Lecture notes
Gaussian processes and sensitivity analysis for computer experiments

François Bachoc
University Paul Sabatier

January 27, 2020

Contents

1 Gaussian process metamodels **2**
1.1 Reminders on Gaussian vectors 2
1.2 Gaussian processes: properties 3
1.3 Gaussian processes: prediction 6
1.4 Gaussian processes: covariance function estimation 11

2 Sensitivity analysis **12**
2.1 Context 12
2.2 ANOVA decomposition 13
2.3 Sobol sensitivity indices 18

Introduction

In these lecture notes we are interested in a function

$$f : [0,1]^d \to \mathbb{R}.$$

This function is unknown (we do not know how $f(x)$ is computed for $x \in [0,1]^d$). We can only observed some values of f of the form

$$(x_1, f(x_1))$$
$$\dots$$
$$(x_n, f(x_n)).$$

Our aim is, essentially, to use these values (our statistical data) to infer the function f, as depicted in Figure 1.

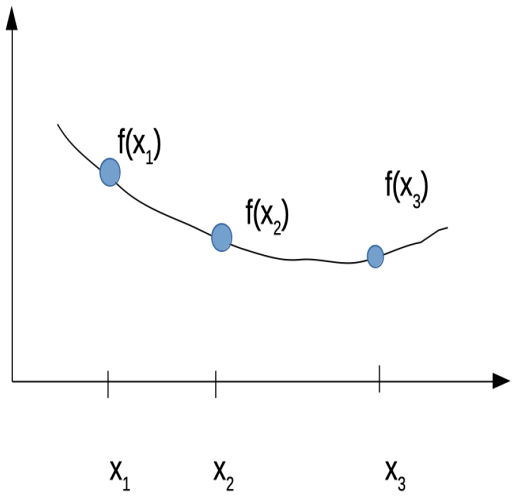

Figure 1: Unknown function and observation points.

Application fields: When f represent a computer model or computer experiment, choosing $x \in [0,1]^d$ correspond to set the values of d simulation parameters. After a (long and costly) computation, the computer model provides the simulation result $f(x)$. We do not know how f works, because f involve, for instance, complex physical models and numerical schemes. Hence there is a limited number of results

$$f(x_1)$$
$$\dots$$
$$f(x_n).$$

Examples: Simulation of aeronautic component, car industry, nuclear engineering. The components of x can correspond to geometry parameters, material characteristics, physical concentrations,... This framework is studied in French companies or institutions such as EDF, CEA, ONERA, AIRBUS,...

These lecture notes will address the two following questions:

- How can we learn f from $(x_1, f(x_1)), \dots, (x_n, f(x_n))$?
- What is the impact of each of the d components of f on the value of $f(x)$?

To address these two questions, the lecture notes are organized into the two main sections:

1. Gaussian process metamodels,
2. Sensitivity analysis.

1 Gaussian process metamodels

1.1 Reminders on Gaussian vectors

Proposition 1 *Let V be a random vector on $\mathbb{R}^n$. The three following assertions are equivalent.*

i) *Any linear combination of V follow a Gaussian distribution. That is, for any fixed $n \times 1$ vector a, there exists $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ such that*

$$a^\top V = \sum_{i=1}^{n} a_i V_i \sim \mathcal{N}(0, \sigma^2).$$

ii) *There exist* $m \in \mathbb{R}^n$ *and a* $n \times n$ *symmetric non-negative definite (SNND) matrix* Σ *such that for all* $n \times 1$ *vector* u*,*

$$\phi_V(u) = \exp\left(\imath u^\top m - \frac{1}{2} u^\top \Sigma u\right),$$

where ϕ_V *is the characteristic function of* V*, with* $\phi_V(u) = \mathbb{E}(e^{iu^\top V})$*.*

iii) *There exist a* $n \times 1$ *vector* m'*, a* $n \times r$ *matrix* K*, with* $r \leq n$ *and a* $r \times 1$ *vector* W*, with independent components with distribution* $\mathcal{N}(0,1)$*, such that*

$$V = m' + KW.$$

Furthermore, with m, Σ *like in* **ii)** *and* m', K *like in* **iii)***, we have*

$$m = m' = \mathbb{E}(V) \quad \textit{(mean vector)}$$

and

$$\Sigma = KK^\top = \mathrm{cov}(V) \quad \textit{(covariance matrix).}$$

We say that V is a Gaussian vector when it satisfies the three previous assertions (we just need to check that one of the three is satisfied to show that V is a Gaussian vector, since the three are equivalent).

1.2 Gaussian processes: properties

Definition 2 (Gaussian process) *Let* $(\Omega, \mathcal{A}, P)$ *be a probability space and*

$$\begin{aligned} Z : (\Omega, \mathcal{A}, P) \times [0,1]^d &\to \mathbb{R} \\ (\omega, x) &\to Z(\omega, x). \end{aligned}$$

We say that Z *is a Gaussian process on* $[0,1]^d$ *when for all* $n \in \mathbb{N}$*, for all* $x_1, \ldots, x_n \in [0,1]^d$*, the function* $\omega \to (Z(\omega, x_1), \ldots, Z(\omega, x_n))$ *is a Gaussian (random) vector.*

For all ω, $x \to Z(\omega, x)$ is a function from $[0,1]^d$ to $\mathbb{R}$. Hence, a Gaussian process is a random function. We call $x \to Z(\omega, x)$ a trajectory, or sample path, of Z (see Figure 1.2).

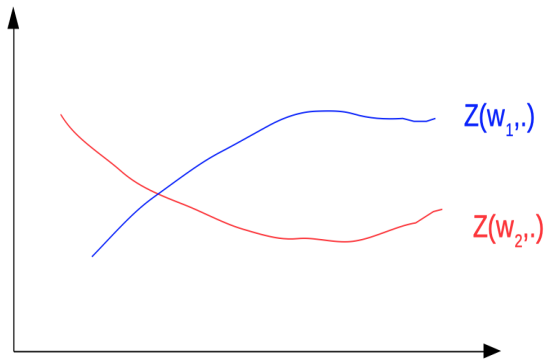

Figure 2: Sample paths of Gaussian processes.

To come back to the topic of the computer model $f : [0,1]^d \to \mathbb{R}$, we say that $f(x) = Z(\omega^\star, x)$ for some probability event $\omega^\star$. That is, the computer model f is a realization of a random function Z. Hence, we are applying the Bayesian framework on the function f. This provides many benefits, as we shall see below. In the sequel, we will often write $Z(x)$ instead of $Z(\omega, x)$.

Example:

Let X_1, X_2 be two independent random variables with distribution $\mathcal{N}(0,1)$. Let $Z(x) = X_1 + cos(x)X_2$ for $x \in [0,1]$. Then Z is a Gaussian process. Indeed, for all $n \in \mathbb{N}$ and $x_1, \dots, x_n$,

$$\begin{pmatrix} Z(x_1) \\ \vdots \\ Z(x_n) \end{pmatrix} = \begin{pmatrix} X_1 + \cos(x_1)X_2 \\ \vdots \\ X_1 + \cos(x_n)X_2 \end{pmatrix} = \begin{pmatrix} 1 & \cos(x_1) \\ \vdots & \\ 1 & \cos(x_n) \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

is a Gaussian vector from **iii)** of Proposition 1.

Definition 3 (mean function) *Let Z be a Gaussian process on $[0,1]^d$. The function*

$$\begin{aligned} m : [0,1]^d &\to \mathbb{R} \\ x &\to \mathbb{E}(Z(x)) \end{aligned}$$

is called the mean function of Z.

Definition 4 (covariance function) *Let Z be a Gaussian process on $[0,1]^d$. The function*

$$\begin{aligned} K : [0,1]^d \times [0,1]^d &\to \mathbb{R} \\ (x,y) &\to \mathrm{cov}(Z(x), Z(y)) \end{aligned}$$

is called the covariance function of Z.

If the function K only depends on $x - y$, that is $x - y = x' - y' \implies K(x,y) = K(x',y')$, then we say that K is stationary. In this case, in abuse of notation, we write $K(x,y) = K(x-y)$.

Proposition 5 *Let Z be a Gaussian process on $[0,1]^d$, with constant mean function and stationary covariance function. Then Z is a stationary process. That is, for all $n \in \mathbb{N}$, for all $x_1, \dots, x_n \in [0,1]^d$, for all $\delta \in \mathbb{R}^d$ such that $x_1 + \delta, \dots, x_n + \delta \in [0,1]^d$, we distribution of*

$$\begin{pmatrix} Z(x_1) \\ \vdots \\ Z(x_n) \end{pmatrix}$$

is the same as that of

$$\begin{pmatrix} Z(x_1 + \delta) \\ \vdots \\ Z(x_n + \delta) \end{pmatrix}.$$

Stationarity means that the distribution is translation invariant, see Figure 3.

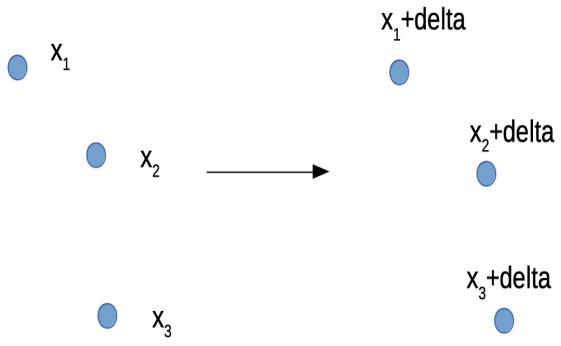

Figure 3: Translation of observation points in the definition of stationarity.

Proof of Proposition 5 Let $n, x_1, \ldots, x_n, \delta$ be as in the statement of the proposition. Since Z is a Gaussian process, the two random vectors in the statement of the proposition are Gaussian vector. Hence, their distributions are characterized by their mean vectors and covariance matrices. Let us show that they are identical between the two Gaussian vectors.

We have, since the mean function is constant, for $i = 1, \ldots, n$,

$$\mathbb{E}(Z(x_i)) = m(x_i) = m(x_i + \delta) = \mathbb{E}(Z(x_i + \delta)).$$

Hence the mean vectors are identical. We have, for $i, j = 1, \ldots, n$,

$$x_i - x_j = (x_i + \delta) - (x_j + \delta).$$

hence, since the covariance function is stationary,

$$K(x_i, x_j) = K(x_i + \delta, x_j + \delta).$$

Hence we have

$$\mathrm{Cov}(Z(x_i), Z(x_j)) = \mathrm{Cov}(Z(x_i + \delta), Z(x_j + \delta)).$$

Hence the two covariance matrices are identical. □

Definition 6 (mean square continuity) *Let Z be a Gaussian process on $[0,1]^d$. We say that Z is mean square continuous if, for all $x_0 \in [0,1]^d$,*

$$\lim_{x \to x_0} \mathbb{E}\left((Z(x) - Z(x_0))^2 \right) = 0.$$

We remark that the definition naturally extends the notion of continuity for functions. The difference between $Z(x)$ and $Z(x_0)$ (measured by the expectation of the square difference) goes to zero as x goes to x_0.

Proposition 7 *Let Z be a Gaussian process on $[0,1]^d$ with mean function m and with covariance function K. Then Z is mean square continuous if and only if m is continuous and, for all $x_0 \in [0,1]^d$, K is continuous at (x_0, x_0).*

Proof of Proposition 7 We have

$$\begin{aligned}
\mathbb{E}\left((Z(x) - Z(x_0))^2 \right) &= \left(\mathbb{E}\left(Z(x) - Z(x_0) \right) \right)^2 + \mathrm{Var}\left(Z(x) - Z(x_0) \right) \\
&= \left(m(x) - m(x_0) \right)^2 + \left(K(x,x) + K(x_0, x_0) - 2K(x, x_0) \right). \qquad (1)
\end{aligned}$$

If m is continuous and if K is continuous at (x_0, x_0), then the two summands in (1) go to zero as $x \to x_0$. This shows the implication "$\Longleftarrow$".

Let us now show the implication "$\Longrightarrow$". Under the assumption of mean square continuity, (1) goes to zero as $x \to x_0$, so the two summands (.) in (1) go to zero (they are both non-negative, the second one being a variance). Hence $(m(x) - m(x_0))^2$ goes to zero and so m is continuous. For $x, y \in [0,1]^d$, we have

$$\begin{aligned}
|K(x,y) - K(x_0, x_0)| &= |K(x,y) - K(x, x_0) + K(x, x_0) - K(x_0, x_0)| \\
&= |\mathrm{Cov}\left(Z(x), Z(y)\right) - \mathrm{Cov}\left(Z(x), Z(x_0)\right) + \mathrm{Cov}\left(Z(x), Z(x_0)\right) - \mathrm{Cov}\left(Z(x_0), Z(x_0)\right)| \\
&= |\mathrm{Cov}\left(Z(x), Z(y) - Z(x_0)\right) + \mathrm{Cov}\left(Z(x) - Z(x_0), Z(x_0)\right)| \\
&\leq |\mathrm{Cov}\left(Z(x), Z(y) - Z(x_0)\right)| + |\mathrm{Cov}\left(Z(x) - Z(x_0), Z(x_0)\right)| \\
&\leq \sqrt{\mathrm{Var}(Z(x))}\sqrt{\mathrm{Var}(Z(y) - Z(x_0))} + \sqrt{\mathrm{Var}(Z(x) - Z(x_0))}\sqrt{\mathrm{Var}(Z(x_0))},
\end{aligned}$$

using the Cauchy-Schwarz inequality. In the above display, the first square root is bounded as $x \to x_0$ because, from the triangle inequality, $|\sqrt{\mathrm{Var}(Z(x))} - \sqrt{\mathrm{Var}(Z(x_0))}| \leq \sqrt{\mathrm{Var}(Z(x) - Z(x_0))}$ that goes to zero by the mean square continuity assumption. The second square root goes to zero as $(x,y) \to (x_0, x_0)$ by the mean square continuity assumption. The third square root also goes to zero by the mean square continuity assumption. The fourth square root is fixed as $(x,y) \to (x_0, x_0)$. Hence, $|K(x,y) - K(x_0, x_0)|$ goes to zero as $(x,y) \to (x_0, x_0)$ and thus K is continuous at (x_0, x_0). □

Definition 8 (mean square differentiability) *A Gaussian process Z on $[0,1]^d$ is differentiable in quadratic mean if there exist d Gaussian processes (defined on the same probability space $(\Omega, \mathcal{A}, P)$),*

$$\frac{\partial Z}{\partial x_1}, \ldots, \frac{\partial Z}{\partial x_d}$$

such that for $k = 1, \ldots, d$, for all $x_0 \in [0,1]^d$, with e_k the k-th base column vector of $\mathbb{R}^d$,

$$e_k = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

where the 1 is at position k, we have

$$\lim_{h \to 0} \mathbb{E}\left(\left(\frac{Z(x_0 + he_k) - Z(x_0)}{h} - \frac{\partial Z}{\partial x_k}(x_0) \right)^2 \right) = 0.$$

Definition 9 (mean square differentiability of higher order) *The definition is by induction. A Gaussian process Z on $[0,1]^d$ is k times mean square differentiable if it is mean square differentiable and if the d Gaussian processes*

$$\frac{\partial Z}{\partial x_1}, \ldots, \frac{\partial Z}{\partial x_d}$$

are $k-1$ times mean square differentiable.

Proposition 10 *Let Z be a Gaussian process on $[0,1]^d$ with mean function m and with covariance function K. If m is k times continuously differentiable on $[0,1]^d$ and if K is $2k$ times continuously differentiable on $[0,1]^d \times [0,1]^d$, then Z is k times differentiable in quadratic mean.*

We do not provide the proof of Proposition 10 in these lecture notes.

1.3 Gaussian processes: prediction

Theorem 11 (Gaussian conditioning theorem (GCT)) *We consider a $(n_1 + n_2) \times 1$ Gaussian vector*

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$$

where Y_1 is of size $n_1 \times 1$ and Y_2 is of size $n_2 \times 1$. We write the mean vector of $(Y_1^\top, Y_2^\top)^\top$ as

$$\begin{pmatrix} m_1 \\ m_2 \end{pmatrix}$$

where m_1 is of size $n_1 \times 1$ and m_2 is of size $n_2 \times 1$. Finally, we write the covariance matrix of $(Y_1^\top, Y_2^\top)^\top$ as

$$\begin{pmatrix} \Sigma_1 & \Sigma_{1,2} \\ \Sigma_{1,2}^\top & \Sigma_2 \end{pmatrix}$$

where Σ_1 is of size $n_1 \times n_1$, $\Sigma_{1,2}$ is of size $n_1 \times n_2$ and Σ_2 is of size $n_2 \times n_2$. We assume that Σ_1 is invertible. Then, conditionally to $Y_1 = y_1$, the random vector Y_2 is a Gaussian vector with mean vector

$$\mathbb{E}\left(Y_2 | Y_1 = y_1\right) = m_2 + \Sigma_{1,2}^\top \Sigma_1^{-1}(y_1 - m_1)$$

and covariance matrix

$$\mathrm{Cov}\left(Y_2 | Y_1 = y_1\right) = \Sigma_2 - \Sigma_{1,2}^\top \Sigma_1^{-1} \Sigma_{1,2}.$$

Proof of Theorem 11

Let

$$\Lambda = \begin{pmatrix} A & D \\ B & C \end{pmatrix}$$

be such that

$$\begin{pmatrix} \Sigma_1 & \Sigma_{1,2} \\ \Sigma_{1,2}^\top & \Sigma_2 \end{pmatrix} = \Lambda\Lambda^\top.$$

Then,

$$\Lambda\Lambda^\top = \begin{pmatrix} A & D \\ B & C \end{pmatrix}\begin{pmatrix} A^\top & B^\top \\ D^\top & C^\top \end{pmatrix} = \begin{pmatrix} AA^\top + DD^\top & AB^\top + DC^\top \\ BA^\top + CD^\top & BB^\top + CC^\top \end{pmatrix} = \begin{pmatrix} \Sigma_1 & \Sigma_{1,2} \\ \Sigma_{1,2}^\top & \Sigma_2 \end{pmatrix}.$$

Then, with

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix},$$

$$m = \begin{pmatrix} m_1 \\ m_2 \end{pmatrix}$$

there exists $\epsilon \sim \mathcal{N}(0, I_{n_1+n_2})$ such that

$$Y = m + \Lambda\epsilon,$$

from Proposition 1. We let

$$G = \begin{pmatrix} A^\top \end{pmatrix}, \qquad H = \begin{pmatrix} B^\top \\ C^\top \end{pmatrix}.$$

Then,

$$Y = m + \begin{pmatrix} G^\top\epsilon \\ H^\top\epsilon \end{pmatrix}.$$

We let span(G) and be the linear space spanned by the columns of G. We define span(H) similarly from H. We remark that span(G) and span(H) are linear subspaces of $\mathbb{R}^{n_1+n_2}$.

We will decompose span(H) the as the sum of its projection on span(G) and of its projection on an orthogonal linear space. This enables to decompose Y_2 as the sum of a function of Y_1 and of something independent to Y_1. We let

$$P_G = G(G^\top G)^{-1}G^\top$$

be the orthogonal projection matrix onto span(G). We write

$$P_G^\perp = I_{n_1+n_2} - P_G$$

for the orthogonal projection matrix onto the orthogonal space to span(G). Then

$$\epsilon = P_G\epsilon + P_G^\perp\epsilon$$

and these two summands are independent because

$$\mathrm{Cov}(P_G\epsilon, P_G^\perp) = P_G\mathrm{Cov}(\epsilon,\epsilon)\left(P_G^\perp\right)^\top = P_G\left(P_G^\perp\right)^\top = P_GP_G^\perp = 0.$$

Then,

$$Y_1 = m_1 + G^\top\epsilon$$

and

$$\begin{aligned} Y_2 &= m_2 + H^\top\epsilon \\ &= m_2 + H^\top P_G\epsilon + H^\top P_G^\perp\epsilon \\ &= m_2 + H^\top G(G^\top G)^{-1}G^\top\epsilon + H^\top P_G^\perp\epsilon \\ &= m_2 + H^\top G(G^\top G)^{-1}(Y_1 - m_1) + H^\top P_G^\perp\epsilon, \end{aligned}$$

and the two above summands are independent as we have seen.

Hence, Y_2 can be written as

$$\text{deterministic function of } \; Y_1 \quad + \quad \text{something independent to } \; Y_1.$$

Hence

$$\mathcal{L}(Y_2|Y_1 = y_1) = \mathcal{N}\left(m_2 + H^\top G(G^\top G)^{-1}(Y_1 - m_1), H^\top P_G^\perp \left(P_G^\perp\right)^\top H\right).$$

Then,

$$\begin{aligned} H^\top G(G^\top G)^{-1} &= \begin{pmatrix} B & C \end{pmatrix} \begin{pmatrix} A^\top \\ D^\top \end{pmatrix} \left(\begin{pmatrix} A & D \end{pmatrix} \begin{pmatrix} A^\top \\ D^\top \end{pmatrix} \right)^{-1} \\ &= \left(BA^\top + CD^\top\right) \left(AA^\top + DD^\top\right)^{-1} \\ &= \Sigma_{1,2}^\top \Sigma_1^{-1}. \end{aligned}$$

Furthermore,

$$\begin{aligned} H^\top P_G^\perp \left(P_G^\perp\right)^\top H &= H^\top P_G^\perp H \\ &= H^\top \left(I_{n_1+n_2} - G(G^\top G)^{-1} G^\top\right) H \\ &= H^\top H - H^\top G(G^\top G)^{-1} G^\top H \\ &= \begin{pmatrix} B & C \end{pmatrix} \begin{pmatrix} B^\top \\ C^\top \end{pmatrix} - \begin{pmatrix} B & C \end{pmatrix} \begin{pmatrix} A^\top \\ D^\top \end{pmatrix} \left(\begin{pmatrix} A & D \end{pmatrix} \begin{pmatrix} A^\top \\ D^\top \end{pmatrix} \right)^{-1} \begin{pmatrix} A & D \end{pmatrix} \begin{pmatrix} B^\top \\ C^\top \end{pmatrix} \\ &= \left(BB^\top + CC^\top\right) - \left(BA^\top + CD^\top\right) \left(AA^\top + DD^\top\right)^{-1} \left(AB^\top + DC^\top\right) \\ &= \Sigma_2 - \Sigma_{1,2}^\top \Sigma_1^{-1} \Sigma_{1,2}. \end{aligned}$$

□

We will see three main applications of the Gaussian conditioning theorem.

(1) Prediction We consider a Gaussian process Y on $[0,1]^d$, with mean function m and covariance function K. We consider n observations of Y

$$\begin{aligned} Y(x_1) = y_1 \quad &(= f(x_1) \text{ for the computer model}), \\ &\vdots \\ Y(x_n) = y_n \quad &(= f(x_n) \text{ for the computer model}). \end{aligned}$$

For all $x \in [0,1]^d$, we want to predict $Y(x)$ (which realization is considered to be $f(x)$). We let R be the $n \times n$ matrix $(K(x_i, x_j))_{i,j=1,\dots,n}$. We write

$$Y^{(n)} = \begin{pmatrix} Y(x_1) \\ \vdots \\ Y(x_n) \end{pmatrix},$$

$$y^{(n)} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix},$$

and

$$m_y = \begin{pmatrix} m(x_1) \\ \vdots \\ m(x_n) \end{pmatrix}.$$

We consider the $n \times 1$ vector

$$r(x) = \begin{pmatrix} K(x, x_1) \\ \vdots \\ K(x, x_n) \end{pmatrix}.$$

Then,

$$\begin{pmatrix} Y^{(n)} \\ Y(x) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} m_y \\ m(x) \end{pmatrix}, \begin{pmatrix} R & r(x) \\ r^\top(x) & K(x,x) \end{pmatrix} \right). \tag{2}$$

Hence, from the GCT,

$$\mathbb{E}\left(Y(x)|Y^{(n)} = y^{(n)}\right) = m(x) + r(x)^\top R^{-1}(y^{(n)} - m_y).$$

we write $\hat{Y}(x) = \mathbb{E}\left(Y(x)|Y^{(n)} = y^{(n)}\right)$. This is a metamodel of f. Indeed, $\hat{Y}(x)$ approximates $f(x)$ and computing $\hat{Y}(x)$ has a negligible cost compared to performing an additional computer experiments in order to computed $f(x)$.

(2) Predictive variance From (2) and the GCT, we obtain

$$\mathrm{Var}\left(Y(x)|Y^{(n)} = y^{(n)}\right) = K(x,x) - r(x)^\top R^{-1} r(x).$$

we write $\hat{\sigma}^2(x) = \mathrm{Var}\left(Y(x)|Y^{(n)} = y^{(n)}\right)$. This is an error indicator of the metamodel because

$$\hat{\sigma}^2(x) = \mathbb{E}\left(\left(Y(x) - \hat{Y}(x)\right)^2\right).$$

This enables us to obtain confidence intervals. Since $\mathcal{L}(Y(x)|Y^{(n)} = y^{(n)}) = \mathcal{N}(\hat{Y}(x), \hat{\sigma}^2(x))$, the confidence interval

$$I(x) = \left[\hat{Y}(x) - 1.96\hat{\sigma}(x), \hat{Y}(x) + 1.96\hat{\sigma}(x)\right]$$

is a 95% confidence interval for $Y(x)$ (conditionally to $Y^{(n)} = y^{(n)}$):

$$P(Y(x) \in I(x)|Y^{(n)} = y^{(n)}) = 0.95.$$

Remark 12 *Assume that there exists $i \in \{1, \ldots, n\}$ such that $x = x_i$. Then*

$$\hat{Y}(x) = y_i$$
$$\hat{\sigma}^2(x) = 0,$$

which means that we know already that $Y(x_i) = y_i$.

Proof of Remark 12
Write

$$R = \begin{pmatrix} \star & r(x) & \star \end{pmatrix}$$

where the three submatrices have dimensions $n \times (i-1)$, $n \times 1$ and $n \times (n-i)$ from left to right. We know that $R^\top R^{-1} = RR^{-1} = I_n$ and thus

$$I_n = \begin{pmatrix} \star \\ r(x)^\top \\ \star \end{pmatrix} R^{-1} = \begin{pmatrix} \star \\ r(x)^\top R^{-1} \\ \star \end{pmatrix}.$$

Hence, with e_i the i-th base column vector of $\mathbb{R}^n$,

$$e_i = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

where the 1 is at position i, we have

$$e_i^\top = r(x)^\top R^{-1}.$$

Hence

$$\begin{aligned}
\hat{Y}(x) &= m(x) + r(x)^\top R^{-1}(y^{(n)} - m_y) \\
&= m(x) + e_i^\top (y^{(n)} - m_y) \\
&= m(x) + (y^{(n)} - m_y)_i \\
&= m(x) + y_i - m(x_i) \\
&= m(x_i) + y_i - m(x_i) \\
&= y_i.
\end{aligned}$$

Furthermore

$$\begin{aligned}
\hat{\sigma}^2(x) &= K(x,x) - r(x)^\top R^{-1} r(x) \\
&= K(x,x) - e_i^\top r(x) \\
&= K(x,x) - (r(x))_i \\
&= K(x,x) - K(x,x_i) \\
&= K(x_i,x_i) - K(x_i,x_i) \\
&= 0.
\end{aligned}$$

□

(3) Conditional covariance function

We still observe

$$Y^{(n)} = \begin{pmatrix} Y(x_1) \\ \vdots \\ Y(x_n) \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = y^{(n)}$$

and we are interested in the joint conditional distribution of

$$\begin{pmatrix} Y(u) \\ Y(v) \end{pmatrix}$$

for $u, v \in [0,1]^d$. From the GCT, this distribution is Gaussian. We also already know the two conditional means and variances. We hence want to find the conditional covariance. Keeping the notation $Y^{(n)}$, $y^{(n)}$, $r(x)$ and R we obtain

$$\begin{pmatrix} Y^{(n)} \\ Y(u) \\ Y(v) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} m_y \\ m(u) \\ m(v) \end{pmatrix}, \begin{pmatrix} R & r(u) & r(v) \\ r(u)^\top & K(u,u) & K(u,v) \\ r(v)^\top & K(v,u) & K(v,v) \end{pmatrix} \right). \tag{3}$$

Hence from the GCT, we obtain

$$\begin{aligned}
\mathrm{Cov}\left(\begin{pmatrix} Y(u) \\ Y(v) \end{pmatrix} \middle| Y^{(n)} = y^{(n)} \right) &= \begin{pmatrix} K(u,u) & K(u,v) \\ K(v,u) & K(v,v) \end{pmatrix} - \begin{pmatrix} r(u)^\top \\ r(v)^\top \end{pmatrix} R^{-1} \begin{pmatrix} r(u) & r(v) \end{pmatrix} \\
&= \begin{pmatrix} K(u,u) & K(u,v) \\ K(v,u) & K(v,v) \end{pmatrix} - \begin{pmatrix} r(u)^\top R^{-1} r(u) & r(u)^\top R^{-1} r(v) \\ r(v)^\top R^{-1} r(u) & r(v)^\top R^{-1} r(v) \end{pmatrix} \\
&= \begin{pmatrix} K(u,u) - r(u)^\top R^{-1} r(u) & K(u,v) - r(u)^\top R^{-1} r(v) \\ K(v,u) - r(v)^\top R^{-1} r(u) & K(v,v) - r(v)^\top R^{-1} r(v). \end{pmatrix}
\end{aligned}$$

We recognize

$$\mathrm{Var}(Y(u)|Y^{(n)} = y^{(n)}) = K(u,u) - r(u)^\top R^{-1} r(u)$$

and

$$\mathrm{Var}(Y(v)|Y^{(n)} = y^{(n)}) = K(v,v) - r(v)^\top R^{-1} r(v).$$

The new formula is

$$\mathrm{Cov}(Y(u), Y(v)|Y^{(n)} = y^{(n)}) = K(u,v) - r(u)^\top R^{-1} r(v),$$

which provides the conditional covariance function of Y. We hence obtain the following result.

Proposition 13 *Conditionally to $Y(x_1) = y_1, \ldots, Y(x_n) = y_n$, the process $Y : [0,1]^d \to \mathbb{R}$ is a Gaussian process, with mean function*

$$x \to \mathbb{E}(Y(x)|Y^{(n)} = y^{(n)}) = \hat{Y}(x) = m(x) + r(x)^\top R^{-1}(y^{(n)} - m_y)$$

and with covariance function

$$(u,v) \to \mathrm{cov}(Y(u), Y(v)|Y^{(n)} = y^{(n)}) = K(u,v|Y^{(n)} = y^{(n)}) = K(u,v) - r(u)^\top R^{-1} r(v).$$

1.4 Gaussian processes: covariance function estimation

In order to apply the GCT, we need to know the mean and covariance functions of Y. In practice (computer experiments and metamodels) we do not know these functions. In these lecture notes (and in many practical uses of Gaussian processes), we use the "plug-in" approach.

1. We estimate m and K by $\hat{m}$ and $\hat{K}$,
2. We apply the GCT, by replacing m and K by $\hat{m}$ and $\hat{K}$ in the equations.

In these lecture notes, we consider a parametric estimation method. Furthermore, for simplicity of exposition, we assume that the mean function is zero and focus on the covariance function.

Definition 14 *A parametric set of covariance functions is a set of the form*

$$\{K_\theta; \theta \in \Theta\}$$

where Θ s a subset of $\mathbb{R}^p$, for $p \in \mathbb{N}$ and where, for all $\theta \in \Theta$, $K_\theta : [0,1]^d \times [0,1]^d$ is a covariance function.

Example: The Matérn 3/2 covariance function for $d = 1$ is given by $\Theta = [0,\infty) \times [0,\infty)$, $\theta = (\sigma^2, \ell)$ and

$$K_{\sigma^2,\ell}(x,y) = \sigma^2 \left(1 + \sqrt{6}\frac{|x-y|}{\ell}\right) e^{-\sqrt{6}\frac{|x-y|}{\ell}}.$$

Definition 15 *An estimator of θ is a function*

$$\hat{\theta} : \bigcup_{n \in \mathbb{N}} \left(\left([0,1]^d \times \mathbb{R} \right)^n \right) \to \Theta.$$

The input space of $\hat{\theta}$ is the space of all possible data sets of the form

$$(x_1, Y(x_1)), \ldots, (x_n, Y(x_n)).$$

The quantity $\hat{\theta}(x_1, y_1, \ldots, x_n, y_n)$ is the estimate of θ after $Y(x_1) = y_1, \ldots, Y(x_n) = y_n$ is observed.

Maximum likelihood: We assume that we know the mean function of Y and that this function is zero. Then, if θ is the true covariance parameter (if K_θ is the covariance function of Y), then $Y^{(n)} = (Y(x_1), \ldots, Y(x_n))$ is a Gaussian vector with mean vector zero and with covariance matrix $R_\theta = (K_\theta(x_i, x_j))_{i,j=1,\ldots,n}$. Hence, $Y^{(n)}$ has a Gaussian density which value of $y^{(n)}$ is

$$f_\theta(y^{(n)}) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(R_\theta)}} e^{-\frac{1}{2}(y^{(n)})^\top R_\theta^{-1} y^{(n)}}.$$

Writing again

$$y^{(n)} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix},$$

for the observations $Y(x_1) = y_1, \ldots, Y(x_n) = y_n$, the maximum likelihood estimator is

$$\hat{\theta}_{\mathrm{ML}}(y^{(n)}) \in \underset{\theta \in \Theta}{\operatorname{argmax}} f_\theta(y^{(n)}).$$

We can simplify this and obtain, with log the Neperian logarithm,

$$\begin{aligned} \hat{\theta}_{\mathrm{ML}}(y^{(n)}) &\in \underset{\theta \in \Theta}{\operatorname{argmax}} \quad \log\left(\frac{1}{(2\pi)^{n/2}\sqrt{\det(R_\theta)}} e^{-\frac{1}{2}(y^{(n)})^\top R_\theta^{-1} y^{(n)}}\right) \\ &\in \underset{\theta \in \Theta}{\operatorname{argmax}} \quad -\frac{1}{2}\log\left(\det(R_\theta)\right) - \frac{1}{2}(y^{(n)})^\top R_\theta^{-1} y^{(n)} \\ &\in \underset{\theta \in \Theta}{\operatorname{argmin}} \quad \log\left(\det(R_\theta)\right) + (y^{(n)})^\top R_\theta^{-1} y^{(n)} \end{aligned}$$

2 Sensitivity analysis

2.1 Context

We consider the computer model

$$\begin{aligned} f : [0,1]^d &\to \mathbb{R} \\ x &\mapsto f(x). \end{aligned}$$

We consider d distributions for the d inputs $x_1, \ldots, x_d : \mathcal{L}_1, \ldots, \mathcal{L}_d$. For $i = 1, \ldots, d$, we assume that the cumulative distribution function $F_{\mathcal{L}_i}$ of $\mathcal{L}_i$ is bijective from $[0,1]$ to $[0,1]$. These distributions can model:

- An uncertainty when we can not know the value that x_i will take in practice (example: the amount of rainfall received by a hydrology computer model).
- An opinion on the values that will be chosen for x_i when a computer experiment will be performed (this is relatively subjective, typically a uniform distribution on $[0,1]$ may be chosen).

Actually, we may assume without loss of generality that, for $i = 1, \ldots, d$, $\mathcal{L}_i$ is the uniform distribution on $[0,1]^d$. Indeed, we can consider:

- The input $\tilde{x}_i = F_{\mathcal{L}_i}(x_i)$ for $i = 1, \ldots, d$. In this case, for $X_i \sim \mathcal{L}_i$ and $\tilde{X}_i = F_{\mathcal{L}_i}(X_i)$, we have, for $t \in [0,1]$, $P(\tilde{X}_i \leq t) = P(F_{\mathcal{L}_i}(X_i) \leq t) = P(X_i \leq F_{\mathcal{L}_i}^{-1}(t)) = F_{\mathcal{L}_i}\left(F_{\mathcal{L}_i}^{-1}(t)\right) = t$. Hence, $\tilde{X}_i$ follows the uniform distribution on $[0,1]^d$.
- The computer model defined by, for $\tilde{x}_1, \ldots, \tilde{x}_d \in [0,1]$, $\tilde{f}(\tilde{x}_1, \ldots, \tilde{x}_d) = f(F_{\mathcal{L}_i}^{-1}(\tilde{x}_1), \ldots, F_{\mathcal{L}_i}^{-1}(\tilde{x}_d)) = f(x_1, \ldots, x_d)$.

Hence, in the rest of this section on sensitivity analysis, we consider

$$f : [0,1]^d \to \mathbb{R}$$

and d independent uniformly distributed on $[0,1]$ random variables $X_1, \ldots, X_d$. The main question that we want to address is: for $i = 1, \ldots, d$, is the random variable $f(X_1, \ldots, X_d)$ strongly or weakly impacted by the value that X_i will take? Answering this question can have the following benefits.

- We can neglect uninfluential variables, and thus simplify the computer model f by freezing some of its inputs (for instance by enforcing their values to $1/2$).
- If a variable X_i is particularly influential, we can invest more resources in the knowledge of its possible values.

2.2 ANOVA decomposition

ANOVA means analysis of variance.

The case of functions of two variables

Proposition 16 *Let $f : [0,1]^2 \to \mathbb{R}$ such that*

$$\int_0^1 \int_0^1 f^2(x_1, x_2) dx_1 dx_2 < +\infty.$$

Then, there exists a unique decomposition of f of the form

$$f(x_1, x_2) = f_0 + f_1(x_1) + f_2(x_2) + f_{1,2}(x_1, x_2),$$

for all $x_1, x_2 \in [0,1]$, where f_0 is a constant and where the functions f_1, f_1 and $f_{1,2}$ satisfy

$$\int_0^1 f_1(x_1) dx_1 = 0,$$

$$\int_0^1 f_2(x_2) dx_2 = 0$$

and for all $x_1, x_2 \in [0,1]$,

$$\int_0^1 f_{1,2}(x_1, x_2) dx_1 = 0,$$

and

$$\int_0^1 f_{1,2}(x_1, x_2) dx_2 = 0.$$

We do not prove Proposition 16 in these lecture notes.

Definition 17 *We call the decomposition of Proposition 16 the ANOVA decomposition of f.*

In this section, we will interpret the space of square summable functions on $[0,1]^2$

$$\{g : [0,1]^2 \to \mathbb{R}; \int_0^1 \int_0^1 g(x_1, x_2)^2 dx_1 dx_2 < +\infty\}$$

as a Hilbert space with inner product given by, for h_1, h_2 from $[0,1]^d$ to $\mathbb{R}$, square summable,

$$\langle h_1, h_2 \rangle = \int_0^1 \int_0^1 h_1(x_1, x_2) h_2(x_1, x_2) dx_1 dx_2.$$

Remark 18 *In Proposition 16, the functions f_0, f_1, f_2 and $f_{1,2}$ are orthogonal.*

Proof of Remark 18 We have

$$\int_0^1 \int_0^1 f_0 f_1(x_1) dx_1 dx_2 = f_0 \int_0^1 f_1(x_1) dx_1 = 0,$$

from Proposition 16. Similarly

$$\int_0^1 \int_0^1 f_0 f_2(x_2) dx_2 = 0.$$

We also have

$$\int_0^1 \int_0^1 f_0 f_{1,2}(x_1, x_2) dx_1 dx_2 = f_0 \int_0^1 \left(\int_0^1 f_{1,2}(x_1, x_2) dx_2 \right) dx_1 = 0$$

from Proposition 16. We also have

$$\int_0^1 \int_0^1 f_1(x_1) f_2(x_2) dx_1 dx_2 = \left(\int_0^1 f_1(x_1) dx_1 \right) \left(\int_0^1 f_2(x_2) dx_2 \right) = 0.$$

Furthermore, again from Proposition 16,

$$\begin{aligned} \int_0^1 \int_0^1 f_1(x_1) f_{1,2}(x_1, x_2) dx_1 dx_2 &= \int_0^1 f_1(x_1) \left(\int_0^1 f_{1,2}(x_1, x_2) dx_2 \right) dx_1 \\ &= \int_0^1 f_1(x_1) 0 dx_1 \\ &= 0. \end{aligned}$$

Finally

$$\begin{aligned} \int_0^1 \int_0^1 f_2(x_2) f_{1,2}(x_1, x_2) dx_1 dx_2 &= \int_0^1 f_2(x_2) \left(\int_0^1 f_{1,2}(x_1, x_2) dx_1 \right) dx_2 \\ &= \int_0^1 f_2(x_2) 0 dx_2 \\ &= 0. \end{aligned}$$

□

In fact, we know the expression of the functions in the ANOVA decomposition.

Proposition 19 *The functions f_0, f_1, f_2 and $f_{1,2}$ in Proposition 16 are given by, for $x_1, x_2 \in [0,1]$,*

$$\begin{aligned} f_0 &= \int_0^1 \int_0^1 f_{1,2}(x_1, x_2) dx_1 dx_2, \\ f_1(x_1) &= \int_0^1 f(x_1, x_2) dx_2 - f_0, \\ f_2(x_2) &= \int_0^1 f(x_1, x_2) dx_1 - f_0 \end{aligned}$$

and

$$f_{1,2}(x_1, x_2) = f(x_1, x_2) - (f_0 + f_1(x_1) + f_2(x_2)).$$

Proof of Proposition 19 Since there is unicity in Proposition 16, we just need to show that the functions of Proposition 19 satisfy the conditions of Proposition 16. Of course we have

$$f(x_1, x_2) = f_0 + f_1(x_1) + f_2(x_2) + f_{1,2}(x_1, x_2),$$

for all $x_1, x_2 \in [0,1]$. For all $x_1, x_2 \in [0,1]$ we also have

$$\int_0^1 f_1(x_1) dx_1 = \int_0^1 \int_0^1 f(x_1, x_2) dx_1 dx_2 - f_0 = f_0 - f_0 = 0.$$

Similarly

$$\int_0^1 f_2(x_2) dx_2 = 0.$$

We also have

$$\begin{aligned} \int_0^1 f_{1,2}(x_1, x_2) dx_2 &= \int_0^1 f(x_1, x_2) dx_2 - f_0 - f_1(x_1) - \int_0^1 f_2(x_2) dx_2 \\ &= \int_0^1 f(x_1, x_2) dx_2 - f_0 - \left(\int_0^1 f(x_1, x_2) dx_2 - f_0 \right) - 0 \\ &= 0. \end{aligned}$$

Similarly

$$\int_0^1 f_{1,2}(x_1, x_2) dx_1 = 0.$$

Remark 20 *We have the following interpretations of the ANOVA decomposition.*

i) f does not depend on x_2 ($f(x_1,x_2) = f(x_1,x_2')$ for all $x_1, x_2, x_2' \in [0,1]$) if and only if, in the ANOVA decomposition of f, we have $f_2 = f_{1,2} = 0$.

ii) f is additive (there exist two functions g_1 and g_2 from $[0,1]$ to $\mathbb{R}$ such that $f(x_1,x_2) = g_1(x_1) + g_2(x_2)$ for all $x_1, x_2 \in [0,1]$) if and only if, in the ANOVA decomposition of f, we have $f_{1,2} = 0$.

Proof of Remark 20 i) If $f(x_1,x_2) = f_0 + f_1(x_1)$, then $f(x_1,x_2) = f(x_1,x_2')$ for all $x_1, x_2, x_2' \in [0,1]$. So $\Longleftarrow$ is proved.

Let us now prove $\Longrightarrow$. For $x_1 \in [0,1]$, let us write $f_{x_1} = f(x_1,0)$. Then

$$\int_0^1 f(x_1,x_2)dx_2 = \int_0^1 f(x_1,0)dx_2 = \int_0^1 f_{x_1}dx_2 = f_{x_1}.$$

Hence, for $x_1, x_2 \in [0,1]$,

$$\begin{aligned} f(x_1,x_2) &= f_{x_1} \\ &= \int_0^1 f(x_1,x_2)dx_2 \\ &= \int_0^1\int_0^1 f(x_1,x_2)dx_1dx_2 + \int_0^1 f(x_1,x_2)dx_2 - \int_0^1\int_0^1 f(x_1,x_2)dx_1dx_2. \end{aligned}$$

The above decomposition is a valid ANOVA decomposition of f as the sum of a constant and as a function of x_1 with mean zero. Hence by unicity of the ANOVA decomposition, $f_2 = f_{1,2} = 0$. □

Proof of Remark 20 ii) Let us prove $\Longleftarrow$. We have $f(x_1,x_2) = f_0 + f_1(x_1) + f_2(x_2)$ so we let, for instance $g_1(x_1) = f_0 + f_1(x_1)$ and $g_2(x_2) = f_2(x_2)$. This proves $\Longleftarrow$.

Let us now prove $\Longrightarrow$. We have for $x_1, x_2 \in [0,1]$, $f(x_1,x_2) = g_1(x_1) + g_2(x_2)$. Hence we may write

$$f(x_1,x_2) = \underbrace{\int_0^1 g_1(x_1)dx_1 + \int_0^1 g_1(x_1)dx_1}_{f_0} + \underbrace{g_1(x_1) - \int_0^1 g_1(x_1)dx_1}_{f_1(x_1)} + \underbrace{g_2(x_2) - \int_0^1 g_2(x_2)dx_2}_{f_2(x_2)} + \underbrace{0}_{f_{1,2}(x_1,x_2)}.$$

We see that

$$\int_0^1 f_1(x_1) = \int_0^1 f_2(x_2) = 0.$$

Hence, f_0, f_1, f_2 and $f_{1,2}$ satisfy the conditions of Proposition 16. Hence by unicity, $f_{1,2} = 0$. □

Following Remark 20 ii), we call $f_{1,2}$ the term corresponding to the interactions of x_1 and x_2 in the function f.

The general case

Proposition 21 (ANOVA decomposition of f) *Let $f : [0,1]^d \to \mathbb{R}$ such that*

$$\int_{[0,1]^d} f^2(x)dx < +\infty.$$

Then, there exists a unique decomposition of f of the form,

$$\begin{aligned} f(x) =& f_0 \\ &+ \sum_{i=1}^d f_i(x_i) \\ &+ \sum_{1\le i<j\le d} f_{i,j}(x_i,x_j) \\ &+ \dots \\ &+ f_{1,\dots,d}(x_1,\dots,x_d) \\ =& \sum_{u\in S} f_u(x_u), \end{aligned}$$

where

- *S is the set of all subsets of $\{1,\dots,d\}$ (including the empty set $\varnothing$ and the entire set $\{1,\dots,d\}$),*
- *$x=(x_1,\dots,x_d)\in[0,1]^d$,*
- *for $u=\varnothing$, $f_u(x_u)$ is a constant f_0, for $u=\{i_1,\dots,i_k\}\neq\varnothing$, with $i_1<\cdots<i_k$, we let $f_u(x_u)=f_{i_1,\dots,i_k}(x_{i_1},\dots,x_{i_k})$,*

such that the functions f_u, for $u\in S$, satisfy

$$\int_0^1 f_u(x_u)dx_i=0,\quad \textit{for all } i\in u,\quad \textit{for all } u\in S.$$

We do not prove Proposition 21 in these lecture notes. In the rest of the section, we let $|u|$ be the cardinality of $u\in S$.

Remark 22 *In the ANOVA decomposition of f, the functions are orthogonal:*

$$\int_{[0,1]^d} f_u(x_u)f_v(x_v)dx=0$$

for $u,v\in S$, $u\neq v$.

Proof of Remark 22 We can find $i\in u$ such that $i\notin v$ (up to exchanging u and v). We have

$$\begin{aligned}\int_{[0,1]^d} f_u(x_u)f_v(x_v)dx &= \int_{[0,1]^{|u\cup v|}} f_u(x_u)f_v(x_v)dx_{u\cup v}\\ &= \int_{[0,1]^{|u\cup v|-1}} f_v(x_v)\left(\int_0^1 f_u(x_u)dx_i\right)dx_{u\cup v\setminus\{i\}}.\end{aligned}$$

Then, by the ANOVA property, $\int_0^1 f_u(x_u)dx_i=0$ and so

$$\int_{[0,1]^d} f_u(x_u)f_v(x_v)dx=0.$$

□

We have similar interpretations as in the two variable case.

Remark 23 *i) For $i=1,\dots,d$, f does not depend on x_i ($f(x)=f(x')$ for all $x,x'\in[0,1]^d$ such that $x_{\{1,\dots,d\}\setminus\{i\}}=x'_{\{1,\dots,d\}\setminus\{i\}}$) if and only if, in the ANOVA decomposition of f, we have $f_u=0$ for $i\in u$ for all $u\in S$.*

ii) f is additive (there exist d functions $g_1,\dots,g_d$ from $[0,1]$ to $\mathbb{R}$ such that $f(x_1,\dots,x_d)=g_1(x_1)+\cdots+g_d(x_d)$ for all $x_1,\dots,x_d\in[0,1]$) if and only if, in the ANOVA decomposition of f, we have $f_u=0$ for $u\in S$ such that $|u|>1$.

Proof of Remark 23 i) If $f_u=0$ for $i\in u$, we see in Proposition 21 that x_i does not appear in the expression of f, so f does not depend on x_i. Hence $\Longleftarrow$ is proved.

Let us now prove $\Longrightarrow$. If f does not depend on x_i, we let, for $x\in[0,1]^d$,

- $\tilde{x}_1=x_1$
- $\vdots$
- $\tilde{x}_{i-1}=x_{i-1}$
- $\tilde{x}_i=x_{i+1}$

- $\vdots$
- $\tilde{x}_{d-1} = x_d$

and we let

$$\tilde{x} = (\tilde{x}_1, \ldots, \tilde{x}_{d-1}) \in [0,1]^{d-1}$$

($\tilde{x}$ implicitly depends on x). We let $\tilde{f}(\tilde{x}_1, \ldots, \tilde{x}_{d-1}) = f(x_1, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_d)$ $(= f(x_1, \ldots, x_d))$. We write $\tilde{S}$ the set of all subsets of $\{1, \ldots, d-1\}$. There exists an ANOVA decomposition of $\tilde{f}$ written

$$\tilde{f}(\tilde{x}) = \sum_{\tilde{u} \in \tilde{S}} \tilde{f}_{\tilde{u}}(\tilde{x}_{\tilde{u}}).$$

For $u = \varnothing$, we let $f_u = \tilde{f}_\varnothing = \tilde{f}_0$. Let us now consider $u \neq \varnothing$. We then write $f_u(x_u) = 0$ for $i \in u$ and $f_u(x_u) = \tilde{f}_{\tilde{u}}(\tilde{x}_{\tilde{u}})$ for $i \notin u$ with

$$u = \{i_1 < \cdots < i_l < i_{l+1} < \cdots < i_k\},$$

$$i_l < i < i_{l+1} \qquad \text{(convention } i_0 = 0 \text{ and } i_{k+1} = \text{d+1)}$$

and

$$\tilde{u} = \{i_1 < \cdots < i_l < i_{l+1} - 1 < \cdots < i_k - 1\}.$$

Let us now show that the $f_u(x_u)$, $u \in S$, satisfy the properties of an ANOVA decomposition. Of course we have, for $x \in [0,1]^d$,

$$f(x_1, \ldots, x_d) = f(x_1, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_d) = \sum_{\tilde{u} \in \tilde{S}} \tilde{f}_{\tilde{u}}(\tilde{x}_{\tilde{u}}) = \sum_{u \in S} f_u(x_u).$$

For $i \in u$ and $j \in u$, we have

$$\int_0^1 f_u(x_u) dx_j = \int_0^1 0 dx_j = 0.$$

Consider now $i \in u$ and $j \notin u$.

If $j < i$.

$$\int_0^1 f_u(x_u) dx_j = \int_0^1 \tilde{f}_{\tilde{u}}(\tilde{x}_{\tilde{u}}) d\tilde{x}_j = 0$$

because $j \in \tilde{u}$ and by definition of the ANOVA decomposition of $\tilde{f}$.

If $i < j$. We have

$$\int_0^1 f_u(x_u) dx_j = \int_0^1 \tilde{f}_{\tilde{u}}(\tilde{x}_{\tilde{u}}) d\tilde{x}_{j-1} = 0$$

because $j - 1 \in \tilde{u}$ and by definition of the ANOVA decomposition of $\tilde{f}$.

Hence, finally, the $f_u(x_u)$, $u \in S$ constitute the ANOVA decomposition of f (by unicity). Hence, in the ANOVA decomposition of f, $f_u(x_u) = 0$ for all $x \in [0,1]^d$ and $u \in S$ with $i \in u$. □

Proof of Remark 23 ii) If $f_u = 0$ for $|u| > 1$ then

$$f(x) = \underbrace{f_0 + f_1(x_1)}_{g_1(x_1)} + \cdots + \underbrace{f_d(x_d)}_{g_d(x_d)}$$

so f is additive. So $\Longleftarrow$ is proved.

Let us now prove $\Longrightarrow$. If f is additive,

$$\begin{aligned} f(x) &= g_1(x_1) + \cdots + g_d(x_d) \\ &= \underbrace{\int_0^1 g_1(x_1)dx_1 + \cdots + \int_0^1 g_d(x_d)dx_d}_{f_0} \\ &\quad + \underbrace{g_1(x_1) - \int_0^1 g_1(x_1)dx_1}_{f_1(x_1)} + \cdots + \underbrace{g_d(x_d) - \int_0^1 g_d(x_d)dx_d}_{f_d(x_d)} \\ &\quad + \sum_{\substack{u\in S \\ |u|>1}} \underbrace{0}_{f_u(x_u)}. \end{aligned}$$

We check that for $i = 1, \ldots, d$,

$$\int_0^1 f_i(x_i)dx_i = \int_0^1 g_i(x_i)dx_i - \int_0^1 g_i(x_i)dx_i = 0.$$

In all the other cases where $|u| > 1$, for $i \in u$,

$$\int_0^1 f_u(x_u) = \int_0^1 f_u(x_u)dx_i = 0.$$

Hence $(f_u)_{u\in S}$ satisfy the conditions of the ANOVA decomposition of f and thus by unicity they constitute the ANOVA decomposition of f. Hence, in the ANOVA decomposition of f, $f_u = 0$ for $u \in S$, $|u| > 1$. □

Similarly as for the two variable case, when $u = \{i_1, \ldots, i_k\}$, we call $f_u(x_u)$ the term corresponding to the interactions of $x_{i_1}, \ldots, x_{i_k}$ in the function f.

2.3 Sobol sensitivity indices

We now write the probabilistic form

$$Y = f(X_1, \ldots, X_d)$$

where each X_i, $i \in \{1, \ldots, d\}$, is uniform on $[0, 1]$ and where $X_1, \ldots, X_d$ are independent. Then

$$f(X_1, \ldots, X_d) = \sum_{u\in S} f_u(X_u)$$

and the random variables in the above sum are decorrelated and have mean zero (except f_0): for $u, v \in S$, $u \neq v$,

$$\mathrm{Cov}(f_u(X_u), f_v X_v) = 0$$

and for $u \neq \varnothing$,

$$\mathbb{E}(f_u(X_u)) = 0.$$

Indeed, we have for $u \neq \varnothing$, with $i \in u$,

$$\mathbb{E}(f_u(X_u)) = \int_{[0,1]^d} f_u(x_u)dx = \int_{[0,1]^{d-1}} \left(\int_0^1 f_u(x_u)dx_i\right) dx_{\{1,\ldots,d\}\setminus\{i\}} = \int_{[0,1]^{d-1}} 0 dx_{\{1,\ldots,d\}\setminus\{i\}} = 0$$

from the properties of the ANOVA decomposition. Hence we have for $u, v \in S$, $u \neq v$,

$$\mathrm{Cov}(f_u(X_u), f_v X_v) = \mathbb{E}(f_u(X_u) f_v X_v) = \int_{[0,1]^d} f_u(x_u) f_v x_v dx = 0$$

from Remark 22.

From the decorrelation, we have

$$\mathrm{Var}\left(f(X_1,\dots,X_d)\right) = \sum_{u\in S} \mathrm{Var}\left(f_u(X_u)\right).$$

This is why the name analysis of variance is used. We write for $u \in S$

$$I_u = \frac{\mathrm{Var}(f_u(X_u))}{\mathrm{Var}(f(X_1,\dots,X_d))}$$

and we call I_u the sensitivity index associated to the group of variables u. More precisely:

- If $u = \{i\}$ for $i = 1,\dots,d$, I_u is called the sensitivity index for the main effect of the variable i.
- If $u = \{i_1,\dots,i_k\}$ with $k > 1$, I_u is called sensitivity index for the interactions of the variables $i_1,\dots,i_k$.

We have

$$\sum_{u\in S} I_u = \frac{\sum_{u\in S} \mathrm{Var}\left(f_u(X_u)\right)}{\mathrm{Var}\left(f(X_1,\dots,X_d)\right)}$$

and thus

$$\boxed{\sum_{u\in S} I_u = 1.}$$

Hence, we have a decomposition of the total variance into fractions that correspond to the main and interaction effects of the variables.

An explicit computation in dimension two For $d = 2$, we consider $f(x_1,x_2) = x_2(x_1+w)$ with $w \in \mathbb{R}$ a fixed constant. The ANOVA decomposition of f is

$$f(x_1,x_2) = \underbrace{\frac{1}{2}\left(\frac{1}{2}+w\right)}_{f_0} + \underbrace{\frac{1}{2}\left(x_1-\frac{1}{2}\right)}_{f_1(x_1)} + \underbrace{\left(\frac{1}{2}+w\right)\left(x_2-\frac{1}{2}\right)}_{f_2(x_2)} + \underbrace{\left(x_1-\frac{1}{2}\right)\left(x_2-\frac{1}{2}\right)}_{f_{1,2}(x_1,x_2)}.$$

Indeed the decomposition is equal to

$$\begin{aligned}\cancel{\frac{1}{4}} + \cancel{\frac{1}{2}}w + \cancel{\frac{1}{2}}x_1 - \cancel{\frac{1}{4}} + \cancel{\frac{1}{2}}x_2 - \cancel{\frac{1}{4}} + wx_2 - \cancel{\frac{1}{2}}w + x_1x_2 - \cancel{\frac{1}{2}}x_1 - \cancel{\frac{1}{2}}x_2 + \cancel{\frac{1}{4}} &= x_1x_2 + wx_2 \\ &= x_2(x_1+w).\end{aligned}$$

Furthermore we check $\int_0^1 f_1(x_1)dx_1 = 0$, $\int_0^1 (f_2(x_2)dx_2 = 0$, $\int_0^1 f_{1,2}(x_1,x_2)dx_2 = 0$ and $\int_0^1 f_{1,2}(x_1,x_2)dx_1 = 0$ for $x_1,x_2 \in [0,1]$ (using $\int_0^1 x dx = 1/2$).

Let us compute all the variances. We will use the formulas, for $j = 1,2$

$$\mathbb{E}(X_j) = \frac{1}{2},\ \mathbb{E}(X_j^2) = \frac{1}{3},\ \mathrm{Var}(X_j) = \frac{1}{12}.$$

We have

$$\begin{aligned}
\mathrm{Var}\left(f(X_1,X_2)\right) &= \mathrm{Var}\left(X_2(X_1+w)\right)\\
&= \mathbb{E}\left(X_2^2(X_1+w)^2\right) - \mathbb{E}\left(X_2(X_1+w)\right)^2\\
&= \mathbb{E}(X_2^2)\mathbb{E}\left((X_1+w)^2\right) - \mathbb{E}^2(X_2)\mathbb{E}^2(X_1+w)\\
&= \frac{1}{3}\left(\mathbb{E}(X_1^2)+w^2+2w\mathbb{E}(X_1)\right) - \frac{1}{4}\left(\frac{1}{2}+w\right)^2\\
&= \frac{1}{9}+\frac{w^2}{3}+\frac{w}{3}-\frac{1}{4}\left(\frac{1}{4}+w^2+w\right)\\
&= \frac{16}{144}+\frac{4w^2}{12}+\frac{4w}{3}-\frac{9}{144}-\frac{3w^2}{12}-\frac{3w}{12}\\
&= \frac{7}{144}+\frac{w^2}{12}+\frac{w}{12}\\
&= \frac{1}{12}\left(\frac{7}{12}+w^2+w\right)\\
&= \frac{1}{12}\left(\left(w+\frac{1}{2}\right)^2+\frac{1}{3}\right).
\end{aligned}$$

Then

$$\begin{aligned}
\mathrm{Var}(f_1(X_1)) &= \mathrm{Var}\left(\frac{1}{2}\left(X_1-\frac{1}{2}\right)\right)\\
&= \frac{1}{4}\mathrm{Var}(X_1)\\
&= \frac{1}{4}\frac{1}{12}.
\end{aligned}$$

Then

$$\begin{aligned}
\mathrm{Var}(f_2(X_2)) &= \mathrm{Var}\left(\left(\frac{1}{2}+w\right)\left(X_2-\frac{1}{2}\right)\right)\\
&= \left(\frac{1}{2}+w\right)^2\frac{1}{12}.
\end{aligned}$$

Then

$$\begin{aligned}
\mathrm{Var}(f_{1,2}(X_1,X_2)) &= \mathrm{Var}\left(\left(X_1-\frac{1}{2}\right)\left(X_2-\frac{1}{2}\right)\right)\\
&= \mathbb{E}\left(\left(X_1-\frac{1}{2}\right)^2\left(X_2-\frac{1}{2}\right)^2\right) - \mathbb{E}^2\left(\left(X_1-\frac{1}{2}\right)\left(X_2-\frac{1}{2}\right)\right)\\
&= \mathrm{Var}(X_1)\mathrm{Var}(X_2) - 0\\
&= \frac{1}{144}.
\end{aligned}$$

Hence,

$$I_1 = \frac{\frac{1}{4}\frac{1}{12}}{\frac{1}{12}\left(\left(w+\frac{1}{2}\right)^2+\frac{1}{3}\right)}$$

and thus

$$\boxed{I_1 = \frac{\frac{1}{4}}{\left(w+\frac{1}{2}\right)^2+\frac{1}{3}}}$$

Also

$$\boxed{I_2 = \frac{\left(w+\frac{1}{2}\right)^2}{\left(w+\frac{1}{2}\right)^2+\frac{1}{3}}.}$$

Finally,

$$I_{1,2} = \frac{\frac{1}{12}\frac{1}{12}}{\frac{1}{12}\left(\left(w+\frac{1}{2}\right)^2+\frac{1}{3}\right)}$$

and thus

$$\boxed{I_{1,2} = \frac{\frac{1}{12}}{\left(w+\frac{1}{2}\right)^2+\frac{1}{3}}}$$

The interpretation is the following:

- If $w \to \pm\infty$, $I_1 \to 0$, $I_2 \to 1$ and $I_{1,2} \to 0$. Indeed, if w is very large then f almost does not depend on the specific value that x_1 is taking in $[0,1]$ because $w \approx w+1$. Hence $f(x_1, x_2) \approx wx_2$.
- I_1 is maximal for $w = -1/2$ with $I_1 = (1/4)/(1/3) = 3/4$. Indeed, $x_1 - 1/2$ gives the most impact to the value of x_1 in $[0,1]$. In particular, the sign is most uncertain ($P(X_1 - 1/2 \geq 0) = P(X_1 - 1/2 \leq 0) = 1/2$).
- For $w = 0$, $I_1 = I_2 = (1/4)/((1/4) + (1/3)) = (3/12)/((3/12) + (4/12)) = 3/7$ and $I_{1,2} = (1/12)/((1/4) + (1/3)) = 1/7$. It is normal that $I_1 = I_2$ because $f(x_1, x_2) = x_1x_2$ is symmetric in x_1, x_2.

Limitation There are 2^d subsets of $\{1, \ldots, d\}$ so 2^d indices I_u, $u \in S$. When d is above, say, 10, this makes too many quantities to estimate/interpret. We thus now study a smaller number of indices that are particularly interpretable.

First order indices and ranking of the variables by order of priority The sensitivity index I_i, for $i = 1, \ldots, d$ is called a first order index.

Proposition 24 *We have for $i = 1, \ldots, d$,*

$$I_i = \frac{\mathrm{Var}\left(\mathbb{E}\left(Y|X_i\right)\right)}{\mathrm{Var}(Y)} = 1 - \frac{\mathbb{E}\left(\mathrm{Var}\left(Y|X_i\right)\right)}{\mathrm{Var}(Y).}$$

Hence, if $i \in \{1, \ldots, d\}$ is such that I_i is the largest, then $\mathbb{E}(\mathrm{Var}(Y|X_i))$ is the smallest. Hence, knowing X_i decrease the variance of Y the most. X_i is thus the prioritary variable, on which we want to make the most effort to reduce the uncertainty.

Proof of Proposition 24 In the ANOVA decomposition of f,

$$\begin{aligned}\mathbb{E}(Y|X_i) &= \sum_{u\in S} \mathbb{E}\left(f_u(X_u)|X_i\right) \\ &= f_0 + f_i(X_i) + \sum_{\substack{u\in S \\ u\neq\varnothing \\ u\neq\{i\}}} \mathbb{E}\left(f_u(X_u)|X_i\right).\end{aligned}$$

For $u \in S$ such that $u \neq \varnothing$ and $u \neq \{i\}$, there exists $j \in \{1, \ldots, d\}$ such that $j \in u$ and $j \neq i$. We have

$$\begin{aligned}\mathbb{E}\left(f_u(X_u)|X_i\right) &= \int_{[0,1]^{d-1}} f_u(x_{\{1,\ldots,d\}\setminus\{i\}}, X_i) dx_{\{1,\ldots,d\}\setminus\{i\}} \\ &= \int_{[0,1]^{d-2}} \underbrace{\left(\int_0^1 f_u(x_{\{1,\ldots,d\}\setminus\{i\}}, X_i) dx_j\right)}_{=0} dx_{\{1,\ldots,d\}\setminus\{i,j\}} \\ &= 0.\end{aligned}$$

Hence $\mathbb{E}(Y|X_i) = f_0 + f_i(X_i)$ and so $\mathrm{Var}(\mathbb{E}(Y|X_i)) = \mathrm{Var}(f_i(X_i))$. This shows the first equation of the proposition. For the second equation, we have for any random variables A and B with finite variances

$$\mathrm{Var}(B) = \mathrm{Var}(\mathbb{E}(B|A)) + \mathbb{E}(\mathrm{Var}(B|A))$$

which is called the law of total variance. Indeed

$$\begin{aligned}
\mathrm{Var}(B) &= \mathbb{E}(B^2) - \mathbb{E}^2(B) \\
&= \mathbb{E}\left(\mathbb{E}\left((B - \mathbb{E}(B|A) + \mathbb{E}(B|A))^2 \middle| A\right)\right) - \mathbb{E}^2\left(\mathbb{E}(B|A)\right) \\
&= \mathbb{E}\left(\mathbb{E}\left((B - \mathbb{E}(B|A))^2 \middle| A\right)\right) + \mathbb{E}\left(\mathbb{E}^2(B|A)\right) + 2\mathbb{E}\left(\mathbb{E}\left((B - \mathbb{E}(B|A))\mathbb{E}(B|A)\middle| A\right)\right) - \mathbb{E}^2\left(\mathbb{E}(B|A)\right) \\
&= \mathbb{E}(\mathrm{Var}(B|A)) + 2\mathbb{E}\left(\mathbb{E}^2(B|A) - \mathbb{E}^2(B|A)\right) + \mathbb{E}\left(\mathbb{E}^2(B|A)\right) - \mathbb{E}^2\left(\mathbb{E}(B|A)\right) \\
&= \mathbb{E}(\mathrm{Var}(B|A)) + \mathrm{Var}(\mathbb{E}(B|A)).
\end{aligned}$$

Hence

$$\frac{\mathrm{Var}\left(\mathbb{E}\left(Y|X_i\right)\right)}{\mathrm{Var}(Y)} = 1 - \frac{\mathbb{E}\left(\mathrm{Var}\left(Y|X_i\right)\right)}{\mathrm{Var}(Y).}$$

□

Total indices and choice of the inputs to freeze For $i = 1, \ldots, d$, we define the total index TI_i of the variable i as

$$\mathrm{TI}_i = \sum_{\substack{u \in S \\ i \in u}} I_u.$$

Proposition 25 *For $i = 1, \ldots, d$,*

$$\mathrm{TI}_i = \frac{\mathbb{E}\left(\mathrm{Var}(Y|X_{-i})\right)}{\mathrm{Var}(Y)} = 1 - \frac{\mathrm{Var}\left(\mathbb{E}(Y|X_{-i})\right)}{\mathrm{Var}(Y)},$$

with $X_{-i} = X_{\{1,\ldots,d\}\setminus\{i\}}$.

If $\mathrm{TI}_i = 0$, then $\mathrm{Var}(Y|X_{-i}) = 0$ and thus Y is a deterministic function of X_{-i}. Hence, $f(X_1, \ldots, X_d)$ does not depend on X_i. Hence if TI_i is very small, Y almost does not depend on X_i and we can freeze X_i (for instance to $1/2$) to decrease the number of input variables fo the computer model f.

Proof of Proposition 25 We have

$$\begin{aligned}
\mathbb{E}(Y|X_{-i}) &= \mathbb{E}\left(\sum_{u \in S} f_u(X_u)|X_{-i}\right) \\
&= f_0 + \sum_{\substack{u \in S \\ u \neq \varnothing \\ i \notin u}} f_u(X_u) + \sum_{\substack{u \in S \\ i \in u}} \int_0^1 f_u(X_{u\setminus\{i\}}, x_i) dx_i \\
&= f_0 + \sum_{\substack{u \in S \\ u \neq \varnothing \\ i \notin u}} f_u(X_u).
\end{aligned}$$

Hence

$$\frac{\mathrm{Var}\left(\mathbb{E}(Y|X_{-i})\right)}{\mathrm{Var}(Y)} = 1 - \mathrm{TI}_i.$$

Hence

$$\mathrm{TI}_i = 1 - \frac{\mathrm{Var}\left(\mathbb{E}(Y|X_{-i})\right)}{\mathrm{Var}(Y)}.$$

Finally we use the law of total variance $\mathrm{Var}(Y) = \mathbb{E}(\mathrm{Var}(Y|X_{-i})) + \mathrm{Var}(\mathbb{E}(Y|X_{-i}))$. □

We call I_i and $\mathrm{TI}_i = 0$, for $i = 1, \ldots, d$ the Sobol sensitivity indices.

Estimation of the sensitivity indices For $i = 1, \dots, d$, for I_i, we want to estimate $\mathrm{Var}(\mathbb{E}(Y|X_i))$.

Proposition 26 *Let $i \in \{1, \dots, d\}$. Let X^A and X^B be independent random vectors with uniform distribution on $[0,1]^d$. We let*

$$Y_i^A = f(X_1^A, \dots, X_{i-1}^A, X_i^A, X_{i+1}^A, \dots, X_d^A)$$

and

$$Y_i^B = f(X_1^B, \dots, X_{i-1}^B, X_i^A, X_{i+1}^B, \dots, X_d^B).$$

We also let

$$X^{A,i} = (X_1^A, \dots, X_{i-1}^A, X_i^A, X_{i+1}^A, \dots, X_d^A)$$

and

$$B^{B,i} = (X_1^B, \dots, X_{i-1}^B, X_i^A, X_{i+1}^B, \dots, X_d^B).$$

Then

$$\mathrm{Var}\left(\mathbb{E}\left(Y|X_i\right)\right) = \mathrm{Cov}(Y_i^A, Y_i^B).$$

Proof of Proposition 26 We have

$$\begin{aligned}
\mathrm{Cov}(Y_i^A, Y_i^B) &= \sum_{\substack{u \in S \\ u \neq \varnothing}} \sum_{\substack{v \in S \\ v \neq \varnothing}} \mathrm{Cov}\left(f_u(X_u^{A,i}), f_v(X_v^{B,i})\right) \\
&= \sum_{\substack{u \in S \\ u \neq \varnothing}} \sum_{\substack{v \in S \\ v \neq \varnothing}} \mathbb{E}\left(f_u(X_u^{A,i}) f_v(X_v^{B,i})\right).
\end{aligned}$$

Assume that in one of the above expectations there exists $j \neq i$ with $j \in u$ or $j \in v$. Assume first that $j \in v$. We have

$$\begin{aligned}
\mathbb{E}\left(f_u(X_u^{A,i}) f_v(X_v^{B,i})\right) &= \mathbb{E}\left(\mathbb{E}\left(f_u(X_u^{A,i}) f_v(X_v^{B,i}) \middle| X_{-j}^{B,i}, X^{A,i}\right)\right) \\
&= \mathbb{E}\left(f_u(X_u^{A,i}) \mathbb{E}\left(f_v(X_v^{B,i}) \middle| X_{-j}^{B,i}, X^{A,i}\right)\right) \\
&= \mathbb{E}\left(f_u(X_u^{A,i}) \int_0^1 f_v(X_{v\setminus\{j\}}^{B,i}, x_j) dx_j\right) \\
&= \mathbb{E}\left(f_u(X_u^{A,i}) 0\right) \\
&= 0.
\end{aligned}$$

Assume then that $j \in u$. We have

$$\begin{aligned}
\mathbb{E}\left(f_u(X_u^{A,i}) f_v(X_v^{B,i})\right) &= \mathbb{E}\left(\mathbb{E}\left(f_u(X_u^{A,i}) f_v(X_v^{B,i}) \middle| X_{-j}^{A,i}, X^{B,i}\right)\right) \\
&= \mathbb{E}\left(f_v(X_v^{B,i}) \mathbb{E}\left(f_u(X_u^{A,i}) \middle| X_{-j}^{A,i}, X^{B,i}\right)\right) \\
&= \mathbb{E}\left(f_v(X_v^{B,i}) \int_0^1 f_u(X_{u\setminus\{j\}}^{A,i}, x_j) dx_j\right) \\
&= \mathbb{E}\left(f_v(X_v^{B,i}) 0\right) \\
&= 0.
\end{aligned}$$

Hence we obtain

$$\begin{aligned}
\mathrm{Cov}\left(Y_i^A, Y_i^B\right) &= \mathrm{Cov}\left(f_i(X_i^{A,i}), f_i(X_i^{B,i})\right) \\
&= \mathrm{Var}(f_i(X_i)) \\
&= \mathrm{Var}\left(\mathbb{E}\left(Y|X_i\right)\right)
\end{aligned}$$

from the proof of Proposition 24. □

Consider then independent vectors

$$\begin{pmatrix} X^{A,i,(1)} \\ X^{B,i,(1)} \end{pmatrix}, \ldots, \begin{pmatrix} X^{A,i,(n)} \\ X^{B,i,(n)} \end{pmatrix}$$

with the same distribution as

$$\begin{pmatrix} X^{A,i} \\ X^{B,i} \end{pmatrix}.$$

Let, for $j = 1, \ldots, n$,

$$Y_i^{A,(j)} = f(X^{A,i,(j)}), \quad Y_i^{B,(j)} = f(X^{B,i,(j)}).$$

Then we let

$$\hat{f}_0 = \frac{1}{2n} \sum_{j=1}^{n} \left(Y_i^{A,(j)} + Y_i^{B,(j)} \right)$$

and

$$\hat{\sigma}^2 = \frac{1}{2n} \sum_{j=1}^{n} \left(\left[Y_i^{A,(j)} \right]^2 + \left[Y_i^{B,(j)} \right]^2 \right) - \hat{f}_0^2.$$

Then the estimator of I_i is

$$\hat{I}_i = \frac{\frac{1}{n} \sum_{j=1}^{n} Y_i^{A,(j)} Y_i^{B,(j)} - \hat{f}_0^2}{\hat{\sigma}^2.}$$

For TI_i we want to calculate $\mathbb{E}(\mathrm{Var}(Y|X_{-i}))$.

Proposition 27 *Let $i \in \{1, \ldots, d\}$ be fixed. Let X^C be a third random vector with uniform distribution on $[0,1]^d$, independent from X^A and X^B. Let*

$$X^{C,i} = \left(X_1^A, \ldots, X_{i-1}^A, X_i^C, X_{i+1}^A, \ldots, X_d^A \right).$$

Let $Y_i^C = f(X^{C,i})$. Then

$$\mathbb{E}\left(\mathrm{Var}\left(Y|X_{-i} \right) \right) = \frac{1}{2} \mathbb{E}\left(\left(Y_i^A - Y_i^C \right)^2 \right).$$

Proof of Proposition 27

We have shown in the proof of Propositioin 25 that

$$\mathrm{Var}\left(\mathbb{E}\left(Y|X_{-i} \right) \right) = \sum_{\substack{u \in S \\ i \notin u}} \mathrm{Var}(f_u(X_u)).$$

Hence

$$\mathbb{E}\left(\mathrm{Var}\left(Y|X_{-i} \right) \right) = \mathrm{Var}(Y) - \sum_{\substack{u \in S \\ i \notin u}} \mathrm{Var}(f_u(X_u)).$$

We have

$$\frac{1}{2} \mathbb{E}\left(\left(Y_i^A - Y_i^C \right)^2 \right) = \frac{1}{2} \left(\mathrm{Var}(Y) + \mathrm{Var}(Y) \right) - \mathrm{Cov}\left(Y_i^A, Y_i^C \right).$$

In order to conclude the proof, it is hence sufficient to prove that

$$\mathrm{Cov}\left(Y_i^A, Y_i^C \right) = \sum_{\substack{u \in S \\ i \notin u}} \mathrm{Var}(f_u(X_u)).$$

We have

$$\mathrm{Cov}\left(Y_i^A, Y_i^C \right) = \sum_{u \in S} \sum_{v \in S} \mathbb{E}\left(f_u(X_u^{A,i}) f_v(X_v^{C,i}) \right)$$

Assume that $i \in u$. We have

$$\begin{aligned}
\mathbb{E}\left(f_u(X_u^{A,i}) f_v(X_v^{C,i})\right) &= \mathbb{E}\left(\mathbb{E}\left(f_u(X_u^{A,i}) f_v(X_v^{C,i}) \middle| X^A_{\{1,\dots,d\}\setminus\{i\}}, X^C\right)\right) \\
&= \mathbb{E}\left(f_v(X_v^{C,i}) \int_0^1 f_u(X^{A,i}_{u\setminus\{i\}}, x_i) dx_i\right) \\
&= \mathbb{E}\left(f_v(X_v^{C,i}) 0\right) \\
&= 0.
\end{aligned}$$

Similarly, for $i \in v$,

$$\begin{aligned}
\mathbb{E}\left(f_u(X_u^{A,i}) f_v(X_v^{C,i})\right) &= \mathbb{E}\left(\mathbb{E}\left(f_u(X_u^{A,i}) f_v(X_v^{C,i}) \middle| X^C_{\{1,\dots,d\}\setminus\{i\}}, X^A\right)\right) \\
&= \mathbb{E}\left(f_u(X_u^{A,i}) \int_0^1 f_v(X^{C,i}_{v\setminus\{i\}}, x_i) dx_i\right) \\
&= \mathbb{E}\left(f_u(X_u^{A,i}) 0\right) \\
&= 0.
\end{aligned}$$

Hence

$$\begin{aligned}
\mathrm{Cov}\left(Y_i^A, Y_i^C\right) &= \sum_{\substack{u\in S \\ i\notin u}} \sum_{\substack{v\in S \\ i\notin v}} \mathbb{E}\left(f_u(X_u^{A,i}) f_v(X_v^{A,i})\right) \\
&= \sum_{\substack{u\in S \\ i\notin u}} \mathrm{Var}\left(f_u(X_u^{A,i})\right)
\end{aligned}$$

because of the decorrelation of the terms in the ANOVA decomposition. □

Consider then independent vectors

$$\begin{pmatrix} X^{A,i,(1)} \\ X^{C,i,(1)} \end{pmatrix}, \dots, \begin{pmatrix} X^{A,i,(n)} \\ X^{C,i,(n)} \end{pmatrix}$$

with the same distribution as

$$\begin{pmatrix} X^{A,i} \\ X^{C,i} \end{pmatrix}.$$

Let, for $j = 1, \dots, n$,

$$Y_i^{C,(j)} = f(X^{C,i,(j)}).$$

Then an estimator of TI_i is

$$\widehat{\mathrm{TI}}_i = \frac{1}{2n} \frac{\sum_{j=1}^n \left(Y_i^{A,(j)} - Y_i^{C,(j)}\right)^2}{\hat{\sigma}^2}$$