Metamodeling course: Chapter 1

O. Roustant

INSA Toulouse, September 2024

Random processes

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space on which all the (real-valued) random variables will be defined. We denote by $L^2(\mathbb{P})$ the Hilbert space of square integrable random variables (defined on Ω).

For a given set $\mathbb{X}$, a *random process* (RP) is a family of random variables $Y(x) : \Omega \to \mathbb{R}$, indexed by $x \in \mathbb{X}$.

We will denote $Y := (Y(x))_{x \in \mathbb{X}}$.

Trajectory, realization or sample path

Let Y be a RP. For a fixed $w \in \Omega$, a *trajectory or realization or sample path* of Y is the function $x \mapsto Y(x)(w)$.

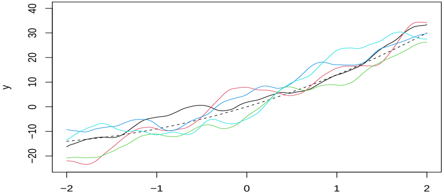


Figure: Five sample paths of a Gaussian process on $\mathbb{X} = [-2, 2]$

Second-order random process, mean, kernel.

Y is a *second-order RP* when all the r.v. $Y(x)$ belong to $L^2(\mathbb{P})$.

By Cauchy-Schwartz inequality, this implies that first moments (expectation) as well as second moments (covariances) are well-defined.

- The *mean* of Y is the function $x \in \mathbb{X} \mapsto \mathbb{E}(Y(x))$.
- The *covariance function* or *kernel* of Y is the function $(x, x') \in \mathbb{X} \times \mathbb{X} \mapsto \mathbb{C}\mathrm{ov}(Y(x), Y(x'))$.

Similarly, the *variance* of Y denotes the function $x \in \mathbb{X} \mapsto k(x, x) = \mathbb{V}\mathrm{ar}(Y(x))$.

Stationarity

- Y is *strongly stationary* if for all locations $x_1, \dots, x_n \in \mathbb{X}$, the law of $(Y(x_1 + h), \dots, Y(x_n + h))$ does not depend on h.

Stationarity

- Y is *strongly stationary* if for all locations $x_1, \dots, x_n \in \mathbb{X}$, the law of $(Y(x_1 + h), \dots, Y(x_n + h))$ does not depend on h.
- Y is *weakly stationary* if for all locations $x_1, \dots, x_n \in \mathbb{X}$, the first two moments of the law of $(Y(x_1 + h), \dots, Y(x_n + h))$ do not depend on h. Equivalently:

$$\mathbb{E}(Y(x)) = m, \qquad k(x, x') = c(x - x')$$

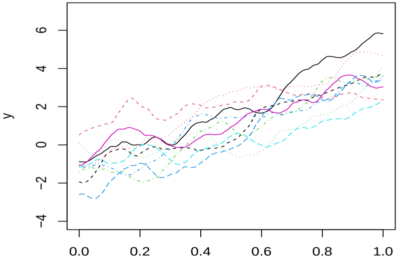
with $m = \mathbb{E}(Y(x_0))$ (for some $x_0 \in \mathbb{X}$) and $c(h) = k(x, x - h)$.

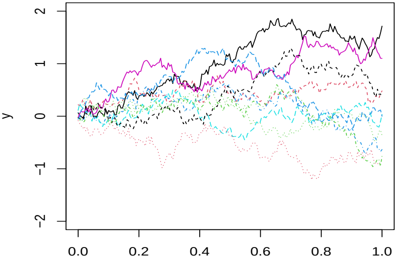
Stationarity

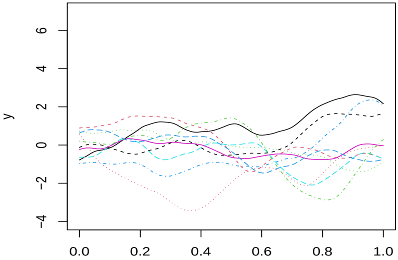
- Y is *strongly stationary* if for all locations $x_1, \ldots, x_n \in \mathbb{X}$, the law of $(Y(x_1 + h), \ldots, Y(x_n + h))$ does not depend on h.
- Y is *weakly stationary* if for all locations $x_1, \ldots, x_n \in \mathbb{X}$, the first two moments of the law of $(Y(x_1 + h), \ldots, Y(x_n + h))$ do not depend on h. Equivalently:
$$\mathbb{E}(Y(x)) = m, \qquad k(x, x') = c(x - x')$$
with $m = \mathbb{E}(Y(x_0))$ (for some $x_0 \in \mathbb{X}$) and $c(h) = k(x, x - h)$.

Obviously, strong stationarity implies weak stationarity.

Only one graph corresponds to a stationary process. Which why?
For the others, why the corresponding random process is non stationary?







Gaussian processes

A random process Y defined on $\mathbb{X}$ is a *Gaussian process* (GP) if for all locations $x_1, \ldots, x_n \in \mathbb{X}$, the random vector $(Y(x_1), \ldots, Y(x_n))$ is a Gaussian vector.

The law of such random vectors is fully characterized by the mean m and the kernel k of Y. We will denote $Y \sim GP(m, k)$.

Some properties of GPs inherited from Gaussian vectors

For GPs, properties of Gaussian vectors directly imply:

- **strong stationarity $\Leftrightarrow$ weak stationarity.**
  Thus, we simply speak of stationary GP.

Some properties of GPs inherited from Gaussian vectors

For GPs, properties of Gaussian vectors directly imply:

- **strong stationarity ⇔ weak stationarity.**
  Thus, we simply speak of stationary GP.
- **independence ⇔ non correlation.**
  If Y is a GP, $Y(x)$ and $Y(x')$ are independent iff $\mathbb{C}\mathrm{ov}(Y(x), Y(x')) = 0$.

Some properties of GPs inherited from Gaussian vectors

For GPs, properties of Gaussian vectors directly imply:

- **strong stationarity ⇔ weak stationarity.**
  Thus, we simply speak of stationary GP.
- **independence ⇔ non correlation.**
  If Y is a GP, $Y(x)$ and $Y(x')$ are independent iff $\mathbb{C}\mathrm{ov}(Y(x), Y(x')) = 0$.
- **a GP is stable by linear mapping.**
  Formally, if Y is a GP, and L is a linear mapping operating on the sample paths of Y, then LY is a GP.

Some properties of GPs inherited from Gaussian vectors

For GPs, properties of Gaussian vectors directly imply:

- **strong stationarity ⇔ weak stationarity.**
  Thus, we simply speak of stationary GP.
- **independence ⇔ non correlation.**
  If Y is a GP, $Y(x)$ and $Y(x')$ are independent iff $\mathbb{C}\mathrm{ov}(Y(x), Y(x')) = 0$.
- **a GP is stable by linear mapping.**
  Formally, if Y is a GP, and L is a linear mapping operating on the sample paths of Y, then LY is a GP.
- **a GP Y conditioned on $Y(x_i) = y_i, (i = 1, \dots, n)$ is still a GP.**
  This is the basis of Gaussian process regression, developed in Chapter 4. By stability of GPs under linearity, this property is true for linear equality constraints (and not only interpolation ones).

Gaussian processes and linear operations

If $Y \sim GP(0, k)$ and L is a linear function acting on the sample paths. of Y, then $LY \sim GP(0, k_L)$ with $k_L(s, t) = L_s L_t k(s, t)$. Here L_s (resp. L_t) means that we apply L on the function $s \mapsto k(s, t)$ (resp. $t \mapsto k(s, t)$).

Gaussian processes and linear operations

If $Y \sim GP(0,k)$ and L is a linear function acting on the sample paths. of Y, then $LY \sim GP(0,k_L)$ with $k_L(s,t) = L_s L_t k(s,t)$. Here L_s (resp. L_t) means that we apply L on the function $s \mapsto k(s,t)$ (resp. $t \mapsto k(s,t)$).

The expression of k_L comes, formally, from the bilinearity of covariance:

$$\mathbb{C}\mathrm{ov}(LY(s), LY(t)) = L_t(\mathbb{C}\mathrm{ov}(LY(s), Y(t)) = L_s L_t \mathbb{C}\mathrm{ov}(Y(s), Y(t))$$

Gaussian processes and linear operations

If $Y \sim GP(0, k)$ and L is a linear function acting on the sample paths. of Y, then $LY \sim GP(0, k_L)$ with $k_L(s,t) = L_s L_t k(s,t)$. Here L_s (resp. L_t) means that we apply L on the function $s \mapsto k(s,t)$ (resp. $t \mapsto k(s,t)$).

The expression of k_L comes, formally, from the bilinearity of covariance:

$$\mathbb{C}\mathrm{ov}(LY(s), LY(t)) = L_t(\mathbb{C}\mathrm{ov}(LY(s), Y(t)) = L_s L_t \mathbb{C}\mathrm{ov}(Y(s), Y(t))$$

Example if L is the derivative. If Y is a 1D centered GP with kernel k (smooth enough), then when it exists, $(Y'(x))_{x \in \mathbb{X}}$ is a centered GP, with kernel:

$$k_{Y'}(s,t) = \mathbb{C}\mathrm{ov}\left(\frac{\partial Y(s)}{\partial s}, \frac{\partial Y(t)}{\partial t}\right) = \frac{\partial}{\partial t}\mathbb{C}\mathrm{ov}\left(\frac{\partial Y(s)}{\partial s}, Y(t)\right) = \frac{\partial}{\partial s}\frac{\partial}{\partial t}\mathbb{C}\mathrm{ov}(Y(s), Y(t)) = \frac{\partial^2 k}{\partial s \partial t}(s,t)$$

Simulation of a GP

A simulation of $Y \sim GP(m, k)$ is possible on a set of discrete locations $X = \{x_1, \dots, x_n\}$.

Simulation of a GP

A simulation of $Y \sim GP(m, k)$ is possible on a set of discrete locations $X = \{x_1, \dots, x_n\}$. Indeed, then the law of $(Y(x_1), \dots, Y(x_n))^\top$ is $\mathcal{N}(m(X), k(X, X))$ where

- $m(X)$ is the vector of size n whose component i is equal to $m(x_i)$
- $k(X, X)$ is the matrix of size n whose coefficient (i, j) is equal to $k(x_i, x_j)$

Obtaining a realization of Y at X, it is thus equivalent to simulating from $\mathcal{N}(m(X), k(X, X))$.

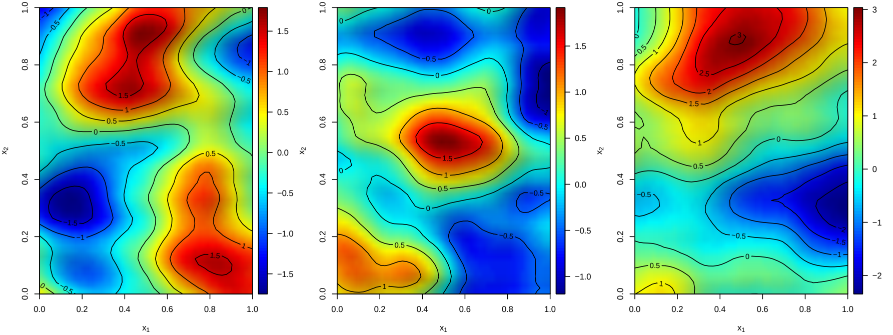


Figure: Simulations of a Gaussian process on $\mathbb{X} = [0,1]^2$ (obtained by simulating a Gaussian vector on a fine grid of $\mathbb{X}$), with kernel $k(x, x'; \ell) = k_1(x_1, x_1'; 2\ell) k_2(x_2, x_2'; \ell)$, where k_1 is a one-dimensional Matérn 5/2 kernel (see Chap. 3) and ℓ is a parameter.

Reminder on Gaussian vectors

$X := (X_1, \dots, X_d)^\top$ is a Gaussian vector iff it is the affine transformation of independent standard Normal random variables: $\exists \mu \in \mathbb{R}^d, A \in \mathbb{R}^{d\times m}$ and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_m)^\top$ where $\varepsilon_1, \dots, \varepsilon_m$ are i.i.d. $\mathcal{N}(0, 1)$, such that

$$X = \mu + A\varepsilon$$

Reminder on Gaussian vectors

$X := (X_1, \dots, X_d)^\top$ is a Gaussian vector iff it is the affine transformation of independent standard Normal random variables: $\exists \mu \in \mathbb{R}^d, A \in \mathbb{R}^{d \times m}$ and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_m)^\top$ where $\varepsilon_1, \dots, \varepsilon_m$ are i.i.d. $\mathcal{N}(0,1)$, such that

$$X = \mu + A\varepsilon$$

The mean of X is equal to μ, and its covariance matrix is

$$\mathbb{C}\mathrm{ov}(X) := \mathbb{E}\left[(X-\mu)(X-\mu)^\top\right] = AA^\top$$

Reminder on Gaussian vectors

$X := (X_1, \dots, X_d)^\top$ is a Gaussian vector iff it is the affine transformation of independent standard Normal random variables: $\exists \mu \in \mathbb{R}^d, A \in \mathbb{R}^{d \times m}$ and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_m)^\top$ where $\varepsilon_1, \dots, \varepsilon_m$ are i.i.d. $\mathcal{N}(0, 1)$, such that

$$X = \mu + A\varepsilon$$

The mean of X is equal to μ, and its covariance matrix is

$$\mathbb{C}\mathrm{ov}(X) := \mathbb{E}\left[(X - \mu)(X - \mu)^\top\right] = AA^\top$$

If $\Gamma := \mathbb{C}\mathrm{ov}(X)$ is invertible, X is called non degenerated.
We denote $X \sim \mathcal{N}(\mu, \Gamma)$.

Density function of the multivariate normal distribution

If $X \sim \mathcal{N}(\mu, \Gamma)$ is a non degenerated Gaussian vector in $\mathbb{R}^d$, then X admits the density function

$$f_X(x) = \frac{1}{(2\pi)^{d/2}|\Gamma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Gamma^{-1}(x-\mu)\right)$$

where $|\Gamma| = \det(\Gamma)$.

Density function of the multivariate normal distribution

If $X \sim \mathcal{N}(\mu, \Gamma)$ is a non degenerated Gaussian vector in $\mathbb{R}^d$, then X admits the density function

$$f_X(x) = \frac{1}{(2\pi)^{d/2}|\Gamma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Gamma^{-1}(x-\mu)\right)$$

where $|\Gamma| = \det(\Gamma)$.

This comes directly from the definition, using the theorem of change of variables.

The level sets of the density function (the sets of $x \in \mathbb{R}^d$ such that $f_X(x) = y$, for a given y) are ellipsoids centered at μ, whose axis are given by the eigenvectors of Γ.

The linear combination property

$X := (X_1, \dots, X_d)^\top$ is a Gaussian vector iff all linear combination of its components follow a (one-dimensional) Normal distribution:

$$\forall t_1, \dots, t_d \in \mathbb{R}, \quad t_1 X_1 + \cdots + t_d X_d \quad \text{follows a Normal distribution}$$

The linear combination property

$X := (X_1, \ldots, X_d)^\top$ is a Gaussian vector iff all linear combination of its components follow a (one-dimensional) Normal distribution:

$$\forall t_1, \ldots, t_d \in \mathbb{R}, \quad t_1 X_1 + \cdots + t_d X_d \quad \text{follows a Normal distribution}$$

This can be proved with the characteristic function of a probability measure.

This is a practical way to show that X is a Gaussian vector. Mind that it is *not sufficient* that $X_1, \ldots, X_d$ are normally distributed.

Stability by linear mapping

A linear mapping of a Gaussian vector is a Gaussian vector. More precisely, if $X \sim \mathcal{N}(\mu, \Gamma)$ is Gaussian vector on $\mathbb{R}^d$, and $L : \mathbb{R}^d \to \mathbb{R}^{d'}$ is a $d \times d'$ matrix, then LX is a Gaussian vector on $\mathbb{R}^{d'}$ with $LX \sim \mathcal{N}(L\mu, L\Gamma L^\top)$

Stability by linear mapping

A linear mapping of a Gaussian vector is a Gaussian vector. More precisely, if $X \sim \mathcal{N}(\mu, \Gamma)$ is Gaussian vector on $\mathbb{R}^d$, and $L : \mathbb{R}^d \to \mathbb{R}^{d'}$ is a $d \times d'$ matrix, then LX is a Gaussian vector on $\mathbb{R}^{d'}$ with $LX \sim \mathcal{N}(L\mu, L\Gamma L^\top)$

The result is obvious from the definition: if $X = \mu + A\varepsilon$, then $LX = L\mu + LA\varepsilon$.

Non-correlation and independence

If $X = (X_1, \dots, X_d)$ is a Gaussian vector, then for all (i, j), the random variables X_i and X_j are independent if and only if $\mathbb{C}\mathrm{ov}(X_i, X_j) = 0$.

Non-correlation and independence

If $X = (X_1, \dots, X_d)$ is a Gaussian vector, then for all (i, j), the random variables X_i and X_j are independent if and only if $\mathbb{C}\mathrm{ov}(X_i, X_j) = 0$.

This is because the probability distribution of X only depends on the mean and the covariances of its components.

The property is FALSE if X is not a Gaussian vector (even if $X_1, \dots, X_n$ are normally distributed!).

Linear and non-linear regression

Let $Y, X_1, \dots, X_d$ be random variables in $L^2(\mathbb{P})$, and $X = (X_1, \dots, X_d)$. Define:

- $\mathbb{E}(Y|X_1, \dots, X_d)$, **the *non-linear regression* of Y on $X_1, \dots X_d$**, as the best approximation of Y by functions of $X_1, \dots, X_d$ in the L^2 sense.

  It is the orthogonal projection of Y onto $L^{\mathbf{2}}(X_{\mathbf{1}}, \dots, X_d)$, the Hilbert space of square integrable random variables: $\mathbb{E}(Y|X_{\mathbf{1}}, \dots, X_d) = h(X)$ where $h(X)$ is such that $\mathbb{E}([Y - h(X)]^{\mathbf{2}})$ is minimal.

Linear and non-linear regression

Let $Y, X_1, \dots, X_d$ be random variables in $L^2(\mathbb{P})$, and $X = (X_1, \dots, X_d)$. Define:

- $\mathbb{E}(Y|X_1, \dots, X_d)$, **the *non-linear regression* of Y on $X_1, \dots X_d$**, as the best approximation of Y by functions of $X_1, \dots, X_d$ in the L^2 sense.

  It is the orthogonal projection of Y onto $L^2(X_1, \dots, X_d)$, the Hilbert space of square integrable random variables: $\mathbb{E}(Y|X_1, \dots, X_d) = h(X)$ where $h(X)$ is such that $\mathbb{E}([Y - h(X)]^2)$ is minimal.

- $\mathbb{E}_L(Y|X_1, \dots, X_d)$, **the *linear regression* of Y on $X_1, \dots X_d$**, as the best approximation of Y by *linear combinations* of $1, X_1, \dots, X_d$ in the L^2 sense.

  It is the orthogonal projection of Y onto the vector space spanned by $1, X_1, \dots, X_d$:
  $\mathbb{E}_L(Y|X_1, \dots, X_d) = \beta_0 + \beta^\top X$ where β_0, β are such that $\mathbb{E}([Y - (\beta_0 + \beta^\top X)]^2)$ is minimal.

Linear and non-linear regression

Let $Y, X_1, \dots, X_d$ be random variables in $L^2(\mathbb{P})$, and $X = (X_1, \dots, X_d)$. Define:

- **$\mathbb{E}(Y|X_1, \dots, X_d)$, the *non-linear regression* of Y on $X_1, \dots X_d$**, as the best approximation of Y by functions of $X_1, \dots, X_d$ in the L^2 sense.

  It is the orthogonal projection of Y onto $L^2(X_1, \dots, X_d)$, the Hilbert space of square integrable random variables: $\mathbb{E}(Y|X_1, \dots, X_d) = h(X)$ where $h(X)$ is such that $\mathbb{E}([Y - h(X)]^2)$ is minimal.

- **$\mathbb{E}_L(Y|X_1, \dots, X_d)$, the *linear regression* of Y on $X_1, \dots X_d$**, as the best approximation of Y by *linear combinations* of $1, X_1, \dots, X_d$ in the L^2 sense.

  It is the orthogonal projection of Y onto the vector space spanned by $1, X_1, \dots, X_d$:
  $\mathbb{E}_L(Y|X_1, \dots, X_d) = \beta_0 + \beta^\top X$ where β_0, β are such that $\mathbb{E}([Y - (\beta_0 + \beta^\top X)]^2)$ is minimal.

If $(Y, X_1, \dots, X_d)$ is a Gaussian vector, then

$$\mathbb{E}(Y|X_1, \dots, X_d) = \mathbb{E}_L(Y|X_1, \dots, X_d)$$

Conditioning of Gaussian vectors

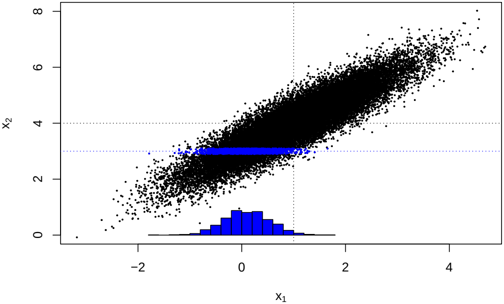


Figure: Illustration of the conditioning of Gaussian vectors on a simulated sample.

Let $U = (V, W) \sim \mathcal{N}(\mu, \Gamma)$ be a Gaussian vector on $\mathbb{R}^d$, where V, W are subvectors of dimension d_V, d_W respectively. Write $\mu = (\mu_V, \mu_W)^\top$ with $\mu_V = \mathbb{E}(V), \mu_W = \mathbb{E}(W)$ and

$$\Gamma = \begin{bmatrix} \Gamma_V & \Gamma_{V,W} \\ \Gamma_{W,V} & \Gamma_W \end{bmatrix}$$

where $\Gamma_V = \mathbb{C}\mathrm{ov}(V), \Gamma_W = \mathbb{C}\mathrm{ov}(W)$, and $\Gamma_{V,W} = \mathbb{C}\mathrm{ov}(V, W) := \mathbb{E}[(V - \mu_V)(W - \mu_W)^\top]$ (similar definition for $\Gamma_{W,V}$).

Then $V|W = w$ is a Gaussian vector on $\mathbb{R}^{d_W}$ with mean and covariance matrix

$$\begin{aligned} \mathbb{E}(V|W = w) &= \mu_V + \Gamma_{V,W}\Gamma_W^{-1}(w - \mu_W) && \text{linear with respect to } w \\ \mathbb{C}\mathrm{ov}(V|W = w) &= \Gamma_V - \Gamma_{V,W}\Gamma_W^{-1}\Gamma_{W,V} && \text{does not depend on } w \end{aligned}$$

Exercise: Proof for centered vectors, in the case $d_V = 1$

Here V, W are centered, and V is a random variable. $W = (W_1, \ldots, W_d)^\top$.

- Let us write $\mathbb{E}_L(V|W) = \beta_0 + \beta_1^\top W$. By using the definition of the orthogonal projection in $L^2(\mathbb{P})$, show that $\mathbb{E}_L(V|W) = \Gamma_{V,W}\Gamma_W^{-1} w$.
- Define $\varepsilon = V - \mathbb{E}_L(V|W)$. Show that $\mathbb{V}\mathrm{ar}(\varepsilon) = \Gamma_V - \Gamma_{V,W}\Gamma_W^{-1}\Gamma_{W,V}$.
- Let us show that linear / non-linear regression of V on W coincide.
  - Prove that (ε, W) is a Gaussian vector.
  - Deduce that ε and W are independent.
  - Deduce that $\mathbb{E}(V|W) = \mathbb{E}_L(V|W)$
- Prove that $V|W = w$ and $\varepsilon + \beta_1 w$ have the same law.
- Conclude.