Computer experiments Exam

INSA Toulouse – ModIA – 2021 October

1) Authorized documents: a single handwritten A4 sheet (recto-verso).
2) Electronic devices (mobile phone, calculator, laptop, etc.) are not allowed.
3) All answers must be justified.

1 Gaussian process modeling [6 pts]

We want to improve the metamodeling of the Ishigami function

$$(x_1, x_2, x_3) \mapsto \sin(x_1) + A\sin^2(x_2) + Bx_3^4 \sin(x_1)$$

with a Gaussian process (GP), by adding extra information. More generally, we consider a 3-dimensional function defined on $\mathbb{R}^3$, of the form:

$$f(x) = f_2(x_2) + f_{1,3}(x_1, x_3)$$

where $x = (x_1, x_2, x_3)$. We also assume that $f_{1,3}$ is an odd function with respect to x_1 :

$$f_{1,3}(-x_1, x_3) = -f_{1,3}(x_1, x_3) \qquad \text{for all } x_1, x_3 \in \mathbb{R}. \tag{1}$$

1. Let us first consider the part $f_{1,3}(x_1, x_3)$. Let Z_0 a centered GP defined on $\mathbb{R}^2$, with kernel k_0. Define:

$$Z(x_1, x_3) = Z_0(x_1, x_3) - Z_0(-x_1, x_3).$$

(a) [1 pt] Give an example of expression for k_0. How can you construct it from 1-dimensional kernels?

A usual choice for a 2-dimensional kernel is to construct if as a tensor product of 1-dimensional kernels. For instance,

$$k_0((x_1, x_3), (x_1', x_3')) = \exp\left(-\frac{(x_1 - x_1')^2}{\theta_1^2}\right) \exp\left(-\frac{(x_3 - x_3')^2}{\theta_3^2}\right)$$

(b) [2 pts] Check that the trajectories of Z satisfy Equation (1). Prove that Z is a Gaussian process, by considering linear combinations.
We have indeed $Z(-x_1, x_3) = Z_0(-x_1, x_3) - Z_0(x_1, x_3) = -Z(x_1, x_3)$.
Now consider a linear combination extracted from Z at points $x^1, \ldots, x^n$. It is a linear combination extracted from Z_0 at points $x^1, \ldots, x^n$ and their symmetric with respect to the first coordinate, i.e. the points $(-x_1^i, x_3^i)$ $(i = 1, \ldots, n)$. Now Z_0 is a GP, so this linear combination is normally distributed. Hence Z is a GP.

(c) [2 pts] Prove that Z is centered, and compute its kernel $k_Z((x_1, x_3), (x_1', x_3'))$ in function of k_0.

As Z_0 is centered, Z is centered by additivity of the expectation.
By using the bilinearity of the covariance function, we have:

$$\begin{aligned} k_Z((x_1,x_3),(x_1',x_3')) &= k_0((x_1,x_3),(x_1',x_3')) + k_0((-x_1,x_3),(-x_1',x_3')) \\ &- k_0((x_1,x_3),(-x_1',x_3')) - k_0((-x_1,x_3),(x_1',x_3')) \end{aligned}$$

2. [1 pt] Finally, what kernel can you propose to model f, given k_0 and an extra 1-dimensional kernel k_2? What GP is corresponding to that kernel?
Let k_2 be another one-dimensional kernel. Then the kernel

$$k_2(x_2,x_2') + k_Z((x_1,x_3),(x_1',x_3'))$$

is valid, as a sum of kernel. We know that it corresponds to a GP

$$Y_2(x_2) + Z(x_1,x_3)$$

where Y_2 is a centered GP of kernel k_2, independent of Z. That GP has exactly the same form as f.

2 Gaussian process regression based on derivatives [4 pts]

Let $Y = (Y(t))_{t\in\mathbb{R}}$ be a centered Gaussian Process (GP) on $\mathbb{R}$ with kernel k. We denote by $Y'(t)$ the derivative of $Y(t)$ at t. We want to adapt kriging in order to account for information on derivatives.
We recall the formula for Gaussian conditioning: if (U_1,U_2) is a centered Gaussian vector $\mathcal{N}(0,\Sigma)$ with $\Sigma = \begin{pmatrix}\Sigma_{1,1} & \Sigma_{1,2}\\ \Sigma_{2,1} & \Sigma_{2,2}\end{pmatrix}$ then $U_2|U_1=u_1$ is a Gaussian vector with:

$$\begin{aligned} \mathbb{E}[U_2|U_1=u_1] &= \Sigma_{2,1}\Sigma_{1,1}^{-1}u_1 \\ \mathrm{Cov}[U_2|U_1=u_1] &= \Sigma_{2,2} - \Sigma_{2,1}\Sigma_{1,1}^{-1}\Sigma_{1,2} \end{aligned}$$

In all the exercise, we consider only one design point $x_1\in\mathbb{R}$ and one new point $x\in\mathbb{R}$ where to predict.

1. [1 pt] Let $s,t\in\mathbb{R}$. Express $\mathrm{Cov}(Y(s),Y'(t))$ with some derivative of k. Same question for $\mathrm{Cov}(Y'(s),Y'(t))$.

   Derivation is a linear operation; By using the bilinearity of the covariance, we have:
   $\mathrm{Cov}(Y(s),Y'(t)) = \frac{\partial k}{\partial t}(s,t)$ and $\mathrm{Cov}(Y'(s),Y'(t)) = \frac{\partial^2 k}{\partial s\partial t}(s,t)$.

2. [1 pt] We admit that the vector $(Y'(x_1),Y(x))$ is centered and Gaussian (which is due to the linearity of derivation). Write its covariance matrix.
   Notice that the vector is Gaussian because derivation is linear and Y is a Gaussian process (hence any linear combination extracted from Y and Y' is normally distributed). It is centered by linearity of the expectation. By applying the result of the previous question, its covariance matrix is:

$$\Sigma = \begin{pmatrix} \frac{\partial^2 k}{\partial s\partial t}(x_1,x_1) & \frac{\partial k}{\partial s}(x_1,x) \\ \frac{\partial k}{\partial t}(x,x_1) = \frac{\partial k}{\partial s}(x_1,x) & k(x,x)\end{pmatrix}$$

3. [1 pt] Deduce the expression below for the kriging mean and kriging variance accounting for derivatives, i.e. the distribution of $Y(x)$ knowing that $Y'(x_1)=d_1$:

$$\begin{aligned} \mathbb{E}[Y(x)|Y'(x_1)=d_1] &= d_1\frac{\partial k}{\partial s}(x_1,x)/\frac{\partial^2 k}{\partial s\partial t}(x_1,x_1) \\ \mathrm{Var}[Y(x)|Y'(x_1)=d_1] &= k(x,x) - \frac{\partial k}{\partial s}(x_1,x)^2/\frac{\partial^2 k}{\partial s\partial t}(x_1,x_1) \end{aligned}$$

   This is a direct application of the formula of GP conditioning with $U_1 = Y'(x_1)$, and $U_2 = Y(x)$.

4. [1 pt] What can you say about the dependence on d_1? From what property does it come from?
   As for usual kriging, this kriging mean formula depends linearly on d_1, and this kriging variance does not depend on d_1. These properties come from Gaussian conditioning.

3 Sensivity analysis (6 pts)

Let us consider the 2-dimensional function:

$$f(x_1, x_2) = x_1 + x_2 + x_1 x_2$$

The aim is to perform a global sensitivity analysis of $f(X_1, X_2)$ where X_1, X_2 are independent uniform random variables, with $X_1 \sim \mathcal{U}\left[-\frac{a}{2}, \frac{a}{2}\right]$ and $X_2 \sim \mathcal{U}[-\frac{1}{2}, \frac{1}{2}]$, and $a > 0$.

We recall that for a uniform random variable $Z \sim \mathcal{U}[s, t]$, we have $\mathbb{E}(Z) = \frac{s+t}{2}$ and $\mathrm{Var}(Z) = \frac{(t-s)^2}{12}$.

1. [2 pts] Show that the ANOVA decomposition of $f(X_1, X_2)$ is simply:

   $$f_0 = 0, \quad f_1(X_1) = X_1, \quad f_2(X_2) = X_2, \quad f_{1,2}(X_1, X_2) = X_1 X_2$$

   We immediately have $\mathbb{E}(X_1) = \mathbb{E}(X_2) = 0$, and by independence $\mathbb{E}(X_1 X_2) = \mathbb{E}(X_1)\mathbb{E}(X_2) = 0$, proving the centering conditions. Furthermore $\mathbb{E}(X_1 X_2 | X_1) = X_1 \mathbb{E}(X_2 | X_1) = X_1 \mathbb{E}(X_2) = 0$, and similarly $\mathbb{E}(X_1 X_2 | X_2) = 0$, proving the non-simplification conditions. By unicity, the ANOVA decomposition of $f(X_1, X_2)$ is then given by:

   $$f_0 = 0, \quad f_1(X_1) = X_1, \quad f_2(X_2) = X_2, \quad f_{1,2}(X_1, X_2) = X_1 X_2.$$

2. [1.5 pts] Compute the partial variances $D_I = \mathrm{Var}(f_I(X_I))$ for $I = \{1\}, \{2\}, \{1, 2\}$ and check that the global variance is $D = \mathrm{Var}(f(X_1, X_2)) = \frac{1}{12}\left(1 + \frac{13}{12}a^2\right)$
   We immediately have $D_1 = \mathrm{Var}(X_1) = \frac{a^2}{12}$ and $D_2 = \mathrm{Var}(X_2) = \frac{1}{12}$.
   Further, using once again the independence of X_1, X_2, we get:

   $$D_{1,2} = \mathrm{Var}(X_1 X_2) = \mathbb{E}(X_1^2 X_2^2) = \mathbb{E}(X_1^2)\mathbb{E}(X_2^2) = \mathrm{Var}(X_1)\mathrm{Var}(X_2) = \frac{a^2}{12} \times \frac{1}{12}$$

   Hence $D = D_1 + D_2 + D_{1,2} = \frac{1}{12}\left(1 + \frac{13}{12}a^2\right)$.

3. [1 pts] Recall that Sobol indices are defined by $S_I = D_I / D$. Compute S_1 and S_2, and check that S_1 (resp. S_2) is an increasing (resp. decreasing) function of a. Interpretation?
   We have $S_1 = \frac{a^2}{1 + \frac{13}{12}a^2} = \frac{1}{\frac{13}{12} + \frac{1}{a^2}}$ which is an increasing function of a. Similarly, $S_2 = \frac{1}{1 + \frac{13}{12}a^2}$ is a decreasing function of a. This is rather logical: Increasing the uncertainty of X_1 increases the importance of X_1 in the output $f(X_1, X_2)$ and reduces the importance of X_2.

We now assume that X_1, X_2 are independent random variables, with $X_1, X_2 \sim \mathcal{U}[0, 1]$.

4. [1,5 pts] By looking at the assumptions of the ANOVA decompositions, explain why in this new situation we *cannot* have $f_1(X_1) = X_1$. Compute $f_1(X_1)$.

We cannot have $f_1(X_1) = X_1$ since for instance $\mathbb{E}(X_1) = \frac{1}{2} \neq 0$. Now we have:

$$\begin{aligned}
f_0 &= \mathbb{E}(f(X)) = \mathbb{E}(X_1) + \mathbb{E}(X_2) + \mathbb{E}(X_1)\mathbb{E}(X_2) = \frac{5}{4} \quad (X_1 \text{ and } X_2 \text{ independent}) \\
f_1(X_1) &= \mathbb{E}(f(X)|X_1) - f_0 \\
&= X_1 + \mathbb{E}(X_2|X_1) + X_1\mathbb{E}(X_2|X_1) - \frac{5}{4} \\
&= X_1 + \mathbb{E}(X_2) + X_1\mathbb{E}(X_2) - \frac{5}{4} \quad (X_1 \text{ and } X_2 \text{ independent}) \\
&= \frac{3}{2}(X_1 - \frac{1}{2})
\end{aligned}$$

4 Bayesian optimization (4 pts)

We want to tune the parameters of a time-consuming algorithm with Bayesian optimization (BO).

1. [2 pts] Write the principle of BO for optimizing a time-consuming function, and write a pseudo-code. To what function can we apply BO here?
   For the first part, we refer to the slides on BO.
   For the second part, we refer to the computer labs: we can apply BO to a cross-validation error criterion.

2. [2 pts] Assume that one parameter is discrete. Some people continue to use BO with the usual squared exponential kernel. Is that reasonable? In your response, consider the two situations: if the parameter is ordinal (e.g. a number of layers in a neural network) or not (e.g. type of kernel in SVM).
   In 1 dimension, the usual squared exponential kernel is based on a distance on $\mathbb{R}$. When the discrete variable is ordinal, a distance between the values (levels) of the variable can make sense (e.g. between 2 layers and 5 layers), but much less when it is non ordinal: what is the distance between the level 'polynomial kernel' and 'sigmoid kernel' in SVM? In theory, there are specific kernels for discrete variables that should be preferred. In the ordinal case, such kernels also account for non-stationarities: the distance between levels 2 and 3 is not necessarily the same than betweens levels 10 and 11.