

Advances in Gaussian Process

François Bachoc

Institut de Mathématiques de Toulouse
Université Paul Sabatier

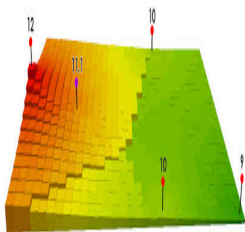
Machine Learning Summer School, Skoltech, Moscow
September 2019

- 1 Introduction to Gaussian processes
- 2 Sequential learning and consistency of stepwise uncertainty reduction strategies
- 3 Gaussian processes under inequality constraints

Gaussian processes in different fields

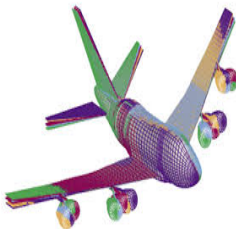
Gaussian processes are studied in different fields :

Geostatistics



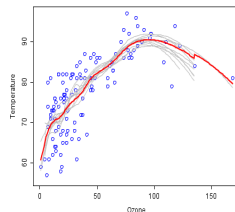
Stein, 99

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

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

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

The function f is a **black box**

- ⇒ Only available through observations
- ⇒ 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

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

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

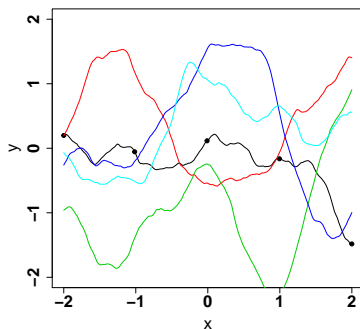
Gaussian processes also tackle these types of observations and are based on a [Bayesian prior on the function \$f\$](#)

⇒ Hence they provide an important benefit for [uncertainty quantification](#)

Gaussian processes as Bayesian prior

Bayesian prior

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



Usefulness

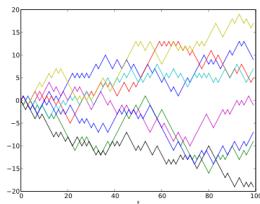
Using the conditional distribution of ξ , given the **observations**, to learn f

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

A **stochastic process** on \mathbb{X} is a function $\xi : \mathbb{X} \rightarrow \mathbb{R}$ 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 $x \rightarrow \xi(\omega_0, x)$ a **realization** of the stochastic process ξ .

\Rightarrow The probability space Ω is the same for all $\xi(\omega, x)$ with $x \in \mathbb{X}$

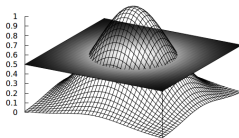
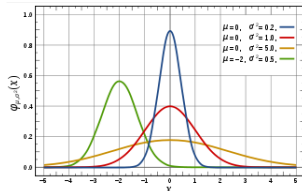
Gaussian variables and vectors

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 \mathbf{V} is a **Gaussian vector** with mean vector \mathbf{m} and invertible covariance matrix \mathbf{R} when its multidimensional probability density function is

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



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

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

⇒ All linear combinations of Gaussian vectors are Gaussian variables

- When \mathbf{R} is not invertible, \mathbf{V} is supported on a lower dimensional linear subspace of \mathbb{R}^n (spanned by the eigenvectors of the non-zero eigenvalues of \mathbf{R})

Definition

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

Mean and covariance functions

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

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

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

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

\Rightarrow A Gaussian process is **characterized** by its mean and covariance functions

Constraints on the covariance function

First, remark that k is symmetric :

$$k(x_1, x_2) = \text{Cov}(\xi(x_1), \xi(x_2)) = \text{Cov}(\xi(x_2), \xi(x_1)) = k(x_2, x_1)$$

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

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

\Rightarrow Hence a second constraint on k

Constraints on the covariance function

Symmetric non-negative definite functions

A function $h : \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}$ 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 $x_1, \dots, x_n \in \mathbb{X}$ and $\lambda_1, \dots, \lambda_n \in \mathbb{R}$:

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

⇒ Covariance functions are SNND

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

Hence, covariance functions can also be called

- kernels
- radial basis functions
- positive definite functions

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 $k_1 : \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}$ 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 $k_2 : \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}$ 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)

\Rightarrow The behavior of the corresponding Gaussian process is **invariant by translation**

Bochner's theorem

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

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

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

\Rightarrow A convenient characterization of stationary covariance functions

Proof of one implication of Bochner's theorem

Assume that \hat{k} takes positive values

For all $x_1, \dots, x_n \in \mathbb{X}$, $\lambda_1, \dots, \lambda_n \in \mathbb{R}$:

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

Hence k is SNND

□

Hence some stationary covariance functions on \mathbb{R}

■ **Exponential covariance function**

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

⇒ 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

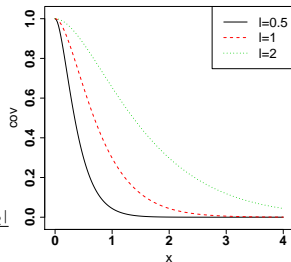
Example of the Matérn $\frac{3}{2}$ covariance function on \mathbb{R}

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

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

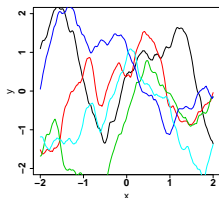


Interpretation

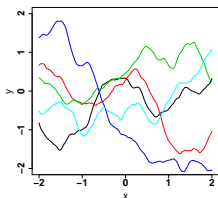
- stationary
- σ^2 corresponds to the order of magnitude of the functions that are realizations of the Gaussian process
- ℓ corresponds to the speed of variation of the functions that are realizations of the Gaussian process

The Matérn $\frac{3}{2}$ covariance function on \mathbb{R} : illustration of ℓ

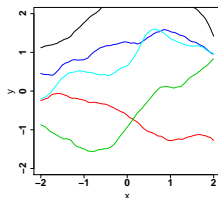
Plot of realizations of a Gaussian process having the Matérn $\frac{3}{2}$ covariance function for $\sigma^2 = 1$ and various values of ℓ



$\ell = 0.5$



$\ell = 1$



$\ell = 2$

Smoothness of the covariance function and Gaussian process

Continuous covariance function \implies 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)



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

Smooth covariance function \implies 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)



- 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 ,

$$\text{Cov}(\xi'(u), \xi'(v)) = \frac{\partial^2 k(u, v)}{\partial u \partial v}$$

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

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

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

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



- k is $2r$ times differentiable

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

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 $r \in \mathbb{N}$,

- $\nu > r$



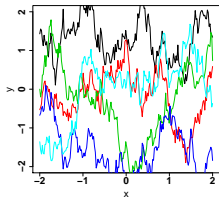
- The trajectories of ξ are **almost surely r times differentiable** on \mathbb{R}



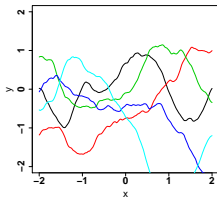
The integer part of ν is the number of derivatives

Illustration of the impact of ν

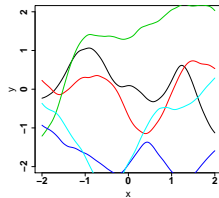
Trajectories of Gaussian processes with mean function 0 and Matérn covariance functions with $\sigma^2 = 1$, $\ell = 1$ and various values of ν



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



$\nu = 3/2$
once differentiable



$\nu = 5/2$
twice differentiable

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 $\phi : \mathbb{X}_1 \rightarrow \mathbb{X}_2$ 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 $x_1, \dots, x_n \in \mathbb{X}_1$ and $\lambda_1, \dots, \lambda_n \in \mathbb{R}$,

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

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



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 ϕ_1, \dots, ϕ_d with $\phi_i(x) = x_i$ for $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ □

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

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

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**

⇒ 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*, forthcoming

Isotropic covariance functions

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

$$k(x_1, x_2) = \psi(\|x_1 - x_2\|), \quad (1)$$

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

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 $k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ 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 $\psi(\sqrt{\cdot})$ 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

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

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 $u = (u_1, \dots, u_d) \in \mathbb{R}^d$ and $v = (v_1, \dots, v_d) \in \mathbb{R}^d$,

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

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

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

⇒ These functions are SNND from the previous results

Interpretation of the parameters :

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

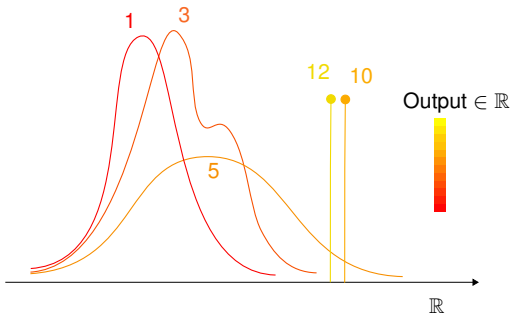
Functional inputs

Consider now that

$$\mathbb{X} = \{\text{square summable functions } f : [0, 1]^p \rightarrow \mathbb{R}\}$$

Motivations

- Inputs of computer models can be **curves** ($p = 1$) or **maps** ($p = 2$). E.g. input power profile of an industry system. Modeling of the system output as a Gaussian process
- Classification of **curves** ($p = 1$) or **images** ($p = 2$). E.g. individual healthy or unhealthy according to EEG



Obtaining covariance functions on functional inputs by limit

Consider functions $f_1, \dots, f_n \in \mathbb{X}$ and $\lambda_1, \dots, \lambda_n \in \mathbb{R}$. Let $(X_a)_{a \in \mathbb{N}}$ be i.i.d. and uniformly distributed on $[0, 1]^p$

We have

$$\sum_{i,j=1}^n \lambda_i \lambda_j e^{-\sqrt{\int_{[0,1]^p} (f_i(x) - f_j(x))^2 dx}} =_{a.s.} \lim_{d \rightarrow \infty} \sum_{i,j=1}^n \lambda_i \lambda_j e^{-\sqrt{\frac{1}{d} \sum_{a=1}^d (f_i(X_a) - f_j(X_a))^2}} \\ \geq 0$$

by considering the **isotropic exponential** covariance function on \mathbb{R}^d with the n input points

- $(f_1(X_1), \dots, f_1(X_d))$
- \dots
- $(f_n(X_1), \dots, f_n(X_d))$

Hence standard covariance functions for functional inputs

The function k defined by, for $f \in \mathbb{X}$ and $g \in \mathbb{X}$,

$$k(f, g) = \sigma^2 \psi \left(\frac{1}{\ell} \sqrt{\int_{[0,1]^p} (f(x) - g(x))^2 dx} \right)$$

is

- the **functional isotropic exponential** covariance function when

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

- the **functional isotropic square exponential** covariance function when

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

- the **functional isotropic Matérn** covariance function when

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

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)

By the same principle as above, the covariance function defined by, for $f \in \mathbb{X}$ and $g \in \mathbb{X}$,

$$k(f, g) = \sigma^2 e^{-\frac{1}{\ell} \int_{[0,1]^D} |f(x) - g(x)| dx}$$

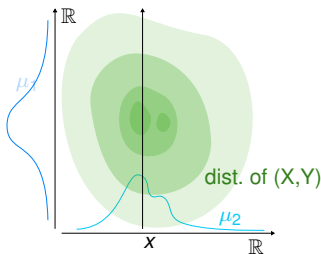
is **SNND**.

Examples of work related to functional inputs

- In [Muehlenstaedt et al., 2017](#), Gaussian processes with functional inputs are considered, for computer experiments
The functional inputs are decomposed on spline basis functions
- In [Morris 2012](#), the above constructions of covariance functions for functional inputs are used
- Related setting : inputs are distributions on \mathbb{R}^p .
For two distributions μ_1, μ_2 on \mathbb{R}^p , the [Wasserstein](#) (or Monge-Kantorovich) distance W_2 between μ_1 and μ_2 is defined by

$$W_2^2(\mu_1, \mu_2) = \inf_{\substack{(X,Y) \text{ random vector on } \mathbb{R}^{2p} \\ X \sim \mu_1 \\ Y \sim \mu_2}} \mathbb{E} \left(\|X - Y\|^2 \right)$$

(optimal transport)



Examples of work based on the Wasserstein distance

■ In



S. Kolouri, Y. Zou, and G. K. Rohde, Sliced wasserstein kernels for probability distributions, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5258-5267.

and



F. Bachoc, F. Gamboa, J.M. Loubes and N. Venet, Gaussian process regression model for distribution inputs, *IEEE Transactions on Information Theory*, 2017.

Gaussian processes are studied on the set of **one-dimensional distributions**

- In dimension 1, with q_1 and q_2 the quantile functions of μ_1 and μ_2 we have the explicit **explicit expression** for the Wasserstein distance

$$W_2^2(\mu_1, \mu_2) = \int_0^1 (q_1(t) - q_2(t))^2 dt$$

⇒ Hence same covariance functions as above, operating on quantile functions

⇒ The Wasserstein distance is **beneficial** because less sensitive to very high probability densities

- In **Koulouri et al 2016** and



F. Bachoc, A. Suvorikova, D. Ginsbourger, J-M. Loubes and V. Spokoiny, Gaussian processes with multidimensional distribution inputs via optimal transport and Hilbertian embedding, *arXiv :1805.00753*

extensions are given to distributions on \mathbb{R}^p with $p \geq 2$

Conclusions

- Covariance function drives the **order of magnitude** and **speed of variation** of the Gaussian process
- On \mathbb{R}^d , smooth covariance function \implies smooth Gaussian process
- Catalog of available SNND functions on \mathbb{R}^d
- Canonical extensions to functional inputs

Topics we did not address

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

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

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

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

and variance

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

Illustration

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

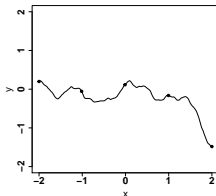
$$\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

Then

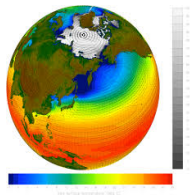
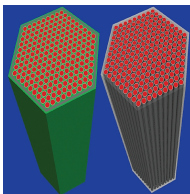
$$\mathbb{E}(Y_2 | Y_1 = y_1) = \mu_2 + \rho(y_1 - \mu_1) \quad \text{and} \quad \text{var}(Y_2 | Y_1 = y_1) = 1 - \rho^2$$

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

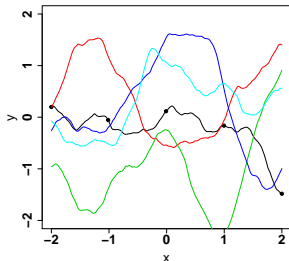
It is a **function interpolation/approximation** problem

Possible methods : polynomial regression, neural networks, splines, RKHS, ...

→ can provide a **deterministic** error bound

Gaussian process model : representing the **deterministic and unknown** function f by a **realization** of a **Gaussian process**.

→ 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 $x_1, \dots, x_n \in \mathbb{X}$

Notations

- Let $\mathbf{Y}_n = (\xi(x_1), \dots, \xi(x_n))^T$ be the observation vector. It is a Gaussian vector
- Let $\mathbf{y}_n = (f(x_1), \dots, f(x_n))^T$ be the observed values
- Let \mathbf{m}_n be the mean vector of \mathbf{Y}_n : $\mathbf{m}_n = (m(x_1), \dots, m(x_n))^T$
- 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 \mathbf{Y}_n :

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

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

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

\implies 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

$$\begin{aligned} m(x) + \mathbf{r}(x)^\top \mathbf{R}^{-1}(\mathbf{y}_n - \mathbf{m}_n) &= m(x) + \mathbf{r}(x)^\top \times \begin{pmatrix} \mathbf{r}(x)^\top \\ * \\ \vdots \\ * \end{pmatrix}^{-1} \times \begin{pmatrix} f(x_1) - m(x_1) \\ \vdots \\ f(x_n) - m(x_n) \end{pmatrix} \\ &= m(x) + (1, 0, \dots, 0) \begin{pmatrix} f(x_1) - m(x) \\ \vdots \\ f(x_n) - m(x_n) \end{pmatrix} = f(x_1) \end{aligned}$$

Conservative extrapolation

Let x be far from x_1, \dots, 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

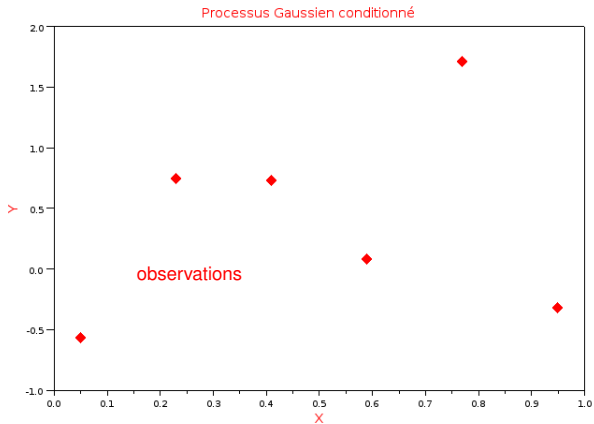


Illustration of Gaussian process prediction

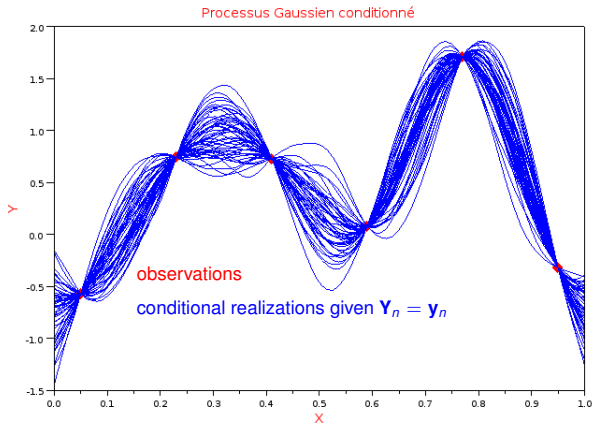


Illustration of Gaussian process prediction

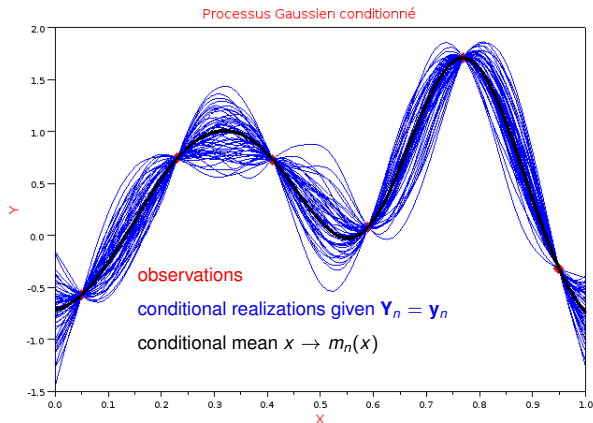
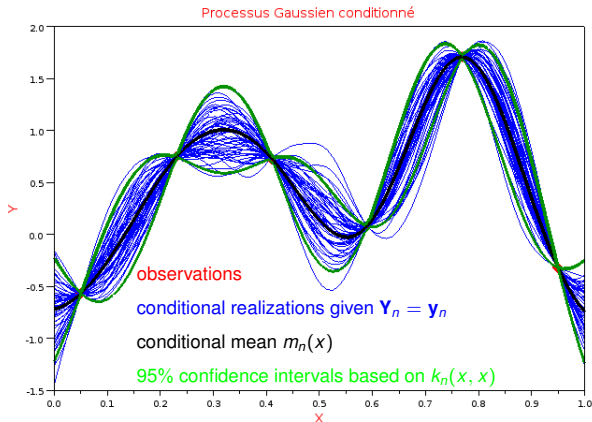


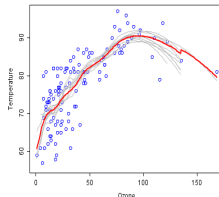
Illustration of Gaussian process prediction



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 \Rightarrow common in machine learning applications
- \Rightarrow E.g. flight delay from flight features



We consider that at x_1, \dots, x_n , we observe

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

$\mathcal{E}_1, \dots, \mathcal{E}_n$ 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

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

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 $u, v \in \mathbb{X}$:

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

\implies 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

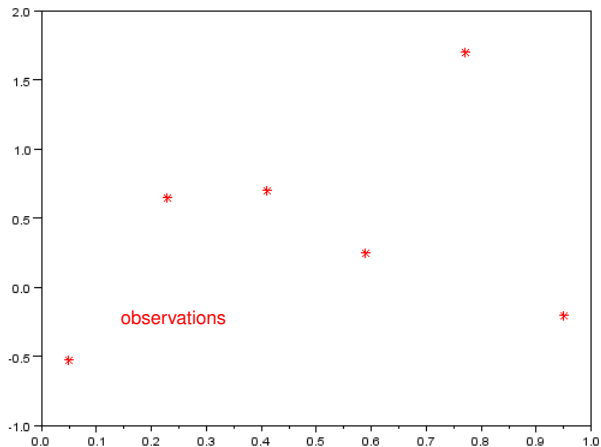


Illustration of Gaussian process prediction with measure error

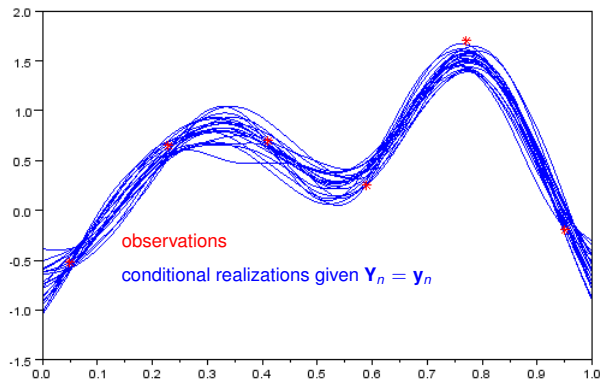


Illustration of Gaussian process prediction with measure error

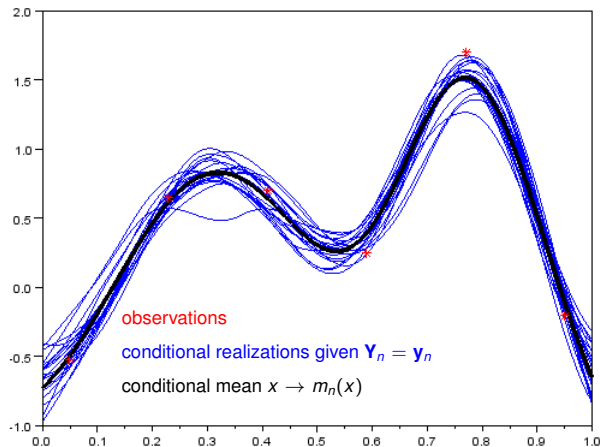
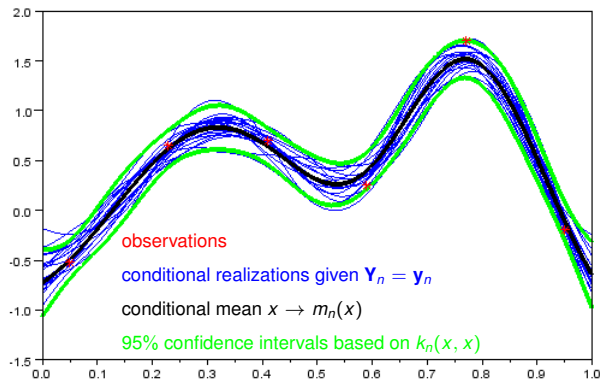


Illustration of Gaussian process prediction with measure error



- 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. Rulli re, 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

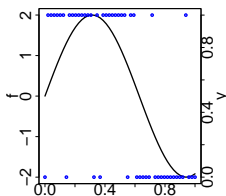
- Gaussian process ξ with realization f
- Observation points x_1, \dots, x_n
- Observation vector

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

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

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

- α large $\Rightarrow \mathbb{P}(Y_i = 1)$ close to 0 or 1 $\Rightarrow Y_i$ 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 $\mathbf{v} = (v_1, \dots, v_n)^\top \in \mathbb{R}^n$,

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

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}
 \implies 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

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

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

approximately following the conditional distribution of \mathbf{V}_n given $\mathbf{Y}_n = \mathbf{y}_n$

⇒ 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 $N \rightarrow \infty$ and for large MCMC budget
- Potentially computationally costly
- Approximations in [Nickisch and Rasmussen, 2008](#) are typically faster (but less guarantees)

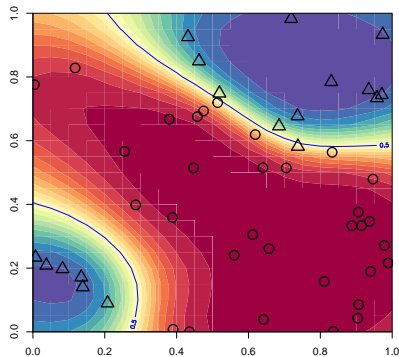


FIGURE: posterior probabilities of 1

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 $x_1, \dots, x_n \in \mathbb{X}$, yielding the Gaussian vector $\mathbf{Y}_n = (\xi(x_1), \dots, \xi(x_n))^\top$.
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

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

Maximum Likelihood

Define \mathbf{C}_θ as the correlation matrix of $\mathbf{Y}_n = (\xi(x_1), \dots, \xi(x_n))^T$ under correlation function c_θ .

The Maximum Likelihood estimator of (σ^2, θ) is

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

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](#)

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

Leave one out estimation

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

and

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

Virtual Leave One Out formula

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

Virtual Leave-One-Out

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



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

Using the virtual Cross Validation formula :

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

and

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

■ 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

Conclusions on Gaussian processes

- Gaussian processes can be defined on any space \mathbb{X} , by using suitable covariance functions
- Setting of direct observations is **favorable** for conditioning \implies 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 \implies Gaussian processes are also a convenient theoretical framework
- Gaussian processes can be used as elementary bricks to construct more complex stochastic processes

1 Introduction to Gaussian processes

2 Sequential learning and consistency of stepwise uncertainty reduction strategies

3 Gaussian processes under inequality constraints

We consider a Gaussian process ξ on a compact $\mathbb{X} \subset \mathbb{R}^d$ with continuous mean function m , continuous covariance function k and continuous sample paths
We consider the setting of **exact direct observations of $\xi(x)$**

Motivation

- When we observe $\xi(x_1), \dots, \xi(x_n)$, the mean and covariance functions become m_n and k_n
- \implies We want to choose x_1, \dots, x_n so that m_n and k_n become **maximally informative** (e.g. $k_n(x, x)$ small, or $k_n(x, x)$ small when $m_n(x)$ is large)

Sequential design

It is more efficient to select x_{i+1} **after** $\xi(x_1), \dots, \xi(x_i)$ are observed

The observation points x_1, \dots, x_n become **random** observation points X_1, \dots, X_n

Definition

A sequence $(X_n)_{n \geq 1}$ of random points in \mathbb{X} will be said to form a (non-randomized) **sequential design** if, for all $n \geq 1$, X_n is \mathcal{F}_{n-1} -measurable, where

$$\mathcal{F}_i = \sigma(X_1, \xi(X_1), \dots, X_i, \xi(X_i))$$

Gaussian measures

- A Gaussian measure ν is a measure on $\mathcal{C}(\mathbb{X})$ corresponding to a Gaussian process with continuous sample paths (see e.g. [Bogachev 98](#)).
- ν is characterized by the mean function m_ν and the covariance function k_ν
- We let $\mathcal{GP}(m_\nu, k_\nu)$ denote the Gaussian measure ν

The conditioning mapping

The conditioning mapping

We let $\text{Cond}_{x_1, z_1, \dots, x_n, z_n}(\nu)$ be the Gaussian measure $\mathcal{GP}(m_{\nu, n}, k_{\nu, n})$ where

$$m_{\nu, n}(x) = m_{\nu}(x) + \mathbf{r}(x)^{\top} \mathbf{R}^{-1}(\mathbf{z}_n - \mathbf{m}_n)$$

and

$$k_n(x, y) = k_{\nu}(x, y) - \mathbf{r}(x)^{\top} \mathbf{R}^{-1} \mathbf{r}(y)$$

with

- $\mathbf{z}_n = (z_1, \dots, z_n)^{\top}$
- \mathbf{R} is the $n \times n$ matrix $[k_{\nu}(x_i, x_j)]$
- $\mathbf{r}(x) = (k_{\nu}(x, x_1), \dots, k_{\nu}(x, x_n))^{\top}$
- $\mathbf{m}_n = (m_{\nu}(x_1), \dots, m_{\nu}(x_n))^{\top}$

A convenient result

For any sequential design of experiment (X_i) , the conditional distribution of ξ (with Gaussian measure ν) given $X_1, \xi(X_1), \dots, X_n, \xi(X_n)$ is $\text{Cond}_{X_1, \xi(X_1), \dots, X_n, \xi(X_n)}(\nu)$

\implies conditioning 'as if' X_1, \dots, X_n were **deterministic**

Let $\nu = \mathcal{GP}(m_\nu, k_\nu)$ be a Gaussian measure and let ξ_ν be a Gaussian process with measure ν

Uncertainty functional

It is a function $\mathcal{H} : \nu \mapsto \mathcal{H}(\nu) \in [0, \infty)$

- Expected improvement (EI) (Mockus 78, Jones et al 98)

$$\mathcal{H}(\nu) = \mathbb{E}(\max_{u \in \mathbb{X}} \xi_\nu(u)) - \max_{u \in \mathbb{X}; k_\nu(u, u) = 0} \mathbb{E}(\xi_\nu(u))$$

- Knowledge gradient (Frazier et al 08, 09)

$$\mathcal{H}(\nu) = \mathbb{E}(\max_{u \in \mathbb{X}} \xi_\nu(u)) - \max_{u \in \mathbb{X}} \mathbb{E}(\xi_\nu(u))$$

- Integrated Bernoulli variance (Bect et al 12, Chevalier et al 14)

$$\mathcal{H}(\nu) = \int_{\mathbb{X}} p_\nu(u)(1 - p_\nu(u)) du$$

with $p_\nu(u) = \mathbb{P}(\xi_\nu(u) \leq T)$ for fixed $T \in \mathbb{R}$

- Variance of excursion volume (Bect et al 12, Chevalier et al 14)

$$\mathcal{H}(\nu) = \text{Var} \left(\int_{\mathbb{X}} \mathbf{1}_{\xi_\nu(u) \leq T} du \right)$$

Let

$$\mathcal{J}_x(\nu) = \mathbb{E} \left(\mathcal{H}(\text{Cond}_{x, \xi_\nu(x)}(\nu)) \right)$$

$\mathcal{J}_x(\nu)$ is the **expected uncertainty** after observing $\xi(x)$

Stepwise Uncertainty Reduction (SUR)

The sequential design (X_i) follows a SUR strategy when

$$X_{i+1} \in \underset{x \in \mathbb{X}}{\operatorname{argmin}} \mathcal{J}_x(\text{Cond}_{X_1, \xi(X_1), \dots, X_i, \xi(X_i)}(\nu_0))$$

with ν_0 the distribution of the Gaussian process ξ

For the examples

Let \mathbb{E}_n , Cov_n and \mathbb{P}_n denote conditional mean, covariance and probability for the distribution of ξ given \mathcal{F}_n

- Expected improvement

$$X_{n+1} \in \operatorname{argmax}_{x \in \mathbb{X}} \mathbb{E}_n \left(\left(\xi(x) - \max_{u \in \mathbb{X}; k_n(u, u)=0} \right)^+ \right)$$

- Knowledge gradient

$$X_{n+1} \in \operatorname{argmax}_{x \in \mathbb{X}} \mathbb{E}_n \left(\max_{u \in \mathbb{X}} \mathbb{E}_n(\xi(u) | \xi(x)) \right)$$

- Integrated Bernoulli variance

$$X_{n+1} \in \operatorname{argmin}_{x \in \mathbb{X}} \mathbb{E}_n \left(\int_{\mathbb{X}} p_{n+1, x}(u) (1 - p_{n+1, x}(u)) du \right)$$

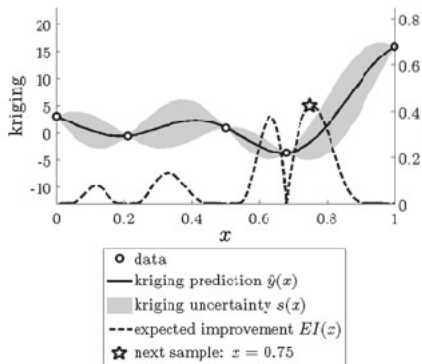
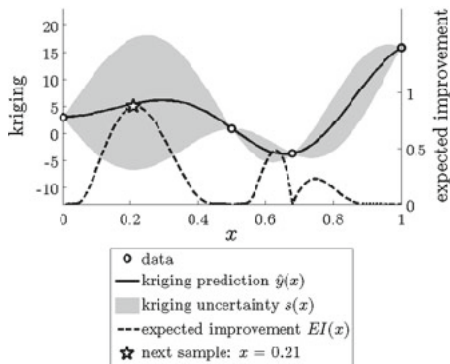
with $p_{n+1, x}(u) = \mathbb{P}_n(\xi(u) \leq T | \xi(x))$

- Variance of excursion volume

$$X_{n+1} \in \operatorname{argmin}_{x \in \mathbb{X}} \mathbb{E}_n \left(\operatorname{Var}_n \left(\int_{\mathbb{X}} \mathbf{1}_{\xi(u) \leq T} du \middle| \xi(x) \right) \right)$$

Illustration of Expected Improvement

(for minimization)



(Figure borrowed from [Viana et al 13, Journal of Global Optimization](#))

- Expected improvement is the most used SUR strategy
 - optimal design (car industry...)
 - optimal fitting of parametric model (chemistry...)
- Integrated Bernoulli variance and Variance of excursion volume are used in failure domain estimation
 - nuclear engineering...
- Knowledge gradient can be used when Expected improvement is used
 - drug discovery...

Remark : Other (non SUR) sequential strategies exist, for instance [Gaussian process upper confidence band \(GP UCB\)](#), for optimization, [Srinivas et al 12](#)

- robotic
- bandit problems

Now we study a joint work with Julien Bect and David Ginsbourger



J. Bect, F. Bachoc and D. Ginsbourger, A supermartingale approach to Gaussian process based sequential design of experiments, *Bernoulli*, forthcoming, 2018

We want to provide general conditions ensuring that

$$\mathcal{H}(\text{Cond}_{X_1, \xi(X_1), \dots, X_n, \xi(X_n)}(\nu_0)) \xrightarrow[n \rightarrow \infty]{a.s.} 0$$

with ν_0 the distribution of the Gaussian process ξ

\Rightarrow **Uncertainty going to zero**

- [Srinivas et al 12](#) provide rates of convergence for the sequential strategy GP-UCB (optimization)
- [Bull 11](#) provides rates of convergence for expected improvement. Here the function f to optimize is deterministic and belongs to the RKHS of k
- [Vazquez and Bect 10](#) prove the consistency of Expected Improvement. They work with covariance functions that are not too smooth and not degenerate (we will improve this point here)
- [Yarotsky 13](#) shows that expected improvement can be inconsistent for *specific* fixed objective functions and covariance functions

Convergence

For any sequential design of experiments (X_i) , a.s. as $n \rightarrow \infty$

- The conditional mean function m_n converges to a random continuous function $m_\infty : \mathbb{X} \rightarrow \mathbb{R}$
- The conditional covariance function k_n converges to a random continuous function $k_\infty : \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}$
- The above convergences are uniform on \mathbb{X} and $\mathbb{X} \times \mathbb{X}$

Proof : the conditional variance is decreasing + martingale arguments

Limit conditioning

Let \mathcal{F}_∞ be the sigma-algebra generated by $\cup_{n \geq 1} \mathcal{F}_n$. Then conditionally to \mathcal{F}_∞ , ξ is a Gaussian process with mean function m_∞ and covariance function k_∞

An 'Ad Hoc' convergence for Gaussian measures

Definition

Let (ν_n) denote a sequence of Gaussian measures. We will say that (ν_n) is an *almost surely convergent sequence of conditional distributions* if

- i) there exists a random Gaussian measure ν_∞ such a.s., as $n \rightarrow \infty$, m_{ν_n} and k_{ν_n} converge to m_{ν_∞} and k_{ν_∞} uniformly on \mathbb{X} and $\mathbb{X} \times \mathbb{X}$;
- ii) there exists a Gaussian process ξ such that, for all $n \in \mathbb{N} \cup \{+\infty\}$, $\nu_n = P(\xi \in \cdot \mid \tilde{\mathcal{F}}_n)$ for some σ -algebra $\tilde{\mathcal{F}}_n \subset \mathcal{F}$.

Two examples

- For any sequential design, the conditional distribution $P_n^\xi = P(\xi \in \cdot \mid \mathcal{F}_n)$ converges almost surely to $P_\infty^\xi = P(\xi \in \cdot \mid \mathcal{F}_\infty)$
- Let $x_\infty \in \mathbb{X}$ such that $k(x_\infty, x_\infty) > 0$. Let (x_i) be a sequence in \mathbb{X} such that $x_i \rightarrow x_\infty$. For each $i \in \mathbb{N} \cup \{+\infty\}$, let $\nu_i = \text{Cond}_{x_i, \xi(x_i)}(P_0^\xi)$. Then (ν_i) is an almost surely convergent sequence of conditional distributions with limit ν_∞ .

Supermartingale property

Definition

The functional \mathcal{H} is said to have the *supermartingale property* if, for any sequential design X_1, X_2, \dots , the sequence $(\mathcal{H}(P_n^\xi))$ is an (\mathcal{F}_n) -supermartingale

The supermartingale property holds for the four examples.

Expected improvement

with $P_{n+1, \xi(x)}^\xi = \text{Cond}_{x, \xi(x)}(P_n^\xi)$

$$\begin{aligned}\mathcal{H}(P_n^\xi) - \mathbb{E}_n[\mathcal{H}(P_{n+1, \xi(X_{n+1})}^\xi)] &= \mathbb{E}_n(\max_{u \in \mathbb{X}} \xi(u)) - \mathbb{E}_n \left(\mathbb{E}_n(\max_{u \in \mathbb{X}} \xi(u) | \xi(X_{n+1})) \right) \\ &\quad - \max_{k_n(u, u)=0} \mathbb{E}_n(\xi(u)) + \mathbb{E}_n \left(\max_{k_n(u, u | \xi(X_{n+1}))=0} \mathbb{E}_n(\xi(u) | \xi(X_{n+1})) \right) \\ &\geq \mathbb{E}_n \left(\max_{k_n(u, u)=0} \mathbb{E}_n(\xi(u) | \xi(X_{n+1})) \right) - \max_{k_n(u, u)=0} \mathbb{E}_n(\xi(u)) \\ &= \max_{k_n(u, u)=0} \xi(u) - \max_{k_n(u, u)=0} \xi(u) \\ &= 0\end{aligned}$$

from law of total variance and since $k_n(u, u | \xi(x)) = \text{Var}_n(\xi(u) | \xi(x)) \leq k_n(u, u)$

Integrated Bernoulli variance

Let $p_{n+1,x,z}(u) = \mathbb{E}_n(\mathbf{1}_{\xi(u) \leq T} | \xi(x) = z)$

$$\begin{aligned}\mathbb{E}_n[\mathcal{H}(P_{n+1,\xi(X_{n+1})}^\xi)] &= \mathbb{E}_n \left(\int_{\mathbb{X}} p_{n+1,X_{n+1},\xi(X_{n+1})}(u) (1 - p_{n+1,X_{n+1},\xi(X_{n+1})}(u)) du \right) \\ &= \int_{\mathbb{X}} \mathbb{E}_n (\text{var}_n(\mathbf{1}_{\xi u \leq T} | \xi(X_{n+1}))) du \\ &\leq \int_{\mathbb{X}} \text{var}_n(\mathbf{1}_{\xi u \leq T}) du \\ &= \mathcal{H}(P_n^\xi)\end{aligned}$$

The convergence result

Let

$$\mathcal{G}(\nu) = \sup_{x \in \mathbb{X}} \left(\mathcal{H}(\nu) - \mathbb{E}(\mathcal{H}(\text{Cond}_{x, \xi_{\nu(x)}}(\nu))) \right)$$

(maximum expected uncertainty reduction)

Theorem

Let \mathcal{H} denote an uncertainty functional with the supermartingale property.
Let (X_n) denote a SUR sequential design for \mathcal{H}

$$X_{n+1} \in \underset{x \in \mathbb{X}}{\operatorname{argmin}} \mathbb{E}(\mathcal{H}(\text{Cond}_{x, \xi(x)}(P_n^\xi)))$$

Then $\mathcal{G}(P_n^\xi) \rightarrow 0$ almost surely. If, moreover,

- i) $\mathcal{H}(P_n^\xi) \rightarrow \mathcal{H}(P_\infty^\xi)$ almost surely
 - ii) $\mathcal{G}(P_n^\xi) \rightarrow \mathcal{G}(P_\infty^\xi)$ almost surely
 - iii) $\mathcal{G}(\nu) = 0 \implies \mathcal{H}(\nu) = 0$
- then $\mathcal{H}(P_n^\xi) \rightarrow 0$ almost surely

Assumptions i) and ii) are continuity assumptions

Assumption iii) is essential, it means

no possible uncertainty reduction with one more observation \implies no uncertainty

- We prove that the general results apply to the four examples
- We introduce the notion of **regular loss function**, where \mathcal{H} is an average loss when estimating a quantity of interest (e.g. maximum of ξ , $\{u \in \mathbb{X} : \xi(u) \leq T\}, \dots$)
- We provide a specific convergence result for regular loss functions, with easier to check assumptions

Summary

- The probabilistic framework of Gaussian processes enables to define expected uncertainties and Stepwise Uncertainty Reduction (SUR) strategies
- We prove convergence of SUR strategies
- **Remark :** Our proof does not rely on showing that (X_i) is almost surely dense in \mathbb{X} . We allow for degenerate or very smooth covariance functions. Sometimes we do not need $\sup_{u \in \mathbb{X}} k_n(u, u) \rightarrow 0$

Two open questions

- When the covariance function is estimated (frequentist or Bayesian)
- Rate of convergence

1 Introduction to Gaussian processes

2 Sequential learning and consistency of stepwise uncertainty reduction strategies

3 Gaussian processes under inequality constraints

This last section is based on the PhD thesis of Andr  s Felipe Lopez Lopera \implies to be defended on September 19

We consider a Gaussian process ξ on $\mathbb{X} = [0, 1]^d$ (with mean function 0 for simplicity) for which we assume that additional information is available :

- $\xi(x)$ belongs to $[\ell, u]$ for $x \in [0, 1]^d$ (**boundedness constraints**)
- $\partial/\partial x_i \xi(x) \geq 0$ for $x \in [0, 1]^d$ and $i = 1, \dots, d$ (**monotonicity constraints**)
- ξ is convex on $[0, 1]^d$ (**convexity constraints**)
- Modifications and/or combinations of the above constraints

Application cases :

- Quantity of interest belongs to \mathbb{R}^+ (energy) or $[0, 1]$ (concentration, energetic efficiency)
- Inputs are known to have positive effects (more input power \rightarrow more output energy)

Generic form of the constraints :

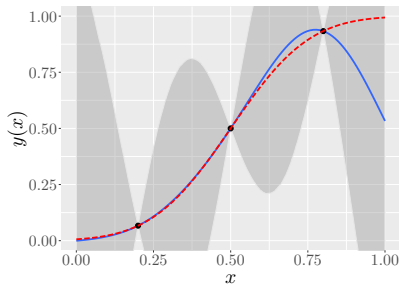
$$\xi \in \mathcal{E}$$

where \mathcal{E} is a set of functions from $[0, 1]^d \rightarrow \mathbb{R}$ so that $P(\xi \in \mathcal{E}) > 0$

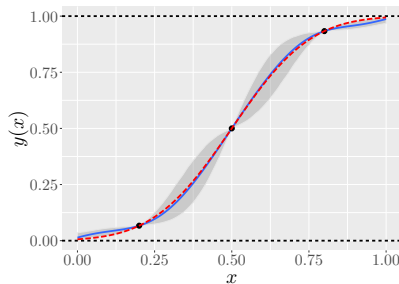
Impact :

- **New stochastic model** : The law of the realization function is $P(\xi \in \cdot | \xi \in \mathcal{E})$
- **New conditional distribution** : Conditional distribution of ξ given $\xi \in \mathcal{E}$ and $\xi(x_1) = y_1, \dots, \xi(x_n) = y_n$
- **New estimation** of the covariance parameters θ in the covariance model $\{k_\theta; \theta \in \Theta\}$

Illustration of constraint benefits



Unconstrained Gaussian process



Constrained Gaussian process

- true function
- predictive mean
- training points
- confidence intervals

Existing work

- For boundedness constraints, it is possible to consider models of the form $y_i = T(\xi(x_i))$ with T bijective from \mathbb{R} to $[\ell, u]$ and ξ a Gaussian process
- For monotonicity and convexity constraints, the approach $P(\xi \in \cdot | \xi \in \mathcal{E})$ has become standard
- \Rightarrow but the constraint $\xi \in \mathcal{E}$ needs to be approximated
- $\xi \in \mathcal{E}$ is replaced by a finite number of constraints on inducing points in



S. Da Veiga and A. Marrel, Gaussian process modeling with inequality constraints, *Annales de la faculté des sciences de Toulouse Mathématiques* 21 (2012) 529-555.



S. Golchi, D. Bingham, H. Chipman and D.A. Campbell, Monotone emulation of computer experiments, *SIAM/ASA Journal on Uncertainty Quantification* 3 (2015) 370-392.

- ξ is replaced by a **finite-dimensional approximation** ξ_m in



H. Maatouk and X. Bay, Gaussian process emulators for computer experiments with inequality constraints, *Mathematical Geosciences* 49(5) (2017) 557-582.

(we follow this latter approach)

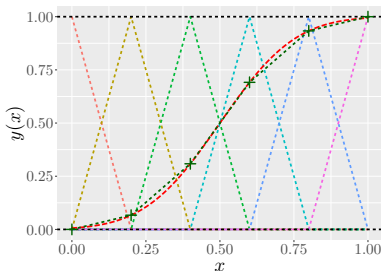
The finite dimensional approximation

Maatouk and Bay 2017 suggest to consider, in dimension $d = 1$,

$$\xi_m(t) = \sum_{j=1}^m E_j \phi_j(t),$$

where

- $E_j = \xi(t_j)$
- $t_1 = 0, t_2 = 1/(m-1), \dots, t_m = 1$
- the ϕ_j are hat functions, $\phi_j(t) = (1 - (m-1)|t - t_j|)^+$ for $j = 1, \dots, m$



The finite dimensional approximation

Computational benefit ([Maatouk and Bay 2017](#)) :

- $\ell \leq \xi_m \leq u \iff \ell \leq E_1, \dots, E_m \leq u$
- ξ_m is a non-decreasing function $\iff E_1 \leq \dots \leq E_m$
- ξ_m is a convex function $\iff E_2 - E_1 \leq \dots \leq E_m - E_{m-1}$

\implies Only a finite number of inequalities \implies guarantee to satisfy the constraints everywhere on $[0, 1]$

Extension to dimension 2

$$\xi_m(t_1, t_2) = \sum_{j_1, j_2=1}^m E_{j_1, j_2} \phi_{j_1}(t_1) \phi_{j_2}(t_2)$$

- Becomes problematic in higher dimension
- We are developing other approaches (cf later)

With the finite-dimensional approximation

$$\xi_m(t) = \sum_{j=1}^m E_j \phi_j(t),$$

we study linear constraints of the form

$$\ell \leq \mathbf{\Lambda} \mathbf{E} \leq \mathbf{u}$$

where

- $\mathbf{E} = (E_1, \dots, E_m)^\top$
- $\mathbf{\Lambda}$ is a $q \times m$ matrix
- ℓ and \mathbf{u} are $q \times 1$ vectors
- boundedness, monotonicity, convexity constraints can be enforced, as well as combinations

\implies After observed values \mathbf{y}_n of $(\xi(x_1), \dots, \xi(x_n))^\top$, the conditional distribution is

$$\mathcal{L}(\mathbf{\Lambda} \mathbf{E} | \mathbf{\Phi} \mathbf{E} = \mathbf{y}_n, \ell \leq \mathbf{\Lambda} \mathbf{E} \leq \mathbf{u}),$$

where $\mathbf{\Phi} = [\phi_j(x_i)]_{i=1, \dots, n, j=1, \dots, m}$ is $n \times m$

Sampling problem

Let \mathbf{M} be the covariance matrix of $\mathbf{E} = (E_1, \dots, E_m)^\top = (\xi(t_1), \dots, \xi(t_m))^\top$

We have

$$\begin{aligned}\mathcal{L}(\boldsymbol{\Lambda}\mathbf{E}|\boldsymbol{\Phi}\mathbf{E} = \mathbf{y}_n) &= \mathcal{N}\left(\boldsymbol{\Lambda}\mathbf{M}\boldsymbol{\Phi}^\top(\boldsymbol{\Phi}\mathbf{M}\boldsymbol{\Phi}^\top)^{-1}\mathbf{y}_n, \boldsymbol{\Lambda}\mathbf{M}\boldsymbol{\Lambda}^\top - \boldsymbol{\Lambda}\mathbf{M}\boldsymbol{\Phi}^\top(\boldsymbol{\Phi}\mathbf{M}\boldsymbol{\Phi}^\top)^{-1}\boldsymbol{\Phi}\mathbf{M}\boldsymbol{\Lambda}^\top\right) \\ &:= \mathcal{N}(\boldsymbol{\Lambda}\boldsymbol{\mu}, \boldsymbol{\Lambda}\boldsymbol{\Sigma}\boldsymbol{\Lambda}^\top)\end{aligned}$$

Hence the sampling problem is to sample

$$\mathbf{V}_n \sim \mathcal{N}(\boldsymbol{\Lambda}\boldsymbol{\mu}, \boldsymbol{\Lambda}\boldsymbol{\Sigma}\boldsymbol{\Lambda}^\top),$$

conditionally to $\ell \leq \mathbf{V}_n \leq \mathbf{u}$

- We take $\boldsymbol{\Lambda}$ injective so that $\mathbf{V}_n \implies \mathbf{E} \implies \xi_m$
- Computing $\operatorname{argmax}_{\tilde{\mathbf{V}}} p_{\mathbf{V}_n}(\tilde{\mathbf{V}}|\ell \leq \mathbf{V}_n \leq \mathbf{u})$ provides the **mode**
- Computing $\mathbb{E}(\mathbf{V}_n|\ell \leq \mathbf{V}_n \leq \mathbf{u})$ provides the **conditional mean**
- Sampling \mathbf{V}_n given $\ell \leq \mathbf{V}_n \leq \mathbf{u}$ provides **conditional samples**

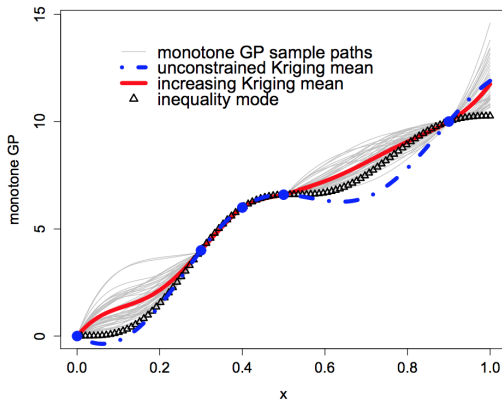


FIGURE: (from [Maatouk and Bay 2017](#)) Illustration of conditional samples with constraints (monotone GP sample paths), conditional mean without constraints (unconstrained Kriging mean), conditional mean with monotonicity constraints (increasing Kriging mean) and mode with monotonicity constraints (inequality mode)

The mode is obtained by solving

$$\hat{\mathbf{v}}_n \in \underset{\substack{\mathbf{v} \in \mathbb{R}^q \\ \ell \leq \mathbf{v} \leq \mathbf{u}}}{\operatorname{argmin}} (\mathbf{V} - \mathbf{\Lambda}\boldsymbol{\mu})^\top (\mathbf{\Lambda}\Sigma\mathbf{\Lambda}^\top)^{-1} (\mathbf{v} - \mathbf{\Lambda}\boldsymbol{\mu})$$

- quadratic function optimization subject to linear inequality constraints
- quite fast algorithms
- corresponds to the (unconstrained) conditional mean $\mathbf{\Lambda}\boldsymbol{\mu}$ if it satisfies the inequality constraints

Sampling $\mathbf{V}_n \sim \mathcal{N}(\mathbf{\Lambda}\boldsymbol{\mu}, \mathbf{\Lambda}\Sigma\mathbf{\Lambda}^\top)$ subject to $\ell \leq \mathbf{V}_n \leq \mathbf{u}$:

- rejection sampling from the mode [Maatouk and Bay 2017](#) (low acceptance rate for q large)

We investigate

- Hastings metropolis
- Gibbs sampling (never rejects) [Taylor and Benjamini 2017](#)
- Minimax tilting [Botev 2017 JRSSB](#)
- Hamiltonian Monte Carlo [Pakman and Paninski 2014 JCGS](#)

and conclude that Hamiltonian Monte Carlo is an efficient sampler in our framework

An application to nuclear engineering

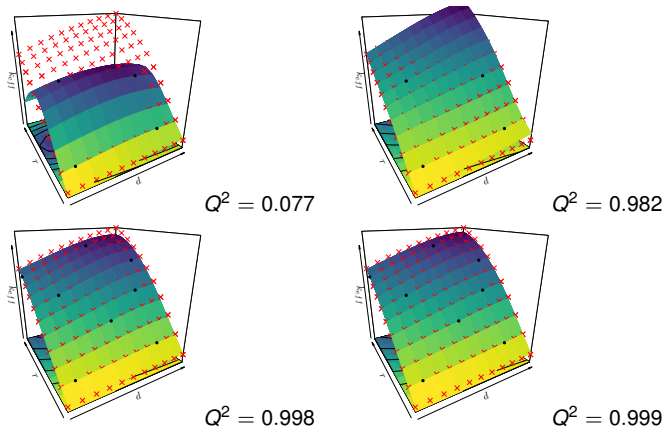


FIGURE: Two dimensional nuclear engineering example. **radius** and **density** of uranium sphere \Rightarrow **criticality coefficient**. **Mononicity constraints**. Left : unconstrained Gaussian process models. Right : constrained Gaussian process models. The Q^2 measures the prediction quality and should be close to 1



A. F. López-Lopera, F. Bachoc, N. Durrande and O. Roustant, Finite-dimensional Gaussian approximation with linear inequality constraints, *SIAM/ASA Journal on Uncertainty Quantification*, forthcoming.

Adaptation to higher dimension

- In dimension $d \geq 5$, say, we can not use the full grid approach
- We aim for a representation

$$\xi_m = \text{function}(E_1, \dots, E_m)$$

so that we keep

$$\xi_m \in \mathcal{E} \iff (E_1, \dots, E_m) \in \mathcal{C}$$

- **Approach 1** : additive Gaussian processes

$$\xi_m(x_1, \dots, x_d) = \sum_{i=1}^d \xi_{m,i}(x_i) + \sum_{\substack{i,j=1,\dots,d \\ i \neq j}} \xi_{m,i,j}(x_i, x_j)$$

with grids in dimensions 1 and 2.

- **Approach 2** : Tensorized grid with less grid points for less important variables

Covariance parameter estimation under constraints

Setting

- For simplicity, let us forget about the finite-dimensional approximation ξ_m (but see the papers)
- We observe the Gaussian process ξ at $x_1, \dots, x_n \in [0, 1]^d$ and let $\mathbf{Y}_n = (\xi(x_1), \dots, \xi(x_n))^\top$
- We assume that ξ has covariance function k
- We consider the model of covariance functions $\{k_\theta; \theta \in \Theta\}$
- The inequality constraints are $\xi \in \mathcal{E}$

The maximum likelihood estimator of θ is

$$\hat{\theta}_{ML} \in \operatorname{argmax}_{\theta \in \Theta} \mathcal{L}_n(\theta)$$

with

$$\mathcal{L}_n(\theta) = \log(p_\theta(\mathbf{Y}_n)) = \log \left(\frac{1}{(2\pi)^{n/2} |\mathbf{R}_\theta|} e^{-\frac{1}{2} \mathbf{Y}_n^\top \mathbf{R}_\theta^{-1} \mathbf{Y}_n} \right)$$

- (it ignores the information $\xi \in \mathcal{E}$)
- explicit expression of \mathcal{L}_n with $O(n^3)$ cost

The constrained maximum likelihood estimator of θ is

$$\hat{\theta}_{cML} \in \operatorname{argmax}_{\theta \in \Theta} \mathcal{L}_{C,n}(\theta)$$

with

$$\mathcal{L}_{C,n}(\theta) = \log(p_{\theta}(\mathbf{Y}_n)) - \log(p_{\theta}(\xi \in \mathcal{E})) + \log(p_{\theta}(\xi \in \mathcal{E} | \mathbf{Y}_n))$$

- The additional terms $\log(p_{\theta}(\xi \in \mathcal{E}))$ and $\log(p_{\theta}(\xi \in \mathcal{E} | \mathbf{Y}_n))$ have no explicit expressions
- They need to be approximated by numerical integration or Monte Carlo : [Genz 1992 JCGS](#), [Botev 2017 JRSSB](#)

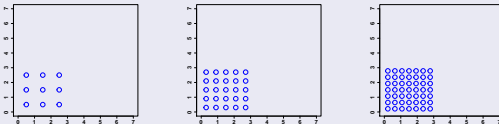
⇒ We aim at comparing $\hat{\theta}_{ML}$ and $\hat{\theta}_{cML}$ asymptotically

Two asymptotic frameworks for covariance parameter estimation

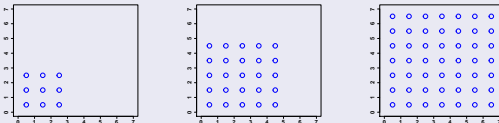
- Asymptotics (number of observations $n \rightarrow +\infty$) is an active area of research
- (case without constraints so far)
- There are **several asymptotic frameworks** because there are several possible **location patterns** for the observation points





Two main asymptotic frameworks

- **fixed-domain asymptotics** : The observation points are dense in a bounded domain



- **increasing-domain asymptotics** : number of observation points is proportional to domain volume \rightarrow unbounded observation domain.



- Consistent estimation is possible for all covariance parameters (that are identifiable in finite-sample). [asymptotic **independence** between observations]
- Asymptotic normality proved for maximum likelihood
 -  Mardia K, Marshall R, Maximum likelihood estimation of models for residual covariance in spatial regression, *Biometrika* 71 (1984) 135-146.
 -  N. Cressie and S.N Lahiri, The asymptotic distribution of REML estimators, *Journal of Multivariate Analysis* 45 (1993) 217-233.
 -  N. Cressie and S.N Lahiri, Asymptotics for REML estimation of spatial covariance parameters, *Journal of Statistical Planning and Inference* 50 (1996) 327-341.
 -  F. Bachoc, Asymptotic analysis of the role of spatial sampling for covariance parameter estimation of Gaussian processes, *Journal of Multivariate Analysis* 125 (2014) 1-35.

- Consistent estimation is **impossible** for some covariance parameters (identifiable in finite-sample), see e.g.



Zhang, H., Inconsistent Estimation and Asymptotically Equivalent Interpolations in Model-Based Geostatistics, *Journal of the American Statistical Association* (99), 250-261, 2004.



Stein M, Interpolation of Spatial Data : Some Theory for Kriging, *Springer, New York, 1999.*

- covariance parameters that **can not** be estimated consistently are called **non-microergodic**
- covariance parameters that **can** be estimated consistently are called **microergodic**
- For instance, consider the set of covariance functions $\{k_\theta, \theta \in (0, \infty)^2\}$ on $[0, 1]$ given by $\theta = (\sigma^2, \alpha)$ and $k_\theta(t_1, t_2) = \sigma^2 e^{-\alpha|t_1 - t_2|}$
 - σ^2 is non-microergodic
 - α is non-microergodic
 - $\sigma^2 \alpha$ is microergodic

⇒ We address fixed-domain asymptotics here

Preservation of consistency

Setting :

- ξ is a Gaussian process on $[0, 1]^d$, $d \in \mathbb{N}$, with mean zero and covariance function k
- $\theta = (\sigma^2, \alpha_1, \dots, \alpha_d)$
- k_θ is the covariance function of the Gaussian process $(x_1, \dots, x_d) \rightarrow \sigma^2 \xi(\alpha_1 x_1, \dots, \alpha_d x_d)$
- ⇒ $k = k_{\theta_0}$ with $\theta_0 = (1, \dots, 1)$
- The constraints are given by the set \mathcal{E} and are **boundedness, monotonicity or convexity**
- $(x_i)_{i \in \mathbb{N}}$ is dense in $[0, 1]^d$

Proposition : preservation of consistency for ML (López-Lopera, Bachoc, Durrande, Roustant 2018)

Assume that the covariance function k satisfy technical conditions (see papers).

Assume $\forall \varepsilon > 0$,

$$P(\|\hat{\theta}_{ML} - \theta_0\| \geq \varepsilon) \xrightarrow{n \rightarrow \infty} 0 \quad (\text{unconditional consistency of ML})$$

Then, we have $P(\xi \in \mathcal{E}) > 0$, and thus

$$P(\|\hat{\theta} - \theta_0\| \geq \varepsilon \mid \xi \in \mathcal{E}) \xrightarrow{n \rightarrow \infty} 0 \quad (\text{conditional consistency of ML})$$

Preservation of consistency

Proposition : preservation of consistency for cML (López-Lopera, Bachoc, Durrande, Roustant 2018)

Assume that the covariance function k satisfy technical conditions (see papers).

Assume that $\forall \varepsilon > 0$ and $\forall M < \infty$, (sufficient condition for unconditional consistency of ML)

$$P\left(\sup_{\|\theta - \theta_0\| \geq \varepsilon} (\mathcal{L}_n(\theta) - \mathcal{L}_n(\theta_0)) \geq -M\right) \xrightarrow{n \rightarrow \infty} 0$$

Then, (sufficient condition for conditional consistency of cML)

$$P\left(\sup_{\|\theta - \theta_0\| \geq \varepsilon} (\mathcal{L}_{C,n}(\theta) - \mathcal{L}_{C,n}(\theta_0)) \geq -M \mid \xi \in \mathcal{E}\right) \xrightarrow{n \rightarrow \infty} 0$$

Consequently (conditional consistency of ML and cML)

$$\hat{\theta}_{ML} \xrightarrow[n \rightarrow \infty]{P|\xi \in \mathcal{E}} \theta_0 \quad \text{and} \quad \hat{\theta}_{cML} \xrightarrow[n \rightarrow \infty]{P|\xi \in \mathcal{E}} \theta_0$$

Setting :

- Gaussian process ξ on $[0, 1]^d$, $d \in \mathbb{N}$, with zero mean function and covariance function k
- Monotonicity, boundedness or convexity constraints (as before)
- $(x_i)_{i \in \mathbb{N}}$ is dense in $[0, 1]^d$
- $\theta = \sigma^2$ and $k_\theta(u_1, u_2) = \sigma^2 k(u_1, u_2)$

Known results

- It is well-known that in this case

$$\sqrt{n} \left(\hat{\sigma}_{ML}^2 - \sigma_0^2 \right) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} N(0, 2\sigma_0^4)$$

Asymptotic normality result 1 : variance estimation

Notation : we write $X_n \xrightarrow[n \rightarrow \infty]{\mathcal{L}|\xi \in \mathcal{E}} L$ when for all bounded measurable function f :

$$\mathbb{E}(f(X_n)|\xi \in \mathcal{E}) \xrightarrow{n \rightarrow \infty} \int f(x) dL(x)$$

Theorem (Bachoc, Lagnoux, López-Lopera 2018)

Under technical conditions on k and the sequence $(x_i)_{i \in \mathbb{N}}$ (see papers), we have

$$\sqrt{n} \left(\hat{\sigma}_{ML}^2 - \sigma_0^2 \right) \xrightarrow[n \rightarrow \infty]{\mathcal{L}|\xi \in \mathcal{E}} N(0, 2\sigma_0^4)$$

and

$$\sqrt{n} \left(\hat{\sigma}_{cML}^2 - \sigma_0^2 \right) \xrightarrow[n \rightarrow \infty]{\mathcal{L}|\xi \in \mathcal{E}} N(0, 2\sigma_0^4)$$

- Same asymptotic distribution as the (unconstrained) maximum likelihood estimator, in the unconstrained case
- No asymptotic impact of the constraints

Asymptotic normality result 2 : Matérn model

Setting :

- Gaussian process ξ on $[0, 1]^d$, $d = 1, 2, 3$, with zero mean function and covariance function k
- Monotonicity, boundedness or convexity constraints (as before)
- $(x_i)_{i \in \mathbb{N}}$ is dense in $[0, 1]^d$
- $\theta = (\sigma^2, \rho) \in (0, \infty)^2$ and

$$k_{\theta, \nu}(x, x') = \sigma^2 K_{\nu} \left(\frac{\|x - x'\|}{\rho} \right) = \frac{\sigma^2}{\Gamma(\nu) 2^{\nu-1}} \left(\frac{\|x - x'\|}{\rho} \right)^{\nu} \kappa_{\nu} \left(\frac{\|x - x'\|}{\rho} \right)$$

- Γ is the Gamma function
- κ_{ν} is the modified Bessel function of the second kind
- $\nu > 0$ (assumed known) is the smoothness parameter : $\nu > r \iff$ corresponding Gaussian process if r times differentiable

In this case :

- σ^2 is non-microergodic
- ρ is non-microergodic
- $\sigma^2 / \rho^{2\nu}$ is microergodic and

$$\sqrt{n} \left(\frac{\hat{\sigma}_{ML}^2}{\hat{\rho}_{ML}^{2\nu}} - \frac{\sigma_0^2}{\rho_0^{2\nu}} \right) \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \mathcal{N} \left(0, 2 \left(\frac{\sigma_0^2}{\rho_0^{2\nu}} \right)^2 \right)$$



C. G. Kaufman and B. A. Shaby, The Role of the Range Parameter for Estimation and Prediction in Geostatistics, *Biometrika* 100 (2013) 473–484.

Asymptotic normality result 2 : Matérn model

We show

Theorem (Bachoc, Lagnoux, López-Lopera 2018)

Under technical conditions on ν and the sequence $(x_i)_{i \in \mathbb{N}}$ (see papers), we have

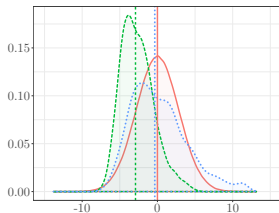
$$\sqrt{n} \left(\frac{\hat{\sigma}_{ML}^2}{\hat{\rho}_{ML}^{2\nu}} - \frac{\sigma_0^2}{\rho_0^{2\nu}} \right) \xrightarrow[n \rightarrow +\infty]{\mathcal{L} | \xi \in \mathcal{E}} \mathcal{N} \left(0, 2 \left(\frac{\sigma_0^2}{\rho_0^{2\nu}} \right)^2 \right)$$

and

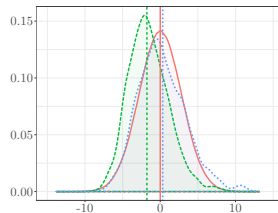
$$\sqrt{n} \left(\frac{\hat{\sigma}_{cML}^2}{\hat{\rho}_{cML}^{2\nu}} - \frac{\sigma_0^2}{\rho_0^{2\nu}} \right) \xrightarrow[n \rightarrow +\infty]{\mathcal{L} | \xi \in \mathcal{E}} \mathcal{N} \left(0, 2 \left(\frac{\sigma_0^2}{\rho_0^{2\nu}} \right)^2 \right)$$

- Same conclusions as for the estimation of a variance parameter

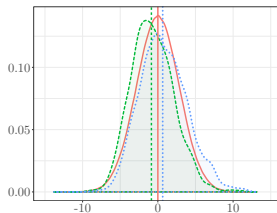
An illustration



$n = 20$



$n = 50$



$n = 80$

FIGURE: An example with the estimation of σ^2 with boundedness constraints. Distribution of $n^{1/2}(\hat{\sigma}^2 - \sigma_0^2)$. $n = 20$ (top left), $n = 50$ (top right) and $n = 80$ (bottom). Green : ML. Blue : cML. Red : Gaussian limit

For consistency :



A. F. López-Lopera, F. Bachoc, N. Durrande and O. Roustant, Finite-dimensional Gaussian approximation with linear inequality constraints, *SIAM/ASA Journal on Uncertainty Quantification*, forthcoming

For asymptotic normality :



F. Bachoc, Agnès Lagnoux and A. F. López-Lopera, Maximum likelihood estimation for Gaussian processes under inequality constraints, *Electronic Journal of Statistics*, forthcoming










Summary








- Inequality constraints correspond to additional information (e. g. physical knowledge)
- Taking them into account can significantly improve the predictions
- With a computational cost (explicit \implies Monte Carlo)
- The constrained maximum likelihood estimator (cML) has similar consistency guarantees as maximum likelihood (ML)
- Asymptotically, we do not see an impact of the constraints and $ML \approx cML$
- For small sample size, cML appears to be beneficial








Ongoing work

- The finite-dimensional approach in higher dimension

Thank you for your attention !

-  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
-  B. Scholkopf and 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
-  R. J. Adler, An introduction to continuity, extrema, and related topics for general Gaussian processes, *IMS*, 1990
-  T. Muehlenstaedt, J. Fruth and O. Roustant , Computer experiments with functional inputs and scalar outputs by a norm-based approach, *Statistics and Computing*, 27 (4) 1083-1097, 2017
-  M. Morris, Gaussian surrogates for computer models with time-varying inputs and outputs, *Technometrics*, 54, 42-50, 2012
-  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
-  V. I. Bogachev, Gaussian measures, *American Mathematical Soc.*, 1998

-  J. B. Mockus, V. Tiesis and A. Zilinskas, The application of Bayesian methods for seeking the extremum, *In Dixon, L. C. W. and Szegö, G. P., editors, Towards Global Optimization*, volume 2, pages 117-129, North Holland, New York, 1978
-  D. Jones, M. Schonlau and W. Welch, Efficient global optimization of expensive black box functions, *Journal of Global Optimization*, 13 :455-492, 1998
-  Frazier, P. I., Powell, W. B., and Dayanik, S., A knowledge-gradient policy for sequential information collection, *SIAM Journal on Control and Optimization*, 47(5) :2410-2439 (2008)
-  Frazier, P. I., Powell, W. B., and Dayanik, S., The knowledge-gradient policy for correlated normal beliefs, *INFORMS Journal on Computing*, 21(4) :599-613, 2009
-  Bect, J., Ginsbourger, D., Li, L., Picheny, V., and Vazquez, E., Sequential design of computer experiments for the estimation of a probability of failure, *Statistics and Computing*, 22 (3) :773-793, 2012
-  Chevalier, C., Bect, J., Ginsbourger, D., Vazquez, E., Picheny, V., and Richet, Y., Fast parallel kriging-based stepwise uncertainty reduction with application to the identification of an excursion set, *Technometrics*, 56(4) :455-465, 2014
-  Srinivas, N., A., K., Kakade, S., and Seeger, M., Information-theoretic regret bounds for Gaussian process optimization in the bandit setting, *IEEE Transactions on Information Theory*, 58 :3250-3265, 2012

-  Taylor, J. and Benjamini, Y., RestrictedMVN : multivariate normal restricted by affine constraints, *CRAN*, 2017
-  Botev, Z. I., The normal law under linear restrictions : simulation and estimation via minimax tilting, *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 79(1), 125-148, 2017
-  Pakman, A. and Paninski, L., Exact Hamiltonian Monte Carlo for Truncated Multivariate Gaussians, *Journal of Computational and Graphical Statistics*, 23(2), 518-542, 2014
-  A. Genz, Numerical Computation of Multivariate Normal Probabilities, *Journal of Computational and Graphical Statistics*, 1, 141-150, 1992
-  Bull, A. D., Convergence rates of efficient global optimization algorithms, *Journal of Machine Learning Research*, 12 :2879-2904, 2011
-  Vazquez, E. and Bect, J., Convergence properties of the expected improvement algorithm with fixed mean and covariance functions, *Journal of Statistical Planning and Inference*, 140(11) :3088-3095, 2010
-  Yarotsky, D., Examples of inconsistency in optimization by expected improvement, *Journal of Global Optimization*, 56(4) : 1773-1790, 2013