méthodologie basée sur les mesures de dépendance HSIC pour l'analyse de sensibilité de second niveau

A. Meynaoui, A. Marrel, B. Laurent-Bonneau

MÉTHODOLOGIE BASÉE SUR LES MESURES DE DÉPENDANCE HSIC POUR L'ANALYSE DE SENSIBILITÉ DE SECOND NIVEAU

Anouar MEYNAOUI [1], Amandine MARREL [1] & Béatrice LAURENT-BONNEAU [2]

[1] *CEA, DEN, DER, SESI, LEMS, 13108 Saint-Paul-lez-Durance, France*
anouar.meynaoui@cea.fr,
amandine.marrel@cea.fr

[2] *INSA Toulouse, France*
laurentb@insa-toulouse.fr

Résumé. Nous nous intéressons à l'analyse de sensibilité des simulateurs numériques dans le cas où les distributions de probabilités des variables d'entrées incertaines sont elles-mêmes méconnues. L'objectif est de quantifier l'impact de ces incertitudes sur les résultats de l'analyse de sensibilité. Pour cela, on propose une méthodologie de type simple boucle Monte-Carlo basée sur les mesures de dépendance de Hilbert-Schmidt et inspirée des techniques de tirage d'importance, cette approche permettant de limiter significativement le nombre d'évaluations du simulateur. Une application numérique est proposée pour illustrer l'ensemble de la méthodologie et tester ses différentes options.

Mots-clés. Analyse de sensibilité, Mesures de dépendance, Critère d'indépendance de Hilbert-Schmidt.

Abstract. We are interested in the sensitivity analysis of numerical simulators in the case where the probability distributions of the uncertain input variables are themselves (even partially) unknown. The objective is to quantify the impact of these uncertainties on the sensitivity analysis results. To achieve it, we propose a single Monte Carlo loop methodology based on Hilbert-Schmidt dependence measures and importance sampling techniques. This approach significantly limits the number of simulator evaluations. A numerical application is proposed to illustrate the whole methodology, while comparing its different options.

Keywords. Sensitivity analysis, Dependence measures, Hilbert-Schmidt independence criterion.

1 Introduction

Les simulateurs numériques (ou codes de calculs) sont essentiels pour comprendre, modéliser et prévoir des phénomènes physiques. Ces simulateurs numériques prennent en entrée un grand nombre de paramètres caractéristiques du phénomène étudié. Ces paramètres peuvent être entachés d'une certaine incertitude. La sortie est donc à son tour entachée

d'une incertitude. Il est donc important de considérer non seulement les valeurs nominales des paramètres d'entrée, mais aussi l'ensemble des valeurs possibles dans le domaine de variation de chaque paramètre (Rocquigny et al., 2008). L'objectif est alors d'évaluer l'impact des incertitudes des entrées sur la variabilité de la sortie. Dans le cadre d'une approche probabiliste, les paramètres d'entrée $X_1, ..., X_d$ et la sortie Y sont considérés comme des variables aléatoires et leurs incertitudes sont modélisées par des distributions de probabilités respectivement notées $\mathbb{P}_{X_1}, ..., \mathbb{P}_{X_d}$ et $\mathbb{P}_Y$. On se place dans le cadre de variables aléatoires d'entrées indépendantes continues caractérisées par des densités de probabilités (ou *pdf* pour *probability density function*).

1.1 Analyse de Sensibilité Globale (ASG) et mesures HSIC

L'Analyse de Sensibilité Globale (Iooss, 2011) notée ASG, vise à déterminer la façon dont la variabilité globale des paramètres en entrée influe sur la valeur de la sortie. Il s'agit plus particulièrement d'identifier et éventuellement quantifier, pour chaque paramètre d'entrée X_k, sa contribution à la variabilité de la sortie Y. L'ASG permet ainsi de séparer les entrées en deux groupes : celles qui influencent considérablement la valeur de la sortie (que l'on appelle aussi variables significatives) et celles dont l'influence sur la sortie peut être négligée. Cette séparation des variables d'entrée en deux groupes est appelée criblage (ou screening).

Parmi les méthodes d'ASG pour les simulateurs numériques, on s'intéresse ici plus particulièrement à celles basées sur des mesures de dépendance, récemment proposées pour l'ASG par Da Veiga (2015). Ces mesures ont pour objectif de quantifier d'un point de vue probabiliste, la dépendance entre la sortie Y et chaque paramètre X_k en entrée. On s'intéresse en particulier au critère d'indépendance de Hilbert-Schmidt (Gretton et al., 2005), noté HSIC pour Hilbert Schmidt Independence Criterion et donné pour tout $k \in \{1, ..., d\}$ par la formule suivante :

$$\begin{aligned} \mathrm{HSIC}(X_k, Y) = \mathbb{E}\left[l_k(X_k, X_k')l(Y, Y')\right] - 2\mathbb{E}\left[\mathbb{E}\left[l_k\left(X_k, X_k'\right) \mid X_k\right]\mathbb{E}\left[l\left(Y, Y'\right) \mid Y\right]\right] \\ + \mathbb{E}\left[l_k\left(X_k, X_k'\right)\right]\mathbb{E}\left[l\left(Y, Y'\right)\right], \end{aligned} \tag{1}$$

où X_k' est une copie indépendante et identiquement distribuée de X_k et Y' est la sortie associée à $\mathbf{X}' = (X_1', ..., X_d')$. Enfin, l et l_k sont des objets mathématiques appelés noyaux reproduisants universels (Aronszajn, 1950) associés respectivement à X_k et Y.

Cette mesure de dépendance qui généralise la notion de covariance entre deux variables aléatoires permet de capturer un très large spectre de formes de dépendance entre les variables et caractérise l'indépendance des variables sous l'hypothèse de noyaux universels. En pratique, les mesures HSIC sont estimées par approche Monte-Carlo. Elles présentent l'avantage d'avoir un faible coût d'estimation (quelques centaines de simulations pour une dizaine d'entrées) et leur estimation pour l'ensemble des entrées ne dépend

pas du nombre d'entrées. Des travaux récents de De Lozzo et Marrel (2016) ont aussi montré l'efficacité de ces mesures HSIC pour réaliser un criblage des variables d'entrée, en associant à ces mesures différents tests statistiques de significativité. Enfin, les mesures HSIC peuvent être étendues à des entrées non scalaires (vectorielles, fonctionnelles ou encore catégorielles).

Dans le cadre de l'ASG, Da Veiga (2015) propose des indices de sensibilité (compris entre 0 et 1) directement dérivés des mesures HSIC. Ces indices permettent de classer les variables d'entrée $X_1, ..., X_d$ par ordre d'influence sur la sortie Y. Ces indices notés $\mathrm{R}^2_{\mathrm{HSIC},k}$ sont définis pour tout $k \in \{1, ..., d\}$ par :

$$\mathrm{R}^2_{\mathrm{HSIC},k} = \frac{\mathrm{HSIC}(X_k, Y)}{\sqrt{\mathrm{HSIC}(X_k, X_k)}\sqrt{\mathrm{HSIC}(Y, Y)}}. \tag{2}$$

Une autre approche pourrait être d'utiliser les p-valeurs[1] des tests statistiques de significativité pour hiérarchiser les variables d'entrées.

1.2 Analyse de Sensibilité Globale de second niveau (ASG2)

Dans certaines applications, les lois de probabilités $\mathbb{P}_{X_1}, ..., \mathbb{P}_{X_d}$ caractérisant les entrées incertaines $X_1, ..., X_d$ du simulateur peuvent elles-mêmes être incertaines. Cette incertitude peut être liée à une divergence d'avis d'expert sur la loi de probabilité à affecter à chacune des entrées ou encore une absence (ou une quantité limitée) d'information pour caractériser cette loi. Dans le cadre d'une démarche probabiliste, ces incertitudes sont modélisées par des lois des probabilités $\mathbb{P}_{\mathbb{P}_{X_1}}, ..., \mathbb{P}_{\mathbb{P}_{X_d}}$ sur un ensemble de lois de probabilités des entrées. Cette incertitude sur les lois des entrées peut considérablement modifier les résultats de l'ASG réalisée par HSIC ou par une autre mesure de dépendance. **Il est alors important d'identifier et quantifier l'impact de l'incertitude sur les lois des entrées sur les résultats de l'ASG, on parle alors d'ASG de 2nd niveau, notée ASG2.** Les résultats d'ASG2 pourront, par la suite, être utilisés pour évaluer si la caractérisation des lois des entrées doit être améliorée et prioriser les efforts de caractérisation sur les entrées dont la loi influe le plus sur les résultats d'ASG.

2 Méthodologie proposée pour l'ASG2

2.1 Construction d'indices de type HSIC de 2nd niveau

La réalisation d'une ASG2 soulève plusieurs problématiques et verrous techniques. Tout d'abord, il est nécessaire de "caractériser" les résultats d'une ASG afin de comparer les

[1]La p-valeur d'un test statistique d'indépendance est la probabilité que, sous l'hypothèse testée (ici l'indépendance), la statistique du test soit supérieure ou égale à la valeur observée sur les données.

résultats de l'ASG obtenus pour différentes lois en entrée. Cette caractérisation consiste à associer à chaque jeu de distributions $\mathbb{P}_{\mathbf{X}} = \mathbb{P}_{X_1} \times ... \times \mathbb{P}_{X_d}$ des entrées, une quantité mesurable $\mathcal{R}$ représentative des résultats de l'ASG. Pour choisir cette quantité d'intérêt, on propose les options suivantes :

- **Classement des entrées $X_1, ..., X_d$ basé sur les indices $\mathrm{R}^2_{\mathrm{HSIC},1}, ..., \mathrm{R}^2_{\mathrm{HSIC},d}$.** Dans ce cas, la quantité d'intérêt $\mathcal{R}$ est une permutation sur l'ensemble $\{1, ..., d\}$, qui vérifie que $\mathcal{R}(k) = j$ si et seulement si la variable X_j est la k-ième dans le classement.

- **Vecteur $R^2_{\mathrm{HSIC}} = (R^2_{\mathrm{HSIC},1}, ..., \mathrm{R}^2_{\mathrm{HSIC},d})$ regroupant les indices de sensibilité des entrées**. Dans ce cas, la quantité d'intérêt $\mathcal{R} = R^2_{\mathrm{HSIC}}$ est un vecteur de d composantes.

- **Vecteur des p-valeurs associé aux d tests d'indépendance par mesure HSIC.** Cette quantité d'intérêt $\mathcal{R}$ est donc un vecteur à d coordonnées, dans $[0, 1]^d$.

Une fois choisie la quantité d'intérêt $\mathcal{R}$ caractéristique de l'ASG, le but est de quantifier l'impact des incertitudes sur $\mathbb{P}_{X_1}, ..., \mathbb{P}_{X_d}$ (modélisées par $\mathbb{P}_{\mathbb{P}_{X_1}}, ..., \mathbb{P}_{\mathbb{P}_{X_d}}$) sur les résultats de l'ASG. Pour cela, on définit des indices de sensibilité mesurant la dépendance entre la quantité $\mathcal{R}$ et les lois $\mathbb{P}_{X_1}, ..., \mathbb{P}_{X_d}$ de chaque entrée : $\mathrm{HSIC}(\mathbb{P}_{X_k}, \mathcal{R}), k = 1...d$. Ces indices sont définis par une formule analogue à (1), grâce à des noyaux reproduisants universels étendus aux lois des entrées (Sriperumbudur et al., 2010) et à la quantité d'intérêt $\mathcal{R}$ en sortie [voir Jiao et Vert (2016) si $\mathcal{R}$ est une permutation et Gretton et al. (2005) si $\mathcal{R}$ est un vecteur]. On parle alors des mesures de dépendance de 2$^{\mathrm{nd}}$ niveau. On en déduit des indices de sensibilité sur le même principe que la formule (2).

2.2 Estimation par approche simple boucle

Pour estimer les d indices de sensibilité de 2$^{\mathrm{nd}}$ niveau, on doit disposer d'un échantillon $(\mathbb{P}^{(i)}_{\mathbf{X}}, \mathcal{R}^{(i)})_{1 \leq i \leq n_1}$ du couple $(\mathbb{P}_{\mathbf{X}}, \mathcal{R})$. Pour cela, on pourrait envisager une approche double boucle Monte-Carlo. La première boucle à n_1 itérations consiste à tirer aléatoirement à chaque itération i une loi $\mathbb{P}^{(i)}_{\mathbf{X}}$ des entrées. Pour calculer la quantité d'intérêt $\mathcal{R}^{(i)}$ associée à la loi $\mathbb{P}^{(i)}_{\mathbf{X}}$, on tire alors un échantillon $(X_1^{(i,j)}, ..., X_d^{(i,j)})_{1 \leq j \leq n_2}$ de taille n_2 des entrées suivant $\mathbb{P}^{(i)}_{\mathbf{X}}$. La seconde boucle consiste à calculer les n_2 sorties $Y^{(i,j)}$ pour $j \in \{1...n_2\}$, où chaque $Y^{(i,j)}$ est associée au vecteur des entrées $(X_1^{(i,j)}, ..., X_d^{(i,j)})$. Enfin, grâce à l'échantillon $\mathcal{E}^{(i)} = (X_1^{(i,j)}, ..., X_d^{(i,j)}, Y^{(i,j)})_{1 \leq j \leq n_2}$, on calcule la quantité d'intérêt $\mathcal{R}^{(i)}$.

Cette approche double boucle Monte-Carlo nécessite au total $n_1 n_2$ simulations du code. Par exemple, si $n_1 = 100$ et $n_2 = 1000$, le calcul des HSIC de 2$^{\mathrm{nd}}$ niveau nécessite un total de 10^5 appels au code. L'approche double boucle Monte-Carlo n'est donc pas envisageable pour des simulateurs coûteux en temps de calcul.

Pour réduire le coût d'estimation des mesures de dépendance HSIC de 2nd niveau (et donc le coût de l'ASG2), nous proposons une méthodologie simple boucle Monte-Carlo. Cette méthodologie est basée sur :

- **Le tirage d'un unique échantillon suivant une unique loi pour chaque entrée, dite loi de référence.** Dans cette étape, on tire un unique échantillon $\overline{\mathbf{X}} = \left(\mathbf{X}^{(j)}\right)_{1\leq j\leq n_2}$ suivant une loi de référence $\overline{\mathbb{P}}_{\mathbf{X}} = \overline{\mathbb{P}}_{X_1} \times ... \times \overline{\mathbb{P}}_{X_d}$. Pour choisir judicieusement cette loi par rapport à l'ensemble des lois de probabilités possibles pour les entrées, on propose trois possibilités : loi mélange et lois barycentriques au sens de la distance de Wasserstein (Bigot et al., 2016) ou de Kullback-Leibler symétrisée (Veldhuis, 2002).

- **La constitution de l'échantillon d'apprentissage.** Cette étape consiste à calculer, au moyen du code de calcul, les sorties $\left(Y^{(j)}\right)_{1\leq j\leq n_2}$ associées à l'échantillon $\overline{\mathbf{X}} = \left(\mathbf{X}^{(j)}\right)_{1\leq j\leq n_2}$. Un échantillon d'entrées/sorties noté $\mathcal{E} = \left(\mathbf{X}^{(i)}, Y^{(j)}\right)_{1\leq j\leq n_2}$ est alors obtenu.

- **La réalisation de plusieurs ASG à partir de $\mathcal{E}$ et d'estimateurs modifiés des mesures de dépendance HSIC.** L'objectif de cette étape est de réaliser les ASG associées aux lois $\mathbb{P}_{\mathbf{X}}^{(i)}, i = 1...n_1$, en utilisant seulement l'échantillon $\mathcal{E}$. L'idée est d'estimer les mesures HSIC (et/ou $\mathrm{R}^2_{\mathrm{HSIC}}$), associées à chaque distribution $\mathbb{P}_{\mathbf{X}}^{(i)}$ à l'aide de l'échantillon $\mathcal{E}$ généré suivant $\overline{\mathbb{P}}_{\mathbf{X}}$, loi différente de $\mathbb{P}_{\mathbf{X}}^{(i)}$ [voir par ex. Cannaméla (2007)]. À l'issue de cette étape, on obtient donc un échantillon $(\mathbb{P}_{\mathbf{X}}^{(i)}, \mathcal{R}^{(i)})_{1\leq i\leq n_1}$.

- **L'estimation de HSIC 2nd niveau.** Dans cette étape, les indices de sensibilité de 2nd niveau sont calculés grâce à l'échantillon $(\mathbb{P}_{\mathbf{X}}^{(i)}, \mathcal{R}^{(i)})_{1\leq i\leq n_1}$ obtenu à l'étape précédente.

3 Conclusion et Perspectives

Une première application de la méthodologie d'ASG2 proposée est réalisée sur un exemple analytique en dimension 3. Les différentes possibilités pour la loi de tirage unique sont testées et comparées. Les premiers résultats montrent que la loi barycentrique au sens de la distance de Kullback-Leibler donne les meilleurs résultats en termes de convergence des estimateurs HSIC de 2nd niveau et de hiérarchisation des distributions de probabilités incertaines. L'intérêt de l'approche simple boucle par rapport à l'approche double boucle est aussi illustré sur ce premier exemple numérique. L'approche simple boucle permet une estimation bien plus précise des HSIC de 1er niveau et, en conséquence, des HSIC (et $\mathrm{R}^2_{\mathrm{HSIC}}$) de 2nd niveau.
Dans la continuité de ces premiers travaux, la méthodologie d'ASG2 développée sera

appliquée à un cas test simulant un accident de perte de débit primaire non protégé pour le réacteur de quatrième génération ASTRID refroidi au sodium (Advanced Sodium Technological Reactor for Industrial Demonstration).

Il serait aussi pertinent dans la suite de ces travaux d'améliorer certains éléments de la méthodologie proposée. En particulier, l'utilisation de plans d'expériences présentant des propriétés optimales de remplissage de l'espace, plans de type *space filling*, permettrait d'améliorer la convergence des estimateurs. L'étude du lien entre la loi de référence (pour tirer l'échantillon unique) et la paramétrisation des HSIC de 2^{nd} niveau (choix du noyau sur les densités de probabilité) pourrait aussi être approfondie, afin d'optimiser la précision des HSIC de 2^{nd} niveau.

Bibliographie

[1] Aronszajn, N. (1950). Theory of Reproducing Kernels. Transactions of the American Mathematical Society, 68(3).

[2] Bigot, J., Gouet, R., Klein, T. et López, A. (2016). Minimax convergence rate for estimating the Wasserstein barycenter of random measures on the real line. Journal of Optimization Theory and Applications.

[3] Cannaméla, C. (2007). Apport des méthodes probabilistes dans la simulation du comportement sous irradiation du combustible à particules. Thèse de doctorat, Paris 7.

[4] Da Veiga, S. (2015). Global sensitivity analysis with dependence measures. Journal of Statistical Computation and Simulation, 85(7):12831305.

[5] De Lozzo, M. et Marrel, A. (2016). New improvements in the use of dependence measures for sensitivity analysis and screening. Journal of Statistical Computation and Simulation, 86(15):30383058.

[6] De Rocquigny, E., Devictor, N. et Tarantola, S. (2008). Uncertainty in industrial practice : a guide to quantitative uncertainty management. John Wiley & Sons.

[7] Gretton, A., Bousquet, O., Smola, A. et Scholkopf, B. (2005). Measuring statistical dependence with Hilbert-Schmidt norms. In Algorithmic Learning Theory, Lecture Notes in Computer Science, volume 3734, pages 6378. Springer.

[8] Iooss, B. (2011). Revue sur l'analyse de sensibilité globale de modèles numériques. Journal de la Société Française de Statistique, 152(1):325.

[9] Jiao, Y. et Vert, J.-P. (2016). The Kendall and Mallows kernels for permutations. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[10] Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schlkopf, B. et Lanckriet, G. R. (2010). Hilbert space embeddings and metrics on probability measures. Journal of Machine Learning Research, 11:15171561.

[11] Veldhuis, R. (2002). The centroid of the symmetrical Kullback-Leibler distance. IEEE signal processing letters, 9(3):9699.