Bayesian approach for the detection of inactive variables in Gaussian process approximation

Enikő Bartók
Université Paris–Saclay, CNRS, CentraleSupélec, Laboratoire des Signaux et Systèmes, France & IFP Énergies Nouvelles, France

Miguel Munoz Zuniga
Nicolas Bonfils
IFP Énergies Nouvelles, France

Emmanuel Vazquez
Université Paris–Saclay, CNRS, CentraleSupélec, Laboratoire des Signaux et Systèmes, France

Gaussian Processes (GPs) are recognized for their effectiveness as metamodels of numerical simulators [6]. They offer a Bayesian framework for supervised learning, allowing the incorporation of prior knowledge about a function through suitable kernel selection [7].

A widely used kernel in GP modeling is the anisotropic Matérn covariance function [7], which can be written as

$$k_{\nu,\sigma,\rho}(x,y) := \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} h_\rho\right)^\nu K_\nu \left(\sqrt{2\nu} h_\rho\right), \quad \text{with } h_\rho = \left(\sum_i^d \frac{(x_i - y_i)^2}{\rho_i^2}\right)^{1/2},$$

and where Γ is the Gamma function, and K_ν is the modified Bessel function of the second kind. The parameters $\nu \in \mathbb{R}^+$, $\sigma \in \mathbb{R}^+$ and $\rho = (\rho_1, \ldots, \rho_d) \in \mathbb{R}^{+d}$ are usually selected using the maximum likelihood approach (see, e.g., [4]). This covariance function is known for its ability to model functions with different degrees of smoothness and variable correlations across different dimensions.

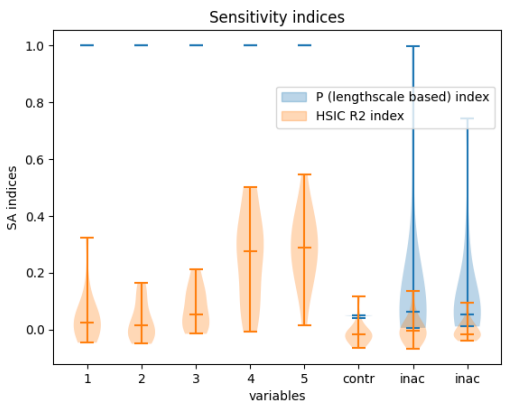
Building upon this framework, our work focuses on identifying inactive variables—those with no influence on the function output—in settings where the number of active variables is small (e.g., fewer than 20) but the overall dimensionality is large (e.g., greater than 50). Specifically, we consider functions $f : \mathbb{R}^d \to \mathbb{R}$, for which there exists function of k inputs, $g : \mathbb{R}^k \to \mathbb{R}$, such that:

$$f(\mathbf{x}) = g(x_{(1)}, x_{(2)}, \ldots, x_{(k)}), \quad \mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d \text{ and } \{(1), ..., (k)\} \subset \{1, ..., d\}.$$

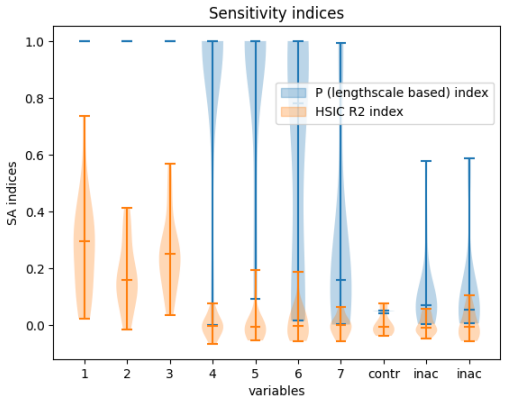
To sequentially identify inactive variables and reduce dimensionality using GPs, a common first idea is to use sensitivity analysis, such as in the work of Marrel et al. [3], where GPs are combined with HSIC (Hilbert-Schmidt Independence Criterion) indices to assess variable importance. Another approach, as demonstrated by Salem et al. [5], relies on the lengthscale parameters $\rho_1, \ldots, \rho_d$ of the GP covariance function $k_{\nu,\sigma,\rho}$. In this method, large values of a lengthscale parameter indicate slow variation of the output with respect to the corresponding variable, signifying that the variable is likely inactive.

Our method builds on the latter approach, relying on the lengthscale parameters and adopting a fully Bayesian framework (see, e.g., [1]). We generate samples from the posterior distribution of the lengthscale parameters using a Metropolis-Hastings algorithm. The main idea of the proposed approach is to introduce an inactive control variable x_{d+1}, which allows us to establish a reference posterior density for the lengthscale parameters of inactive variables. To determine whether a given variable is active, a significance level α is first fixed, and a threshold t_α is computed such that the posterior probability $\mathbb{P}_n(\rho_{d+1} > t_\alpha) \geq 1 - \alpha$, where ρ_{d+1} is the lenghscale parameter of the control variable x_{d+1}. Then, we introduce indices $P_i = \mathbb{P}_n(\rho_i \leq t_\alpha)$, which reflect the probability that the variable x_i is active.

Initial comparisons between our method and R^2_{HSIC} indices [2] demonstrate promising results (see, e.g., Figure 1).



(a) $f_1(\mathbf{x}) = \frac{1}{3}\sum_{i=1}^{5}\left(x_{(i)} + 2.2\prod_{i<j=2}^{5} x_{(i)}x_{(j)}\right)$



(b) $f_2(\mathbf{x}) = 6x_{(1)} + 4x_{(2)} + 5.5x_{(3)} + 3x_{(1)}x_{(2)} + 2.2x_{(1)}x_{(3)} + 1.4x_{(2)}x_{(3)} + x_{(4)} + 0.5x_{(5)} + 0.2x_{(6)} + 0.1x_{(7)}$

Figure 1: Distributions of the lengthscale-based indices P_i (blue, $\alpha = 5\%$) and the R^2_{HSIC} indices [2] (orange), providing a comparison of variable importance, with a focus on distinguishing between active $(1, \ldots, k)$, control (contr), and inactive (inac) variables, for 20 repetitions of random uniform designs of size $n = 30$. Functions f_1 and f_2 have $k = 5$ and $k = 7$ active variables, respectively, within an overall dimension of $d = 50$. Two randomly selected inactive variables from the set of $d - k$ are also represented.

References

[1] R. Benassi, J. Bect, and E. Vazquez. Robust gaussian process-based global optimization using a fully bayesian expected improvement criterion. In *Learning and Intelligent Optimization: 5th Int. Conf., LION 5, Rome, Italy*. Springer, 2011.

[2] S. Da Veiga. Global sensitivity analysis with dependence measures. *J. Statistical Computation and Simulation*, 85(7):1283–1305, 2015.

[3] A. Marrel, B. Iooss, and V. Chabridon. The ICSCREAM methodology: Identification of penalizing configurations in computer experiments using screening and metamodel—applications in thermal hydraulics. *Nuclear Science and Engineering*, 196(3):301–321, 2022.

[4] S. J. Petit, J. Bect, P. Feliot, and E. Vazquez. Parameter selection in gaussian process interpolation: an empirical study of selection criteria. *SIAM/ASA J. on Uncertainty Quantification*, 11(4):1308–1328, 2023.

[5] 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*, 7(4):1369–1397, 2019.

[6] T. J. Santner, B. J Williams, and W. I.r Notz. *The Design and Analysis of Computer Experiments*. Springer, 2018.

[7] M. L. Stein. *Interpolation of spatial data: some theory for kriging*. Springer, 1999.

[Enikő Bartók; IFPEN — Université Paris–Saclay — Laboratoire des Signaux et Systèmes]
[`eniko.bartok@ifpen.fr` –]

[Emmanuel Vazquez; Université Paris–Saclay — Laboratoire des Signaux et Systèmes]
[`emmanuel.vazquez@centraleusupelec.fr` –]