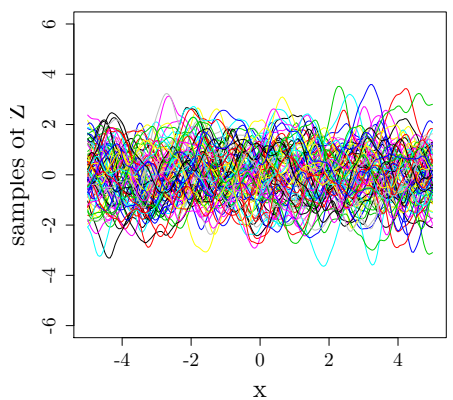
$$(\sigma^2, \theta) = (3, 3) \qquad (\sigma^2, \theta) = (1, 0.5) \qquad (\sigma^2, \theta) = (1, 50)$$

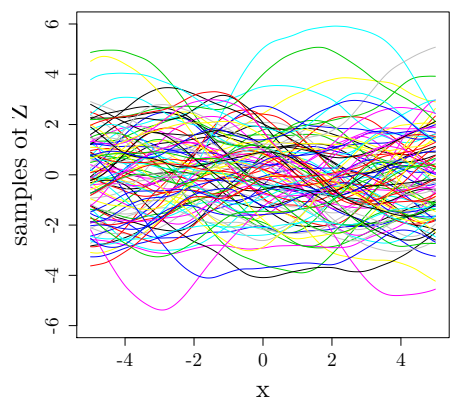








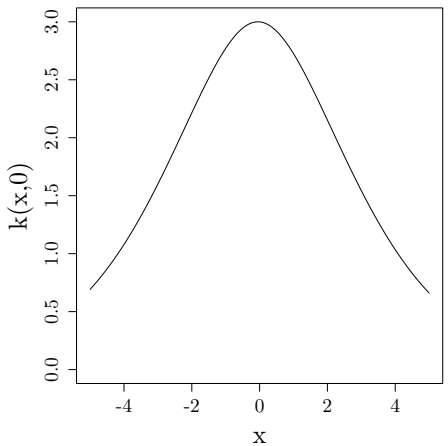


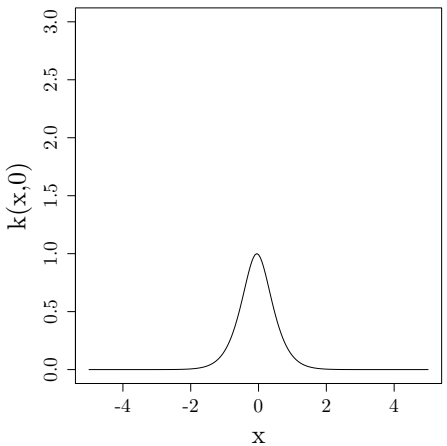


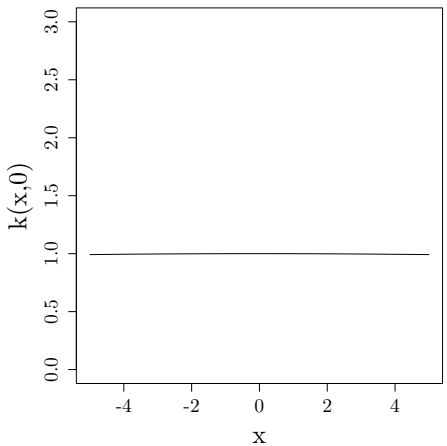
Exercice:

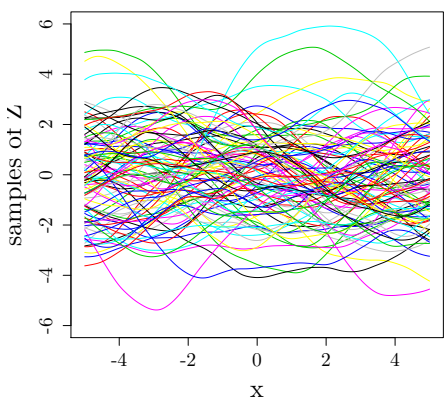
Answer is:

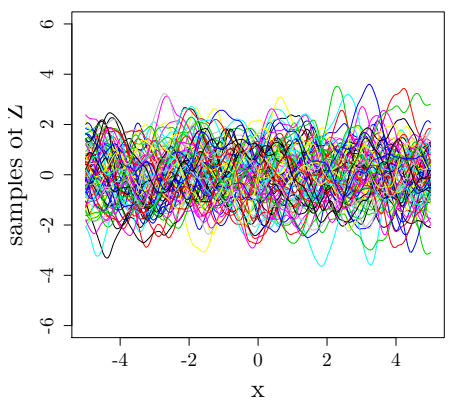
$(\sigma^2, \theta) = (3, 3)$ $(\sigma^2, \theta) = (1, 0.5)$ $(\sigma^2, \theta) = (1, 50)$

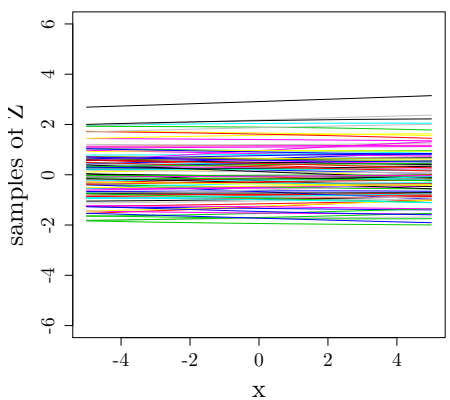












In higher dimension one can introduce one lengthscale parameter per dimension. The usual Euclidean distance between two points $||x - y|| = (\sum (x_i - y_i)^2)^{1/2}$ is thus replaced by

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

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

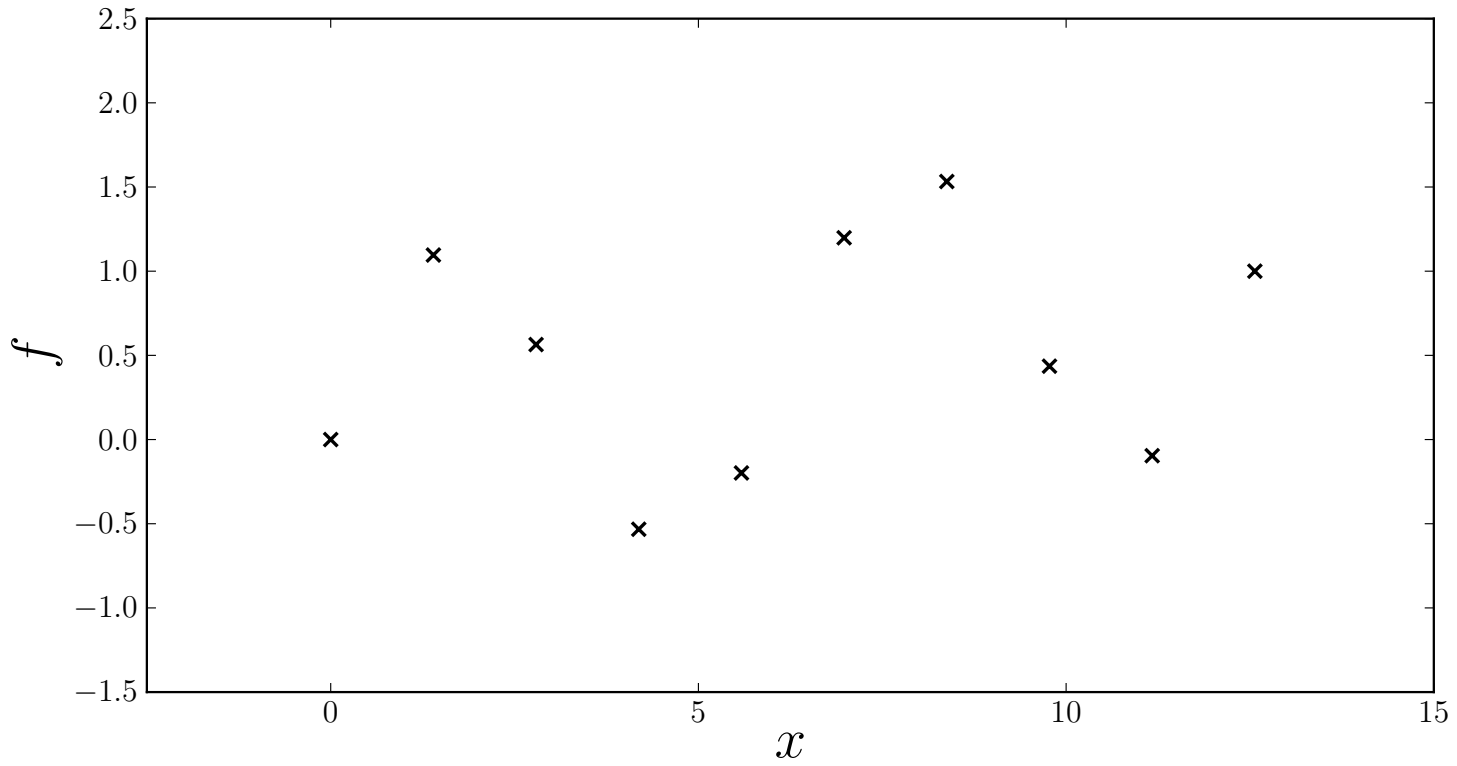
Here is a list of the most common kernels:

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

Once again we can look at sample paths.

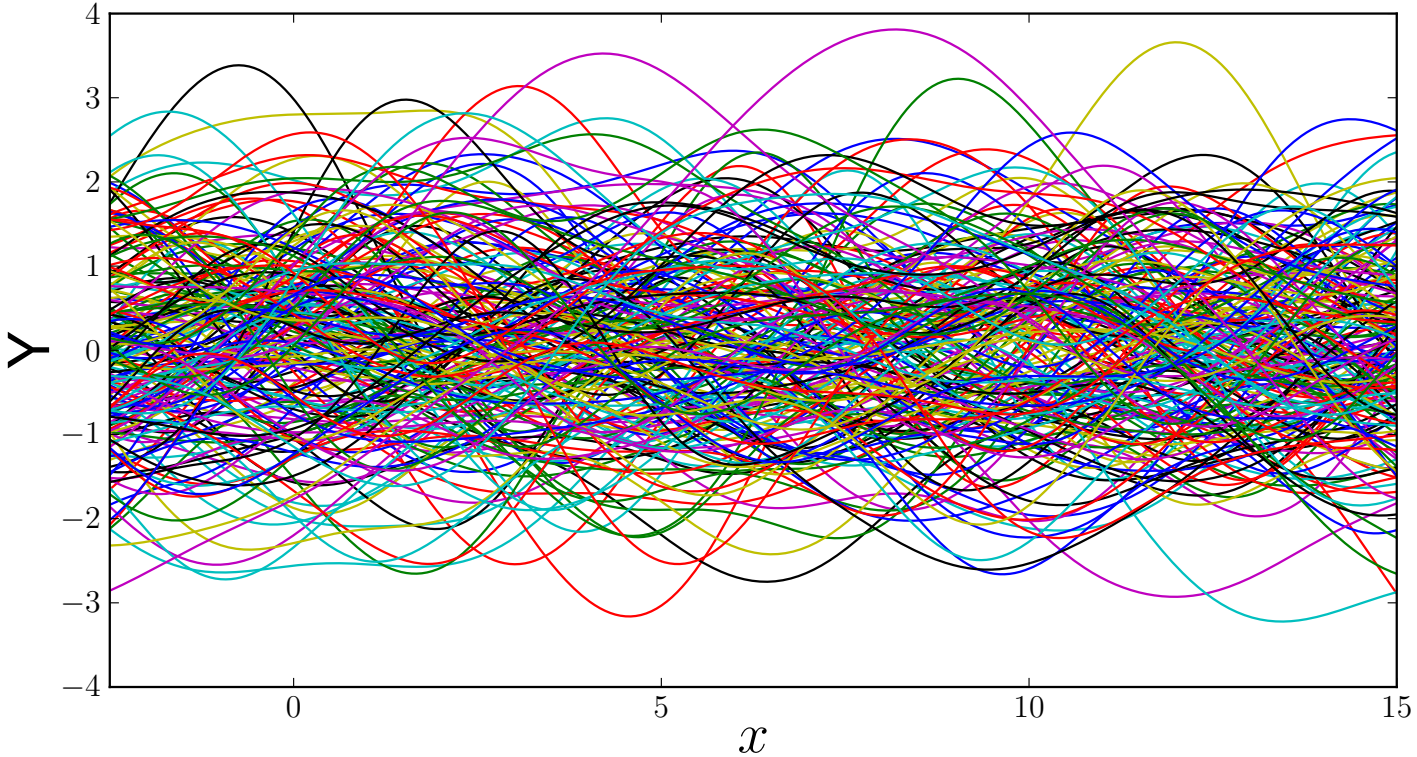
3. Gaussian process regression

We assume we have observed a function f for 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 $Z \sim \mathcal{N}(0, k)$:



What would be the next step?

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

Exercice

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 $c(X_1, X_1)$.

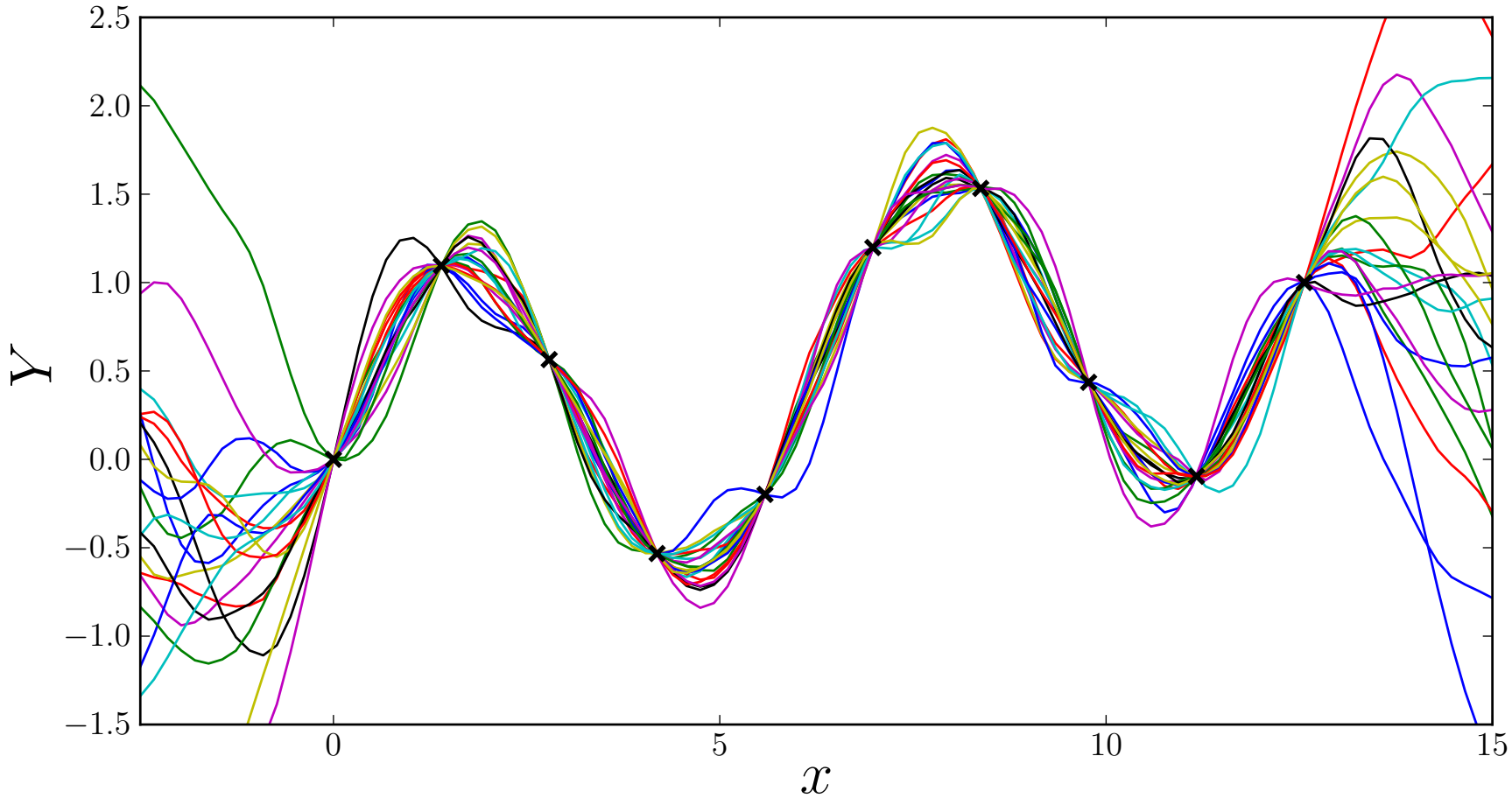
Solution

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

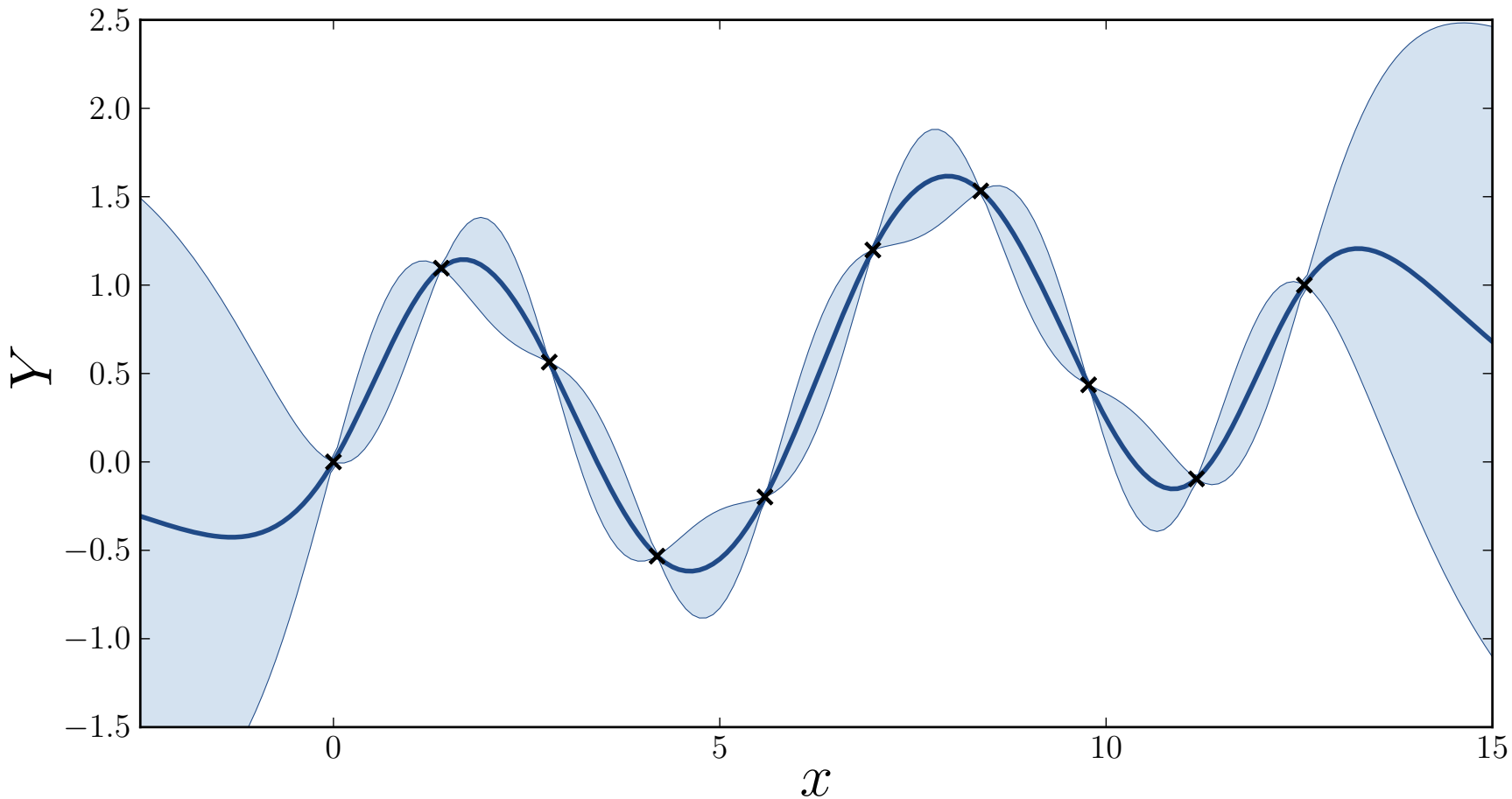
$$\begin{aligned} m(x) &= \mathrm{E}[Z(x)|Z(X)=F] \\ &= k(x,X)k(X,X)^{-1}F \\ c(x,y) &= \mathrm{cov}[Z(x),Z(y)|Z(X)=F] \\ &= k(x,y) - k(x,X)k(X,X)^{-1}k(X,y) \end{aligned}$$

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.

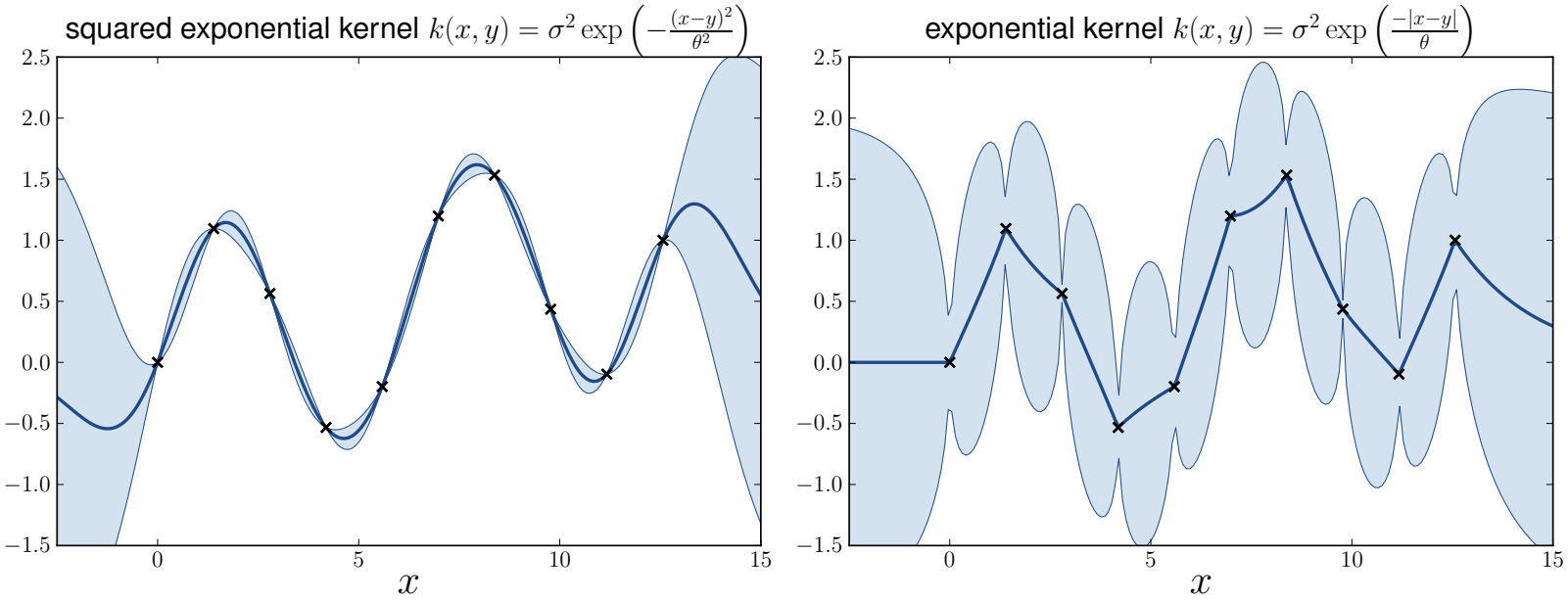


A few remarkable properties of GPR models

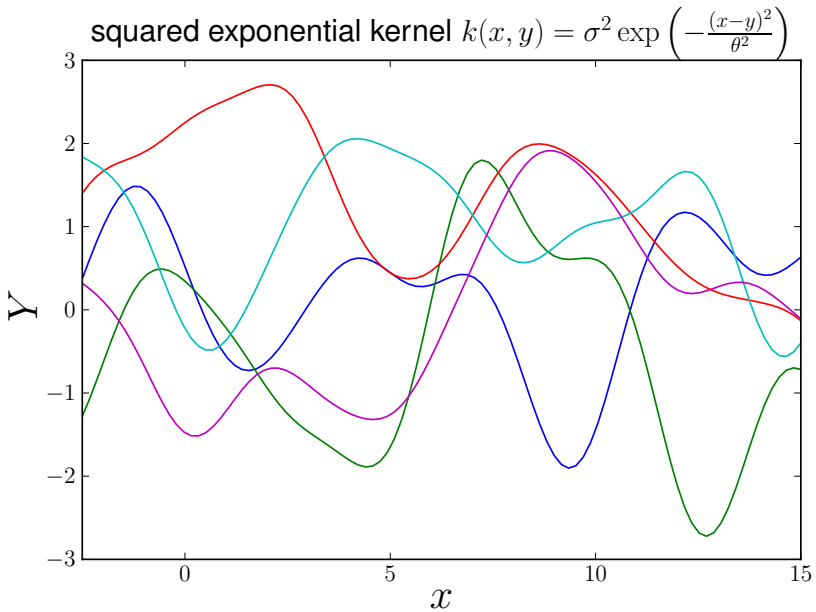
- They interpolate the data-points
- The prediction variance does not depend on the observations
- The mean predictor does not depend on the variance
- They (usually) come back to zero when we are far away from the observations.

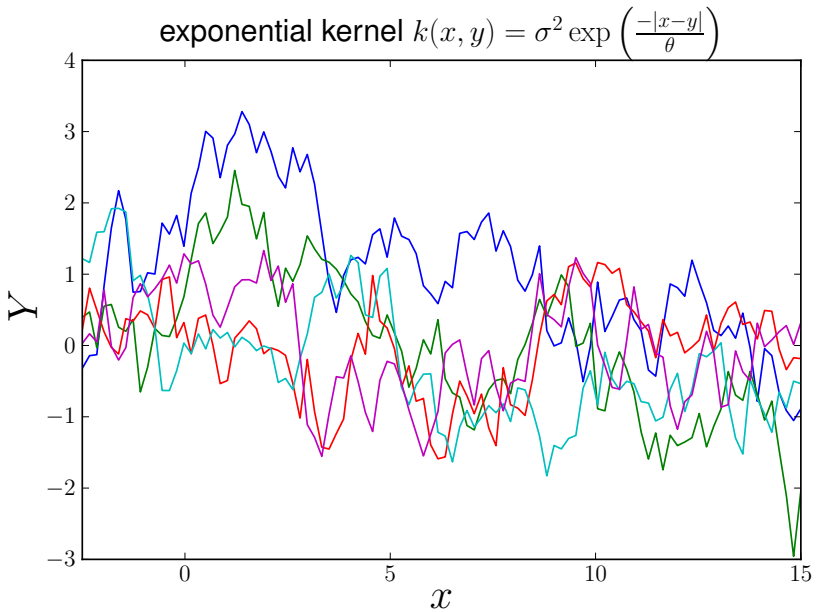
Can you prove them?

Changing the the kernel has a huge impact on the model:



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



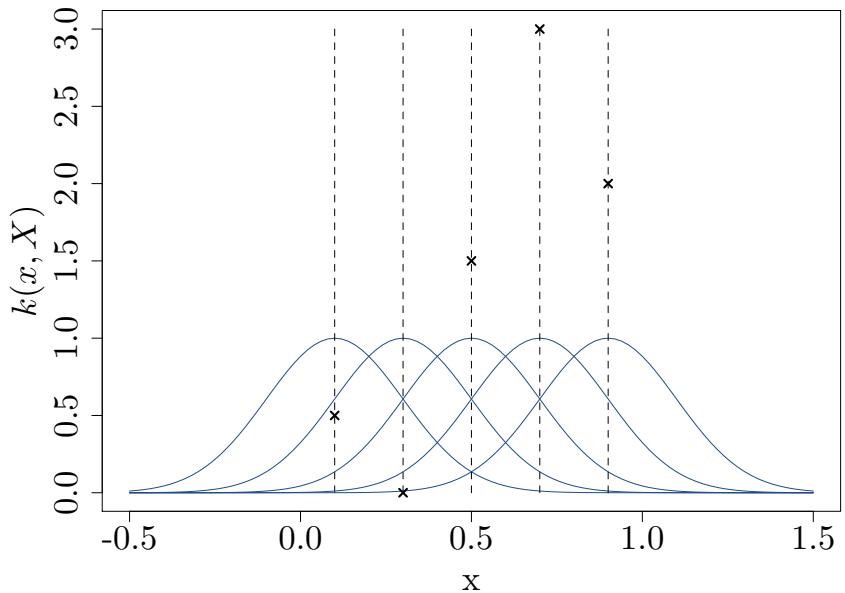


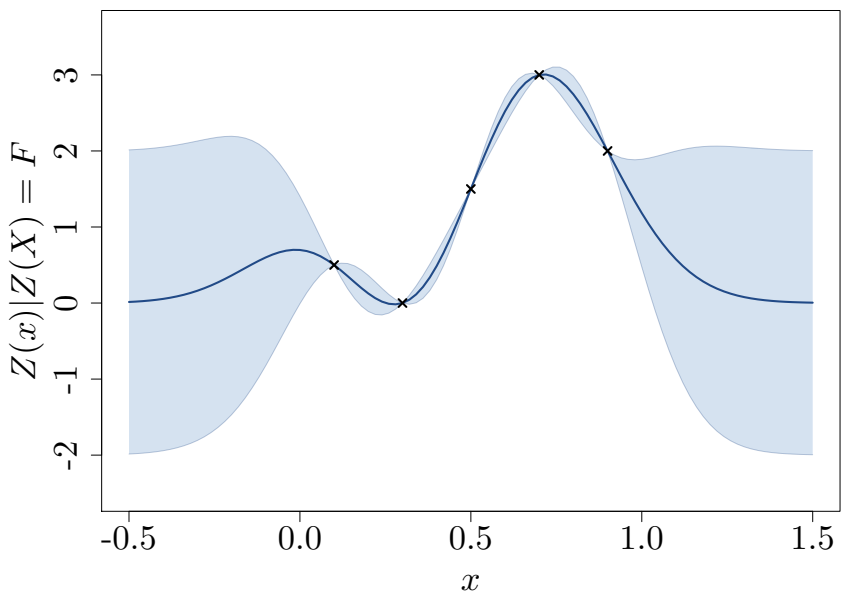
The best predictor can be seen either as a linear combination of

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

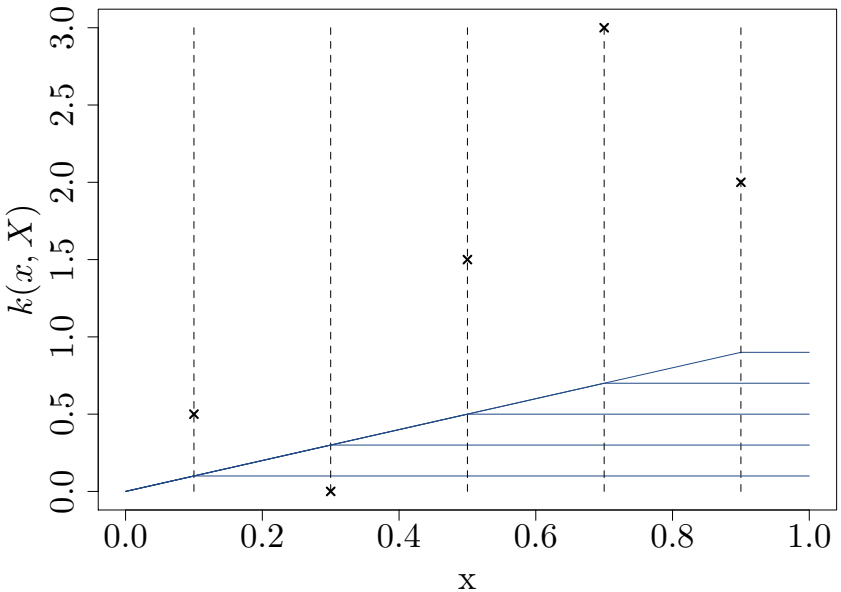
The later is interesting to understand the model shape and behaviour.

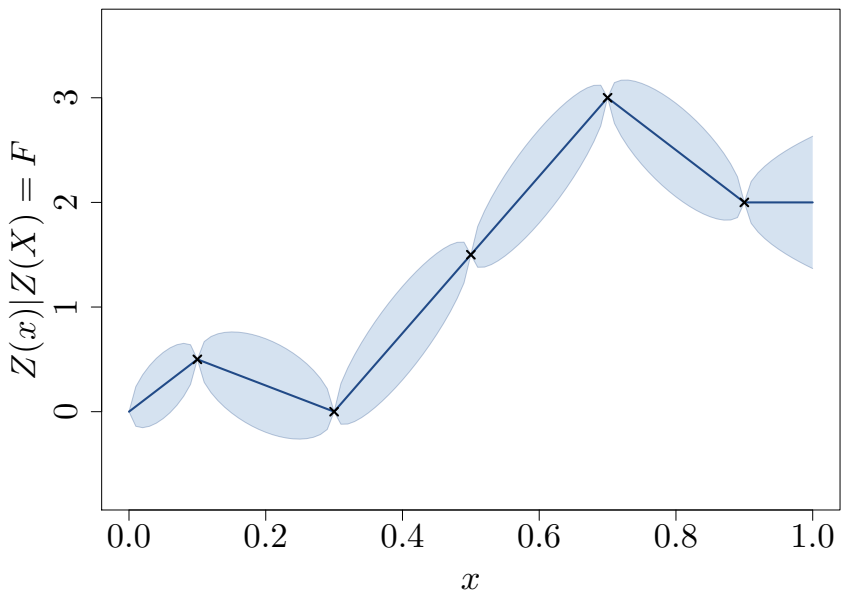
For example, we have for a squared exponential kernel





and for a Brownian kernel:





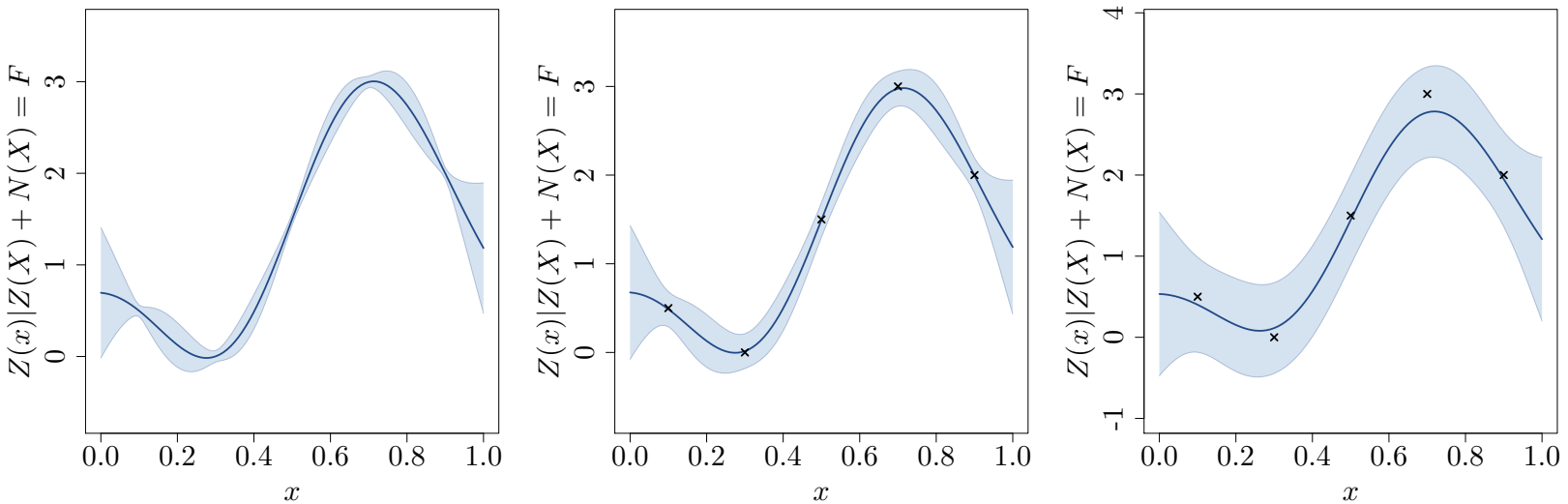
We are not always interested in models that interpolate the data.
For example, if there is some observation noise: $F = f(X) + \varepsilon$.

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

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

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

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



The values of τ^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).