Surrogate models and Gaussian Process regression – lecture 2/5

Kriging and Gaussian Process Regression

Mines St-Étienne – Majeure Data Science – 2016/2017

Nicolas Durrande (durrande@emse.fr)

Gaussian Process Regression

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 Σ : $\Sigma_{i,j} = \mathrm{cov}(Y_i, Y_j)$

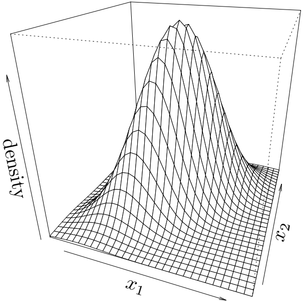
Property:

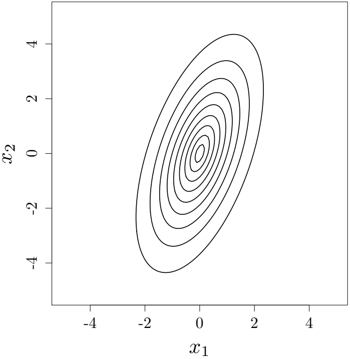
A covariance matrix is

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

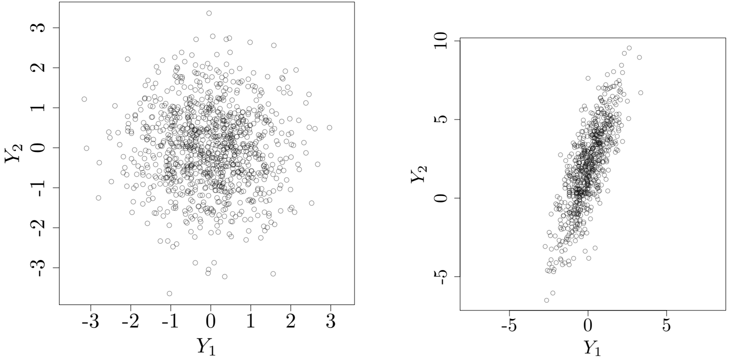
The density of a multivariate Gaussian is:

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

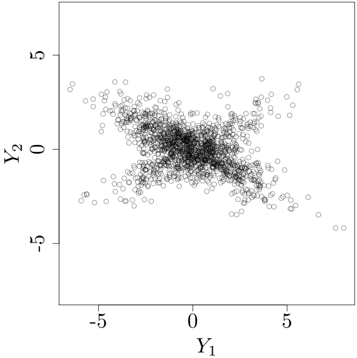




Example



Counter example



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

Exercise

Let:

- ε be a vector of n independent random variables with distribution $\mathcal{N}(0,1)$.
- W be a matrix of size $n \times n$.
- b be a vector of length n.

We define the random vector Y by

$$Y = W\varepsilon + b.$$

Questions:

1. Compute the expectation of Y.
2. Compute the covariance matrix of Y.
3. Deduce from the above a method to generate samples from a Gaussian vector with arbitrary distribution $\mathcal{N}(\mu, \Sigma)$.

Conditional distribution

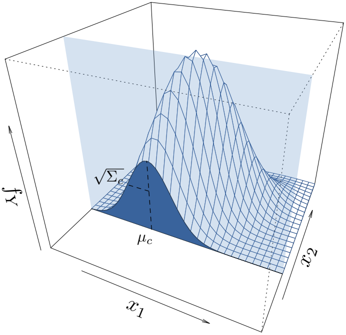
Let (Y, Z) be a Gaussian vector (Y and Z may both be vectors) with mean $(\mu_Y, \mu_Z)^t$ and covariance matrix

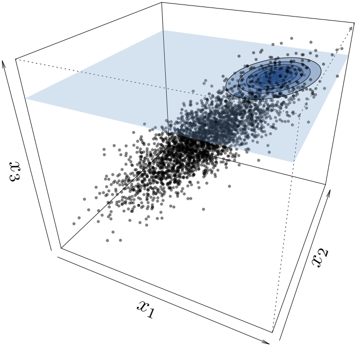
$$\begin{pmatrix} \mathrm{cov}(Y, Y) & \mathrm{cov}(Y, Z) \\ \mathrm{cov}(Z, Y) & \mathrm{cov}(Z, Z) \end{pmatrix}.$$

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

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

Here is a graphical illustration of this property:





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:

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

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

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

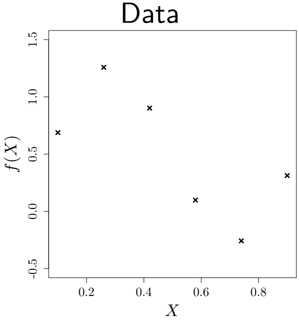
The distribution of a GP is fully characterised by:

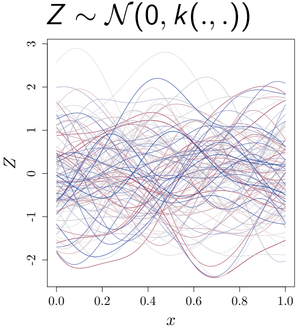
- its mean function m defined over D
- its covariance function (or kernel) k defined over $D \times D$: $k(x, y) = \text{cov}(Z(x), Z(y))$

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

Gaussian processes Regression

Suppose that we have:





Exercise

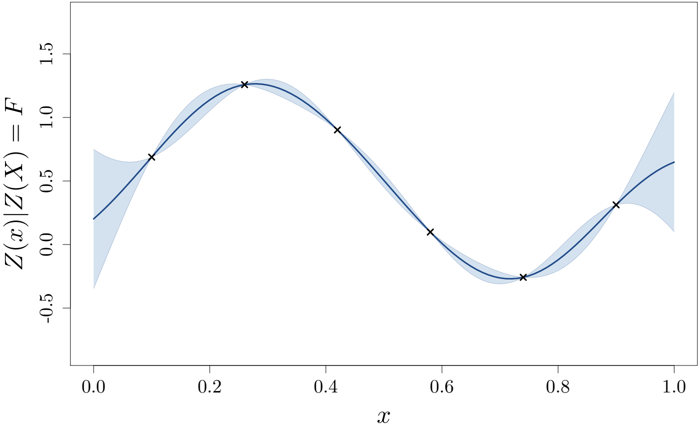
1. What is the conditional distribution of $Z(x)|Z(X) = F$?
2. Compute the conditional mean m and covariance $c(.,.)$.

Solution

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

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

We finally obtain



where the blue area corresponds to 95% confidence intervals: $m(x) \pm 1.96\sqrt{c(x,x)}$.

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
- The mean predictor (usually) come back to zero when for predictions far away from the observations.

Exercise

- Prove the first three items.
- Prove the last item for a squared exponential kernel:

$$k(x, y) = \exp\left(-(x-y)^2\right).$$

Kernels

For a given set of observations, the model is fully determined by the kernel. 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.

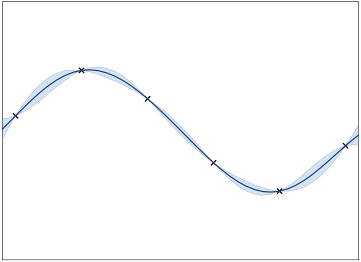
Proving that a function is psd is often intractable. However there are a lot of functions that have already been proven to be psd:

constant	$k(x, y) = 1$
white noise	$k(x, y) = \delta_{x,y}$
Brownian	$k(x, y) = \min(x, y)$
exponential	$k(x, y) = \exp\left(-\lvert x - y\rvert\right)$
Matern 3/2	$k(x, y) = \left(1 + \lvert x - y\rvert\right) \exp\left(-\lvert x - y\rvert\right)$
Matern 5/2	$k(x, y) = \left(1 + \lvert x - y\rvert + 1/3\lvert x - y\rvert^2\right) \exp\left(-\lvert x - y\rvert\right)$
squared exp.	$k(x, y) = \exp\left(-(x - y)^2\right)$
$\vdots$	

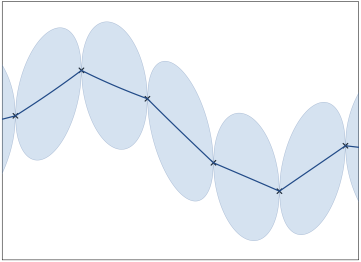
When k is a function of $x - y$, the kernel is called **stationary**.

Changing the kernel has a huge impact on the model:

Gaussian kernel:

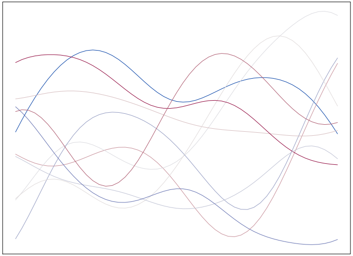


Exponential kernel:

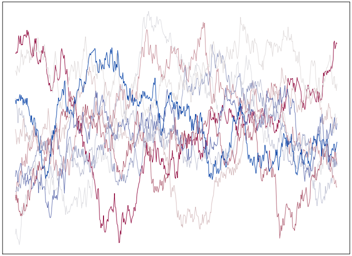


This is because changing the kernel means changing the prior on f

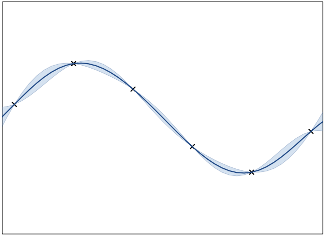
Gaussian kernel:



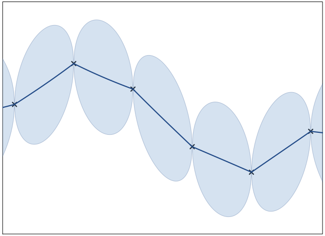
Exponential kernel:



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 ?
- ...

We have seen that one kernel gives one model. However, one can include some scaling parameters into the kernels to improve their adequacy to the data:

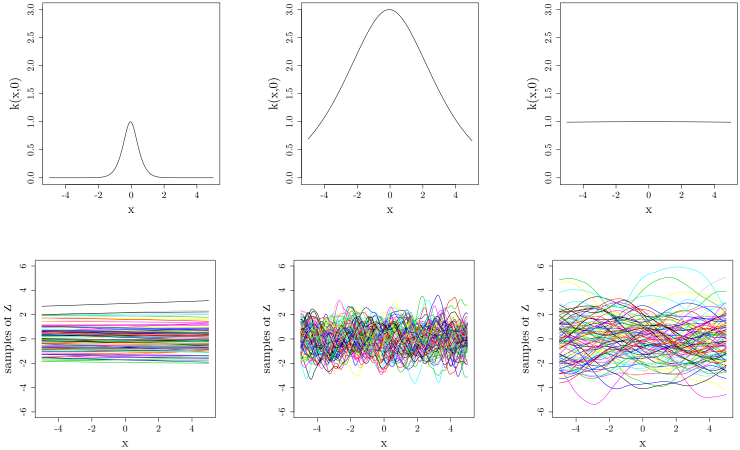
Exercise:
If Z is a GP $\mathcal{N}(0, k(.,.))$, can you detail the distribution of $Y(x) = \sigma Z(x/\theta)$

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

Exercise:

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

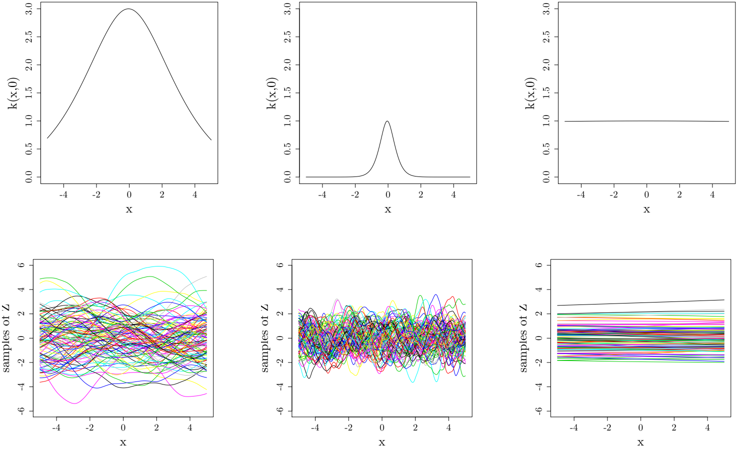
$$(\sigma^2, \theta) = (3, 3) \qquad (\sigma^2, \theta) = (1, 0.5) \qquad (\sigma^2, \theta) = (1, 50)$$



Exercise:

Answer is:

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



Adding these parameters to usual kernels gives

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

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

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

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

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

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

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

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

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, it is common to work with the concentrated log-likelihood:

$$\ell(\sigma^2, \theta) = \log(|k(X,X)|) + F^t k(X,X)^{-1} F$$

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

Example

We consider 100 sample from a Matern 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.

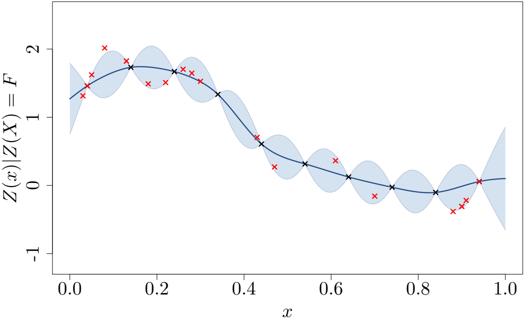
Model validation

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 interesting question now is to know how to get a good model. To do so, we will need to answer the following questions:

- What is a good model?
- How to measure it?

The idea is to introduce new data and to compare the model prediction with reality



Since GPR models provide a mean and a covariance structure for the error they both have to be assessed.

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:

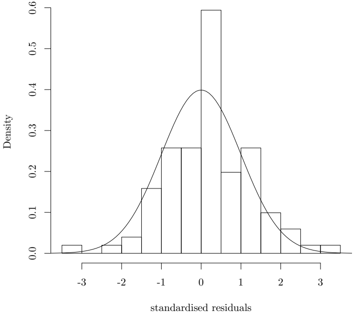
$$\begin{aligned} &\text{Mean Square Error} && MSE = \mathrm{mean}((F_t - m(X_t))^2) \\ &\text{A “normalised” criterion} && Q_2 = 1 - \frac{\sum (F_t - m(X_t))^2}{\sum (F_t - \mathrm{mean}(F_t))^2} \end{aligned}$$

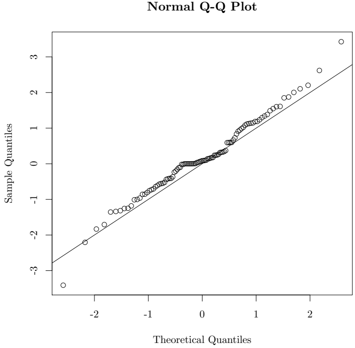
On the above example we get $MSE = 0.038$ and $Q_2 = 0.95$.

The predicted distribution can be tested by normalising 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 $\mathcal{N}(0, 1)$:





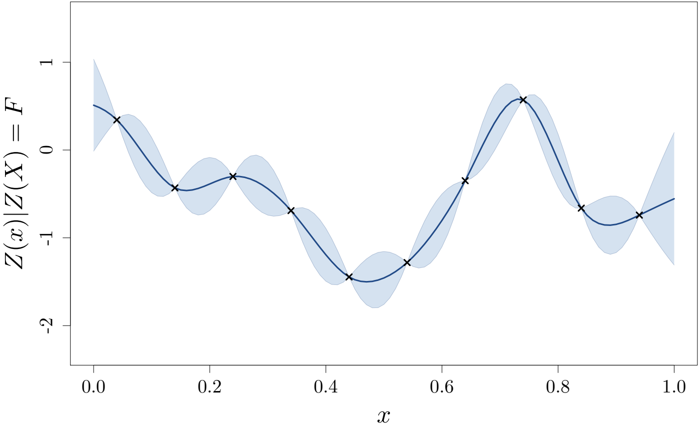
When no test set is available, another option is to consider cross validation methods such as leave-one-out.

The steps are:

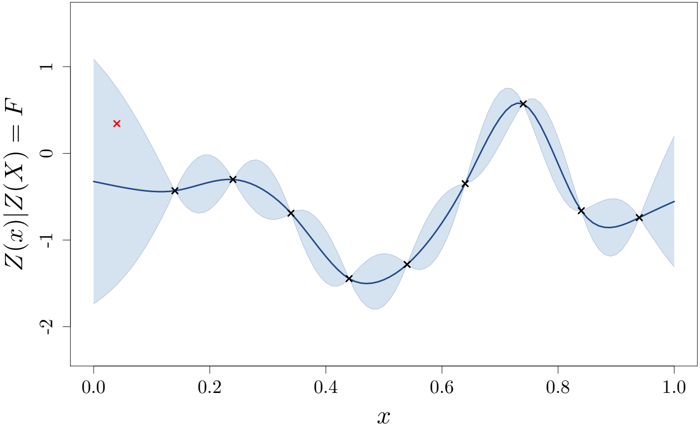
1. build a model based on all observations except one
2. compute the model error at this point

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

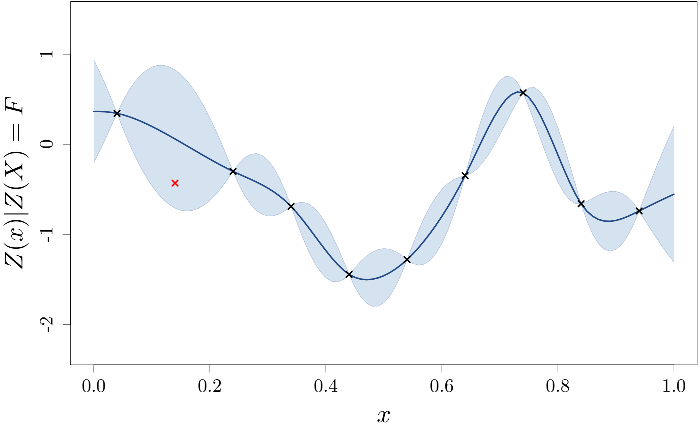
Model to be tested:



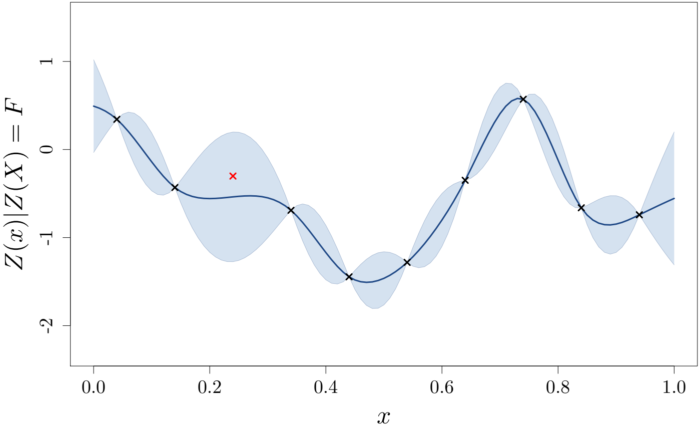
Step 1:



Step 2:



Step 3:



On this example, we obtain $MSE = 0.24$ and $Q_2 = 0.34$.

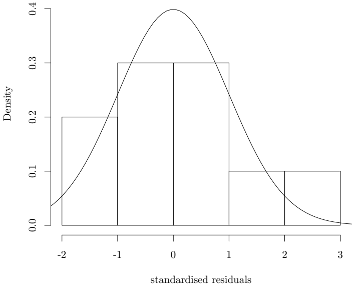
Why doesn't the model perform as good previously?

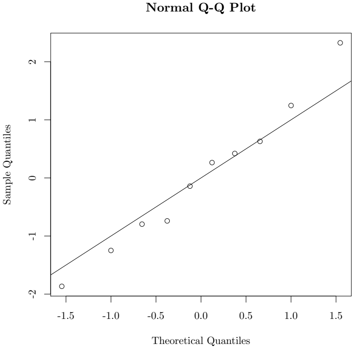
On this example, we obtain $MSE = 0.24$ and $Q_2 = 0.34$.

Why doesn't the model perform as good previously?

It turns out that the error is always computed at the 'worst' location!

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 standardised independently. We obtain:


Density
standardised residuals


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles