MASCOT-NUM 2022
Clermont Ferrand, France
7-9 June 2022

Sensitivity to statistical estimation uncertainties and probabilistic model identification

C. SURGET
Université Clermont Auvergne - SIGMA Clermont - ONERA

Supervisor(s): S. Dubreuil (ONERA), J. Morio (ONERA), C. Mattrand (SIGMA Clermont), J-M. Bourinet (SIGMA Clermont) and N. Gayton (SIGMA Clermont)

PhD expected duration: May 2021 - May 2024

Address: ONERA - Centre de Toulouse, 2 Av. Edouard Belin, 31000 Toulouse

E-mail: charles.surget@onera.fr

Abstract: Let us consider a black-box function $\phi : \mathcal{X} \to \mathbb{R}$. Its input $\mathbf{X}$ is a continuous random vector, defined on the domain $\mathcal{X} \subseteq \mathbb{R}^d$, drawn according to a given probability density function (PDF) $f_{\mathbf{X}}$. Its output Y is a random variable such that $\mathrm{Y} = \phi(\mathbf{X})$. An example of application is a finite element analysis with random geometrical or material parameters and whose output would be a stress level, a displacement or any other quantity of interest (QoI). One could be interested in assessing a given expectation of a particular function τ of Y such as a mean or a probability of failure. The actual expectation $\mathbb{E}_{f_{\mathbf{X}}}$ can be estimated by Monte Carlo Simulation (MCS):

$$\mathbb{E}_{f_{\mathbf{X}}}\left[\tau\left(\phi(\mathbf{X})\right)\right] = \int_{\mathcal{X}} \tau\left(\phi\left(\mathbf{x}\right)\right) f(\mathbf{x}) \mathrm{d}\mathbf{x} \tag{1}$$

$$\approx \frac{1}{N_{\mathbf{X}}} \sum_{j=1}^{N_{\mathbf{X}}} \tau\left(\phi\left(\mathbf{X}^{(j)}\right)\right). \tag{2}$$

In an industrial context, the knowledge of the input probability distribution would often be restricted to a $N_{\mathbf{D}}$-sample, with $N_{\mathbf{D}}$ assumed to be small. This database $\tilde{\mathbf{D}} := \left(\mathbf{D}^{(1)}, ..., \mathbf{D}^{(N_{\mathbf{D}})}\right)$ is defined on the domain $\mathcal{D} \subseteq \mathbb{R}^{d \times N_{\mathbf{D}}}$ and gathers $N_{\mathbf{D}}$ random vectors $\mathbf{D}^{(i)}$ that are independent and identically distributed (i.i.d.) from a true, but unknown, PDF $f_{\mathbf{X}}$. Realizations of $\mathbf{D}^{(i)}$ are assumed to be obtained from expensive experimental tests. Thus, the expectation in equation (1) can be expressed as follows:

$$\mathbb{E}_{\hat{f}_{\mathbf{X}|\tilde{\mathbf{D}}}}\left[\tau\left(\phi(\mathbf{X})\right)\right] = \int_{\mathcal{X}} \tau\left(\phi\left(\mathbf{x}\right)\right) \hat{f}_{\mathbf{X}|\tilde{\mathbf{D}}}(\mathbf{x}) \mathrm{d}\mathbf{x}, \tag{3}$$

where $\mathbf{X}^{(j)} \overset{i.i.d.}{\sim} \hat{f}_{\mathbf{X}|\tilde{\mathbf{D}}}$ and $\hat{f}_{\mathbf{X}|\tilde{\mathbf{D}}}$ is the joint PDF estimate of the true PDF $f_{\mathbf{X}}$. The estimator in equation (2) is then subject to two epistemic uncertainty sources. The first one comes from the identification of the joint PDF whereas the second one comes from the MCS sample. The originality of the proposed approach is to take into consideration both of these statistical sources of uncertainty. Nonetheless, the difficulty lies in their dependence.

The main objective here is to assess the variance of the given estimator with respect to those two sources of uncertainty. This variance can be reduced by increasing the data and MCS sample sizes, respectively $N_{\mathbf{D}}$ and $N_{\mathbf{X}}$, i.e. by performing new mechanical tests and running new simulations. Experimental tests and black-box evaluations however have a non-negligible cost for complex industrial applications. Hence, it is desirable to determine how the budget should be allocated: is it better to add new tests or new simulations ?

The proposed work focuses on the test-simulation trade-off, succeeding G. Sarazin's work [1]. The novelty is to tackle this issue by means of a sensitivity analysis in order to determine the contribution of each uncertainty source in the variance of the QoI. A Bootstrap resampling method is here used to generate

new databases, which serve in turns to estimate several PDFs $\hat{f}_{\mathbf{X}|\tilde{\mathbf{D}}}$. This allows us to deal with the statistical uncertainty whilst estimating the PDF $f_{\mathbf{X}}$. Sobol' indices [2] are then computed and a Pick-Freeze approach [3] is used in order to reduce the computational burden. The relevance of the method is illustrated on academic examples.

References

[1] Gabriel Sarazin. *Analyse de sensibilité fiabiliste en présence d'incertitudes épistémiques introduites par les données d'apprentissage.* PhD thesis, Toulouse, ISAE, 2021.

[2] Ilya M Sobol. Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and computers in simulation*, 55(1-3):271–280, 2001.

[3] Fabrice Gamboa, Alexandre Janon, Thierry Klein, A Lagnoux, and Clémentine Prieur. Statistical inference for sobol pick-freeze monte carlo method. *Statistics*, 50(4):881–902, 2016.

Short biography – Charles Surget obtained a master's degree in mechanical engineering from SIGMA Clermont in 2021 and then started his PhD thesis in May 2021. The research work is funded by Université Clermont Auvergne and ONERA. The framework of the thesis is to consider a bi-level of uncertainties obtained from database identifications and Monte Carlo Simulations. This issue is of a major concern in an industrial context. The goal is to suggest an approach able to propose a well-chosen test-simulation trade-off.