Analyses de Sensibilité

Airbus Central Research & Technology
regis.lebrun@airbus.com

3 février 2020

Analyse de sensibilité

Analyse de sensibilité

Rappel sur la propagation d'incertitudes

Etant donnés :

- Un vecteur aléatoire $\boldsymbol{X}$ prenant ses valuers dans $\mathbb{R}^n$ (les sources d'incertitude)
- Une fonction mesurable $g : \mathbb{R}^n \to \mathbb{R}^p$ (le modèle numérique)

On souhaite obtenir de l'information sur la distribution de $\boldsymbol{Y} = g(\boldsymbol{X})$ (d'où le terme de propagation des incertitudes) :

- des moments $\mathbb{E}[h(\boldsymbol{Y})]$ pour différentes fonctions d'intérêt h
- comme cas particulier, la probabilité d'événements particuliers $\mathbb{P}(\boldsymbol{Y} \in B)$

Trois questions importantes pour le concepteur :

- Hiérarchiser les sources d'incertitude vis-à vis de la distribution de $\boldsymbol{Y}$, plus précisément du critère d'intérêt sur cette distribution ;
- Réduire la dimension de l'espace des paramètres incertains ;
- Modifier la fonction g de manière informée si le critère sur $\boldsymbol{Y}$ n'est pas satisfaisant.

Analyse de sensibilité

Définition

L'indice de Sobol d'ordre k associé aux variables $(X_{i_1}, \ldots, X_{i_k})$ donne la part de variance de Y due à la meilleure approximation de Y comme fonction de $(X_{i_1}, \ldots, X_{i_k})$:

$$S_{i_1,\ldots,i_k} = \frac{\mathbf{Var}\left[\mathbb{E}\left[Y|X_{i_1}, \ldots, X_{i_k}\right]\right]}{\mathbf{Var}\left[Y\right]} \tag{1}$$

L'indice de Sobol total d'ordre k associé aux variables $(X_{i_1}, \ldots, X_{i_k})$ donne la part de variance de Y due aux contributions cumulées de tous les indices de Sobol associés à des groupes de variables ayant au moins une variable en commun avec $(X_{i_1}, \ldots, X_{i_k})$:

$$S^T_{i_1,\ldots,i_k} = \sum_I S_I, \quad I \subset \{1, \ldots, n\}, I \cap \{i_1, \ldots, i_k\} \neq \emptyset \tag{2}$$

Décomposition de Sobol d'une fonction intégrable sur $[0,1]^n$

Pour calculer (1) et (2), on utilise le résultat général dû à Sobol :
Si g est intégrable sur $[0,1]^n$, elle admet une unique décomposition du type :

$$g(x_1,\dots,x_n) = g_0 + \sum_{i=1}^{i=n} g_i(x_i) + \sum_{1\le i<j\le n} g_{i,j}(x_i,x_j) + \cdots + g_{1,\dots,n}(x_1,\dots,x_n) \tag{3}$$

où $g_0 = cst$ et les fonctions de la décomposition sont orthogonales entre elles par rapport à la mesure de Lebesgue sur $[0,1]^n$:

$$\int_{[0,1]^n} g_{i_1,\dots,i_s}(x_{i_1},\dots,x_{i_s}) g_{j_1,\dots,j_k}(x_{j1},\dots,x_{j_k}) d\boldsymbol{x} = 0 \tag{4}$$

dès lors que $(i_1,\dots,i_s) \neq (j_1,\dots,j_k)$.

Utiliser la décomposition de Sobol pour calculer les indices de Sobol

On aimerait décomposer le modèle g selon la décomposition de Sobol ... mais :

1. Les entrées de g ne sont pas sur $[0,1]^n$: dans le cas général, $Y = g(\boldsymbol{X})$ où $\boldsymbol{X}$ est défini sur $\mathbb{R}^n$.

$\Longrightarrow$ Si on pose

$$\boldsymbol{U} = (F_{\mathbf{1}}(X_{\mathbf{1}}), \dots, F_n(X_n))^t) = \phi^{-\mathbf{1}}(\boldsymbol{X}) \tag{5}$$

alors on montre que $\boldsymbol{U}$ a une loi jointe de marginales uniformes et de copule celle de $\boldsymbol{X}$.
En posant $Y = g(\boldsymbol{X}) = g \circ \phi(\boldsymbol{U})$, alors **on peut utiliser la décomposition de Sobol sur** $g \circ \phi$.

1. Les indices de Sobol par rapport aux U_i sont-ils les mêmes que ceux par rapport aux X_i ?

$\Longrightarrow$ Rappel : Si $\boldsymbol{U} = \psi(\boldsymbol{X})$ où ψ est un difféomorphisme et $Y = g(\boldsymbol{X})$ alors :

$$\mathbb{E}[Y|\boldsymbol{U}] = \mathbb{E}[Y|\boldsymbol{X}] \tag{6}$$

En effet : $\mathbb{E}[Y|\boldsymbol{U}] = \mathbb{E}[Y|\psi(\boldsymbol{X})]$ est le projeté orthogonal au sens $L^{\mathbf{2}}$ de Y sur l'espace engendré par $\psi(\boldsymbol{X})$, ie celui engendré par $\boldsymbol{X}$, d'où l'égalité des variables aléatoires (7).
Comme la transformation ϕ (5) agit composante par composante ($U_i \leftrightarrow X_i$) alors on a l'égalité des indicateurs de type :

$$\mathbf{Var}\left[\mathbb{E}\left[Y|U_{i_{\mathbf{1}}}, \dots, U_{i_k}\right]\right] = \mathbf{Var}\left[\mathbb{E}\left[Y|X_{i_{\mathbf{1}}}, \dots, X_{i_k}\right]\right] \tag{7}$$

d'où **l'égalité des indices de Sobol par rapport aux** U_i **et aux** X_i.

Interprétation probabiliste de la décomposition de Sobol

Supposons, sans perdre en généralité, que les variables X_i soient à support $[0,1]$. Alors en décomposant le modèle g selon la décomposition de Sobol (3), on écrit :

$$Y = g(\boldsymbol{X}) = g_0 + \sum_{i=1}^{i=n} g_i(X_i) + \sum_{1 \leq i < j \leq n} g_{i,j}(X_i, X_j) + \cdots + g_{1,\ldots,n}(X_1, \ldots, X_n) \tag{8}$$

On introduit les variables aléatoires $Z_i = g_i(X_i)$, $Z_{i,j} = g_{i,j}(X_i, X_j)$ etc. **La condition d'orthogonalité (4) des $g_{i_1,\ldots,i_k}$ par rapport à la mesure de Lebesgue sur $[0,1]^n$ se traduit par la décorrélation de ces variables si les X_i sont indépendantes, ce qu'on suppose désormais**. Ainsi Y se décompose sous la forme d'une somme :

$$\boldsymbol{Y = g(X) = g_0 + \sum_{i=1}^{i=n} Z_i + \sum_{1 \leq i < j \leq n} Z_{i,j} + \cdots + Z_{1,\ldots,n}} \tag{9}$$

où $Z_{i_1,\ldots,i_s} \perp Z_{j_1,\ldots,j_k}$ (ie $\mathbb{E}\left[Z_{i_1,\ldots,i_s} . Z_{j_1,\ldots,j_k}\right] = 0$).

Calcul des Indices de Sobol I

Grâce à la décomposition probabiliste (9), on calcule $\mathbb{E}[Y]$ et $\mathbf{Var}[Y]$ aisément :

$$\begin{cases} \mathbb{E}[Y] = Z_0 + \sum_{i=1}^{i=n} \underbrace{\mathbb{E}[Z_i]}_{=0 \text{ car } \perp 1} + \sum_{1\le i<j\le n} \underbrace{\mathbb{E}\left[Z_{i,j}\right]}_{=0 \text{ car } \perp 1} + \cdots + \underbrace{\mathbb{E}[Z_{1,\ldots,n}]}_{=0 \text{ car } \perp 1} \\ \mathbb{E}\left[Y^2\right] = \sum_{I\neq J} \underbrace{\mathbb{E}[Z_I Z_J]}_{=0 \text{ car } Z_I \perp Z_J} + \sum_I \mathbb{E}\left[Z_I^2\right], \quad I, J \subset \{1,\ldots,n\} \end{cases}$$

$$\Longrightarrow \mathbf{Var}[Y] = \sum_{i=1}^{i=n} V_i + \sum_{1\le i<j\le n} V_{i,j} + \cdots + V_{1,\ldots,n} \tag{10}$$

où $V_{i_1,\ldots,i_k} = \mathbf{Var}\left[Z_{i_1,\ldots,i_k}\right] = \mathbf{Var}\left[g_{i_1,\ldots,i_k}(X_{i_1},\ldots,X_{i_k})\right]$.

Les indices de Sobol (1) et (2) s'expriment en fonction des $V_{i_1,\ldots,i_k}$:

$$S_{i_1,\ldots,i_k} = \frac{\mathbf{Var}\left[\mathbb{E}\left[Y|X_{i_1},\ldots,X_{i_k}\right]\right]}{\mathbf{Var}[Y]} = \frac{\sum_I V_I}{\mathbf{Var}[Y]} \text{ où } I \subset \{i_1,\ldots,i_k\} \tag{11}$$

$$S^T_{i_1,\ldots,i_k} = \frac{\sum_I \mathbf{Var}\left[\mathbb{E}[Y|X_I]\right]}{\mathbf{Var}[Y]} = \frac{\sum_I \sum_J V_J}{\mathbf{Var}[Y]} \tag{12}$$

$$\text{où} \quad I \subset \{1,\ldots,n\}, I \cap \{i_1,\ldots,i_k\} \neq \emptyset, J \subset I \tag{13}$$

Calcul des Indices de Sobol II

Indices de Sobol : Moyens de calcul dans OpenTURNS

OpenTURNS propose d'évaluer les indices de Sobol via :

- de l'échantillonnage des entrées / sortie :
  - SaltelliSensitivityAlgorithm
  - JansenSensitivityAlgorithm
  - MauntzKucherenkoSensitivityAlgorithm
  - MartinezSensitivityAlgorithm
  - FAST (méthode spectrale)

  Ces algorithmes permettent de calculer les indices de Sobol du premier et du second ordre, ainsi que les indices totaux du premier ordre.
- une décomposition en chaos fonctionnel (polynomial) à l'aide de la classe FunctionalChaosRandomVector qui permet d'exploiter statistiquement le résultat d'un algorithme de décomposition en chaos fonctionnel : calcul de moyenne, variance, **indices de Sobol de tout ordre et les indices de Sobol totaux de tout ordre** (à la demande).

Estimation par échantillonnage I

Soit $\boldsymbol{X}$ un vecteur aléatoire de dimension d à composantes indépendantes. On note :

$$\boldsymbol{A} = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n_X} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n_X} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{1,2} & \cdots & a_{n,n_X} \end{pmatrix}, \quad \boldsymbol{B} = \begin{pmatrix} b_{1,1} & b_{1,2} & \cdots & b_{1,n_X} \\ b_{2,1} & b_{2,2} & \cdots & b_{2,n_X} \\ \vdots & \vdots & \vdots & \vdots \\ b_{n,1} & b_{1,2} & \cdots & b_{n,n_X} \end{pmatrix}$$

deux échantillons de $\boldsymbol{X}$ de taille n et :

$$\boldsymbol{C}^i = \begin{pmatrix} b_{1,1} & b_{1,2} & \cdots & a_{1,i} & \cdots & b_{1,n_X} \\ b_{2,1} & b_{2,2} & \cdots & a_{2,i} & \cdots & b_{2,n_X} \\ \vdots & \vdots & \cdots & \vdots & \ddots & \vdots \\ b_{n,1} & b_{1,2} & \cdots & a_{n,i} & \cdots & b_{n,n_X} \end{pmatrix}, \quad \boldsymbol{E}^i = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & b_{1,i} & \cdots & a_{1,n_X} \\ a_{2,1} & a_{2,2} & \cdots & b_{2,i} & \cdots & a_{2,n_X} \\ \vdots & \vdots & \cdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{1,2} & \cdots & b_{n,i} & \cdots & a_{n,n_X} \end{pmatrix}$$

les matrices de mélange basées sur $\boldsymbol{A}$ et $\boldsymbol{B}$ intervenant dans l'estimation. On note également :

$$\begin{array}{llll} \boldsymbol{Y}^{AB} = g([\boldsymbol{A};\boldsymbol{B}]) & \boldsymbol{Y}_k^{AB} = g([\boldsymbol{A}_k;\boldsymbol{B}_k]) & \boldsymbol{Y}^{A} = g(\boldsymbol{A}) & \boldsymbol{Y}_k^{A} = g(\boldsymbol{A}_k) \\ \boldsymbol{Y}^{B} = g(\boldsymbol{B}) & \boldsymbol{Y}_k^{B} = g(\boldsymbol{B}_k) & \boldsymbol{Y}^{C} = g(\boldsymbol{C}) & \boldsymbol{Y}_k^{C} = g(\boldsymbol{C}_k) \\ \boldsymbol{Y}^{E} = g(\boldsymbol{E}) & \boldsymbol{Y}_k^{E} = g(\boldsymbol{E}_k) & & \end{array}$$

On donne l'expression des différents estimateurs des indices du premier ordre $\hat{S}$ et du premier ordre total $\hat{S}^T$ disponibles dans OpenTURNS.

Estimation par échantillonnage II

Estimateur de Saltelli

Les estimateurs des indices du premier ordres et des indices totaux du permier ordre s'écrivent :

$$S = \frac{\mathbf{Cov}\left[\boldsymbol{Y^B}, \boldsymbol{Y^E}\right]}{\mathbf{Var}\left[Y^{AB}\right]} \rightarrow \hat{S} = \frac{\frac{1}{n}\sum_{k=1}^{n} Y_k^B Y_k^E - \left(\frac{1}{n}\sum_{k=1}^{n} Y_k^A\right)^2}{\frac{1}{n}\sum_{k=1}^{n}(Y_k^{AB})^2 - \left(\frac{1}{n}\sum_{k=1}^{n} Y_k^{AB}\right)^2}$$

$$\hat{S}^T = 1 - \frac{\frac{1}{n}\sum_{k=1}^{n} Y_k^A Y_k^E - \left(\frac{1}{n}\sum_{k=1}^{n} Y_k^A\right)^2}{\frac{1}{n}\sum_{k=1}^{n}(Y_k^{AB})^2 - \left(\frac{1}{n}\sum_{k=1}^{n} Y_k^{AB}\right)^2}$$

Ils sont convergents, asymptotiquement normaux de variances asymptotiques :

$$\sigma^2 = \frac{\mathbf{Var}\left[\left(Y_i^B - \mathbb{E}\left[Y^B\right]\right)\left(Y_i^E - \mathbb{E}\left[Y^E\right]\right) - S\left(Y_i^{AB} - \mathbb{E}\left[Y^{AB}\right]\right)^2\right]}{\mathbf{Var}\left[Y^{AB}\right]^2}$$

$$\sigma_T^2 = \frac{\mathbf{Var}\left[\left(Y_i^A - \mathbb{E}\left[Y^A\right]\right)\left(Y_i^E - \mathbb{E}\left[Y^E\right]\right) - S^{-X}\left(Y_i^{AB} - \mathbb{E}\left[Y^{AB}\right]\right)^2\right]}{\mathbf{Var}\left[Y^{AB}\right]^2}$$

où on a posé $S^{-X} = \frac{\mathbf{Cov}\left[\boldsymbol{Y^A}, \boldsymbol{Y^E}\right]}{\mathbf{Var}\left[Y^{AB}\right]}$.

Estimation par échantillonnage III

Estimateur de Jansen

Les estimateurs des indices du premier ordres et des indices totaux du permier ordre s'écrivent :

$$S = \frac{\mathbf{Var}\left[Y^{AB}\right] - \mathbf{Var}\left[Y^{E} - Y^{B}\right]}{\mathbf{Var}\left[Y^{AB}\right]} \rightarrow \hat{S} = 1 - \frac{\frac{1}{2n}\sum_{k=1}^{n}\left(Y_k^E - Y_k^B\right)^2}{\frac{1}{n}\sum_{k=1}^{n}(Y_k^{AB})^2 - \left(\frac{1}{n}\sum_{k=1}^{n} Y_k^{AB}\right)^2}$$

$$\hat{S}^T = \frac{\frac{1}{n}\sum_{k=1}^{n}\left(Y_k^A - Y_k^E\right)^2}{\mathbf{Var}\left[Y^{AB}\right]} = \frac{\frac{1}{n}\sum_{k=1}^{n}\left(Y_k^A - Y_k^E\right)^2}{\frac{1}{n}\sum_{k=1}^{n}(Y_k^{AB})^2 - \left(\frac{1}{n}\sum_{k=1}^{n} Y_k^{AB}\right)^2}$$

Ils sont convergents, asymptotiquement normaux de variances asymptotiques :

$$\sigma^2 = \frac{1}{4}\frac{\mathbf{Var}\left[\left(Y_i^E - \mathbb{E}\left[Y^E\right] - Y_i^B + \mathbb{E}\left[Y^B\right]\right)^2 - J\left(Y_i^{AB} - \mathbb{E}\left[Y^{AB}\right]\right)^2\right]}{\mathbf{Var}\left[Y^{AB}\right]^2}$$

$$\sigma_T^2 = \frac{\mathbf{Var}\left[\left(Y_i^A - \mathbb{E}\left[Y^A\right] - Y_i^E + \mathbb{E}\left[Y^E\right]\right)^2 - J^{-X}\left(Y_i^{AB} - \mathbb{E}\left[Y^{AB}\right]\right)^2\right]}{\mathbf{Var}\left[Y^{AB}\right]^2}$$

où on a posé $J = \frac{\mathbf{Var}\left[Y^E - Y^B\right]}{\mathbf{Var}\left[Y^{AB}\right]}$.

Estimation par échantillonnage IV

Estimateur de Mauntz-Kucherenko

Les estimateurs des indices du premier ordres et des indices totaux du permier ordre s'écrivent :

$$S = \frac{\mathbf{Cov}\left[\boldsymbol{Y^B}, \boldsymbol{Y^E} - \boldsymbol{Y^A}\right]}{\mathbf{Var}\left[Y^{AB}\right]} \rightarrow \hat{S} = \frac{\frac{1}{n}\sum_{k=1}^{n} Y_k^B (Y_k^E - Y_k^A)}{\frac{1}{n}\sum_{k=1}^{n} (Y_k^{AB})^2 - \left(\frac{1}{n}\sum_{k=1}^{n} Y_k^{AB}\right)^2}$$

$$\hat{S}^T = 1 - \frac{\frac{1}{n}\sum_{k=1}^{n} Y_k^A Y^E - \left(\frac{1}{n}\sum_{k=1}^{n} Y_k^A\right)^2}{\frac{1}{n}\sum_{k=1}^{n} (Y_k^{AB})^2 - \left(\frac{1}{n}\sum_{k=1}^{n} Y_k^{AB}\right)^2}$$

Ils sont convergents, asymptotiquement normaux de variances asymptotiques :

$$\sigma^2 = \frac{\mathrm{Var}\left[\left(Y_i^B - \mathbb{E}\left[Y^B\right]\right)\left(\left(Y_i^E - \mathbb{E}\left[Y^E\right]\right) - \left(Y_i^A - \mathbb{E}\left[Y^A\right]\right)\right) - M\left(Y_i^{AB} - \mathbb{E}\left[Y^{AB}\right]\right)^2\right]}{\mathbf{Var}\left[Y^{AB}\right]^2}$$

$$\sigma_T^2 = \frac{\mathbf{Var}\left[\left(Y_i^A - \mathbb{E}\left[Y^A\right]\right)\left(Y_i^E - \mathbb{E}\left[Y^E\right]\right) - S^{-X}\left(Y_i^{AB} - \mathbb{E}\left[Y^{AB}\right]\right)^2\right]}{\mathbf{Var}\left[Y^{AB}\right]^2}$$

où on a posé $M = \frac{\mathbf{Cov}\left[\boldsymbol{Y^B}, \boldsymbol{Y^E} - \boldsymbol{Y^A}\right]}{\mathbf{Var}\left[Y^{AB}\right]}$.

Estimation par échantillonnage V

Estimateur de Martinez

Les estimateurs des indices du premier ordres et des indices totaux du permier ordre s'écrivent :

$$S = \rho_n\left(Y_k^B, Y_k^E\right) \to \hat{S} = \frac{\frac{1}{n}\sum_{k=1}^{n}\left(Y_k^B - \mathbb{E}\left[Y^B\right]\right)\left(Y_k^E - \mathbb{E}\left[Y^E\right]\right)}{\sqrt{\left(\frac{1}{n}\sum_{k=1}^{n}(Y_k^B - \mathbb{E}\left[Y^B\right])^2\right)\left(\frac{1}{n}\sum_{k=1}^{n}(Y_k^E - \mathbb{E}\left[Y^E\right])^2\right)}}$$

$$\hat{S}^T = 1 - \frac{\frac{1}{n}\sum_{k=1}^{n}\left(Y_k^A - \mathbb{E}\left[Y^A\right]\right)\left(Y_k^E - \mathbb{E}\left[Y^E\right]\right)}{\sqrt{\left(\frac{1}{n}\sum_{k=1}^{n}(Y_k^A - \mathbb{E}\left[Y^A\right])^2\right)\left(\frac{1}{n}\sum_{k=1}^{n}(Y_k^E - \mathbb{E}\left[Y^E\right])^2\right)}}$$

Ils sont convergents, asymptotiquement normaux de variances asymptotiques :

$$\sigma^2 = -\frac{\mathbf{Cov}\left[\boldsymbol{Y^B}, \boldsymbol{Y^E}\right](\mathbf{Cov}\left[\boldsymbol{x}; \boldsymbol{y}\mathbf{Var}\left[\boldsymbol{Y^E}\right] + \boldsymbol{z}\mathbf{Var}\left[\boldsymbol{Y^B}\right]\right]}{\mathbf{Var}\left[Y^B\right]\mathbf{Var}\left[Y^E\right]^2} + \frac{\mathbf{Var}\left[x\right]}{\mathbf{Var}\left[Y^B\right]\mathbf{Var}\left[Y^E\right]}$$

$$\sigma_T^2 = \frac{M^2\left(\mathbf{Var}\left[z\mathbf{Var}\left[Y^A\right]\right] + \mathbf{Var}\left[y\mathbf{Var}\left[Y^E\right]\right]\right)}{4\mathbf{Var}\left[Y^A\right]^2\mathbf{Var}\left[Y^E\right]^2} -$$

$$\frac{\mathbf{Cov}\left[\boldsymbol{Y^A}, \boldsymbol{Y^E}\right](\mathbf{Cov}\left[\boldsymbol{x}; \boldsymbol{y}\mathbf{Var}\left[\boldsymbol{Y^E}\right] + \boldsymbol{z}\mathbf{Var}\left[\boldsymbol{Y^A}\right]\right]}{\mathbf{Var}\left[Y^A\right]\mathbf{Var}\left[Y^E\right]^2} + \frac{\mathbf{Var}\left[x\right]}{\mathbf{Var}\left[Y^A\right]\mathbf{Var}\left[Y^E\right]}$$

Estimation par échantillonnage VI

Estimation LHS, Quasi-Monte Carlo

- Les estimateurs précédents restent convergents dès lors que les matrices $\boldsymbol{A}$ et $\boldsymbol{B}$ sont orthogonales
- Cette propriété est obtenue automatiquement lors d'un échantillonnage de Monte Carlo mais ni avec LHS ni avec QMC
- L'astuce consiste à générer simultanément les matrices $\boldsymbol{A}$ et $\boldsymbol{B}$ comme un échantillon de taille n d'un vecteur $\tilde{\boldsymbol{X}} = (\boldsymbol{X}, \boldsymbol{X}')$ où $\boldsymbol{X}'$ est une copie de $\boldsymbol{X}$ indépendante de $\boldsymbol{X}$.

Cas test Ishigami I

On étudie le comportement des estimateurs sur le cas-test suivant (Ishigami) :

- Le vecteur d'entrée $\boldsymbol{X}$ est de dimension 3, à composantes indépendantes de lois marginales uniformes sur $[-\pi, \pi]$;
- La grandeur d'intérêt Y est liée à $\boldsymbol{X}$ par :

$$Y = \sin(X_1) + a \sin(X_2)^2 + b X_3^4 \sin(X_1) \text{ avec } a = 7, b = \frac{1}{10}$$

ce qui conduit aux indices de Sobol du premier ordre et totaux du premier ordre suivants :

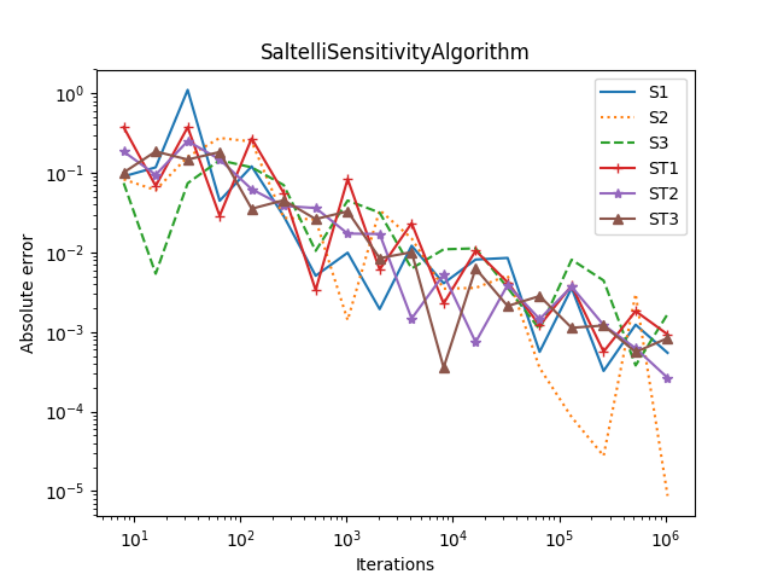
$$S_1 = \frac{36(b\pi^4 + 5)^2}{100b^2\pi^8 + 360b\pi^4 + 225a^2 + 900} \qquad S_2 = \frac{45a^2}{20b^2\pi^8 + 72b\pi^4 + 45a^2 + 180}$$

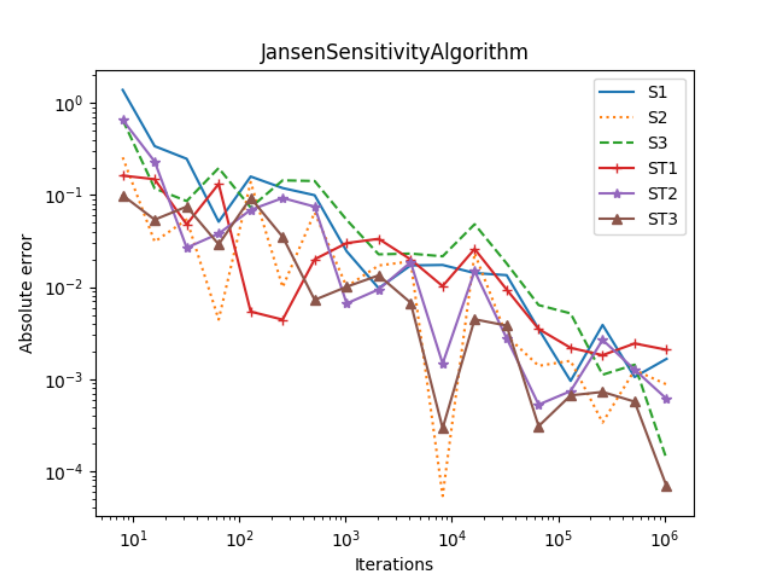
$$S_3 = 0$$

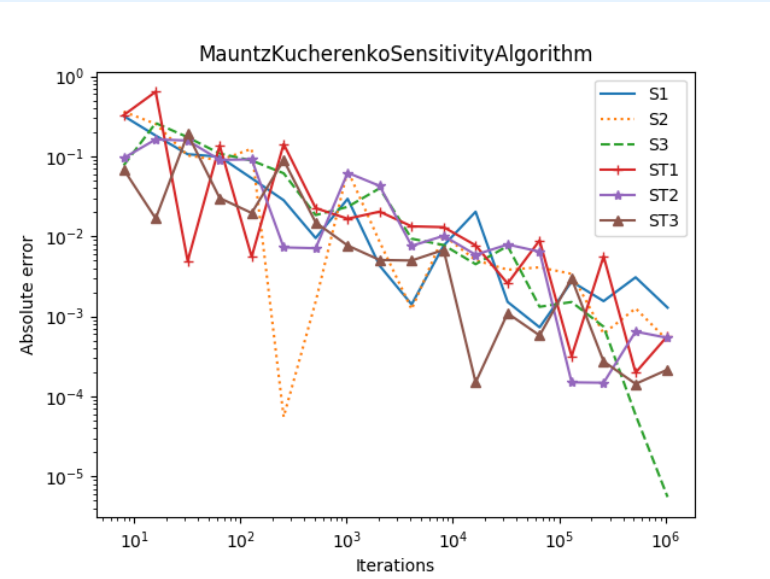
$$S_1^T = \frac{20b^2\pi^8 + 72b\pi^4 + 180}{20b^2\pi^8 + 72b\pi^4 + 45a^2 + 180} \qquad S_2^T = \frac{45a^2}{20b^2\pi^8 + 72b\pi^4 + 45a^2 + 180}$$

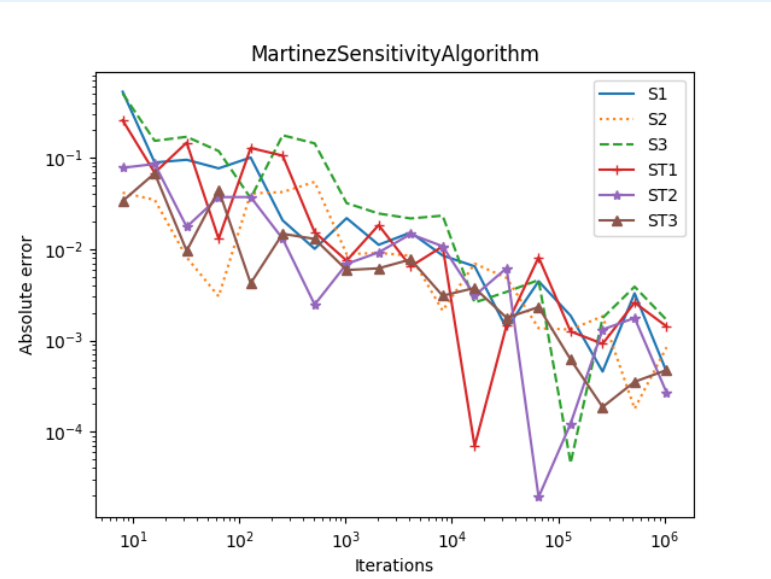
$$S_3^T = \frac{64b^2\pi^8}{100b^2\pi^8 + 360b\pi^4 + 225a^2 + 900}$$

Cas test Ishigami II

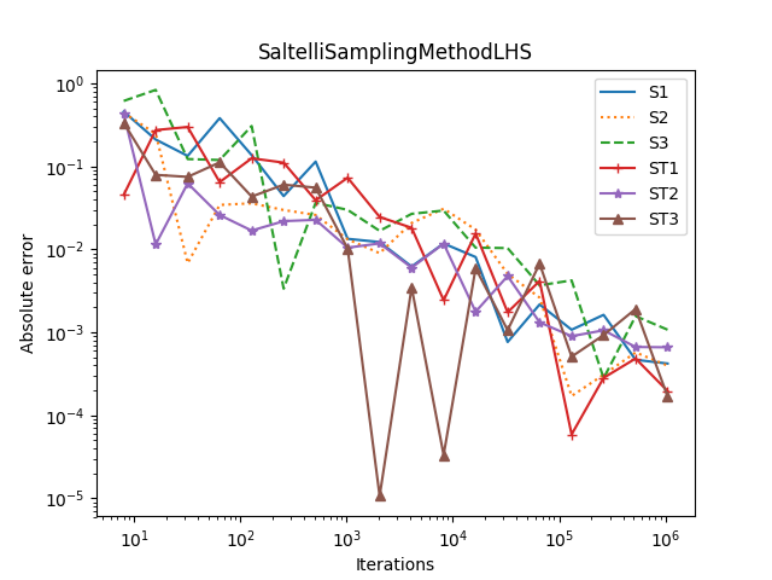


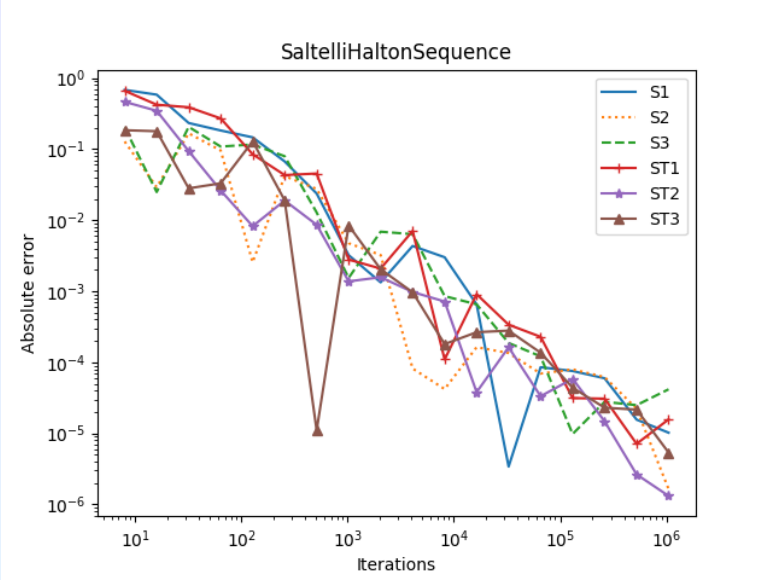


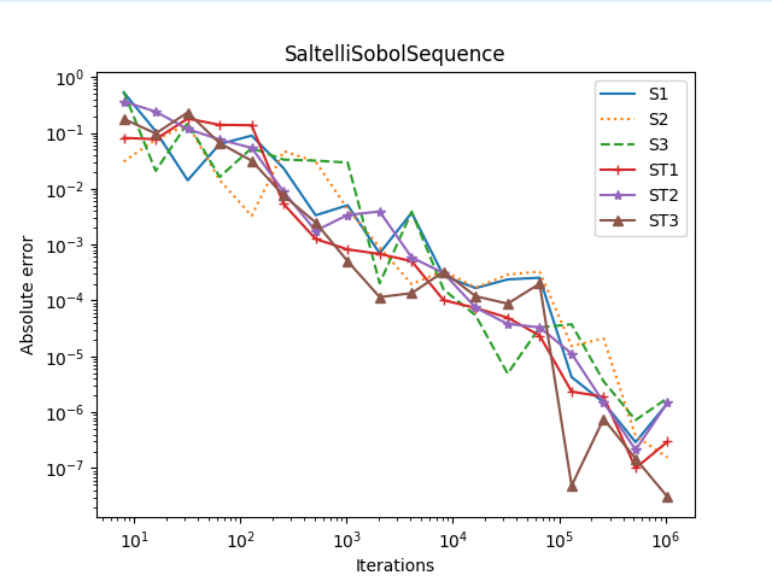


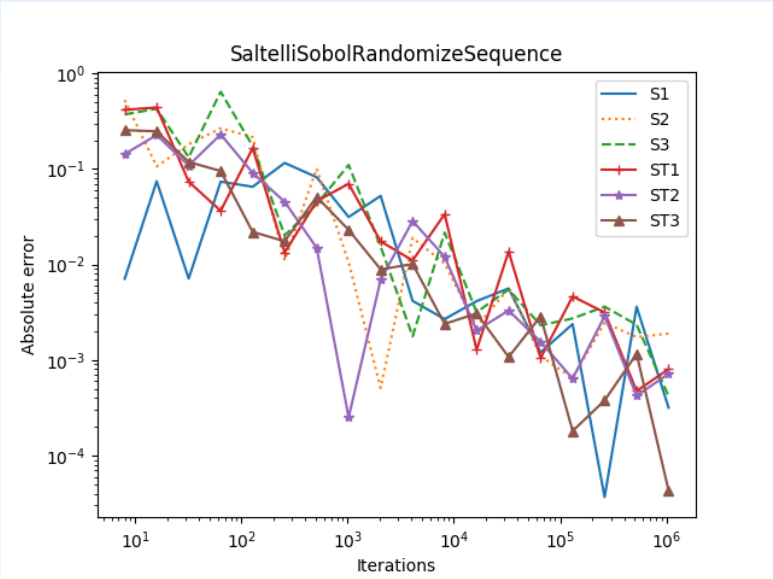


Cas test Ishigami III









Facteurs d'importance, modèle affine I

Si le modèle g est affine : SRC

Si $\boldsymbol{Y = \alpha_0 + \sum_i \alpha_i X_i}$, alors on définit le Standard Regression Coefficient (SRC) :

$$SRC_i^2 = \alpha_i^2 \frac{\mathbf{Var}\left[X_i\right]}{\mathbf{Var}\left[Y\right]} \quad \text{sign}(SRC_i) = \text{sign}(\alpha_i) \tag{14}$$

Si les variables X_i sont décorrélées, les coefficients SRC^2 constituent une décomposition de la variance de Y : $\sum_{i=1}^{n} SRC_i^2 = 1$, et si de plus les X_i sont indépendantes, les coefficients SRC^2 coïncident avec les indices de Sobol à l'ordre 1 : $\boldsymbol{SRC^2(Y \mid X_i) = S_i(Y \mid X_i)}$.

Facteurs d'importance, modèle affine II

Si le modèle g est monotone par marginale : SRRC

Si $Y = g(\boldsymbol{X})$ avec les X_i indépendantes et si g est monotone par rapport à chaque X_i, en posant $\boldsymbol{U} = (F_{\mathbf{1}}(X_{\mathbf{1}}), \ldots, F_n(X_n))^t = \phi^{-\mathbf{1}}(\boldsymbol{X})$, on a la relation suivante sur les rangs Z de Y et U_i des X_i :

$$Z = F_Y(Y) = F_Y \circ g \circ \phi(\boldsymbol{U})$$

Si on suppose **de plus** que :

$$\boldsymbol{Z = \sum_i \alpha_i U_i} \tag{15}$$

alors on définit le Standard Rank Regression Coefficient (SRRC) :

$$SRRC(Y \mid X_i) = SRC(Z \mid U_i) = \alpha_i \sqrt{\frac{\mathbf{Var}\left[U_i\right]}{\mathbf{Var}\left[Z\right]}} = \mathrm{sign}(\alpha_i)\sqrt{S_i(Z \mid U_i)}$$

Donc $SRRC^2$ **est un indice de Sobol** à l'ordre 1 calculé sur les rangs des X_i et de Y.

SRC et SRRC : Moyens de calcul dans OpenTURNS

La classe CorrelationAnalysis permet le calcul de **SRC et SRRC sur la base d'un échantillonnage de $\boldsymbol{X}$ et de Y.**

Analyse de sensibilité

Facteurs d'importance historiques I

Cumul quadratique (tendance centrale)

$Y = g(\boldsymbol{X})$ est approché par son **approximation de Taylor à l'ordre 1 au point moyen** $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{X}]$:

$$Y = g(\boldsymbol{\mu}) + < \boldsymbol{\nabla} g(\boldsymbol{\mu}), (\boldsymbol{X} - \boldsymbol{\mu}) >= g(\boldsymbol{\mu}) + \sum_i (X_i - \mu_i) \left. \frac{\partial g}{\partial x_j} \right|_{\boldsymbol{\mu}} \tag{16}$$

Sous cette **hypothèse de linéarité du modèle au point moyen**, on calcule :

$$\mathbf{Var}\left[Y\right] \quad = \quad {}^t\boldsymbol{\nabla} g(\boldsymbol{\mu}).\mathbf{Cov}\left[\boldsymbol{X}\right].\boldsymbol{\nabla} g(\boldsymbol{\mu}) = \sum_{i,j} \left. \frac{\partial g}{\partial x_i} \right|_{\boldsymbol{\mu}} \mathrm{Cov}\left[X_i, X_j\right] . \left. \frac{\partial g}{\partial x_i} \right|_{\boldsymbol{\mu}} \tag{17}$$

On définit le **facteur d'importance de** X_i :

$$\boldsymbol{FI(X_i)} = \frac{\left(\sum_j \left. \frac{\partial g}{\partial x_j} \right|_{\boldsymbol{\mu}} \mathbf{Cov}\left[\boldsymbol{X_i}, \boldsymbol{X_j}\right] \right) \left. \frac{\partial g}{\partial x_i} \right|_{\boldsymbol{\mu}}}{\mathbf{Var}\left[\boldsymbol{Y}\right]} \tag{18}$$

Si les X_i sont décorrélées, alors (18) se simplifie en :

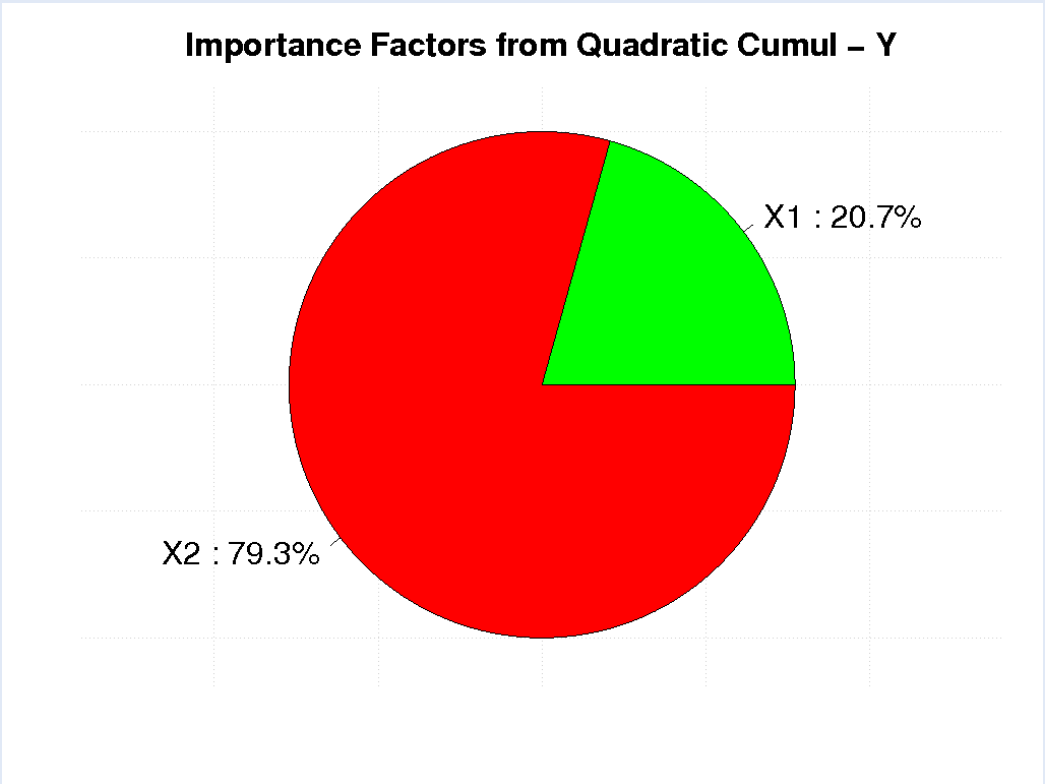
$$\boldsymbol{FI(X_i)} = \left(\left. \frac{\partial g}{\partial x_j} \right|_{\boldsymbol{\mu}} \right)^2 \frac{\mathbf{Var}\left[\boldsymbol{X_i}\right]}{\mathbf{Var}\left[\boldsymbol{Y}\right]} = SRC^2(Y \mid X_i)$$

et si de plus elles sont indépendantes, $\boldsymbol{FI(X_i) = S_i(Y \mid X_i)}$. **Dans le cas des X_i indépendantes et d'un modèle linéaire, les *FI* du cumul quadratique sont des indices de Sobol à l'ordre 1.**

Facteurs d'importance historiques II

Cumul quadratique : Moyens de calcul dans OpenTURNS

La classe **QuadraticCumul** permet le calcul des facteurs d'importance *FI* ainsi que leur tracé sous forme de camembert : getImportanceFactors() et drawImportanceFactors().



Facteurs d'importance historiques III

FORM / SORM : facteurs d'importance en régime d'événements rares

Le modèle $Y = g(\boldsymbol{X})$ est plongé via une transformation isoprobabiliste $\boldsymbol{U} = T(\boldsymbol{X})$ dans l'espace standard des variables sphériques décorrélées U_i.
Dans cet espace, le **modèle est linéarisé au point de conception** P^*. Si $\beta = ||\boldsymbol{OP}^*||$ et $\boldsymbol{\alpha} = \frac{1}{\beta}\boldsymbol{OP}^*$, alors :

$$Y \simeq M = \beta - < \boldsymbol{\alpha}, \boldsymbol{U} >= \beta - \sum_i \alpha_i U_i \tag{19}$$

La **littérature ne prenait en compte jusqu'à récemment que des espaces standards gaussiens** dans lesquels les **facteurs d'importance de FORM** sont définis par :

$$\boldsymbol{FI(X_i) = \alpha_i^2} \tag{20}$$

Cette définition est **généralisable au cas des espaces standards sphériques**. Dans cet espace, les variables U_i sont décorrélées (indépendantes dans le cas gaussien), et de marginales centrées réduites. Donc on montre que, pour le modèle linéarisé (19) :

$$SRC^2(Y \mid U_i) = \frac{\alpha_i^2 \mathbf{Var}\left[U_i\right]}{\mathbf{Var}\left[M\right]} = \alpha_i^2 \tag{21}$$

car $\mathbf{Var}\left[U_i\right] = 1$ et $\mathbf{Var}\left[M\right] = \sum_i \alpha_i^2 \mathbf{Var}\left[U_i\right] = \sum_i \alpha_i^2 = 1$.
Donc les **facteurs d'importance de FORM sont des indices SRC** pour le modèle linéarisé au point de conception, et également **des indices de Sobol d'ordre 1 dans le cas gaussien**.
Attention ! : les FI sont des $SRC^2(Y \mid \boldsymbol{U_i})$ et que dire dans le cas où les X_i sont corrélées qui associe U_i à une combinaison linéaire de plusieurs X_j ? ? ?

Facteurs d'importance historiques IV

FORM / SORM : Implémentation dans OpenTURNS

Dans OpenTURNS, la transformation isoprobabiliste utilisée est :

- celle de Nataf Généralisée lorsque la copule de $\boldsymbol{X}$ est elliptique : l'espaces standard sphérique pas nécessairement gaussien et les U_i sont décorrélées (et indépendantes uniquement dans le cas gaussien).
- celle de Rosenblatt lorsque la copue de $\boldsymbol{X}$ n'est pas elliptique : l'espace standard est gaussien, et les U_i sont indépendantes.

Pour remédier au problème de la non bijection entre U_i et X_i (i.e. margianle à marginale), OpenTURNS ramène le point de conception dans l'espace elliptique des Z_i corrélés (suppression de l'étape de décorrélation qui mélange les composantes entre elles) où :

$$\boldsymbol{Z}^* = (\boldsymbol{E}^{-1} \circ \boldsymbol{F}_1(\boldsymbol{X}_1^*), \ldots, \boldsymbol{E}^{-1} \circ \boldsymbol{F}_n(\boldsymbol{X}_n^*))^t \tag{22}$$

où E est la CDF centrée réduite de même type que la copule de $\boldsymbol{X}$ dans le cas elliptique ou gaussienne dans le cas non elliptique.
La transformation étant faite composante à composante, $Z_i \leftrightarrow X_i$ et **OpenTURNS définit le facteur d'importance de X_i** comme :

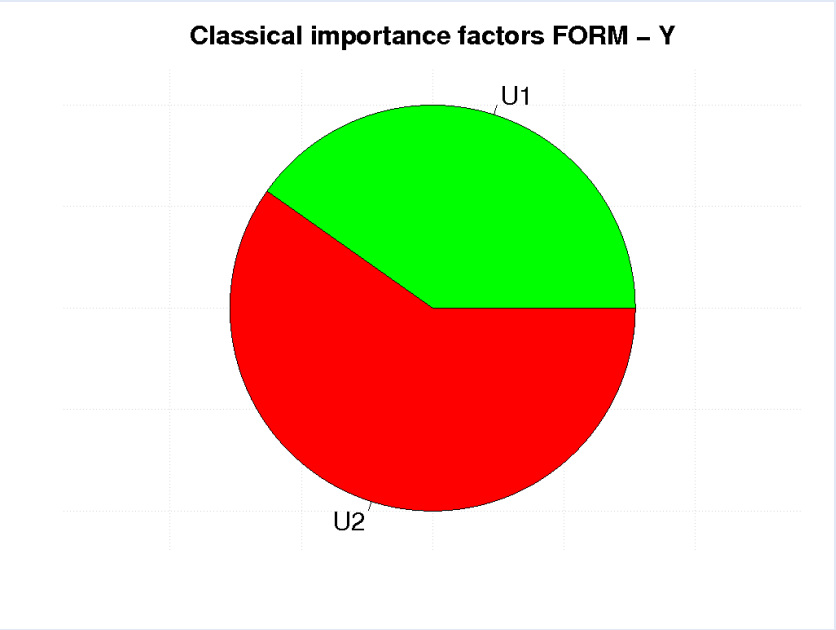
$$\boldsymbol{FI}(\boldsymbol{X_i}) = \left(\frac{\boldsymbol{Z_i^*}}{||\boldsymbol{Z}^*||} \right)^2 \tag{23}$$

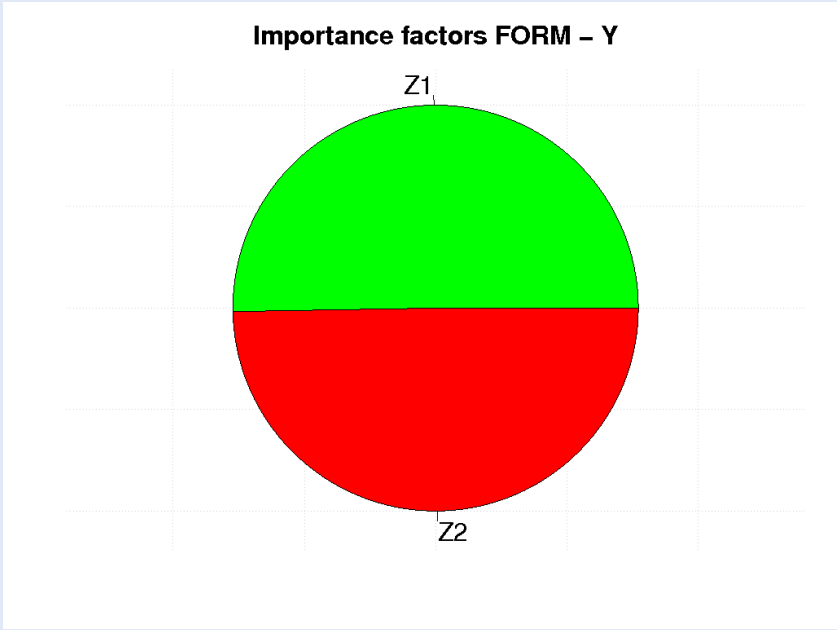
Il y a concordance des définitions (20) et (23) dans le cas de variables à copule gaussienne de corrélation identité.

Facteurs d'importance historiques V

Moyens de calcul dans OpenTURNS

Les classes **FORMResult** et **SORMResult** permettent le calcul des facteurs d'importance *FI* ainsi que leur tracé sous forme de camembert : getImportanceFactors() et drawImportanceFactors().





Analyse de sensibilité

Corrélation I

Si le modèle est linéaire : Pearson

Si on suppose que $\boldsymbol{Y = \sum_i \alpha_i X_i}$, on définit la **corrélation de Pearson** entre Y et X_i comme :

$$\boldsymbol{\rho(Y, X_i) = \frac{\mathbf{cov}\,[Y, X_i]}{\sqrt{\mathbf{Var}\,[X_i]\mathbf{Var}\,[Y]}}} \tag{24}$$

Dans le cadre linéaire, **si les X_i sont indépendantes**, alors on montre que :

$$\rho^2(Y, X_i) = SRC_i^2 = S(Y \mid X_i)$$

Si le modèle est monotone : Spearman

Si on suppose que le modèle est monotone et que **les rangs sont linéaires** (15) , on définit la **corrélation des rangs de Spearman** entre Y et X_i comme :

$$\boldsymbol{\rho_S(Y, X_i) = \rho(F_Y(Y), F_i(X_i))}$$

De même, on montre que dans le cas de **variables indépendantes** :

$$\rho_S^2(Y, X_i) = SRRC^2(Y \mid X_i) = SRC^2(Z \mid U_i) = S(Z \mid U_i)$$

Corrélation II

Si le modèle est linéaire : PCC

Si on suppose que $Y = \sum_i \alpha_i X_i$, on définit le coefficient ***PCC$_i$*** (**Partial Correlation Coefficient**) comme étant la corrélation de Pearson entre $(Y - \mathbb{E}[Y|X_i])$ et $(Y - \mathbb{E}[Y|X_{\sim i}])$ où $X_{\sim i} = \{X_1, \ldots, X_n\} \backslash \{X_i\}$. On a donc :

$$PCC_i = \mathrm{sgn}(\alpha_i) \frac{\sum_{j \neq i} \alpha_j \mathbf{Cov}\left[\boldsymbol{X_i}, \boldsymbol{X_j}\right]}{\sqrt{\mathbf{Var}\left[X_i\right]} \sqrt{\mathbf{Var}\left[Y - X_i\right]}} \tag{25}$$

Ce coefficient est nul dès lors que X_i est décorrelé de $X_{\sim i}$.
Hors cadre linéaire, le coefficient *PCC$_i$* perd de sa signification.

Coefficients de corrélation linéaire sur les rangs : PRCC

Cadre d'un modèle monotone entre Y et chaque X_i.
La monotonie de la relation entre Y et les X_i rend linéaire la relation entre les rangs des réalisations des Y et les rangs des réalisations des X_i.
Le coefficient ***PRCC$_i$*** (**Partial Rank Correlation Coefficient**) est l'équivalent du coefficient PCC_i calculé sur les rangs des variables.

Pearson, Spearman, PCC, PRCC : Moyens de calcul dans OpenTURNS

La classe **CorrelationAnalysis** permet le calcul de toutes ces grandeurs.

Analyse de sensibilité

Sensibilités en régime d'événements rares I

FORM / SORM

Les **facteurs de sensibilité** qui quantifient la dépendance de l'incertitude sur $Y = g(\boldsymbol{X})$ aux paramètres de la loi de $\boldsymbol{X}$ peuvent être lus dans deux échelles de risque différentes :

- comme **sensibilités sur la probabilité de défaillance** $\frac{\partial \mathbb{P}(Z>s)}{\partial \boldsymbol{\lambda}}$,
- comme **sensibilités sur l'indice de fiabilité** $\frac{\partial \beta}{\partial \boldsymbol{\lambda}}$.

Dans l'espace standard, on peut calculer le point moyen dans l'espace de dépassement de seuil $\mathcal{D}$:

$$\boldsymbol{u}^* = \mathbb{E}\left[\boldsymbol{U} | \boldsymbol{U} \in \mathcal{D}\right] \tag{26}$$

et en dériver des facteurs d'importance comme :

$$\boldsymbol{FI_{MC}(X_i)} = \left(\frac{\boldsymbol{u_i^*}}{||\boldsymbol{U^*}||}\right)^2 \tag{27}$$

Par symétrie de la loi sphérique de l'espace standard, $\boldsymbol{U}^*$ est sur la droite portée par le cosinus directeur du point de conception.
Dans le cas où l'approximation FORM est exacte, les 2 facteurs d'importances (20) et (27) coïncident.
Implémentation dans OpenTURNS
On veut calculer les facteurs d'importance de X_1 et X_2 sur l'événement $Y > s$ avec $s \in \mathbb{R}$ et $Y = 1/(\exp(-2X_1) + exp(-4X_2))$ où X_1 et X_2 sont iid $\mathcal{N}(0,1)$.

Sensibilités en régime d'événements rares II

Listing 1 – monteCarloSensitivity.py

```
import openturns as ot
import openturns.viewer as otv

# Physical model
f = ot.MemoizeFunction(ot.SymbolicFunction(["x1", "x2"], ["1.0 / (exp(-2*x1)+exp(-4*x2))"
    ]))
########################################
# Mandatory for sensitivity analysis #
########################################
f.enableHistory()
# Input distribution
dist_x = ot.Normal(2)
# Input random vector
X = ot.RandomVector(dist_x)
# Output random vector
Y = ot.CompositeRandomVector(f, X)
# Event
s = 4.0
test = ot.Greater()
E = ot.Event(Y, test, s)
# Monte Carlo
algo = ot.ProbabilitySimulationAlgorithm(E, ot.MonteCarloExperiment())
algo.setMaximumOuterSampling(10000)
algo.setMaximumCoefficientOfVariation(0.0)
algo.run()
result = algo.getResult()
# Probability
```

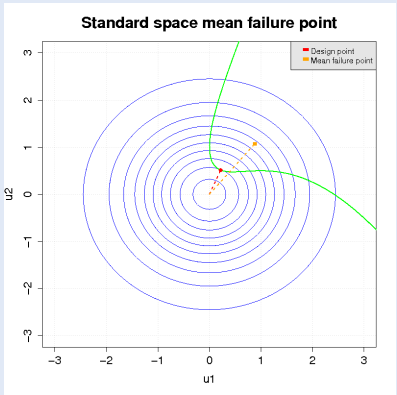
Sensibilités en régime d'événements rares III

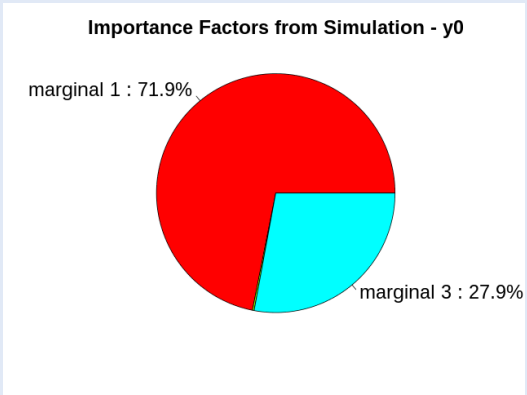
```
p = result.getProbabilityEstimate()
print("Probabilite Monte Carlo=", p)
sensitivity = ot.SimulationSensitivityAnalysis(result)
print("Point moyen de l'evenement=", sensitivity.computeMeanPointInEventDomain())
print("Facteurs d'importance=", sensitivity.computeImportanceFactors())
graph = sensitivity.drawImportanceFactorsRange()
view = otv.View(graph)
view.save("Monte_carlo_sensitivity.png")
view.close()
```

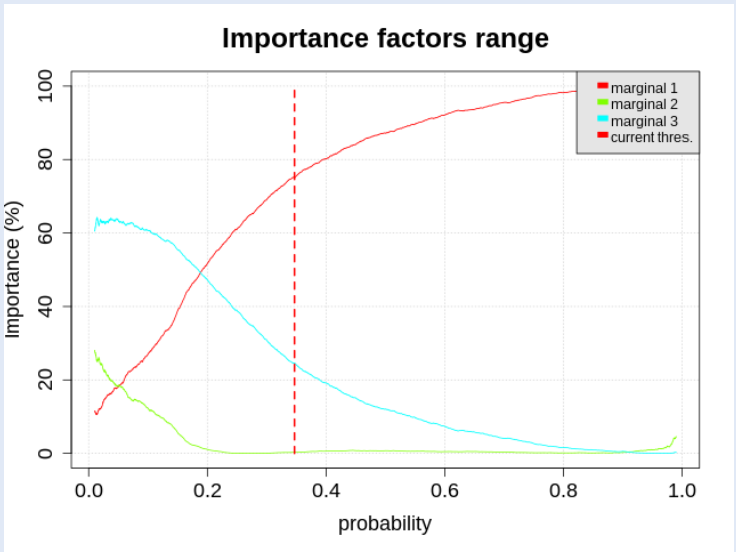
```
Probabilite Monte Carlo= 0.07180000000000028
Point moyen de l'evenement= [1.35094,1.11694]
Facteurs d'importance= [X0 : 0.593972, X1 : 0.406028]
```

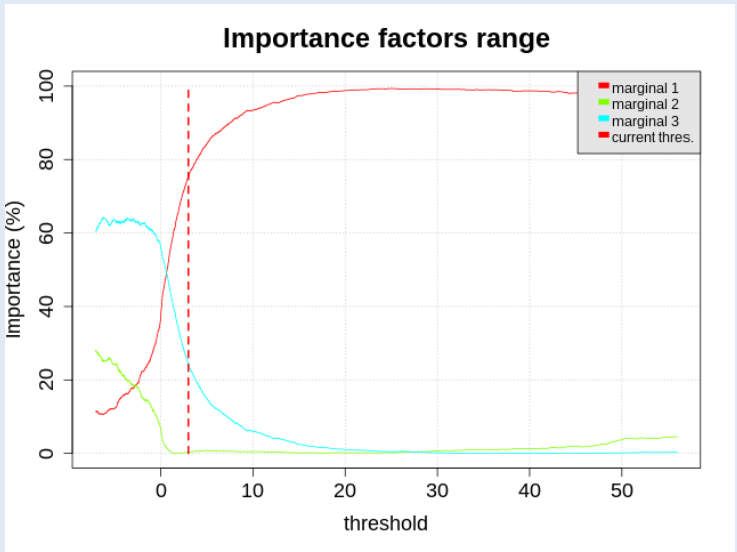
Sensibilités en régime d'événements rares IV

Monte Carlo









Analyse de sensibilité

Contexte (rappel)

Soient :

- g une fonction mesurable de $\mathbb{R}^d$ dans $\mathbb{R}$
- $\boldsymbol{X}$ un vecteur aléatoire de dimension d
- F sa fonction de répartition supposée absolument continue de densité p.

On définit la variable aléatoire $Y = g(\boldsymbol{X})$ et on souhaite quantifier l'intensité de la dépendance entre Y et un sous-vecteur $\tilde{\boldsymbol{X}}$ de $\boldsymbol{X}$ par une grandeur scalaire, une mesure de dépendance.
La notion de divergence de Csiszár [Csiszar1963] permet de construire de telles mesures de dépendance.
Dans beaucoup d'applications, l'objectif est de hiérarchiser l'influence des composantes X_i de $\boldsymbol{X}$ sur Y, éventuellement pour fixer les variables peu influentes à une valeur typique et réduire la dimension de la relation entre $\boldsymbol{X}$ et Y.

Quantifier une dépendance I

Comment quantifier l'influence de X_i sur $Y = g(\boldsymbol{X})$?

- L'approche «classique» : via une décomposition de la variance de Y en termes de variances conditionnelles $\mathbf{Var}\left[\mathbb{E}\left[Y|X_i\right]\right]$ et de termes complémentaires, sous l'hypothèse d'indépendance des X_i (indices de Sobol, [Sobol2001]). Formellement, pour une variable aléatoire Y de variance finie, l'indice de Sobol de Y par rapport à X_i est défini par :

$$S_{X_i}^Y = \frac{\mathbf{Var}\left[\mathbb{E}\left[Y|X_i\right]\right]}{\mathbf{Var}\left[Y\right]}$$

  ce qui s'écrit également :

$$S_{X_i}^Y = 1 - \frac{\mathbb{E}\left[\mathbf{Var}\left[Y|X_i\right]\right]}{\mathbf{Var}\left[Y\right]}$$

  Sous cette forme, il apparaît que cet indice de sensibilité revient à comparer la variance de Y à la valeur moyenne de la variance de Y sachant X_i.

- Lorsque les composantes de $\boldsymbol{X}$ sont dépendantes, une extension des indices de Sobol a été proposée dans [Caniou2012]. Cette extension reste cependant adhérente à une décomposition de la variance, qui n'est pas forcément le critère de dépendance d'intérêt.

Quantifier une dépendance II

- Dans [Liu2005], les auteurs proposent de considérer l'entropie relative de la loi de (X_i, Y) par rapport à la loi produit des lois marginales de X_i et Y comme mesure de sensibilité de Y par rapport à X_i. La même idée est reprise dans [Borgonovo2016], mais en utilisant cette fois la variation totale entre ces deux lois.
- Prolongeant ces idées, dans [DaVeiga2013] l'auteur propose de définir les indices de sensibilité subordonnés à une divergence D_f par

$$S_i^f = \mathbb{E}\left[D_f(Y||Y|X_i)\right] = \int_{\mathbb{R}}\int_{\mathbb{R}} f\left(\frac{p_Y(y)}{p_{Y|X_i}(y|x)}\right) p_{Y|X_i}(y|x)dy\, p_{X_i}(x)dx$$

$$= \int_{\mathbb{R}^2} f\left(\frac{p_{X_i}(x)p_Y(y)}{p_{X_i,Y}(x,y)}\right) p_{X_i,Y}(x,y)\, dxdy = D_f(X \otimes Y||(X,Y))$$

C'est cette extension que nous détaillons. Mais qu'est-ce que D_f ? ? ?

Analyse de sensibilité

Définition [Csiszar1963] I

Définition

Soient P et Q deux mesures de probabilité définies sur un espace Ω telles que P soit absolument continue par rapport à Q. Soit f une fonction convexe positive définie au minimum sur $\mathbb{R}^+$ telle que $f(1) = 0$, la f-divergence de Csiszár de Q par rapport à P est définie par :

$$D_f(P||Q) = \int_\Omega f\left(\frac{P}{Q}\right) dQ \tag{28}$$

Si P et Q sont absolument continues par rapport à une mesure de référence μ, de densités p et q, alors :

$$D_f(P||Q) = \int_\Omega f\left(\frac{p(x)}{q(x)}\right) q(x) d\mu(x) \in [0, +\infty] \tag{29}$$

Quelques cas usuels :

- μ est la mesure de comptage sur $\mathbb{N}$: $D_f(P||Q) = \sum_{k=0}^{\infty} f\left(\frac{p(x_k)}{q(x_k)}\right) q(x_k)$
- μ est la mesure de Lebesgues sur $\mathbb{R}$: $D_f(P||Q) = \int_{\mathbb{R}} f\left(\frac{p(x)}{q(x)}\right) q(x) dx$
- P et Q peuvent être des distributions sur $\mathbb{R}^n$

$\longrightarrow$ c'est précisément la situation qui nous intéresse

Exemples

Non usuel	Formule	Générateur $f(u)$	$f(0) + f^*(0)$
Variation totale	$\frac{1}{2}\int \lvert p(x) - q(x)\rvert dx$	$\frac{1}{2}\lvert u-1\rvert$	1
Kullback-Liebler	$\int p(x) \log \frac{p(x)}{q(x)} dx$	$-\log u$	∞
Hellinger (carré)	$\int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx$	$\left(\sqrt{u}-1\right)^2$	2
Chi-2 Pearson	$\int \frac{(p(x)-q(x))^2}{p(x)} dx$	$(u-1)^2$	∞

Table – Exemples de f-divergences usuelles.

Propriétés

- Unicité : $\forall (P,Q), D_{f_1}(P||Q) = D_{f_2}(P||Q) \Leftrightarrow \exists c \in \mathbb{R}, f_1(u) - f_2(u) = c(u-1)$
- Symétrie : On pose $f^* : u \mapsto uf(1/u)$ la fonction $*$-conjuguée de f. On a $\forall (P,Q), D_f(P||Q) = D_{f^*}(Q||P)$ et $\forall (P,Q), D_{f^*}(P||Q) = D_f(P||Q) \Leftrightarrow \exists c \in \mathbb{R}, f^*(u) - f(u) = c(u-1)$
- Plage de valeurs : $0 = f(1) \leq D_f(P||Q) \leq f(0) + f^*(0)$ avec égalité sur la borne inférieure si $P = Q$ (équivalence si f est strictement convexe en 1) et égalité sur la borne supérieure si $P \perp Q$ (équivalence si $f(0) + f^*(0) < \infty$), ie $\exists \mathcal{A}, Q(\mathcal{A}) > 0, P(\mathcal{A}) = Q(\mathcal{A}^c) = 0$.
- Convexité : $\forall \lambda \in [0,1], \quad D_f(\lambda P_1 + (1-\lambda)P_2 || \lambda Q_1 + (1-\lambda)Q_2) \leq \lambda D_f(P_1||Q_1) + (1-\lambda) D_f(P_2||Q_2)$

La caractérisation de la plage de valeurs est le résultat principal pour l'analyse de dépendance.

Ainsi, si X_i' et Y' sont deux variables aléatoires indépendantes et telles que X_i' et X_i ont la même loi, Y' et Y la même loi, alors en notant P la loi de (X_i', Y') et Q celle de (X_i, Y), pour f strictement convexe en 1 on a $S_i^f = 0$ si et seulement si Y est indépendant de X_i et $S_i^f = f(0) + f^*(0)$ si et seulement si Y est une fonction de X_i seule.

Analyse de sensibilité

Sensibilités basées sur les divergences et copule

- La relation (29) montre que les indices de sensibilité (28) peuvent se reformuler en termes de copule :

$$S_i^f = \int_{[0,1]^2} f\left(\frac{1}{c_{X_i,Y}(u,v)}\right) c_{X_i,Y}(u,v)\, dudv = \int_{[0,1]^2} f^*\left(c_{X_i,Y}(u,v)\right)\, dudv \qquad (30)$$

- On construit un estimateur $\hat{S}_i^f$ de S_i^f en remplaçant $c_{X_i,Y}$ par un estimateur $\hat{c}_{X_i,Y}$ dans (30)
- Nous proposons d'utiliser l'estimateur non-paramétrique $\hat{c}_{X_i,Y}$ basé sur les copules de Bernstein.

Copules de Bernstein, [Sancetta2004]

Définition

Soit $\boldsymbol{m} = (m_1, \ldots, m_d) \in \mathbb{N}^{*d}$ un multi-indice et $\alpha : [0,1]^d \to [0,1]$ une fonction coïncidant avec une copule sur la grille $\mathcal{G}_{\boldsymbol{m}} = \left\{ \frac{0}{m_1}, \ldots, \frac{m_1}{m_1} \right\} \times \ldots \times \left\{ \frac{0}{m_d}, \ldots, \frac{m_d}{m_d} \right\}$. La copule de Bernstein associée à α et $\mathcal{G}_{\boldsymbol{m}}$ est définie par :

$$\forall \boldsymbol{u} \in [0,1]^d, C_{\alpha,\mathcal{G}_{\boldsymbol{m}}}(\boldsymbol{u}) = \sum_{i_1=0}^{m_1} \ldots \sum_{i_d=0}^{m_d} \alpha\left(\frac{i_1}{m_1}, \ldots, \frac{i_d}{m_d}\right) \prod_{j=1}^{d} P_{i_j,m_j}(u_j) \tag{31}$$

où $P_{a,b}$ est le polynôme de Bernstein donné par $\forall x \in [0,1], P_{a,b}(x) = \frac{b!}{a!(b-a)!} x^a (1-x)^{b-a}$.

Dans la suite on se restreint à $m_1 = \ldots = m_d = m$.

Estimation par copule de Bernstein I

Définition

Soit $(\boldsymbol{X}_k)_{k=1,\dots,n}$ un n-échantillon d'un vecteur aléatoire de dimension n et $(\boldsymbol{U}_k)_{k=1,\dots,n}$ l'échantillon des rangs normalisés défini par $U_{i,k} = F_{i,n}(X_{i,k})$ où $F_{i,n}$ est la fonction de répartition empirique de $(X_{i,k})_{k=1,\dots,n}$.
On appelle copule empirique de $(\boldsymbol{X}_k)_{k=1,\dots,n}$ la fonction de répartition de la loi discrète uniforme sur $(\boldsymbol{U}_k)_{k=1,\dots,n}$.

Lemme

Soit C_n la copule empirique associée à l'échantillon $(\boldsymbol{X}_k)_{k=1,\dots,n}$ et $\mathcal{G}_m$ une grille de $[0,1]^d$. La trace de C_n sur $\mathcal{G}_m$ est la trace d'une copule si et seulement si m divise n.

Quitte à supprimer jusqu'à $m-1$ points de l'échantillon, on suppose dans la suite que m divise n.

Définition

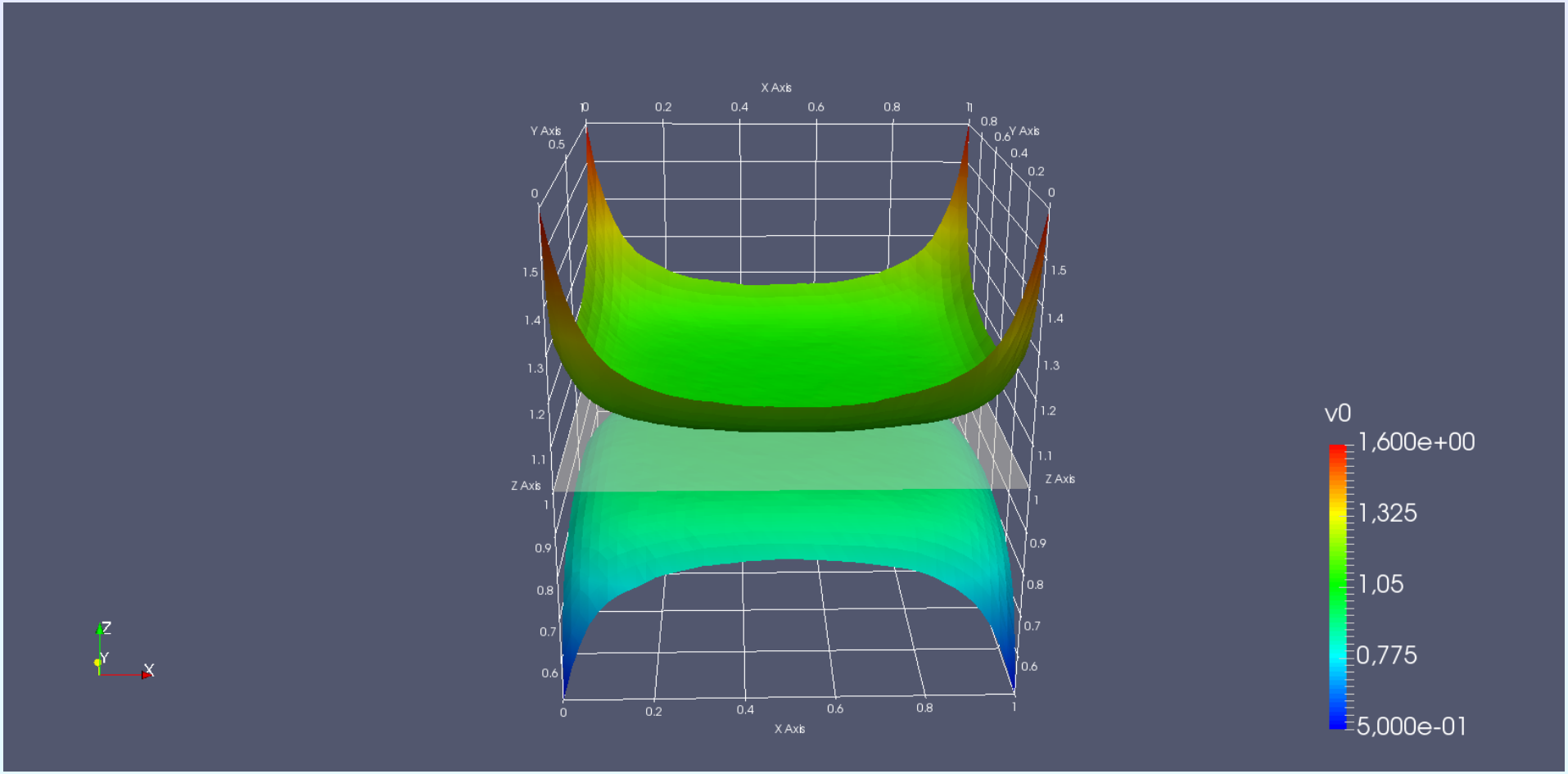
Soit C_n la copule empirique associée à un échantillon de copule absolument continue C de densité c, $\mathcal{G}_m$ une grille et $C_{C_n,\mathcal{G}_m}$ la copule de Bernstein associée à C_n. On appelle estimateur de Bernstein de c la densité $\hat{c}_{n,m}$ de $C_{C_n,\mathcal{G}_m}$.

Propriétés de l'estimateur de Bernstein

On suppose que c est dérivable sur $[0,1]^d$.

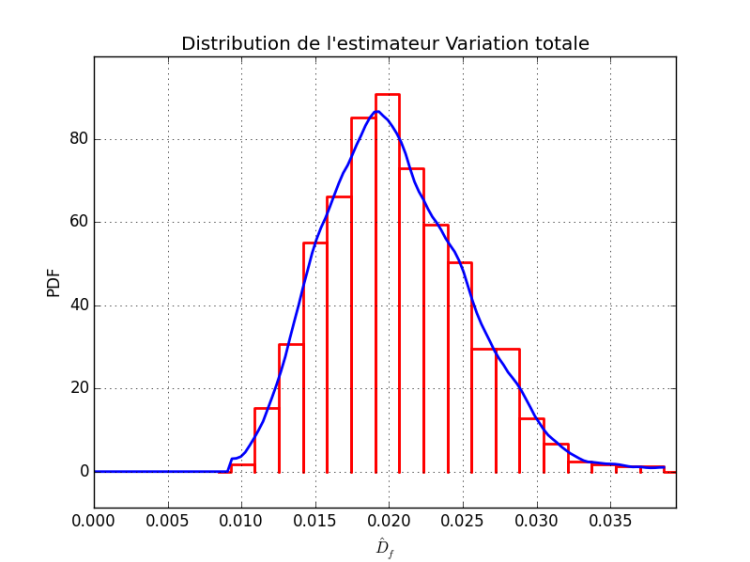
- $\forall \boldsymbol{u} \in [0,1]^d, \mathbb{E}\left[\hat{c}_{n,m}(\boldsymbol{u}) - c(\boldsymbol{u})\right] = \mathcal{O}(m^{-1})$
- $\forall \boldsymbol{u} \in [0,1]^d, \exists j, u_j \in \{0,1\} \Rightarrow \mathbf{Var}\left[\hat{c}_{n,m}(\boldsymbol{u})\right] = \mathcal{O}\left(\frac{m^d}{n}\right)$ sinon $\mathbf{Var}\left[\hat{c}_{n,m}(\boldsymbol{u})\right] = \mathcal{O}\left(\frac{m^{d/2}}{n}\right)$
- Quand $m, n \to \infty$, $\hat{c}_{n,m}(\boldsymbol{u}) \to c(\boldsymbol{u})$ en moyenne quadratique si $\frac{m^{d/2}}{n} \to 0$ pour $\boldsymbol{u} \in]0,1[^d$ et si $\frac{m^d}{n} \to 0$ sinon.
- Le choix optimal de m est $m = \mathcal{O}\left(n^{\frac{2}{d+4}}\right)$ pour $\boldsymbol{u} \in]0,1[^d$, $m = \mathcal{O}\left(n^{\frac{1}{d+2}}\right)$ sinon.
- Si $m > d+2$, $z_{n,m}(\boldsymbol{u}) = \hat{c}_{n,m}(\boldsymbol{u}) - \mathbb{E}\left[\hat{c}_{n,m}\right]$ converge quand $n \to \infty$ vers un processus Gaussien centré de covariance $\mathbb{E}\left[z_{m,n}(\boldsymbol{u}) z'_{m,n}(\boldsymbol{v})\right]$.

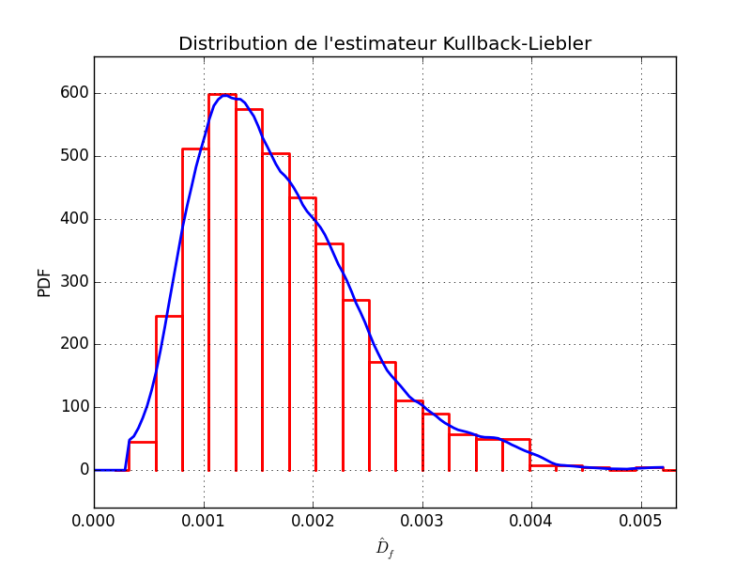
Estimation de la copule indépendante

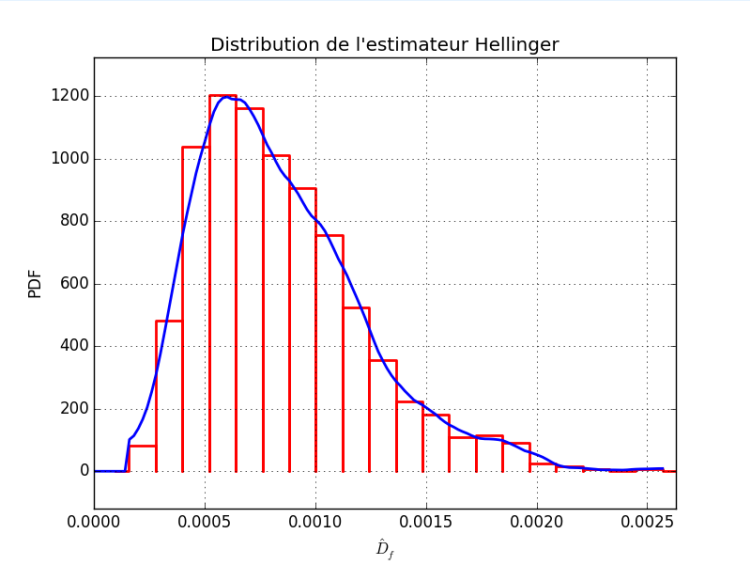


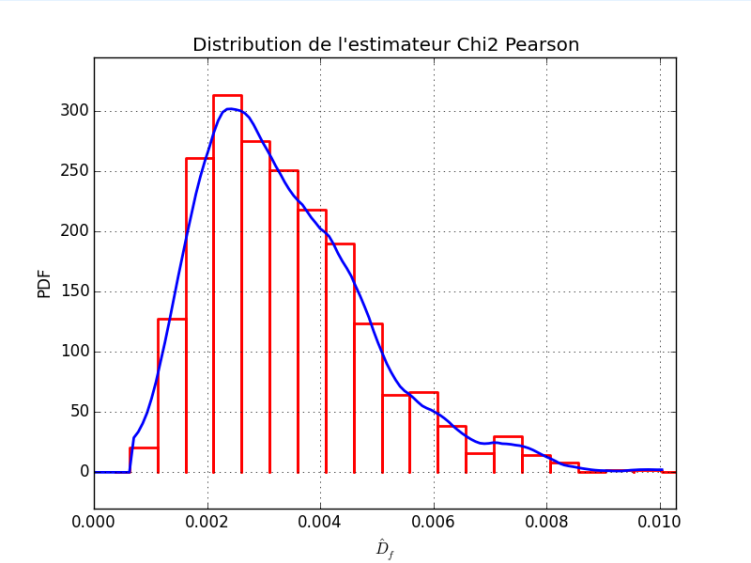
Bande de confiance à 95%, $n = 1000$

Estimation de divergences de Csizar









Estimateur des indices de sensibilité, $n = 1000$, 10^4 répétitions iid.

Analyse de sensibilité

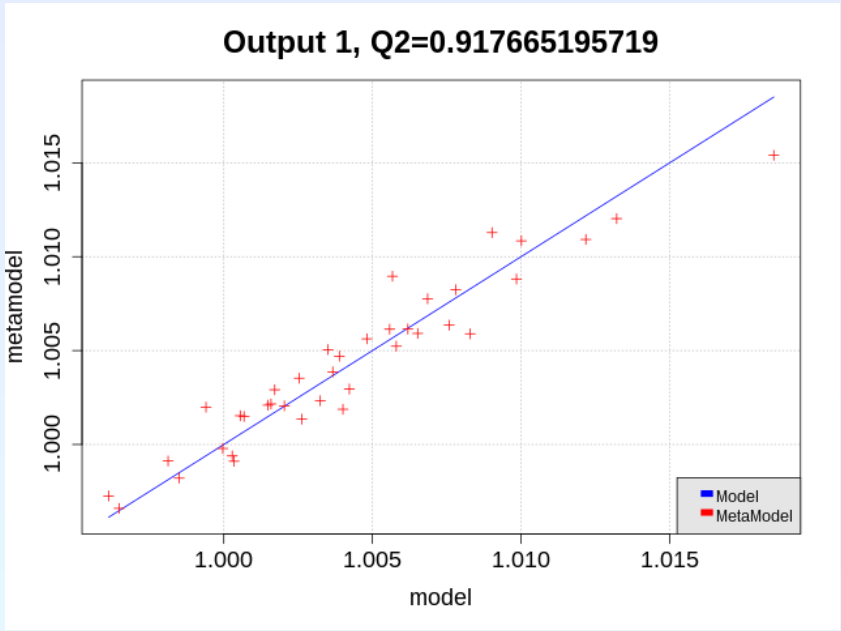
Analyse en aveugle d'une base de données de coefficients aérodynamiques

- On s'intéresse à une fonction boîte noire de $\mathbb{R}^{24}$ dans $\mathbb{R}^{12}$
- Cette fonction est connue via une base de données de taille $n = 377$
- Aucune information n'est connue sur la distribution des entrées constituant la base
- L'objectif est d'identifier pour chaque sortie les contributeurs principaux.
- On présente l'étude de la première sortie

Hiérarchisation par indices de Sobol

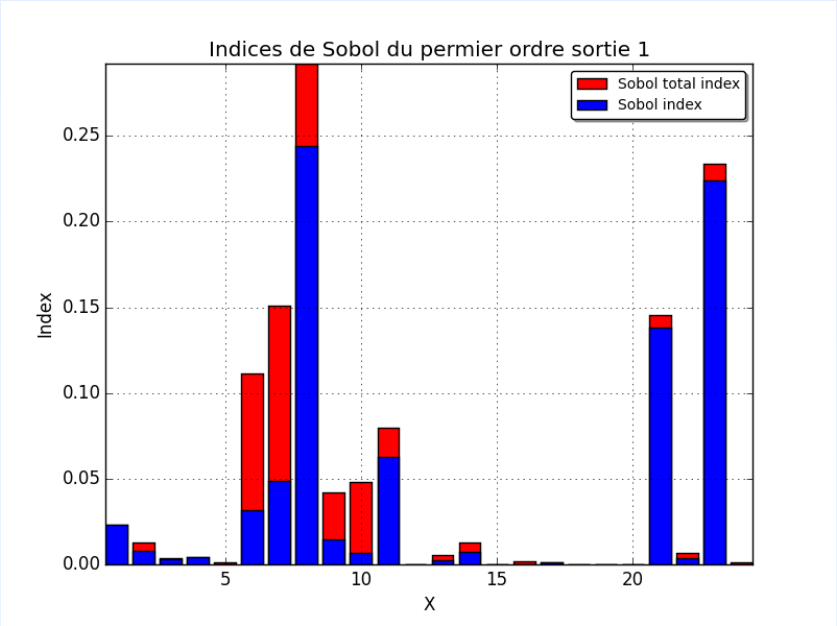
- On teste l'hypothèse d'indépendance des entrées via leur corrélation des rangs : on ne peut pas rejeter l'hypothèse au seuil de 95%
- On estime de manière non-paramétrique les distributions marginales
- On utilise la base des polynômes orthonormés spécifique à ces lois marginales
- On construit un méta-modèle par chaos polynomial pénalisé (Least Angle Regression Stepwise + Corrected Leave One Out) sur 90% de la base, on valide le méta-modèle sur les 10% restant
- On exploite ce méta-modèle pour calculer les indices de Sobol et les indices de Sobol totaux du premier ordre.

Qualité du méta-modèle



Validation du modèle de chaos polynomial

Indices de Sobol



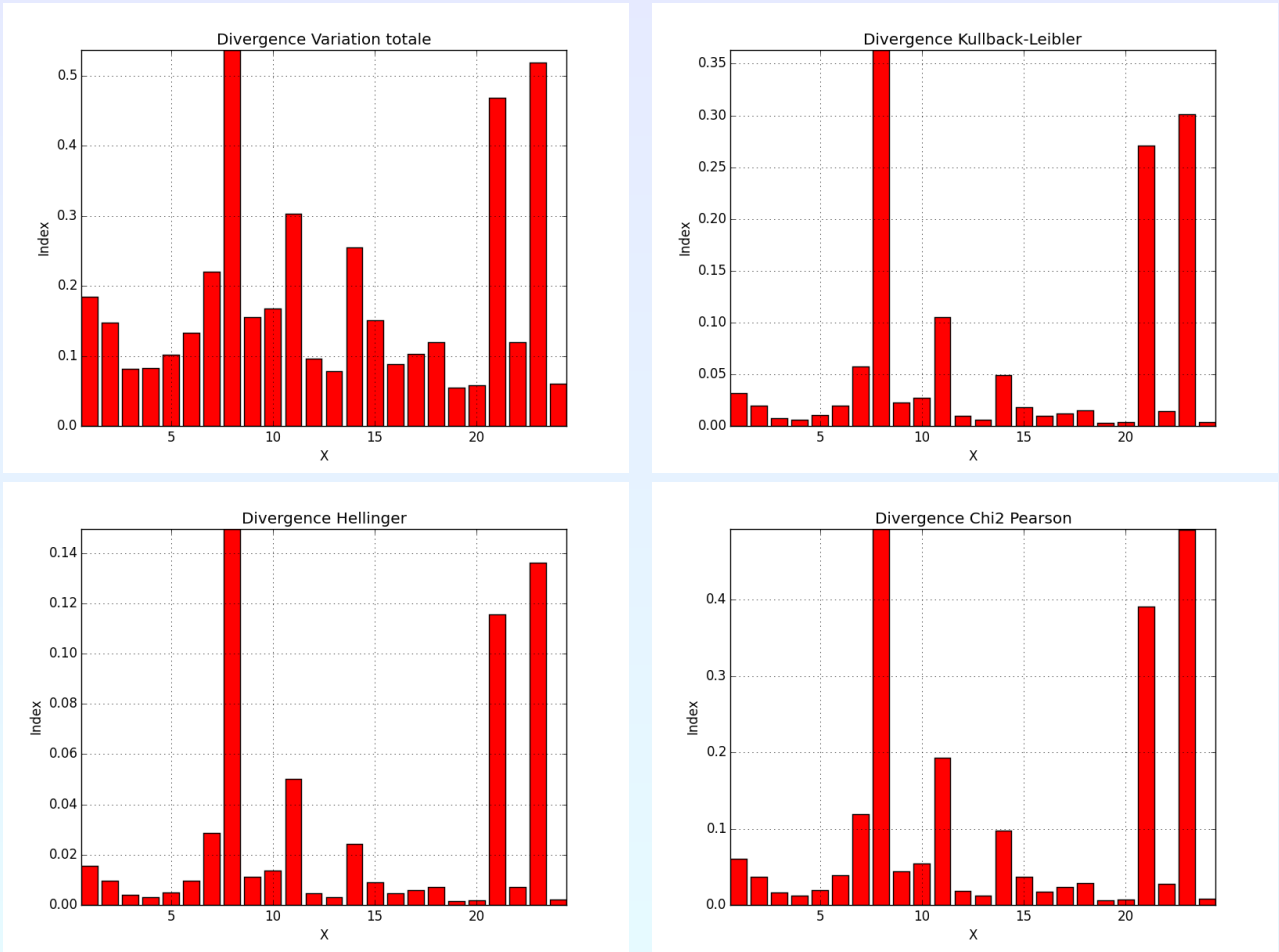
Contributions des entrées à la variance de la sortie 1

La lecture de ce graphique montre qu'il est sans doute important de garder les entrées 6, 7, 8, 11, 21 et 23, et sans doute anodin de supprimer les entrées 3, 4, 5, 12, 13, 15, 16, 17, 18, 19, 20, 22 et 24, soit une division par au moins 2 de la dimension d'entrée.

Hiérarchisation par divergence de Csiszár

- On choisit les types de divergence qu'on souhaite utiliser. Ce choix est dicté par l'application, en l'absence d'expérience sur ces nouveaux indices on explore toutes les divergences présentées dans le tableau 1.
- Pour chaque entrée du modèle, on estime la copule de sa loi jointe avec la première sortie à l'aide de l'estimateur de Bernstein.
- On utilise la densité de la copule estimée pour calculer chacune des quatre divergences via une méthode de quadrature adaptative.
- Pour chaque entrée jugée peu influente on vérifie si la copule estimée est incluse dans la région de confiance ponctuelle à 95% de la copule indépendante.

Divergences estimées I

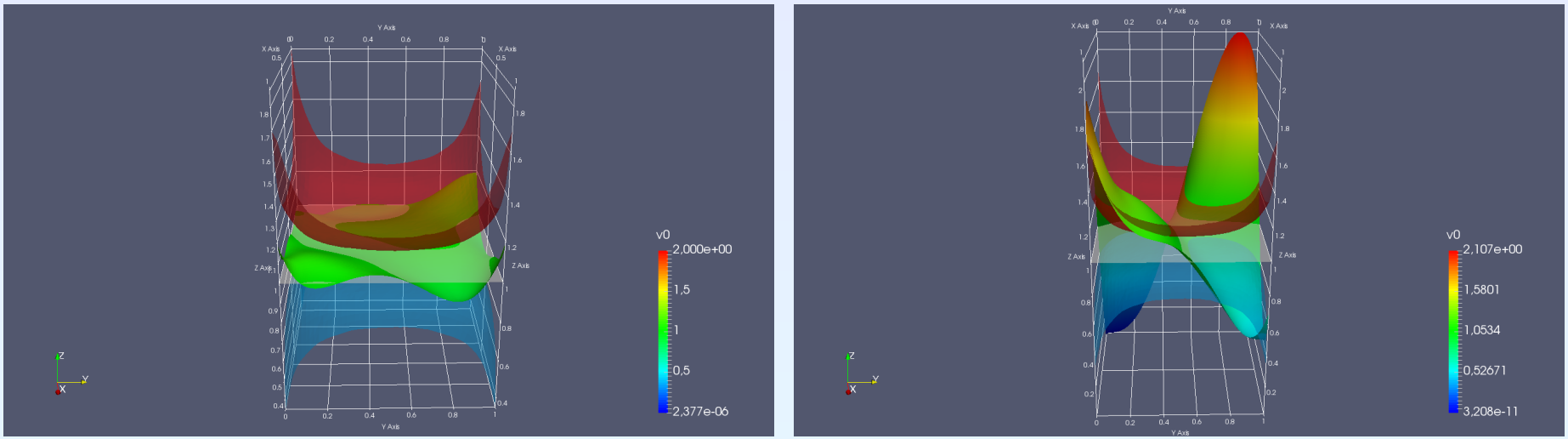


Estimateur des indices de sensibilité Variation totale (haut gauche), Kullback-Leibler (haut droite), Hellinger (bas gauche), Chi2 Pearson (bas droite).

Divergences estimées II

- Mise à part la divergence en variation totale, les autres divergences donnent une hiérarchisation essentiellement conforme à celle obtenue via les indices de Sobol
- La divergence en variation totale, correspondant aux indices introduits par Borgonovo, ont tendance à ne rien considérer comme négligeable : il existe toujours un événement prenant une probabilité très différente selon qu'on considère la sortie comme indépendante de l'entrée considérée ou liée via la copule estimée.
- L'avantage essentiel de ne nécessité aucune hypothèse d'indépendance entre les entrées pour le calcul et l'interprétation
- L'inconvénient principal de ne pas former une partition d'une grandeur globale.
- Il reste à se forger une expérience dans la manipulation de ces nouveaux indices...

Copules estimées



Copule jointe estimée de (X_{19}, Y_1) (gauche) et de (X_8, Y_1) (droite)

Ces graphiques montrent qu'on ne peut pas rejeter l'hypothèse que Y_1 est indépendant de X_{19}, alors que Y est manifestement fortement dépendant de X_8, cette dépendance étant essentiellement une co-monotonie croissante puisque le voisinage de la première diagonale reçoit l'essentiel de la masse probabiliste.

Analyse de sensibilité

Quoi de neuf ?

Sur la définition des indices de sensibilité :

- L'espérance d'une divergence entre la loi de la sortie et la loi conditionnelle de la sortie sachant une entrée est une divergence entre lois bidimentionnelles partageant les mêmes lois marginales
- C'est en fait une divergence entre la copule indépendante et la copule liant la sortie à l'une des entrées
- On peut ramener l'estimation de ces indices à celle d'une densité de copule bivariée

Sur le calcul des indices de sensibilité :

- La définition d'un estimateur basé sur l'estimateur de Bernstein de copules absolument continues
- Son calcul effectif via une quadrature bidimensionnelle
- Sa mise en œuvre effective sur un cas-test industriel

Pistes de recherche :

- Poursuivre l'exploration des liens entre estimation de ces indices et estimation de densité de copule. Pertinence de la norme L_2 ?
- Propriétés asymptotiques ou à horizon fini des estimateurs proposés
- Interprétation des indices : partition d'une grandeur, définition d'une échelle absolue
- ...

Références I


Emanuele Borgonovo and Elmar Plischke.
Sensitivity analysis : A review of recent advances.
European Journal of Operational Research, 248(3) :869–887, 2016.


Yann Caniou.
Global sensitivity analysis for nested and multiscale modelling.
Theses, Université Blaise Pascal - Clermont-Ferrand II, November 2012.


Imen Csiszár.
Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der egodizität von markoffschen ketten.
Publ. Math. Inst. Hungar. Acad. Sci., 8 :85–107, 1963.


Sébastien Da Veiga.
Global Sensitivity Analysis with Dependence Measures.
working paper or preprint, November 2013.


Huibin Liu, Wei Chen, and Agus Sudjianto.
Relative Entropy Based Method for Probabilistic Sensitivity Analysis in Engineering Design.
Journal of Mechanical Design, 128(2) :326–336, 2006.

Références II

Alessio Sancetta and Stephen Satchell.
The bernstein copula and its applications to modeling and approximations of multivariate distributions.
Econometric Theory, 20(03) :535–562, 2004.

I. M. Sobolá.
Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates.
Math. Comput. Simul., 55(1-3) :271–280, February 2001.