Numerical stability of Sobol' indices estimation formula

Michael, Baudin
EDF Lab Chatou, France

Khalid, Boumhaout
EDF Lab Chatou, France

Thibault, Delage
EDF Lab Chatou, France

Bertrand, Iooss
EDF Lab Chatou, France

Jean-Marc, Martinez
CEA Saclay, France

Variance-based sensitivity analysis has become a common practice when using computer models in engineering studies (Ferretti et al., 2016). The so-called Sobol' indices express the share of the model output variance that is due to a given model input or input combination and write for instance (Sobol, 1993; Saltelli, 2002)

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

where $G(X)$ 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 a direct estimation (without additional modeling), several sampling-based formulas have been proposed in the literature (see Prieur and Tarantola, 2017, for a recent review), but have shown some instabilities from a numerical point of view, delivering different behavior in different cases.

We focus on the 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:

- Sobol-Saltelli estimator (Sobol, 1993; Saltelli, 2002):

$$\hat{V}_i = \frac{1}{n-1}\sum_{k=1}^{n} G(\mathbf{B}^{(k)})G(\mathbf{A}_{B(i)}^{(k)}) - \hat{G_1}^2 \; ; \; \hat{V}_{-i} = \frac{1}{n-1}\sum_{k=1}^{n} G(\mathbf{A}^{(k)})G(\mathbf{A}_{B(i)}^{(k)}) - \hat{G_0}^2 \,, \quad (2)$$

  where $\mathbf{A}_{B(i)}$ is a re-sampled matrix, where all columns come from $\mathbf{A}$ except column i which comes from $\mathbf{B}$, and where the two square means and the variance are estimated by

$$\hat{G_0}^2 = \left[\frac{1}{n}\sum_{k=1}^{n} G(\mathbf{A}^{(k)})\right]^2 \,, \; \hat{G_1}^2 = \frac{1}{n}\sum_{k=1}^{n} G(\mathbf{A}^{(k)})G(\mathbf{B}^{(k)}) \,, \; \hat{V} = \frac{1}{n-1}\sum_{k=1}^{n} G(\mathbf{A}^{(k)})^2 - \frac{n}{n-1}\hat{G_0}^2 \,. \quad (3)$$

- Mauntz estimator (Mauntz, 2002):

$$\hat{V}_i = \frac{1}{n-1}\sum_{k=1}^{n} G(\mathbf{B}^{(k)})\left(G(\mathbf{A}_{B(i)}^{(k)}) - G(\mathbf{A}^{(k)})\right) \; ; \; \hat{V}_i^{\mathrm{tot}} = \frac{1}{n-1}\sum_{k=1}^{n} G(\mathbf{A}^{(k)})\left(G(\mathbf{A}^{(k)}) - G(\mathbf{A}_{B(i)}^{(k)})\right) \,. \quad (4)$$

- Jansen estimator (Jansen, 1999):

$$\hat{V}_i = \hat{V} - \frac{1}{2n-1}\sum_{k=1}^{n}\left(G(\mathbf{B}^{(k)}) - G(\mathbf{A}_{B(i)}^{(k)})\right)^2 \; ; \; \hat{V}_i^{\mathrm{tot}} = \frac{1}{2n-1}\sum_{k=1}^{n}\left(G(\mathbf{A}^{(k)}) - G(\mathbf{A}_{B(i)}^{(k)})\right)^2 \,. \quad (5)$$

- Martinez estimator (Martinez, 2011): By noticing that

$$S_i = \rho\left(G(\mathbf{B}), G(\mathbf{A}_{B(i)})\right) \text{ and } S_i^{\text{tot}} = 1 - \rho\left(G(\mathbf{A}), G(\mathbf{A}_{B(i)})\right) \tag{6}$$

  where 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).

Remark 1: The denominator of the indices, the model variance V, can be estimated from several ways. We restrict our study to the one of Eq. (3).

Remark 2: The same $n(d+2)$ evaluations are needed for applying the four estimators (2), (4), (5) and (6). A direct comparison between them, using a bootstrap technique to obtain confidence intervals, is possible in practice if $\mathbf{A}$ and $\mathbf{B}$ are i.i.d samples.

Remark 3: For the Martinez estimator, asymptotic confidence intervals are approximated via a Fisher's transformation applied to the sample correlation coefficients $\hat{S}_i$ and $\hat{S}_i^{\text{tot}}$ from Eq. (6). For the classical 95% confidence level, we have:

$$\texttt{Prob}(S_i \in [\tanh(\frac{1}{2}\ln\frac{1+\hat{S}_i}{1-\hat{S}_i} - \frac{1.96}{\sqrt{n-3}}), \tanh(\frac{1}{2}\ln\frac{1+\hat{S}_i}{1-\hat{S}_i} + \frac{1.96}{\sqrt{n-3}})]) \simeq 0.95\ , \tag{7}$$

$$\texttt{Prob}(S_i^{\text{tot}} \in [1 - \tanh(\frac{1}{2}\ln\frac{2-\hat{S}_i^{\text{tot}}}{\hat{S}_i^{\text{tot}}} + \frac{1.96}{\sqrt{n-3}}), 1 - \tanh(\frac{1}{2}\ln\frac{2-\hat{S}_i^{\text{tot}}}{\hat{S}_i^{\text{tot}}} - \frac{1.96}{\sqrt{n-3}})] \simeq 0.95\ . \tag{8}$$

It is only valid under Gaussian hypothesis of the output variable distribution. Current works aim at extending this result to non-Gaussian distribution (Touati, 2016).

In this communication, two pathological issues of the estimators' behavior are studied:

1. Non-centered output. In this case, we show that the Sobol-Saltelli and Mauntz estimators are subject to a non-negligible bias, while the other estimators are insensitive to this effect.
2. Small sensitivity indices. In this case, the numerical precision obtained for the Sobol' indices depend on the conditioning of each estimator formula. Indeed, when the terms are close to zero, differences between products (as in the Sobol-Saltelli estimator) are more sensitive than products of differences.

Numerical studies will illustrate all these effects for the different estimators, demonstrating that the Martinez estimator is particularly robust.

References:

F. Ferretti, A. Saltelli and S. Tarantola (2016), Trends in sensitivity analysis practice in the last decade, *Science of the Total Environment*, 568:666-670.

M.J.W. Jansen (1999), Analysis of variance designs for model outputs, *Computer Physics Communication* 117:35-43.

J-M. 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, Paris, France.

W. Mauntz (2002). *Global sensitivity analysis of general nonlinear systems.* Master's thesis, Imperial College.

C. Prieur and S. Tarantola (2017). Variance-based sensitivity analysis: Theory and estimation algorithms. In: *Springer Handbook on UQ*, R. Ghanem, D. Higdon and H. Owhadi (Eds).

A. Saltelli (2002), Making best use of model evaluations to compute sensitivity indices, *Computer Physics Communication*, 145:580-297.

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

T. Touati (2016), Confidence intervals for Sobol' indices. *Proceedings of the SAMO 2016 Conference*, Réunion Island, France.

[Bertrand Iooss; EDF R&D, 6 Quai Watier, 78401 Chatou, France]
[`bertrand.iooss@edf.fr` – `http://www.gdr-mascotnum.fr/doku.php?id=iooss1`]