Mini-course 1: lecture
Introduction to Gaussian processes

François Bachoc

Institut de Mathématiques de Toulouse
Université Paul Sabatier
Institut universitaire de France (IUF)

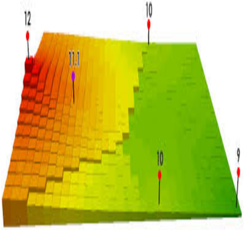
Workshop Gaussian processes and related topics
Toulouse
July 2025

Outline of the course

Gaussian processes in different fields

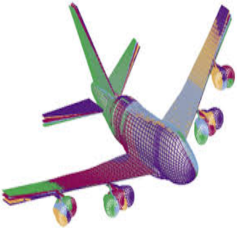
Gaussian processes are studied in different fields :

geostatistics



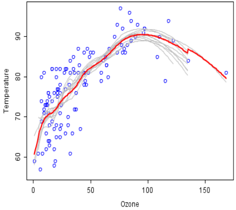
Stein, 99

computer experiments



Santner et al, 03

machine learning



Rasmussen and Williams, 06

Common ground but also

- Different type of data
- Different algorithms
- Different theoretical focus
- Different vocabulary

Canonical goal : learning an unknown function

We are interested in learning a fixed unknown function

$$\begin{aligned} f: \mathbb{X} &\to \mathbb{R} \\ x &\mapsto f(x) \end{aligned}$$

- $\mathbb{X}$: input space (no assumption so far)
- x : input parameter
- $f(x)$: quantity of interest

The function f is a black box

$\Longrightarrow$ Only available through observations

$\Longrightarrow$ No or few a priori information available

Examples :

- Geostatistics : x is a two-dimensional position and $f(x)$ is a pollutant concentration
- Computer experiments : x is a simulation parameter and $f(x)$ is a simulation result
- Machine learning : x is a set of flight features and $f(x)$ is a delay time

Various types of observations of f

Regression

- Exact observations : We observe $f(x_1), \dots, f(x_n)$
- Noisy observations : We observe $f(x_1) + \epsilon_1, \dots, f(x_n) + \epsilon_n$
  f can be interpreted as a conditional expectation

Binary classification

- We observe $Y_1, \dots, Y_n$ where, for $i = 1, \dots, n$, $Y_i \in \{0, 1\}$ and

$$\mathbb{P}(Y_i = 1) = \phi(f(x_i)),$$

  with ϕ strictly increasing from $(-\infty, \infty)$ to $(0, 1)$
  E.g. logistic function $\phi(t) = e^t/(1 + e^t)$

And more : multiclass classification, f gives the intensity of a point process,...

The role of Gaussian processes

The previous types of observations can be tackled by several statistics or machine learning algorithms

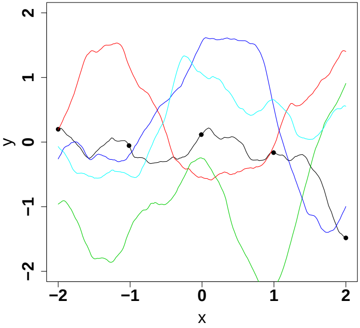
- Kernel smoothing
- Random forests
- Neural networks
- and many more

Gaussian processes also tackle these types of observations and are based on a Bayesian prior on the function f

$\Longrightarrow$ Hence they provide an important benefit for uncertainty quantification

Bayesian prior

Modeling the **black box function** f as a **single realization** of a Gaussian process $x \to \xi(x)$ on the domain $\mathbb{X}$



Usefulness

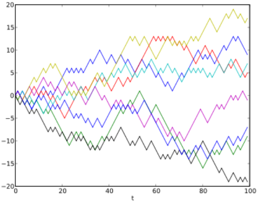
Using the conditional distribution of ξ, given the **observations**, to learn f

A quick summary

Gaussian processes provide a Bayesian prior over unknown functions, that enables to address various machine learning problems, with the benefit of uncertainty quantification

Stochastic processes

A stochastic process on $\mathbb{X}$ is a function $\xi : \mathbb{X} \to \mathbb{R}$ such that $\xi(x)$ is a random variable for all $x \in \mathbb{X}$.
Alternatively a stochastic process is a function on $\mathbb{X}$ that is random

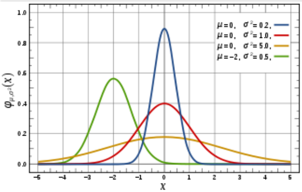


Probability space

We explicit the randomness of $\xi(x)$ by writing it $\xi(\omega, x)$ with ω in a probability space Ω.
For a given ω_0, we call the function $x \to \xi(\omega_0, x)$ a realization of the stochastic process ξ.
$\Longrightarrow$ The probability space Ω is the same for all $\xi(\omega, x)$ with $x \in \mathbb{X}$

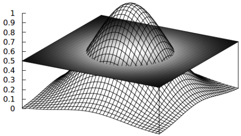
A random variable X on $\mathbb{R}$ is a Gaussian variable with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$ when its probability density function is

$$f_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$



A n-dimensional random vector $\boldsymbol{V}$ is a Gaussian vector with mean vector $\boldsymbol{m}$ and invertible covariance matrix $\boldsymbol{R}$ when its multidimensional probability density function is

$$f_{\boldsymbol{m},\boldsymbol{R}}(\boldsymbol{v}) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{det(\boldsymbol{R})}} \exp\left(-\frac{1}{2}(\boldsymbol{v}-\boldsymbol{m})^\top \boldsymbol{R}^{-1}(\boldsymbol{v}-\boldsymbol{m})\right)$$



Characterization by mean and variance

E.g. for Gaussian variables : μ and σ^2 are both parameters of the probability density function and the mean and variances of it. That is $\int_{-\infty}^{+\infty} x f_{\mu,\sigma^2}(x)dx = \mu$ and $\int_{-\infty}^{+\infty} (x-\mu)^2 f_{\mu,\sigma^2}(x)dx = \sigma^2$

A random variable X that is constant equal to μ is said to be a Gaussian variable with mean μ and variance $\sigma^2 = 0$

A n-dimensional random vector $\boldsymbol{V}$ is a Gaussian vector with mean vector $\boldsymbol{m}$ and covariance matrix $\boldsymbol{R}$ when, for any fixed $n \times 1$ vector $\boldsymbol{\lambda}$, $\boldsymbol{\lambda}^\top \boldsymbol{V}$ is a Gaussian variable with mean $\boldsymbol{\lambda}^\top \boldsymbol{m}$ and variance $\boldsymbol{\lambda}^\top \boldsymbol{R} \boldsymbol{\lambda}$

- This definition holds whether or not $\boldsymbol{R}$ is invertible

$\Longrightarrow$ All linear combinations of Gaussian vectors are Gaussian variables

- When $\boldsymbol{R}$ is not invertible, $\boldsymbol{V}$ is supported on a lower dimensional linear subspace of $\mathbb{R}^n$

Definition

A stochastic process ξ on $\mathbb{X}$ is a Gaussian process when for all $x_1, ..., x_n \in \mathbb{X}$, the random vector $(\xi(x_1), ..., \xi(x_n))$ is a Gaussian vector

Mean and covariance functions

- The mean function of a Gaussian process ξ is the function

$$\begin{aligned} m \colon \mathbb{X} &\to \mathbb{R} \\ x &\mapsto \mathbb{E}(\xi(x)) \end{aligned}$$

- The covariance function of a Gaussian process ξ is the function

$$\begin{aligned} k \colon \mathbb{X} \times \mathbb{X} &\to \mathbb{R} \\ (x_1, x_2) &\mapsto Cov(\xi(x_1), \xi(x_2)) \end{aligned}$$

$\Longrightarrow$ A Gaussian process is characterized by its mean and covariance functions

Constraints on the covariance function

First, remark that k is symmetric :

$$k(x_1, x_2) = Cov(\xi(x_1), \xi(x_2)) = Cov(\xi(x_2), \xi(x_1)) = k(x_2, x_1)$$

Second, let ξ be a Gaussian process on a set $\mathbb{X}$, with covariance function k
Consider $x_1, \ldots, x_n \in \mathbb{X}$ and $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ to be fixed
We have

$$\begin{aligned} 0 \leq Var\left(\sum_{i=1}^{n} \lambda_i \xi(x_i)\right) \\ = \sum_{i,j=1}^{n} \lambda_i \lambda_j \, Cov(\xi(x_i), \xi(x_j)) \\ = \sum_{i,j=1}^{n} \lambda_i \lambda_j k(x_i, x_j) \end{aligned}$$

$\Longrightarrow$Hence a second constraint on k

Constraints on the covariance function

Symmetric non-negative definite functions

A function $h : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ is symmetric non-negative definite (SNND) if

- For any $x_1, x_2 \in \mathbb{X}$:

$$h(x_1, x_2) = h(x_2, x_1)$$

- For any $x_1, \ldots, x_n \in \mathbb{X}$ and $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$:

$$\sum_{i,j=1}^{n} \lambda_i \lambda_j h(x_i, x_j) \geq 0$$

$\Longrightarrow$Covariance functions are SNND

Alternatively, for any $x_1, \ldots, x_n \in \mathbb{X}$, the $n \times n$ covariance matrix $\boldsymbol{R} = [k(x_i, x_j)]_{i,j=1,\ldots,n}$ of the Gaussian vector $(\xi(x_1), \ldots, \xi(x_n))$ is symmetric non-negative definite

Hence, covariance functions can also be called

- kernels
- radial basis functions
- non-negative definite functions

Existence of Gaussian processes

Theorem

- Let $\mathbb{X}$ be any set
- Let m be any function from $\mathbb{X}$ to $\mathbb{R}$
- Let k be any SNND function from $\mathbb{X} \times \mathbb{X}$ to $\mathbb{R}$

Then there exists a Gaussian process ξ on $\mathbb{X}$ with mean function m and covariance function k

Proof : Kolmogorov extension theorem □

Hence

- To create a Gaussian process it is sufficient to create a mean and covariance function
- Any function can be a mean function
- The crux is thus to create SNND functions

Next :

1. Creation of covariance (SNND) functions and interplay with behavior of the Gaussian process
2. Given a mean and covariance function $\longrightarrow$ conditional distribution of the Gaussian process given observations
3. Estimating the mean and covariance functions

Two extreme covariance functions

Let $\mathbb{X}$ be any set

Constant covariance function

Let the function $k_1 : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ be defined by, for any $x_1, x_2 \in \mathbb{X}$,

$$k_1(x_1, x_2) = 1$$

Then k_1 is *SNND*
A Gaussian process ξ with mean zero and covariance function k_1 is constant :

$$\text{for all } x \in \mathbb{X}, \xi(x) = X,$$

where $X \sim \mathcal{N}(0, 1)$

White noise covariance function

Let the function $k_2 : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ be defined by, for any $x_1, x_2 \in \mathbb{X}$,

$$k_2(x_1, x_2) = \mathbf{1}_{\{x_1 = x_2\}}$$

Then k_2 is *SNND*
A Gaussian process ξ with mean zero and covariance function k_2 is composed of independent Gaussian values

Covariance functions on $\mathbb{R}^d$

Let $\mathbb{X} = \mathbb{R}^d$

Stationarity

A covariance function k is stationary when for any $x_1, x_2 \in \mathbb{R}^d$:

$$k(x_1, x_2) = k(x_1 - x_2)$$

(slight abuse of notation)

$\Longrightarrow$ The behavior of the corresponding Gaussian process is invariant by translation

Bochner's theorem

Consider a continuous function $k : \mathbb{R}^d \to \mathbb{R}$ with Fourier transform $\hat{k}$, such that the inverse Fourier relation holds :

$$\text{for all } x \in \mathbb{R}^d,\ k(x) = \int_{\mathbb{R}^d} \hat{k}(\omega) e^{\mathrm{i}\omega^\top x} d\omega$$

Then k is SNND if and only if $\hat{k}$ takes positive values

$\Longrightarrow$ A convenient characterization of stationary covariance functions

Proof of one implication of Bochner's theorem

Assume that $\hat{k}$ takes positive values
For all $x_1, ..., x_n \in \mathbb{X}$, $\lambda_1, ..., \lambda_n \in \mathbb{R}$:

$$\begin{aligned}
\sum_{i,j=1}^{n} \lambda_i \lambda_j k(x_i, x_j) &= \sum_{i,j=1}^{n} \lambda_i \lambda_j k(x_i - x_j) \\
&= \sum_{i,j=1}^{n} \lambda_i \lambda_j \int_{\mathbb{R}^d} \hat{k}(\omega) e^{\mathrm{i}\omega^\top (x_i - x_j)} d\omega \\
&= \int_{\mathbb{R}^d} \hat{k}(\omega) \left(\sum_{i,j=1}^{n} \lambda_i \lambda_j e^{\mathrm{i}\omega^\top x_i} e^{-\mathrm{i}\omega^\top x_j} \right) d\omega \\
&= \int_{\mathbb{R}^d} \hat{k}(\omega) \left(\sum_{i,j=1}^{n} \lambda_i e^{\mathrm{i}\omega^\top x_i} \overline{\lambda_j e^{\mathrm{i}\omega^\top x_j}} \right) d\omega \\
&= \int_{\mathbb{R}^d} \hat{k}(\omega) \left| \sum_{i=1}^{n} \lambda_i e^{\mathrm{i}\omega^\top x_i} \right|^2 d\omega \\
&\geq 0
\end{aligned}$$

Hence k is SNND □

Hence some stationary covariance functions on $\mathbb{R}$

- **Exponential covariance function**

$$k(x_1, x_2) = \sigma^2 e^{-|x_1 - x_2|/\ell}$$

$\Longrightarrow$ parametrized by variance σ^2 and correlation length ℓ
(positive Fourier transform)

- **Square exponential (or Gaussian) covariance function**

$$k(x_1, x_2) = \sigma^2 e^{-(x_1 - x_2)^2/\ell^2}$$

(positive Fourier transform)

- **Matérn covariance function**

$$k(x_1 - x_2) = \frac{\sigma^2}{\Gamma(\nu)2^{\nu-1}} \left(\frac{2\sqrt{\nu}|x_1 - x_2|}{\ell} \right)^{\nu} K_{\nu} \left(\frac{2\sqrt{\nu}|x_1 - x_2|}{\ell} \right)$$

  - $\nu > 0$ is called the smoothness parameter
  - Γ is the Gamma function
  - K_{ν} is the modified Bessel function of the second kind

The Fourier transform $\hat{k}$ is of the form, for $\omega \in \mathbb{R}$,

$$\hat{k}(\omega) = \frac{a}{(b + \omega^2)^{\nu+1/2}} \geq 0,$$

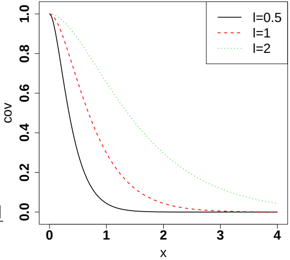
where $a \geq 0$ and $b > 0$ depend on σ^2, ℓ, ν but not on ω

The Matérn $\frac{3}{2}$ ($\nu = 3/2$) covariance function, for a Gaussian process on $\mathbb{R}$, is parameterized by

- A variance parameter $\sigma^2 > 0$
- A correlation length parameter $\ell > 0$

The Matérn formula is simplified to

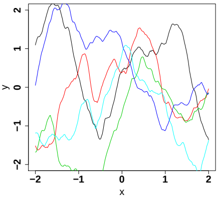
$$k(x_1, x_2) = \sigma^2 \left(1 + \sqrt{6}\frac{|x_1 - x_2|}{\ell} \right) e^{-\sqrt{6}\frac{|x_1 - x_2|}{\ell}}$$



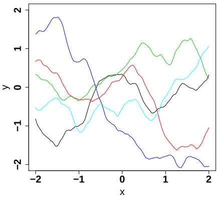
Interpretation

- stationary
- σ^2 corresponds to the order of magnitude of the functions that are realizations of the Gaussian process
- ℓ corresponds to the speed of variation of the functions that are realizations of the Gaussian process

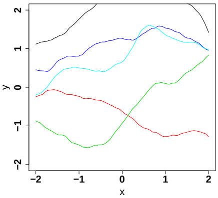
Plot of realizations of a Gaussian process having the Matérn $\frac{3}{2}$ covariance function for $\sigma^2 = 1$ and various values of ℓ



$\ell = 0.5$



$\ell = 1$



$\ell = 2$

Smoothness of the covariance function and Gaussian process

Continuous covariance function $\Longrightarrow$ continuous Gaussian process :

Proposition (see e.g. Adler, 1990)

Let ξ be a Gaussian process on $\mathbb{R}$ with mean function 0 and covariance function k
Then

- k is continuous (+ mild technical assumptions)

$\Longrightarrow$

- The trajectories of ξ are almost surely continuous on $\mathbb{R}$

Smooth covariance function $\Longrightarrow$ smooth Gaussian process :

Proposition (see e.g. Adler, 1990)

Let ξ be a Gaussian process on $\mathbb{R}$ with mean function 0 and covariance function k
Then, for $r \in \mathbb{N}$,

- k is $2r$ times differentiable (+ mild technical assumptions)

$\Longrightarrow$

- The trajectories of ξ are almost surely r times differentiable on $\mathbb{R}$

The covariance function k needs to be twice as much differentiable as ξ, because it can be shown that, with ξ' the derivative of ξ,

$$Cov\left(\xi'(u), \xi'(v)\right) = \frac{\partial k(u, v)}{\partial u \partial v}$$

Using properties of Fourier transform :

Proposition

Let k be a stationary covariance function with Fourier transform $\hat{k}$, such that the inverse Fourier transform relation holds

$$\text{for all } x \in \mathbb{R}^d,\ k(x) = \int_{\mathbb{R}^d} \hat{k}(\omega) e^{\mathrm{i}\omega^\top x} d\omega$$

Then, for $r \in \mathbb{N}$,

- The Fourier transform $\hat{k}$ verifies $\int_{\mathbb{R}} \omega^{2r} \hat{k}(\omega) < +\infty$

$\Longrightarrow$

- k is $2r$ times differentiable

Fourier transform decays quickly at infinity $\Longrightarrow$ covariance function is smooth $\Longrightarrow$ Gaussian process is smooth

Smoothness of the Matérn model

Recalling that the Fourier transform of Matérn is

$$\hat{k}(\omega) = \frac{a}{(b + \omega^2)^{\nu + 1/2}} \geq 0,$$

we obtain

Proposition

Let ξ be a Gaussian process on $\mathbb{R}$ with mean function 0 and covariance function k of the Matérn class with parameters $\sigma^2 \geq 0$, $\ell > 0$ and $\nu > 0$. Then, for $r \in \mathbb{N}$,

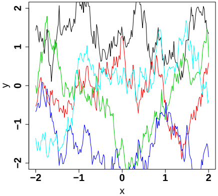
- $\nu > r$

$\Longrightarrow$

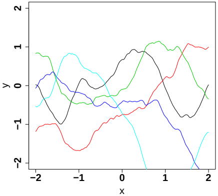
- The trajectories of ξ are almost surely r times differentiable on $\mathbb{R}$

$\Longrightarrow$ The integer part of ν is the number of derivatives

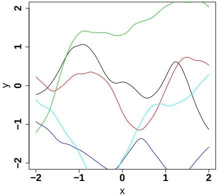
Trajectories of Gaussian processes with mean function 0 and Matérn covariance functions with $\sigma^2 = 1$, $\ell = 1$ and various values of ν



$\nu = 1/2$
continuous, not differentiable



$\nu = 3/2$
once differentiable



$\nu = 5/2$
twice differentiable

Product and mapping of kernels

Proposition (product of SNND functions)

Let k_1 and k_2 be two SNND functions on $\mathbb{X}$ (here can be any space)
Then $k_1 k_2$ is SNND on $\mathbb{X}$

See e.g. Scholkopf and Smola, 06

Proposition (kernel mapping)

Let k_2 be a SNND function on a set $\mathbb{X}_2$. Let $\phi : \mathbb{X}_1 \to \mathbb{X}_2$ be any function. Let k_1 be defined on $\mathbb{X}_1 \times \mathbb{X}_1$ by, for $u, v \in \mathbb{X}_1$,

$$k_1(u, v) = k_2(\phi(u), \phi(v))$$

Then k_1 is SNND

Proof : For $x_1, \ldots, x_n \in \mathbb{X}_1$ and $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$,

$$\sum_{i,j=1}^{n} \lambda_i \lambda_j k_1(x_i, x_j) = \sum_{i,j=1}^{n} \lambda_i \lambda_j k_2(\phi(x_i), \phi(x_j)) \geq 0$$

since k_2 is SNND and $\phi(x_1), \ldots, \phi(x_n) \in \mathbb{X}_2$ □

Tensorization to create covariance functions on $\mathbb{R}^d$

Proposition (tensorization)

Let $k_1, \dots, k_d$ be SNND functions on $\mathbb{R}$. Let k be defined on $\mathbb{R}^d \times \mathbb{R}^d$ as

$$k(u, v) = k_1(u_1, v_1) \times \dots \times k_d(u_d, v_d)$$

for $u = (u_1, \dots, u_d) \in \mathbb{R}^d$ and $v = (v_1, \dots, v_d) \in \mathbb{R}^d$.
Then k is SNND

Proof : Application of the two previous propositions with mapping functions $\phi_1, \dots, \phi_d$ with $\phi_i(x) = x_i$ for $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ $\square$

Standard tensorized covariance functions

The function k defined by, for $u = (u_1, \dots, u_d) \in \mathbb{R}^d$ and $v = (v_1, \dots, v_d) \in \mathbb{R}^d$,

$$k(u, v) = \sigma^2 \prod_{i=1}^{d} \psi(|u_i - v_i| / \ell_i)$$

is

- the tensorized exponential covariance function when
$$\psi(t) = e^{-t}$$
- the tensorized square exponential covariance function when
$$\psi(t) = e^{-t^2}$$
- the tensorized Matérn covariance function when
$$\psi(t) = \frac{1}{\Gamma(\nu) 2^{\nu-1}} \left(2\sqrt{\nu} t\right)^{\nu} K_{\nu}\left(2\sqrt{\nu} t\right)$$

Interpretation of the parameters :

- σ^2 is the variance and is interpreted as before
- For $i = 1, \dots, d$, ℓ_i is the correlation length for the variable i
- ℓ_i small means that variable i is important
$\Longrightarrow$ Allows variable ranking and screening

M. Ben Salem, F. Bachoc, O. Roustant, F. Gamboa and L. Tomaso, Gaussian Process based dimension reduction for goal-oriented sequential design, *SIAM/ASA Journal on Uncertainty Quantification*, 7(4) (2019) 1369-1397

Isotropic covariance functions

We want to create covariance functions on $\mathbb{R}^d$ of the form, for $x_1, x_2 \in \mathbb{R}^d$,

$$k(x_1, x_2) = \psi(||x_1 - x_2||), \tag{1}$$

with $\psi : \mathbb{R}^+ \to \mathbb{R}$

We have a characterization of the functions ψ for which we obtain an SNND function for all $d \in \mathbb{N}$

Theorem (Shoenberg, 38)

Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ defined by (1) where ψ is not constant. Then the following statements are equivalent

1. k is SNND for all $d \in \mathbb{N}$
2. ψ is of the form

$$\psi(t) = \int_0^{+\infty} e^{-\omega t^2} d\mu(\omega),$$

   with a non-negative measure μ on $\mathbb{R}^+$, not concentrated at 0
3. $\psi(\sqrt{.})$ is completely monotone on $[0, \infty)$ and not constant. A function g on $[0, \infty)$ is completely monotone if

$$(-1)^r g^{(r)}(t) \geq 0 \quad \text{for } r \in \mathbb{N} \text{ and } t \in [0, \infty)$$

Standard isotropic covariance functions

The function k defined by, for $u \in \mathbb{R}^d$ and $v \in \mathbb{R}^d$,

$$k(u, v) = \sigma^2 \psi(||u - v||/\ell)$$

is

- the isotropic exponential covariance function when

$$\psi(t) = e^{-t}$$

- the isotropic square exponential covariance function when

$$\psi(t) = e^{-t^2}$$

- the isotropic Matérn covariance function when

$$\psi(t) = \frac{1}{\Gamma(\nu) 2^{\nu - 1}} \left(2\sqrt{\nu} t\right)^{\nu} K_{\nu} \left(2\sqrt{\nu} t\right)$$

Interpretation of the parameters :

- σ^2 is the variance and is interpreted as before
- ℓ is the correlation length, controls how fast covariance changes with distance (in any direction)

Geometric anisotropy

The function k defined by, for $u = (u_1, \ldots, u_d) \in \mathbb{R}^d$ and $v = (v_1, \ldots, v_d) \in \mathbb{R}^d$,

$$k(u, v) = \sigma^2 \psi \left(\sqrt{\sum_{i=1}^{d} \frac{(u_i - v_i)^2}{\ell_i^2}} \right)$$

is

- the geometric anisotropic exponential covariance function when

$$\psi(t) = e^{-t}$$

- the geometric anisotropic square exponential covariance function when

$$\psi(t) = e^{-t^2}$$

- the geometric anisotropic Matérn covariance function when

$$\psi(t) = \frac{1}{\Gamma(\nu) 2^{\nu-1}} \left(2\sqrt{\nu} t\right)^{\nu} K_{\nu} \left(2\sqrt{\nu} t\right)$$

$\Longrightarrow$ These functions are SNND from the previous results

Interpretation of the parameters :

- σ^2 is the variance and is interpreted as before
- For $i = 1, \ldots, d$, ℓ_i is the correlation length for the variable i
- ℓ_i small means that variable i is important
  $\Longrightarrow$ Allows variable ranking and screening

Conclusions

- Covariance function drives the order of magnitude and speed of variation of the Gaussian process
- On $\mathbb{R}^d$, smooth covariance function $\Longrightarrow$ smooth Gaussian process
- Catalog of available SNND functions on $\mathbb{R}^d$

Topics we did not address

- Covariance functions for functional or distributional inputs
- Covariance functions on character strings
- Covariance functions on a manifold (e.g. the sphere in climate sciences)
- Covariance functions on neural network architectures
- ...

Next : Conditional distribution given observations (with a fixed given covariance function)

Gaussian conditioning theorem

Theorem

Let $(\boldsymbol{Y}_1, \boldsymbol{Y}_2)^\top$ be a $(n_1 + n_2) \times 1$ Gaussian vector with mean vector $(\boldsymbol{m}_1^\top, \boldsymbol{m}_2^\top)^\top$ and covariance matrix

$$\begin{pmatrix} \boldsymbol{R}_1 & \boldsymbol{R}_{1,2} \\ \boldsymbol{R}_{1,2}^\top & \boldsymbol{R}_2 \end{pmatrix}$$

Then, conditionaly on $\boldsymbol{Y}_1 = \boldsymbol{y}_1$, $\boldsymbol{Y}_2$ is a Gaussian vector with mean

$$\mathbb{E}(\boldsymbol{Y}_2 | \boldsymbol{Y}_1 = \boldsymbol{y}_1) = \boldsymbol{m}_2 + \boldsymbol{R}_{1,2}^\top \boldsymbol{R}_1^{-1} (\boldsymbol{y}_1 - \boldsymbol{m}_1)$$

and variance

$$var(\boldsymbol{Y}_2 | \boldsymbol{Y}_1 = \boldsymbol{y}_1) = \boldsymbol{R}_2 - \boldsymbol{R}_{1,2}^\top \boldsymbol{R}_1^{-1} \boldsymbol{R}_{1,2}$$

Illustration

Let $(Y_1, Y_2)^\top$ be a 2×1 Gaussian vector with mean vector $(\mu_1, \mu_2)^\top$ and covariance matrix

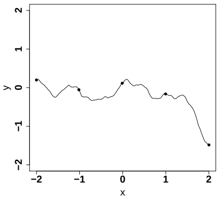
$$\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

Then

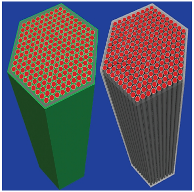
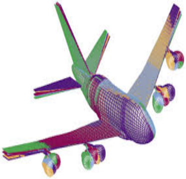
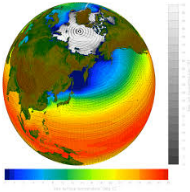
$$\mathbb{E}(Y_2 | Y_1 = y_1) = \mu_2 + \rho(y_1 - \mu_1) \quad \text{and} \quad var(Y_2 | Y_1 = y_1) = 1 - \rho^2$$

The case of exact observations

We can obtain exact observations of the function f



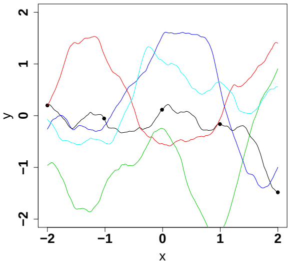
Typical example : $f(x)$ is the result of a deterministic computer experiment with simulation parameters x

Reminder of the Bayesian model

It is a function interpolation/approximation problem
Possible methods : polynomial regression, neural networks, splines, RKHS, ...
$\longrightarrow$ can provide a deterministic error bound

Gaussian process model : representing the deterministic and unknown function f by a realization of a Gaussian process.
$\longrightarrow$ gives a stochastic error bound



Bayesian statistics

In statistics, a Bayesian model generally consists in representing a deterministic and unknown number/vector by the realization of a random variable/vector (the prior)

Gaussian process prediction

- We let ξ be the Gaussian process on $\mathbb{X}$, with mean function m and covariance function k
- ξ is observed at $x_1, ..., x_n \in \mathbb{X}$

Notations

- Let $\mathbf{Y}_n = (\xi(x_1), ..., \xi(x_n))^\top$ be the observation vector. It is a Gaussian vector
- Let $\mathbf{y}_n = (f(x_1), ..., f(x_n))^\top$ be the observed values
- Let $\mathbf{m}_n$ be the mean vector of $\mathbf{Y}_n$: $\mathbf{m}_n = (m(x_1), \ldots, m(x_n))^\top$
- Let $\mathbf{R}$ be the $n \times n$ covariance matrix of $\mathbf{Y}_n$: $R_{i,j} = k(x_i, x_j)$
- Let $x \in \mathbb{X}$ be a new input point for the Gaussian process ξ. We want to predict $\xi(x)$
- Let $\mathbf{r}(x)$ be the $n \times 1$ covariance vector between $\mathbf{Y}_n$ and $\xi(x)$: $r(x)_i = k(x_i, x)$

Then the Gaussian conditioning theorem gives the conditional mean function of ξ given the observed values in Y_n :

$$m_n(x) := \mathbb{E}(\xi(x)|\mathbf{Y}_n = \mathbf{y}_n) = m(x) + \mathbf{r}(x)^\top \mathbf{R}^{-1}(\mathbf{y}_n - \mathbf{m}_n)$$

We also have the conditional covariance function, for $u, v \in \mathbb{X}$:

$$k_n(u, v) := Cov(\xi(u), \xi(v)|\mathbf{Y}_n = \mathbf{y}_n) = k(u, v) - \mathbf{r}(u)^\top \mathbf{R}^{-1}\mathbf{r}(v)$$

$\Longrightarrow$ Conditionally to $\mathbf{Y}_n = \mathbf{y}_n$, ξ is a Gaussian process with mean function m_n and covariance function k_n

Gaussian process prediction : interpretation

Exact interpolation of known values

Assume $x = x_1$. Then, $R_{1,i} = k(x_1, x_i) = k(x, x_i) = r(x)_i$. Thus

$$m(x) + \mathbf{r}(x)^\top \mathbf{R}^{-1}(\mathbf{y}_n - \mathbf{m}_n) = m(x) + \mathbf{r}(x)^\top \times \begin{pmatrix} \mathbf{r}(x)^\top \\ * \\ \vdots \\ * \end{pmatrix}^{-1} \times \begin{pmatrix} f(x_1) - m(x_1) \\ \vdots \\ f(x_n) - m(x_n) \end{pmatrix}$$

$$= m(x) + (1, 0, \ldots, 0) \begin{pmatrix} f(x_1) - m(x) \\ \vdots \\ f(x_n) - m(x_n) \end{pmatrix} = f(x_1)$$

Conservative extrapolation

Let x be far from $x_1, ..., x_n$. Then, we generally have $r(x)_i = k(x_i, x) \approx 0$. Thus

$$m_n(x) = m(x) + \mathbf{r}(x)^\top \mathbf{R}^{-1}(\mathbf{y}_n - \mathbf{m}_n) \approx m(x)$$

and

$$k_n(x, x) = k(x, x) - \mathbf{r}(x)^\top \mathbf{R}^{-1} \mathbf{r}(x) \approx k(x, x)$$

$\Rightarrow$ conservative

Illustration of Gaussian process prediction

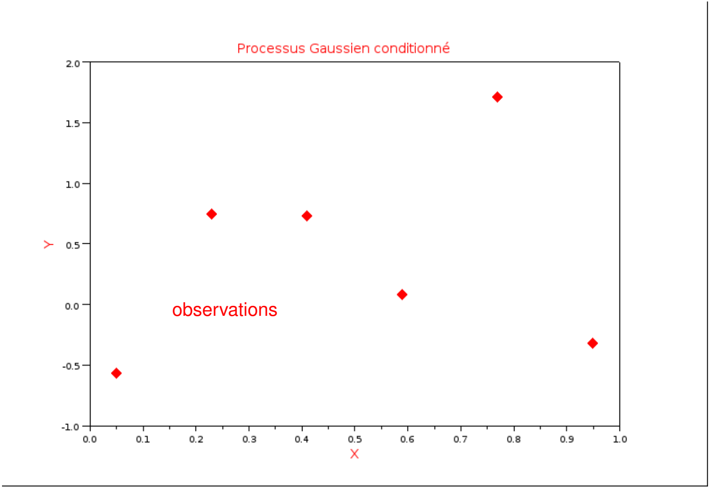


Illustration of Gaussian process prediction

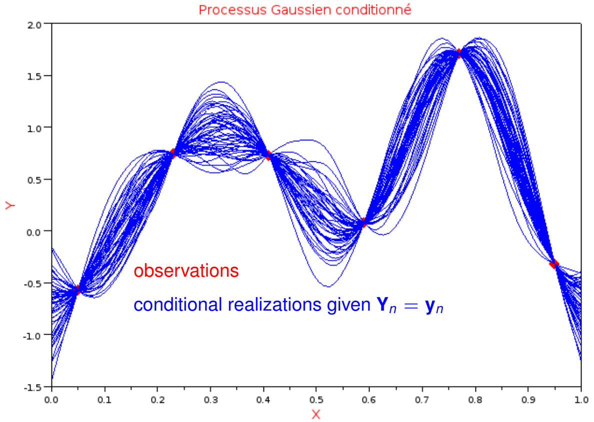


Illustration of Gaussian process prediction

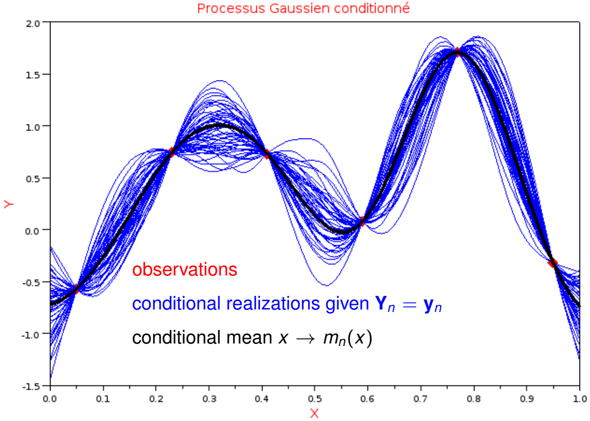
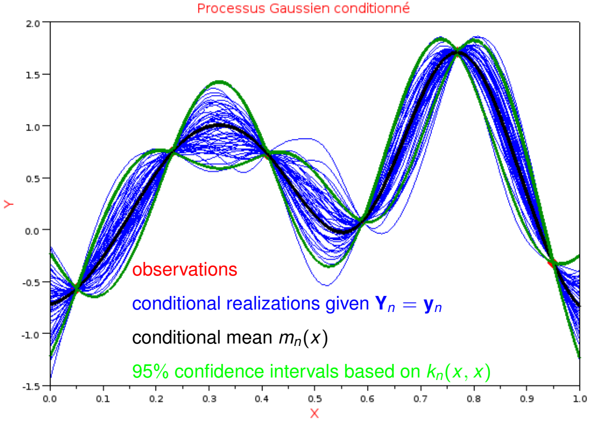


Illustration of Gaussian process prediction

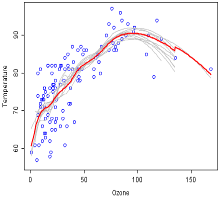


Gaussian process prediction with noisy observations

It can be desirable not to reproduce the observed values exactly :

- when same x can give different observed values $\Longrightarrow$ common in machine learning applications

$\Longrightarrow$ E.g. flight delay from flight features



We consider that at $x_1, ..., x_n$, we observe

$$\mathbf{Y}_n = \begin{pmatrix} \xi(x_1) + \mathcal{E}_1 \\ \vdots \\ \xi(x_n) + \mathcal{E}_n \end{pmatrix}$$

$\mathcal{E}_1, ..., \mathcal{E}_n$ are independent and are Gaussian variables, with mean 0 and variance τ^2

- We let $\mathbf{y}_n$ be the realization of $\mathbf{Y}_n$

$$\mathbf{y}_n = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} f(x_1) + \epsilon_1 \\ \vdots \\ f(x_n) + \epsilon_n \end{pmatrix}$$

Gaussian process prediction with noisy observations

Then the Gaussian conditioning theorem still gives the conditional mean of $\xi(x)$ given the observed values in $\mathbf{y}_n$:

$$m_n(x) := \mathbb{E}(\xi(x)|\mathbf{Y}_n = \mathbf{y}_n) = m(x) + \mathbf{r}(x)^\top (\mathbf{R} + \tau^2 \mathbf{I}_n)^{-1} (\mathbf{y}_n - \mathbf{m}_n)$$

We also have the conditional covariance, for $u, v \in \mathbb{X}$:

$$k_n(u,v) := Cov(\xi(u), \xi(v)|\mathbf{Y}_n = \mathbf{y}_n) = k(u,v) - \mathbf{r}(u)^\top (\mathbf{R} + \tau^2 \mathbf{I}_n)^{-1} \mathbf{r}(v)$$

$\Longrightarrow$ Conditionally to $\mathbf{Y}_n = \mathbf{y}_n$, ξ is a Gaussian process with mean function m_n and covariance function k_n

Illustration of Gaussian process prediction with measure error

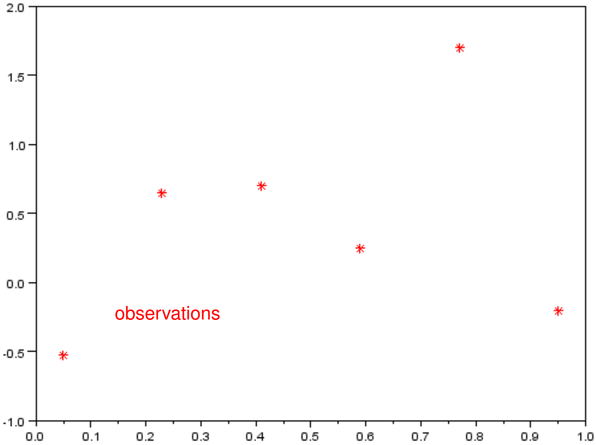


Illustration of Gaussian process prediction with measure error

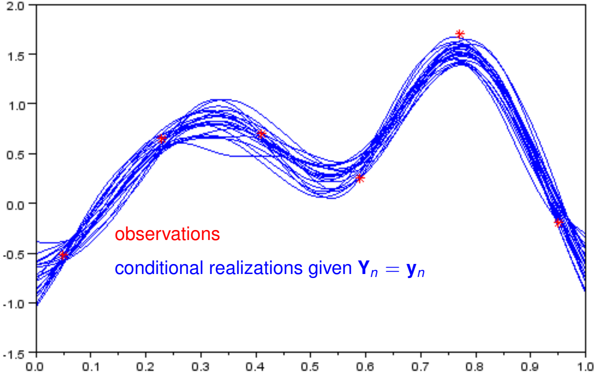
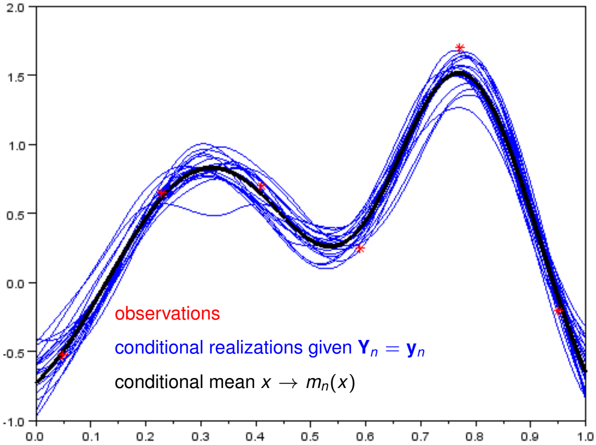
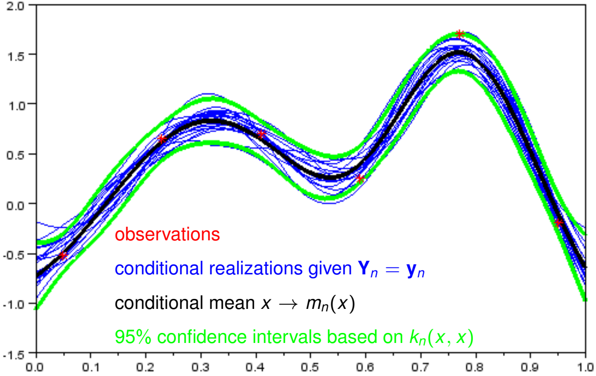


Illustration of Gaussian process prediction with measure error




2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
observations
conditional realizations given $\mathbf{Y}_n = \mathbf{y}_n$
conditional mean $x \to m_n(x)$
95% confidence intervals based on $k_n(x, x)$

Remarks

- The conditioning takes the same form, independently of the input space $\mathbb{X}$
- The computation cost for an exact implementation is
  - $O(n^2)$ in storage and $O(n^3)$ in computation, once, offline
  - $O(n^2)$ in computation for each new x, online
- Exist various works when n very large

  Aggregation of submodels :

  B. van Stein, H. Wang, W. Kowalczyk, T, Bäck, and M. Emmerich, Optimally weighted cluster kriging for big data regression, *In International Symposium on Intelligent Data Analysis*, pages 310-321, Springer, 2015

  D. Rullière, N. Durrande, F. Bachoc and C. Chevalier, Nested Kriging predictions for datasets with a large number of observations, *Statistics and Computing*, 28(4), 849-867, 2018

  Inducing points :

  J. Hensman, N. Fusi, N.D. Lawrence, Gaussian Processes for Big Data, *Uncertainty in Artificial Intelligence conference*, paper Id 244, 2013
- Works well with integrals and derivatives (remains Gaussian)

Gaussian process classification model

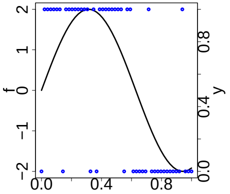
- Gaussian process ξ with realization f
- Observation points $x_1, \ldots, x_n$
- Observation vector

$$\mathbf{Y}_n = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \in \{0,1\}^n$$

with for $i = 1, \ldots, n$

$$\mathbb{P}(Y_i = 1 | \xi = f) = \frac{e^{\alpha f(x_i)}}{1 + e^{\alpha f(x_i)}}$$

- α large $\Longrightarrow \mathbb{P}(Y_i = 1)$ close to 0 or 1 $\Longrightarrow Y_i$ almost deterministic given $\xi = f$



Step 1 : conditional distribution of Gaussian vector given observations

- Let

$$\mathbf{V}_n = \begin{pmatrix} \xi(x_1) \\ \vdots \\ \xi(x_n) \end{pmatrix}$$

- Let $\mathbf{y}_n$ be the observed realization of $\mathbf{Y}_n$
- Then, conditionally to $\mathbf{Y}_n = \mathbf{y}_n$, $\mathbf{V}_n$ has density ϕ_n given by, for $\mathbf{v} = (v_1, \ldots, v_n)^\top \in \mathbb{R}^n$,

$$\phi_n(\mathbf{v}) = (\text{constant not depending on } \mathbf{v}) \times \mathcal{N}(\mathbf{v}|\mathbf{m}_n, \mathbf{R}) \times \prod_{i=1}^{n} \left(\mathbf{1}_{\{y_i=1\}} \frac{e^{\alpha v_i}}{1+e^{\alpha v_i}} + \mathbf{1}_{\{y_i=0\}} \frac{1}{1+e^{\alpha v_i}} \right)$$

  with
  - $\mathcal{N}(\mathbf{v}|\mathbf{m}_n, \mathbf{R})$ the Gaussian density at $\mathbf{v}$ with mean vector $\mathbf{m}_n$ and covariance matrix $\mathbf{R}$ $\Longrightarrow$ density of $\mathbf{V}_n$
- The conditional density ϕ_n is non-Gaussian
- Sampling from ϕ_n or approximating ϕ_n is the difficult part
- MCMC procedures, Laplace approximation, EM algorithm, ...

H. Nickisch and C. E. Rasmussen, Approximations for binary Gaussian process classification, *Journal of Machine Learning Research*, 9 : 2035-2078, 2008

Conditional distribution

Step 2 : Classification after $\mathbf{V}_n$ is sampled from ϕ_n
Assumes that $\mathbf{v}_n$ is a conditional realization of $\mathbf{V}_n$ given $\mathbf{Y}_n = \mathbf{y}_n$ (density ϕ_n)

- Conditionally to $\mathbf{Y}_n = \mathbf{y}_n$ and $\mathbf{V}_n = \mathbf{v}_n$, ξ is a Gaussian process with mean function m_n (depends on $\mathbf{v}_n$) and covariance function k_n
- Conditionally to $\mathbf{Y}_n = \mathbf{y}_n$ and $\mathbf{V}_n = \mathbf{v}_n$, $\xi(x)$ is Gaussian with mean $m_n(x)$ (depends on $\mathbf{v}_n$) and variance $k_n(x, x)$
- Consider a new observation $Y_x \in \{-1, 1\}$ such that

$$\mathbb{P}(Y_x = 1 | \xi = f) = \frac{e^{\alpha f(x)}}{1 + e^{\alpha f(x)}}$$

- Then, conditionally to $\mathbf{Y}_n = \mathbf{y}_n$ and $\mathbf{V}_n = \mathbf{v}_n$,

$$\mathbb{P}(Y_x = 1 | \mathbf{Y}_n = \mathbf{y}_n, \mathbf{V}_n = \mathbf{v}_n) = \int_{-\infty}^{+\infty} \mathcal{N}(v | m_n(x), k_n(x, x)) \frac{e^{\alpha v}}{1 + e^{\alpha v}} dv$$

- One-dimensional integral can be computed explicitly
- Things are again Gaussian and simpler

An example of purely Monte Carlo classification

- **Step 1 :** obtain N realizations

$$\mathbf{v}_n^{(1)}, \ldots, \mathbf{v}_n^{(N)}$$

  approximately following the conditional distribution of $\mathbf{V}_n$ given $\mathbf{Y}_n = \mathbf{y}_n$
  $\Longrightarrow$ Potentially costly MCMC here
- Each realization $\mathbf{v}_n^{(i)}$ provides a conditional mean function $m_n^{(i)}$
- **Step 2 :** average classifications

$$\mathbb{P}(Y_x = 1 | \mathbf{Y}_n = \mathbf{y}_n) \approx \frac{1}{N} \sum_{i=1}^{N} \int_{-\infty}^{+\infty} \mathcal{N}(v | m_n^{(i)}(x), k_n(x,x)) \frac{e^{\alpha v}}{1 + e^{\alpha v}} dv$$

Remarks :

- There can be convergence guarantees as $N \to \infty$ and for large MCMC budget
- Potentially computationally costly
- Approximations in Nickisch and Rasmussen, 2008 are typically faster (but less guarantees)

Illustration

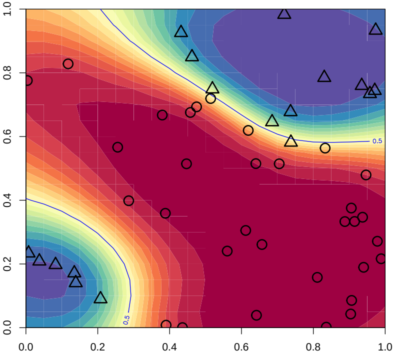


Figure – posterior probabilities of 1

Covariance function estimation

Parameterization

Covariance function model $\{\sigma^2 c_\theta, \sigma^2 \geq 0, \theta \in \Theta\}$ for the Gaussian Process ξ

- σ^2 is the variance parameter
- θ is the multidimensional correlation parameter. c_θ is a stationary correlation function
- We want to choose the covariance function k of the form $\sigma^2 c_\theta$
- Assume mean function is 0 for simplicity

Estimation

ξ is observed at $x_1, ..., x_n \in \mathbb{X}$, yielding the Gaussian vector $\mathbf{Y}_n = (\xi(x_1), ..., \xi(x_n))^\top$.
Estimators $\hat{\sigma}^2(\mathbf{Y}_n)$ and $\hat{\theta}(\mathbf{Y}_n)$

"Plug-in" Gaussian process prediction

1. Estimate the covariance function
2. Assume that the covariance function is fixed and carry out the conditioning studied before

Maximum Likelihood for estimation

Explicit Gaussian likelihood function for the observation vector $\mathbf{Y}_n$

Maximum Likelihood

Define $\mathbf{C}_\theta$ as the correlation matrix of $\mathbf{Y}_n = (\xi(x_1), ..., \xi(x_n))^\top$ under correlation function c_θ.
The Maximum Likelihood estimator of (σ^2, θ) is

$$(\hat{\sigma}^2_{ML}, \hat{\theta}_{ML}) \in \underset{\sigma^2 \geq 0, \theta \in \Theta}{\operatorname{argmin}} \frac{1}{n} \left(\ln \left(|\sigma^2 \mathbf{C}_\theta| \right) + \frac{1}{\sigma^2} \mathbf{Y}_n^\top \mathbf{C}_\theta^{-1} \mathbf{Y}_n \right)$$

Remarks :

- Needs to be optimized numerically
- Cost $O(n^3)$ in time per evaluation of likelihood
- Existing work to approximate when n is large, e.g. Gramacy and Apley 2015

Cross Validation for estimation

- $m_{n,\theta}^{(-i)} = \mathbb{E}_{\sigma^2,\theta}(\xi(x_i)|\xi(x_1),...,\xi(x_{i-1}),\xi(x_{i+1}),...,\xi(x_n))$
- $\sigma^2(c_{n,\theta}^{(-i)})^2 = var_{\sigma^2,\theta}(\xi(x_i)|\xi(x_1),...,\xi(x_{i-1}),\xi(x_{i+1}),...,\xi(x_n))$

Leave one out estimation

$$\hat{\theta}_{CV} \in \underset{\theta \in \Theta}{\operatorname{argmin}} \sum_{i=1}^{n} (\xi(x_i) - m_{n,\theta}^{(-i)})^2$$

and

$$\frac{1}{n} \sum_{i=1}^{n} \frac{(\xi(x_i) - m_{n,\hat{\theta}_{CV}}^{(-i)})^2}{\hat{\sigma}_{CV}^2 (c_{n,\hat{\theta}_{CV}}^{(-i)})^2} = 1 \Leftrightarrow \hat{\sigma}_{CV}^2 = \frac{1}{n} \sum_{i=1}^{n} \frac{(\xi(x_i) - m_{n,\hat{\theta}_{CV}}^{(-i)})^2}{(c_{n,\hat{\theta}_{CV}}^{(-i)})^2}$$

Let $\mathbf{C}_\theta$ be the correlation matrix of $\mathbf{Y}_n = (\xi(x_1), ..., \xi(x_n))^\top$ with correlation function c_θ

Virtual Leave-One-Out

$$\xi(x_i) - m_{n,\theta}^{(-i)} = \frac{\left(\mathbf{C}_\theta^{-1}\mathbf{Y}_n\right)_i}{\left(\mathbf{C}_\theta^{-1}\right)_{i,i}} \quad \text{and} \quad (c_{n,\theta}^{(-i)})^2 = \frac{1}{(\mathbf{C}_\theta^{-1})_{i,i}}$$

O. Dubrule, Cross Validation of Kriging in a Unique Neighborhood, *Mathematical Geology*, 1983.

Using the virtual Cross Validation formula :

$$\hat{\theta}_{CV} \in \underset{\theta \in \Theta}{\operatorname{argmin}} \frac{1}{n} \mathbf{Y}_n^\top \mathbf{C}_\theta^{-1} diag(\mathbf{C}_\theta^{-1})^{-2} \mathbf{C}_\theta^{-1} \mathbf{Y}_n$$

and

$$\hat{\sigma}_{CV}^2 = \frac{1}{n} \mathbf{Y}_n^\top \mathbf{C}_{\hat{\theta}_{CV}}^{-1} diag(\mathbf{C}_{\hat{\theta}_{CV}}^{-1})^{-1} \mathbf{C}_{\hat{\theta}_{CV}}^{-1} \mathbf{Y}_n$$

Some references on covariance function estimation

- Practical aspects of cross validation

F. Bachoc, Cross Validation and Maximum Likelihood estimation of hyper-parameters of Gaussian processes with model misspecification, *Computational Statistics and Data Analysis*, 66 55-69, 2013

H. Zhang and Y. Wang, Kriging and cross-validation for massive spatial data, *Environmetrics*, 21(3/4) :290-304, 2010

- Theory on maximum likelihood and cross validation

F. Bachoc, Asymptotic analysis of covariance parameter estimation for Gaussian processes in the misspecified case, *Bernoulli*, 24(2), 1531-1575, 2018

C.G. Kaufman and B. A. Shaby, The role of the range parameter for estimation and prediction in geostatistics, *Biometrika*, 100(2), 473-484, 2013

General conclusions

- Gaussian processes can be defined on any space $\mathbb{X}$, by using suitable covariance functions
- Setting of direct observations is favorable for conditioning $\Longrightarrow$ benefit of Gaussian processes
- Indirect observations (e.g. Gaussian process classification) are computationally more challenging.
- But the Gaussian process still brings simplifications
- Gaussian variables, vectors and processes come with many existing theoretical results $\Longrightarrow$ Gaussian processes are also a convenient theoretical framework
- Gaussian processes can be used as elementary bricks to construct more complex stochastic processes

Thank you for your attention !

Bibliography

- M. Stein, Interpolation of spatial data, some theory for Kriging, *Springer*, 1999
- T. J. Santner, B. J. Williams and W. I. Notz, The design and analysis of computer experiments, *Springer Science & Business Media*, 2003
- C.E. Rasmussen and C.K.I. Williams, Gaussian Processes for Machine Learning, *The MIT Press, Cambridge*, 2006
- R. J. Adler, An introduction to continuity, extrema, and related topics for general Gaussian processes, *IMS*, 1990
- B. Scholkopf amd A. J. Smola, Learning with kernels : support vector machines, regularization, optimization, and beyond, *MIT press*, 2006
- I. J Schoenberg, Metric spaces and completely monotone functions, *Annals of Mathematics*, 811-841, 1938
- H. Nickisch and C. E. Rasmussen, Approximations for binary Gaussian process classification, *Journal of Machine Learning Research*, 9 : 2035-2078, 2008
- R. B. Gramacy and D. W. Apley, Local Gaussian process approximation for large computer experiments, *Journal of Computational and Graphical Statistics*, 24(2), 561-578, 2015