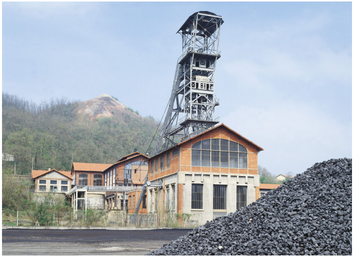
An introduction to Kriging metamodels

Didier Rullière
Preliminary version, please indicate any typo to didier.rulliere@emse.fr

December 2020 - PART I



picture: mining headframe (chevalement) at Saint-Etienne

Majeure Science des données, UP4

Acknowledgements

This course is
an overview of Kriging metamodeling and Gaussian Process Regression

This material is partly recycled from previous classes by Nicolas Durrande [2], Roldolphe Le Riche [5], Xavier Bay and many others, thanks a lot !

All errors are mine, do not hesitate to tell me.

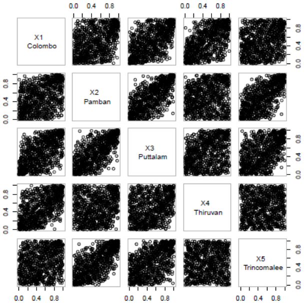
Outline

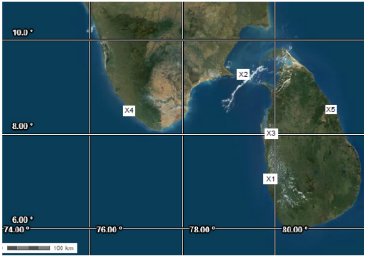
1. Introduction
2. The Statistical Approach
   - Simple Kriging
   - Other Kriging techniques
3. The Gaussian Process Approach
   - Random and Gaussian Processes
   - Gaussian process regression
   - Kriging noisy data

Introduction

Introduction example : rainfall data

An example of rainfall data in Sri Lanka





- How to predict rainfall somewhere, if it is only measured on few specific sites ?

QUIZZ Which sites exhibits more correlation ?

QUIZZ Is this in link with spatial distance between sites ?

QUIZZ How would you do to predict between two sites ?

The Origins of Kriging

- ... ok about rain but...
- How to predict gold concentration somewhere, if it is only measured on few specific sites ?

QUIZZ Who is this guy ?

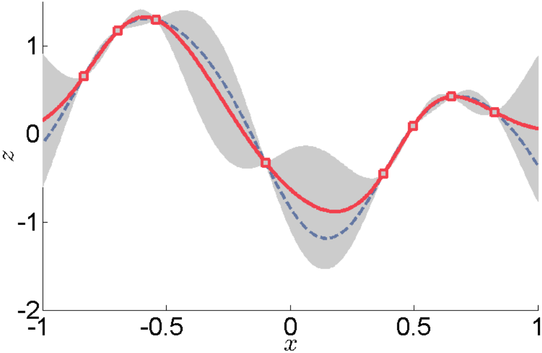
1. Danie Spline
2. Danie Krige
3. Danie Kernel

QUIZZ Where is this mining engineer in the picture ?

A. South Africa
B. Bermuda
C. Couriot Mine in Saint-Etienne

Kriging ?

"In statistics, originally in geostatistics, kriging or Gaussian process regression is a method of interpolation for which the interpolated values are modeled by a Gaussian process governed by prior covariances (...)". Wikipedia (citation and curve)



Mathematical formalization by Georges Matheron (Ecole des Mines de Paris, student of Paul Lévy) in *Mémoires du BRGM*.

Many possible applications

Use with computer experiments

Kriging is most often used in the context of expensive experiments (simulators)

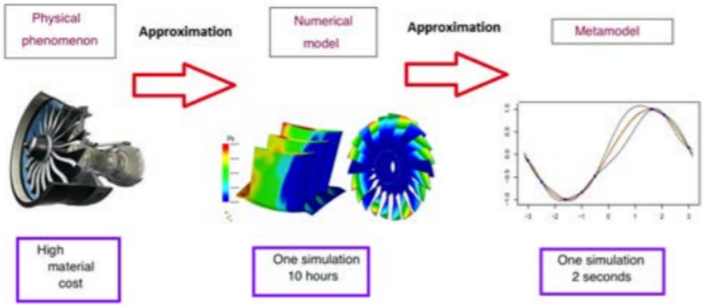


Illustration from your previous lecture Design of Experiment.

Many possible domains

Geostatistic (climate, mining)
Industry (crash tests, computer experiments)
Insurance (mortality tables, Economic Scenario Generator, nested simulations).

Context

Observations

Each experiment can be seen as a function of the input parameters.

input parameters $\in \chi \longrightarrow \boxed{\text{(computer/physical/...) experiment}} \longrightarrow$ output $\in \mathbb{R}$

so that $y = f(x)$ where f is a **costly to evaluate function**.

In the following, we will assume that

- $x \in \chi$: There are d input variables. Usually (but not necessarily) χ is $\mathbb{R}^d$.
- $y \in \mathbb{R}$: The output is a scalar. But extensions to GP regression with multiple outputs exist.

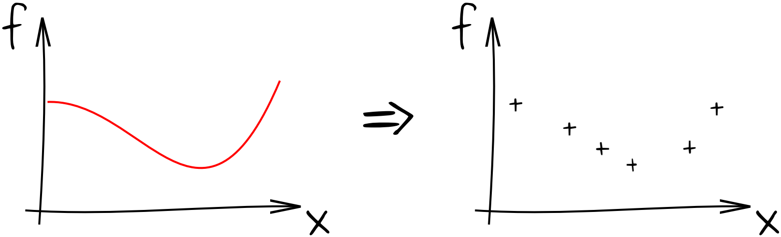
The interpolation problem

How to predict the output value for some new input parameters ?

f costly

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

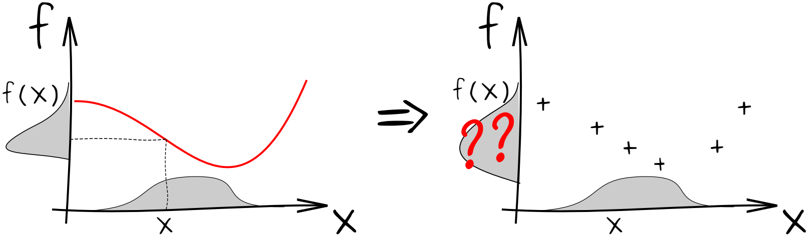
1. Representing the function is not possible...



f costly

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

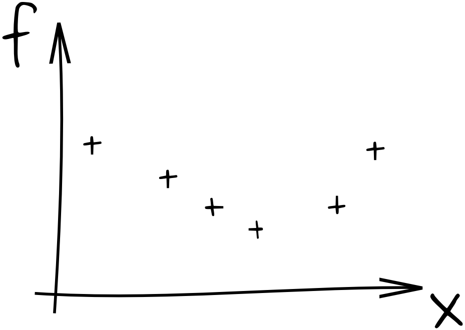
2. Uncertainty propagation is not possible...



f costly

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

3. Optimisation is also tricky...



4. Computing integrals is not possible...
5. Sensitivity analysis is not possible...

Need of a metamodel

Metamodel

Need to replace the costly f by a metamodel

- that can give a mean interpolation
- that can also measure the uncertainty associated with this interpolation

The presented one

Here we present Kriging metamodels, also known as Gaussian Process Regression (GPR) under some Gaussian assumptions.
Many other metamodels exist : splines, Inverse Distance Weighting, decomposition in basis functions, etc.
They are sometimes related to Kriging.

The Statistical Approach

Notations

Observations

set of possible sites :	$\chi = \mathbb{R}^d$	(e.g. rainfall $\chi = \mathbb{R}^2$)
n observations sites :	$\mathbb{X} = (x_1, \ldots, x_n) \in \chi^n$	(e.g. n city locations)
n observed responses :	$\mathbf{Y}_{\mathbb{X}} = (Y_{x_1}, \ldots, Y_{x_n})^\top \in \mathbb{R}^n$	(e.g. annual rainfall quantity)
indices :	$I = \{1, \ldots, n\}$	

Quantity of interest

One new prediction site :	$x \in \chi$	(e.g. one new city location)
Unknown response at this site :	$Y_x \in \mathbb{R}$	(e.g. rainfall to be predicted)

Assumptions

- all Y_{x_i} are random variables with finite mean and finite variance
- Covariances matrix $\mathbf{K} = (K_{ij})_{i,j \in I}$ and vector $\mathbf{k}_x = (k_i(x))_{i \in I}$ are known.
  where $K_{ij} = \mathrm{Cov}\left[Y_{x_i}, Y_{x_j}\right]$ covariance between responses,
  and $k_i(x) = \mathrm{Cov}\left[Y_{x_i}, Y_x\right]$ covariance with target Y_x.

Simple Kriging : the model

The idea

Most natural idea : your prediction is a linear combination of observed responses.

The Simple Kriging Model

One assumes $\mathbf{Y}_{\mathbb{X}}$ and Y_x centered : $\forall x, \mathrm{E}\left[Y_x\right] = 0$. Define a predictor $M(x)$ as

$$M(x) = \sum_{i=1}^{n} \alpha_i(x) Y_{x_i} \tag{1}$$

where weights $\boldsymbol{\alpha}_x = (\alpha_i(x))_{i=1..n}$ are minimizing

$$\Delta(x) = \mathrm{E}\left[(M(x) - Y_x)^2\right] \ . \tag{2}$$

QUIZZ Check that unbiasedness holds : $\mathrm{E}\left[M(x)\right] = \mathrm{E}\left[Y_x\right]$

Simple Kriging : calculations (1)

QUIZZ Step 1 : Develop $\Delta(x)$, express it as a function of covariances $\mathbf{K}$ and $\mathbf{k}_x$

Recall that $\mathbf{k}_x$ is the covariance vector between Y_x and the vector $\mathbf{Y}_{\mathbb{X}}$, and $\mathbf{K}$ is the covariance matrix of $\mathbf{Y}_{\mathbb{X}}$. Using $M(x) = \boldsymbol{\alpha}_x^\top \mathbf{Y}_{\mathbb{X}}$, let us develop

$$\Delta(x) = \mathrm{E}\left[(M(x) - Y_x)^2 \right] .$$

do your calculations here :

Simple Kriging : calculations (2)

Step 2 : find the weights $\boldsymbol{\alpha}_x$ that minimize $\Delta(x)$

Now let us minimize on $\boldsymbol{\alpha}_x$

$$\Delta(x) = \boldsymbol{\alpha}_x^\top \mathbf{K} \boldsymbol{\alpha}_x - 2\boldsymbol{\alpha}_x^\top \mathbf{k}_x + \text{constant}$$

do your calculations here :

Simple Kriging : Result (1)

Optimal weights

This leads to the vector of weights

$$\boldsymbol{\alpha}_x = \mathbf{K}^{-1}\mathbf{k}_x$$

where $\mathbf{k}_x$ is the covariance vector between Y_x and the vector $\mathbf{Y}_{\mathbb{X}}$, and $\mathbf{K}$ is the covariance matrix of $\mathbf{Y}_{\mathbb{X}}$.

Predictor and variance

From that follows the expression of $M(x)$ and $\Delta(x)$:

$$\left\{ \begin{array}{lcl} M(x) & = & \mathbf{k}_x^\top \mathbf{K}^{-1} \mathbf{Y}_{\mathbb{X}} \\ \Delta(x) & = & \sigma_x^2 - \mathbf{k}_x^\top \mathbf{K}^{-1} \mathbf{k}_x \end{array} \right.$$

Notice that $\Delta(x)$ does not depend on observed responses $\mathbf{Y}_{\mathbb{X}}$.

Simple Kriging : Result (1)

Results remain valid for q prediction points. Given a specific instance $\mathbf{Y}_{\mathbb{X}} = \mathbf{y}$, we get :

Simple Kriging

One assumes that $\mathbf{Y}_{\mathbb{X}}$ and Y_x are centered. Kriging mean corresponds to the Best Linear Unbiased Predictor of Y_x given $\mathbf{Y}_{\mathbb{X}} = \mathbf{y}$, and Kriging variance to the mean square error $\Delta(x)$:

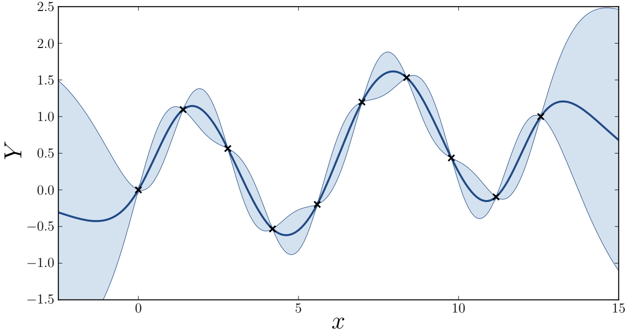
$$\left\{ \begin{array}{rcl} m(x) & = & \mathbf{k}_x^\top \mathbf{K}^{-1}\mathbf{y} \\ v(x) & = & \sigma_x^2 - \mathbf{k}_x^\top \mathbf{K}^{-1}\mathbf{k}_x \end{array} \right. \tag{3}$$

where $\mathbf{K} = \mathrm{Cov}\left[\mathbf{Y}_{\mathbb{X}}, \mathbf{Y}_{\mathbb{X}}\right]$ is $n \times n$ covariance matrix, and $\mathbf{k}_x = \mathrm{Cov}\left[\mathbf{Y}_{\mathbb{X}}, Y_{\mathbf{x}}\right]$ is a $n \times q$ covariance matrix.

QUIZZ At home, for $x, x' \in \chi$, determine $\Delta(x, x') = \mathrm{E}\left[(M(x) - Y_x)(M(x') - Y_{x'})\right]$, compare with $c(x, x')$ in the next section.

Simple Kriging : illustration

It can summarized by a mean function $m(x)$ and 95% confidence intervals corresponding to the variance $v(x)$ (under a distribution assumption).



QUIZZ The kriging predictor is interpolating $m(x_i) = Y_{x_i}$ for all i, why ?

Ordinary Kriging (1)

One assumes Y_{x_i}, $i \in I$ and Y_x have the same unknown mean μ. The predictor $M(x)$ writes as previously :

$$M(x) = \sum_{i=1}^{n} \alpha_i(x) Y_{x_i} \tag{4}$$

but unbiasedness condition $\mathrm{E}\left[M(x)\right] = \mathrm{E}\left[Y_{x_i}\right]$ implies $\sum_{i \in I} \alpha_i(x) = 1$.

QUIZZ Find the weights minimizing $\Delta(x) = \mathrm{E}\left[(Y_x - M(x))^2\right]$, subject to $\sum_{i \in I} \alpha_i(x) = 1$

Ordinary Kriging (2)

Using a Lagrange multiplier, we minimize in $\boldsymbol{\alpha}_x$

$$\Delta(x) - 2\lambda(\mathbf{1}^\top \boldsymbol{\alpha}_x - 1) = \boldsymbol{\alpha}_x^\top \mathbf{K} \boldsymbol{\alpha}_x - 2\boldsymbol{\alpha}_x^\top \mathbf{k}_x + \sigma_x^2 - 2\lambda(\mathbf{1}^\top \boldsymbol{\alpha}_x - 1) \tag{5}$$

after few calculation this gives

Ordinary Kriging

Under the assumption $\mathrm{E}\left[Y_{x_i}\right] = \mathrm{E}\left[Y_x\right] = \mu$, for all $i \in I$, Ordinary Kriging mean and variance are

$$\begin{cases} m(x) &= \boldsymbol{\alpha}_x^\top \mathbf{y} \\ v(x) &= \boldsymbol{\alpha}_x^\top \mathbf{K} \boldsymbol{\alpha}_x - 2\boldsymbol{\alpha}_x^\top \mathbf{k}_x + \sigma_x^2 \end{cases} \tag{6}$$

with $\boldsymbol{\alpha}_x = \mathbf{K}^{-1} \left(\mathbf{k}_x + \underbrace{\left(\frac{1 - \mathbf{k}_x^\top \mathbf{K}^{-1} \mathbf{1}}{\mathbf{1}^\top \mathbf{K}^{-1} \mathbf{1}} \right)}_{=\lambda} \cdot \mathbf{1} \right)$.

Ordinary Kriging can be seen as a Simple Kriging on residuals, with :

$$\begin{cases} \hat{\mu} &= \left(\mathbf{1}^\top \mathbf{K}^{-1} \mathbf{1}\right)^{-1} \mathbf{1}^\top K^{-1} \mathbf{Y}_{\mathbb{X}} \\ m(x) &= \hat{\mu} + \mathbf{k}_x^\top \mathbf{K}^{-1} (\mathbf{Y}_{\mathbb{X}} - \hat{\mu} \mathbf{1}) \end{cases} \tag{7}$$

Universal Kriging

Consider given matrices of factors, e.g. $F(\mathbb{X}) = (\mathbf{1}, \mathbb{X})$ and $F(x) = (1, x)$.
The universal Kriging predictor writes

$$M(x) = F(x)^\top \beta + \sum_{i=1}^{n} \alpha_i(x) Y_{x_i} \tag{8}$$

The vector β does not depend on x. One can show (*Sacks et al.*, 1989) :

Universal Kriging

The optimal coefficients β and $\alpha(x)$ are the same as those obtained by :

1. doing a linear regression $\mathbf{Y}_\mathbb{X} = F(\mathbb{X})^\top \beta + \epsilon$ to estimate the β_i's

$$\hat{\beta} = \left(F(\mathbb{X})^\top \mathbf{K}^{-1} F(\mathbb{X}) \right)^{-1} F(\mathbb{X})^\top \mathbf{K}^{-1} \mathbf{Y}_\mathbb{X}$$

2. then doing a Simple Kriging on residuals

... so that no other results are needed ☺.

QUIZZ What happens when $F(\mathbb{X}) = \mathbf{1}$?

Advantages of the Statistical Approach

Some advantages of the “statistical approach” (compared to other approaches)

Pro

- General : only requires random variables with two moments, no Gaussian assumption, manipulate only finite vectors
- Can be extended with other regression techniques : penalizations (LASSO, ridge), cross effects, quadratic terms, link functions...

$$M(x) = \sum_i \alpha_i(x) Y_{x_i} - \lambda \left| \sum_i \alpha_i(x) \right|$$

$$M(x) = f\left(Y_{x_1}, \ldots, Y_{x_n}, \boldsymbol{\alpha}\right)$$

- Can be nested using other estimators

$$M(x) = \sum_i \alpha_i(x) M_i(x)$$

Cons

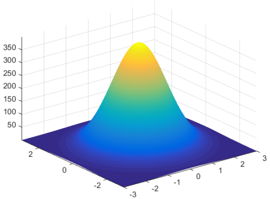
- Interpretation : No direct interpretation as a conditional process
- Theroretical : Conditional quantities sometimes hard to derive

The Gaussian Process Approach

Gaussian Vectors

A Gaussian Vector $\mathbf{Y}$ with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ is a random vector with density

$$f_{\mathbf{Y}}(y_1, \ldots, y_d) = \frac{1}{\sqrt{(2\pi)^d \det \boldsymbol{\Sigma}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \qquad (9)$$



- Non-degenerate if $\boldsymbol{\Sigma}$ definite positive : $\forall \mathbf{a}$ non zero, $\mathbf{a}^\top \boldsymbol{\Sigma} \mathbf{a} > 0$.
- Linear combinations of components of $\mathbf{Y}$ are Gaussian,
- thus components Y_i are Gaussian, $i = 1, \ldots, d$ (reverse not true).

Conditional Gaussian Vectors

Let $\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{bmatrix}$ be Gaussian with mean $\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}$ and covariance $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$, then

Conditional Gaussian Vector

The conditional distribution of $\mathbf{Y}_1$ given $\mathbf{Y}_2 = \mathbf{y}_2$ is Gaussian with mean and covariance

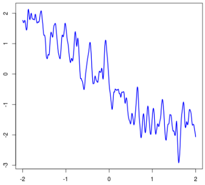
$$\left\{ \begin{array}{lcl} \boldsymbol{\mu}_{2|1} & = & \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{y}_2 - \boldsymbol{\mu}_2) \\ \boldsymbol{\Sigma}_{2|1} & = & \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}. \end{array} \right. \tag{10}$$

Random Process

A random process is a set of RV's indexed by $x \in \chi$

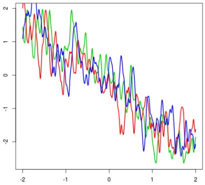
random event $\omega \in \Omega$
(e.g., weather)

$\Longrightarrow$



Repeat the random event ($3\times$)

$\Longrightarrow$



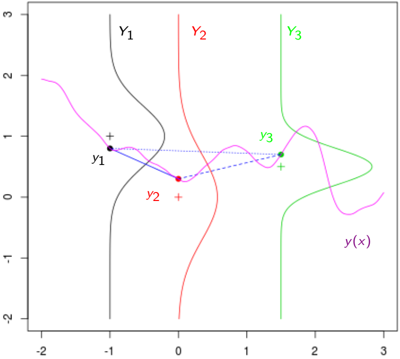
This creates 3 trajectories $y(x)$'s. They are different, yet bear strong similarities.

Gaussian Process (1) : definition

Gaussian Process : one possible definition

A stochastic process is Gaussian $\Longleftrightarrow$ all finite subvectors are Gaussian

- implies that for any $x \in \chi$, Y_x is a Gaussian RV (reverse not true).
- implies that any finite linear combination of some Y_x's is Gaussian.



Gaussian Process (2) : characterisation

For such a Gaussian Process (GP), we denote

$$k(x, x') = \mathrm{Cov}\left[Y_x, Y_{x'}\right].$$

Gaussian Process (2) : characterisation

The distribution of a GP is fully characterised by :

- its mean function $\mu : \chi \to \mathbb{R}$:

$$\mu(x) = \mathrm{E}\left[Y_x\right]$$

- its covariance function, or *kernel*, $k : \chi \times \chi \to \mathbb{R}$:

$$k(x, x') = \mathrm{Cov}\left[Y_x, Y_{x'}\right]$$

In particular, $\forall \mathbb{X} = \begin{pmatrix} x_1 \\ \cdots \\ x_n \end{pmatrix} \in \chi^n$, $\mathbf{Y}_{\mathbb{X}} = \begin{pmatrix} Y_{x_1} \\ \cdots \\ Y_{x_n} \end{pmatrix} \sim \mathcal{N}(\mu(\mathbb{X}), \mathbf{K})$,

where $\mathbf{K} = (K_{ij})_{i,j \in I}$, $K_{ij} = \mathrm{Cov}(Y_{x_i}, Y_{x_j}) = k(x_i, x_j)$.

Gaussian Process (3) : covariance function

Conditions

... but conditions hold (detailed later) for the covariance function $k(.,.)$!

QUIZZ Should $k(x,x') = k(x',x)$? why ?

QUIZZ For any $\boldsymbol{a}$, variance of the random variable $\boldsymbol{a}^\top \mathbf{Y}_{\mathbb{X}}$? consequence on $k(.,.)$?

One example (for the moment)

The *Gaussian kernel*, or *Squared Exponential (SE) covariance function* :

$$k(x,x') = \sigma^2 \exp\left(-\frac{1}{2\theta^2}||x-x'||_2^2\right)$$

has two parameters, the variance σ^2 and the lengthscale θ.

Matrix notations for kernels

for two vectors $\mathbf{u} \in \chi^q$, $\mathbf{v} \in \chi^n$, we often use the matrix notation :

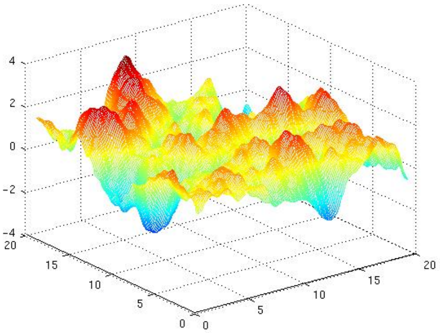
$$k(\mathbf{u},\mathbf{v}) = (k(u_i,u_j))_{i=1,\dots,q;j=1,\dots,n} \quad \in \mathbb{R}^{q\times n}$$

- $\mathbf{K} = k(\mathbb{X},\mathbb{X}) \in \mathbb{R}^{n\times n}$,
- $\mathbf{k}_x = k(\mathbb{X},x) = k(x,\mathbb{X})^\top \in \mathbb{R}^{n\times 1}$.

Gaussian Process (4) : random fields

On previous illustrations $x \in \mathbb{R}$, so that trajectories are functions $\mathbb{R} \to \mathbb{R}$.

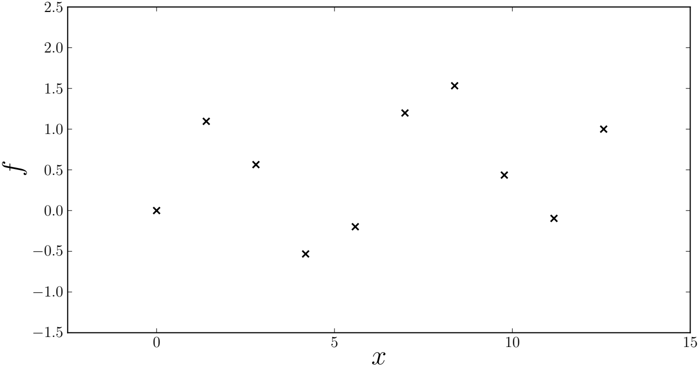
When $x \in \mathbb{R}^d$, with $d > 1$, trajectories are functions $\mathbb{R}^d \to \mathbb{R}$.



Nothing is changed, but we sometimes call the process a Gaussian Random Field.

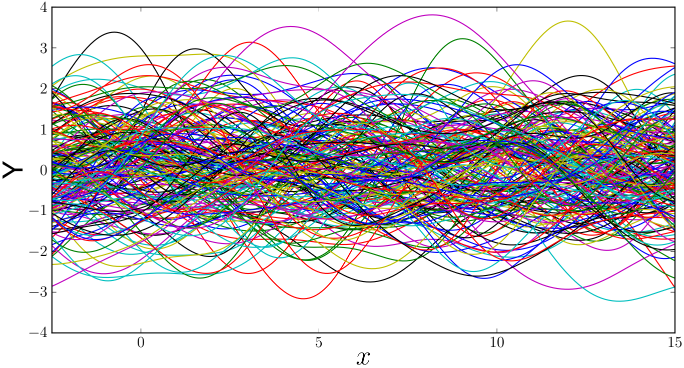
Gaussian process regression

Assume we have observed a function $f()$ over a set of points $X = (x_1, \ldots, x_n)$:



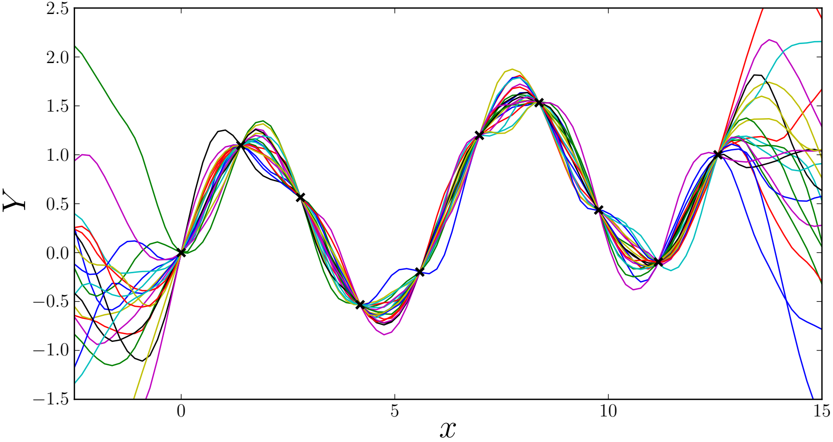
The vector of observations is $\mathbf{y} = f(\mathbb{X})$, i.e. $y_i = f(x_i)$.

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

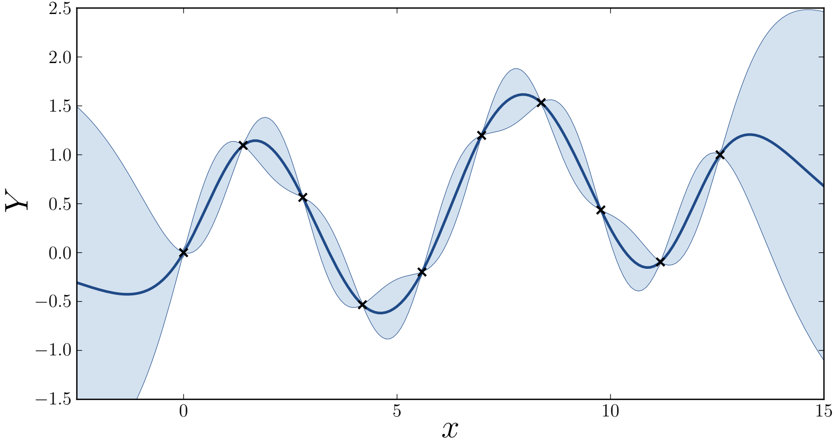


(here $\mu(x) = 0$)

If we remove all the samples that do not interpolate the observations we obtain :



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



Kriging equations (1/2)

The conditional distribution can be obtained analytically :

By definition, $(Y_x, \mathbf{Y}_{\mathbb{X}})$ is multivariate normal. Formulas on the conditioning of Gaussian vectors, in Equation (10), give the distribution of $Y_x|\mathbf{Y}_{\mathbb{X}} = \mathbf{y}$. It is $\mathcal{N}(m(.), c(.,.))$ with :

$$\begin{aligned} m(x) &= \mathrm{E}\left[Y_x|\mathbf{Y}_{\mathbb{X}}{=}\mathbf{y}\right] \\ &= \mu(x) + k(x,\mathbb{X})k(\mathbb{X},\mathbb{X})^{-1}(\mathbf{y} - \mu(\mathbb{X})) \\ c(x,x') &= \mathrm{Cov}\left[Y_x, Y_{x'}|\mathbf{Y}_{\mathbb{X}}{=}\mathbf{y}\right] \\ &= k(x,x') - k(x,\mathbb{X})k(\mathbb{X},\mathbb{X})^{-1}k(\mathbb{X},x') \end{aligned}$$

Simple Kriging, Gaussian case

For a centered process, when $\mu(x) = \mu(\mathbb{X}) = 0$, the simple Kriging predictor in Equation (3) corresponds to

$$\begin{cases} m(x) & = & \mathrm{E}\left[Y_x|\mathbf{Y}_{\mathbb{X}}{=}\mathbf{y}\right] \\ v(x) & = & \mathrm{V}\left[Y_x|\mathbf{Y}_{\mathbb{X}}{=}\mathbf{y}\right] \end{cases}$$

Kriging equations (2/2)

Summary

The distribution of $Y_x|\mathbf{Y}_{\mathbb{X}} = \mathbf{y}$ is $\mathcal{N}(m(.), c(., .))$ with mean and covariance

$$\begin{cases} m(x) & = & \mu(x) + k(x,\mathbb{X})k(\mathbb{X},\mathbb{X})^{-1}(\mathbf{y} - \mu(\mathbb{X})) \\ c(x,x') & = & k(x,x') - k(x,\mathbb{X})k(\mathbb{X},\mathbb{X})^{-1}k(\mathbb{X},x') \end{cases}$$

- $k(\mathbb{X},\mathbb{X}) = [k(x_i, x_j)]$: covariance matrix between observed outputs, sometimes named Gram matrix. It is of size $n \times n$.
- $k(x,\mathbb{X}) = [k(x, x_i), \dots, k(x, x_n)]$: $n \times 1$ covariance vector between observed output and target Y_x.

Remarks

- It is a Gaussian distribution : gives confidence intervals, can be sampled, this is actually how the previous slides were generated.
- It is Bayesian : $Y_x|\mathbf{Y}_{\mathbb{X}} = \mathbf{y}$ is the posterior distribution of Y_x once $\mathbf{Y}_{\mathbb{X}} = \mathbf{y}$ is observed.
- It is named *Gaussian Process Regression*, often identified with the term *Kriging*.

Properties

A few remarkable properties of Gaussian Process Regression (GPR) models

- They (can) interpolate the data-points
- The prediction variance does not depend on the observations **y**
- The mean predictor does not depend on the scale of variances parameters
- They (usually) come back to the a priori trend $\mu(x)$ when we are far away from the observations.

proofs left as exercise

Kriging of noisy data

An important special case, noisy data $\mathbf{Z}_{\mathbb{X}} = \mathbf{Y}_{\mathbb{X}} + \epsilon_{\mathbb{X}}$.
where $\varepsilon(.) \sim \mathcal{N}(0, n(.,.))$ independent of $Y(.)$. Then,

$$\begin{cases} \text{Cov}(Y_{x_i} + \epsilon_{x_i}, Y_{x_j} + \epsilon_{x_j}) & = & k(x_i, x_j) + n(x_i, x_j) \\ \text{Cov}(Y_x, Y_{x_i} + \epsilon_{x_i}) & = & k(x, x_i) \end{cases}$$

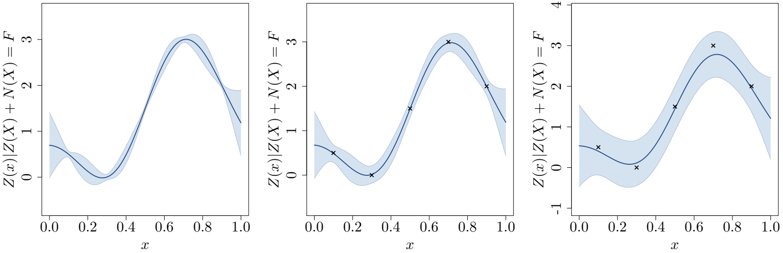
The expressions of GPR with noise become (just apply Gaussian vector conditioning with the above), when $x \notin \mathbb{X}$ and ϵ_x, $\epsilon_{x'}$, $\epsilon_{\mathbb{X}}$ mutually independent.

$$\begin{aligned} m(x) &= \mathrm{E}\left[Z_x | \mathbf{Z}_{\mathbb{X}}{=}\mathbf{z}\right] \\ &= \mu(x) + k(x, \mathbb{X})\left(k(\mathbb{X}, \mathbb{X}) + n(\mathbb{X}, \mathbb{X})\right)^{-1}(\mathbf{z} - \mu(\mathbb{X})) \\ c(x, x') &= \mathrm{Cov}\left[Z(x), Z(x') | Z(\mathbb{X}){=}\mathbf{z}\right] \\ &= k(x, x') - k(x, \mathbb{X})\left(k(\mathbb{X}, \mathbb{X}) + n(\mathbb{X}, \mathbb{X})\right)^{-1} k(\mathbb{X}, x') \end{aligned}$$

Remarks

- this is the same distribution as the one of $Y_x | \mathbf{Z}_{\mathbb{X}} = \mathbf{z}$, $x \notin \mathbb{X}$.
- usually $n(\mathbb{X}, \mathbb{X})$ diagonal, called nugget effect.
- can be used to help the inversion of $k(\mathbb{X}, \mathbb{X})$

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



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

Kriging with noise kernel (nugget) does not interpolate the data.

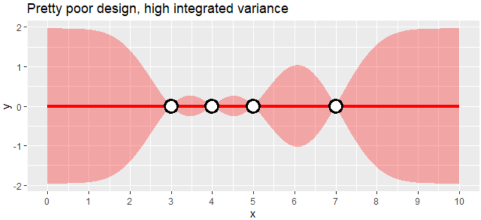
A small τ^2 (e.g., 10^{-10}) often used to make the covariance matrix invertible (more on regularization of GPs in [6]).

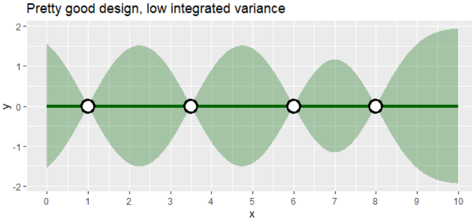
Kriging based Design

Write here $k(\mathbb{X},\mathbb{X}) = \mathrm{Cov}\left[\mathbf{Y}_{\mathbb{X}}, \mathbf{Y}_{\mathbb{X}}\right] = \mathbf{K}$ and $k(\mathbb{X},x) = \mathrm{Cov}\left[\mathbf{Y}_{\mathbb{X}}, Y_x\right] = k(x,\mathbb{X})^\top = \mathbf{k}_x$.

Lower prediction variance is better : optimize $\mathbb{X}$ in order to minimize the sum of prediction variances over χ for a given measure μ.

$$IMSE(\mathbb{X}) = \int_{\chi} v(x) d\mu(x) = \int_{\chi} \left(\sigma_x^2 - k(x,\mathbb{X})k(\mathbb{X},\mathbb{X})^{-1}k(x,\mathbb{X})\right) d\mu(x)$$


Pretty poor design, high integrated variance


Pretty good design, low integrated variance

Other criterions : maximizing entropy or Mutual Information *Krause et al.*, 2008.

to be continued...

in the rest of the lecture, we will detail more results on

- Kernels, covariance functions.
- (Hyper)-parameters estimation.

Some references I

[1] Abrahamsen, P. (1997). A review of gaussian random fields and correlation functions.

[2] Durrande, N. and Le Riche, R. (2017). Introduction to Gaussian Process Surrogate Models. Lecture at 4th MDIS form@ter workshop, Clermont-Fd, France. HAL report cel-01618068.

[3] Hensman, J., Fusi, N., and Lawrence, N. D. (2013). Gaussian processes for big data. *arXiv preprint arXiv :1309.6835*.

[4] Krause, A., Singh, A., and Guestrin, C. (2008). Near-optimal sensor placements in gaussian processes : Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9(Feb) :235–284.

[5] Le Riche, R. (2014). Introduction to Kriging. Lecture at mnmuq2014 summer school, Porquerolles, France. HAL report cel-01081304.

[6] Le Riche, R., Mohammadi, H., Durrande, N., Touboul, E., and Bay, X. (2017). A Comparison of Regularization Methods for Gaussian Processes. slides of talk at siam conference on optimization op17 and accompanying technical report hal-01264192, Vancouver, BC, Canada. `https://www.emse.fr/~leriche/op17_R_LeRiche_slides_v2.pdf` and `https://hal.archives-ouvertes.fr/hal-01264192`.

Some references II

[7] López-Lopera, A. F., Bachoc, F., Durrande, N., and Roustant, O. (2018). Finite-dimensional gaussian approximation with linear inequality constraints. *SIAM/ASA Journal on Uncertainty Quantification*, 6(3) :1224–1255.

[8] Rasmussen, C. E. (2003). Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer.

[9] Roustant, O., Ginsbourger, D., and Deville, Y. (2012). DiceKriging, DiceOptim : Two R packages for the analysis of computer experiments by kriging-based metamodeling and optimization. *Journal of Statistical Software*, 51(1).

[10] Rullière, D., Durrande, N., Bachoc, F., and Chevalier, C. (2018). Nested kriging predictions for datasets with a large number of observations. *Statistics and Computing*, 28(4) :849–867.

[11] Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989). Design and analysis of computer experiments. *Statistical science*, pages 409–423.