Fully Bayesian approach for the estimation of (first-order) Sobol indices
Benoit Jan, Julien Bect and Emmanuel Vazquez
SUPELEC, Gif-sur-Yvette, France

Context
We consider the problem of estimating the first-order Sobol indices of a function that represents the output of a computer code. It is well known that Monte Carlo estimators of Sobol indices require many evaluations of the computer code (see, e.g., [1]). When running the code is time- or resource-consuming, it has become common practice [2, 4] to use a metamodel. A natural question is then to quantify the error of approximation of the Sobol indices that is made when using the metamodel instead of the computer code. We focus in this work on the case of a Gaussian process-based (kriging) metamodel [3, 4], where the posterior distribution of the Sobol indices provides an elegant answer to the above concern. Algorithms for drawing samples from this posterior distribution have been proposed in [3, 4]. We argue that the use of a plug-in approach for the parameters of the covariance function is a dangerous practice in this setting. We propose to adopt instead a fully Bayesian approach.

First-order Sobol index
Let \( X \) be a rectangle of \( \mathbb{R}^d \), and consider a random variable \( X \) uniformly distributed on \( X \). The first-order Sobol indices of a function \( f : X \subset \mathbb{R}^d \to \mathbb{R} \) are defined by
\[
S_i(f) = \frac{\text{var}(f(X_1, \ldots, X_i, \ldots, X_d))}{\text{var}(f(X))}, \quad i = 1, \ldots, d.
\]
This quantity can be approximated by several Monte-Carlo estimators (see, e.g., [1]).

Gaussian process, posterior distributions
When \( f \) is the output of a time- or resource-consuming computer code, MC estimators become too expensive.

Classical idea: we put a prior about \( f \) in the form of a Gaussian Process (GP) \( \xi \), characterized by a covariance function with a parameter vector \( \theta \).

Allowed the estimation of the posterior distributions of Sobol indices based on a few evaluations \( \mathcal{F}_n \) of \( f \) (by calculating the Sobol indices of conditional sample paths).

Frequentist coverage properties of credible intervals
Evaluation of the frequentist coverage probabilities of several Bayesian credible intervals (i.e. the proportion of times when these intervals include the true values of the Sobol indices, when \( f \) varies in a certain class of functions).

Frequentist probabilities estimated from the estimation of the posterior distributions of Sobol indices of 1500 sample paths of a given Gaussian process.

Observations chosen on maximum LHS designs
Restricted Maximum Likelihood (REML) vs Posterior Mean (PM) vs Fully Bayesian approach (FB).

Choice of the covariance parameters
- Two possible approaches
  - Plug-in approach: a single parameter \( \theta \) is used to simulate all the sample paths (\( \theta \) is assumed to be known, or is estimated by maximum likelihood or cross-validation, for instance).
  - Fully Bayesian approach: a prior \( p_{\theta} \) on the covariance parameter is chosen, and for a random sample \( \xi_1, \ldots, \xi_N \) from the posterior distribution \( p_{\theta} \), this is used to simulate the sample paths.

The fully Bayesian approach makes it possible to take into account the uncertainty about \( \theta \) for the estimation of the posterior distribution of Sobol indices.

Results on a benchmark 2D function
Let \( g \) be the function of Sobol, defined on \([0, 1]^d\) by
\[
g(x) = \prod_{j=1}^{d} \left( 1 + a_j x_j \right)^{n_j} a_0 > 0, \forall x,
\]
(a classic benchmark function for Sobol indices estimation) for which Sobol indices can be obtained in a closed form.

- Estimation of the posterior distributions of the Sobol indices of \( g \) using a plug-in approach (REML estimation) and a fully Bayesian approach.

Results on a 7-dimensional function
- We consider a costly computer code simulating the yield of a power converter, with seven numerical inputs.
- Estimation of Sobol indices based on ten simulations of the code chosen on a maximum latin hypercube; the restricted maximum likelihood estimate (REML) was used for the plug-in approach.

Conclusions
- The fully Bayesian approach gives more conservative results than the plug-in one when the number of observations \( n \) is small; both approaches give similar results when the number of observations is getting moderately large.
- The posterior probabilities of the credible intervals extracted from the posterior distributions for both approaches are consistent with their frequentist coverage probabilities when \( n \) is large.
- Questions that could be addressed in the future: how sensitive is the fully Bayesian approach to the choice of the prior \( p_{\theta} \) for the estimation of Sobol indices? Can we propose a "default" prior well fitted for this application?

References