Intervalles de confiance pour les indices de Sobol

Taieb Touati [1]

[1] *UPMC-LSTA, 4 Place Jussieu, 75005 Paris*
[1] *taiebtouati.tt@gmail.com*

Résumé. Nous présentons une extension de la méthode de Martinez au cas non Gaussien. En effet, ce cas de figure peut altérer la précision de l'intervalle de confiance fourni par l'approximation de Fisher, les deux points suivant seront étudiés: (Cette présentation reprend le travail présenté au SAMO2016: Touati (2016), la fonction R correspondante à la méthode: SobolTouati() est disponible dans le Package Sensitivity)

1. Intervalles de confiances asymptotiques. Dans ce cas, nous donnons un intervalle de confiance pour les indices de Sobol dans un cas général basé sur un résultat bien connu concernant le coefficient de corrélation.

2. Intervalles de confiances non-asymptotiques. Dans ce cas, nous comparons plusieurs méthodes pour améliorer la méthode de Martinez tout en maintenant l'approche d'approximation d'une part et avec une approche de Bootstrap d'autre part.

Mots-clés. Analyse de sensibilité, Indices de Sobol, Coefficient de corrélation, Bootstrap.

Abstract. In this communication, the extension of the Martinez method to non Gaussian distribution is studied. Indeed, non Gaussianity can distort the Fisher's confidence interval, and the outcome can be quite misleading. The two following points will be discussed: (This presentation takes up the work presented at the SAMO2016: Touati (2016), the R implementation of the method: SobolTouati() is available in the Package Sensitivity)

1. Asymptotic confidence intervals. In this case, we give an asymptotic confidence interval for Sobol' indices in a general case based on a well known result concerning the correlation coefficient.

2. Non asymptotic confidence intervals. In this case, we compare several methods to improve the Martinez method while keeping the approximation approach on the one hand and with a Bootstrapping approach on the other hand.

Keywords. Sensitivity Analysis, Sobol indices, Correlation coefficient, Bootstrapping.

1 Etat de l'art et contexte

Lorsque l'on étudie les interactions entre les variables, aller au-delà des régressions est crucial pour plus de précision et de robustesse.

En effet, l'étude de l'interdépendance entre les variances de chacune des composantes du modèle ajoute un plus large éventail de techniques d'interprétation et de prévision.
La littérature définit ce segment d'analyse comme l'analyse de sensibilité basée sur la variance considérée comme l'un des modèles informatiques les plus utilisés dans des études d'ingénierie(Ferretti et al., 2016).
l'indice de Sobol: Sobol (1993); Iooss et al., (2015) est défini par:

$$S_i = \frac{V_i}{V} = \frac{\mathrm{Var}[E(f(X)|X_i)]}{\mathrm{Var}[f(X)]} \text{ and } S_i^{\mathrm{tot}} = \frac{V_i^{\mathrm{tot}}}{V} = 1 - \frac{V_{-i}}{V} = 1 - \frac{\mathrm{Var}[E(f(X)|X_{-i})]}{\mathrm{Var}[f(X)]}\ , \quad (1)$$

$f(X)$ est un modéle numérique, $X = (X_1, \ldots, X_d) \in R^d$ sont les entrées du modèle (Variables aléatoires indépendantes), $i = 1, \ldots, d$, et X_{-i} est le vecteur d'entrée en enlevant X_i. S_i,L'indice de Sobol de premier ordre, ne prenant en compte que le seul effet de X_i, et S_i^{tot}, est L'indice de Sobol total, tenant compte de tous les effets de X_i en incluant ses interractions avec d'autres entrées.
Pour u un sous-espace de $\{1, 2, ..., d\}$ nous considérons la partition: $X = X_u \cup X_{\bar{u}}$, où $\bar{u}$ est le complémentaire de u dans $\{1, 2, ..., d\}$.

Comme dans Iooss et al. (2016) ,on choisit d'étudier les estimateurs qui fournissent $(\hat{S}_i, \hat{S}_i^{\mathrm{tot}})$, estimateurs de $(S_i, S_i^{\mathrm{tot}})$, en utilisant deux copies d'entrées indépendantes $\mathbf{A}$ and $\mathbf{B}$, matrices avec n lignes (Taille de l'échantillon) et d colonnes.

Nous donnons une importance particulière pour l'estimateur de Martinez qui établit les indices de Sobol comme un coefficient de corrélation.
Les propriétés mathématiques du coefficient de corrélation empirique conduisent à explorer plus en profondeur les propriétés de cet estimateur et à construire ensuite les intervalles de confiance asymptotiques correspondant.
Janon et al. (2014) ont étudié la normalité asymptotiques et l'efficacité de deux estimateurs des indices de Sobol, une comparaison sera faite en terme de longueur de l'intervalle et de probabilité de couverture.

2 Intervalles de confiance asymptotiques

Estimateur de Martinez (Martinez, 2011): En remarquant que

$$S_i = \rho\left(f(\mathbf{B}), f(\mathbf{A}_{B(i)})\right) \text{ et } S_i^{\mathrm{tot}} = 1 - \rho\left(f(\mathbf{A}), f(\mathbf{A}_{B(i)})\right) \quad (2)$$

$A_{B(u)} = A_u \cup B_{\bar{u}}$, u est sous-ensemble $\{1, 2, ..., d\}$ (Pour l'estimateur Martinez u=i, $i = (1, \ldots, d)$.

ρ est le coefficient de corrélation linéaire,les indices de Sobol peuvent être estimés par la bien formule empirique de ρ bien conditionnée (*i.e.* en utilisant le produit de différences).

Pour l'estimateur de Martinez les intervalles de confiances sont approximés en utilisant la trasnformée de Fisher appliquée aux coefficients de corrélation empiriques $\hat{S}_i$ et $\hat{S}_i^{\text{tot}}$ déduits de Eq 2.
Ceci est vrai seulement si la variable de sortie est Gaussienne. Les intervalles de confiance à 95% obtenues par la méthode de Martinez sont décrits dans Iooss et al. (2016).

En se basant sur le fait que les indices Sobol sont interprétés comme des coefficients de corrélation, nous fournissons deux résultats asymptotiques. Cette méthodologie est analogue à la démonstration élaborée par Lehman (1999).
Nous explicitons une formule pour la variance asymptotique comme fonction polynomiale du coefficient de corrélation.
On suppose que (Y,Z) est un couple de variables aléatoires réelles de carrés intégrables. R_n est le coefficient de corrélation empirique du couple (Y,Z) et ρ est le coefficient de corrélation théorique. C_n,$\sigma_n(Y)$ et $\sigma_n(Z)$ sont respectivement la covariance empirique et la variance empirique. La premier théorème concerne la normalité asymptotique du triplet $\{C_n, \sigma_n(Y), \sigma_n(Z)\}$. Si K est la matrice de covariance formée après l'application du théorème de la limite centrale au triplet $\{C_n, \sigma_n(Y), \sigma_n(Z)\}$.
La normalité asymptotique R_n donne:

$$\sqrt{n}(R_n - \rho) \to \mathcal{N}(0, \tau^2) \tag{3}$$

τ^2 est une fonction polynomiale de ρ, cette expression facilite l'implémentation de la méthode.
$\tau^2 = P(\rho)$ avec:

$$P(x) = Ax^2 + Bx + C \tag{4}$$

A,B and C dépendent des coefficients de K.

3 Intervalles de confiance non-asymptotiques

Bishara et al. (2016) ont récemment proposé plusieurs alternatives à la méthode de Fisher pour calculer les intervalles de confiance lorsque les données ne sont pas distribuées selon la loi normale. Les méthodes sont classées en deux groupes principaux: Transformation des données et Bootstrapping. Concerant les méthodes de transformation de données, la meilleure performance a été réalisée par "Speraman rank-order" and "rank inverse normal transformation". Parmi les procédés de Bootstrap Efron et al. (1994), bénéficiant du mérite de conserver l'échelle initiale des données brutes, une méthode a une probabilité de couverture particulèrement adéquate avec des intervalles précis.
Les travaux de Beasly et al. (2007) ont servi de fondement à cette approche dans laquelle le temps de calcul a été réduit, rendant les calculs plus faciles pour les échantillons de plus grande taille.

Les comparaisons entre les deux approches sont faites en termes de probabilité de couverture et de longueur d'intervalle de confiance. Des études numériques illustreront tous ces effets pour les différentes méthodes, démontrant qu'avec la méthode asymptotique, nous avons une probabilité de couverture plus précise par rapport à l'approche de Martinez. Les résultats suggèrent que pour l'échantillon non gaussien, il est préférable d'éviter l'intervalle de confiance de Fisher en faveur d'alternatives plus robustes (non asymptotiques ou asymptotiques).

Bibliographie

[1]A J.Bishara and J B.Hittner (2016), Confidence intervals for correlations when data are not normal. *Behavior Research Methods, in press.*

[2] WH.Beasley, L.DeShea, LE.Toothaker, JL.Mendoza, DE.Bard, and JL.Rodgers (2007), Bootstrapping to test for nonzero population correlation coefficients using univariate sampling. *Psychological methods American Psychological Association.*, 12:414.

[3] B.Efron and RJ.Tibshirani (1994), An introduction to the bootstrap.*CRC press*

[4] B.Iooss and P.Lemaître (2015), A review on global sensitivity analysis methods. *Uncertainty Management in Simulation-Optimization of Complex Systems.* 101–122, Springer.

[5] B. Iooss, M.Baudin, K.Boumhaout, T.Delage, J.M Martinez (2016), Numerical stability of Sobol indices estimation formula. *Proceedings of the SAMO 2016 Conference.*, La Réunion, France.

[6] Janon, Alexandre and Klein, Thierry and Lagnoux, Agnes and Nodet, Maëlle and Prieur, Clémentine (2014), Asymptotic normality and efficiency of two sobol index estimators. *ESAIM: Probability and Statistics.*

[7]I. Sobol (1993), Sensitivity estimates for non linear mathematical models. Mathematical Modelling and Computational Experiments, 1:407-414.

[8]E.L Lehman(1999), Elements of large-sample theory. Springer Science & Business Media.

[9] T. Touati (2016), Confidence intervals for Sobol indices.Proceedings of the SAMO 2016 Conference, La Réunion, France.