Analyse de sensibilité de variables dépendantes par les indices de Shapley

B. Broto (1), L. Clouvel (3), M. Depecker (1), F. Bachoc (2), JM. Martinez (3)

(1) CEA/DRT, (3) CEA/DEN, (2) Inst. Math. Toulouse.

GdR MaNu - 7 décembre 2018

1 Introduction

- Typologie des incertitudes en simulation numérique
- Analyse de sensibilité globale par modélisation probabiliste des incertitudes
- Cas de variables indépendantes - Indices de Sobol

2 Prise en compte des dépendances

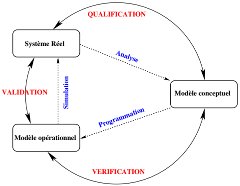
- Indices de Shapley
- Relations entre indices Sobol/Shapley (variables indépendantes)
- Application au modèle linéaire gaussien
- Application au modèle additif par blocs de variables indépendants
- Cas d'étude en physique des réacteurs

Incertitudes au coeur du processus de Vérification et Validation des Outils de Calcul Scientifique

MODÈLE CONCEPTUEL	
Aléatoire	**Epistémique**
• variabilité *naturelle* : régimes turbulents, géométrie d'une brèche, signal sismique • pas de valeur unique • irréductibles : générateurs aléatoires de variables, champs, processus	• méconnaissance : sections efficaces, modèles physiques, lois de fermeture, lois de comportement • on ne sait pas quelle est la bonne valeur, la *bonne* loi • réductibles : assimilation de données (mesures), calibration Bayésienne adaptée (*connaissances* a priori)

MODÈLE OPÉRATIONNEL := OCS	
Statistique	**Numérique**
• calculs Monte Carlo • réductibles : plus de simulations/particules, méthodes de réduction de variance	• méthodes, discrétisations, seuils de convergence • réductibles : finesse/adaptativité du maillage, de la convergence

- CONSENSUS sur la modélisation probabiliste des incertitudes aléatoires, statistiques et épistémiques (interprétation subjective) lorsqu'elles portent sur des constantes physiques
- CRITIQUES sur l'utilisation de variables aléatoires pour représenter les incertitudes épistémiques des modèles physiques
- VERROUS SCIENTIFIQUES : transposition (changements d'échelles géométrie/physique), interactions entre incertitudes épistémiques et numériques
- INTÉRÊT DE L'ANALYSE DE SENSIBILITÉ sur les modèles conceptuel et opérationnel



W. Oberkampf (2004)
Society of Computer Simulation (1979)

Incertitudes modélisées par des variables aléatoires

- ANALYSE DE SENSIBILITÉ : quantifier l'influence des paramètres, des variables incertaines (entrées) sur les grandeurs calculées (sorties) par l'OCS
  - aide à à la vérification du modèle opérationnel et à la validation (calibration) du modèle conceptuel
- Pour une MODÉLISATION PROBABILISTE des incertitudes, l'AS est obtenue par décomposition de la variance des grandeurs calculées par l'OCS sur la base des variables aléatoires
  - représentation de l'incertitude par la variance
- FORMALISATION
  - $X = (X_1, \ldots, X_p)$ un vecteur aléatoire réel modélisant les incertitudes
  - $Y = \varphi(X)$ la variable aléatoire à analyser associée à une des grandeurs calculées par l'OCS
- QUANTIFICATION de l'influence des variables X_i sur Y par des indices de sensibilité
  - variables X_i indépendantes $\rightarrow$ indices de Sobol
  - variables X_i dépendantes $\rightarrow$ indices de Shapley
- L'AS fondée sur l'ANOVA (ANalysis Of VAriance) utilise les ESPÉRANCES CONDITIONNELLES $\mathbb{E}(Y|X_{u\subseteq[1:p]})$ modélisant chacune la meilleure approximation de Y par une fonction de X_u, vecteurs de composantes $X_{i\in u}$:

$$\mathbb{E}(Y|X_u) = \arg\min_{f(X_u)} \mathbb{E}_{X_{-u}}[(Y - f(X_u))^2]$$

Indices de Sobol

- $Y = \varphi(X_1, \ldots, X_p)$, indépendance des X_i et $\mathbb{E}(Y^2) < \infty$
- Décomposition Fonctionnelle (Hoeffding)

$$Y = \sum_{u \subseteq [1:p]} \varphi_u(X_u)$$

- Unicité si $\varphi_\emptyset = \mathbb{E}(Y)$ et $\varphi_u(X_u)$ mutuellement non corrélées

$$\varphi_u(X_u) = \mathbb{E}(Y|X_u) - \sum_{v \subset u} \varphi_v(X_v)$$

  - $\varphi_u(X_u)$ fonction qui modélise l'effet dû aux interactions éventuelles entre les composantes de X_u
- Décomposition Fonctionnelle de la variance en notant $\sigma_u^2 = \mathrm{Var}(\varphi_u^2(X_u))$:

$$\mathrm{Var}(Y) = \sum_{u \subseteq [1:p]} \sigma_u^2 = \sigma^2$$

- Indices de Sobol (I. M. Sobol 2001) : S_u^i quantifie la part de la variance due aux interactions entre les composantes de X_u et l'indice dit complet S_u^c caractérise la part de la variance due au groupe X_u et S_u^t la part de la variance due au groupe X_u et aux interactions avec les autres variables X_{-u}

$$\begin{aligned}
S_u^i &= \frac{\mathrm{Var}(\varphi_u(X_u))}{\sigma^2} = \frac{\sigma_u^2}{\sigma^2} \Rightarrow \sum_{u \subseteq [1:p]} S_u^i = 1 \\
S_u^c &= \frac{\mathrm{Var}(\mathbb{E}(Y|X_u))}{\sigma^2} = \frac{\sum_{v \subseteq u} \sigma_v^2}{\sigma^2} \\
S_u^t &= \frac{\sum_{v \cap u \neq \emptyset} \sigma_v^2}{\sigma^2}, \quad \Rightarrow S_u^i + S_{-u}^t = 1
\end{aligned}$$

  - si Y a une forme additive : $\varphi(X) = \sum_{i=1}^p \varphi_i(X(i)) \Rightarrow S_{\{i\}}^i = S_{\{i\}}^c = S_{\{i\}}^t$
  - et si $Y = \varphi_u(X_u) + \varphi_{-u}(X_{-u}) \Rightarrow S_u^c + S_{-u}^c = 1$

Transposition des travaux de Shapley (théorie des jeux) à l'AS

- Dans la décomposition fonctionnelle, on perd l'orthogonalité entre les $\varphi_u(X_u)$
  - due à la dépendance entre les X_u
  - pas de formule explicite pour calculer les $\varphi_u(X_u)$
- Indices de Sobol généralisés issus des travaux de G. Chastaing (2013)
  - aux S_u on rajoute des termes de covariances
  - préservation de $\sum_{u\subseteq[1:p]} S_u = 1$
  - mais interprétation difficile car certains S_u peuvent être négatifs.
- Indices proposés par A. Owen (2014) à partir des travaux de L. S. Shapley (1953, théorie des jeux)
  - diffusés dans la communauté *incertitudes* (GdR MascotNum) par C. Prieur et B. Iooss (2017)
  - travaux actuels de B. Broto (thésard CEA) pour le cas paramétrique *linéaire gaussien* et dans le cas où on ne dispose que d'un n-échantillon i.i.d. (GdR MascotNum, soumis Mathematics and Computers in Simulation)
  - en neutronique, mise en oeuvre par L. Clouvel (thésarde CEA) sur les incertitudes sur les données nucléaires (fluence cuve, Physor 2018, M&C2019)

Indices de Shapley proposés par Owen pour les études en AS

- Evaluation des joueurs dans un jeu à p joueurs
- L'ensemble des coalitions/équipes $u \subseteq [1:p]$ au nombre $\leq 2^p$
- Fonction de gain $c(u)$ définie pour l'ensemble des coalitions avec $c(\emptyset) = 0$
- Valeur de Shapley η_i d'un joueur i est la valeur moyenne du gain qu'il apporte lorsqu'il intègre les équipes u constituées par les autres joueurs

$$\eta_i \quad = \quad \frac{1}{p} \sum_{u \subseteq -i} \binom{p-1}{|u|}^{-1} (c(u \cup \{i\}) - c(u)) \Rightarrow \sum_{i=1}^{p} \eta_i = c([1:p])$$

- Pour l'AS, A. Owen (2014) propose comme fonction de gain la part de la variance de Y apportée par la coalition X_u (comme Sobol)

$$c(u) \quad := \quad \frac{\mathrm{Var}(\mathbb{E}(Y|X_u))}{\mathrm{Var}(Y)} \Rightarrow \sum_{i=1}^{p} \eta_i = 1$$

- Complexité combinatoire : cas général, 2^p calculs/estimations du gain
  - reformulation du calcul des indices en considérant l'ensemble des permutations des joueurs $\rightarrow$ estimateurs par tirages Monte Carlo des permutations (E. Song)
  - solution analytique dans le cas linéaire gaussien, réduction du coût calculatoire en présence de groupes de variables indépendants (B. Broto)

Dans le cas de variables indépendantes

- Dans ce cas, l'indice de Shapley s'exprime en fonction des indices de Sobol

$$\eta_i = \sum_{u \subseteq [1:p] \,|\, u \ni i} \frac{S_u^i}{|u|}, \quad |u| = \texttt{cardinal de } u$$

    - lien avec l'indice total de Sobol $S_{\{i\}}^t = \sum_{u \subseteq [1:p] \,|\, u \ni i} S_u^i$ intégrant toutes les éventuelles interactions entre X_i et les autres variables (sans la pondération par $1/|u|$)

- La valeur de l'indice de Shapley se situe alors entre celle de l'indice de Sobol du 1er ordre et celle de l'indice de sensibilité totale :

$$S_{\{i\}}^i \leq \eta_i \leq S_{\{i\}}^t$$

- L'égalité entre les 3 indices est obtenue si le modèle est de forme additive :

$$\left(\begin{array}{ll} Y & = \sum_{i=1}^{p} \varphi_i(X_i) \\ X_i & \texttt{indépendantes} \end{array} \right) \Rightarrow S_{\{i\}}^i = \eta_i = S_{\{i\}}^t, \quad i = 1, \ldots, p$$

- Dans le cas général, par rapport aux indices de Sobol, les indices de Shapley ne permettent pas de distinguer entre la part due à la corrélation de celle due aux interactions (lorque le modèle est non linéaire)

Indices de Shapley d'un modèle linéaire gaussien

- Cas particulier du modèle linéaire gaussien

$$Y = \boldsymbol{\beta}^T X, \quad X \sim \mathcal{N}(\boldsymbol{\mu}, \Gamma)$$

- Expression analytique des indices de Shapley en fonction
  - des paramètres du modèle $\boldsymbol{\beta}^T$
  - de la matrice des variances, covariances Γ de la loi multivariée X

$$\eta_i = \frac{1}{p(\boldsymbol{\beta}^T \Gamma \boldsymbol{\beta})} \sum_{u \subseteq -i} \binom{p-1}{|u|}^{-1} [\boldsymbol{\beta}_{-u}^T (\Gamma_{-u,-u} - \Gamma_{-u,u} \Gamma_{u,u}^{-1} \Gamma_{u,-u}) \boldsymbol{\beta}_{-u} - \boldsymbol{\beta}_v^T (\Gamma_{v,v} - \Gamma_{v,u+i} \Gamma_{u+i,u+i}^{-1} \Gamma_{u+i,v}) \boldsymbol{\beta}_v]$$

- A noter la même complexité combinatoire : 2^p termes à évaluer
- Algorithme LG-Shapley (B. Broto)
  - scripts R, Scilab
  - intégrée dans URANIE (plate-forme incertitudes du CEA/DEN)

Réduction de la complexité combinatoire

- En présence plusieurs groupes de variables indépendants, une méthode (B. Broto) permet de réduire considérablement la complexité calculatoire dans le cas d'un modèle additif
- Soient p variables avec k groupes indépendants de $n_{j=[1:k]}$ variables. Partition $[1:p] = [u_1, u_2, \ldots, u_k]$

$$Y = \sum_{j=1}^{k} \varphi_j(X_{u_j})$$

- L'analyse de sensibilité des p variables se ramène à k analyses de sensibilité découplées

$$\begin{aligned} \eta_{i \in u_j}^{(j)} &= \texttt{indice de Shapley de } X_{i \in u_j} \texttt{ de } \varphi_j(X_{u_j}) \\ \eta_i &= \eta_i^{(j)} \frac{\mathrm{Var}(\varphi_j(X_{u_j}))}{\mathrm{Var}(Y)} \end{aligned}$$

- Le nombre d'évaluations se réduit considérablement à $\sum_j 2^{n_j}$ comparé à 2^p
- Exemple : pour 2 groupes de 10 variables, réduction du nombre de calculs $2 \times 2^{10}/2^{20} \simeq 2 \times 10^{-3}$

Interprétation des indices de Shapley sur un exemple à 2D

- Toute AS d'un modèle linéaire gaussien à 2D se ramène à l'AS du modèle

$$\begin{aligned} X &\sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right) \\ Y &= \beta_1 X_1 + \beta_2 X_2 \end{aligned}$$

- Valeurs des indices de Shapley

$$\eta_1 = \frac{\beta_1^2 + \rho\beta_1\beta_2 - \rho^2(\beta_1^2 - \beta_2^2)/2}{\beta_1^2 + \beta_2^2 + 2\rho\beta_1\beta_2}, \quad \eta_2 = 1 - \eta_1$$

- Les effets des sensibilités *locales* (β_1, β_2) s'estompent lorsque le coefficient de corrélation tend vers 1 :

$$\forall \beta_1, \beta_2 \quad \lim_{\rho \to 1} \eta_1 = \eta_2 = 0.5$$

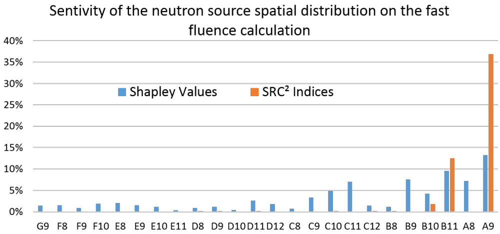
  - Si $\rho_{X_1, X_2} = 1$, l'AS par les indices de Shapley des 2 modèles ...

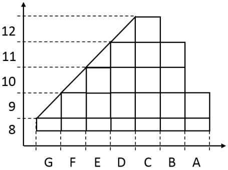
$$\begin{aligned} Y_1 &= 1 \times X_1 + 0 \times X_2 \\ Y_2 &= 0 \times X_1 + 1 \times X_2 \end{aligned}$$

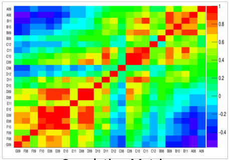
  - ... affectera la même valeur aux 2 indices de Shapley $\eta_1 = \eta_2 = 0.5$
- Par construction, deux variables fortement corrélées sont indiscernables par Shapley
  - déroutant mais logique ! !

Corrélations des incertitudes sur les données nucléaires

- Mise en oeuvre dans les études sur le vieillissement de la cuve des REP (L. Clouvel, CEA/SERMA)
  - Traitement de la dosimétrie du programme de surveillance des effets de l'irradiation (projet PRFLU)
  - Analyse de sensibilité du flux neutronique aux incertitudes sur des données nucléaires avec prise en compte de leurs corrélations
  - Incertitudes sur la distribution spatiale de la source neutronique représentées par une loi normale multivariée
  - Plan d'expériences numériques suivi d'une approximation linéaire validée $Y = \varphi(X) \simeq \mathbb{E}(Y) + \beta^T(X - \mathbb{E}(X))$






Correlation Matrix

- L. Clouvel, P. Mosca, J.M. Martinez, Uncertainty propagation of double-differential scattering cross section in fast fluence calculation for the PWR surveillance capsule, Physor2018

⇒ L. Clouvel, P. Mosca, J.M. Martinez, Shapley effects for sensitivity analysis of PWR power distribution in fast fluence calculation, M&C2019

- Etude menée via la plateforme Uranie (DEN) après intégration de l'algorithme LG-Shapley

Quelques références

B. Broto, F. Bachoc, M. Depecker, and J.M. Martinez, *Sensitivity indices for independent groups of variables*, hal-01680687 (2018).

N. Benoumecharia and K. Elie-Dit-Cosaque, *Shapley effects for sensitivity analysis with dependent inputs : bootstrap and kriging-based algorithms*, hal-01677501 (2018).

L. Clouvel, P. Mosca, and J.M. Martinez, *Uncertainty propagation of double-differential scattering cross section in fast fluence calculation for the pwr surveillance capsules*, april 2018.

B. Iooss and C. Prieur, *Analyse de sensibilité avec entrées dépendantes : estimation par échantillonnage et par métamodèles des indices de shapley*, 49èmes Journées de la SFdS, Avignon (2017).

______, *Shapley effects for sensitivity analysis with dependent inputs : comparaisons with sobol indices, numerical estimation and applications*, hal-01556303v3 (2018).

A. Owen and C. Prieur, *On shapley value for measuring importance of dependent inputs*, arXiv :1610.02080v3 (2017).

A. Owen, *Sobol's indices and shapley value*, SIAM/ASA Journal on Uncertainty Quantification **2(1)** (2014).

E. Song, B. Nelson, and J. Staum, *Shapley effects for global sensitivity analysis : theory and computation*, SIAM/ASA Journal on Uncertainty Quantification **4** (2016).