

Surrogate models and Gaussian Process regression – lecture 3/5

Kriging with trend and/or noisy observations

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

Nicolas Durrande (durrande@emse.fr)

Outline of the lecture:

1. Approximation

- ▶ GPR with noisy observations

2. GPR in practice

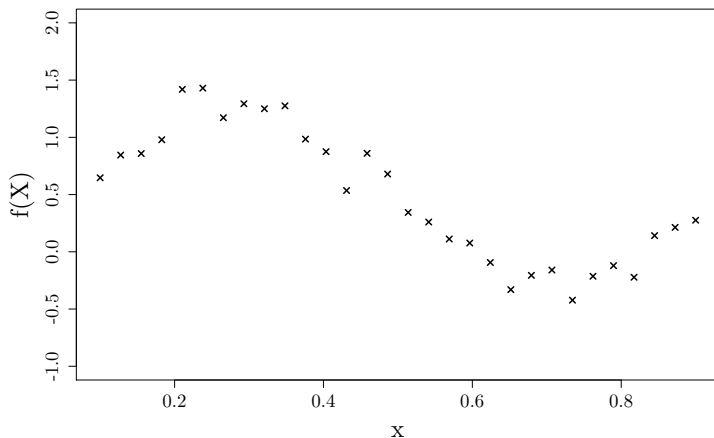
- ▶ Overall Steps for GPR
- ▶ Numerical stability
- ▶ Numerical complexity

3. GPR with trend

- ▶ Ordinary kriging
- ▶ Universal kriging

Approximation

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



Exercise:

Let f be a function of interest and let F be a vector of “noisy” observations of f at some input locations X :

$$F = f(X) + \varepsilon \quad \text{with } \varepsilon \sim \mathcal{N}(0, \tau^2 Id).$$

Let Z be a Gaussian Process corresponding that can be used as prior distribution for f .

1. Compute the conditional mean of $Z(x)|Z(X) + \varepsilon = F$
2. Compute the conditional covariance function

Solution:

1. The conditional mean is

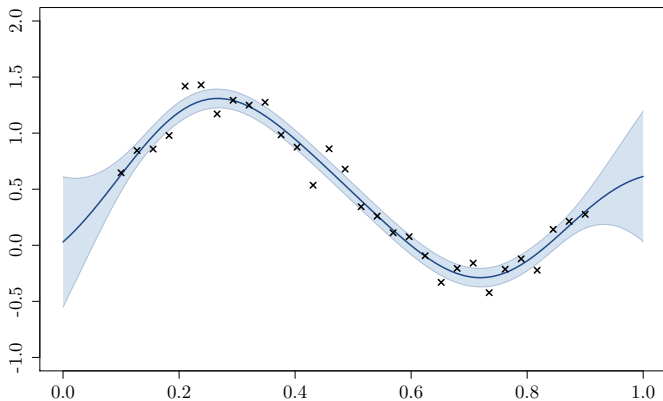
$$\begin{aligned}m(x) &= \mathbb{E}[Z(x)|Z(X) + \varepsilon = F] \\&= k(x, X)(k(X, X) + \tau^2 Id)^{-1}F\end{aligned}$$

2. The conditional variance is

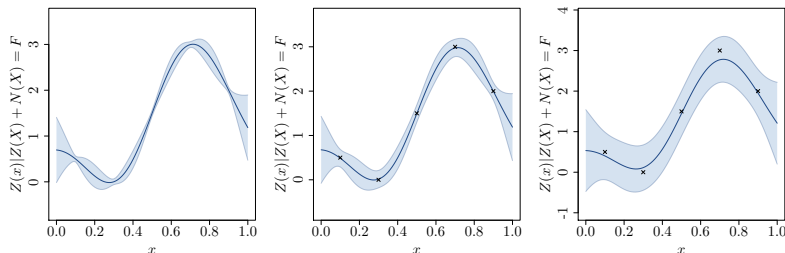
$$\begin{aligned}c(x, y) &= \text{cov}[Z(x), Z(y)|Z(X) + \varepsilon = F] \\&= k(x, y) - k(x, X)(k(X, X) + \tau^2 Id)^{-1}k(X, y)\end{aligned}$$

Note that is si straightforward to generalize for a noise that is not i.i.d (as long as the noise is Gaussian distributed).

We obtain the following model



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.

GPR in practice

The various steps for building a GPR model are:

1. Create a DoE

- ▶ What is the overall evaluation budget?
- ▶ What is my model for?

2. Choose a kernel

3. Estimate the parameters

- ▶ Maximum likelihood
- ▶ Cross-validation
- ▶ Multi-start

4. Validate the model

- ▶ Test set
- ▶ Leave-one-out to check mean and confidence intervals
- ▶ Leave- k -out to check predicted covariances

Remarks

- It is very common to iterate over steps 2, 3 and 4

In practice, the following errors may appear:

- Error: the matrix is not invertible
- Error: the matrix is not positive definite

Covariance matrices are positive semi-definite. Null eigenvalues arise if one information is repeated.

Example

For $X = (0.1, 0.1, 0.4, 0.6, 0.8)$, the covariance of a squared exponential kernel with parameters $\sigma^2 = 1$, $\theta = 0.2$ is:

$$k(X, X) = \begin{pmatrix} 1.00 & 1.00 & 0.32 & 0.04 & 0.00 \\ 1.00 & 1.00 & 0.32 & 0.04 & 0.00 \\ 0.32 & 0.32 & 1.00 & 0.61 & 0.14 \\ 0.04 & 0.04 & 0.61 & 1.00 & 0.61 \\ 0.00 & 0.00 & 0.14 & 0.61 & 1.00 \end{pmatrix}$$

The first two columns are the same, so the matrix is not invertible.

It is particularly interesting to look at the eigenvectors associated with null eigenvalues. On the previous example, this eigenvector is

$$P_0 = \left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, 0, 0, 0 \right)^t$$

Now, 2 situations can be distinguished

- (A) The observations are compatible with model: $P_0^t F = 0$.
 - ▶ the model is appropriate and one observation can be removed without any loss of information.
- (B) The observations **are not** compatible with model: $P_0^t F \neq 0$.
 - ▶ the model is not appropriate and it should be modified. For example, observation noise can be added.

In both cases, the suggested actions will make the matrix invertible.

In practice, invertibility issues may arise if observations points are close-by.

This is specially true if

- the kernel corresponds to very regular sample paths (squared-exponential for example)
- the range (or length-scale) parameters are large

On the other hand, adding a (very) small observation noise can be of great help.

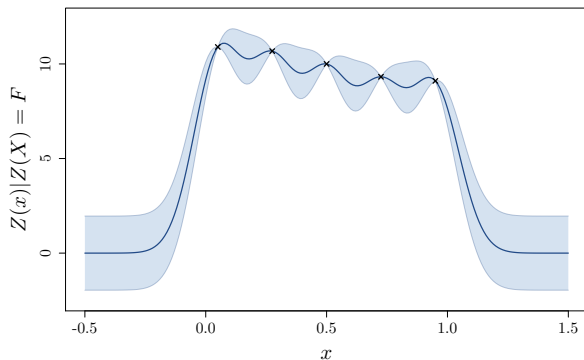
In order to avoid numerical problems during optimization, one can:

- impose a maximum bound to length-scales
- impose a minimal bound for noise variance
- choose a Matérn kernel

GPR with trend

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

This behaviour is not always wanted

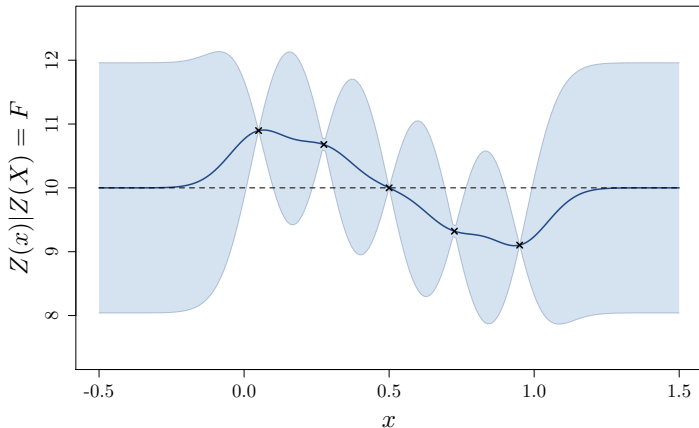


If the trend $t(\cdot)$ is known, the usual formulas for multivariate normal conditional distribution apply:

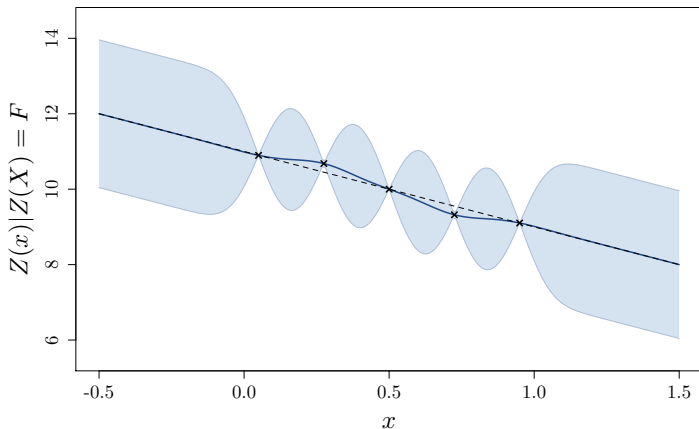
$$\begin{aligned}m(x) &= E[Z(x)|Z(X)=F] \\&= t(x) + k(x, X)k(X, X)^{-1}(F - t(X)) \\c(x, y) &= \text{cov}[Z(x), Z(y)|Z(X)=F] \\&= k(x, y) - k(x, X)k(X, X)^{-1}k(X, y)\end{aligned}$$

We can see that the trend is first subtracted and added in the end.

In the previous example, we can consider that trend is constant
 $t(x) = 10$:



We can also try a linear trend $t(x) = 11 - 2x$:

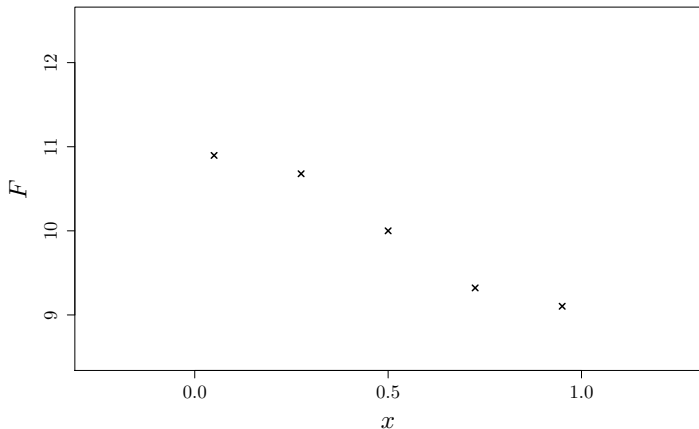


In practice, the trend is often unknown... The question is then how to estimate it.

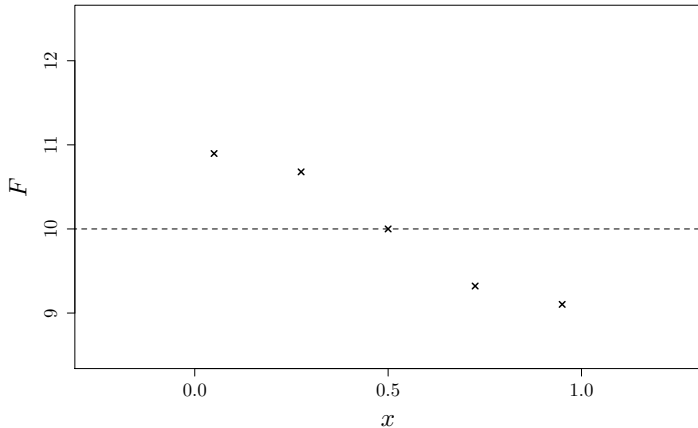
We will distinguish:

- **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

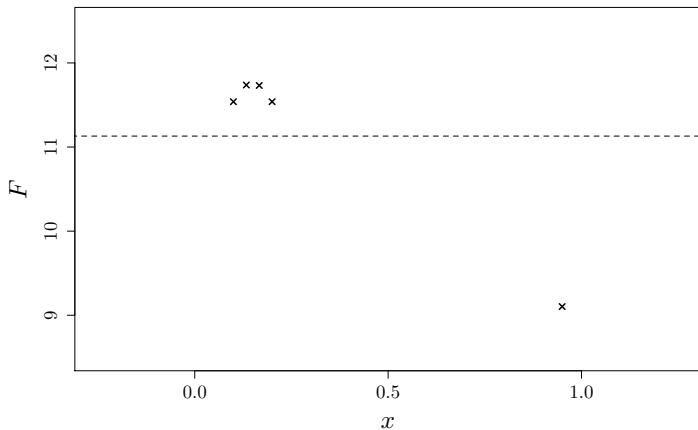
We will first focus on **ordinary kriging**. We thus need to estimate a constant:



The idea of considering $t(x) = \text{mean}(F)$ looks all right on this example...



but not on this one.



Any other idea?

We have considered maximum likelihood estimation for the kernel's parameter... why not doing the same thing here ?

Exercise

1. Compute the maximum likelihood estimation \hat{t} of t . A few hints
 - ▶ consider the log-likelihood
 - ▶ take the derivative
 - ▶ find where it is null
2. What can we recognize in this expression ?

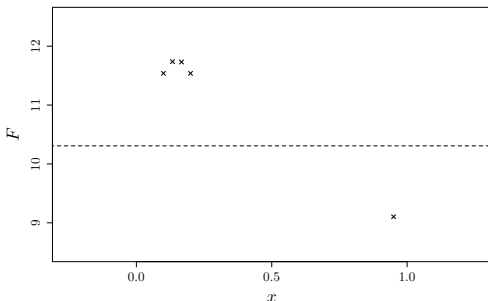
We recall that the likelihood is

$$L(t) = \frac{1}{(2\pi)^{n/2} |k(X, X)|^{1/2}} \exp \left(-\frac{1}{2} (x - t\mathbf{1})^t k(X, X)^{-1} (x - t\mathbf{1}) \right)$$

Solution

1. We obtain $\hat{t} = \frac{\mathbf{1}^t k(X, X)^{-1} F}{\mathbf{1}^t k(X, X)^{-1} \mathbf{1}}$
2. It can be seen as an orthogonal projection $t = \frac{\langle \mathbf{1}, F \rangle}{\langle \mathbf{1}, \mathbf{1} \rangle}$ for a inner product given by $k(X, X)^{-1}$.

On the previous example
we obtain $t = 10.3$:



Under the hypothesis $F = Z(X)$, the estimation

$$\hat{t} = \frac{\mathbf{1}^t k(X, X)^{-1} F}{\mathbf{1}^t k(X, X)^{-1} \mathbf{1}} \text{ is a sample from } T = \frac{\mathbf{1}^t k(X, X)^{-1} Z(X)}{\mathbf{1}^t k(X, X)^{-1} \mathbf{1}}.$$

The distribution of T is Gaussian with moments:

$$\mathbb{E}[T] = \frac{\mathbf{1}^t k(X, X)^{-1} \mathbb{E}[Z(X)]}{\mathbf{1}^t k(X, X)^{-1} \mathbf{1}} = t$$

$$\text{var}[T] = \frac{\mathbf{1}^t k(X, X)^{-1} \text{var}[Z(X)] k(X, X)^{-1} \mathbf{1}}{(\mathbf{1}^t k(X, X)^{-1} \mathbf{1})^2} = \frac{1}{\mathbf{1}^t k(X, X)^{-1} \mathbf{1}}$$

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

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

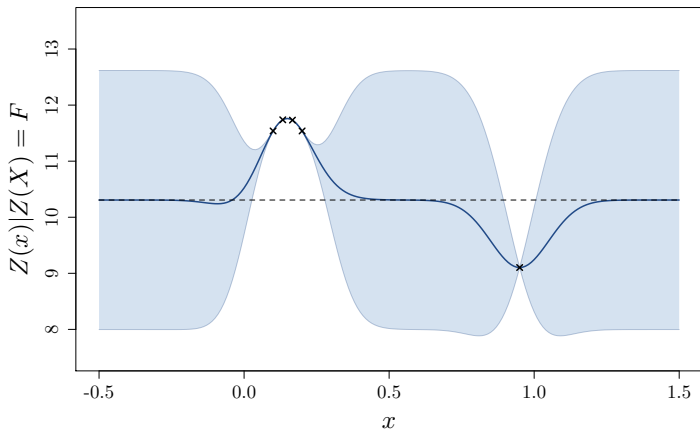
Regarding the **model variance**, it must account for the estimator's variance. We will use the law of total Variance :

$$\text{var}[X] = E[\text{var}(X|Y)] + \text{var}[E(X|Y)]$$

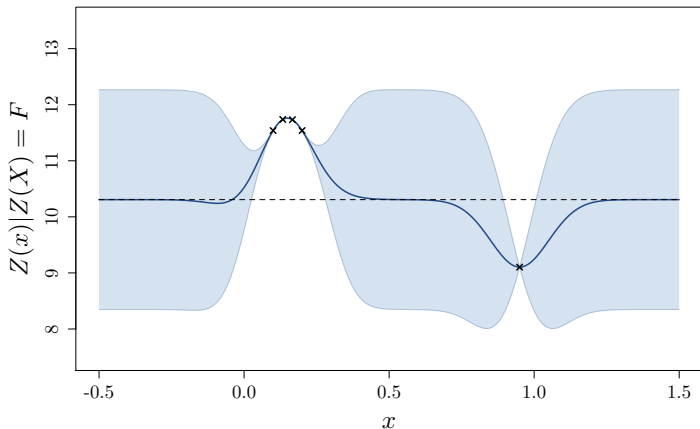
If we apply this to the GPR variance prediction we get:

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

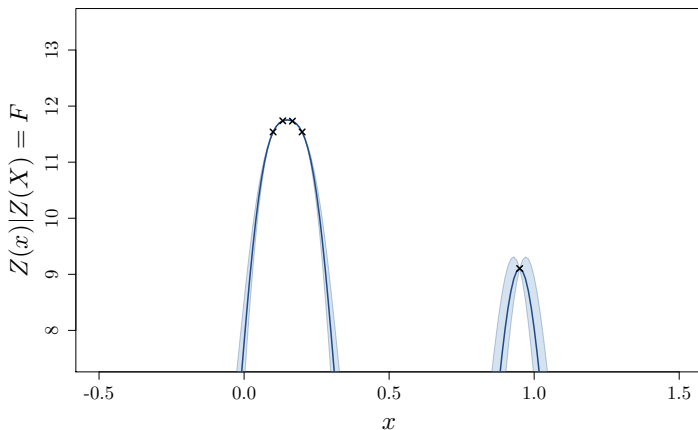
On the previous example we obtain:



We would have obtain this the mean were considered known.



it can be compared with simple kriging



If the trend is not constant but linear, quadratic, etc. it is interesting to consider the following probabilistic model for the prior:

$$Z(x) = Y(x) + \sum_i \beta_i h_i(x)$$

where the $h_i(x)$ are basis functions and the β_i are unknown scalars.

As previously, we can consider the maximum likelihood estimator

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

where H is the matrix of general term $H_{i,j} = h_j(X_i)$.

The final equations are very similar to ordinary kriging:

Universal kriging

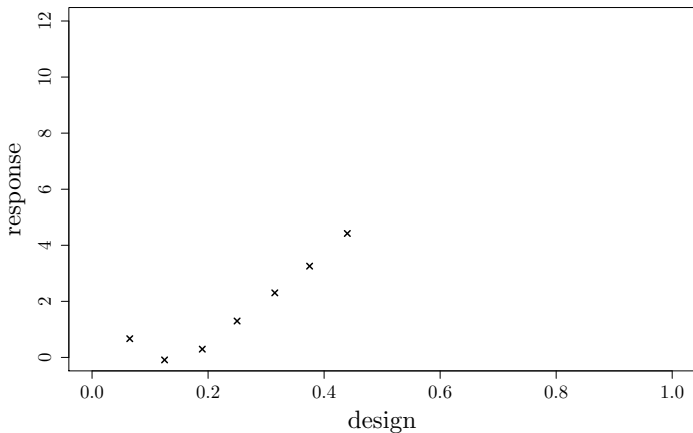
$$\begin{aligned}m(x) &= h(x)^t \hat{\beta} - k_x K^{-1} (F - h(X)^t \hat{\beta}) \\c(x, y) &= k(x, y) - k_x K^{-1} k_y^t \\&\quad + (h(x)^t + k_x K^{-1} H)^t (H^t K^{-1} H)^{-1} (h(y)^t + k_y K^{-1} H)\end{aligned}$$

where $k_x = k(x, X)$ and $K = k(X, X)$.

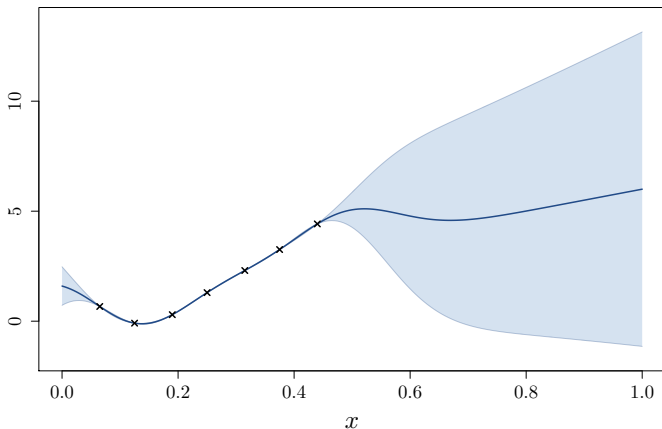
Remarks

- Ordinary kriging is a special case of universal kriging with only one constant basis function.
- The model always interpolates whatever $\hat{\beta}$ is.
- the trend part can be seen as generalised least square (regression with correlated residuals)

We consider the following example



Universal kriging model with linear trend: $h_1(x) = 1$, $h_2(x) = x$.



It can be compared to simple kriging with known trend

