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 $\mathcal{X}$ be a rectangle of $\mathbb{R}^d$, and consider a random variable X uniformly distributed on $\mathcal{X}$.

The first-order Sobol indices of a function $f : \mathcal{X} \subset \mathbb{R}^d \to \mathbb{R}$ are defined by

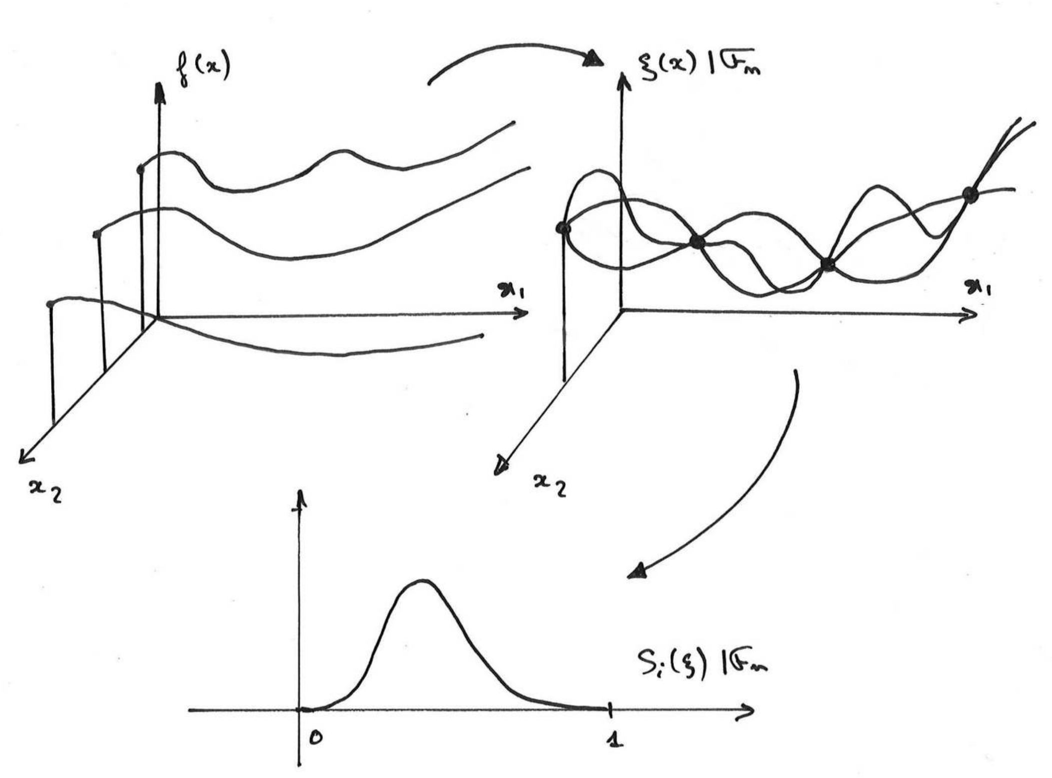
$$S_i(f) = \frac{\mathsf{var}[\mathsf{E}(f(X)|X_i)]}{\mathsf{var}[f(X)]}, \quad i = 1, \ldots, d.$$

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) ξ, characterized by a covariance function with a parameter vector θ
- Allows 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)



Choice of the covariance parameters

- Two possible approaches
  - Plug-in approach: a single parameter θ is used to simulate all the sample paths (θ is assumed to be known, or is estimated by maximum likelihood or cross-validation, for instance)
  - Fully Bayesian approach: a prior π_0 on the covariance parameter is chosen, and a random sample $\theta_1, \ldots, \theta_J$ from the posterior distribution π_n of θ is used to simulate the sample paths
- The fully Bayesian approach makes it possible to take into account the uncertainty about θ for the estimation of the posterior distribution of Sobol indices

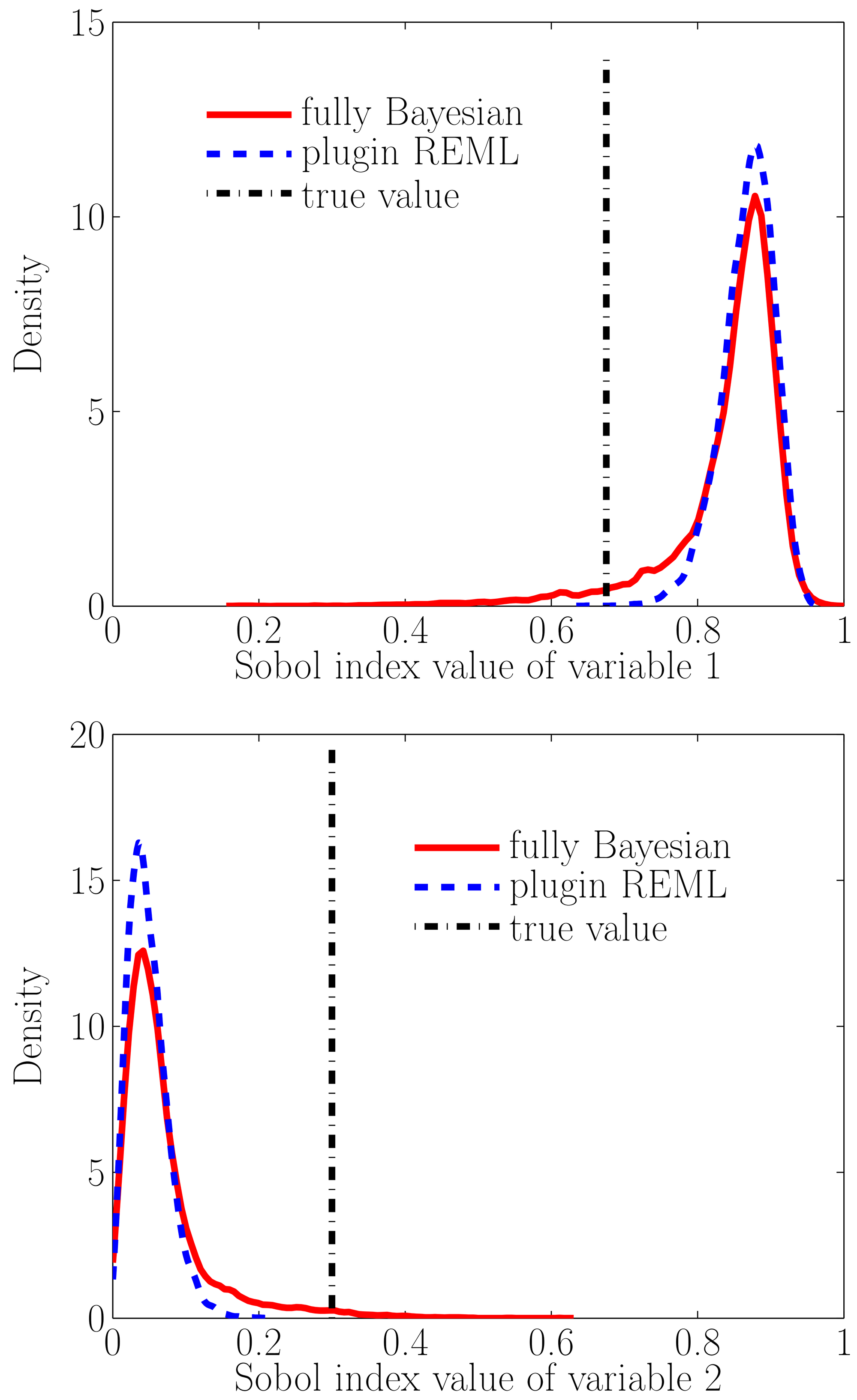
Results on a benchmark 2D function

Let g be the function of Sobol, defined on $[0,1]^d$ by

$$g(x) = \prod_{k=1}^{d} \frac{|4x_k - 2| + a_k}{1 + a_k}, \quad a_k > 0, \forall k,$$

(a classic benchmark function for Sobol indices estimation) for which Sobol indices can be obtained is 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:



Posterior distributions of the first-order Sobol indices of g based on 10 evaluations (LHS)
(g function of Sobol with parameters $a_1 = 1$ and $a_2 = 2$, top figure = first input, bottom figure = second input)

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 maximin LHS designs
- Restricted Maximum Likelihood (REML) vs Posterior Mean (PM) vs Fully Bayesian approach (FB)

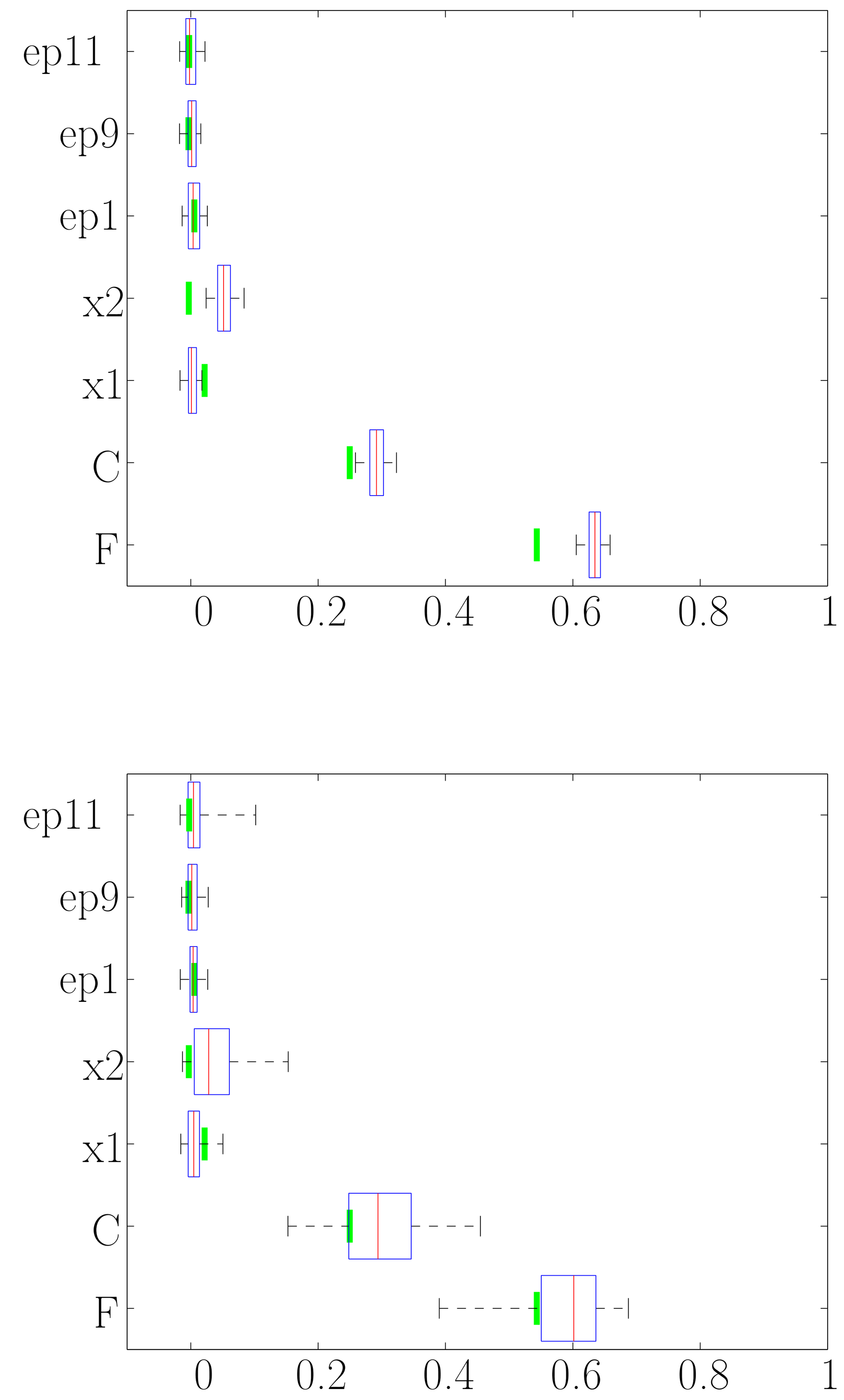
n	10			40		
approach	PM	REML	FB	PM	REML	FB
CI 95%	71.6	76.0	88.9	94.0	94.2	94.5
CI 90%	65.5	70.0	83.1	88.5	88.8	89.1
CI 80%	55.9	60.8	73.4	78.1	78.2	79.1
CI 50%	33.0	36.6	45.0	48.6	50.0	50.3

Frequentist coverage probabilities of several credible intervals in the case of a 2-dimensional Gaussian process
(isotropic Matérn covariance function on $[0,1]^2$, with $\sigma^2 = 1$, $\nu = 5/2$, and $\rho = 1$)

Analysis. The fully Bayesian approach is more conservative than the two plug-in ones (it gives larger credible intervals) when n is small. The frequentist probabilities of the different credible intervals estimated for both the plug-in and the fully Bayesian approaches are similar when n is large, and are close to the posterior probabilities of those intervals.

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 maximin latin hypercube; the restricted maximum likelihood estimate (REML) was used for the plug-in approach



Sobol indices distributions on a 7D function
(output of a computer code simulating the yield of a power converter
top figure = plug-in approach, bottom figure = fully Bayesian approach)
green bars = estimations by Monte-Carlo directly on f

Conclusions

- The fully Bayesian approach gives more conservative results than the plug-in one when the number of observations 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 π_0 for the estimation of Sobol indices? Can we propose a "default" prior well fitted for this application?

References

[1] Alexandre Janon, Thierry Klein, Agnès Lagnoux, Maëlle Nodet, and Clémentine Prieur. Asymptotic normality and efficiency of two sobol index estimators. http://hal.inria.fr/hal-00665048/, February 2012.

[2] Alexandre Janon, Maëlle Nodet, and Clémentine Prieur. Uncertainties assessment in global sensitivity indices estimation from metamodels. *International Journal for Uncertainty Quantification*, May 2012.

[3] Amandine Marrel, Bertrand Iooss, Béatrice Laurent, and Olivier Roustant. Calculations of sobol indices for the gaussian process metamodel. *Reliability Engineering and System Safety*, 94(3):742–751, 2009.

[4] Curtis B. Storlie, Laura P. Swiler, Jon C. Helton, and Cedric J. Sallaberry. Implementation and evaluation of nonparametric regression procedures for sensitivity analysis of computationally demanding models. *Reliability Engineering and System Safety*, 94(11):1735–1763, 2009.