A minimum variance unbiased (generalized) estimator of total sensitivity indices: an illustration to a flood risk model

MATIEYENDOU LAMBONI[a,b]
[a] University of Guyane (UG), Department of Science and Technology (DFRST), 97346 Cayenne, French Guiana
[b] 228-UMR Espace-Dev, 97323 Cayenne Cedex, French Guiana

I. Objective Variance-based sensitivity analysis [1-2] and multivariate sensitivity analysis [3-5] aim at apportioning the variability of the model output(s) into input factors and their interactions. Sobol's total index, which accounts for the effects of interactions, is often used for selecting the most influential parameters. In this paper, we propose a generalized and optimal estimator of the variance of the total effect (non-normalized total sensitivity index- TSI). The generalized and optimal estimator of the non-normalized TSI makes use of p-fold sets of input values to obtain the TSI estimates. When $p = 2$, we obtain the Jansen's estimator. An illustration to a flood model shows that we can improve the TSI estimations using p larger than 2.

II. Methods Let $Y = f(\mathbf{X})$ be a model output with $\mathbf{X} = (X_1, \ldots, X_d)$, d independent input factors (A1). Under assumption $\mathbb{E}\left(f^2(\mathbf{X})\right) < +\infty$ (A2), we have the Hoeffding decomposition:

$$f(\mathbf{X}) \;=\; \sum_{u \subseteq \{1,2,\ldots d\}} f_u(X_u)\,, \tag{1}$$

where $f_\emptyset = \mathbb{E}\left[f(\mathbf{X})\right]$, $f_j(X_j) = \mathbb{E}\left[f(\mathbf{X})|X_j\right] - f_\emptyset$, and $\mathbb{E}\left[f_u(X_u)\right] = 0$.
It is shown in [2,6] that the non-normalized TSI of a set of inputs $\mathbf{X}_u = (X_j, j \in u)$, is also defined as follows:

$$D_u^{tot} = \mathbb{E}\left(f(\mathbf{X}) - \mathbb{E}\left[f(\mathbf{X})|\mathbf{X}_{\sim u}\right]\right)^2 \,. \tag{2}$$

Definition 1 Let us consider independent samples $\mathbf{X}_u^{(1)}, \ldots, \mathbf{X}_u^{(p)}$ from the measure $\mu(\mathbf{X}_u)$, $\mathbf{X}_u^{(1)'}, \ldots, \mathbf{X}_u^{(p)'}$ from $\mu(\mathbf{X}_u)$ and $\mathbf{X}_{\sim u} = (X_j, j \in \{1, 2, \ldots, d\}\backslash u)$ from $\mu(\mathbf{X}_{\sim u})$. We define a kernel function as:

$$K\left(\mathbf{X}_u^{(1)}, \ldots, \mathbf{X}_u^{(p)}, \mathbf{X}_{\sim u}\right) = \frac{p-1}{p^2} \sum_{l=1}^{p} \left(\sum_{j=1}^{p} c_j^{(l)} [f(\mathbf{X}_u^{(l)}, \mathbf{X}_{\sim u}) - f(\mathbf{X}_u^{(j)}, \mathbf{X}_{\sim u})] \right)^2 \,, \tag{3}$$

with $c_j^{(l)} = \frac{1}{p-1}$ if $j \neq l$ and 0 otherwise (A3).
If we define $\sigma_{l,1}^2 = \mathbb{C}ov\left[K\left(\mathbf{X}_u^{(1)}, \ldots, \mathbf{X}_u^{(l)}, \mathbf{X}_u^{(l+1)}, \ldots, \mathbf{X}_u^{(p)}, \mathbf{X}_{\sim u}\right), K\left(\mathbf{X}_u^{(1)}, \ldots, \mathbf{X}_u^{(l)}, \mathbf{X}_u^{(l+1)'}, \ldots, \mathbf{X}_u^{(p)'}, \mathbf{X}_{\sim u}\right)\right]$, then it satisfies $\sigma_{l,1}^2 = \mathbb{V}\left(\mathbb{E}\left[K\left(\mathbf{X}_u^{(1)}, \ldots, \mathbf{X}_u^{(p)}, \mathbf{X}_{\sim u}\right) |\mathbf{X}_u^{(1)}, \ldots, \mathbf{X}_u^{(l)}, \mathbf{X}_{\sim u}\right]\right)$ ([7]).

Theorem 1 Let $Y = f(\mathbf{X})$ and consider independent samples $\left(\mathbf{X}_{i,u}^{(1)}, \mathbf{X}_{i,\sim u}\right), \ldots, \left(\mathbf{X}_{i,u}^{(p)}, \mathbf{X}_{i,\sim u}\right)$ from $\mu(\mathbf{X})$ with $i = 1, 2, \ldots, m$. Under assumptions A1, A3 ($\forall\, j, l = 1, 2, \ldots, p$, $c_j^{(l)} = \frac{1}{p-1}$ if $j \neq l$ and 0 otherwise), A4 ($\mathbb{E}\left[f^4(\mathbf{X})\right] < +\infty$), and A5 ($2 \leq p$), we have:

i) the optimal, unbiased estimator of D_u^{tot} for a given p is:

$$\widehat{D}_u^{tot} = \frac{p-1}{mp^2} \sum_{i=1}^{m} \sum_{l=1}^{p} \left(\sum_{j=1}^{p} c_j^{(l)} [f(\mathbf{X}_{i,u}^{(l)}, \mathbf{X}_{i,\sim u}) - f(\mathbf{X}_{i,u}^{(j)}, \mathbf{X}_{i,\sim u})] \right)^2 \,; \tag{4}$$

ii) some properties of $\widehat{D}_u^{tot}$ are:

$$m\mathbb{E}\left(\widehat{D}_u^{tot} - D_u^{tot}\right)^2 = \sigma_{p,1}^2, \qquad \sqrt{m}\left(\widehat{D}_u^{tot} - D_u^{tot}\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma_{p,1}^2\right) \,. \tag{5}$$

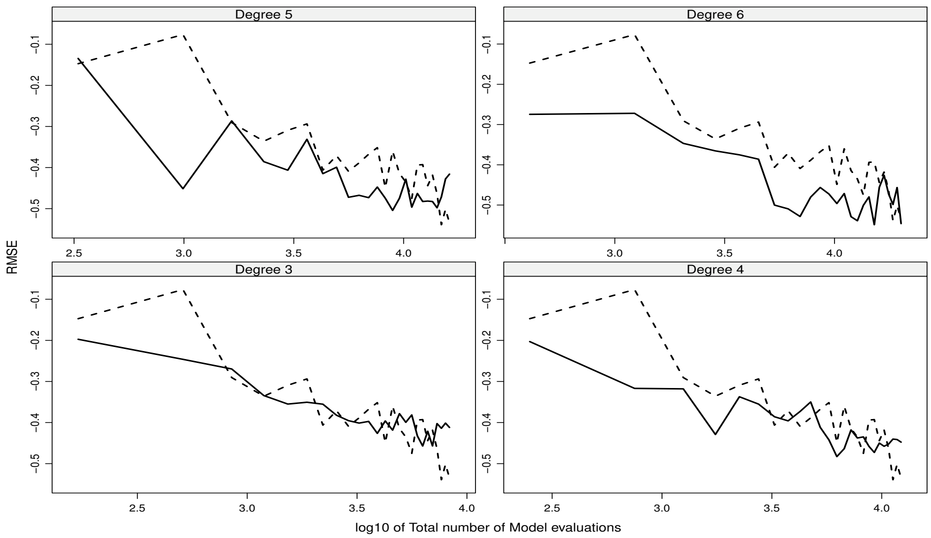

Figure 1: Log-RMSEs against the total number of model runs (in $\log_{10}$) for four values of the degree p = 3, 4, 5, 6. For each degree, we show the corresponding RMSE (solid line) and the RMSE for Jansens estimator (dashed line).

Proof The kernel $K(\cdot)$ is symmetric with respect to its first argument $(\mathbf{X}_u^{(1)}, \ldots, \mathbf{X}_u^{(p)})$ and we have $\mathbb{E}\left[K(\mathbf{X}_u^{(1)}, \ldots, \mathbf{X}_u^{(p)}, \mathbf{X}_{\sim u})\right] = D_u^{tot}$. The ponits i) and ii) are obtained using the properties of U-statistics. (see [8] for more details). $\square$

III. Results and Conclusions To illustrate our approach, we consider a flood model that simulates the height of a river compared to the height of a dyke [2,9]. The model includes 8 input factors. We compared the TSI estimates for four different values of p = 3, 4, 5, 6 to those for Jansen's estimator ($p = 2$), using Sobol's design. Figure 1 shows the average of the root mean square errors (RMSEs) of the 8 inputs against the total number of model runs for each degree p = 3, 4, 5, 6 compared to $p = 2$. It comes out that the degree $p = 6$ provides the best estimations.

References:

- [1] Saltelli A, Ratto M, Andres T, Campolongo F, Cariboni J, Gatelli D, Salsana M, Tarantola S (2008) Global sensitivity analysis - The primer. Wiley.

-[2] Higdon, D., Ghanem, R. and Owhadi H. (2016) Handbook of Uncertainty Quantication. Springer.

-[3] Lamboni M, Makowski D, Lehuger S, Gabrielle B, Monod H (2009) Multivariate global sensitivity analysis for dynamic crop models. Fields Crop Reasearch 113: 312 - 320.

-[4] Lamboni M, Monod H, Makowski D (2011) Multivariate sensitivity analysis to measure global contribution of input factors in dynamic models. Reliability Engineering and System Safety 96: 450 - 459.

-[5] Gamboa F, Janon A, Klein T, Lagnoux A (2014) Sensitivity indices for multivariate outputs. Comptes Rendus de lAcadmie des Sciences p In press.

-[6] Lamboni M (2013) New way of estimating total sensitivity indices. In: Proceedings of the 7th International Conference on Sensitivity Analysis of Model Output (SAMO 2013), Nice, France.

-[7] Ferguson TS (1996) A Course in Large Sample Theory. Chapman-Hall, New York.

-[8] Lamboni M (2016) Global sensitivity analysis: a generalized, unbiased and optimal estimator of total-effect variance (submitted).

-[9] Iooss B Revue sur l'analyse de sensibilité globale de modèles numériques, Journal de la Société Française de Statistique 152 (2011) 1 - 23.

[M. Lamboni; UG, DFRST, Cayenne-French Guiana, matieyendou.lamboni@gmail.com]