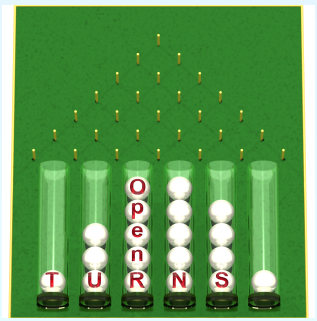
Analyses de Sensibilité et de Corrélation / Facteurs d'importance Implémentation dans Open TURNS

Anne DUTFOY (EDF R&D)
MRI
anne.dutfoy@edf.fr

8 juillet 2010



Sensibilités / Corrélations / Facteurs d'importance dans Open TURNS

Sensibilités / Corrélations / Facteurs d'importance

Contexte

Etudes de traitement des incertitudes

L'objectif des études de traitement des incertitudes est de propager les incertitudes entâchant certaines données d'entrée d'un modèle numérique jusqu'à un grandeur dite d'*intérêt*.

On considère : $Y = f(X_1, \ldots, X_n)$ où :

- g est un **modèle** (code de calcul, expression analytique, ...)
- $(X_1, \ldots, X_n)$ est l'ensemble **des paramètres incertains** que l'on modélise par une loi de probabilité
- Y est la **grandeur d'intérêt** évaluée par le modèle, supposée ici scalaire.

Une fois les incertitudes quantifiées et propagées jusqu'au critère final (tendance centrale, dépasssement de seuil), la dernière étape de la **Méthodologie Globale de Traitement des Incertitudes** est de hiérarchiser les sources d'incertitudes entre elle.

Trois notions

Les **facteurs d'importance** se rattachent à la **décomposition de la variance** et permettent d'expliquer la variabilité de Y en fonction des variabilités des entrées X_i.
Les **corrélations** sont des indicateurs de **relation linéaire** entre Y et les entrées X_i.
Les **sensibilités** sont des indicateurs de **dépendance locale** reposant sur l'évaluation de **gradients** par rapport à des grandeurs caractéristiques des entrées.
Selon la **structure du modèle** f, les indices sont définis différemment.

Sensibilités / Corrélations / Facteurs d'importance

Facteurs d'importance

Décomposition de la variance

De manière très générale, si $Y = f(\underline{X})$, alors on peut décomposer la variance comme suit :

$$Var(Y) = \sum_i \text{Var}\left(\mathbb{E}\left(Y|X_i\right)\right) + \sum_{i \neq j} \text{Var}\left(\mathbb{E}\left(Y|X_i, X_j\right)\right) + \cdots + \underbrace{\text{Var}\left(\mathbb{E}\left(Y|X_1, \ldots, X_n\right)\right)}_{=0} \tag{1}$$

Indices de Sobol

L'**indice de Sobol d'ordre** k donne la variance résiduelle de Y lorsque les entrées $X_{i_1}, \ldots, X_{i_k}$ sont connues (et donc plus aléatoires) :

$$S_{i_1,\ldots,i_k} = \frac{\text{Var}\left(Y|X_{i_1}, \ldots, X_{i_k}\right)}{\text{Var}\left(Y\right)} \tag{2}$$

L'**indice total de Sobol d'ordre** k donne la variance de Y due aux entrées $X_{i_1}, \ldots, X_{i_k}$ lorsque toutes les autres sont connues (et donc plus aléatoires) :

$$S^T_{i_1,\ldots,i_k} = \frac{\sum_I \text{Var}\left(Y|X_I\right)}{\text{Var}\left(Y\right)}, \; I \subset \{1, \ldots, n\} - \{i_1, \ldots, i_k\} \tag{3}$$

Facteurs d'importance

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

Pour calculer (2) et (3), on utilise le résultat général dû à Sobol :
Si f est intégrable sur $[0,1]^n$, elle admet une unique décompositin du type :

$$f(x_1,\dots,x_n) = f_0 + \sum_{i=1}^{i=n} f_i(x_i) + \sum_{1\le i<j\le n} f_{i,j}(x_i,x_j) + \cdots + f_{1,\dots,n}(x_1,\dots,x_n) \quad (4)$$

où $f_0 = cst$ et les fonctions de la décomposition sont orthogonales entre elles par rapport à la mesure de Lebesgue sur $[0,1]^n$:

$$\int_0^1 f_{i_1,\dots,i_s}(x_{i_1},\dots,x_{i_s}) f_{j_1,\dots,j_k}(x_{j1},\dots,x_{j_k}) d\underline{x} = 0 \quad (5)$$

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

Facteurs d'importance

Comment utiliser ce résultat ?

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

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

$\Longrightarrow$ Si on pose

$$\underline{U} = (F_1(X_1), \ldots, F_n(X_n))^t) = \phi^{-1}(\underline{X}) \tag{6}$$

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

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

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

$$\mathbb{E}(Y|\underline{U}) = \mathbb{E}(Y|\underline{X}) \tag{7}$$

En effet : $\mathbb{E}(Y|\underline{U}) = \mathbb{E}(Y|\psi(\underline{X}))$ est le projeté orthogonal au sens L_2 de Y sur l'espace engendré par $\psi(\underline{X})$, ie celui engendré par $\underline{X}$, d'où l'égalité des variables aléatoires (8).
Comme la transformation ϕ (6) agit composante par composante ($U_i \leftrightarrow X_i$) alors on a l'égalité des indicateurs de type :

$$\mathrm{Var}\left(\mathbb{E}\left(Y|U_{i_1}, \ldots, U_{i_k}\right)\right) = \mathrm{Var}\left(\mathbb{E}\left(Y|X_{i_1}, \ldots, X_{i_k}\right)\right) \tag{8}$$

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

Facteurs d'importance

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 f selon la décomposition de Sobol (4), on écrit :

$$Y = f(\underline{X}) = f_0 + \sum_{i=1}^{i=n} f_i(X_i) + \sum_{1 \leq i < j \leq n} f_{i,j}(X_i, X_j) + \cdots + f_{1,\ldots,n}(X_1, \ldots, X_n) \quad (9)$$

La condition d'orthogonalité (5) des $f_{i_1,\ldots,i_k}$ par rapport à la mesure de Lebesgue sur $[0,1]^n$ s'interprète comme un calcul d'expérance si les X_i sont indépendantes.
$\Longrightarrow$ On suppose donc que dans la suite, les X_i sont **indépendantes**.
Conclusion :
Y se décompose sous la forme d'une somme de **variables aléatoires orthogonales** entre elles :

$$\boldsymbol{Y = f(\underline{X}) = Z_0 + \sum_{i=1}^{i=n} Z_i + \sum_{1 \leq i < j \leq n} Z_{i,j} + \cdots + Z_{1,\ldots,n}} \quad (10)$$

où $Z_0 = cst$ et $Z_{i_1,\ldots,i_s} \perp Z_{j_1,\ldots,j_k}$ (ie $\mathbb{E}\left(Z_{i_1,\ldots,i_s}.Z_{j_1,\ldots,j_k}\right) = 0$).

Facteurs d'importance

Calcul des Indices de Sobol

Grâce à la décomposition probabiliste (10), on calcule $\mathbb{E}(Y)$ et $\mathrm{Var}(Y)$ aisément :

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

$$\Longrightarrow \mathrm{Var}(Y) = \sum_{i=1}^{i=n} V_i + \sum_{1 \le i < j \le n} V_{i,j} + \cdots + V_{1,\ldots,n} \qquad (11)$$

où $V_{i_1,\ldots,i_k} = \mathrm{Var}\left(Z_{i_1,\ldots,i_k}\right) = \mathrm{Var}\left(f_{i_1,\ldots,i_k}(X_{i_1},\ldots,X_{i_k})\right)$.

Facteurs d'importance

Calcul des Indices de Sobol

Les indices de Sobol (2) et (3) s'expriment en fonction des $V_{i_1,\dots,i_k}$:

$$S_{i_1,\dots,i_k} = \frac{\mathrm{Var}\left(\mathbb{E}\left(Y|X_{i_1},\dots,X_{i_k}\right)\right)}{\mathrm{Var}\left(Y\right)} = \frac{\sum_{I\subset\{i_1,\dots,i_k\}} V_I}{\mathrm{Var}\left(Y\right)} \tag{12}$$

$$S^T_{i_1,\dots,i_k} = \frac{\sum_I \mathrm{Var}\left(\mathbb{E}\left(Y|X_I\right)\right)}{\mathrm{Var}\left(Y\right)} = \frac{\sum_I \sum_J V_J}{\mathrm{Var}\left(Y\right)} \tag{13}$$

$$\text{où} \quad I \subset \{1,\dots,n\}, I \cap \{i_1,\dots,i_k\} \neq \emptyset, J \subset I \tag{14}$$

Facteurs d'importance

Si le modèle f est affine : SRC

Si $Y = \alpha_0 + \sum_i \alpha_i X_i$, avec les X_i **indépendantes**, alors on définit le **Standard Regression Coefficient (SRC)** :

$$SRC_i = \frac{\alpha_i^2 \mathrm{Var}(X_i)}{\mathrm{Var}(Y)} \tag{15}$$

Donc SRC **est un indice de Sobol** à l'ordre 1 de X_i : $\boldsymbol{SRC(Y/X_i) = S_i(Y/X_i)}$.

Si le modèle f est *monotone par marginale* : SRRC

Si $Y = f(\underline{X})$ avec les X_i **indépendantes** et si f est monotone par rapport à chaque X_i, en posant $\underline{U} = (F_1(X_1), \ldots, F_n(X_n))^t) = \phi^{-1}(\underline{X})$, on a :

$$Z = F_Y(Y) = F_Y \circ f \circ \phi(\underline{U})$$

Si on suppose **de plus** que :

$$Z = \sum_i \alpha_i U_i \tag{16}$$

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

$$SRRC(Y/X_i) = SRC(Z/U_i) = \frac{\alpha_i^2 \mathrm{Var}(U_i)}{\mathrm{Var}(Z)} = S_i(Z/U_i)$$

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

Facteurs d'importance

Indices de Sobol : Moyens de calcul dans Open TURNS

Open TURNS propose d'évaluer les indices de Sobol via :

- de l'**échantillonnage** des entrées / sortie :
  - classe ***CorrelationAnalysis*** : calcul des **indices S_I d'ordre 1, 2 et 3 et des indices totaux d'ordre 1** par la méthode d'échantillonnage de Sobol. La programmation de la méthode dans Open TURNS est parallèle (si l'évaluation de la fonction sur un échantillon a été codée de manière à bénéficier de plusieurs coeurs de calcul, l'algorithme en profite automatiquement : par exemple, en codant un wrapper multithread).
  - classe ***SensitivityAnalysis*** : calcul des **indices S_I d'ordre 1 et des indices totaux d'ordre 1** via le package *Sensitivity version 1.3.1 de R par la mï¿½thode de Saltelli.*
- 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).

SRC et SRRC : Moyens de calcul dans Open TURNS

La classe ***CorrelationAnalysis*** permet le calcul de **SRC et SRRC.**

Facteurs d'importance

Cumul quadratique

$Y = f(\underline{X})$ est approché par son **approximation de Taylor à l'ordre 1 au point moyen** :

$$Y = f(\overline{\underline{X}}) + < \underline{\nabla} f(\overline{\underline{X}}), (\underline{X} - \overline{\underline{X}}) >= f(\overline{\underline{X}}) + \sum_i (X_i - \overline{X}_i) \left. \frac{\partial f}{\partial X_j} \right|_{\overline{\underline{X}}} \tag{17}$$

Sous cette **hypothèse de linéarité du modèle au point moyen** $\overline{\underline{X}}$, on calcule :

$$\mathrm{Var}\,(Y) \quad = \quad {}^t\underline{\nabla} f(\overline{\underline{X}}).\underline{\underline{\mathrm{Cov}}}\,[\underline{X}]\,.\underline{\nabla} f(\overline{\underline{X}}) = \sum_{i,j} \left. \frac{\partial f}{\partial X_i} \right|_{\overline{\underline{X}}} \mathrm{Cov}\,[X_i, X_j]\,.\, \left. \frac{\partial f}{\partial X_i} \right|_{\overline{\underline{X}}} \tag{18}$$

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

$$\boldsymbol{FI(X_i)} = \frac{\left(\sum_j \left. \frac{\partial f}{\partial X_j} \right|_{\overline{\underline{X}}} \mathbf{Cov}\,[\boldsymbol{X_i}, \boldsymbol{X_j}] \right) \left. \frac{\partial f}{\partial X_i} \right|_{\overline{\underline{X}}}}{\mathrm{Var}\,(\boldsymbol{Y})} \tag{19}$$

Si les X_i sont indépendantes entre elles, alors (19) se simplifie en :

$$\boldsymbol{FI(X_i)} = \left(\left. \frac{\partial f}{\partial X_j} \right|_{\overline{\underline{X}}} \right)^2 \frac{\mathrm{Var}\,(\boldsymbol{X_i})}{\mathrm{Var}\,(\boldsymbol{Y})} = SRC(Y/X_i) = S_i(Y/X_i)$$

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

Cumul quadratique : Moyens de calcul dans Open TURNS

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

FORM / SORM

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

$$Y \simeq M = \beta - < \underline{\alpha}, \underline{U} >= \beta - \sum_i \alpha_i U_i \tag{20}$$

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 :

$$\boldsymbol{FI(X_i) = \alpha_i^2} \tag{21}$$

Cette définition est **généralisable au cas des espaces standards sphériques**. Dans cet espace, les variables U_i sont au moins *décorrélées* (indépendantes dans le cas gaussien), et de *marginales centrées réduites*. Donc on montre que, pour le modèle linéarisé (20) :

$$SRC(Y/U_i) = \frac{\alpha_i^2 \mathrm{Var}(U_i)}{\mathrm{Var}(M)} = \alpha_i^2 \tag{22}$$

car $\mathrm{Var}(U_i) = 1$ et $\mathrm{Var}(M) = \sum_i \alpha_i^2 \mathrm{Var}(U_i) = \sum_i \alpha_i^2 = 1$.
Donc les **facteurs d'importance de FORM sont des indices de Sobol** d'ordre 1 pour le modèle linéarisé au point de conception.
Attention ! : les FI sont des $SRC(Y/U_i)$ 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

FORM / SORM : Implémentation dans Open TURNS

Dans Open TURNS, la transformation isoprobabiliste utilisée est :

- celle de Nataf Généralisée lorsque la copule de $\underline{X}$ 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 $\underline{X}$ n'est pas elliptique : l'espace standard est gaussien, et les U_i sont indépendantes.

Pour remédier au problème de la non bijection entre les U_i et les X_i, Open TURNS 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 :

$$\underline{Z}^* = (E^{-1} \circ F_1(X_1^*), \ldots, E^{-1} \circ F_n(X_n^*))^t \tag{23}$$

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

$$FI(X_i) = \left(\frac{Z_i^*}{||\underline{Z}^*||} \right)^2 \tag{24}$$

Il ya concordance des définitions (21) et (24) dans le cas de variables indépendantes ou de copule gaussienne. La version d'OT proposera les 2 définitions.

Facteurs d'importance

FORM : Moyens de calcul dans Open TURNS

Les classes ***FORMResult*** et ***SORMResult*** permettent le calcul des facteurs d'importance *FI* ainsi que leur tracé sous forme de camembert : *getImportanceFactors() et drawImportanceFactors().*
Dans la prochaine version d'Open TURNS, l'utilisateur spécifiera à l'aide d'un flag la définition voulue.

Sensibilités / Corrélations / Facteurs d'importance

Corrélation

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 :

$$\boldsymbol{\rho(Y, X_i) = \frac{\mathrm{cov}\,[Y, X_i]}{\sqrt{\mathrm{Var}\,(X_i)\,\mathrm{Var}\,(Y)}}} \tag{25}$$

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

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

Si le modèle est *monotone* : Spearman

Si on suppose que le modèle est monotone et que **les rangs sont linéaires** (16) , 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** :

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

Corrélation

Si le modèle est linéaire : PCC

Si on suppose que $Y = \sum_i \alpha_i X_i$, on définit le coefficient PCC_i (**Partial Pearson Correlation Coefficient**) quantifie la corrélation directe de Y à X_i une fois supprimée les dépendances linéaires de X_i et de Y par rapport aux autres X_j pour $j \neq i$.

- il évalue, à l'aide de régression linéaire, la moyenne de X_i par rapport aux X_j pour $j \neq i$ (partie de X_i qui n'est pas expliquée par les X_j) : $\tilde{X}_i = \sum_{k \neq i} \tilde{b}_k X_k$
- il évalue, à l'aide de régression linéaire,la moyenne de Y par rapport aux X_j pour $j \neq i$ (partie de Y qui n'est pas expliquée par les X_j) : $\tilde{Y} = \sum_{k \neq i} \tilde{a}_k X_k$
- il calcule le coefficient de corrélation linéaire (de Pearson) entre les résidus issus de Y et ceux issus de X_i : $\rho_{(Y-\tilde{Y}),(X_i-\tilde{x}_i)}$, estimé à partir de l'échantillon aléatoire.

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.

Corrélation

Pearson, Spearman, PCC, PRCC : Moyens de calcul dans Open TURNS

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

Sensibilités / Corrélations / Facteurs d'importance

Sensibilités

FORM / SORM

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

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

Seules les sensibilités par rapport aux paramètres des lois marginales sont disponibles dans OpenTURNS.

Sensibilités

Monte Carlo

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

$$\underline{u}^* = \mathbb{E}\left(\underline{u} | \underline{u} \in \mathcal{D}\right) \tag{26}$$

et en dériver des ***facteurs d'importance*** comme :

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

Par symétrie de la loi sphérique de l'espace standard, $\underline{U}^*$ 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 (21) et (27) coïnccident.

Monte Carlo : Implémentation dans Open TURNS

Prochainement dans Open TURNS.