Confidence intervals for Sobol' indices

Taieb,Touati
Pierre and Marie Curie University, France

When studying the interactions between variables, going beyond regressions is crucial for more precision and accuracy. In fact, studying the interdependence between the variances of each of the model's components adds a wider array of interpretation and forecasting techniques. The literature defines this analysis segment as variance based sensitivity analysis which is regarded as one of the most frequently used computer models in engineering studies (Ferretti et al., 2016). The model's output variance that is caused by a specific model input or a combination of more than one input (Sobol, 1993; Iooss et al., 2015).

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

where $f(X)$ is the computer model, $X = (X_1, \ldots, X_d) \in \mathbb{R}^d$ are the model inputs (independent random variables), $i = 1, \ldots, d$, and X_{-i} is the input vector except X_i. S_i, the first-order Sobol' index, only includes the sole effect of X_i, while S_i^{tot}, the total Sobol' index, takes into account all the effects of X_i including its interaction effects with other inputs. For u a subset of $\{1, 2, ..., d\}$ we consider the partition: $X = X_u \cup X_{\bar{u}}$, where $\bar{u}$ is the complement of u in $\{1, 2, ..., d\}$.

As in Iooss et al. (2016) ,we chose to study estimators which provide $(\hat{S}_i, \hat{S}_i^{\mathrm{tot}})$, estimates of $(S_i, S_i^{\mathrm{tot}})$, by using two independent input designs $\mathbf{A}$ and $\mathbf{B}$, matrices with n rows (sample size) and d columns. the following set We focus especially on the Martinez estimator that sets Sobol indices as correlation coefficient. The mathematical properties of the empirical correlation coefficient lead to explore more thoroughly the properties of this estimator and build thereafter confidence intervals. Martinez estimator (Martinez, 2011): By noticing that

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

where $A_{B(u)} = A_u \cup B_{\bar{u}}$, u a subset of $\{1, 2, ..., d\}$ (for Martinez estimator u=i, $i = 1, \ldots, d$).
ρ is the linear correlation coefficient, the Sobol' indices can be estimated using the well-conditioned empirical formula of ρ (*i.e.* using the product of differences).

For the Martinez estimator, asymptotic confidence intervals are approximated by using Fisher's transformation applied to the sample correlation coefficients $\hat{S}_i$ and $\hat{S}_i^{\mathrm{tot}}$ from Eq 2. It is only valid under Gaussian hypothesis of the output variable distribution. The classical 95% confidence intervals obtained by the Martinez method are described in Iooss et al. (2016).

Based on the fact that the Sobol indices are interpreted as correlation coefficients, we give two asymptotic results which will be applied for Sobol indices. This methodology is analogue to the demonstration given by Lehman (1999). We provide a formula for the asymptotic variance as a polynomial function of the correlation coefficient.
We assume that (Y,Z) is a squared integrable couple of random variables. R_n is the empirical correlation coefficient of (Y,Z) and ρ the theoretical correlation coeffcient. C_n,$\sigma_n(Y)$ and $\sigma_n(Z)$ mean respectively the empirical covariance and the empirical variances. The first theorem concerns the asymptotic normality of the triplet $\{C_n, \sigma_n(Y), \sigma_n(Z)\}$. If there K is the covariance matrix forming after applying the central limit theorm to the triplet $\{C_n, \sigma_n(Y), \sigma_n(Z)\}$. the asymptotic normality of R_n gives:

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

τ^2 is a polynomial function of ρ, this can facilitate the implementation of the method. $\tau^2 = P(\rho)$ where:

$$P(x) = Ax^2 + Bx + C \qquad (4)$$

A,B and C depends on the coefficients of K.

Remark

Bishara et al. (2016) gives recently several alternatives to Fisher's method to compute confidence intervals when data are not normal. The methods are classified in two main groups: Transforming data and Bootstrapping. For the transforming data methods the best performance was performed by the well kown Speraman rank-order and the rank inverse normal transformation. Among the bootstrapping methods Efron et al. (1994), which have the merit of conserving the original scale of raw data, an observed imposed bootstrap had an adequate coverage probability with precise intervals comparing to other Bootstrap methods.
The work of Beasly et al. (2007) served as a foundation for this method in which computing time has been reduced making computations easier for larger samples.

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:

1. Asymptotic confidence intervals. In this case, through the methodology described in Remark 2 we give an asymptotic confidence interval for Sobol' indices in a general case.
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. We base this study on the methodology described in remark 3. Comparisons are made in terms of coverage probability and confidence interval length.

Numerical studies will illustrate all these effects for the different methods, demonstrating that with the asymptotic method we have more accurate coverage probability comparing to the Martinez approach. The results suggest that sample non Gaussianity can justify avoidance of the Fisher's interval in favor of more robust alternative (Non-asymptotic or asymptotic).

References:

F. Ferretti, A. Saltelli and S. Tarantola (2016), Trends in sensitivity analysis practice in the last decade, *Science of the Total Environment*, in press.

JM. Martinez (2011), Analyse de sensibilité globale par décomposition de la variance, *Presentation in "Journée des GdR Ondes & Mascot Num"*, 13 janvier 2011, Institut Henri Poincaré.

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

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

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

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

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

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.

B. Iooss, M.Baudin, K.Boumhaout, T.Delage, J.M Martinez (2016), Numerical stability of Sobolâ indices estimation formula. *Submitted to SAMO 2016 Conference.*, La Réunion, France.

[Taieb Touati; UPMC, 4 Place Jussieu, 75005 Paris, France]
[`taieb.touati@etu.upmc.fr` – `https://cran.r-project.org/web/packages/sensitivity/index.html`]