Étude de sensibilité sur un quantile de sortie

Thomas Browne [1] & Jean-Claude Fort [2]

[1] *EDF R&D Chatou - Université Paris-Descartes ; thomas .ga.browne@gmail.com*
[2] *Université Paris-Descartes ; fort43@gmail.com*

Résumé. Dans un contexte de codes numériques avec des entrées aléatoires, il est fréquent d'avoir recours à une analyse de sensibilité afin de déterminer l'influence respective d'une entrée X du code sur la sortie d'intérêt Y. Le *gosa* (*goal-oriented sensitivity analysis* pour " analyse de sensibilité orientée dans un but ") introduit par N. Rachdi dans sa thèse (2011), propose en amont de l'étude de se focaliser sur une caractéristique de la distribution de la sortie $\theta(Y)$ (moyenne, quantile, probabilité d'échec etc...), qui serait définie par une stratégie propre à l'utilisateur, et de quantifier l'influence de chaque entrée X sur cette dernière. Dans notre cas, $\theta(Y)$ est un quantile de la distribution de Y. L'idée est d'évaluer la variabilité de $\theta(Y \mid X)$ ainsi que son écart par rapport à $\theta(Y)$. Dans cet objectif un indice d'analyse de sensibilité basé sur des fonctions de contraste a été défini (Fort et al., 2013) : il compare un "écart moyen" par un contraste bien choisi entre $\theta(Y \mid X)$ et $\theta(Y)$. Nous proposons dans cet exposé un estimateur statistique pour l'indice dont nous prouvons également la consistance. Une extension de ces indices dans le cas de contrôles de détection de défauts (courbes de probabilités de détection) est présentée.

Mots-clés. Analyse de sensibilité, Quantiles, Estimateurs à noyaux ...

Abstract. In the context of numerical codes, it is often required to use sensistivity analysis in order to assess the influence of each random input X over the output Y. *Gosa* (*goal-oriented sensitivity analysis*), introduced by N. Rachdi in his PhD thesis (2011), states that one must first focus on one probability feature $\theta(Y)$ of Y's distribution (such as its mean, quantile, or a probability of failure etc...), which would be chosen regarding a relevant strategy. The wish is to evaluate the impact of each input over $\theta(Y)$. In our case, $\theta(Y)$ is the α-level quantile of Y, where $\alpha \in]0,1[$, and we study the variability of the conditionnal feature $\theta(Y \mid X)$. To this effect, a sensitivity analysis index based on contrast functions, has been set (Fort et al., 2013). The latter quantifies a distance between $\theta(Y \mid X)$ and $\theta(Y)$. In the presentation we introduce a statistical estimator for this index. Its consistency is also proved. At the end, we propose an extension of the index to the case of defect detection where the output is no longer a scalar value but a random cumulative distribution function : the PoD-curve.

Keywords. Sensitivity analysis, Quantiles, Kernel-based estimators...

1 Introduction

Le souhait est de réaliser une analyse de sensibilité sur une sortie scalaire Y d'un code numérique : nous nous concentrons sur des indices permettant de quantifier l'influence d'une entrée aléatoire sur un quantile de Y. On appelle f la fonction du code et $(X_1, ..., X_d)$ l'ensemble des entrées : $Y = f(X_1, ..., X_d)$. Nous présentons par la suite les indices de sensibilité par contrastes $S^i_{c_\alpha}(Y)$,avec $i \in \{1, ..., d\}$, sur lesquels nous basons toute l'étude, ainsi qu'un estimateur par noyaux. Le souhait à long terme de ce travail est d'étendre ces indices au cas de contrôles de détection de défauts où la sortie d'intérêt n'est plus un scalaire mais une fonction de répartition aléatoire : la courbe de *PoD* (*Probability of detection*).

2 Goal oriented-sensitivity analysis (*gosa*)

N. Rachdi a introduit l'idée d'analyse de sensisbilité orientée dans un but (*gosa* pour *goal-oriented sensitivity analysis*) au cours de sa thèse [1]. Sa théorie suggère que lorsqu'une analyse de sensibilité est requise sur Y, il est important de s'interroger avant tout sur la quantité d'intérêt de la distribution de Y. La principale idée est que, dans la plupart des cas, seulement un petit nombre (sinon un) de propriété de Y est pertinent pour l'étude : l'utilisateur peut s'intéresser principalement à la moyenne de la sortie $\mathbb{E}[Y]$, à un quantile $q^\alpha(Y)$, avec $\alpha \in]0, 1[$, ou à une probabilité d'échec, $P_f = \mathbb{P}(Y < 0)$. Par exemple dans le cas où f simulerait le bénéfice tiré d'une certaine opération, P_f s'imposerait comme une quantité primordiale. Le *gosa* propose de s'interroger sur l'influence de chaque entrée sur P_f. Selon notre approche, nous nous posons la question suivante : si l'une des entrées X_i est fixée à une valeur déterministe x_i, cela a-t-il un impact sur P_f ? On illustre en Figure (1) un test numérique avec trois entrées (X_1, X_2, X_3) où chacune d'entre elles est successivement fixée à vingt valeurs différentes ($X_i = x_i^j$, $j \in \{1, ..., 20\}$). Pour une entrée X_i on calcule $\mathbb{P}(Y < 0 \mid X_i = x_i^j)$ et on observe sa variation selon les différentes valeurs de x_i^j). Comme X_3 est l'entrée dont la variabilité se propage le plus à travers la quantité conditionnée, on peut facilement conclure que X_3 a la plus haute influence sur P_f. De la même manière, on voit aussi que X_2 n'a pas d'influence significative.

3 Analyse de sensibilité par fonction de contraste : cas du quantile

Avant toute chose, rappelons-nous qu'une fonction de contraste simple ψ est une fonction positive à deux entrées réelles y et θ, et c'est une fonction convexe de $(y - \theta)$. On définit son contraste moyen Ψ_Y par rapport à une variable aléatoire Y par :

$$\forall \theta \in \mathbb{R} \quad \Psi_Y(\theta) = \mathbb{E}\left[\psi(Y, \theta)\right].$$

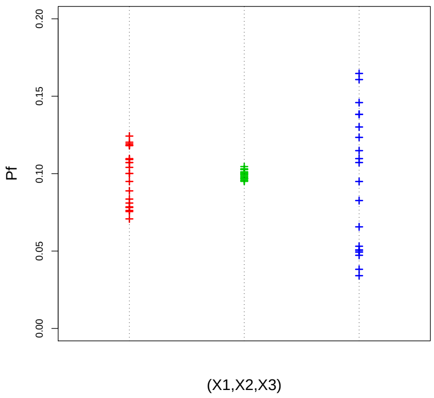

Figure 1: Probabilités d'échec P_f condtionnées à chaque entrée : en rouge, $\mathbb{P}\left(Y \leq 0 \mid X_1 = x_1^j\right)$, en vert $\mathbb{P}\left(Y \leq 0 \mid X_2 = x_2^j\right)$ et en bleu $\mathbb{P}\left(Y \leq 0 \mid X_3 = x_3^j\right)$, à chaque fois pour $j = 1, ..., 20$.

On aboutit alors à la définition d'une propriété θ^* de la distribution de Y (telle que sa moyenne, son quantile ...) avec :

$$\theta^* := \arg\min_{\theta \in \mathbb{R}} \Psi_Y(\theta).$$

Il est évident que la nature de θ^* dépend uniquement du choix de ψ. Remarquons qu'avec $\psi = c_\alpha$, avec $\alpha \in]0, 1[$, tel que :

$$\forall y, \theta \in \mathbb{R} \quad c_\alpha(y, \theta) = (y - \theta)(\mathbf{1}_{y \leq \theta} - \alpha),$$

on obtient $\theta^* = q^\alpha(Y)$. Dans [2] les auteurs définissent les indices suivants, pour $i \in \{1, ..., d\}$:

$$S^i_{c_\alpha}(Y) = \min_{\theta \in \mathbb{R}} \mathbb{E}\left[c_\alpha(Y, \theta)\right] - \mathbb{E}_{X_i}\left[\min_{\theta \in \mathbb{R}} \mathbb{E}\left[c_\alpha(Y, \theta) \mid X_i\right]\right],$$

En utilisant le fait que $q^\alpha(Y) = \arg\min_{\theta \in \mathbb{R}} \mathbb{E}\left[c_\alpha(Y, \theta)\right]$, on a :

$$S^i_{c_\alpha}(Y) = \mathbb{E}\left[c_\alpha(Y, q^\alpha(Y))\right] - \mathbb{E}\left[c_\alpha\left(Y, q^\alpha\left(Y \mid X_i\right)\right)\right],$$

où $q^\alpha\left(Y \mid X_i\right)$ est la quantile conditionnel de Y par X_i. Dans la dernière expression, on peut voir que $S^i_{c_\alpha}(Y)$ quantifie l'influence de fixer l'entrée aléatoire X_i à une valeur déterministe sur $q^\alpha(Y)$. $S^i_{c_\alpha}(Y)$ est positif et on le normalise (en divisant par $\mathbb{E}\left[c_\alpha(Y, q^\alpha(Y))\right]$) de manière à en tirer l'interprétation suivante : si $S^i_{c_\alpha}(Y) \simeq 0$ on considère que X_i n'a pas d'influence sur $q^\alpha(Y)$ et si $S^i_{c_\alpha}(Y) \simeq 1$, X_i détermine à lui seul la valeur de $q^\alpha(Y)$. À partir d'un N-échantillon $\left(Y^1, ..., Y^N\right)$, où $N \in \mathbb{N}$ et pour $j \in \{1, ..., N\}$, $Y^j = f\left(X_1^j, ..., X_d^j\right)$, on propose l'estimateur à noyau suivant pour $S^i_{c_\alpha}(Y)$:

$$\widehat{S^i_{c_\alpha}(Y)} = \frac{1}{N} \sum_{j=1}^{N} c_\alpha\left(Y^j, \widehat{q^\alpha}(Y)\right) - \frac{1}{hN^2} \sum_{k=1}^{N} \min_{\theta \in \mathbb{R}} \frac{1}{f_i(X_i^k)} \left[\sum_{j=1}^{N} c_\alpha\left(Y^j, \theta\right) K\left(\frac{X_i^k - X_i^j}{h}\right)\right]$$

où K est un noyau, $h > 0$ la fenêtre (positive), f_i la densité de X_i et $\widehat{q^\alpha}(Y)$ l'estimateur empirique classique de $q^\alpha(Y)$.

4 Propriétés de l'estimateur

Sous les conditions $h(N) \underset{N \longrightarrow +\infty}{\longrightarrow} 0$ and $h(N) \times N \underset{N \longrightarrow +\infty}{\longrightarrow} +\infty$, on prouve la convergence presque-sûre de l'estimateur :

$$\lim_{N \longrightarrow +\infty} \widehat{S^i_{c_\alpha}(Y)} = S^i_{c_\alpha}(Y) \quad \text{a.s.}$$

ainsi que le résultat suivant :

$$\sqrt{h(N) \times N} \left(\widehat{S^i_{c_\alpha}(Y)} - S^i_{c_\alpha}(Y) - \mathcal{B}(N) \right) \underset{N \longrightarrow +\infty}{\overset{\mathcal{L}}{\longrightarrow}} \mathcal{N}(0, \mathcal{V})$$

avec $\lim_{N \longrightarrow +\infty} \mathcal{B}(h) = 0$ et $\mathcal{V} > 0$, qui nous permet de construire un intervalle de confiance sur $S^i_{c_\alpha}(Y)$. Des tests numériques ont été réalisés afin de justifier la pertinence de ces indices. De plus ils confirment leurs propriétés asymptotiques.

Bibliographie

[1] N. Rachdi (2011), Statistical Learning and Computer Experiments, Institut de Mathématiques de Toulouse, France.
[2] J-C. Fort, T. Klein et N. Rachdi (2013), New sensitivity analysis subordinated to a contrast, *Communication in Statistics : Theory and Methods.*