Analyse de sensibilité aux incertitudes d'estimation statistique et à l'identification d'un modèle probabiliste

Charles Surget[a,b], Sylvain Dubreuil[a], Jérôme Morio[a], Cécile Mattrand[b], Jean-Marc Bourinet[b], Nicolas Gayton[b]

[a] *ONERA/DTIS, F-31055 Toulouse, France*
[b] *SIGMA Clermont, Institut Pascal, F-63000 Clermont–Ferrand, France*

Keywords: Small-data, Analyse de sensibilité, Compromis essai-simulation

Au sein des algorithmes traitant de la quantification d'incertitudes, le système étudié est bien souvent modélisé par une fonction boîte noire ϕ où l'entrée est un vecteur aléatoire continu $\mathbf{X}$, de densité $f_{\mathbf{X}}$, à valeurs dans $\mathbb{R}^d$ et la sortie est une variable aléatoire Y à valeurs dans $\mathbb{R}$ telle que $\mathrm{Y} = \phi(\mathbf{X})$. Un intérêt spécifique est accordé à l'espérance d'une fonction particulière de Y telle qu'une moyenne ou une probabilité de défaillance. Ces espérances peuvent être estimées par approche Monte-Carlo, telle que $\mathbb{E}\left[\phi\left(\mathbf{X}\right)\right]$ dans l'équation suivante :

$$\widehat{\mathbb{E}}\left[\phi(\mathbf{X})\right] = \frac{1}{N_{\mathbf{X}}} \sum_{i=0}^{N_{\mathbf{X}}} \phi\left(\mathbf{X}_i\right) \text{ avec } \mathbf{X}_i \sim f_{\mathbf{X}}.$$

Dans un cadre réaliste industriel, la connaissance de la loi de probabilité de l'entrée $\mathbf{X}$ peut être restreinte à un $N_{\mathbf{D}}$-échantillon, i.e. une base de données, noté $\widetilde{\mathbf{D}}$ et défini tel que $\widetilde{\mathbf{D}} := (\mathbf{D}_1, ..., \mathbf{D}_{N_{\mathbf{D}}})$ est à valeurs dans $\mathbb{R}^{d \times N_{\mathbf{D}}}$. Les vecteurs aléatoires $\mathbf{D}_i$ sont indépendants et identiquement distribués de densité $f_{\mathbf{X}}$ inconnue et $N_{\mathbf{D}}$ est supposé limité. L'estimateur $\widehat{\mathbb{E}}\left[\phi(\mathbf{X})\right]$ subit alors deux sources d'incertitudes épistémiques : une première est issue de l'identification de la densité jointe $f_{\mathbf{X}}$ tandis qu'une seconde provient de l'estimation de Monte-Carlo suite à un échantillonnage selon la loi identifiée. L'originalité de cette étude réside dans la dépendance des deux sources d'incertitudes au sein d'un contexte *small-data*. Ainsi, dans l'objectif de réduire la variance de l'estimateur, est-il préférable d'augmenter le nombre d'essais, i.e. la base de données, ou bien d'augmenter le nombre de réalisations générées selon la loi identifiée, i.e. le nombre de simulations de Monte-Carlo ? C'est dans cette démarche de compromis essai-simulation que s'inscrivent ces travaux de thèse, succédant au travail de G. Sarazin [1]. La singularité de l'approche proposée consiste à modéliser cette question comme un problème d'analyse de sensibilité globale afin de déterminer les parts de variance de l'estimateur liées à l'échantillon de données initial et à celle de l'échantillon Monte-Carlo. Les variables d'entrée du vecteur aléatoire $\mathbf{X}$ sont supposées indépendantes. Un ré-échantillonnage par *Bootstrap* est pratiqué afin de générer la variabilité d'inférence des densités marginales. Les indices de Sobol' [2] sont alors évalués pour déterminer les contributions respectives des deux sources d'incertitudes sur la variance de l'estimateur. La pertinence de la méthode est illustrée sur des exemples académiques et une approche Pick-Freeze [3] est proposée pour réduire le nombre de calculs.

References

[1] SARAZIN, Gabriel. Analyse de sensibilité fiabiliste en présence d'incertitudes épistémiques introduites par les données d'apprentissage. 2021. Thèse de doctorat. Institut Supérieur de l'Aéronautique et de l'Espace Toulouse.

[2] SOBOL', Ilya M. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Mathematics and computers in simulation, 2001, vol. 55, no 1-3, p. 271-280.

[3] GAMBOA, Fabrice, JANON, Alexandre, KLEIN, Thierry, et al. Statistical inference for Sobol pick-freeze Monte Carlo method. Statistics, 2016, vol. 50, no 4, p. 881-902.