Statistical modeling and its applications Exam

Mines St-Etienne – 21st January 2015

No document allowed except an A4 sheet with hand written remarks from your own hand

Exercise 1 (5 pts)

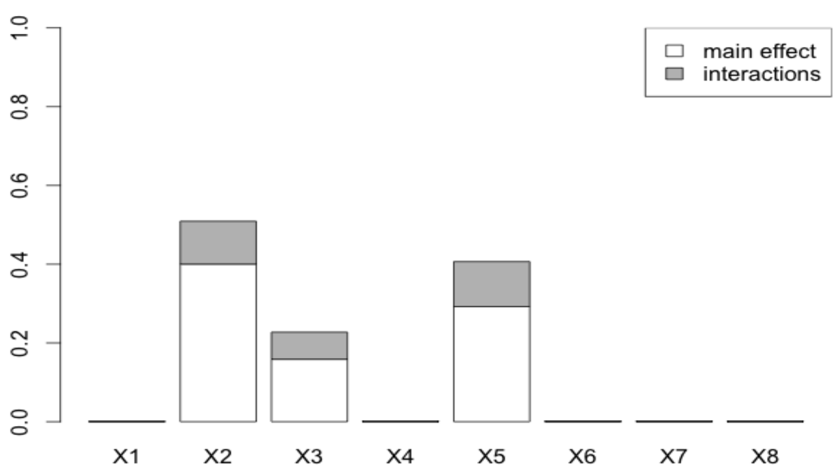
We consider the function $f(x_1, x_2) = e^{x_1} \times x_2$. The aim is to perform a global sensitivity analysis of $f(X_1, X_2)$ when X_1, X_2 are random variables following the uniform distribution on $[-1/2, 1/2]$.

1. [0.5 pt] Show that $E(X_2) = 0$. We denote by $m_1 := E(e^{X_1})$ (do not compute it).
2. [2 pts] Derive the Sobol-Hoeffding decomposition of $f(X_1, X_2)$.
3. [2,5 pts] We denote $v_1 = \text{var}\left(e^{X_1}\right)$, $v_2 = \text{var}(X_2)$. (Again: do not compute them).
   (a) Express the variance D of $f(X_1, X_2)$ as a function of m_1, v_1 and v_2.
   (b) Deduce from Question 2 that the Sobol indices of the main effects are given by: $S_1 = 0$, $S_2 = \frac{m_1^2}{m_1^2 + v_1}$.
   (c) Deduce, with a simple argument, the expression of the Sobol index of the interaction, $S_{1,2}$.

Exercise 2 (3 pts)

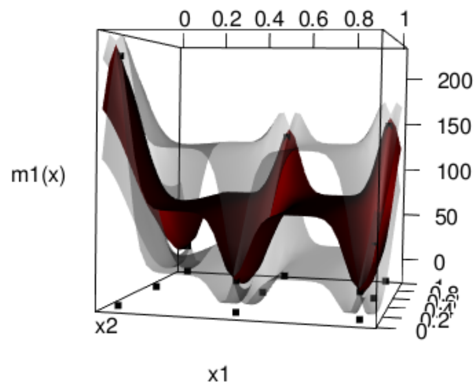
We consider a metamodel depending on 8 input variables $X_1, ..., X_8$. In the figure below, we show the result of a global sensitivity analysis on it.

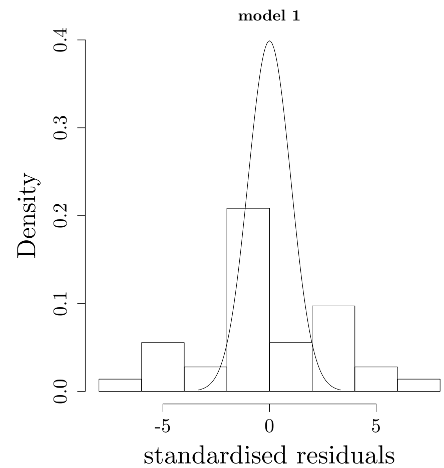
1. [1 pt] Interpret the results: What are the most influential input variables, the inactive ones? Is it true that there is no interaction between X_2 and X_4 ?
2. [2 pts] Explain how you can draw the main effect $x_2 \mapsto \mu_2(x_2)$ from N simulations of $X_1, ..., X_8$? Make a figure.

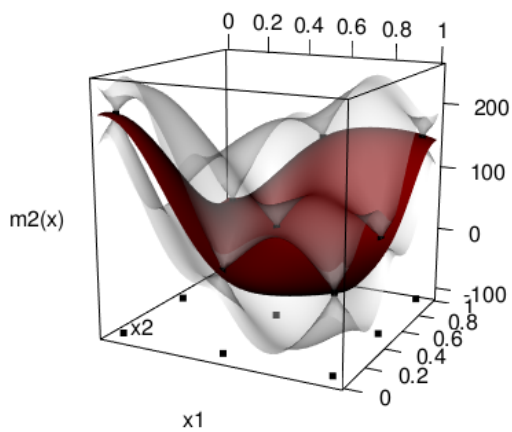


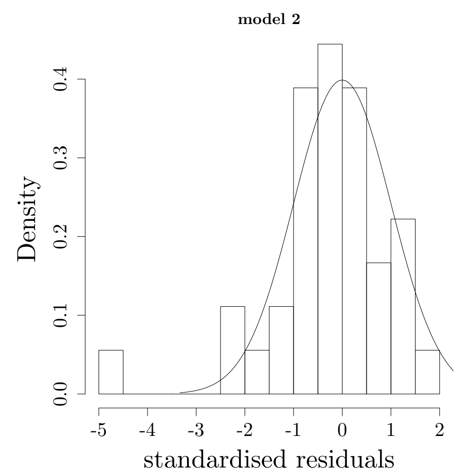
Exercise 3 (5 pts)

We are interested in two Gaussian process regression models based on 9 observations of a 2-dimensional function (see below). The only difference between the models is in the choice of the kernel parameters. A test set of 36 points is introduced to asses the models quality. The Q_2 criteria of m_1 and m_2 are respectively 0.43 and 0.69.





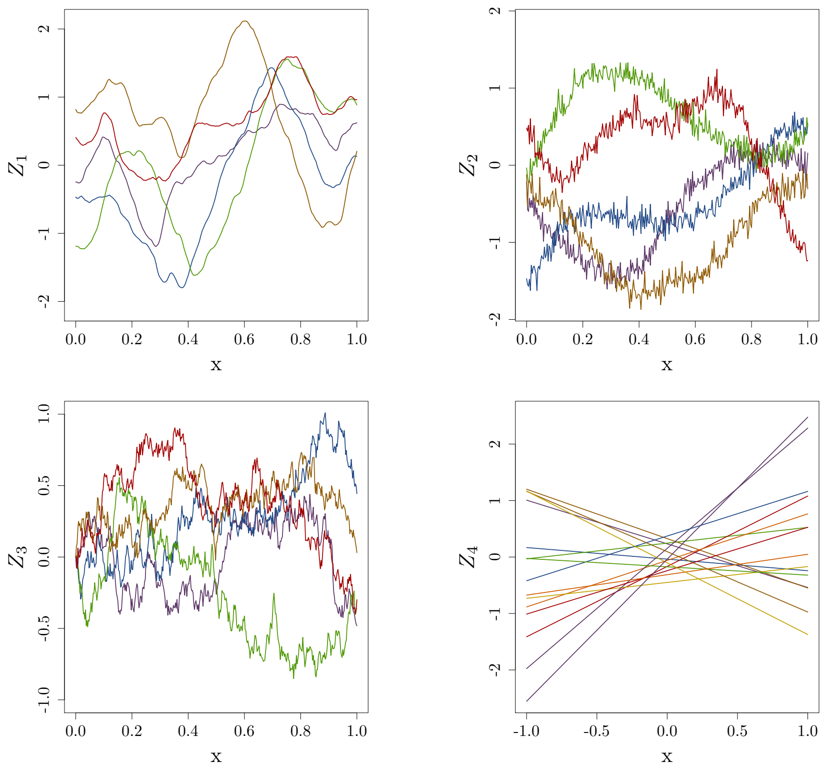




1. [1 pt] Give the name and expression of a kernel that may have been used.
2. [0.5 pt] For one model, the lengthscales are $(\theta_1, \theta_2) = (0.05, 0.5)$ and for the other they are $(\theta_1, \theta_2) = (0.25, 0.25)$. Find which is which and justify your answer.
3. [1.5 pt] How are the standardised residuals computed?
4. [1 pts] Which model seems to be the best ? Justify your answer.
5. [1 pt] Is there a trend in the models? If yes, is it possible to say if it is known or if it has been estimated?

Exercise 4 (2 pts)

Suggest a kernel (or a combination of usual kernels) for the following centred Gaussian process samples:



Exercise 5 (3 pts)

ANOVA kernels over $\mathbb{R}^d \times \mathbb{R}^d$ are kernels of the form: $k_{anova}(x, y) = \sigma^2 \prod_{i=1}^{d} (1 + k_i(x_i, y_i))$, where the k_i can be any kernel over $\mathbb{R} \times \mathbb{R}$.

1. [1 pt] Prove that k_{anova} is a valid covariance function.
2. [1 pt] Write down the expression of a Gaussian process regression model based on k_{anova} and show that it can be seen as the sum of 2^d submodels. Show that the submodels can be interpreted as a conditional Gaussian process.
3. [1 pt] Give a condition on the k_i such that the submodel mean predictors correspond to the terms of the Sobol-Hoeffding decomposition of the mean predictor based on k_{anova}.

Exercise 6 (2 pts)

Let $\mathcal{H}$ be a RKHS of functions over $[0, 1]$ such that the derivative evaluation in 0 : $L : f \mapsto f'(0)$ is continuous.

1. [1 pt] Show that the Riesz theorem applies and compute the representer of L.
2. [1 pt] What is the reproducing kernel of the subspace $\mathcal{H}_0 = \{f \in \mathcal{H} \text{ such that } \frac{\mathrm{d}f}{\mathrm{d}x}(0) = 0\}$?